Bayes Linear Bayes Network Models for Medical Diagnosis and Prognosis

Wael Abdulateef Jasim Al-Taie

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics & Statistics & Physics
Newcastle University
Newcastle upon Tyne
United Kingdom

March 2020

I dedicate this thesis to Almighty Allah, my mum and dad, wife and lovely children Dalal, Mustafa and Yousif.

Acknowledgements

First of all, I thank Almighty Allah for helping and supporting me to finish my thesis. I would like to thank my supervisor Dr. Malcolm Farrow for his kindness and compassion. I really appreciate his help and I have never seen a person who is more patient than him in my entire life. He keeps on motivating and guiding me especially during the frustrating moments of my study.

I would like to thank my parents, my father Mr. Abdulateef Jasim, and lovely mother Mrs. Maysoon Thanoon, for helping me with everything during my research journey and your constant prayers and supplications for my success in life. Truly, I am nothing without them. Many thanks and appreciation to my sweetheart and lovely wife Mrs. Iman Tareq for her kindness and assistance during my PhD. I would like to thank my siblings, Mr. Mohammed Abdulateef, Mr. Omar Abdulateef, Mrs. Asmma Abdulateef and Alyaa Abdulateef. Besides, big thanks to all my relatives and cousins.

Special thanks to my lecturers in Newcastle University, and all the staff of School of mathematics, Statistics and Physics, such as Dr. Colin Gillespie, Professor Robin Henderson, Professor Richard Boys, Dr. Kevin Wilson, Dr. Daniel Henderson, Dr. Andrew Golightly, Dr. Jian Shi, Dr. David Walshaw, Dr. Peter Avery, Dr. Ged Cowburn, Mr. John Nicholson, Dr. Michael Beaty and Dr. George Stagg.

To those whose names that are not mentioned here but have significantly helped me during my study, I do apologise for such shortwriting. Undoubtedly, your contributions will always be fondly remembered.

I would also like to thank my friends in Newcastle University, especially Muhammad Irfan Bin Abdul Jalal, Muhammad Safwan Bin Ibrahim, Juliana Iworikumo Consul, Stephen Johnson, Maryam Kashia Garba, Amit Seta, Cetin Can Evirgen, Yameng Ji, Thomas Bland and Dimitrios Chiotis.

I would like to acknowledge the financial support that I have received from the Ministry of Higher Education and Scientific Research in my great country Iraq.

Abstract

In this thesis, we develop the application of Bayes linear kinematics and Bayes linear Bayes graphical models to problems in medical diagnosis and prognosis. In medical diagnosis or prognosis, we might use information from a number of covariates to make inferences about the underlying condition, prediction about survival or simply a prognostic index. The covariates may be of different types, such as binary, ordinal, continuous, interval censored and so on. The covariates and the variable of interest may be related in various ways. We may wish to be able to make inferences when only a subset of the covariates is observed so relationships between covariates must be modelled. In the standard Bayesian framework, such a case might suggest the use of Markov chain Monte Carlo (MCMC) methods to integrate over the distribution of the missing covariate values but this may be impractical in routine use. We propose an alternative method, using Bayes linear kinematics within a Bayes linear Bayes model in which relationships between the variables are specified through a Bayes linear structure rather than a fully specified joint probability distribution. This is much less computationally demanding, easily allows the use of subsets of covariates and does not require convergence of a MCMC sampler. In earlier work on Bayes linear Bayes models, a conjugate marginal prior has been associated with each covariate. We relax this requirement and allow non-conjugate marginal priors by using one-dimensional numerical integration. We compare this approach with one using conjugate marginal priors and with a Bayesian analysis using MCMC and a fully specified joint prior distribution. We illustrate our methods with an application to prognosis for patients with non-Hodgkin's lymphoma in which we treat the linear predictor of the lifetime distribution as a latent variable and use its expectation, given whatever covariates are available, as a prognostic index.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Real world data often involve multiple variables and need complex models to reach realistic conclusions. As we encounter widely applicable models, we often need advanced computational methods to fit them. Therefore, methods such as Markov Chain Monte Carlo (MCMC), which allow sampling from the posterior distribution when there is no analytical solution are often used. Bayes linear Bayes models and Bayes linear kinematics (Goldstein and Shaw, 2004) offer an alternative approach which is computationally much simpler.

This thesis addresses the methodology of using Bayes linear Bayes network models in the context of medical diagnosis and prognosis problems. The main aim of this thesis is to construct a Bayes linear Bayes prognostic network. This can be done by relating T to a latent prognostic index.

A Bayes linear analysis (Goldstein and Wooff, 2007) differs from a full Bayesian analysis in that only first and second order moments are specified in the prior. Posterior (termed *adjusted*) moments are then calculated when data are observed. The introduction of Bayes linear kinematics and Bayes linear Bayes models (Goldstein and Shaw, 2004) extends Bayes linear methods to allow the incorporation of observations of types which are not readily accommodated in a straightforward Bayes linear analysis. For example, beliefs about certain unknown quantities might be updated by full conditional Bayesian inference when observations are made on conditionally Poisson or binomial variables and then information can be propagated between these unknowns, or to other unknowns, via

a Bayes linear belief structure. This approach avoids the need for computationally intensive methods such as Markov chain Monte Carlo which are often required in standard Bayesian analyses.

In routine clinical use, in diagnosis or prognosis, the use of methods such as MCMC is not ideal. The methods are computationally demanding and require attention to issues such as convergence. We aim in this thesis to investigate a method which does not have these drawbacks and which can be used even when only a subset of covariates is available.

This proposed method is based on the new idea of using the non-conjugate prior update to construct Bayes linear kinematic prognostic index values. In this way, we construct a Bayes linear kinematic network.

1.2 Bayesian network models and why they are important

A Bayesian network (BN) is a representation of the joint probability distribution of a number of variables which makes use of conditional independence relationships among the variables. We can represent a Bayesian network as a directed acyclic graph (DAG).

Bayesian network models can be useful by combining expert knowledge with the theory of probabilities. There are many reasons why these models are useful and important. Firstly, they are graphical models, so we can represent the relationships between the nodes or vertices clearly, intuitively and in an attractive way. These relationships can often be represented as cause and effect, but this is not always the case. Secondly, these models also can represent more complex problems in a simple graph with dependence relationships. Thirdly, because of the rapid development of computer languages and softwares, we can learn from "big data" and even construct very large and complex networks.

In this thesis, we investigate developing Bayesian methods for selecting, fitting and using models with appropriate conditional independence structures, i.e. graphical models, in the context of medical diagnosis and prognosis problems. So we fit some survival models such as a Weibull distribution to a data set on patients with non-Hodgkin lymphoma, with missing data values for some covariates. That leads us to follow the advice of Farrow (2003) to elicit the structure of the covariance matrix. In some cases, especially when the covariates are a sequence of measurements taken over time, it might be appropriate to use a generalised autoregression model (Pourahmadi, 1999; Daniels and Pourahmadi, 2002).

Bayesian networks may be constructed using expert judgement. However, we may wish to construct a network by inference from historical data.

In order to make inference in Bayesian networks, we need first to learn about the structure of the network. Kułaga (2006) considered that, when he had a small number of variables such as 5, he can manage all the potential models and then calculate the posterior probability for them and choose the best one. However, he mentioned that when the number of variables increases, the number of models will increase at an exponential rate. Therefore, the solution for this problem is to use Markov chain Monte Carlo (MCMC) methods. Also he explained the idea of a Markov blanket for the BN which is defined as a set of nodes that separates a target node from the rest of the nodes in the network which includes its parents, its children and other nodes sharing a child. He then defined this object for a dynamic structure.

Husmeier et al. (2005) gave some insight about how we can learn a BN from complete and incomplete data. They used a Metropolis Hastings algorithm to construct a BN in computational molecular biology and bioinformatics, such as sequence alignment, molecular evolution and genetic networks. Scutari and Denis (2014) gave many examples of BN in the real world. One example used data for medical diagnosis to predict the human body composition which forms the body weight: *bone, fat* and *lean*.

Efficient algorithms are available for information propagation "inference" within certain classes of BN, where the conditional distributions are all (finite) categorical or all Gaussian. Inference using networks with other conditional distributions can be more difficult. Furthermore, the problem of using data to inform the *construction* of a BN (*"network learning"*), particularly the structure of the network, remains challenging. Heckerman and Chickering (1995) use a score metric to describe learning Bayesian network from gathering knowledge and statistical data.

1.3 Bayesian networks in medical diagnosis and prognosis

An important application since the early days of Bayesian networks has been in medical diagnosis. Diagnosis can be viewed as a decision problem and Bayesian networks can assist physicians in making the right decisions, diagnosing the disease early and choosing the most appropriate treatment and thereby improving the outcomes for patients in terms

of health and, in some cases, survival. Therefore, Bayesian networks are powerful tools for helping physicians to make important decisions that lead to the correct treatment with low risk for patients. Similarly, among patients with a particular disease, the prognosis may vary according to various risk factors. A Bayesian network (BN) can be used to improve the efficient use of information in making a prognosis and informing decisions on treatment.

In survival analysis, Langseth (1998) constructed a Bayesian network for survival times using a proportional hazard model. The results showed that his network is useful for qualitative observations. See also (Kjaerulff and Madsen, 2013).

Verdurmen (2003) proposed a model to predict whether clients are likely defaulters at any time during the loan time. He demonstrated a Bayesian network with an exponential survival model. He compared his method to a proportional hazards model and showed that his model can represent much more complex functions than the semi-parametric hazards model.

Jiang et al. (2014) developed a new Bayesian network with high dimensional data to predict patient survival. They developed a new algorithm for Bayesian networks that was used to predict the survival of a patient separately each year. Also, their results showed that their method was better than a proportional hazard model for several reasons such as that their algorithm can deal with data with high dimensions. Kraisangka and Druzdzel (2014) used a BN to interpret a proportional hazards model. Then they compared the accuracy of their BN for the proportional hazards model with Kaplan-Meier estimates and with a BN learned from data. The results showed that constructing a BN from a proportional hazards model is more accurate than the other methods, even if they have a small number of data recorded.

Bayesian decision networks can combine probabilistic models under uncertainty and utilities to help the users make decisions that maximise the expected utility. See Korb and Nicholson (2004). Bayesian decision making requires specification of two elements. One is "beliefs": a probability distribution over the possible outcomes, or, at least, sufficient judgments about the uncertain outcomes to be able to evaluate the necessary expectations. The other is a utility function over the possible outcomes. Gosling (2014) briefly mentions elicitation of utility functions from patients and other people, but is largely concerned with elicitation of the probabilistic beliefs of experts rather than utilities. However, in a decision-making context, the utility element can not be ignored and, in practice, this requires the use of some structure which, as far as possible, does not impose assumptions

but makes the results usable and interpretable.

Therefore, it is important to consider beliefs and utilities together. Often, if the outcome involves different *attributes*, then the utility function in this case will be a *multi-attribute* utility function. For instance, years of life, cost in time and inconvenience to the patient, *e*tc. As a result, this requires us to consider using the *joint* distribution of outcomes, hence necessitating the elicitation of a belief structure involving the dependencies. According to Gosling et al. (2013), the prior elicitation of such dependencies can be difficult due to several reasons such as differences in the experimental methods that are used to measure the outcomes and quantities of interest might be on different scales (e.g, between body mass index Kg/m^2 and blood pressure measurements $mmHg$). However, analysis of the decision problem and the associated utility function can show that decisions can be sensitive to beliefs about dependencies.

The methodology for imprecision in multi-attribute utility functions developed by Farrow and Goldstein (2010) leads to an overall utility function involving a linear combination of various marginal utilities and various products of marginal utilities. Therefore, evaluation of expected utility requires the evaluation of expectations of these quantities. The expectation of a product requires consideration of dependence between the stochastic quantities involved. Farrow and Goldstein (2010) did not explicitly consider this but an extension of the methodology to allow this and, furthermore, to deal with imprecise specification of these expectations seems to be within reach.

1.4 Bayes linear and Bayes linear Bayes methods

Farrow and Goldstein (2006) were motivated in their decision analytic work by problems in the design of experiments using a Bayes linear approach to statistical inference. In the Bayes linear approach, probability distributions are not fully specified but only certain moments are required. See Farrow and Goldstein (1993); Goldstein and Wooff (2007). In recent years an extension of Bayes linear methods, using Bayes linear kinematics and Bayes linear Bayes graphical models, suggested by Goldstein and Shaw (2004), has allowed the combination of Bayes linear structures describing the dependencies between quantities with explicit use of observable quantities with non-Gaussian distributions. The original idea in Goldstein and Shaw (2004) has been developed and applied in a number of papers, including Wilson and Farrow (2010); Gosling et al. (2013) and Wilson et al. (2013). We consider that Bayes linear analysis gives a good approximation to full-Bayes analysis while

in Bayes linear analysis we do not need to specify the prior in a probabilistic way, but we need to specify only the first and the second moments.

1.5 Project aims

The main aims of the project are

- Develop Bayesian methods for selecting, fitting and using models with appropriate conditional independence structures, i.e. graphical models, in the context of medical diagnosis and prognosis problems. In addition, we are looking for improvements to some existing methods.
- Investigate methods for a wider class of conditional distributions, e.g. a survival distribution.
- Build probabilistic models for diagnosis and prognosis with various Bayesian network learning algorithms to help the physicians and others to make decisions about their patients more accurately and efficiently.
- Construct a Bayes linear kinematic network which can be used when we observe only some of the covariates. Develop methods for incorporating different kinds of covariates in such a network.
- Make comparisons between different methods to construct Bayes linear kinematic prognostic networks.

1.6 Outline of the thesis

The remainder of the thesis has the following structure. In Chapter 2 we describe the data which will be used for illustration in the thesis. These include data on survival for patients with Non-Hodgkin lymphoma and leukemia. We give an overview of the explanatory variables in these data sets.

Chapter 3 reviews the basic ideas of Bayesian inference and Markov Chain Monte Carlo (MCMC) methods which are used to compute posterior distributions. We also give an introduction to generalised linear models and particularly the logistic regression model

with the logit link function. We discuss variable selection methods and in particular Bayesian variable selection methods. We illustrate the missing data problem and data augmentation and give an example involving lung transplant data. We use the logistic regression model to fit the data where the response variable represents whether the lung is used for transplant or not.

In Chapter 4, we introduce probabilistic graphical models, concentrating on directed acyclic graphs, and give some algorithms that are well known in the field of graphical models. We give some important definitions and concepts related to Bayesian networks. We explain some important and useful methods to construct Bayesian networks and learn from the data. We introduce a method called the *arc deletion method* which depends on finding the most optimal network using MCMC methods.

Chapter 5 deals with survival analysis with some important aspects and definitions related to our work. An important feature in survival analysis is censoring. We give an explanation for the most familiar survival models such as proportional hazard models, piecewise constant hazard models and accelerated failure time models. We also discuss prognostic indices and how we compute them. We mention also in this chapter some parametric distributions in survival such as the exponential and Weibull distributions. Part of this chapter also deals with Bayesian inference in survival analysis using MCMC techniques and how to make inference about the coefficients in various models. We give an example using the Non-Hodgkin lymphoma data which involves some missing data and show how to deal with this kind of problem. We use `rjags`, (Plummer, 2013) a package in R, (R Core Team, 2018) to do the analysis.

In Chapter 6, we introduce Bayes linear methods. We start the chapter by giving some definitions and theory related to Bayes linear methods. Then we explain Bayes linear kinematics with some aspects such as the issue of commutativity and the use of multiple updates in Bayes linear kinematics with a number of examples. We use the idea of transforming the parameters. We introduce a novel feature which uses the non-conjugate marginal updates in order to find the posterior mean and variance, before the information is propagated through other unknown quantities within a Bayes linear structure. In this chapter, we also give different types of examples such as using binomial observations and Poisson observations and compute the results with the posterior means and variance using full-Bayes analysis.

Chapter 7 describes two sorts of problems. The first is illustrated using the leukaemia example in which we use Bayes linear kinematics with non-conjugate prior updates to

compute the posterior moments for the model parameters. Then we compare different types of methods including log-mode and lognormal forms of Bayes linear kinematics and full Bayes methods. Secondly, we describe the application of Bayes linear kinematics to prognostic index calculation in survival. We illustrate this using the non-Hodgkin lymphoma data. We introduce a novel method that uses a Bayes linear Bayes prognostic network with different sorts of variables such as binary, ordinal, unordered categorical and interval censored variables. We use an offline learning model to determine values for some parameters that we need to calculate the Bayes linear Bayes prognostic index values. We find that the prognostic index values from the Bayes linear Bayes model and the prognostic index values that are calculated from MCMC methods are similar. We give some results and graphs to represent the relationships between the prognostic values for both Bayes analysis and Bayes linear methods. Our prototype prognostic network produces prognostic index values using all, or only some, of the possible covariates almost instantly and has the potential to be used, for example, as a Web-based calculator.

Chapter 8 describes simulation experiments to compare the methods that we use in this thesis. We give three different examples with different ranges of ages and sexes and compare two methods for dealing with categorical variables: the direct and indirect methods.

In Chapter 9, we provide some conclusions and propose some future work in this area.

Chapter 2

Example data sets

2.1 Introduction

In this chapter, we will look in detail at the two data sets that we have used for illustration in this thesis, the non-Hodgkin's lymphoma and leukemia data sets. We give general information about the data and some important definitions for the covariates within each data set.

2.2 Scotland and Newcastle Lymphoma Group (SNLG) data

2.2.1 Background of SNLG data set

In 1979, there was formed a group called the Scotland and Newcastle Lymphoma Group (SNLG) that built up a database on about 18,000 patients with lymphoma within Northern England and Scotland. The process of collecting the data is called *Population Adjusted Clinical Epidemiology* (PACE) and this process was used by the Northern Regional Haematology Group (NRHG). See Proctor and Taylor (2000). The period of time that they needed to collect the data was about 10 years, from 1992 to 2002. The lymphoma group includes specialists from different disciplines such as medicine, pathology, surgery, radiology and clinical oncology. The collected data have been used by various groups of people working in individual centres in order to improve the choice of treatment, such

as chemotherapy or radiotherapy, for the patients. In the thesis, we will focus on non-Hodgkin's lymphoma (NHL) which is one of the most common sorts of cancer. Particularly, we use a subset of the SNLG data set.

2.2.2 Non-Hodgkin's Lymphoma (NHL)

Non-Hodgkin lymphoma is a type of cancer that begins in the cells of the immune system. The immune system fights infections and other diseases. The lymphatic system is regarded as a part of the immune system. The lymphatic system includes the following

- **Lymph vessels:** the lymphatic system has a network of lymph vessels. Lymph vessels branch into all the tissues of the body.
- **Lymph:** the lymph vessels carry clear fluid called lymph. Lymph contains white blood cells, particularly lymphocytes such as B cells and T cells.
- **Lymph nodes:** lymph vessels are connected to small, round masses of tissue called lymph nodes. Groups of lymph nodes are found in the neck, chest, underarms, groin and abdomen. Lymph nodes store white blood cells. They trap and remove bacteria or other harmful substances that may be in the lymph.

See Freedman et al. (2012).

There are more than 12,000 people diagnosed with NHL in the UK every year. The chance of developing the disease increases as people get older and most cases occur in people aged over 65 years with slightly more men than women. See NHS (2018).

2.2.3 Diffuse large B-cell lymphoma

Diffuse large B-cell lymphoma (DLBCL) is the most common sort of non-Hodgkin lymphoma. It is a cancer of blood cells called lymphocytes. Nowadays, the number of patients with the illness in the USA and Europe is approximately 15-20 cases for every 100,000 people, (Martelli et al., 2013). DLBCL is not just one disease. There are a number of different types of DLBCL. The most common type of it is described as the "not otherwise specified" form or DLBCL-NOS. See Miranda et al. (2013).

2.2.4 SNLG data set

The NHL data set incorporates variables that are widely used by clinicians in choosing the appropriate therapy for patients, (Lucas et al., 1998). The relevance of most of these variables is supported by literature on prognostic factors in NHL. First, the information that can be extracted from the clinician about NHL is divided into three groups:

- Pre-treatment information, (*i.e.* information that is required for treatment selection).
- Treatment information, (*i.e.* the various treatment alternatives).
- Post-treatment information, (*i.e.* side effects, and early and long-term treatment results for the disease).

The most important pre-treatment variables are the variable "Clinical Stage", which expresses severity of the disease according to a common clinical classification, and histological classification, which stands for the assessment by a pathologist of tumour tissue obtained from a biopsy. The most important post-treatment variables include the variable "early result", being the endoscopically verified result of the treatment, six to eight weeks after treatment. Possible outcomes are:

- *Complete remission*, i.e. tumour cells are no longer detectable.
- *Partial remission*, some tumour cells are detectable, no change or progressive disease.

Another important post-treatment variable is "3-year result", which represents the patient either surviving three years following treatment or not.

2.2.5 Non-Hodgkin Lymphoma Example: General overview of the covariates

In this section, we describe the covariates in the non-Hodgkin lymphoma data set in detail. We have 14 prognostic variables that have been selected from the clinical research by Professor Proctor, Dr. Sieniawski and Mrs White (Sieniawski et al., 2009). The dependent-variable is survival time with censoring indicator coded as "1" for death and

Stage	Description
I	Lymphoma is discovered in one lymph node site.
II	Lymphoma is discovered in two or more lymph node regions and on the same side of the diaphragm.
III	Lymphoma is discovered in lymph node regions and on both sides of the diaphragm.
IV	Diffuse or disseminated involvement of one or more distant extranodal organs with or without associated lymph node involvement.

Table 2.1: Ann Arbor staging process

"0" for a censored observation. Here, we give more information about the covariates which include binary, continuous, categorical and interval censored variables.

- **Age:** This variable represents the patient's age at diagnosis. It has mean 62 years and standard deviation 14.2 years. This variable is regarded as a continuous variable.
- **Sex:** This is a binary variable, taking the value 1 for male and 2 for female. In our data we have 704 male and 687 female. It seems that the disease is slightly more common in the male than the female.
- **Clinical Stage:** This is an ordinal variable with 4 levels. It represents the way that the doctor can discover the lymphoma in the body of the patient, giving you the number of places that show the lymphoma. See Cancer Research UK (2018a). Knowing the stage of the illness will help the doctors to make an accurate decision about the suitable treatment that the patient needs. The staging process used here is Ann Arbor Staging (Carbone et al., 1971) which is widely used. The categories are ordered from I to IV with the earlier category (I) refering to the least extent of spread and the latter category (IV) refering to the greatest extent of spread. We coded the stages with the values 1,2,3 and 4. See Table 2.1.
- **ECOG:** This is the Eastern Cooperative Oncology Group performance status, (Oken et al., 1982). It is an ordinal variable. It has 5 states from 0 to 5 where the status 5 refers to the death of the patient. So, in our case, we restricted this to 0 to 4. Table 2.2 shows the definitions for ECOG.

Performance scale	Description
0	The patient is fully active and has no performance restrictions.
1	The patient has limited restriction to do strenuous physical activity and he or she has the ability to perform the light work.
2	The patient can take care of himself. However, he will be unable to perform any work activities.
3	The patient has a limited capacity to take care of himself and confined to bed or chair with more than 50% of waking hours.
4	The patient is completely unable to perform any work activities and can not take care of himself and confined to bed or chair.

Table 2.2: ECOG performance

- **Serum Lactate Dehydrogenase (LDH):** Although this variable is actually continuous, it is often categorised and represented as an ordinal variable. Studies show the importance of this variable in the prognosis of non-Hodgkin's lymphoma. See Yadav et al. (2016). The survival time of the patient has been negatively related with the levels of Serum LDH and statistical analyses show that individuals with lower levels of LDH, tend to have longer survival times, (Ferraris et al., 1979). For more information about LDH in the SNLG data set, see Consul (2016). In the SNLG data set, LDH is actually recorded as an interval censored variable. Observations within the normal range are simply recorded as "normal". Observations outside the normal range are recorded as the actual values.

- **Haemoglobin (HB):** This variable is coded as a continuous variable. The measurements of HB are in grams (g) per deciliter (dl) g/dl. It is a protein which is located in the red blood cells and it carries the oxygen from the lungs to the body's tissues and returns carbon dioxide from the tissues back to the lungs. The normal range of HB depends on the age and sex of the person. Table 2.3 represents the normal range for different groups, (Longmore et al., 2014).

- **White Blood Cell (WBC):** This variable is treated as a continuous variable on $(0, \infty)$. White blood cells are also called leukocytes and they are the cells of the immune system that are involved in protecting the body against both infectious

Patient's Group	HB (g/dl)
Adult man	13.8 to 17.2
Adult woman	12.1 to 15.1
Pregnant woman	11 to 12
Children	11 to 16

Table 2.3: Normal range for HB

disease and foreign invaders. See Maton (1997). The range of the WBC in this thesis is between 1.1 and 27.2, where 1 unit is $50 \times 10^9/l$.

2.2.6 Binary variables

Some covariates in the NHL data set are represented as binary variables, such as serum albumin, blood urea nitrogen, etc. In fact, these variables are coded in different ways but the way we put it in the model is using the values either -1 or 1. Below is the list of all the binary variables with a brief description of each variable.

- **Serum Albumin (Albumin):** This is a binary variable. Albumin is considered to be the most abundant protein in the blood plasma for humans and is produced in the liver. Low albumin indicates liver disease and high albumin indicates dehydration. The albumin concentration in blood is 35-55 g/l for the normal range. Any observation outside the above range is a sign of abnormality. In the SNLG data, albumin is categorised as either normal or abnormal.

- **Blood Urea Nitrogen (urea):** This variable is also binary. Urea measures the amount of nitrogen in the blood that comes from the waste product urea. The normal range for urea nitrogen in blood is 5 to 20 mg/dl, see Hosten (1990). The values outside the above range are considered to be abnormal. In the SNLG data, urea is categorised as either normal or abnormal.

- **Alkaline Phosphatase (AP):** Alkaline phosphatase is an important component in hard tissue formation, highly expressed in mineralised tissue cells. See Golub and Boesze-Battaglia (2007). The normal range for AP for those aged over 16 years is 36-113 IU/l. Different age groups have different AP values. Any value outside the normal range is regarded as abnormal. In the SNLG data, it is recorded as a binary variable with value 1 refering to normal and value 2 refering to abnormal.

Variable	No. missing	Percentage missing
Albumin	97	6.97
Urea	51	3.67
Ap	78	5.61
Extranod	1	0.07
Bulk	109	7.84
Marrow	196	14.09
Bsy	13	0.93

Table 2.4: The percentage of missing values for several covariates in NHL

- **Extranodal without Bone Marrow (extranod):** This happens when the lymphoma spreads outside the lymph nodes. See Brooks (2008). The variable is recorded as either "present" or "absent".

- **Bulk Disease (Bulk):** This is to measure whether the patient has bulk disease or not. It describes the tumours which are very large in size, also called bulky tumours. See Pfreundschuh et al. (2008). This is a binary variable.

- **Bone Marrow Involvement (marrow):** Bone marrow is the soft tissue inside the bones where blood cells are made. See El-Galaly et al. (2012). The variable records whether or not the patient has shown evidence of lymphoma disease that is in bone marrow.

- **B-symptoms (Bsy):** The patient with non-Hodgkin lymphoma may have some symptoms such as sweating at night, temperature that goes and returns without any infection, losing weight (more than one tenth of the total weight) and unexplained itching. See Cancer Research UK (2018b). These symptoms are called B-symptoms. The presence and absence of B-symptoms has an important significance in prognosis exactly the same as in the staging of NHL. The variable records the presence or absence of B-symptoms.

2.2.7 Missing data

Several of the covariates have missing values for some patients. This is summarised in Table 2.4.

2.3 Leukemia example

2.3.1 Introduction

In this thesis, we use also a data set on patients with leukemia. These data are taken from the North West Leukemia Register in the UK in order to investigate the leukemia survival time for 1043 patients between 1982 and 1998 where 879 died and 164 were censored. See Henderson et al. (2002).

2.3.2 General overview of the covariates

In this data set, we have some covariates which we believe might have an effect on the survival times for the individuals. These covariates are age, sex, white blood cell count (WBC) and a measure of the deprivation of the area of residence which is called the Townsend score (Townsend et al., 1988). We give the censoring indicator "1" for death and "0" for censored data. The covariates are

- **Age:** This variable represents the age of the patient in years.
- **Sex:** This is the sex of the patient. We coded the variable to be "1" for the male patient and "-1" for the female patient. We have 547 (52%) female and 496 (48%) male.
- **White blood cell (WBC):** See Section 2.2.5 for more details about white blood cell count at the time of diagnosis (with 1 unit= $50 \times 10^9/l$).
- **Deprivation score (Depscore):** This variable measures the deprivation for the residential area of the patient. We use the Townsend deprivation index (TDI) (Townsend et al., 1988). The scale of the variable is from -7 to 10 with lower values indicating more severe deprivation. Alston et al. (2007) mentioned that TDI can vary by region and they found that deprivation affected cancer rates.

2.4 Summary

In this chapter, we have given some general and useful information about the data sets that will be used in the thesis. An overview of the different sorts of covariates has been

given in both cases, SNLG and leukemia.

Chapter 3

Bayesian inference and Generalised Linear Models (GLMs)

3.1 Introduction

In this chapter, we will explain in detail what Bayesian inference is and illustrate generalised linear models by introducing one of the most common models that is used widely in medical studies, which is the logistic regression model. In Section 3.2, we give an introduction to Bayesian inference, henceforth abbreviated to BI. In Section 3.3, we explain some important methods to calculate some summary statistics related to the posterior distribution (e.g. the posterior mean and the posterior variance) using various methods. In Section 3.4, we explain Markov Chain Monte Carlo methods (MCMC) which are very widely used in Bayesian statistics. In Section 3.4.4, we give an explanation of the Gibbs sampler, a common method in MCMC, which depends on the calculation of the full conditional distribution for the parameter of interest given all the variables in the model. Section 3.4.8 demonstrates the use of another method in MCMC which is called the Metropolis-Hastings method. This method involves generating samples from a proposal distribution in order to evaluate the posterior distributions. In Section 3.5, we illustrate the generalised linear model (GLM) with some common link functions related with GLM. In Section 3.7, we demonstrate different kinds of variable selection methods, focussing on Bayesian variable selection methods. We introduce one type of prior which is used with variable selection and is called a spike and slab prior, with more details about this prior discussed in Section 3.7.5. In Section 3.8, we define the missing data mechanism

with different sorts of missingness. We give a brief introduction to data augmentation in Section 3.9. In Section 3.10, we have an example on lung transplants. In this example we fit a logistic model as the response variable is whether the lung is used for transplant or not. The lung transplant example also illustrates the missing data problem. Finally we give the summary of Chapter 3 in Section 3.11.

3.2 Introduction to Bayesian inference

Bayesian inference has become widely used in different disciplines, such as medicine, biology, clinical trails, bioinformatics, survival analysis, physics and so on.

The idea behind doing Bayesian inference is to infer about the "unknown" quantity of interest, say θ, in the model and learn about it from the data. Therefore, there is always uncertainty which is associated with this parameter and we represent this in the form of the *joint* probability density for all unobserved quantities. See Gamerman and Lopes (2006).

In a Bayesian context, we always describe the uncertainty in the values of the unknown quantities in terms of probability distributions which represent beliefs about the values. Therefore, we assign probabilities to the values of those unknown quantities, say θ. The distribution before observing data is known as the *prior* distribution. If this is a continuous distribution, we can write its density as $\pi(\theta)$. After observing the data, y, we update our prior belief and find the *posterior* probability distributions $\pi(\theta \mid y)$. The information that has come from the data is found in the *likelihood function*, $\mathrm{L}(\theta \mid y)$ which refers to the sampling density for the data. Multiplying the prior density with the likelihood function, gives the posterior density as follows

$$\pi(\theta \mid y) = \frac{\pi(\theta)\mathrm{L}(\theta \mid y)}{\int \pi(\theta)\mathrm{L}(\theta \mid y)\ d\theta}. \tag{3.1}$$

The integral in the denominator, known as the *normalising constant*, is often intractable. In such cases we need to use numerical methods in order to evaluate posterior distributions. The most common sort of numerical method in recent years is *Markov chain Monte Carlo* (MCMC) methods. See, for example, Gilks et al. (1996). Often MCMC may be implemented using software such as "BUGS" (Spiegelhalter et al., 1996).

3.3 Numerical integration methods

In Chapter 6 we will discuss Bayes linear kinematics and Bayes linear Bayes graphical models. We will see that it is necessary to revise our mean and variance of an underlying unknown quantity when we observe and associated variable. We will deal with cases where the observable variables are not Gaussian and, in particular, in Section 6.7, we will introduce a new method which involves using a non-conjugate prior distribution for the underlying quantity. We will need to use numerical integration to evaluate the revised moments. Therefore, in the section, we review some methods of numerical integration which may be used.

Suppose that we need to calculate the posterior mean of θ as follows

$$\mathrm{E}(\theta \mid y) = \frac{\int \theta \pi(\theta) \mathrm{L}(\theta \mid y) d\theta}{\int \pi(\theta) \mathrm{L}(\theta \mid y) d\theta}$$

and the posterior variance of θ will be

$$\mathrm{Var}(\theta \mid y) = \frac{\int \theta^2 \pi(\theta) \mathrm{L}(\theta \mid y) d\theta}{\int \pi(\theta) \mathrm{L}(\theta \mid y) d\theta} - \Big[\mathrm{E}(\theta \mid y)\Big]^2.$$

Sometimes we can calculate all these integrals analytically especially when the prior distribution is conjugate (if the posterior distribution and the prior belong to the same family of distributions, then the prior is called a conjugate prior) to the likelihood but this is not always the case. These integrals are often complicated when we deal with the problem of non-conjugate prior distributions especially when there is more than a small number of parameters in the analysis. Therefore, we need numerical methods to solve problems of this kind.

3.3.1 Trapezoidal rule

The main idea of using numerical integration is to calculate an approximate solution for definite integrals. One of these numerical methods is called the trapezoidal rule. See Jones et al. (2014). This method gives an approximate value for an integral between two values.

$x_i = a + i\Delta x$	$f(x_i)$	evaluate
$x_0 = a + 0\Delta x$	$f(0)$	0.00000
$x_1 = a + 1\Delta x$	$f(0.05)$	0.00214
$x_2 = a + 2\Delta x$	$f(0.10)$	0.00729
$x_3 = a + 3\Delta x$	$f(0.15)$	0.01382
$x_4 = a + 4\Delta x$	$f(0.20)$	0.02048
$x_5 = a + 5\Delta x$	$f(0.25)$	0.02637
$x_6 = a + 6\Delta x$	$f(0.30)$	0.03087
$x_7 = a + 7\Delta x$	$f(0.35)$	0.03364
$x_8 = a + 8\Delta x$	$f(0.40)$	0.03456
$x_9 = a + 9\Delta x$	$f(0.45)$	0.03369
$x_{10} = a + 10\Delta x$	$f(0.50)$	0.03125

Table 3.1: Evaluate the functions in order to compute trapezoidal rule

Suppose we have the integral

$$I = \int_a^b f(x)dx.$$

This can be done by dividing the interval between a and b into n subintervals of width Δx. So, $\Delta x = (b-a)/n$. Then, to calculate the trapezoidal approximation, we have

$$T = \frac{\Delta x}{2}\Big[f(x_0) + 2f(x_1) + \cdots + 2f(x_{n-1}) + f(x_n)\Big]$$

where $x_i = a + i\Delta x$. Now let us explain the method with a simple example.

Suppose we need to calculate the following integral, $\int_0^{0.5} \theta^2(1-\theta)^3 d\theta$. We have in this case $a = 0$, $b = 0.5$ and also give $n = 10$ for example. Then $\Delta x = 0.05$, and we have the values shown in Table 3.1.

Therefore, the numerical integration for the above integral gives $T = 0.01092$ following the calculations from Table 3.1. Now the exact solution for this integral 0.01094. As a result, there is an absolute error 2×10^{-05} between the two values. Thus the solution using the trapezoidal rule is close to the exact value. We notice that we can obtain a more accurate result by increasing the number of subintervals, say $n = 100$.

3.3.2 Laplace approximation method

It is very important in Bayesian framework to calculate the integrals of the form

$$I \approx \frac{\int g(\theta)\mathrm{L}(\theta|\underline{y})\pi(\theta)d\theta}{\int \mathrm{L}(\theta|\underline{y})\pi(\theta)d\theta} \tag{3.2}$$

where $g(\theta)$ is an arbitrary function of θ and $\mathrm{L}(\theta|\underline{y})$ is the likelihood function and $\pi(\theta)$ is the prior density. We can write (3.2) as

$$I \approx \frac{\int g(\theta)e^{\{\ell(\theta|\underline{y})+f(\theta)\}}d\theta}{\int e^{\{\ell(\theta|\underline{y})+f(\theta)\}}d\theta} \tag{3.3}$$

where $\ell(\theta|\underline{y}) = \log f(y_1, ..., y_n|\theta)$ is called the log likelihood function and $f(\theta) = \log\{\pi(\theta)\}$ is the log of the prior density $\pi(\theta)$.

For instance, if θ has one-dimension, then $g(\theta) = \theta$ gives us the posterior mean of the distribution. More generally, when we have $g(\theta) = \theta^p$, we can gain the pth moment of the posterior distribution. See Press (2009). Now the denominator of (3.3) is called the normalising constant which is sometimes analytically intractable. So, we need approximation methods to evaluate both integrals in (3.3).

The Laplace approximation is one of the analytical methods that is useful to compute the integrals in (3.3). This method of approximation was introduced by Tierney and Kadane (1986). It depends on the normal approximation in order to calculate the posterior mean and posterior variance and so on. Moreover, recent developments have led to the use of the integrated nested Laplace approximation (INLA) method which is a very efficient method to give accurate approximations for the posterior marginals in seconds or in minutes while using MCMC methods needs more time to run. See Rue et al. (2009).

Suppose that we are interested in calculating the expectation in (3.3). So we can rewrite it in the following expression

$$\mathrm{E}\{g(\theta)|\underline{y}\} = \frac{\int \exp\{-nk^*(\theta)\}d\theta}{\int \exp\{-nk(\theta)\}d\theta} \tag{3.4}$$

where n is the number of data points,

$$-nk(\theta) = \log\{\mathrm{L}(\theta|\underline{y})\} + \log\{\pi(\theta)\}$$

and

$$-nk^*(\theta) = \log\{g(\theta)\} + \log\{\mathrm{L}(\theta|\underline{y})\} + \log\{\pi(\theta)\}.$$

Now, we use Taylor expansions for k and k^* to find the modes $\hat{\theta}$ and $\hat{\theta}^*$ respectively.

$$-k(\hat{\theta}) = \max_{\theta}\{-k(\theta)\} \qquad \text{and} \qquad -k^*(\hat{\theta}^*) = \max_{\theta}\{-k^*(\theta)\}$$

and retain these to the quadratic terms. For instance, to estimate the denominator, we have

$$\int \exp\{-nk(\theta)\}d\theta \approx \sqrt{2\pi}\sigma n^{-1/2}\exp\{-nk(\hat{\theta})\}$$

where $\sigma = \Big[-\frac{d^2}{d\theta^2}k(\theta)|_{\theta=\hat{\theta}}\Big]^{-1/2}$. Likewise, we can do the same thing for the numerator.

We obtain

$$\mathrm{E}\{g(\theta)|\underline{y}\} \approx \Big(\frac{\sigma^*}{\sigma}\Big)\frac{g^*(\hat{\theta})\mathrm{L}^*(\hat{\theta}|\underline{y})\pi^*(\hat{\theta})}{\mathrm{L}(\hat{\theta}|\underline{y})\pi(\hat{\theta})} \tag{3.5}$$

where, $\sigma^* = \Big[-\frac{d^2}{d\theta^2}k^*(\theta)|_{\theta=\hat{\theta}}\Big]^{-1/2}$.

Now let us give an example to illustrate how to use the Laplace method to obtain the posterior moments.

Suppose we have a Poisson likelihood function, so that

$$\mathrm{L}(\theta|\underline{y}) = \prod_{i=1}^{n} f(y_i|\theta) \propto \theta^{n\bar{y}}e^{-n\theta}.$$

Then our conjugate prior for θ is a gamma density with two parameters, the shape parameter a and the rate parameter b, so $\theta \sim \mathrm{Gamma}(a, b)$ and

$$\pi(\theta) \propto \theta^{a-1}e^{-b\theta} \qquad a > 0 \quad , \quad b > 0$$

So, the posterior density $\pi(\theta|\underline{y})$ will be

$$\pi(\theta|\underline{y}) \propto \theta^{a+n\bar{y}-1}e^{-(b+n)\theta}.$$

Therefore, let $a^* = a + n\bar{y}$ and $b^* = b + n$. It is obvious that the exact posterior mean is

$$\mathrm{E}(\theta|\underline{y}) = \frac{a^*}{b^*}$$

Now, we are interested in finding the approximation of the posterior mean using the Laplace method. We have

$$\mathrm{E}(\theta|\underline{y}) \approx \frac{\int \theta e^{\{\ell(\theta|\underline{y})+f(\theta)\}}d\theta}{\int e^{\{\ell(\theta|\underline{y})+f(\theta)\}}d\theta}$$

which can be written as

$$\mathrm{E}(\theta|\underline{y}) \approx \frac{\int \exp\{-nk^*(\theta)\}d\theta}{\int \exp\{-nk(\theta)\}d\theta}$$

where

$$\begin{aligned}
-nk(\theta) =& \log\{\mathrm{L}(\theta|\underline{y})\} + \log\{\pi(\theta)\} \\
=& n\bar{y}\log(\theta) - n\theta + (a-1)\log(\theta) - b\theta \\
=& (n\bar{y} + a - 1)\log(\theta) - (b+n)\theta \\
=& (a^* - 1)\log(\theta) - b^*\theta
\end{aligned}$$

and

$$\begin{aligned}
-nk^*(\theta) =& \log(\theta) + \log\{\mathrm{L}(\theta|\underline{y})\} + \log\{\pi(\theta)\} \\
=& a^*\log(\theta) - b^*\theta.
\end{aligned}$$

Now, we need to find the modes $\hat{\theta}$ and $\hat{\theta}^*$ respectively as follow

$$\frac{d}{d\theta}\Big[-nk(\theta)\Big] = \frac{a^*-1}{\theta} - b^* = 0$$

so,

$$\hat{\theta} = \frac{a^*-1}{b^*}.$$

and

$$\frac{d}{d\theta}\Big[-nk^*(\theta)\Big] = \frac{a^*}{\theta} - b^* = 0$$

so,

$$\hat{\theta}^* = \frac{a^*}{b^*}.$$

By substituting $\hat{\theta}$ in $\{-nk(\theta)\}$ and $\hat{\theta}^*$ in $\{-nk^*(\theta)\}$, we obtain

$$\begin{aligned}-nk(\theta) =&(a^*-1)\log\left(\frac{a^*-1}{b^*}\right)-b^*\left(\frac{a^*-1}{b^*}\right)\\ =&(a^*-1)\left[\log\left(\frac{a^*-1}{b^*}\right)-1\right].\end{aligned}$$

Similarly,

$$\begin{aligned}-nk^*(\theta) =&a^*\log\left(\frac{a^*}{b^*}\right)-b^*\left(\frac{a^*}{b^*}\right)\\ =&a^*\left[\log\left(\frac{a^*}{b^*}\right)-1\right].\end{aligned}$$

We should also find the second derivative of the Taylor expansions in order to find σ and σ^*. So

$$\frac{d^2}{d\theta^2}\Big[-nk(\theta)\Big]=-\frac{(a^*-1)}{\theta^2}$$

and

$$\sigma=\Big[-\frac{d^2}{d\theta^2}nk(\theta)|_{\theta=\hat{\theta}}\Big]^{-1/2}=\frac{\sqrt{a^*-1}}{b^*}.$$

Likewise,

$$\frac{d^2}{d\theta^2}\Big[-nk^*(\theta)\Big]=-\frac{a^*}{\theta^2}$$

so,

$$\sigma^*=\Big[-\frac{d^2}{d\theta^2}nk^*(\theta)|_{\theta=\hat{\theta}^*}\Big]^{-1/2}=\frac{\sqrt{a^*}}{b^*}.$$

Thus,

$$\left(\frac{\sigma^*}{\sigma}\right)=\frac{\sqrt{a^*}}{\sqrt{a^*-1}}.$$

So, our approximate posterior mean is calculated exactly as in (3.5)

$$\mathrm{E}(\theta|\underline{y})\approx\left(\frac{\sigma^*}{\sigma}\right)\frac{\exp\{-nk^*(\hat{\theta}^*)\}}{\exp\{-nk(\hat{\theta})\}}$$

$$\mathrm{E}(\theta|\underline{y}) \approx \frac{\sqrt{a^*}}{\sqrt{a^*-1}} . \frac{\exp\left\{a^*\left[\log\left(\frac{a^*}{b^*}\right)-1\right]\right\}}{\exp\left\{(a^*-1)\left[\log\left(\frac{a^*-1}{b^*}\right)-1\right]\right\}}$$

For example, let $a = 2$, $b = 3$, $n = 10$, $\bar{y} = 5$. We obtain $\mathrm{E}(\theta|\underline{y}) \approx 4.00013$ while the exact posterior mean $\mathrm{E}(\theta|\underline{y}) = a^*/b^* = 4$. Therefore, the absolute error (representing the difference between the exact value and the approximate value) is 0.00013, which is very small.

3.4 Markov Chain Monte Carlo methods

3.4.1 Introduction

Markov Chain Monte Carlo (MCMC) techniques have become the most popular methods for evaluating posterior distributions. These techniques allow sampling from the posterior distribution of the unknown parameters in the model when there is no analytical solution. The idea of using MCMC was originally proposed by Metropolis et al. (1953) as an efficient method for simulation. There are two main algorithms in MCMC that are used in the majority of cases, the *Gibbs sampler* (Geman and Geman, 1984; Gelfand and Smith, 1990) and *Metropolis-Hastings* algorithms (Hastings, 1970).

The use of Monte Carlo methods in Bayesian statistics has dramatically increased since the early 1990s. The basic idea of methods of this kind is to draw random samples from probability distributions and, when the number of draws becomes large, Monte Carlo methods give good approximations to properties of the distributions. These methods are used when there is no analytic solution or there is a difficulty in finding the numerical solution. Therefore, we obtain an approximate solution using these methods. See Jackman (2009); Lesaffre and Lawson (2012); Gelman et al. (2014). In the following subsections, we will demonstrate the most common direct and indirect sampling methods to evaluate the posterior summaries in Bayesian statistics.

In Chapters 6, 7 and 8, when we introduce new ideas in Bayes linear kinematics and Bayes linear Bayes models, we will compare our results using these ideas in examples with real data and in simulation experiments with results obtained using standard Bayesian

analysis and MCMC methods.

We use software called `rjags` (Plummer, 2013) to fit the models and do the analysis as well as using `R` (R Core Team, 2018).

3.4.2 Monte Carlo integration

In this subsection, we explain the direct sampling method for Monte Carlo integration. Suppose that we have the function $f(\theta)$ and our target is to find $\delta = \mathrm{E}\{g(\theta)\} = \int g(\theta)f(\theta)d\theta$. So, if $\theta_1, ..., \theta_n \overset{iid}{\sim} f(\theta)$, we have

$$\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n} g(\theta_i) \tag{3.6}$$

and that converges to $\mathrm{E}\{g(\theta)\}$ with probability 1 as $n \to \infty$, using the strong law of large numbers. In the case of Bayesian inference, $f(\theta)$ is the posterior distribution and δ is the posterior mean of $g(\theta)$. As a result, in order to compute the posterior mean, we need just a sample of size n from the posterior distribution. See Carlin and Louis (2008).

3.4.3 Importance sampling

Geweke (1989) suggested an important method for sampling indirectly from the posterior (target) distribution. Let $\pi(\theta|\underline{y}) \propto \mathrm{L}(\theta|\underline{y})\pi(\theta)$ be the target distribution and suppose that we wish to find the approximate mean for it. So

$$\mathrm{E}\left[g(\theta)|\underline{y}\right] = \frac{\int g(\theta)\mathrm{L}(\theta|\underline{y})\pi(\theta)d\theta}{\int \mathrm{L}(\theta|\underline{y})\pi(\theta)d\theta}$$

where $\mathrm{L}(\theta|\underline{y})$ is the likelihood function and $\pi(\theta)$ is the prior density.

Now, suppose we can easily sample from a density $s(\theta)$ and we define $w(\theta)$ as an importance weight, $w(\theta) = \mathrm{L}(\theta|\underline{y})\pi(\theta)/s(\theta)$. We have

$$\begin{aligned}\mathrm{E}\left[g(\theta)|\underline{y}\right] &= \frac{\int g(\theta)w(\theta)s(\theta)d\theta}{\int w(\theta)s(\theta)d\theta} \\ &\approx \frac{\frac{1}{n}\sum_{i=1}^{n} g(\theta_i)w(\theta_i)}{\frac{1}{n}\sum_{i=1}^{n} w(\theta_i)}\end{aligned}$$

where $\theta_i \overset{iid}{\sim} s(\theta)$ and this $s(\theta)$ stands for the importance function. We also notice that if $s(\theta)$ is a good approximation to the posterior density then all the weights will be approximately equal. So the algorithm for importance sampling can be written as follows.

Algorithm 1: Importance sampling algorithm for calculating the posterior mean of a distribution

1 for $i = 1$ to n do.
2 sample $\theta_i \sim s(\theta)$.
3 $w_i \leftarrow \mathrm{L}(\theta|\underline{y})\pi(\theta_i)/s(\theta_i)$
4 end for
5

$$\mathrm{E}\left[g(\theta)|\underline{y}\right] \approx \frac{\frac{1}{n}\sum_{i=1}^{n} g(\theta_i)w_i}{\frac{1}{n}\sum_{i=1}^{n} w(\theta_i)}.$$

See Carlin and Louis (2008); Jackman (2009).

3.4.4 The Gibbs sampler

In this section, we demonstrate the basic idea of using the Gibbs sampler method and how to implement it in practice.

Assume that our model has n parameters, say $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_n)'$. This method involves generating samples from the *full conditional* distributions, $\pi(\theta_i \mid \theta_{j\neq i}, \underline{y})$, where $i = 1, 2, ..., n$ and the observed data are $\underline{y}$.

The algorithm starts with assigning some initial values $(\theta_2^{(0)}, ..., \theta_n^{(0)})$ and then we apply the algorithm as follows

Algorithm 2: Gibbs sampling algorithm

1 Initialise $\theta_i, i = 1, ..., n$.
2 For $k = 1, ..., K$.
3 Sample $\theta_1^{(k)}$ from $\pi(\theta_1 \mid \theta_2^{(k-1)}, \theta_3^{(k-1)}, ..., \theta_n^{(k-1)}, \underline{y})$.
4 Sample $\theta_2^{(k)}$ from $\pi(\theta_2 \mid \theta_1^{(k)}, \theta_3^{(k-1)}, ..., \theta_n^{(k-1)}, \underline{y})$.
$\vdots$
5 Sample $\theta_n^{(k)}$ from $\pi(\theta_n \mid \theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)}, ..., \theta_{n-1}^{(k)}, \underline{y})$.
6 Change counter k to $k + 1$, and return to step 2.

Let $\underline{\theta}^{(k)} = (\theta_1^{(k)}, ..., \theta_n^{(k)})'$. Then the sequence ... $\underline{\theta}^{(k-1)}, \underline{\theta}^{(k)}, \underline{\theta}^{(k+1)}$ forms a Markov chain. As $k \to \infty$, the distribution of $\underline{\theta}^{(k)}$ tends to the joint distribution of $\theta_1, ..., \theta_n$, known as the target distribution, which in Bayesian statistics, is typically the posterior distribution. See, e.g., Gilks et al. (1996).

3.4.5 Burn-In and convergence in MCMC samples

It is important to check whether the distribution of our sampled values is close to the stationary distribution of the Markov chain. Therefore, "burn-in" is the process of removing the initial values which are related to the non-stationary part of the Markov chain. We can visualise the convergence of the samples and that can be done by looking at the trace plots of these random samples against the iterations of samples in the model particularly if two or more parallel chains are use. For instance, if the burn-in is 1000 iterations, and we need to make the number of iteration 10000, then we are determining 11000 of the generated samples in order to give a summary of the posterior distribution of the parameter of interest.

There are other tools that can help to assess the convergence of MCMC chains such as diagnostic statistics. Some of these statistics are available when using the software `rjags` in the "CODA" package which was written by Plummer et al. (2006).

We give a brief introduction to two statistics that have been used widely to assess convergence. The first statistic is called the Brooks-Gelman-Rubin (BGR) statistic. See Brooks and Gelman (1998); Gelman and Rubin (1992). This statistic deals with two types of variability when we have multiple chains running. The first one is the variability of the observations within each chain and the second is the variability between the chains. If the variability between the chains is relatively small compare to the variability within each chain, then the chains are judged to have converged to the posterior distribution. See Hosmer et al. (2013).

Now before we illustrate the second statistic, let us take the case when we assume that we have k chains where $j = 1, ..., k$. Each chain gives sampled values with $i = 1, ..., n$ denoted as μ_{ij}. Therefore, the variance of the parameter values for one chain j is

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\mu_{ij} - \bar{\mu}_{.j})^2$$

where $\bar{\mu}_{.j}$ is the mean of the sampled parameter values in the chain j. We define the

variability within the chain as W which represents the mean of the the variances of all of the chains

$$W = \frac{1}{k}\sum_{j=1}^{k} s_j^2.$$

and we denote the variability between the chains as B, which is given by

$$B = \frac{1}{k-1}\sum_{j=1}^{k} n(\mu_{.j} - \bar{\mu}_{..})^2,$$

where $\bar{\mu}_{..}$ is the mean of all the sample value from all the chains. So we can compute the expected marginal posterior variance for the parameter as follows

$$\hat{S}_\theta = \frac{n-1}{n}W + \frac{1}{n}B.$$

As we can see, this $\hat{S}_\theta$ refers to the weighted mean of the two variances W and B.

Now, we introduce the second statistic which is called the "effective sample size" which was suggested by Spiegelhalter et al. (2002) and can be computed as

$$n_{\text{effective}} = nk\frac{\hat{S}_\theta}{B} \tag{3.7}$$

and this formula has been defined as a function in `rjags` as `effectiveSize()`. So for example, we might consider the mixing of the chains to be satisfactory if the variability between the chain is lower than the expected posterior variability and that makes the quantity in (3.7) larger. Gelman and Hill (2007) suggested that the effective sample size should be at least 100 samples in order to conclude that we have obtained sufficient MCMC samples.

3.4.6 Thinning

Thinning is the process of discarding all-but-every k-th sample from a sequence of MCMC samples of the posterior distribution. See Link and Eaton (2011). To illustrates the idea of thinning in Monte Carlo methods, assume we have the following situation.

We have generated samples which are dependent. The posterior mean $\mathrm{E}(\theta|\underline{y})$ of the unknown quantity is approximated by the sample mean $\bar{\theta}$ and the accuracy of this approximation is measured by the Monte Carlo variance of $\bar{\theta}$. This Monte Carlo variance

will be larger than it would be given a sample of independent draws from the posterior distribution. This is the case when the value of the samples from successive iterations are positively autocorrelated.

Now the autocorrelation across iterations can be reduced by thinning the chain. That will give us the sampled values from iterations $k, 2k, 3k, ...$ where k is an integer, $k > 1$. Thinning gives a sample of n/k values and increased Monte Carlo variance. However, when we have a positive auto correlation, this increase can be small. There are cases where time-consuming computations are done on each sampled value after it is collected. In such cases it may be more computationally efficient to increase n and then thin using $k > 1$ before executing these post-sample computations. For instance, the computation of the posterior predictive means. Thinning is also important to assess the convergence and when we have a problem with storage space which is nowadays not as likely to be a problem as computers generally have very big storage spaces. All these issues are discussed by Geyer (1992).

3.4.7 Example: normal random sample

Suppose we have a random sample from a normal distribution with mean μ and variance σ^2. Hence, $Y_i \mid \mu, \tau \sim N(\mu, 1/\tau)$, where $\tau = 1/\sigma^2$ and $i = 1, 2, ..., n$. We use a semi-conjugate prior for μ and τ which is represented as follows

$$\mu \sim N(a, 1/b) \qquad \text{and} \qquad \tau \sim Ga(c, d)$$

with μ and τ independent, so that $\pi(\mu, \tau) = \pi(\mu)\pi(\tau)$. Applying Bayes theorem here, the joint posterior density for μ and τ is

$$\pi(\mu, \tau \mid y) \propto \pi(\mu, \tau)\mathrm{L}(\mu, \tau \mid y) \tag{3.8}$$

where $\mathrm{L}(\mu, \tau \mid y)$ is the likelihood function. To find the full conditional distribution for μ, we need to condition on the second parameter which is τ here, to obtain $f(\mu \mid \tau, y)$. Likewise, if we condition on μ, we get $f(\tau \mid \mu, y)$. Therefore, the FCD (which stands for full conditional distribution) for μ and τ can be summarised respectively in the following expressions

$$\mu \mid \tau, y \sim N\left(\frac{ab + n\bar{y}\tau}{b + n\tau}, \frac{1}{b + n\tau}\right)$$

and

$$\tau \mid \mu, y \sim Ga\Big(c + \frac{n}{2}, d + \frac{n}{2}\big\{s^2 + (\bar{y} - \mu)^2\big\}\Big).$$

where $s^2 = \sum(y_i - \bar{y})^2/n$ and $\bar{y} = \sum y_i/n$. In order to implement our algorithm, we need to give initial values for μ and τ. So, for example we could use our prior means to be starting values for the Gibbs algorithm, where $\mu^{(0)} = a$ and $\tau^{(0)} = c/d$. We obtain new values $\tau^{(k)}$ and $\mu^{(k)}$ from $\tau^{(k-1)}$ and $\mu^{(k-1)}$ by successive generation of values

$$\mu^{(k)} \sim N\left(\frac{ab + n\bar{y}\tau^{(k-1)}}{b + n\tau^{(k-1)}}, \frac{1}{b + n\tau^{(k-1)}}\right)$$

$$\tau^{(k)} \sim Ga\Big(c + \frac{n}{2}, d + \frac{n}{2}\big\{s^2 + (\bar{y} - \mu^{(k)})^2\big\}\Big).$$

A `R` function is written to illustrate how to generate samples from the posterior distribution of μ and τ. See Appendix A.3.1. So the trace plots (which show the history of the sampled parameter value across the iterations of the chain and therefore where the chain has been exploring) and the autocorrelation plot for μ and τ using a Gibbs algorithm are shown in Figure 3.1.

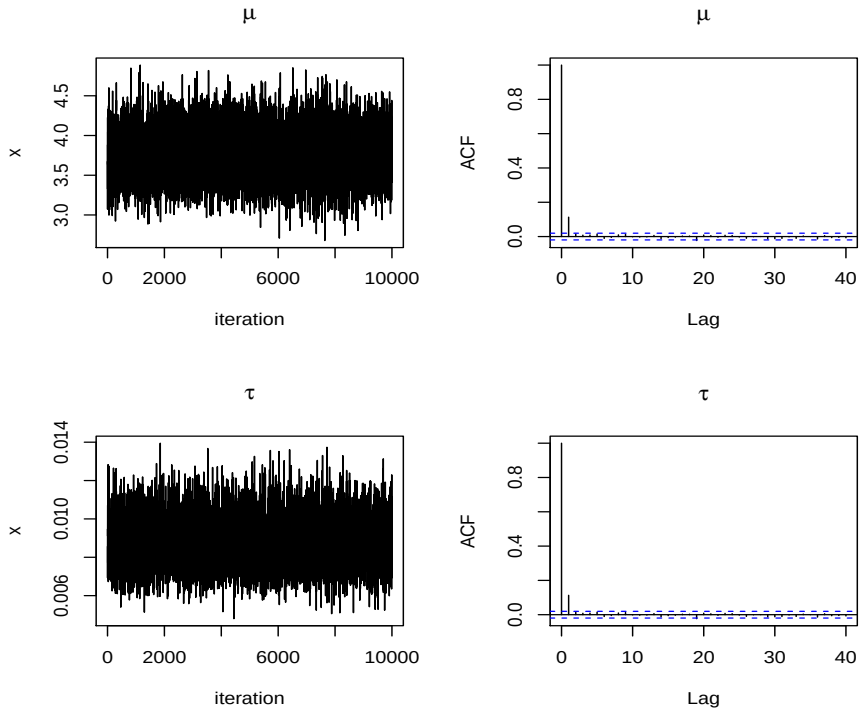


Figure 3.1: Trace plots and the autocorrelation plots for μ and τ

As we can see from the trace plots in Figure 3.1, there is good mixing for both parameters and the chain has converged, so these chains are stationary. We notice from the trace plots that there is not any long term trend in our chains. As a result the local average value of μ and τ in the chain is roughly constant. Similarly, we can see in the autocorrelation plots, that the samples that we generated from MCMC using the Gibbs sampler are almost independent samples.

3.4.8 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm arose when the algorithm of Metropolis et al. (1953) was generalised by Hastings (1970). It is one of the MCMC algorithms that has been used widely in the Bayesian framework. The aim of this algorithm is to sample realisations from the posterior distribution $\pi(\theta|\underline{y})$. It is useful particularly when it is difficult to sample from the FCDs.

Suppose we wish to sample realisations from the posterior density $\pi(\theta|\underline{y})$ and all of the FCDs are non-standard. Furthermore, suppose that we have a proposal distribution with density $q(\theta^*|\theta)$, which can be easily sampled. This distribution can help us to propose new values θ^* from the current value θ. So the algorithm can be written as in Algorithm 3.

Algorithm 3: Metropolis-Hastings algorithm

1 sample θ^* from the proposal distribution $q(\theta^*|\theta^{(t-1)})$.
2

$$A = \frac{\pi(\theta^*|\underline{y})q(\theta^*|\theta^{(t-1)})}{\pi(\theta^{(t-1)}|\underline{y})q(\theta^{(t-1)}|\theta^*)} \tag{3.9}$$

3 $\alpha = \min(A, 1)$.
4 sample $U \sim \text{Unif}(0, 1)$.
5 **if** $U \leq \alpha$ **then**
6 $\quad \theta^{(t)} = \theta^*$.
7 else
8 $\quad \theta^{(t)} = \theta^{(t-1)}$.
9 end if

3.4.9 Metropolis within Gibbs algorithm

If we have a posterior distribution with FCDs, it may be that some of these FCDs can be sampled directly and others cannot. In the latter case we can use Metropolis-Hastings updates. This type of algorithm is called Metropolis within Gibbs. This algorithm uses each of the full conditional distribution in turn. In each case we either sample directly or apply the Metropolis-Hastings update for the FCDs when direct sampling is difficult.

The selection of a suitable proposal distribution in a Metropolis-Hastings scheme for the whole collection of unknowns could be cumbersome. Nevertheless, it is possible sometimes to sample from the FCDs for a subset of θ. Suppose that the full conditional distribution for the j^{th} component of θ is written as:

$$\pi(\theta_j \mid \theta_1, \theta_2, ..., \theta_{j-1}, \theta_{j+1}, .., \theta_n, y) = \pi(\theta_j \mid \theta_{-j}, \underline{y}) \qquad j = 1, ..., n.$$

A Metropolis within Gibbs algorithm is given by Algorithm 4.

Algorithm 4: Metropolis-Hastings Algorithm: Component-wise Transitions.

1. For $j = 1, ..., n$. Initialise the chain with $\theta^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_n^{(0)}\right)'$.
2. Obtain a new value $\theta^{(j)} = \theta_1^{(j)}, \theta_2^{(j)}, ..., \theta_n^{(j)}$ from $\theta^{(j-1)}$ using consecutive values generated from proposal distributions:

   $\theta_1^{(j)} \sim \pi\left(\theta_1 \mid \theta_2^{(j-1)}, \theta_3^{(j-1)}, ..., \theta_n^{(j-1)}, \underline{y}\right)$ using a Metropolis-Hastings step with proposal $q_1\left(\theta_1^* \mid \theta_1^{(j-1)}\right)$

   $\theta_2^{(j)} \sim \pi\left(\theta_2 \mid \theta_1^{(j-1)}, \theta_3^{(j-1)}, ..., \theta_n^{(j-1)}, \underline{y}\right)$ using a Metropolis-Hastings step with proposal $q_2\left(\theta_2^* \mid \theta_2^{(j-1)}\right)$

   $\vdots$

   $\theta_n^{(j)} \sim \pi\left(\theta_n \mid \theta_1^{(j-1)}, \theta_2^{(j-1)}, ..., \theta_n^{(j-1)}, \underline{y}\right)$ using a Metropolis-Hastings step with proposal $q_n\left(\theta_n^* \mid \theta_n^{(j-1)}\right)$
3. Set j+1 and return to step 2

For the j^{th} component of θ, if it is feasible to sample from a FCD of a known form,

then we could do so.

3.5 Generalised linear model

3.5.1 Introduction

In Chapter 6, in the context of Bayes linear kinematics and Bayes linear Bayes models, we will consider how non-Gaussian observable variables can be linked to corresponding variables in an underlying linear structure. This involves ideas related to those of generalised linear models. Therefore, in this section, we introduce some basic ideas of generalised linear models.

In a normal linear model, we typically write $Y_i \sim N(\mu_i, \sigma^2)$ where

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{3.10}$$

where $\beta_0, \beta_1, ..., \beta_p$ are the regression coefficients of the model and x_{ij} is the value of the covariate j for observation i. So we can write the regression model as $Y_i = \mu_i + \varepsilon_i$, where ε_i $(i = 1, ..., n)$ have a normal distribution with a constant variance σ^2 and they are independent, *i.e.* $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. As a result, the requirements in the linear model assume that the response variable must be continuous and normally distributed. These requirements are not valid in many cases such as in social science research where the outcome variable has dichotomous, ordinal or nominal outcomes. In such cases we can often use generalised linear models (GLMs).

The GLMs are more complicated than linear models. Generally speaking, there is not any closed form of the posterior distribution with these kinds of models so typically MCMC methods are used to sample from the posterior distributions.

3.5.2 Linear predictors and link functions

In the linear regression model, we have

$$\text{E}(Y_i) = \mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

Model	Link function $g(\theta)$	Error distribution
Linear regression	$\eta = \theta$	Normal distribution
Logistic regression	$\eta = \log\left(\frac{\theta}{1-\theta}\right)$	Binomial distribution
Probit regression	$\eta = \Phi^{-1}(\theta)$	Binomial distribution
Poisson regression	$\eta = \log(\theta)$	Poisson distribution
Complementary log-log	$\eta = \log\left[-\log(1-\theta)\right]$	Binomial distribution

Table 3.2: Most common link functions with corresponding with their generalised linear models (adapted from Lynch, 2007).

Now, we define and introduce the linear predictor as a linear combination of the model parameters β in the following form

$$\eta_i = g\left[\mathrm{E}(Y_i)\right] = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{3.11}$$

or, in matrix notation,

$$\eta = g\left[\mathrm{E}(Y)\right] = g(\mu) = X\beta.$$

As we can see, in the linear model $\mu_i = \eta_i$ whereas in a generalised linear model there is a link function that links the mean and the the linear function: $\eta_i = g(\mu_i)$ where $g(\cdot)$ is a known function called the *link function.* The link function must be monotonic and differentiable. We notice that (3.11) does not have an error term and that is because the expected value of Y_i is the linear predictor.

In GLMs there are two important features: the conditional distribution of the response variable Y, which need not be normal but it could be a member of the exponential family and the link function which relates the mean of Y to the linear predictor. See Dobson and Barnett (2008) or Faraway (2006).

Some common GLMs with their link functions are given in Table 3.2

3.6 Bayesian analysis for a logistic regression model

Suppose that we have the simple logistic regression model

$$\eta_i = g(x_i, \beta) = \beta_0 + \beta_1 x_i. \tag{3.12}$$

and that we use the logit link function, $\eta_i = \log[\theta_i/(1-\theta_i)]$, where

$$\theta(x_i) = \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right).$$

where $\theta(x_i)$ represents the probability of the event for subject i who has the covariate x_i. In the Bayesian framework, we should specify the prior density for the model parameters. If the prior distribution for β_0 and β_1 is a bivariate normal distribution which is a common selection (see for instance, Section 14.8 of Gelman et al., 2008) then our prior for the intercept and the slope can be represented as

$$\underline{\beta} = (\beta_0, \beta_1)' \sim N(\underline{m}, V).$$

Now, in order to apply Bayes' theorem, we need to determine the likelihood function for the parameters β_0, β_1 given the observed response variable y. So

$$\begin{aligned}
\mathrm{L}(\beta_0, \beta_1|\underline{y}) &= \prod_{i=1}^{n} \theta(x_i)^{y_i} \left[1-\theta(x_i)\right]^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i} \left[1-\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)\right]^{1-y_i}.
\end{aligned}$$

The quantity $\pi^{(1)}(\beta_0, \beta_1|\underline{y})$ is called the posterior density for the parameters given the data y which have been observed and can be written as

$$\pi^{(1)}(\beta_0, \beta_1|\underline{y}) = \frac{\pi^{(0)}(\beta_0, \beta_1)\mathrm{L}(\beta_0, \beta_1|y)}{f(y)} \tag{3.13}$$

where

$$f(y) = \int\int \pi^{(0)}(\beta_0, \beta_1)\mathrm{L}(\beta_0, \beta_1|y) d\beta_0 d\beta_1.$$

Unfortunately, the expression in (3.13) is very difficult to evaluate, in particular the integral in the denominator. So there are several methods which have been suggested to compute the posterior density. One of them is use the numerical methods of approximation such as a Laplace approximation or a quadrature method. Other methods that are widely used in Bayesian inference are Monte Carlo methods such as MCMC methods and the purpose of using those methods is to sample from the posterior distribution. See Section 3.4.

3.7 Variable selection methods

3.7.1 Introduction

In most studies, we look for the prediction of a response variable using covariates in order to explain the response. This relationship is unknown between the response and the covariates. In a Bayesian analysis, if we have no particular reason to wish to avoid using some covariates and we do not wish specifically to make inferences about the hypothesis that a covariate has no effect, there is not usually any need to consider removing covariates. This is particularly true when the sample size is large compared to the number of covariates (eg Hoeting et al., 1999). However, in some cases, for example when the sample size is small compared to the number of covariates or when costs associated with measurements or computations are important, we may wish to select only a subset of covariates to use.

In this thesis, we are interested in variable selection because of its relevance to structure learning in Bayesian networks. In the lung transplant example in Section 3.10 we have many covariates and address the question of whether we need all of these covariates or not. In this example, making a quick decision rather than waiting for all measurements to become available has some benefits as the lungs deteriorate over time.

3.7.2 Bayesian variable selection methods

In the last 20 years, we have seen many approaches to tackle the area of Bayesian variable selection (BVS). In this section we will concentrate on the principles of BVS methods. We might use selection approaches when we have some uncertainty about the statistical model. Assume that we have n models $M = (M_1, ..., M_n)$ for data Y. Under M_n, $Y \sim p(Y|\beta_n, M_n)$ where β_n is a vector of unknown parameters corresponding to the covariates in M_n.

Here, we need to use Bayesian techniques to assign a prior probability distribution $p(\beta_n|M_n)$ and $p(M_n)$ for each model. This specification can be recognised as a *three* stage hierarchical mixture model to generate the data Y. See Chipman et al. (2001).

1. Sampling the model M_j from $p(M_1), ..., p(M_n)$.
2. Sampling the parameter vector β_j from $p(\beta_j|M_j)$.
3. Sampling the data Y from $p(Y|\beta_j, M_j)$.

Hence the probability that the "true" model was really M_j, conditioning on having observed Y is the posterior model probability

$$p(M_j \mid Y) = \frac{p(M_j)p(Y \mid M_j)}{\sum_i p(M_i)p(Y \mid M_i)}.$$

We can write the marginal or (integrated) likelihood function of the model M_j averaged over the possible values of model parameters as follows:

$$p(Y \mid M_j) = \int p(Y \mid M_j, \beta_j) f(\beta_j \mid M_j) d\beta_j.$$

Depending on these posterior probabilities, the relative probability or posterior odds between two models M_1 and M_2 is

$$\frac{p(M_1 \mid Y)}{p(M_2 \mid Y)} = \frac{p(M_1)}{p(M_2)} . \frac{\int p(Y \mid M_1, \beta_{M_1}) p(\beta_{M_1} \mid M_1) d\beta}{\int p(Y \mid M_2, \beta_{M_2}) p(\beta_{M_2} \mid M_2) d\beta}. \tag{3.14}$$

So, (3.14) demonstrates how we can use the data Y through the Bayes factor

$$\frac{\int p(Y \mid M_1, \beta_{M_1}) p(\beta_{M_1} \mid M_1) d\beta}{\int p(Y \mid M_2, \beta_{M_2}) p(\beta_{M_2} \mid M_2) d\beta}$$

to update the prior odds $p(M_1)/p(M_2)$ to obtain the posterior odds. See Chipman et al. (2001).

In other words,

$$\text{posterior odds} = \text{prior odds} \times \text{Bayes factor}.$$

3.7.3 Bayesian variable selection using Zellner's g-prior

In variable selection, we should choose the prior distribution with care. In the linear regression model, $Y = X\beta + \epsilon$, we need to construct a family of priors in this model. So, Zellner (1986) proposed a particular prior that belongs to the conjugate normal-gamma family called a g-prior. In this prior

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad , \quad \beta \mid \sigma^2 \sim N\left(\beta_a, \frac{g}{\sigma^2}(X^T X)^{-1}\right). \tag{3.15}$$

where β_a is the prior mean of β and the prior variance−covariance matrix of β is a scalar multiple g of the Fisher information matrix and that depends on the observed data through the design matrix X.

3.7.4 Bayesian variable selection using reversible jump Markov chain Monte Carlo

A reversible jump Markov chain Monte Carlo algorithm in Bayesian variable selection can produce an MCMC chain which moves within the model space. See Green (1995). So, it moves from model M_n to model M_{n^*} by passing the choice of regression coefficients because the Metropolis-Hastings algorithm depends on $f(y \mid M_n)$. Therefore after selecting the model M_{n^*}, we need to sample the parameters β_{n^*} .To generalise this method, the joint sampling of the parameters β_n and the model M_n should be used with a Metropolis-Hastings approach. A proposal (β_{n^*},M_{n^*}) is generated from the proposal distribution $q(\beta_{n^*}, M_{n^*} \mid \beta_n, M_n)$ (the combination of values from the space of parameter vectors and model identifiers) given the current value (β_n,M_n). We accept the proposal with probability

$$\alpha = \min\left(1, \frac{f(\beta_{n^*}, M_{n^*} \mid y) q(\beta_n, M_n \mid \beta_{n^*}, M_{n^*})}{f(\beta_n, M_n \mid y) q(\beta_{n^*}, M_{n^*} \mid \beta_n, M_n)}\right)$$

$$= \min\left(1, \frac{f(y \mid \beta_{n^*}, M_{n^*}) f(\beta_{n^*} \mid M_{n^*}) f(M_{n^*}) q(\beta_n, M_n \mid \beta_{n^*}, M_{n^*})}{f(y \mid \beta_n, M_n) f(\beta_n \mid M_n) f(M_n) q(\beta_{n^*}, M_{n^*} \mid \beta_n, M_n)}\right)$$

where the second line follows from Bayes' theorem. The proposal is done in 2 steps with

Step 1: A proposal for M_{k^*}.

Step 2: A proposal for β_{k^*} which implies that
$q(\beta_{n^*}, M_{n^*} \mid \beta_n, M_n) = q(M_{n^*} \mid \beta_n, M_n) q(\beta_{n^*} \mid \beta_n, M_{n^*}, M_n)$.

The challenge is to ensure that the detailed balance condition holds which means that the move from model (β_k, M_k) to model (β_{k^*}, M_{k^*}) should be as easy as the opposite move. However, it is not immediately clear how to guarantee this condition when the dimensions of the model changes which occurs with variable selection. According to Chib and Greenberg (1995), reversibility is guaranteed by the reversible jump MCMC approach. See also Green (1995); Lesaffre and Lawson (2012).

3.7.5 Spike and slab priors

3.7.5.1 Introduction

A method that combines variable selection with inference for regression parameters, makes use of the variable selection priors known as *spike and slab priors*. These types of priors are defined as a mixture of two distributions, spike and slab distributions where the spike prior has a mass concentrated on zero and the slab prior has a possibly uniform distribution over the range of that prior. See Walli (2010). Mitchell and Beauchamp (1988) suggested these priors for BVS in normal linear regression models. They defined them as a mixture of a Dirac measure concentrated at zero and a uniform diffuse component. Therefore we can write the prior as

$$\pi(\beta_i \mid \delta_i) = \delta_i \pi_{slab}(\beta_i) + (1 - \delta_i)\pi_{spike}(\beta_i)$$

George and McCulloch (1993) proposed an alternative spike and slab prior that can be easily implemented in Gibbs sampler. This prior has the form

$$\beta_i \mid \delta_i \sim (1 - \delta_i)N(0, \sigma_i^2) + \delta_i N(0, c_i^2 \sigma_i^2)$$

where c_i is large, $c_i > 1$ and $\Pr(\delta_i = 1) = 1 - \Pr(\delta_i = 0) = p_i$. In subsection 3.7.5.2 we will illustrate two Bayesian variable selection (BVS) methods using spike and slab priors. We will assume that the intercept β_0 has a diffuse prior, $\beta_0 \sim N(0, \sigma_{\beta_0}^2)$ with $\sigma_{\beta_0}^2$ large.

3.7.5.2 Gibbs variable selection using spike and slab priors

Dellaportas et al. (2002) proposed another method for BVS based on a spike and slab prior. In their method which is called Gibbs variable selection (GVS), they introduced variable indicators in the model. That is, the linear predictor of the model is equal to

$$Y_i = \alpha + \sum_{k=1}^{d} \delta_k \beta_k x_k$$

where β_k are the regression coefficients. The joint density of y, β, δ is

$$f(y, \beta, \delta) = f(y, \beta_k, \beta_{(k)}, \delta) = f(y \mid \beta, \delta) f(\beta_k \mid \beta_{(k)}, \delta) f(\beta_{(k)} \mid \delta) f(\delta)$$

Here, $\beta_{(k)}$ is the vector of regression coefficients excluding β_k. After removing the constant term, the full conditional distribution for the regression parameters is $f(\beta_k \mid y, \beta_{(k)}, \delta) \propto f(y \mid \beta, \delta) f(\beta_k \mid \delta_k)$.

Kuo and Mallick (1998) assumed that the prior of β is independent of δ, so that $f(\beta_k \mid \beta_{(k)}, \delta) = f(\beta_k \mid \beta_{(k)})$. Now, $f(y \mid \beta, \delta)$ only contains β_k when $\delta_k = 1$. Removing the constant terms then yields two expressions:

$$f(\beta_k \mid y, \beta_{(k)}, \delta) = \begin{cases} f(y \mid \beta, \delta) f(\beta_k \mid \beta_{(k)}) & \text{if} \quad \delta_k = 1 \\ f(\beta_k \mid \beta_{(k)}) & \text{if} \quad \delta_k = 0 \end{cases}$$

For GVS, it is also assumed that β_k depends only on δ_k, i.e., $f(\beta_k \mid \beta_{(k)}, \delta) = f(\beta_k \mid \delta_k)$. Dellaportas et al. (2002) suggested taking

$$f(\beta_k \mid \delta_k) = (1 - \delta_k) N(\mu_{0k}, \tau_{0k}^2) + \delta_k N(\mu_{1k}, \tau_{1k}^2) \tag{3.16}$$

for suitable choices of $\tau_{0k}^2 < \tau_{1k}^2$. Notice that $f(y \mid \beta, \delta)$ contains β_k only when $\delta_k = 1$ and, combined with the prior in (3.16), the full conditional density for β_k becomes

$$f(\beta_k \mid y, \beta_{(k)}, \delta) = \begin{cases} f(y \mid \beta, \delta) N(\mu_{1k}, \tau_{1k}^2) & \text{if} \quad \delta_k = 1 \\ N(\mu_{0k}, \tau_{0k}^2) & \text{if} \quad \delta_k = 0 \end{cases}$$

where the distribution $N(\mu_{0k}, \tau_{0k}^2)$ is the prior when $f(\beta_k \mid \delta_k = 0)$ and is called a *pseudo-prior*, and the distribution $N(\mu_{1k}, \tau_{1k}^2)$ is the prior when $f(\beta_k \mid \delta_k = 1)$. Finally, the full conditional density for δ_k is Bernoulli with success probability $\pi_k/(1 + \pi_k)$ with the odds π_k equal to

$$\frac{f(\delta_k = 1 \mid \delta_{(k)}, \beta, y)}{f(\delta_k = 0 \mid \delta_{(k)}, \beta, y)} = \frac{f(y \mid \beta, \delta_k = 1, \delta_{(k)}) f(\beta \mid \delta_k = 1, \delta_{(k)}) f(\delta_k = 1 \mid \delta_{(k)})}{f(y \mid \beta, \delta_k = 0, \delta_{(k)}) f(\beta \mid \delta_k = 0, \delta_{(k)}) f(\delta_k = 0 \mid \delta_{(k)})}.$$

See Lesaffre and Lawson (2012).

3.7.5.3 Stochastic Search Variable Selection using spike and slab priors

The Stochastic Search Variable Selection (SSVS) method was suggested by George and McCulloch (1993) for variable selection in linear regression. The linear predictor is given

by

$$Y_i = \alpha + \sum_{k=1}^{d} \beta_k x_k$$

where $\beta_1, ..., \beta_d$ are assumed to have a mixture prior of spike and slab Gaussian components. The spike element is a normal distribution concentrated closely around zero, representing the real absence of the variable in the model. The slab component has a large variance to allow for the "nonzero" coefficients to spread over a larger range of values. This kind of separation is being regulated by two tuning parameters τ_k and c_k, where $\tau_k^2 > 0$ is the variance in the spike component and $c_k^2\tau_k^2 > 0$ is the variance in the slab component.

The components of the SSVS hierarchical prior are as follows:

$$\begin{aligned}
\beta_k \mid \sigma^2_{\beta_k} &\sim N(0, \sigma^2_{\beta_k}) \\
\sigma^2_{\beta_k} \mid \tau^2_{0k}, \tau^2_{1k}, \delta_k &\sim (1-\delta_k)\psi_{\tau^2_{0k}}(.) + \delta_k \psi_{\tau^2_{1k}}(.), \\
\sigma^2 &\sim IG(a_0, b_0), \\
\delta_k \mid \omega_k &\sim Bern(\omega_k), \\
\omega_k &\sim U(0,1).
\end{aligned}$$

with $a_0 = \nu_\delta/2$ and $b_0 = \frac{\nu_\delta \psi_\delta}{2}\ \delta_k$. Moreover, $\tau^2_{0k} = \tau^2_k$ and $\tau^2_{1k} = c^2_k\tau^2_k$. Here, δ_k indicates the component of the mixture (for $\delta_k = 0$ the k^{th} regressor is "practically zero"); $\psi_x(.)$ is the Kronecker delta concentrated at point x; ν_δ and ξ_δ possibly depend on δ and ω_k is the prior probability that β_k is nonzero. For details see Lesaffre and Lawson (2012). SSVS can be extended to GLMs without much difficulty (George et al., 1996).

3.8 Missing data

3.8.1 Introduction

In this section, we introduce some important definitions and assumptions that relate to incomplete data. There is extensive literature which tackles the problem of missing data from two points of view, the frequentist and Bayesian perspectives, such as Little and Rubin (2014); Daniels and Hogan (2008); Raghunathan (2015); Molenberghs and Kenward (2007).

In real life applications, we do not always observe all of the planned data. Hence, missing data are common in many applications. For example, suppose that we are aiming to do the analysis for data that have some covariates with missing values. In this case, reducing the number of cases by deleting those with missing values will not be an ideal thing to do as our inference might be adversely affected. In this thesis, we will use Bayesian inference to deal with the missing data problem.

In the case of completely observed data, it is not necessary to build the model for the covariates. However, when there are missing covariate values, it is important to build a model for the relationships among the covariates in the study. This is called a missing data model. See Zhao (2010). We denote the observed data as Y_{obs}, the missing values as Y_{miss}, so $Y = (Y_{\text{obs}}, Y_{\text{miss}})$ and θ is the parameter vector of interest.

3.8.2 Missing data mechanism

Suppose that we have the missing data indicator I which can take the value 1 if Y is observed and 0 otherwise. So the joint probability distribution of (Y, I) is

$$f(Y, I|\theta, \psi) = f(Y|\theta)f(I|Y, \psi).$$

The conditional probability distribution for I given Y and the unknown quantity ψ represents the missing data mechanism. In order to obtain the distribution of the observed data, we need to integrate out the distribution of Y_{miss} as follows

$$f(Y_{\text{obs}}, I|\theta, \psi) = \int f(Y_{\text{obs}}, Y_{\text{miss}}|\theta)f(I|Y_{\text{obs}}, Y_{\text{miss}}, \psi)dY_{\text{miss}}.$$

We can classify the missing data mechanism into different types of missingness termed *missing at random* (MAR), *missing completely at random* (MCAR) and *missing not at random* (MNAR). See Little and Rubin (2014); Tian et al. (2009).

3.8.3 Missing at random (MAR)

The missing data are said to be missing at random (MAR) if I is conditionally independent of the missing values given the observed values, that is, if

$$f(I|Y_{\text{obs}}, Y_{\text{miss}}, \psi) = f(I|Y_{\text{obs}}, \psi).$$

Then the likelihood function is

$$\begin{aligned} f(Y_{\text{obs}}, I|\theta, \psi) &= \int f(I|Y_{\text{obs}}, \psi) f(Y_{\text{obs}}, Y_{\text{miss}}|\theta) dY_{\text{miss}} \\ &= f(I|Y_{\text{obs}}, \psi) f(Y_{\text{obs}}|\theta). \end{aligned}$$

3.8.4 Missing completely at random (MCAR)

Sometimes we need to make a stronger assumption than MAR. We say that the missing data are missing completely at random (MCAR) if the distribution of I does not depend on either the missing or observed values, that is

$$f(I|Y_{\text{obs}}, Y_{\text{miss}}, \psi) = f(I|\psi).$$

Notice that, in Bayesian inference, it is not usually necessary to assume MCAR. It is usually sufficient to have MAR. The MAR assumption is more plausible when we observe a large number of variables since the observed values are then more likely to provide enough information to make missingness conditionally independent of the missing values.

3.8.5 Missing not at random (MNAR)

In the missing data mechanism, sometimes the MAR and MCAR conditions do not hold. Then the probability of some quantity being missing depends upon unobserved data. Therefore there is no simple way to represent the joint distribution. This type of mechanism is called missing not at random (MNAR). See Molenberghs and Verbeke (2005); Daniels and Hogan (2008).

3.8.6 Missing data and Bayesian inference

In Bayesian inference, missing data can be treated like other unknown quantities and we can integrate over their possible values to make inferences about model parameters. This is often conveniently done using MCMC methods. However, if the data are missing not at random (MNAR) then the data might not be informative about all model parameters.

3.9 Data augmentation (DA)

3.9.1 Introduction

There are some models that have intractable likelihood functions which lead to more difficult calculations. To make things more simple, we introduce extra variables which are called auxiliary variables. In fact, these variables are not observed but if they were observed that would make the likelihood more straightforward. Then we can treat these auxiliary variables as if they were missing data. As a result, we can define the data augmentation (DA) as the addition of unobserved auxiliary variables to the observed data.

Tanner and Wong (1987) proposed the use of the DA method in the Bayesian context in order to compute the posterior distribution. They introduce the term *"data augmentation"*. See Imai and Van Dyk (2005).

3.10 Lung transplant example

3.10.1 Introduction

As an example we consider some data on lung transplants. See Andreasson et al. (2016, 2017). A lung transplant is surgery to remove a person's diseased lung and replace it with a healthy lung from a deceased donor. Lung transplants are used for people who are likely to die from lung disease within 1 to 2 years. Lung transplants are not carried out frequently in the UK, mainly due to the lack of available donors. During 2013−2014 there were 198 lung transplants performed in England. See NHS (2016). In our data we have 30 covariates, with $\{X_1, ..., X_{15}\}$ representing measurements on the inflammatory proteins before ex vivo lung perfusion (EVLP) and $\{X_{16}, ..., X_{30}\}$ representing the measurements afterwards. Increasing the inflammatory proteins will deteriorate the lungs and as a result the lungs will no longer be used in transplant. The study concerned the use of EVLP. This is a technique for assessing and potentially reconditioning human donor lungs previously unacceptable for clinical transplantation with the potential to dramatically push the limits of organ acceptability. See Andreasson et al. (2014).

The response variable Y is a binary variable indicating whether or not the lung was

used.

Suppose that we adopt the following model. We have a number of lungs $n = 41$ and we have 30 covariates that might predict the use of the lungs. For illustration, we choose some of these covariates. See the `rjags` model specification in Appendix A.3.2.

Then we suppose that Y_i is an observation from the Bernoulli(θ_i) distribution. Let the probability that the lung is used be

$$\Pr(\text{lung } i \text{ used}) = \theta_i \quad \text{and} \quad \Pr(\text{lung } i \text{ not used}) = 1 - \theta_i.$$

Then we have the logistic model as follows:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \sum_{j=1}^{30} \beta_j(x_{ij} - \bar{x}_j) = \underline{x}_i^{'}\underline{\beta},$$

where $\underline{\beta} = (\beta_0, \beta_1, ..., \beta_{30})^{'}$ and $\underline{x}_i = (1, x_{i1} - \bar{x}_1, ..., x_{i30} - \bar{x}_{30})^{'}$, x_{ij} is the value of the covariate j for lung i and $\bar{x}_j = \sum_{i=1}^{n} x_{i,j}/n$. The reason why we subtract the mean is because it is easier to construct a prior by thinking about the middle of the covariates rather than one of the ends.

Therefore the likelihood function for $\theta = (\theta_1, ..., \theta_n)^{'}$ is

$$\mathrm{L}(\theta \mid \underline{y}) = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}.$$

We have

$$\frac{\theta_i}{1-\theta_i} = e^{\underline{x}_i^{'}\underline{\beta}} \quad \text{So} \quad \theta_i = \frac{e^{\underline{x}_i^{'}\underline{\beta}}}{1+e^{\underline{x}_i^{'}\underline{\beta}}} \quad \text{and} \quad (1-\theta_i) = (1+e^{\underline{x}_i^{'}\beta})^{-1}.$$

Therefore, the likelihood of β can be written as

$$\begin{aligned}\mathrm{L}(\beta \mid x, \underline{y}) &= \prod_{i=1}^{n}\left[\frac{e^{\underline{x}_i^{'}\underline{\beta}}}{1+e^{\underline{x}_i^{'}\underline{\beta}}}\right]^{y_i}\left(\frac{1}{1+e^{\underline{x}_i^{'}\underline{\beta}}}\right)^{1-y_i} \\ &= \frac{\prod_{i=1}^{n}\left[e^{\underline{x}_i^{'}\underline{\beta}}\right]^{y_i}}{\prod_{i=1}^{n}\left(1+e^{\underline{x}_i^{'}\underline{\beta}}\right)}\end{aligned}$$

parameters	mean	S.D.
β_0	-0.5	1.3
β_1	0	3.16
β_2	0	3.16
β_3	0	3.16
β_4	0	3.16
β_5	0	3.16
β_6	0	3.16
β_7	0	3.16
β_8	0	3.16

Table 3.3: The prior summaries for some regression coefficients

We give $\underline{\beta}$ a multivariate normal prior distribution

$$\underline{\beta} \sim N(\underline{\mu}, \Sigma)$$

where $\underline{\mu}$ is the prior mean vector and Σ is the prior variance and covariance matrix. We set the prior mean for β_0 for example, in the following way.

Suppose that we consider a lung where $x_{ij} = \bar{x}_j$ for $j = 1, ..., 30$. Suppose that we assess the probability that the lung will be used as P_0. Then we use the logit function to calculate the prior mean for β_0, we have

$$\mathrm{E}(\beta_0) = \log\left(\frac{\mathrm{P}_0}{1-\mathrm{P}_0}\right)$$

Now, to elicit the prior standard deviation, we can elicit assessments of the lower and upper quartiles for the proportion of such lungs which will be used. Let these be P_{01} and P_{03} respectively. Then

$$\log\left(\frac{\mathrm{P}_{03}}{1-\mathrm{P}_{03}}\right) - \log\left(\frac{\mathrm{P}_{01}}{1-\mathrm{P}_{01}}\right) = 1.35\sqrt{\mathrm{Var}(\beta_0)}$$

since $\Phi(1.35/2) = 0.75$. In this way we obtain $\mathrm{E}(\beta_0)$ and $\mathrm{Var}(\beta_0)$.

By considering a lung where $x_{i1} = x_1^*$ but $x_{ij} = \bar{x}_j$ for $j = 2, ..., 30$, we can, in a similar way, obtain a prior mean and variance for $\beta_0 + \beta_1(x_1^* - \bar{x}_1)$ and hence a prior mean for β_1 and, if we judge β_0 and β_1 to be independent a priori, a prior variance for β_1. We can assess prior means and variances for $\{\beta_2, ..., \beta_{30}\}$ similarly. The prior moments for $\beta_0, ..., \beta_8$ are given in Table 3.3.

In the example, some of the covariate values are missing, so we need a missing data model. For instance, the first covariate X_1 has some missing data. We adopt a normal model for it:

$$X_1 \sim N(m_1, V_1)$$

with prior distributions for

$$m_1 \sim N(m_{01}, V_{01}) \quad \text{and} \quad \frac{1}{V_1} \sim \text{Gamma}(c_1, d_1).$$

For the second covariate X_2, we also adopt a normal distribution, but the mean of this covariate depends on x_1 so that

$$X_2 \sim N(m_2, V_2)$$

where

$$m_2 = \beta_{02} + \beta_{21}(x_1 - m_1),$$

$$\beta_{02} \sim N(m_{02}, V_{02}) \quad , \quad \beta_{21} \sim N(m_{21}, V_{21}),$$

$$\text{and} \quad \frac{1}{V_2} \sim \text{Gamma}(c_2, d_2),$$

and so on. So we impose the order of the covariates in this example. In general, $X_j \sim N(m_j, V_j)$ where $m_j = \beta_{0,j} + \sum_{k=1}^{j-1} \beta_{j,k}(x_k - m_k)$ with $\beta_{0,j} \sim N(m_{0j}, V_{0j})$ and $\beta_{j,k} \sim N(m_{jk}, V_{jk})$.

3.10.2 Computing the posterior distribution in the lung transplant example

The `rjags` software is used to compute the posterior distribution. A burn in of 6000 samples and then an additional 100000 Gibbs samples were used. Table 3.4 gives the posterior summaries of the regression coefficient corresponding to the priors that were mentioned in Section 3.10.1. Plots of the densities for these priors and posterior are given in Figure 3.2.

Note that some of these posterior distributions are centred close to zero. So it may be that we can simplify the model and the corresponding Bayesian network by setting some of these coefficients to zero and omitting the corresponding arcs from the Bayesian

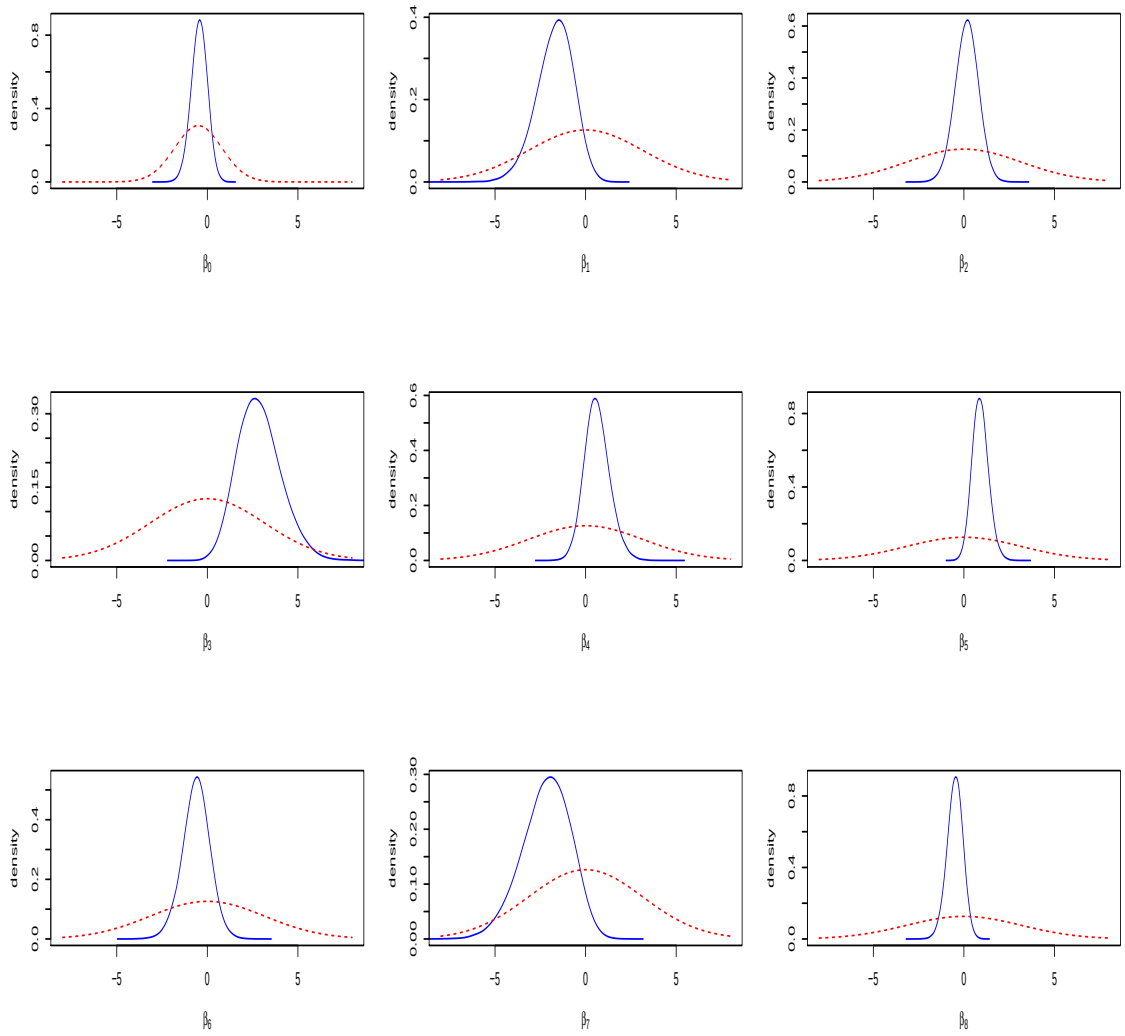


Figure 3.2: Posterior and prior densities for coefficients in the lung transplant example (dashed red: prior, solid blue: posterior).

parameters	mean	S.D.	95% credible interval
β_0	-0.444	0.451	(-1.346 , 0.458)
β_1	-1.693	1.023	(-3.739 , 0.353)
β_2	0.160	0.642	(-1.124 , 1.444)
β_3	2.875	1.202	(0.471 , 5.279)
β_4	0.631	0.699	(-0.767 , 2.029)
β_5	0.906	0.452	(-0.461 , 2.274)
β_6	-0.620	0.770	(-2.160 , 0.920)
β_7	-2.177	1.336	(-4.849 , 0.495)
β_8	-0.493	0.440	(-1.373 , 0.387)

Table 3.4: The posterior summaries for some regression coefficients

network.

The comparison of the prior and posterior standard deviation shows that all of the prior standard deviations were bigger than the corresponding posterior standard deviations.

In fact, in this example we have many covariates with few observations, so we could use the idea of variable selection. See Section 3.7. One possibility is that, before doing EVLP, we might decide to use the lung straight away, on the basis of $\{X_1, ..., X_{15}\}$ or a subset of them.

Therefore, we need to use variable selection here because we might not need all the covariates to decide whether to use the lung or not. There are two reasons why we prefer to use variable selection in this example. One of them is making the calculations of the probabilities easily and the other relates to time saved when we just measure some of the covariates rather than all of them. As the lungs deteriorate over time, it may be better to make a decision quickly rather than spend time making further observations.

3.10.3 Prior and posterior predictive distribution

The prior predictive distribution $f_0(y|x)$ is defined as the distribution of a new observation which is marginalised over the prior and can be written as

$$f_0(y \mid x) = \int f(y|\beta, x)\pi(\beta)d\beta.$$

where $\pi(\beta)$ is the prior density for β. The posterior predictive distribution $f_1(y|x)$ is

$$f_1(y \mid x) = \int f(y|\beta, x)\pi(\beta|\underline{y})d\beta.$$

where $\pi(\beta|\underline{y})$ is the posterior density for β.

The probability that the lung is used when $x = x^*$ is given as

$$\begin{aligned}\Pr(y^* = 1 \mid x^*) &= \mathrm{E}_{\beta|y,x}[\Pr(y^* = 1 \mid x^*, \beta)] \\ &= \mathrm{E}_{\beta|y,x}\left(\frac{e^{x^{*'}\beta}}{1 + e^{x^{*'}\beta}}\right).\end{aligned}$$

where $\underline{x}^* = (1, x_1^*, ..., x_{30}^*)'$ and $\underline{\beta} = (\beta_0, \beta_1, ..., \beta_{30})'$.

We can use an approximation method such as Monte Carlo integration in order to find the posterior predictive probability. The posterior predictive probability depends on the posterior distribution for β and the new observation of the covariates x_i^*.

Therefore, the posterior predictive probability is approximately

$$\Pr(y^* = 1 \mid x^*) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{e^{x^{*'}\beta^{(i)}}}{1 + e^{x^{*'}\beta^{(i)}}}\right)$$

where $\beta^{(1)}, ..., \beta^{(N)}$ are the samples drawn from the posterior distribution of β, $\pi(\beta \mid \underline{y}, x)$. We notice that our posterior predictive probabilities are close to the observed values of y. See Figure 3.3.

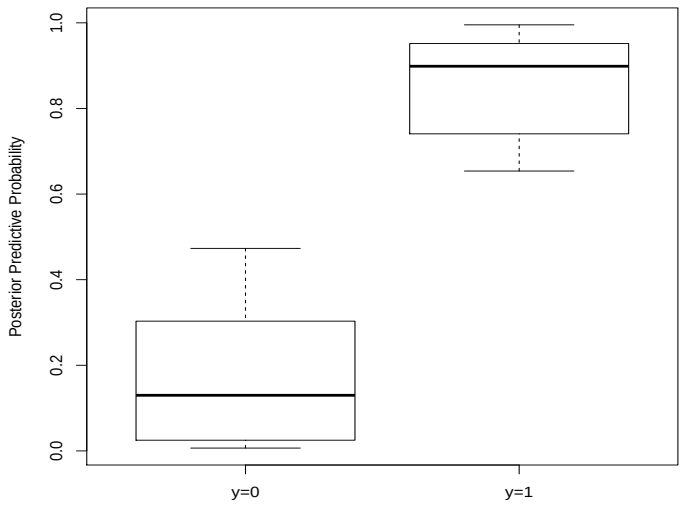


Figure 3.3: Boxplot of the posterior predictive probability that $Y = 1$ for the lung transplant example.

3.11 Summary

In this chapter, we have given a general description of Bayesian analysis with some important terms and definitions that relate to it. We have investigated some numerical integration methods such as the trapezoidal rule and Laplace approximation method as we need them to compute the posterior mean and variance. We have discussed in this chapter using different sorts of Monte Carlo integration, such as importance sampling, Gibbs sampling and Metropolis Hastings. We mentioned some statistics that we need to test the convergence of the samples in MCMC. We gave a motivational example to illustrate the idea of using a Gibbs sampling algorithm to generate realisations from the posterior distribution. We gave an introduction to GLM with the most common link functions such as logit and probit. We have given some background about Bayes analysis for logistic regression and some variable selection methods with different sorts of prior such as spike and slab prior. We also demonstrated some concepts about missing data and data augmentation because, in real life, we do not always observe all the observations from the experiments. We used a logistic regression model to fit the data in the lung transplant example. We obtained the posterior means and variances for all the coefficients in the model and compared them with the prior distribution that we elicited. We investigated the variable selection method from the Bayesian point of view for the lung example and we found that it is important to use some of the covariates in this example rather than all of them because in this case the lung deteriorates. We calculated the posterior predictive probability for the lung being used.

Chapter 4

Bayesian networks

4.1 Introduction

Probabilistic graphical models (PGMs) are graphical representations of the problem under research. See Pearl (1988); Lauritzen (1996); Koller and Friedman (2009). These models have been used in a variety of applications for decades because they can be useful to combine expert knowledge with the theory of probability. These models also have very important aspects. First of all, we can visualise these models in an attractive way. Secondly, these models also can represent complex problems in a simple graph. Finally, we can learn from the data and even construct a very large complex network because of the rapid development of computer software. One of the most familiar sorts of PGMs that we are interested in is called Bayesian networks (BNs). See Mateo Cerdán (2010).

This chapter introduces the methodology of Bayesian networks including methods for learning both the structure and parameters from data. We will tackle the possibility of using different methods for constructing Bayesian networks based on Markov Chain Monte Carlo (MCMC) methods. The idea is to obtain the most optimal configuration of the network and we review methods for choosing network structures.

We start in Section 4.2 by introducing some notation, definitions and important concepts in Bayesian Networks. In Section 4.3, we give the main point for comparison between regression models and BN models. There are two fundamental concepts in developing BNs from data: parameter learning and structure learning. So we give a brief demonstration about what they are in Section 4.4 and Section 4.5. We explain in detail two algorithms to learn about the structure of the network, Grow-Shrink and Hill-Climbing, using an `R`

package called "`bnlearn`". See Scutari and Ness (2012).

In this chapter, we also describe different types of BNs such as the categorical, Gaussian and hybrid and other types of Bayesian networks. In Section 4.11 we propose a method called *arc deletion* in order to choose the most likely configuration. This depends on an imposed ordering of the nodes. We apply this method to the non-Hodgkin lymphoma data set.

4.2 The methodology of Bayesian networks

In this section, we will give some important definitions and notation related to graphical models.

Bayesian networks (BNs) are very effective and flexible models to represent the probabilistic relationship between variables. Bayesian networks also have different names: Recursive graphical models, Bayesian belief networks, belief networks etc. The graph $G = (V, E)$ of a BN consists of a set of variables, nodes or vertices $V = \{X_1, X_2, ..., X_n\}$ and a set E of directed edges, arcs or arrows between these variables where the directed edges represent the dependence relations between the nodes. If there is an arc from X_i to X_j, then X_i is called a parent of X_j and X_j is called a child of X_i. If we define a path between two nodes in the network, say A and B (*i.e.* $A \to B$), then the following path is not allowed in BN, $A \to B \to A$ because it is cyclic graph. As a result, we refer to a directed acyclic graph (DAG) as one representation of a BN. Therefore, we can write the joint probability distribution function for all the variables as follows:

$$\Pr(X_1, ..., X_n) = \prod_{i=1}^{n} \Pr(X_i \mid \text{parents}(\text{X}_\text{i})) \tag{4.1}$$

where parents(X_i) is the set of parents of X_i.

In order to construct a network we need to proceed in two stages. At the *qualitative level*, we need to choose the graph (G) of the BN which satisfies the relationship between the variables in terms of conditional dependence and independence relations. At the *quantitative level*, the local probabilities are determined for the marginal distributions at root nodes (a root node is defined as a node which has no parents) and the conditional probability distributions for the other variables. As a result we can calculate the joint probability distribution (global distribution) as a product of all these marginal and con-

ditional distributions in the model. For more information, see Jensen (1996); Neapolitan (2003); Korb and Nicholson (2004) and Scutari and Denis (2014).

Other important terms in BNs are ascendant and descendant. To explain the relationship between the two concepts, suppose we have three variables X_1, X_2 and X_3. Then, if there is a directed path from X_1 to X_2 and there is a directed path from X_2 to X_3, we called X_1 an ascendant of X_2 and X_3 a descendant of X_2.

As a result, in a BN, each variable is conditionally independent of its non-descendants given its parents.

4.2.1 Causality in Bayesian networks

In this subsection, we give an explanation of a causal model in Bayesian networks. The main feature of BNs is the directed edges between the vertices. Some networks can represent *cause and effect* relationships between the nodes. So the causal model is defined as a set of vertices K of a directed acyclic graph (DAG), where each vertex in the graph matches to a different element of K, such as the example shown in Figure 4.1. In some cases the relationship can be interpreted as a cause and effect relationship, where the nodes refer to the variables in the model and the links indicate the direct causal influence between the vertices. See Pearl and Verma (1995).

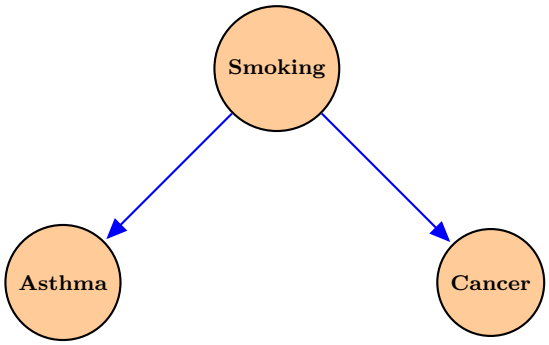


Figure 4.1: Causal network example.

From the network in Figure 4.1 we can infer that it could be that Asthma and Cancer are the effect of a patient's smoking or we can say that Smoking causes Cancer and Asthma. It is sometimes more complicated to construct BNs with causal relationships because we have to give the right meaning to the relations. Otherwise it does not make sense from the scientific or medical point of view. See Margaritis (2003).

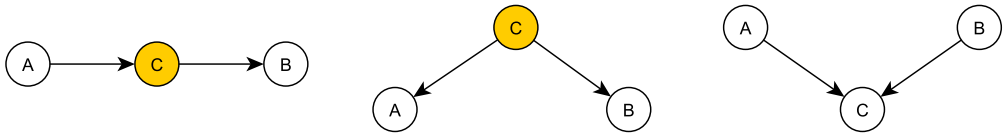


Figure 4.2: D-separation (directed acyclic graph). Left: serial connection, Middle: diverging connection, Right: converging connection.

4.2.2 D-separation

There are different ways to represent the relationships between the variables in a BN. Some variables are directly related to each other and the rest are indirectly related with other nodes in the network. Pearl (2000) introduces an important definition which is related to a DAG called "d-separation" or *directed separation*. It can be defined as follows. Suppose that we have two nodes such as A and B in a causal network, then for all the paths between A and B, there is a middle node C which separates them in the following cases,

- if there is a serial connection or diverging connection when we observed the intermediate node C. Or
- converging connection when there is no evidence about C.

See Jensen (1996); Korb and Nicholson (2004).

In order to explain Pearl's idea of separation in a DAG, let us take the following simple example. Suppose that we have three disjoint sets of nodes, for example, A, B and C in a DAG (G). Then C is said to d-separate A from B, if $A \perp\!\!\!\perp B \mid C$, *i.e.* $\Pr(A, B \mid C) = \Pr(A \mid C)\Pr(B \mid C)$, where $A \perp\!\!\!\perp B \mid C$ means that A and B are conditionally independent given C and $\Pr(A \mid C)$ is the conditional probability of A given C and no evidence for its descendants.

We can represent the d-separation model in the DAGs in Figure 4.2.

The DAG on the left of Figure 4.2 can be represented mathematically as follows

$$\Pr(A, B|C) = \frac{\Pr(A, B, C)}{\Pr(C)} = \frac{\Pr(A)\Pr(C|A)\Pr(B|C)}{\Pr(C)} = \Pr(A|C)\Pr(B|C).$$

since, $\Pr(A|C) = \frac{\Pr(A)\Pr(C|A)}{\Pr(C)}$. So, C separates A and B.

Likewise, we can check if the node C separates A and B in the DAG in the middle of

Figure 4.2 in the following expression

$$\Pr(A, B|C) = \frac{\Pr(A, B, C)}{\Pr(C)} = \frac{\Pr(C)\Pr(A|C)\Pr(B|C)}{\Pr(\mathrm{C})} = \Pr(A|C)\Pr(B|C).$$

So, again, C separates A and B.

However, in the third case, on the right of Figure 4.2, there is no such simplification and C does not separate A and B.

4.2.3 Markov blanket

A Markov blanket is defined as a set of nodes that separates a target node from the rest of the nodes in the network which includes its parents, its children and other nodes sharing a child. Figure 4.3 represents an example of a Markov blanket of node E when the nodes {C,D} are the parents of the node E and the nodes {G,H} are the children of it and the node F is the children's other parent. The rest of the nodes such as {A,B,I} are conditionally independent of E given the blanket $\{C, D, F, G, H\}$. See Scutari and Denis (2014) and Kułaga (2006).

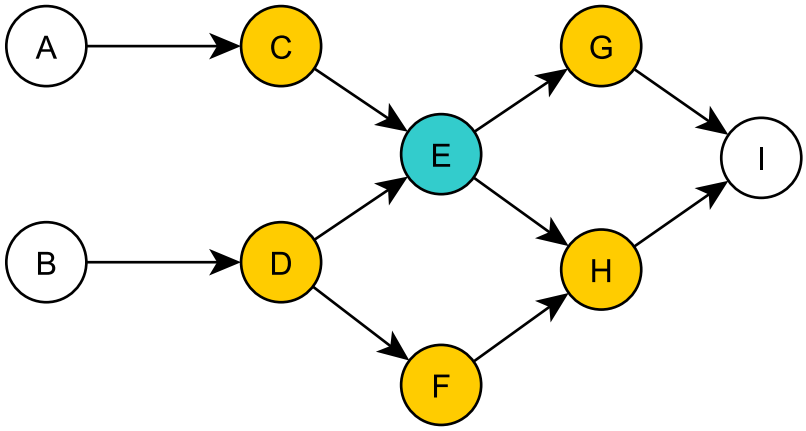


Figure 4.3: Markov blanket of node E.

4.3 Comparison of Bayesian networks with regression models

The point of a BN is that, unlike a standard regression model where we do not specify a distribution for the covariates, only a conditional distribution for the dependent variable given the covariates, in a BN we specify the joint distribution of all variables so that we can use it even when some are not observed. Specifying this full joint distribution allows us to make predictions when only some variables are observed. There are some studies which show that using BN models is preferable and gives an accurate results compared with other regression models. See Witteveen et al. (2018); Gevaert et al. (2006). According to Sesen et al. (2013), they illustrated that BNs are an efficient tool in survival studies and able to predict the survival time for the patient more precisely.

4.4 Bayesian network parameter learning

4.4.1 Introduction

In constructing BNs, we often wish to learn about the values of parameters from data. If we have a given DAG, G with P variables defined, then inference about the parameter θ when we have observed some data $D = \{x_1, x_2, ..., x_P\}$ is called parameter learning. In the case where all of the nodes in the BN are discrete or categorical variables with a finite number of possible values, then we usually use the conditional probability table (CPT) and this type of BN is called a multinomial Bayesian network. The main task here is to make inference about all the values in the CPT when we have given a certain structure. If we have the case where the CPT is unknown, then we can learn from the observed data in order to produce the CPT. There are many algorithms that deal with parameter learning in the case that we have complete data and the case when we observe one or some of the variables in the data sets. See Ji et al. (2015).

We can learn about the parameters using either frequentist or Bayesian techniques. Most of the material in this thesis deals with Bayesian methods. Therefore, we focus in this chapter on parameter learning from the Bayesian point of view.

4.4.2 Parameter learning with complete data set

In the case of learning about the parameters in a BN when we have complete data, there is extensive literature. As a simple case suppose that we have a multinomial Bayesian network. Suppose that we need to determine our uncertainty about the parameters $\theta = \{\theta_1, \theta_2, ..., \theta_P\}$ of a multinomial probability distributions using Bayesian inference, such that $\sum_{p=1}^{P} \theta_p = 1$. Therefore, the likelihood function in this case will be

$$\mathrm{L}(\theta \mid D) = \prod_p \theta_p^{n_p}$$

where $n_1, ..., n_p$ are non-negative integers such that $\sum_{p=1}^{P} n_p = n$. We need to choose a suitable prior distribution $\pi(\theta)$. The most appropriate and conjugate prior to use is a Dirichlet prior which can be described by a set of hyperparameters $\alpha_1, \alpha_2, ..., \alpha_P$, so that

$$\pi(\theta) \propto \prod_p \theta_p^{\alpha_p - 1}.$$

As this is a conjugate prior, the posterior distribution will also be a Dirichlet distribution with the hyperparameters $\{n_1 + \alpha_1, n_2 + \alpha_2, ..., n_P + \alpha_P\}$ as follows

$$\pi(\theta \mid D) \propto \prod_p \theta_p^{n_p + \alpha_p - 1}.$$

See Buntine (1991) and Koller and Friedman (2009).

4.4.3 Parameter learning with incomplete data set

We can learn about the parameter values from the data after we specified a suitable network and that can be done by computing the conditional probability distributions.

Many methods in statistical data analysis require complete data in order to obtain the results. However, in real life, this condition does not always hold. Therefore, structure learning of BN and parameter learning by some methods in this case is analytically intractable. See Riggelsen (2006). However, MCMC methods may be used.

We will focus on using MCMC methods to learn about the parameters in different models with various types of missingness.

4.5 Bayesian network structure learning

4.5.1 Introduction

The main idea behind a Bayesian network structure is the way that we can summarise the conditional independence relationships among the variables graphically. In addition to independencies, a BN structure can sometimes be interpreted as cause and effect relations through the direction of edges. See Section 4.2.1. So, the parent node represents the "direct cause" and the child node represents the "effect node". See Margaritis (2003).

In some cases, the structure of a BN is chosen subjectively using expert opinion. However, in other cases, we may wish to choose a suitable structure based on analysis of data. In addition, learning a network may also take into account the prior information about the independencies of the variables in the problem (for instance, obtained from research or accumulated knowledge).

In the following sections, we explain in detail how we can develop Bayesian network structure and learn about it from the data. The determination of BN structure is one of the most challenging problems that people need to investigate. For example, some causal BN can not easily be determined by experts and the structure and variables might be changed when we add new data. We will use different types of algorithms in this chapter in order to construct Bayesian networks. However, we concentrate on finding BN structure using some Bayesian approaches and numerical techniques such as MCMC.

We will consider algorithms for constructing Bayesian network structure. For instance, determining network structure is a combination of imposing an ordering of the nodes and subsequent arc deletion. Bayesian methods such as Markov Chain Monte Carlo schemes (MCMC) are used to pick the most likely configuration.

Some authors describe the selection of the network structure with greatest posterior probability. Heckerman and Chickering (1995) describe the combination of prior knowledge from experts with observed data in order to construct a BN with higher posterior probability.

Suppose we have given been given data and we are aiming to find the posterior probability distribution for the structure of the network S and suppose that S^* is the most

likely structure that is supported by data, then

$$S^* = \text{argmax}_S \{\Pr(S \mid D)\} \tag{4.2}$$

where S^* is the best structure and D is the data. We apply Bayes' theorem to (4.2) in order to find the posterior distribution for S,

$$\Pr(S \mid D) \propto \Pr(D \mid S)\Pr(S)$$

where the likelihood function $\Pr(D \mid S)$ can be evaluated by integrating out θ, as in the following expression

$$\Pr(D \mid S) = \int_\Theta \Pr(D \mid \theta, S)\Pr(\theta \mid S)d\theta. \tag{4.3}$$

For more details, see Husmeier et al. (2005).

4.5.2 Inferring causality

It is sometimes of interest to make inferences about whether one or more variables cause other variables in the sense that a change in one or more variables brings about a change in one or more other variables. Such a relationship is, of course, a stronger statement than association or probabilistic dependence. A long and widely held view is that causality can only be inferred from the results of controlled experiments in which the values of some variables are deliberately changed and the changes in other variables are measured. More recently the availability of network structure learning algorithms has led to the idea that their use might allow inference about causality from observational data. However, such an inference depends on being sure that there are no unobserved , hidden, variables which might account for observed associations. See, for example, Section 7.1.4 of Jensen and Nielsen (2007).

4.6 Bayesian networks for categorical variables

4.6.1 Introduction

As we know there are various types of variables that might be included in Bayesian networks. For example, we have binary variables which can only take two values such as

the variable sex can only take male or female. Categorical variables can take more than two values. We can divide categorical variables into two groups, ordinal variables and nominal variables. The ordinal variables are the variables that we can describe in ordered categories such as patient's condition (excellent, good, fair, poor). Nominal variables are classified as unorderd categories such as eye colours (brown, blue, black, green). So, constructing a BN for categorical variables requires that all the variables should contain categorical data and the network is then described as a categorical BN. This discrete BN can take the form of a conditional probability table (CPT).

4.6.2 Motivational example for categorical Bayesian network

Suppose we have the following Bayes network containing four nodes which is adopted from Jensen (1996). They are Cloudy (C), Sprinkler (S), Rain (R) and Wet grass (W). All four nodes are binary with two possibilities "TRUE" or "FALSE". Suppose that we have the structure shown in Figure 4.4.

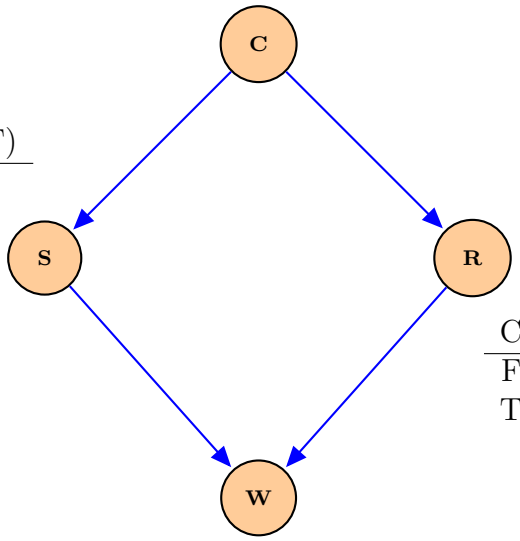
Therefore, the joint probability distribution for the 4 variables in the network will be

$$\Pr(\mathrm{C,S,R,W}) = \Pr(\mathrm{C})\Pr(\mathrm{S} \mid \mathrm{C})\Pr(\mathrm{R} \mid \mathrm{C})\Pr(\mathrm{W} \mid \mathrm{R,S}).$$

So the probability that it is cloudy is 0.5. That is $\Pr(\mathrm{C}) = 0.5$ and $\Pr(\text{not C}) = 1 - \Pr(\mathrm{C}) = 0.5$. As we use a directed acyclic graph, we know that the two nodes Rain and Sprinkler depend on whether it was cloudy or not. So, the node Rain has four possible conditional probabilities which are

$$\begin{aligned} \Pr(\mathrm{R} \mid \mathrm{C}) &= 0.8, \\ \Pr(\mathrm{R} \mid \neg\mathrm{C}) &= 0.2, \\ \Pr(\neg\mathrm{R} \mid \mathrm{C}) &= 0.2, \\ \Pr(\neg\mathrm{R} \mid \neg\mathrm{C}) &= 0.8. \end{aligned}$$

Likewise, the node Sprinkler also has four possible conditional probabilities as it depends



P(C = F)	P(C = T)
0.5	0.5

C	P(S = F)	P(S = T)
F	0.5	0.5
T	0.9	0.1

C	P(R = F)	P(R = T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W = F)	P(W = T)
F	F	0.8	0.2
T	F	0.2	0.8
F	T	0.2	0.8
T	T	0.2	0.8

Figure 4.4: A simple Bayesian network, adapted from Jensen (1996).

on the node Cloudy as well and we can write these probabilities as

$$\begin{aligned}\Pr(\mathrm{S} \mid \mathrm{C}) &= 0.1,\\ \Pr(\mathrm{S} \mid \neg\mathrm{C}) &= 0.5,\\ \Pr(\neg\mathrm{S} \mid \mathrm{C}) &= 0.9,\\ \Pr(\neg\mathrm{S} \mid \neg\mathrm{C}) &= 0.5.\end{aligned}$$

The last node in our network is Wet grass and because this node depends on both Rain and Sprinkler, this node has 8 potential conditional probabilities which are shown in Figure 4.4.

4.7 An introduction to the `R` package "`bnlearn`"

4.7.1 Introduction

The name of the `R` package "`bnlearn`" (Scutari, 2010) is an abbreviation of "**B**ayesian **n**etwork **learn**ing". It is a `R` package that is used to learn about the structure of Bayesian networks, estimate the parameters from frequentist and Bayesian perspectives, and make some inference about the unknown quantities. There are many different algorithms that we can use to learn about the structure. We will use one of the constraint-based structure learning algorithms called the "Grow-Shrink (GS)" algorithm. We also use one of the score-based structure learning algorithms called the "Hill Climbing (HC)" algorithm. See Scutari (2010). In the following sections, we will give more details about each algorithm with an illustrated example.

4.7.2 Grow-Shrink algorithm (GS) in bnlearn package

Margaritis (2003) suggested an algorithm called Grow-Shrink (GS) which depends on the Grow-Shrink Markov blanket algorithm which we can describe as a simple Markov blanket detection algorithm to learn about the structure of a Bayesian network. The main idea of this algorithm is based on finding the structure for each Markov blanket say for example, MB(b) in the network, where the node $b \in V$. Then for each b the GS algorithm works to determine MB(b) in two phases: the grow phase and shrink phase. See Edera et al. (2014).

Algorithm 5: Grow-Shrink Algorithm, adapted from Margaritis (2003).

1 $\Omega \leftarrow \emptyset$.
2 While $\exists\ F \in \mathcal{Z} - \{E\}$ such that $F \not\perp E \mid \Omega$, do $\Omega \leftarrow \Omega \cup \{F\}$. [Growing phase]
3 While $\exists\ F \in \Omega$ such that $F \perp E \mid \Omega - \{F\}$, do $\Omega \leftarrow \Omega - \{F\}$. [Shrinking phase]
4 $\mathbf{B}(E) \leftarrow \Omega$.

In **Algorithm 5**, we start the growing phase with the empty set Ω. Then we add variables to Ω unless they are dependent on E given the current contents of Ω. In this case, we might add some variables which are actually outside the blanket which can be identified and removed from the Bayesian network at the shrinking phase. See Margaritis (2003).

4.7.3 Hill-Climbing algorithm (HC) in bnlearn package

This algorithm is one of the score-based structure algorithms which simply learn about the structure of a BN based on heuristic optimisation methods. We can assign a network score for each selected BN that reflects its goodness of fit and then the algorithm attempts to maximise this score. See Scutari (2010). An example of this type of algorithm is a *greedy search* algorithm such as the *Hill-Climbing* algorithm. The algorithm usually starts with no arcs. At each iteration we can add an arc, delete an arc or reverse an arc provided that we do not create a directed cycle. The most common score which is used in this algorithm is the Bayesian Information Criterion (BIC).

Then we choose the structure that gives us the highest values of the score function. The algorithm ends when no further increase can be made.

Algorithm 6: Hill-Climbing algorithm, adapted from Scutari (2010).

1 Select a network structure N which is usually empty but that is not necessary.
2 Calculate the score of N, defined as $Score_N = Score(N)$.
3 Put $maxscore = Score(N)$.
4 If $maxscore$ increases, repeat the following steps
5 (a) for every possible arc addition, deletion or reversal not resulting in a cyclic network:
6 (i) Calculate the score of the modified network N^*, $Score_{N^*} = Score(N^*)$.
7 (ii) If $Score_{N^*} > Score_N$, set $N = N^*$ and $Score_N = Score_{N^*}$.
8 (b) Set the new value of $Score_N$ in order to update $maxscore$.
9 Return the DAG N.

			Accommodation type					
			apartment			house		
			age			age		
work	tenure	response	<30	31-45	>45	<30	31-45	>45
skilled	rent	yes	18	15	6	34	10	2
		no	15	13	9	28	4	6
	own	yes	5	3	1	56	56	35
		no	1	1	1	12	21	8
unskilled	rent	yes	17	10	15	29	3	7
		no	34	17	19	44	13	16
	own	yes	2	0	3	23	52	49
		no	3	2	0	9	31	51
office	rent	yes	30	23	21	22	13	11
		no	25	19	40	25	16	12
	own	yes	8	5	1	54	191	102
		no	4	2	2	19	76	61

Table 4.1: Danish do-it-yourself

Another possibility is to use the Laplace approximation score function in order to compute the posterior distribution for the parameters in the model structure. Using a Laplace approximation can provide us with more efficient results but it is less accurate as it uses approximate integrations. See Needham et al. (2007); Chickering and Heckerman (1997).

Clearly the Hill-Climbing algorithm could be adopted to use other scores such as the expectation of a utility function.

4.7.4 Motivational example

In this example, the data represent a sample of employed men in Denmark aged between 18 and 67. See (Hand et al., 1994). They were asked a question about whether they had carried out work on their home. The response variable is yes or no, with four categorical explanatory variables: Age: under 30, 31-45 and over 45, Accommodation type: apartment or house, Tenure: rent or own, Work of respondent: skilled, unskilled, office. These data are represented in Table 4.1.

We use the Grow-Shrink and Hill-Climbing algorithms with the BIC score function to construct a Bayesian network for this example. The resulting DAGs are shown in Figures 4.5 and 4.6.

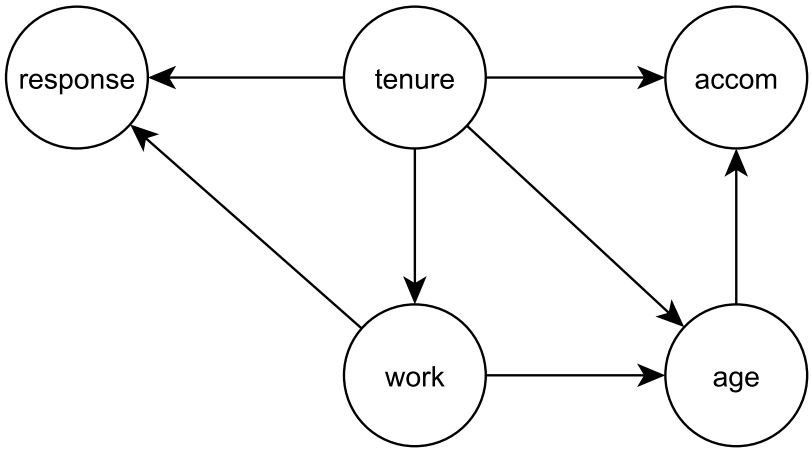


Figure 4.5: Bayesian network structure learning based on Grow-Shrink algorithm.

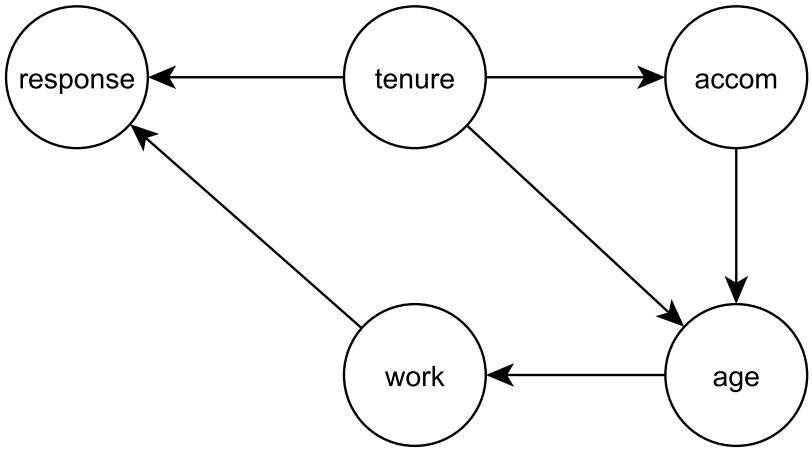


Figure 4.6: Bayesian network structure learning based on Hill-Climbing algorithm.

4.8 Bayesian networks for Gaussian variables

4.8.1 Learning the parameters in Gaussian Bayesian network

In this section, we discuss cases where we do not have a multinomial distribution network. For instance, suppose we have a network where all the nodes are continuous random variables and a multivariate normal distribution is considered to be appropriate. Let us explain the idea how we can model a simple Bayesian network with three nodes A, B and C. Suppose we have decided to use the network in Figure 4.7.

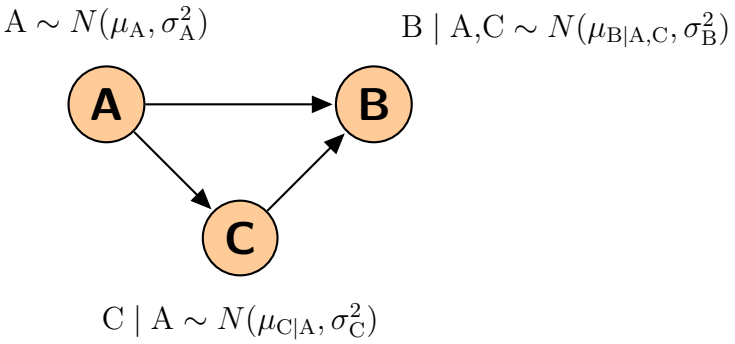


Figure 4.7: Gaussian Bayesian network for three variables.

Then, we have $\mathrm{E(A)} = \mu_\mathrm{A} = \beta_{0,\mathrm{A}}$, since $\mathrm{A} = \beta_{0,\mathrm{A}} + \varepsilon_\mathrm{A}$ and $\mathrm{Var(A)} = \mathrm{Var}(\varepsilon_\mathrm{A}) = \sigma_\mathrm{A}^2$. Similarly, the mean and variance for the node C given that we have observed A are

$$\mathrm{E(C \mid A)} = \mu_{\mathrm{C|A}} = \beta_{0,\mathrm{C}} + \beta_{\mathrm{A,C}}(\mathrm{A} - \mu_\mathrm{A}).$$

and

$$\mathrm{Var(C \mid A)} = \sigma_{\mathrm{C|A}}^2 = \beta_{\mathrm{A,C}}^2\sigma_\mathrm{A}^2 + \sigma_\mathrm{C}^2.$$

and so on for the node B. In order to learn about the parameters in this network, we need to make inference about the coefficients $\beta_{0,\mathrm{A}}, \beta_{0,\mathrm{B}}, \beta_{0,\mathrm{C}}, \beta_{\mathrm{A,C}}, \beta_{\mathrm{A,B}}, \beta_{\mathrm{C,B}}$ and the conditional variances $\sigma_\mathrm{A}^2, \sigma_\mathrm{B}^2, \sigma_\mathrm{C}^2$ in this simple model. We could use MCMC methods to fit this model and evaluate the posterior mean and variance where we can use a multivariate normal prior distribution for all $\underline{\beta} = (\beta_{0,\mathrm{A}}, \beta_{0,\mathrm{B}}, \beta_{0,\mathrm{C}}, \beta_{\mathrm{A,C}}, \beta_{\mathrm{A,B}}, \beta_{\mathrm{C,B}})'$ and writing $\tau_G = 1/\sigma_G^2$ for $G = \mathrm{A,B,C}$, we give τ_G a gamma prior distribution with some parameters α_G and λ_G.

4.9 Other sorts of Bayesian networks

We have discussed simple multinomial and Gaussian networks represented by directed graphs. There are other kinds of Bayesian networks. Here we briefly discuss some of the most common types.

4.9.1 Hybrid Bayesian networks

In BNs, if we combine discrete variables, continuous variables and any other type of variables such as interval censored variables in the network, such a network is called a hybrid BN. Unfortunately, learning about the structure and the inference about the parameters in hybrid BN needs special methods as the inferential problems for these networks are less tractable. See Scutari and Denis (2014).

4.9.2 Dynamic Bayesian network models

The main difference between BN models and dynamic Bayesian network (DBN) models is that the latter incorporate time series structure. We suppose in DBN that each time slice is dependent on the previous time. There is extensive literature on DBN such as Murphy and Russell (2002); Russell and Norvig (2016).

4.9.3 Influence diagrams

An influence diagram is a Bayesian network that includes decision nodes and a node representing a value or utility function. See Kjaerulff and Madsen (2005). So we can select the optimal decision based on maximising the expected utility function. The random variables in an influence diagram can be represented with circles while the decision nodes are represented with rectangles. See Howard and Matheson (2005).

4.9.4 Chain graphs

This type of graph is a mixture of directed and undirected networks, where the directed networks represent the Bayesian networks and the undirected graphs represent Markov networks. See Lauritzen and Wermuth (1989); Studenỳ (1998); Buntine (1995). Graphs

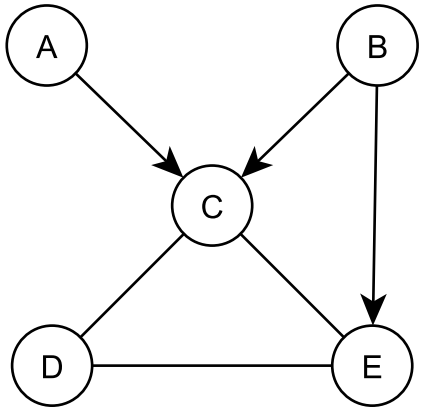


Figure 4.8: Chain graph for 5 variables {A,B,C,D,E}.

of this type represent more complex probability distributions. Figure 4.8 represents an example of a chain graph with 5 variables in the network. The components of this chain are {A},{B} and {C,D,E}.

4.10 Information propagation in Bayesian networks

In this section, we describe briefly the problem of making inferences about the unknown quantities in Bayesian networks using algorithms such as information propagation and the Lauritzen-Spiegelhalter algorithm. See Lauritzen and Spiegelhalter (1988).

The Lauritzen-Spiegelhalter algorithm for a categorical network exploits the structure of the network in computing the marginal distributions. The idea of this algorithm is that the we have a structured joint tree which is basically a join tree. Then the information is propagated through this joint tree. See Lepar and Shenoy (1998).

Now, for the Gaussian networks, all the variables in the network have Gaussian distributions. Therefore, the joint probability distribution for all of these variables is a multivariate normal distribution. See Section 4.8 for an example of a Gaussian network. We can write the joint probability distribution in a Gaussian network as a product of the conditional distributions when each conditional density is independent normal as follows

$$f(x_k|x_1, ..., x_{k-1}) \sim N(\mu_k, \tau_k)$$

where $\mu_k = m_k + \sum_{j=1}^{k-1} \beta_{jk}(x_j - \bar{x}_j)$. Notice that m_k refers to the unconditional mean

of x_k and $\tau_k = 1/V_k$, where V_k is the conditional variance of x_k given we have observed $x_1, ..., x_{k-1}$. See Geiger and Heckerman (1994).

However, in more complicated cases, simple tractable methods are not usually available. We can either use computationally intensive numerical methods or approximations. See Needham et al. (2007); Wilkinson (2007).

In this thesis, Chapter 7, we propose a new method for such networks. This network is called a Bayes linear Bayes prognostic network.

4.11 Proposed technique to construct a Bayesian network

The main idea for this method is to construct a BN using arc deletion and an imposed ordering of the nodes.

We use as an example, a subset of the variables in the non-Hodgkin lymphoma data. So we are starting our model by assuming that the observational covariates such as Sex and Age are independent and the lifetime distribution T depends on both of them.

The assumption of independence is acceptable since Sex and Age are always observed so that we are always conditioning on both of them.

Bayesian methods such as Markov Chain Monte Carlo (MCMC) schemes are used to pick the most likely configuration. The algorithm starts with:

Step 1: Fit the life time distribution which in this case the Weibull distribution.

Step 2: We then introduce the **presence indicator** I with $I = 1$ if the arc is present and 0 otherwise.

Step 3: We have the *product* of those indicators. E.g.

$$I_A(1 - I_B)I_C$$

where A and C are present, B is absent. In general, for a network with N nodes, the indicator for a configuration C is $I_c = \prod_{i=1}^{n} I_i^{I_i}(1 - I_i)^{1-I_i}$ where I_i is the indicator for node i.

Step 4: Finally we calculate the posterior mean of the indicators which is the posterior probability that those coefficients are **non-zero.** We also calculate the posterior mean of a product of indicators which is the posterior probability of the corresponding network structure.

See Appendix A.4.1 and Appendix A.4.2 for `rjags` specification and `R` code to apply this algorithm.

4.11.1 Example: non-Hodgkin lymphoma

We apply this method to the non-Hodgkin lymphoma example, with the possibility of fitting all the covariates at once and we find the most likely configuration as follows.

For illustration, we use six variables in the non-Hodgkin lymphoma example which are Age, Sex, T, Wbc, Hb and Albumin. As we can see from the original BN in Figure 4.9, we have 14 edges among all the nodes. After applying the MCMC approach, we obtain a BN which is shown in Figure 4.10 which has 8 edges representing the relationships between the nodes. This is because the posterior probabilities of some of the coefficients are very close to zero. As a result, we dropped some of the edges.

We suppose that the log of survival lifetime has a normal distribution with

$$\log(T_i) \mid x_{i,1}, x_{i,2} \sim N(\mu_i, \sigma^2)$$

where

$$\mu_i = \beta_{0t} + \beta_{x_1,t} x_{1,t} + \beta_{x_2,t} x_{2,t}$$

and

$$\begin{aligned} \beta_{0t} &\sim N(\mu_{0t}, \sigma^2_{0t}) \\ \beta_{x_1,t} &\sim N(\mu_{1t}, \sigma^2_{1t}) \\ \beta_{x_2,t} &\sim N(\mu_{2t}, \sigma^2_{2t}). \end{aligned}$$

Suppose the indicator for Sex is $X_{i,1}$. Then

$$X_{i,1} \sim Bern(p) \qquad , \qquad p \sim Beta(2,3)$$

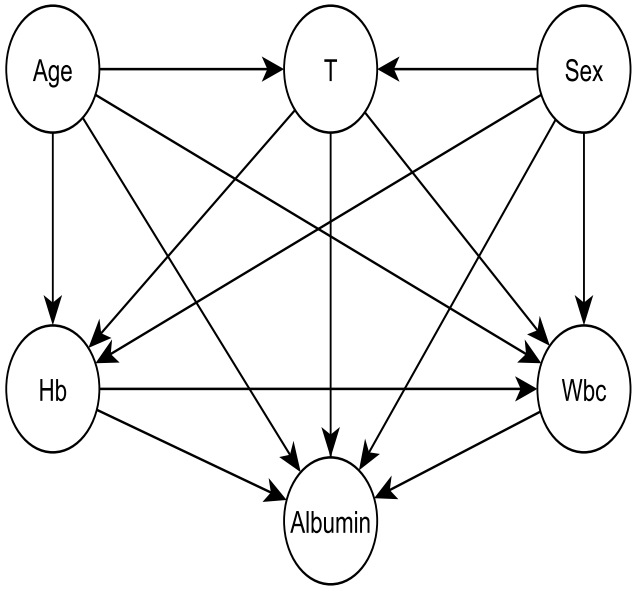


Figure 4.9: Fully-connected (apart from Age and Sex) Bayesian network for non-Hodgkin lymphoma data with imposed ordering of the nodes.

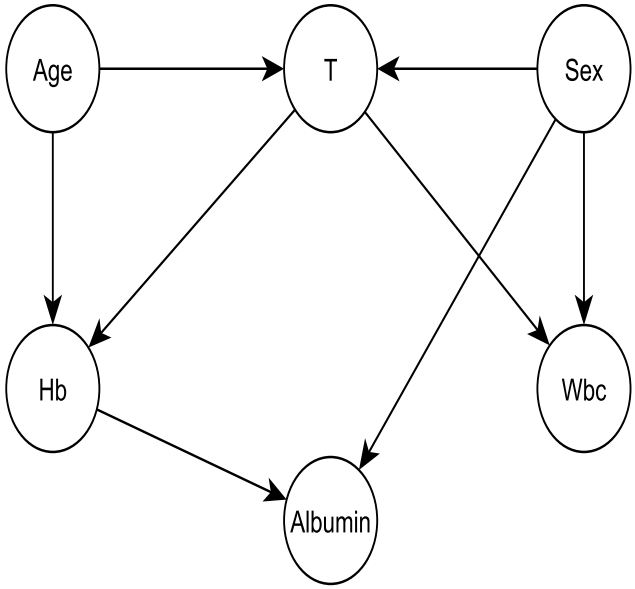


Figure 4.10: Most likely configuration which depends on the posterior probability of the coefficients which are non-zero.

Arcs	Net1	Net2	Net3	Net4	Net5	Net6
Age → Wbc	0	0	1	0	1	0
Hb → Wbc	0	0	0	0	1	1
T → Wbc	1	1	0	1	0	1
T → Hb	1	0	1	1	1	1
Age → Hb	1	1	0	0	0	1
Sex → Hb	0	0	0	0	0	0
Age → T	1	1	1	1	1	1
Sex → T	1	1	1	1	1	1
Sex → Wbc	1	1	1	1	1	1
T → Albumin	0	0	0	0	0	0
Age → Albumin	0	0	0	0	0	0
Sex → Albumin	1	1	1	1	1	1
Hb → Albumin	1	1	1	1	1	1
Wbc → Albumin	0	0	0	0	0	0

Table 4.2: The posterior probabilities for the first six most likely configurations based on the original network that have been chosen from all possible configurations.

Suppose that the Age is represented by $X_{i,2}$. Then

$$X_{i,2} \sim N(\mu_{age}, \sigma^2_{age}) \quad , \quad \mu_{age} \sim N(60, 10)$$

$$\tau_{age} = \frac{1}{\sigma^2_{age}} \quad , \quad \tau_{age} \sim \text{Gamma}(3, 2)$$

The bar chart in Figure 4.11 demonstrates the contribution of each arc in the network by taking the mean of all the nodes which are represented in Table 4.2.

From Figure 4.11, we notice that we have eight arcs with posterior probability greater than 0.5. For instance, there is an arc from Age to T since the posterior probability for this arc is 1, etc.

Finally, we conclude that there is not an arc from T to Albumin since the posterior probability for this arc is 0.

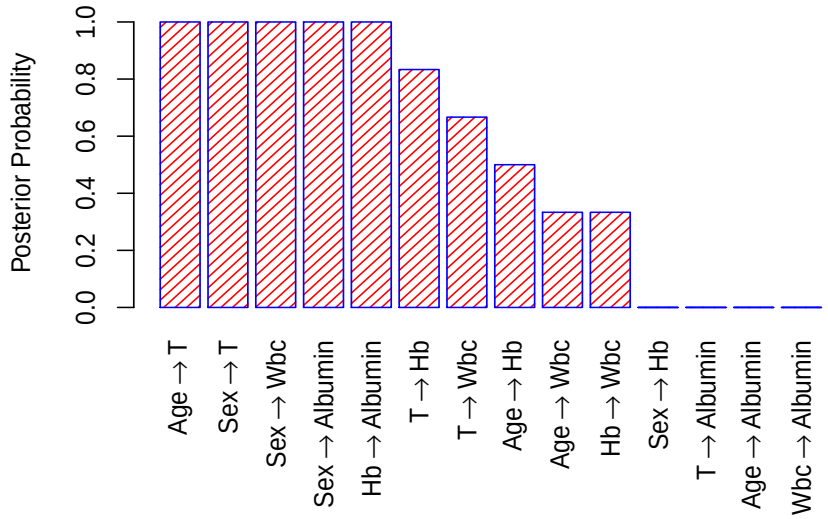


Figure 4.11: Bar chart representing the posterior probabilities that the arcs are present.

4.12 Summary

In this chapter, we have demonstrated some methodology about probabilistic graphical models, especially Bayesian networks. We have defined some concepts that relate to BNs. We compared Bayesian networks with the other models such as regression models. We explained in detail the two key phases in constructing Bayesian networks from data which are parameter learning and structure learning. We described the `R` package "`bnlearn`" which is used to learn about the structure of the network and make inference about the parameters using both frequentist and Bayesian analyses. We explained two algorithms that can be used to construct Bayesian networks, the Grow-Shrink and the Hill-Climbing algorithms and we gave a motivational example to apply these algorithms. We talked in brief about different sorts of Bayesian networks. We demonstrated the use of MCMC to select Bayesian network structure in order to choose the most likely configuration in terms of the posterior probability. This method can reduce the number of nodes or edges and that led to making the calculation for the network simpler than for the fully connected network. We applied this method to the non-Hodgkin lymphoma example and obtained the new proposed configuration.

Chapter 5

Survival analysis

5.1 Introduction

In this chapter, we explain some basic features of survival analysis and some important definitions that are related to our work. Generally speaking, in survival, we regard the starting point as "fixed" and we observe the time until some end point. For example, birth to death, time from cancer remission to recurrence and the time from first heart attack to second. An important aspect for this type of analysis is *censoring*. For more details, see Clark et al. (2003); Aalen (2008); Collett (2015); Ibrahim et al. (2001); Cox and Oakes (1984); Moore (2016).

In Section 5.2, we give a short outline of the general background of survival models including some features of those models and the important definitions that we need to deal with that type of data. In Section 5.3, we explain some important features that relate to survival analysis. We give a general overview of the most common survival models which for example allow the hazard function to be related to some predictive variables. These kinds of models include proportional hazard models which we illustrate in Section 5.5. In Section 5.6, we explain prognostic indices and how we can calculate an index based on a survival analysis. In Section 5.7, we demonstrate with details the most familiar parametric models in survival analysis, such as the exponential and Weibull distributions. In Section 5.8, we refer to using Bayesian inference in survival analysis using MCMC techniques to make inference about the coefficients of the covariates in the model. In Section 5.9, we use Bayesian survival analysis and particularly a `R` package called `rjags` with the model specification to make inference about the coefficients and apply that to the leukemia data

set and show some results.

5.2 General background on survival analysis

There are some reasons why the survival models are different from standard regression models. The main reason is that survival distributions are restricted to $(0, \infty)$ and typically not symmetric but are positively skewed while often, in other regression models we assume that the data follow the Gaussian distribution. Moreover, survival data often include censored observations. There are two possibilities to deal with this sort of data. First, we can resolve the problem of asymmetry in survival data by transforming the data using, for example a logarithmic transformation and secondly, we can adopt other suitable distributions that fit the survival data such as the exponential and Weibull distributions.

5.3 Some important aspects of survival data

5.3.1 Censored time

Let T be the time from a well-defined point known as the starting point until the occurrence of an event of interest. Then we refer to T as either a *survival* or *failure* time. These survival times are often censored and it is known as a *censored* time C. There are *three* common types of censoring in survival analysis.

Right censoring (happening frequently in survival analysis) occurs when we do not observe T but know that $(T \geq C)$. It might occur often for different reasons, such as the patient is still alive after the study ends or failure to keep in contact with him/her because he/she moved to another country.

Left censoring (less common than right censoring in survival) occurs when we do not observe T but we know that the event occurred before a particular time (*recruitment*) $(T < C)$. For example, suppose that a study was conducted to investigate the time to tumour recurrence following surgical resection of the original tumour from its primary site. During periodic six-month surveillance to detect tumour recurrence, certain patients were found to be positive for new tumour masses at the original or metastatic sites. Since such recurrence might have occurred before the patients attended the follow-up sessions, their actual recurrence time should therefore be less than six months. See Collett (2015).

Interval censoring was explained in depth in Zhang and Sun (2010). It occurs when T is known to lie between two times C_1 and C_2, $(C_1 < T < C_2)$ but the precise value is unknown. This type of censoring is more unusual in survival data. For methods to deal with these data, see Andreas (2011).

5.3.2 Independent and non-informative censoring

There are a large number of statistical methods that use failure time data with the assumption that the censoring is *noninformative* of the failure time. This means that the observation that the patient is censored at time c, can tell us only that $T > c$. Suppose that we have the possible censoring time C_i, so the noninformative censoring can be achieved by saying that C_i is independent of T_i, $i = 1, 2, ..., n$. So if we have a group of patients who have the same values of prognostic variables, the patient that has a censored survival time c, should be considered as representative of all other patients in that group who survive to time c. See Collett (2015); Klein and Moeschberger (2005); Kalbfleisch and Prentice (2011) for more detail.

5.4 Survival function, hazard function and cumulative hazard function

Let T_i be the survival time which is measured from the start date, for example the date of the diagnosis of patient i. The lifetime distribution function is

$$F_i(t) = \Pr(T_i < t).$$

The survival function is

$$S_i(t) = \Pr(T_i > t) = 1 - F_i(t)$$

which represents the probability that patient i will survive at least until time t. Notice that, since T is non-negative, therefore, $S_i(0) = 1$ and $\lim_{t\to\infty} S_i(t) = 0$.

The lifetime probability density function is

$$f_i(t) = \frac{d}{dt} F_i(t)$$

The hazard function, sometimes known as the "instantaneous failure (death) rate", is

$$h_i(t) = \frac{f_i(t)}{S_i(t)}. \tag{5.1}$$

We can specify the relationship between the hazard function and the survival function in (5.1). We can rewrite it as

$$h_i(t) = \frac{f_i(t)}{S_i(t)} = \frac{F_i'(t)}{S_i(t)} = -\frac{S_i'(t)}{S_i(t)}.$$

So

$$H_i(t) = \int_0^t h_i(u)du = -\int_0^t \frac{S_i'(u)}{S_i(u)}du = -\log S_i(t).$$

Hence,

$$S_i(t) = \exp[-H_i(t)]$$

where $H_i(t)$ is the cumulative hazard function, which measures the sum of the risks that the patients face between 0 and t.

5.5 Survival models

5.5.1 Proportional hazard models

In survival analysis, we might be interested in building models which allow the hazard function to be related to some explanatory or predictive variables. Therefore, the hazard function $h_i(t)$ depends upon the values of the variables that we measured or observed for the patient i. For instance, if we have three variables, (X_1, X_2, X_3), the values taken for patient i are, $(x_{i,1}, x_{i,2}, x_{i,3})$. If X_3 is the age in years of the patient at diagnosis, so for patient i, we might have $x_{i,3} = 65$.

The proportional hazard model was suggested by Cox (1972). It is a regression model which is commonly use in medical research to investigate the relationship between the survival time for patients and one or more predictive variables. The proportional hazard assumption is very common in survival data analysis. Although the Cox regression model involves a semi-parametric model, we can also have a parametric proportional hazards model. This model can be formulated as follows. Suppose we have values of the variables

$(x_{i,1}, x_{i,2}, ..., x_{i,p})$ taken from the patient at diagnosis. Then the hazard of death of patients at a specific time depends upon $(x_{i,1}, x_{i,2}, ..., x_{i,p})$. We can write the proportional hazard model as

$$h_i(t) = h_0(t) \exp(\eta_i)$$

where $h_0(t)$ is the baseline hazard which is the same for all patients and η_i is a linear predictor (combination) of the p explanatory variables in $\underline{x}_i$ where $\underline{x}_i = (x_{i,1}, ..., x_{i,p})'$. So,

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$$

The linear predictor η_i is also called the *risk score* or *prognostic index* of the patient i. The coefficients $(\beta_0, ..., \beta_p)$ have unknown values and we use our data to make inference about these values.

Suppose we have two patients i and j, who have different $x-$values. Therefore, the hazard function for patient i is

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p})$$

and, the hazard function for patient j is

$$h_j(t) = h_0(t) \exp(\beta_1 x_{j,1} + \beta_2 x_{j,2} + \cdots + \beta_p x_{j,p}).$$

As a result, the ratio of hazards for patients i and j is

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t) \exp(\beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p})}{h_0(t) \exp(\beta_1 x_{j,1} + \beta_2 x_{j,2} + \cdots + \beta_p x_{j,p})}$$

$$= \frac{\exp(\beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p})}{\exp(\beta_1 x_{j,1} + \beta_2 x_{j,2} + \cdots + \beta_p x_{j,p})}$$

$$= \exp(\beta_1[x_{i,1} - x_{j,1}] + \beta_2[x_{i,2} - x_{j,2}] + \cdots + \beta_p[x_{i,p} - x_{j,p}]) \tag{5.2}$$

Based on (5.2), it is obvious that $h_i(t)/h_j(t)$ does not depend on t.

In order to use a parametric hazard model, we should give $h_0(t)$ a particular functional form which may involve one or more unknown parameters. For example, if we use a Weibull distribution, then the hazard function is

$$h_i(t) = \gamma \lambda_i t^{\gamma - 1}$$

where $\lambda_i = \exp(\eta_i)$ and where $\eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$ and $\beta_0, ..., \beta_p$ and γ are unknown parameters.

Directed acyclic graphs (DAGs) are a very attractive and flexible way to represent the (in)dependence relationships between the variables in the model. Figure 5.1, shows the case when we know all the values of the parameters in the survival model (the coefficients of the predictive variables and any other unknown parameters). The lifetime variable T is *stochastically* dependent on the linear predictor η which in turn depends *deterministically* upon the predictive variables $(X_1, X_2, \cdots, X_p)$. The double ring around η indicates that it has deterministic dependence on its *parents*.

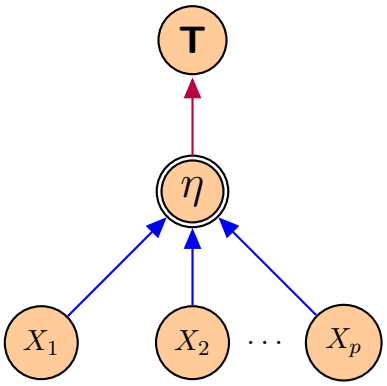


Figure 5.1: Basic survival model

5.5.2 Piecewise constant hazard model

We will use a piecewise constant hazard model for the leukaemia data later in Chapter 7 to illustrate Bayes linear kinematics. We are interested in making inferences about the model parameters using different methods such as full Bayes and non-conjugate prior methods.

We can define a piecewise constant hazard model (PCH) as a model in which the time is divided into disjoint intervals, and then we specify a constant hazard in each interval. However, those hazards are allowed to be different from interval to interval.

So, we choose fixed time points $s_0, s_1, ..., s_k$, where $s_0 = 0$ and $s_k \to \infty$. We can also define the k^{th} interval as $[s_{k-1}, s_k)$. Therefore, the baseline hazard function for interval $s_{k-1} \leq t < s_k$ will be

$$h_0(t) = \lambda_{0,k}$$

and the hazard function for patient i is

$$h_i(t) = \lambda_{i,k} = \phi_{i,k}\lambda_{0,k} = e^{\eta_{i,k}}.$$

where $\eta_{i,k} = \underline{x}_i^{'}\underline{\beta}_k$ is the linear predictor, $\underline{x}_i^{'} = (1, x_{i,1}, ..., x_{i,J})$ and $\underline{\beta}_k = (\beta_{k,0}, ..., \beta_{k,J})^{'}$.

As a result, the integrated hazard function $H_i(t) = \int_0^t h_i(z)dz$ will be as follows

$$H_i(t) = \sum_{r:s_r<t} \lambda_{i,r}(s_r - s_{r-1}) + \lambda_{i,k}(t - s_{k-1}),$$

for $r = (1, ..., k-1)$.

Now, we can write the survival function and the probability density function for patient i at time $s_k \leq t < s_{k+1}$ conditioning on $T \geq s_k$ respectively as

$$S_i(t \mid T \geq s_{k-1}) = \exp[-\lambda_{i,k}(t - s_{k-1})],$$

and

$$f_i(t \mid T \geq s_{k-1}) = \lambda_{i,k} \exp[-\lambda_{i,k}(t - s_{k-1})].$$

If we fix β_k for all k and $k = 1, ..., K$ then this is a proportional hazards model.

5.5.3 Accelerated failure time model

An accelerated failure time or accelerated life model is the same as the usual linear regression models except that the response variable in the accelerated failure model is just the log of the survival times. See Zhou (2015); Wei (1992).

This model has a survival function which differs from the survival function in a proportional hazard model. In a proportional hazard model, we scale the hazard function while, in an accelerated life model, we scale time in the following method.

Suppose that we have the base line survival function $S_0(t)$. Then we assume that the survival function for patient i takes the form

$$S_i(t) = S_0(\zeta_i t)$$

where ζ_i is a positive constant called the acceleration factor for patient i. Therefore, we can make ζ_i depend on the covariates for each patient. So, in order to specify the model

we need to specify the baseline S_0 and constants ζ_i. We set $\log(\zeta_i) = \eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$ where $x_{i,1}, ..., x_{i,p}$ are the covariate values for patient i.

5.6 Prognostic index

5.6.1 Introduction

We use prognostic indices to predict the outcome in patients with a certain disease. The value of the prognostic index depends on the clinical information about patients. A prognostic index can be useful to make a decision about the appropriate treatment for the patients. Henderson et al. (2001) mentioned the significance of using the prognostic indices in different situations. These included its use in the selection of treatments by clinicians, especially for a fatal disease, and its importance for the patients and their families to know and think about the future scenario in the remaining years for their patients by supporting and giving them the hope for the best in their lives.

The prognostic index is defined mathematically as a linear predictor based on the explanatory variables in the model. In a proportional hazards model, the prognostic index of patient i represents the logarithm of the multiplier of the hazard function of patient i. (*i.e.*, if h_i is the hazard function, then $h_i(t) = h_0(t)e^{\tilde{\eta}_i}$, where $h_0(t)$ is the baseline hazard, so the prognostic index $\tilde{\eta}_i = \log[h_i(t)/h_0(t)]$. A greater value for it corresponds to worse prognosis.

5.6.2 Computing the prognostic index

A prognostic index is an index of the prediction of the survival time for the patients with a certain disease. In order to calculate this index, we might fit one of the survival models, say a Weibull distribution. Suppose that we have a survival lifetime distribution, $T_i \sim \text{Weibull}(\alpha, e^{\eta_i})$, where α is the shape parameter in the model and $\lambda_i = e^{\eta_i}$ is the scale parameter. So, $\eta_i = \log(\lambda_i)$ is the prognostic index for patient i, where $(i = 1, 2, ..., n)$. We can write the prognostic index in the following way

$$\eta_i = \beta_0 + \sum_{j=1}^{J} \beta_i x_{i,j} \tag{5.3}$$

where $(\beta_0, \beta_1, \ldots, \beta_J)$ are the regression coefficients of interest and $x_{i,j}$ are the covariates in the model.

Given a set of data on past patients, we can find the posterior distribution of the coefficients $\beta_0, \beta_1, ..., \beta_J$. Since the index η is linear in these coefficients, the predictive mean of η for a new patient with given covariate values can be found using the posterior means of $\beta_0, ..., \beta_J$.

To make the index more interpretable for users, we can convert it to a $[0, 100]$ scale. If we have a large data set of past patients from the same or a similar population, we can compute the index for all patients in this data set. Then, if these past index values, in increasing order, are $\eta^{(1)}, ..., \eta^{(n)}$ and the values for a new patient is η^*, we base a transformed index on the rank within this data set $I = 100[R(\eta^*) + 0.5]/(n + 1)$ where $R(\eta^*) = \max(j : \eta^{(j)} < \eta^*)$ where $\eta^{(0)} \to -\infty$ and $\eta^{(n+1)} \to \infty$.

Alternatively, if the distribution of past η values is approximately normal, perhaps after transformation, we might find the sample mean m_η and sample standard deviation S_η and calculate $I = 100\Phi\left([\eta^* - m_\eta]/S_\eta\right)$, where $\Phi()$ is the standard normal cumulative distribution function.

5.7 Parametric models in survival analysis

5.7.1 Exponential survival model

The exponential distribution has an aspect that its hazard function does not depend on t. It is a special case of the Weibull distribution. Suppose we have observations $(t_1, t_2, ..., t_n)'$ which are independent and identically distributed (i.i.d.) from the survival model. This model is an exponential model with one parameter λ which has the probability density function, $f(t_i \mid \lambda) = \lambda e^{-\lambda t_i}$. Let $(\delta_1, \delta_2, ..., \delta_n)'$ be the censoring indicators, where $\delta_i = 0$ if t_i is a censored observation and $\delta_i = 1$ if t_i is a survival time observation. The survival function for an exponential distribution is $S(t_i \mid \lambda) = e^{-\lambda t_i}$. The hazard function is

$$h(t_i \mid \lambda) = \frac{f(t_i \mid \lambda)}{S(t_i \mid \lambda)} = \frac{\lambda e^{-\lambda t_i}}{e^{-\lambda t_i}} = \lambda$$

The likelihood for patients $i = 1, \ldots, n$ with death or right censoring times $t_1, \ldots, t_n$

then takes the form

$$\mathrm{L}(\theta, D) = \prod_{i=1}^{n} [f_i(t_i)]^{\delta_i} [S_i(t_i)]^{1-\delta_i} \tag{5.4}$$

where $\delta_i = 1$ for a patient whose death time t_i is observed and $\delta_i = 0$ if the lifetime is censored at time t_i.

We can rewrite the likelihood function in (5.4) as

$$\begin{aligned} \mathrm{L}(\theta, D) &= \prod_{i=1}^{n} [h_i(t_i)]^{\delta_i} [S_i(t_i)] \\ &= \lambda^{n_d} e^{-\lambda n \bar{t}} \end{aligned} \tag{5.5}$$

where $n_d = \sum_{i=1}^{n} \delta_i$ and $n\bar{t} = \sum_{i=1}^{n} t_i$.

5.7.2 Weibull survival model

We will use a Weibull model with Bayes linear kinematics in a Bayes linear kinematic prognostic network in Chapter 7.

The Weibull distribution has an additional parameter, α, called the shape parameter. The survival function for a Weibull distribution is

$$S(t_i \mid \lambda) = \exp(-\lambda t^{\alpha}).$$

Therefore the probability density function is

$$f(t_i \mid \lambda) = \alpha \lambda t^{\alpha - 1} \exp(-\lambda t^{\alpha}),$$

and the hazard function is

$$h(t_i \mid \lambda) = \alpha \lambda t^{\alpha - 1}.$$

5.8 Bayesian inference in survival analysis

5.8.1 Introduction

There is literature that deals with using frequentist methods in survival analysis. However, we are interested in this thesis to use Bayesian methods such as MCMC method. We can also use Bayesian inference in survival analysis in order to make some inferences about the unknown parameters in the model. We can fit different sorts of models in survival such as exponential, Weibull and Weibull mixture models, etc. We use `rjags` to fit all these models with different types of data sets such as the non-Hodgkin lymphoma data and leukemia data. In the following sections we illustrate how we can find the posterior distribution for the parameters of interest using Bayesian methodology. For further information, see, for example, Ibrahim et al. (2001).

5.8.2 Bayesian analysis for exponential lifetime distribution

The likelihood function for the exponential distribution is given by (5.5).

The conjugate prior for λ is a gamma density with two parameters. The first parameter α is called the *shape* parameter and the second parameter β is called the *rate* parameter. The prior density is

$$\pi(\lambda \mid \alpha, \beta) \propto \lambda^{\alpha-1} e^{-\lambda\beta} \tag{5.6}$$

Hence, combining the prior density and the likelihood in (5.5) and (5.6) respectively, we obtain the posterior density as follows

$$\pi(\lambda \mid t, \delta, \alpha, \beta) \propto \pi(\lambda \mid \alpha, \beta) \mathrm{L}(\lambda \mid t, \delta)$$

$$\pi(\lambda \mid t, \delta, \alpha, \beta) \propto \lambda^{\alpha+d-1} e^{-\lambda(\beta+\sum_{i=1}^{n} t_i)} \tag{5.7}$$

Based on (5.7), the posterior density is a gamma (α_1, β_1) distribution, where $\alpha_1 = \alpha + d$ and $\beta_1 = \beta + \sum_{i=1}^{n} t_i$. The posterior mean and posterior variance are given respectively as

$$\mathrm{E}_1(\lambda \mid t, \delta, \alpha, \beta) = \frac{\alpha + d}{\beta + \sum_{i=1}^{n} t_i}$$

and

$$\mathrm{Var}_1(\lambda \mid t, \delta, \alpha, \beta) = \frac{\alpha + d}{(\beta + \sum_{i=1}^{n} t_i)^2}.$$

Likewise, in order to calculate the posterior *predictive* density of a future observation t_f, we have

$$f(t_f \mid t, \delta, \alpha, \beta) = \int_0^\infty f(t_f \mid \lambda)\pi(\lambda \mid t, \delta, \alpha, \beta)d\lambda$$

where $f(t_f \mid \lambda) = \lambda e^{-\lambda t_f}$. Therefore, the posterior predictive density is

$$\begin{aligned} f(t_f \mid t, \delta, \alpha, \beta) &= \frac{(\beta + \sum_{i=1}^{n} t_i)^{\alpha+d}}{\Gamma(\alpha + d)} \int_0^\infty \lambda e^{-\lambda t_f} \lambda^{\alpha+d-1} e^{-\lambda(\beta + \sum_{i=1}^{n} t_i)} d\lambda \\ &= \frac{(\beta + \sum_{i=1}^{n} t_i)^{\alpha+d}}{\Gamma(\alpha + d)} \int_0^\infty \lambda^{\alpha+d} e^{-\lambda(\beta + \sum_{i=1}^{n} t_i + t_f)} d\lambda \\ &= \frac{(\beta + \sum_{i=1}^{n} t_i)^{\alpha+d}}{\Gamma(\alpha + d)} \times \frac{\Gamma(\alpha + d + 1)}{\left(\beta + \sum_{i=1}^{n} t_i + t_f\right)^{\alpha+d+1}} \\ &= (\alpha + d) \times \frac{(\beta + \sum_{i=1}^{n} t_i)^{\alpha+d}}{\left(\beta + \sum_{i=1}^{n} t_i + t_f\right)^{\alpha+d+1}} \\ &\propto \left(\beta + \sum_{i=1}^{n} t_i + t_f\right)^{-(\alpha+d+1)} \end{aligned} \tag{5.8}$$

where (5.8) represents the *kernel* of the posterior predictive density when $t_f > 0$.

5.8.3 Bayesian analysis for Weibull lifetime distribution

Suppose X is a design matrix with dimension $n \times (j+1)$ where the i^{th} row of X indicates the covariate values of patient i, $X_i = (1, x_{i,1}, x_{i,2}, ..., x_{i,j})$, n is the number of patients and j is the number of covariates that are used in the model. Let T be the survival times, $T = (t_1, t_2, ..., t_n)'$. The censoring indicator δ_i (as in Section 5.8), represents whether a patient's death time t_i is observed or right censored.

Suppose that the lifetime random variable T has a Weibull distribution with two

parameters (γ, λ) and the data we use are subject to right censoring. Therefore, the probability density function for a patient i is given by

$$f(t_i \mid \lambda_i, \gamma) = \lambda_i \gamma t_i^{\gamma-1} \exp(-\lambda_i t_i^{\gamma}).$$

Let the number of patients whose death time t_i is observed be n_d and the number of patients who had right censoring be n_c. The scale parameter λ_i of a patient i depends on the covariates,

$$\lambda_i = \exp(\eta_i) = \exp(x_i^{'}\underline{\beta})$$

where $\underline{\beta} = (\beta_0, \beta_1, ..., \beta_p)^{'}$, β_0 is the intercept and β_j is the regression coefficient for the j^{th} covariate.

The linear predictor η_i for the i^{th} patient is given by

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$$

So the survival function for the i^{th} patient is

$$S(t_i \mid \lambda_i, \gamma) = \exp(-\lambda_i t_i^{\gamma})$$

The likelihood function for the observed and censored data is written as

$$\begin{aligned} \mathrm{L} &= \left[\prod_{i \in d} f_i(t_i \mid \lambda_i, \gamma)\right]\left[\prod_{i \in c} S_i(t_i \mid \lambda_i, \gamma)\right] \\ &= \left[\prod_{i \in d} \lambda_i \gamma t_i^{\gamma-1}\right]\left[\prod_{\forall i} \exp(-\lambda_i t_i^{\gamma})\right] \\ &= \gamma^{n_d}\left[\prod_{i \in d} t_i^{\gamma-1}\lambda_i\right] \exp\left[-\lambda_i \sum_{i \in d \cup c} t_i^{\gamma}\right] \end{aligned}$$

where c and d are sets representing the censoring and the observed times respectively.

Then the *log likelihood* function is

$$\sum_{i \in d}\Big[\log \lambda_i + (\gamma - 1)\log t_i\Big] + n_d \log \gamma - \sum_{i=1}^{n} \lambda_i t_i^{\gamma}.$$

To apply Bayesian inference, we need the prior density for the unknown parameters.

In this case, suppose that γ and λ have independent prior distribution.

Our prior distribution for the regression coefficient β will have a multivariate normal prior distribution $\beta \sim N_{p+1}(\underline{\mu}, V)$ and we use a gamma prior distribution for γ. The multivariate normal prior density is

$$(2\pi)^{(p+1)/2}|V|^{-1/2}\exp\left\{-\frac{1}{2}\Big[(\beta-\underline{\mu})'V^{-1}(\beta-\underline{\mu})\Big]\right\}.$$

Here $\underline{\mu}$ is the prior *mean* vector $\underline{\mu} = (\mu_0, \mu_1, ..., \mu_p)'$ and V is a covariance matrix with dimension $(p+1)\times(p+1)$.

Suppose $\gamma \sim \text{Gamma}(a, b)$, so the prior density will be

$$\begin{aligned}\pi(\gamma \mid a, b) &= \frac{b^a}{\Gamma(a)}\gamma^{a-1}\exp(-b\gamma)\\ &\propto \gamma^{a-1}\exp(-b\gamma).\end{aligned}$$

The joint posterior density for γ and β is given by

$$\pi(\gamma, \beta \mid D) \propto \pi(\gamma, \beta)L(\gamma, \beta \mid D).$$

So

$$\begin{aligned}\pi(\gamma, \beta \mid D) \propto &\gamma^{a+n_d-1}\exp\left\{\sum_{i=1}^{n}\Big[\delta_i x_i'\beta + \delta_i(\gamma-1)\log t_i - t_i^{\gamma}\exp(x_i'\beta)\Big]\right.\\ &\left.- b\gamma - \frac{1}{2}(\beta-\underline{\mu})'V^{-1}(\beta-\underline{\mu})\right\}\end{aligned} \tag{5.9}$$

We notice that (5.9) does not have a closed form, so there is need for numerical integration or MCMC methods. As a result, a Metropolis-Hastings algorithm should be used to evaluate the posterior distribution for γ and β. See Consul (2016).

5.8.4 Example: inference about the two parameters of Weibull distribution in the non-Hodgkin lymphoma example

In this Bayesian analysis for the Weibull distribution, we use a Metropolis-within-Gibbs algorithm and apply it to the non-Hodgkin lymphoma data.

Suppose the random variable T has a Weibull distribution with two parameters α and

λ. Then the likelihood function will be

$$\mathrm{L} = (\alpha\lambda)^{n_d} \exp\left[(\alpha-1)\sum_{i=1}^{n}\delta_i \log(t_i)\right] \exp\left[-\lambda\sum_{i=1}^{n} t_i^{\alpha}\right]$$

where δ_i is the event indicator. Let $\underline{t} = (t_1, ..., t_n)'$ and $\underline{\delta} = (\delta_1, ..., \delta_n)'$.

Now, in order to make inference about α and λ, we need to specify the prior for the two quantities. So, we suppose that α and λ are independent. In addition, $\alpha \sim \text{Gamma}(a, b)$ and $\lambda \sim \text{Gamma}(r, s)$.

Therefore, the posterior distribution will be

$$\begin{aligned}\pi(\alpha, \lambda|\underline{t}, \underline{\delta}) &\propto \pi(\alpha)\pi(\lambda)\ \mathrm{L} \\ &\propto \alpha^{a-1}\exp[-b\alpha]\ \lambda^{r-1}\exp[-s\lambda]\ (\alpha\lambda)^{n_d}\exp\left[(\alpha-1)\sum_{i=1}^{n}\delta_i\log(t_i)\right]\exp\left[-\lambda\sum_{i=1}^{n}t_i^{\alpha}\right]\end{aligned}$$

Then the full conditional distribution (FCD) of α is

$$\begin{aligned}\pi(\alpha|\lambda, \underline{t}, \underline{\delta}) &= \frac{\pi(\alpha, \lambda|\underline{t}, \underline{\delta})}{\pi(\lambda|\underline{t}, \underline{\delta})} \\ &\propto \pi(\alpha, \lambda|\underline{t}, \underline{\delta}) \\ &\propto \alpha^{a+n_d-1}\exp[-b\alpha]\exp\left[(\alpha-1)\sum_{i=1}^{n}\delta_i\log(t_i)\right]\exp\left[-\lambda\sum_{i=1}^{n}t_i^{\alpha}\right]\end{aligned}$$

and the FCD of λ is

$$\pi(\lambda|\alpha, \underline{t}, \underline{\delta}) \propto \lambda^{r+n_d-1}\exp\left[-\lambda\left(\sum_{i=1}^{n}t_i^{\alpha} + s\right)\right]$$

Now, suppose, for example, we fix α. Then the posterior distribution $\pi(\lambda|\alpha, \underline{t}, \underline{\delta}) \sim \text{Gamma}(r + n_d, \sum_{i=1}^{n} t_i^{\alpha} + s)$.

We notice that, the posterior distribution of λ has a closed form. However, the posterior distribution of α does not have a closed form, so we can use MCMC methods to draw samples from that distribution using a Metropolis within Gibbs algorithm. A `R` function is written in Appendix A.5.1 to generate samples from the posterior distribution of α and λ.

To sample α in this case, we have to use a proposal distribution of α, say $\alpha^* \sim N(\alpha, \sigma_\alpha^2)$

with acceptance probability A where

$$A = \frac{\pi(\alpha^*|\lambda, \underline{t}, \underline{\delta})}{\pi(\alpha|\lambda, \underline{t}, \underline{\delta})}$$
$$= \frac{\alpha^{*(a+n_d)-1} \exp[-b\alpha^*] \exp\left[(\alpha^* - 1)\sum_{i=1}^n \delta_i \log(t_i)\right] \exp\left[-\lambda \sum_{i=1}^n t_i^{\alpha^*}\right]}{\alpha^{a+n_d-1} \exp[-b\alpha] \exp\left[(\alpha - 1)\sum_{i=1}^n \delta_i \log(t_i)\right] \exp\left[-\lambda \sum_{i=1}^n t_i^{\alpha}\right]}$$

The results in Figure 5.2 show that our samples for both parameters using Metropolis with Gibbs algorithm are mixing well and the chains converged.

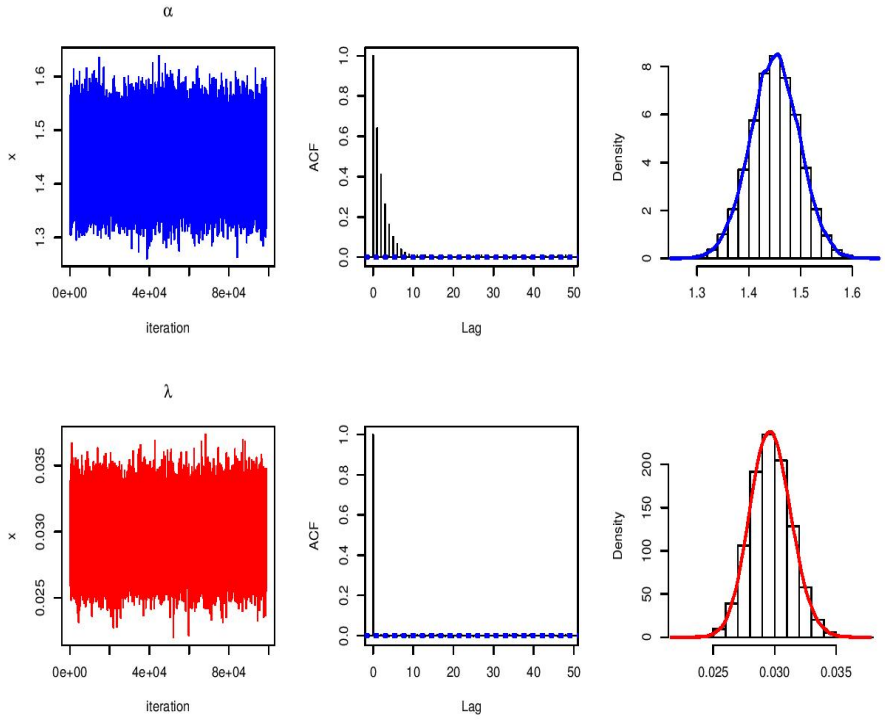


Figure 5.2: Trace plots, the autocorrelation plots and posterior densities for α and λ.

5.9 Bayesian survival analysis using `rjags`

5.9.1 Introduction

In this section, we explain how to fit a lifetime distribution using `rjags`. The MCMC algorithm uses data augmentation (see Section 3.9) to deal with censored lifetimes. The censored lifetimes are sampled at each iteration of the Gibbs sampler. While, in this case,

this is less computationally efficient than writing code to evaluate the exact likelihood, it is convenient especially in more complicated models.

5.9.2 Leukaemia example

As an example we use the leukemia data set, and an exponential distribution for the lifetime T_i. So

$$T_i \sim \exp(\lambda_i)$$

where $i = 1, ..., n$. Therefore, the probability density function (pdf) for the lifetime distribution T_i is

$$f(t_i|\lambda) = \lambda e^{-\lambda t_i}.$$

The data were collected by North-West Leukemia Register in the UK for $n = 1043$ patients from 1982 to 1998. These are right censored data where 879 patients died and 164 were censored. The variable of interest in this study is the time in days until a patient dies.

There are four covariates in this study. See Section 2.3.2. We code them as follows.

1. The age, A_i in years of patient i. We define $x_{i1} = A_i - 60$.
2. The sex of the patient. We have, $x_{i2} = -1$ if the patient is female and $x_{i2} = 1$ if the patient is male.
3. White blood cell count (WBC) W_i at the time of diagnosis with 1 unit$= 50 \times 10^9/l$. We use $x_{i3} = W_i - 8$.
4. The Townsend score, used directly as the covariate x_{i4}.

The time for patient i is t_i and the event indicator is δ_i where $\delta_i = 1$ if it is an observed death time and $\delta_i = 0$ if it is a censoring time.

5.9.3 Model specification for leukemia data

The model specification for the exponential survival time for the leukemia data has been written in `rjags` in Appendix A.5.2.

We have then 4 coefficients in this model, $\underline{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$, where β_0 is the intercept, β_1 is the coefficient of age , β_2 is the coefficient of sex, β_3 is the coefficient of wbc and β_4 is the coefficient of the deprivation score in our model. These coefficients are given independent normal prior distributions. The prior means, prior standard deviations, posterior means and posterior standard deviations are given in Table 5.1. The likelihood is given in (5.4) and (5.5).

5.9.4 Results

We fitted the exponential survival model to the leukemia data set with 100000 iterations and two chains. Figure 5.3 shows the trace plots and the densities for the coefficients.

Table 5.1 shows the posterior means and standard deviation for these coefficients. We can say that all the coefficients in this model show good mixing. Therefore, the local averages of all $\underline{\beta} = (\beta_0, \beta_{age}, \beta_{sex}, \beta_{wbc}, \beta_{depscore})'$ in the chains are roughly constant.

	Prior		Posterior	
Parameter	Mean	SD	Mean	SD
β_0	-6.90	0.12	-6.52	0.048
β_{age}	0.040	0.030	0.039	0.002
β_{sex}	0.050	0.150	0.148	0.061
β_{wbc}	0.080	0.173	0.004	0.001
$\beta_{depscore}$	0.120	0.110	0.021	0.009

Table 5.1: Prior and posterior means and standard deviations for each of the coefficients in the exponential survival model.

Now, suppose we consider new patients. That means we are interested in plotting the survival function for these particular patients, for example, male and age 63, etc. Therefore, we have λ for that patient denoted λ^*. If the posterior median for λ^* is λ^*_m and the lower and upper limits of the 95% interval are $\lambda^*_{0.025}$ and $\lambda^*_{0.975}$ respectively, then the corresponding quantities for the survival probability at time t are $\exp(-\lambda^*_m t), \exp(-\lambda^*_{0.025} t)$ and $\exp(-\lambda^*_{0.975} t)$. These are plotted against t in Figure 5.4. We apply this technique with eight different patients. The values of the covariates for each patient are presented in Table 5.2.

Figure 5.4 shows the predictive survival probability for these patients. For instance, we use the patients 1 and 2 in Table 5.2 to produce the graph on the top left of Figure 5.4 and we use patients 3 and 4 in the Table 5.2 to produce the graph on the top right

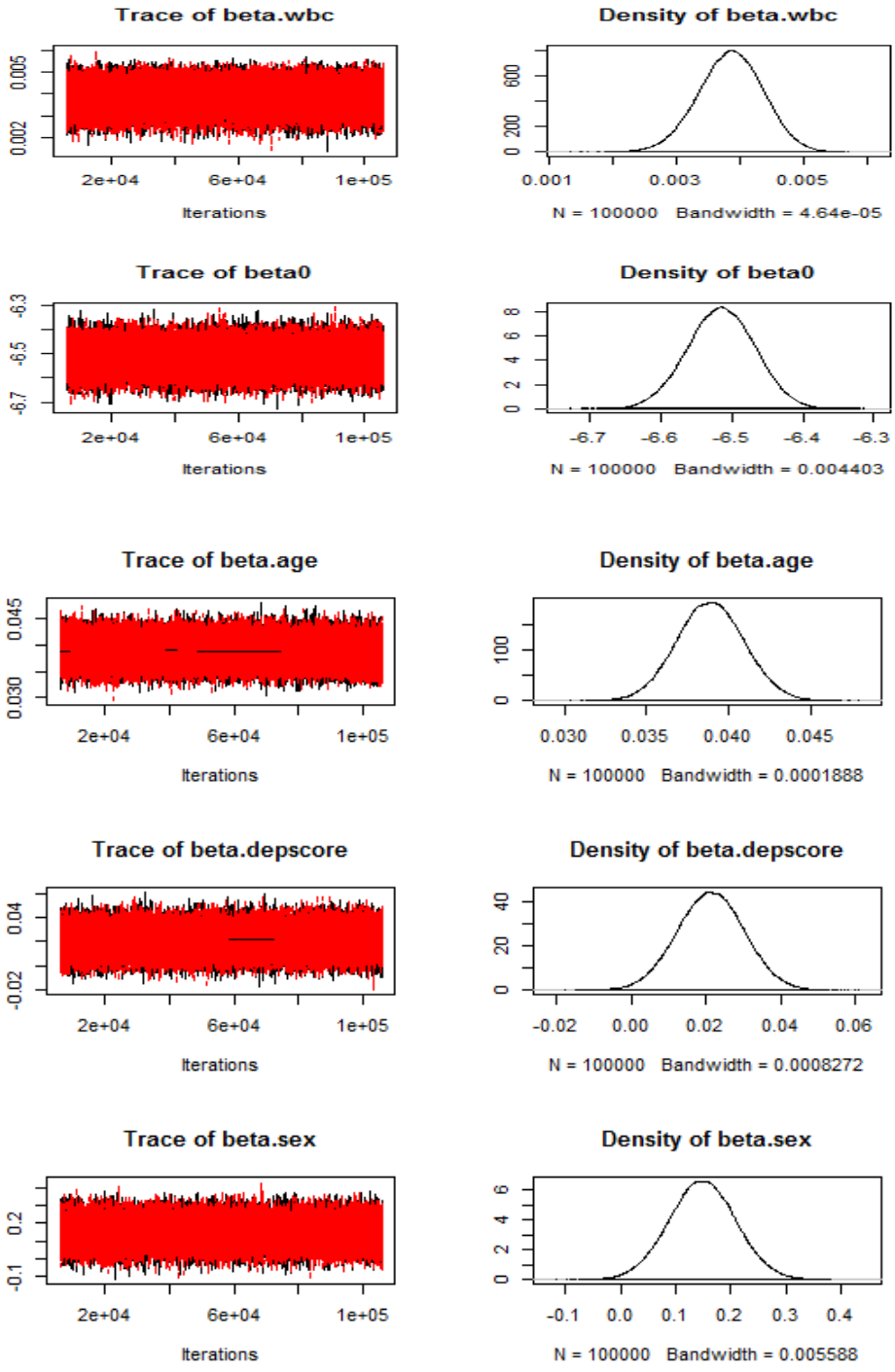


Figure 5.3: Trace plots and the densities for the coefficients, $\beta_0, \beta_{age}, \beta_{sex}, \beta_{wbc}, \beta_{depscore}$.

Patient	Age	Sex	WBC	Deprivation
1	63	male	6.8	2.02
2	63	female	197	7.66
3	61	male	13.3	-1.96
4	83	female	160	-2.59
5	48	male	1.4	-1.7
6	87	female	1.4	-3.47
7	61	male	3.8	4.35
8	84	female	30.5	4.35

Table 5.2: Eight different new patients in the leukaemia example.

of Figure 5.4, etc. We notice that we have narrow credible intervals. See Appendix A.5.3 for the `rjags` code.

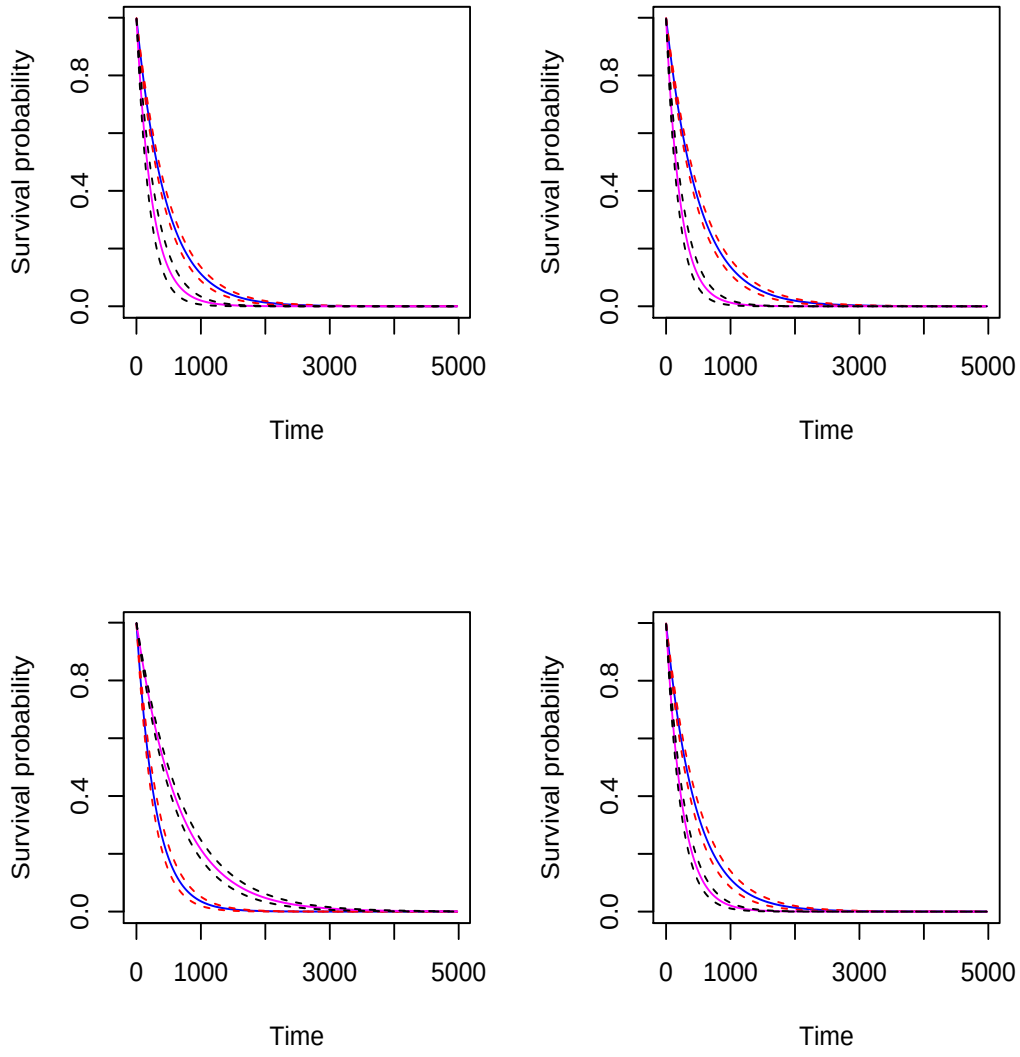


Figure 5.4: Predictive survival probability for eight different patients in the leukemia example. Top left: Patient 1 (blue) and patient 2 (pink). Top right: Patient 3 (blue) and patient 4 (pink). Bottom left: Patient 5 (blue) and patient 6 (pink). Bottom right: Patient 7 (blue) and patient 8 (pink).

5.10 Summary

We have given general background information on survival analysis in this chapter. We illustrated some useful model that relate the survival lifetime distribution to some covariates in the model. These models are proportional hazard models, piecewise constant hazard model and accelerated failure time model. We discussed prognostic indices which are used to predict the outcome in patients with a certain disease. We showed for example, how to calculate this index by fitting the Weibull lifetime distribution. In this chapter, we demonstrated how we can calculate the posterior distribution for the parameters in the survival analysis for exponential and Weibull distributions. We used `rjags` to compute all the posterior means and variances for the parameters of interest in the model. We showed some results and graphs which show that the sampler mixed well and the chains converged. We found that for different new patients in the leukaemia example, we predicted the survival probability with narrow credible intervals.

Chapter 6

Bayes linear kinematics and Bayes linear Bayes graphical models

6.1 Introduction

In Chapter 7, we will describe some novel applications of Bayes linear kinematics to survival data. In this chapter we describe and illustrate Bayes linear methods which are the basis for Bayes linear kinematics. We then describe and illustrate Bayes linear kinematics and Bayes linear Bayes models and introduce some novel developments of the theory.

In Bayesian analysis, we need to specify the prior distribution as our prior beliefs. Therefore, we need to specify our prior distribution with uncertainty and that can be expressed using probability. After observing some data, we calculate the likelihood function and finally, compute the posterior density which is proportional to the prior density, multiplied by the likelihood. With many dimensions in our analysis, probably, we need intensive calculations in order to obtain the results and so we depend on numerical integration methods. One common method that is used in many different fields is Markov chain Monte Carlo (MCMC), which requires intensive and often time-consuming calculations. However, the Bayes linear kinematics (BLK) method can obtain the results faster than MCMC. Furthermore, the BLK method depends only on a second order prior specification, which does not require any assumption about artificial probability distributions. In this chapter, we explain Bayes linear methods in Section 6.2. In Section 6.3, we describe the methods of Bayes linear kinematics. We introduce a novel feature which is the use of

non-conjugate marginal updates in Section 6.7, so we can compute the posterior moments. In Section 6.10, we consider some special types of observational variables, especially those which may be relevant to our prognostic index application. Finally, we give some theory about the BLK direct and BLK indirect methods which we shall use later in Chapter 7.

6.2 Bayes linear methods

6.2.1 Basic theory

In the standard Bayesian paradigm, we should specify the full joint prior distribution for all unknown quantities. By using Bayes' theorem, we update our prior beliefs by conditioning on the observations and then calculating the posterior distributions. A Bayes linear analysis is distinct from the full Bayesian approach in that we only need to specify the first and the second-order moments for the prior and then calculate posterior moments. For instance, if we have a random quantity X then we specify the prior expectation and variance of X respectively as follows $\mathrm{E}_0(\mathrm{X})$ and $\mathrm{Var}_0(\mathrm{X})$. Furthermore, for two quantities X_1 and X_2, we also need to specify a prior covariance $\mathrm{Cov}_0(\mathrm{X}_1, \mathrm{X}_2)$.

Suppose that we have two vectors $\alpha = (\alpha_1, ..., \alpha_p)'$ and $\beta = (\beta_1, ..., \beta_k)'$ where α is the observed quantities and β is the inferential quantities. Assume that we have made a full second-order prior specification for the set $A = \alpha \cup \beta$. Bayes linear methods (Goldstein and Wooff, 2007) suggest a way to update beliefs about β by a linear fitting on α which can be done using the Bayes linear updating equations for $\beta \mid \alpha$

$$\begin{aligned} \mathrm{E}_1(\beta) &= \mathrm{E}_0(\beta) + \mathrm{Cov}_0(\beta, \alpha)\mathrm{Var}_0^{-1}(\alpha)[\alpha - \mathrm{E}_0(\alpha)] \\ \mathrm{Var}_1(\beta) &= \mathrm{Var}_0(\beta) - \mathrm{Cov}_0(\beta, \alpha)\mathrm{Var}_0^{-1}(\alpha)\mathrm{Cov}_0(\alpha, \beta) \end{aligned} \tag{6.1}$$

where $\mathrm{E}_1(\beta)$ and $\mathrm{Var}_1(\beta)$ are the adjusted expectation and adjusted variance for $\beta \mid \alpha$.

Alternatively, we can express the relationship as

$$\beta = \mathrm{E}_0(\beta) + H_{\beta|\alpha}\Big[\alpha - \mathrm{E}_0(\alpha)\Big] + U_{\beta|\alpha} \tag{6.2}$$

where $H_{\beta|\alpha} = \mathrm{Cov}_0(\beta, \alpha)\mathrm{Var}_0^{-1}(\alpha)$ and the random vector $U_{\beta|\alpha}$ has mean zero and variance $\mathrm{Var}(U_{\beta|\alpha}) = \mathrm{Var}_0(\beta) - \mathrm{Cov}_0(\beta, \alpha)\mathrm{Var}_0^{-1}(\alpha)\mathrm{Cov}_0(\alpha, \beta)$.

If $\mathrm{Var}_0(\alpha)$ is not invertible, we need to use a suitable generalised inverse, for instance

a Moore-Penrose inverse. The idea of using a Moore-Penrose inverse was introduced by Moore and Barnard (1935) and Penrose (1955) to obtain the inverse of the matrix even when the matrix is rectangular or singular. So, this sort of inverse can be defined using the following four properties in the definition

Definition The Moore–Penrose inverse of the $p \times n$ matrix V is the $n \times p$ matrix which is denoted by V^+ and that satisfies the conditions

$$\begin{aligned} VV^+V &= V, \\ V^+VV^+ &= V^+, \\ (VV^+)' &= VV^+, \\ (V^+V)' &= V^+V. \end{aligned}$$

Schott (2016) mentioned that this Moore–Penrose inverse has one important aspect that can distinguish it from other generalised inverses. He showed that the Moore–Penrose inverse is uniquely defined.

6.2.2 Bayes linear adjusted expectation

Suppose that we are interested in an unknown quantity θ. Our prior expectation for θ is $\mathrm{E}_0(\theta)$. We observe a collection of data $\underline{y} = (y_1, y_2, \ldots, y_n)'$. Consider that $\mathrm{E}(\theta \mid \underline{y})$, our adjusted expectation of θ, is a *linear* function of the observed data $\underline{y}$ which can be represented as

$$\mathrm{E}(\theta \mid \underline{y}) = \alpha + \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_n y_n$$

We choose $\alpha, \beta_1, \beta_2, ..., \beta_n$ in order to minimise the following quantity

$$\mathrm{E}\Big[(\theta - \alpha - \beta_1 y_1 - \beta_2 y_2 - \cdots - \beta_n y_n)^2\Big]$$

So, the adjusted expectation for $\theta \mid \underline{y}$ is $\mathrm{E}(\theta \mid \underline{y})$.

Now, let $A = \alpha + \beta' Y$. To find the constant α and the vector β', we need to minimise the loss function as follows

$$\mathrm{L} = \mathrm{E}\Big[\big(\theta - \mathrm{A}\big)^2\Big]$$

Hence, we can write A as

$$A = \alpha + \beta^{'}Y = \alpha + \beta^{'}\Big[Y - \mathrm{E(Y)} + \mathrm{E(Y)}\Big] = \alpha + \beta^{'}\mathrm{E(Y)} + \beta^{'}[\mathrm{Y} - \mathrm{E(Y)}]$$
$$= \alpha^{*} + \beta^{'}[Y - \mathrm{E(Y)}]$$

where $\alpha^{*} = \alpha + \beta^{'}\mathrm{E(Y)}$.

Now,

$$\mathrm{E}\Big[(\theta - \mathrm{A})^2\Big] = \Big[\mathrm{E}(\theta - \mathrm{A})\Big]^2 + \mathrm{Var}(\theta - \mathrm{A})$$
$$= \Big[\mathrm{E}(\theta - \alpha^{*} - \beta^{'}[\mathrm{Y} - \mathrm{E(Y)}])\Big]^2 + \mathrm{Var}(\theta - \alpha^{*} - \beta^{'}[\mathrm{Y} - \mathrm{E(Y)}])$$
$$= (\mathrm{E}_0(\theta) - \alpha^{*})^2 + \mathrm{Var}(\theta - \beta^{'}\mathrm{Y}).$$

If we choose $\mathrm{E}_0(\theta) = \alpha^{*}$ then the first term will be zero. We need to differentiate the second term with respect to β.

$$\frac{\partial}{\partial\beta}\mathrm{Var}(\theta - \beta^{'}\mathrm{Y}) = \frac{\partial}{\partial\beta}\Big[\mathrm{Var}(\theta) + \beta^{'}\mathrm{Var(Y)}\beta - 2\mathrm{Cov}(\theta, \mathrm{Y})\beta\Big]$$
$$= 2\beta^{'}\mathrm{Var(Y)} - 2\mathrm{Cov}(\theta, \mathrm{Y})$$

Equating this to 0, we obtain $\beta^{'} = \mathrm{Cov}(\theta, \mathrm{Y})\mathrm{Var}^{-1}(\mathrm{Y})$. The adjusted expectation of $\theta \mid \underline{y}$ is

$$A = \alpha^{*} + \beta^{'}[Y - \mathrm{E(Y)}]$$

where $\alpha^{*} = \mathrm{E}_0(\theta)$ and $\beta^{'} = \mathrm{Cov}(\theta, \mathrm{Y})\mathrm{Var}^{-1}(\mathrm{Y})$. Hence,

$$A = \mathrm{E}_1(\theta) = \mathrm{E}_0(\theta) + \mathrm{Cov}(\theta, \mathrm{Y})\mathrm{Var}^{-1}(\mathrm{Y})[\mathrm{Y} - \mathrm{E(Y)}].$$

6.2.3 Bayes linear adjusted variance

In this subsection, we illustrate the derivation of the adjusted variance in Bayes linear analysis. This adjusted variance can be obtained from the error terms from the Bayes linear fit $\mathrm{R}(Y|D)$, where $D = (D_1, ..., D_n)^{'}$ represents the subset of the values that have been observed.

Therefore,

$$\mathrm{R}(Y|D) = Y - \mathrm{E}(Y|D)$$

where $\mathrm{E}(Y|D) = \mathrm{E}(Y) + \mathrm{Cov}(Y, D)\mathrm{Var}^{-1}(D)[D - \mathrm{E}(D)]$. The adjusted quantity also has two important properties:

$$\mathrm{E}\left[\mathrm{R}(Y|D)\right] = 0$$

and

$$\mathrm{Cov}\left[\mathrm{R}(Y|D), \mathrm{E}(Y|D)\right] = 0.$$

Hence, we can write $Y = \mathrm{R}(Y|D) + \mathrm{E}(Y|D)$, so that

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathrm{R}[Y|D]) + \mathrm{Var}(\mathrm{E}[Y|D]).$$

We also have

$$\mathrm{Var}(Y|D) = \mathrm{Var}(\mathrm{R}[Y|D]) = \mathrm{E}\Big[(Y - \mathrm{E}[Y|D])^2\Big].$$

To evaluate $\mathrm{Var}(Y|D)$, we need to specify our prior variances and covariances so that

$$\mathrm{Var}(Y|D) = \mathrm{Var}(Y) - \mathrm{Cov}(Y, D)\mathrm{Var}^{-1}(D)\mathrm{Cov}(D, Y). \tag{6.3}$$

For more detail, see Goldstein and Wooff (2007).

6.2.4 Bayes linear approach: motivational example and comparison with full-Bayes analysis

Suppose we have a simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Suppose that we want to predict the height (Y) in inches of a student from the student's shoe size (X) using this relationship. Suppose that we have collected some data $(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)$ from a class of students. Therefore, we can write the regression model in terms of

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \text{and} \qquad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

So, we can write our linear regression model in matrix form

$$Y = X\underline{\beta} + \underline{\epsilon}$$

where X represents the design matrix, $\underline{\beta}$ refers to the parameters in the model and the error term $\epsilon \sim N(0, \sigma^2)$. We assume that σ^2 is known.

We have

$$\mathrm{E}(Y) = X\mathrm{E}(\underline{\beta})$$

and

$$\mathrm{Var}(Y) = X\mathrm{Var}(\beta)X^{'} + \mathrm{Var}(\epsilon)$$

To apply Bayes theorem, we need to specify our prior distribution.

Suppose that our prior beliefs are such that

$$\mathrm{E}_0(\beta) = \begin{bmatrix} \mathrm{E}(\beta_0) \\ \mathrm{E}(\beta_1) \end{bmatrix} = \begin{bmatrix} 55.6667 \\ 1.6667 \end{bmatrix}$$

and

$$\mathrm{Var}_0(\beta) = \begin{bmatrix} \mathrm{Var}(\beta_0) & \mathrm{Cov}(\beta_0, \beta_1) \\ \mathrm{Cov}(\beta_0, \beta_1) & \mathrm{Var}(\beta_1) \end{bmatrix} = \begin{bmatrix} 36.5 & -4 \\ -4 & 0.5 \end{bmatrix}$$

and it is known that $\sigma^2 = 2$.

We can update our beliefs about β_0 and β_1 after observing some data, in this case $n = 52$ students. The data are shown in Figure 6.1. Bayes theorem is used to update these beliefs and the posterior distribution is calculated.

The posterior mean vector and the posterior variance matrix are summarised respectively as follows:

$$\mathrm{E}(\beta|y) = \begin{bmatrix} \mathrm{E}(\beta_0|y) \\ \mathrm{E}(\beta_1|y) \end{bmatrix} = \begin{bmatrix} 55.7072 \\ 1.5972 \end{bmatrix}$$

and

$$\mathrm{Var}(\beta|y) = \begin{bmatrix} \mathrm{Var}(\beta_0|y) & \mathrm{Cov}(\beta_0, \beta_1|y) \\ \mathrm{Cov}(\beta_0, \beta_1|y) & \mathrm{Var}(\beta_1|y) \end{bmatrix} = \begin{bmatrix} 0.7122 & -0.0848 \\ -0.0848 & 0.0107 \end{bmatrix}.$$

Now to do Bayes linear method, we start with the Bayes linear equations

$$\mathrm{E}^*(\beta|y) = \mathrm{E}_0(\beta) + \mathrm{Cov}_0(\beta, y)\mathrm{Var}_0^{-1}(y)[y - \mathrm{E}_0(y)]$$

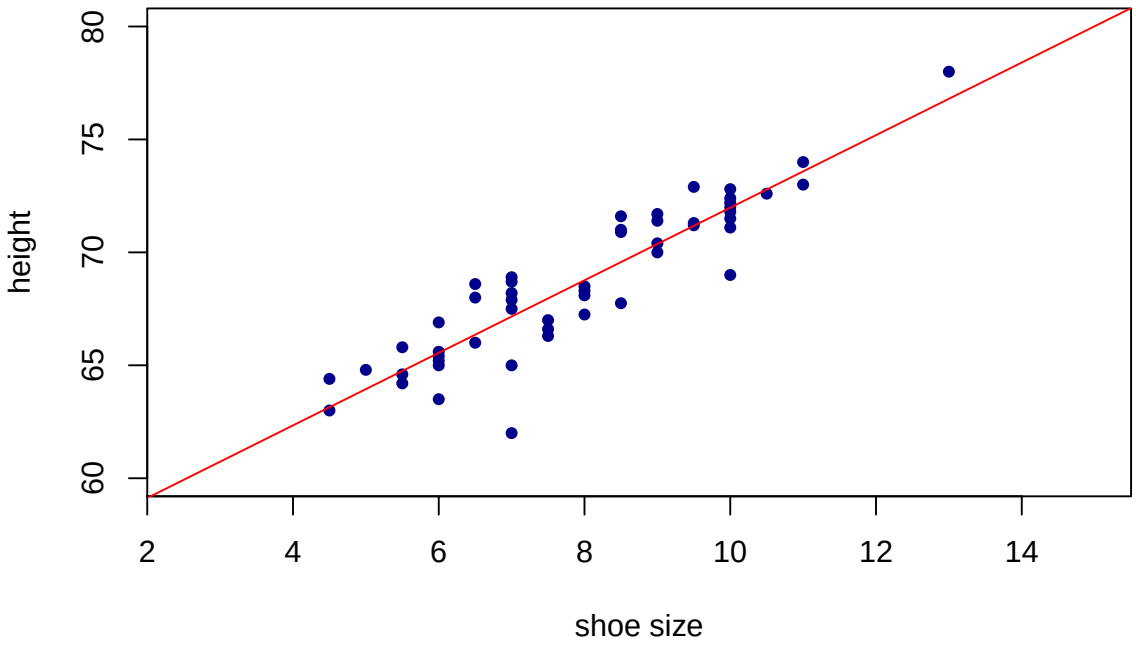


Figure 6.1: Plot of shoe-size and height data.

and

$$\mathrm{Var}^*(\beta|y) = \mathrm{Var}_0(\beta) - \mathrm{Cov}_0(\beta, y)\mathrm{Var}_0^{-1}(y)\mathrm{Cov}_0(y, \beta).$$

In order to calculate the posterior moments $[\mathrm{E}^*(\beta|y), \mathrm{Var}^*(\beta|y)]$ in a Bayes linear approach, we have to specify our partial prior belief about $\underline{\beta}$.

We use the same prior mean and prior variance as in the full-Bayes analysis,

Hence,

$$\mathrm{E}_0(Y) = X\mathrm{E}_0(\beta)$$

where,

$$X = \begin{bmatrix} 1 & 7.0 \\ 1 & 10.0 \\ \vdots & \vdots \\ 1 & 7.0 \end{bmatrix} \qquad \text{and} \qquad \mathrm{E}_0(\beta) = \begin{bmatrix} 55.667 \\ 1.667 \end{bmatrix}.$$

We also have $\mathrm{Cov}_0(\beta, y) = \mathrm{Var}_0(\beta)X'$ and $\mathrm{Var}_0(y) = X\mathrm{Var}_0(\beta)X' + \sigma^2 I_n$, where, I_n is the identity matrix with dimension 52×52.

Substituting all of this information into the Bayes linear equations we obtain posterior

moments as follows

$$\mathrm{E}^*(\beta|y) = \begin{bmatrix} 55.7072 \\ 1.5972 \end{bmatrix}$$

and

$$\mathrm{Var}^*(\beta|y) = \begin{bmatrix} 0.7122 & -0.0848 \\ -0.0848 & 0.0107 \end{bmatrix}.$$

As we can see from the results that we have obtained from applying Bayes analysis and from using Bayes linear method are exactly the same for the posterior mean vector and posterior variance and covariance matrix for β. For further information, see Appendix A.6.1.

6.3 Bayes linear kinematics

6.3.1 Probability kinematics

Jeffrey (1965) supplied a method known as probability kinematics for revising a probability specification which depends on new probabilities over a partition.

Suppose that we have a partition $K = (K_1, ..., K_n)$ and corresponding probabilities $\mathrm{Pr}_0(K_i) = p_i$ and $\sum_{i=1}^n p_i = 1$. Suppose that we have obtained some information which can cause us to update our probabilities of these events to $\mathrm{Pr}_1(K_1), ..., \mathrm{Pr}_1(K_n)$. Then we can impose the condition that, for any future event L

$$\mathrm{Pr}_0(L \mid K_i) = \mathrm{Pr}_1(L \mid K_i), \qquad \forall_i. \tag{6.4}$$

Therefore, the new marginal probability of L can be found by probability kinematics on $\mathrm{Pr}_1(K_1), ..., \mathrm{Pr}_1(K_n)$ as

$$\mathrm{Pr}_1(L) = \sum_{i=1}^{n} \mathrm{Pr}_0(L \mid K_i)\mathrm{Pr}_1(K_i)$$

However, successive revisions of this kind might not necessarily be commutative. There is some good literature which addresses the condition for commutativity. See Field (1978); Diaconis and Zabell (1982) and a simple example was given by Wilson (2011) explaining the case when we have a lack of commutativity.

6.3.2 Bayes linear kinematics

In a full-Bayes analysis, we need to specify the full joint prior density for all the unknown quantities. Then our beliefs adjusted by observing the data can be represented by the posterior distribution and we can obtain the posterior means and posterior variances.

Bayes linear kinematics is defined as a special form of Bayes linear analysis where, instead of observing $\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)'$ as we mentioned in Section 6.2, we simply update our beliefs about this set by obtaining some information. Then those changes in our beliefs can be propagated through other unknown quantities within a Bayes linear structure.

Named after probability kinematics proposed by Jeffrey (1965), Bayes linear kinematics was suggested by Goldstein and Shaw (2004) with the idea that, instead of observing α directly in equation (6.1), we observe some information about another quantity I and that changes our beliefs about α to $\mathrm{E}(\alpha|I)$ and $\mathrm{Var}(\alpha|I)$. Then we wish to propagate these updates to β.

In order to propagate these changes in our beliefs, we could use a full-Bayes analysis which requires a full probabilistic specification and more intensive calculations such as MCMC methods. However, we can use Bayes linear kinematics by adjusting the expectation vector and variance matrix based on (6.2) and therefore, we can adjust our mean and variance of $A = \alpha \cup \beta$ as

$$\begin{aligned} \mathrm{E}_{\alpha|I}(A) &= \mathrm{E}_0(A) + \mathrm{Cov}_0(A,\alpha)\mathrm{Var}_0^{-1}(\alpha)\Big[\mathrm{E}(\alpha \mid I) - \mathrm{E}_0(\alpha)\Big] \\ \mathrm{Var}_{\alpha|I}(A) &= \mathrm{Var}_0(A) - W_0(A,\alpha)\mathrm{Var}_0(\alpha)W_0'(A,\alpha) + W_0(A,\alpha)\mathrm{Var}(\alpha \mid I)W_0'(A,\alpha) \\ &= \mathrm{Var}_0(A) - W_0(A,\alpha)\left[\mathrm{Var}_0(\alpha) - \mathrm{Var}(\alpha \mid I)\right]W_0'(A,\alpha), \end{aligned} \tag{6.5}$$

where $W_0(A,\alpha) = \mathrm{Cov}_0(A,\alpha)\mathrm{Var}_0^{-1}(\alpha)$. Therefore, the equations (6.5) are called Bayes linear kinematics equations.

6.3.3 Commutativity

Now suppose we wish to make multiple updates. First we receive information I_α and that can update our beliefs about A to $Q_1(A; I_\alpha)$, where $Q_1(A; I_\alpha) = [\mathrm{E}(A|I_\alpha), \mathrm{Var}(A|I_\alpha)]$ using equations (6.5). Now afterwards, we observe some information I_β which changes

our beliefs about A to $Q_2(A; I_\alpha, I_\beta)$. Again we can use Bayes linear kinematics to gain $Q_2(A; I_\alpha, I_\beta)$.

Consider the opposite case of observing the two parts of information. Firstly, we observe I_β and update our beliefs about A to $Q_1(A; I_\beta)$ and we apply Bayes linear kinematics to gain $Q_1(A; I_\beta)$. Then later we receive information I_α which changes our beliefs about α to $Q_2(A; I_\beta, I_\alpha)$. Then we use Bayes linear kinematics to gain $Q_2(A; I_\beta, I_\alpha)$.

Now for commutativity of these two updates we should have

$$Q_2(A; I_\alpha, I_\beta) = Q_2(A; I_\beta, I_\alpha)$$

There are some necessary and sufficient conditions for a unique commutative solution in Bayes linear kinematic updates which were introduced by Goldstein and Shaw (2004). In this paper, they proposed a sufficient condition which states that, if

$$\mathrm{Var}(\alpha|I_\alpha) < \mathrm{Var}_0(\alpha) \qquad \text{or} \qquad \mathrm{Var}(\beta|I_\beta) < \mathrm{Var}_0(\beta), \tag{6.6}$$

then, there is a unique commutative solution. If this condition holds, the Bayes linear kinematic update equations can be written as

$$\begin{aligned} \mathrm{E}_{(2)}(A) = \mathrm{Var}_{(2)}(A)\Big[&\mathrm{Var}_1^{-1}(A; I_\alpha)\mathrm{E}_1(A; I_\alpha) + \mathrm{Var}_1^{-1}(A; I_\beta)\mathrm{E}_1(A; I_\beta) \\ &- \mathrm{Var}_0^{-1}(A)\mathrm{E}_0(A)\Big], \end{aligned} \tag{6.7}$$

and

$$\mathrm{Var}_{(2)}(A) = \Big[\mathrm{Var}_1^{-1}(A; I_\alpha) + \mathrm{Var}_1^{-1}(A; I_\beta) - \mathrm{Var}_0^{-1}(A)\Big]^{-1}. \tag{6.8}$$

The equations (6.7) and (6.8) give the commutative solution even if we swap the updates in those equations.

6.3.4 Multiple updates in Bayes linear kinematics

Suppose we have n collections of random quantities, $Q_1, ..., Q_n$ where

$$Q_i = (Q_{i1}, ..., Q_{in_i})'$$

for $i = 1, ..., n$. We define a full second order prior specification for $Q = Q_1 \cup ... \cup Q_n$ and put in the expression $W_0(Q) = \Big[\mathrm{E}_0(Q), \mathrm{Var}_0(Q)\Big]$ and we receive some information D_i and that changes our beliefs about Q_i to become $W_1(Q_i|D_i) = \Big[\mathrm{E}_1(Q_i|D_i), \mathrm{Var}_1(Q_i|D_i)\Big]$. To obtain the Bayes linear kinematic update for Q which depends on (6.5), we have

$$\mathrm{E}_1(Q|D_i) = \mathrm{E}_0(Q) + \mathrm{Cov}_0(Q, Q_i)\mathrm{Var}_0^{-1}(Q_i)\Big[\mathrm{E}_1(Q_i|D_i) - \mathrm{E}_0(Q_i)\Big] \tag{6.9}$$

and

$$\begin{aligned}\mathrm{Var}_1(Q|D_i) =& \mathrm{Var}_0(Q) - \mathrm{Cov}_0(Q, Q_i)\mathrm{Var}_0^{-1}(Q_i)\mathrm{Cov}_0(Q_i, Q) \\ &+ \mathrm{Cov}_0(Q, Q_i)\mathrm{Var}_0^{-1}(Q_i)\mathrm{Var}_1(Q_i|D_i)\mathrm{Var}_0^{-1}(Q_i)\mathrm{Cov}_0(Q_i, Q).\end{aligned} \tag{6.10}$$

These Bayes linear kinematic changes might not have a unique commutative solution. Goldstein and Shaw (2004) give conditions to make these updates have a unique solution. We deal with Q_i as scalar. Therefore, the sufficient condition will be

$$\mathrm{Var}_0^{-1}(Q_i)\mathrm{Var}_1(Q_i|D_i) < 1 \tag{6.11}$$

for all $i = 1, ..., n$. As a consequence, when the condition in (6.11) holds, the solution is

$$\mathrm{E}_n(Q|D) = \mathrm{Var}_n(Q|D)\left[\sum_{i=1}^{n}\mathrm{Var}_1^{-1}(Q|D_i)\mathrm{E}_1(Q|D_i) - (n-1)\mathrm{Var}_0^{-1}(Q)\mathrm{E}_0(Q)\right] \tag{6.12}$$

and

$$\mathrm{Var}_n(Q|D) = \left[\sum_{i=1}^{n}\mathrm{Var}_1^{-1}(Q|D_i) - (n-1)\mathrm{Var}_0^{-1}(Q)\right]^{-1} \tag{6.13}$$

where $\underline{D} = (D_1, ..., D_n)'$, or alternatively we can write (6.12) and (6.13) in general in the following form

$$\mathrm{P}(\underline{X} \mid I) = \sum_{j=1}^{J}\mathrm{P}(\underline{X} \mid I_j) - (J-1)\mathrm{P}(\underline{X}), \tag{6.14}$$

$$\mathrm{P}(\underline{X} \mid I)\mathrm{E}(\underline{X} \mid I) = \sum_{j=1}^{J}\mathrm{P}(\underline{X} \mid I_j)\mathrm{E}(\underline{X} \mid I_j) - (J-1)\mathrm{P}(\underline{X})\mathrm{E}(\underline{X}). \tag{6.15}$$

where, $[\mathrm{P}(\underline{X} \mid I)]^{-1}$ is the adjusted variance and $\mathrm{E}(\underline{X} \mid I)$ is the adjusted expectation, $\underline{I} = (I_1, ..., I_J)'$ and $\mathrm{P}(\underline{X}) = \mathrm{Var}(\underline{X})^{-1}$ is the prior precision matrix.

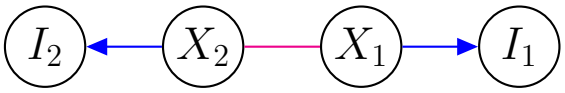


Figure 6.2: Bayes linear Bayes graphical model with two variables

6.4 Bayes linear Bayes graphical models

A Bayes linear Bayes graphical model is a combination of fully Bayesian and Bayes linear graphical models allowing conditioning on marginal distributions of any form and to take advantage of Bayes linear kinematics to involve full conditional updates within Bayes linear adjustments. See Goldstein and Shaw (2004).

In complex models, experts often make full marginal probabilistic specifications, but they are not able to assess the full joint probability distribution for all unknown quantities in the model. Goldstein and Shaw (2004) developed a formalism for updating their beliefs about these quantities which depends on Bayes linear kinematics. They introduced a directed graphical model which is a directed graph $G = (V, E)$ where $V = (X_1, X_2, ..., X_J)$ is a collection of nodes and E is a collection of edges. A Bayes graphical model is the model when the generalised conditional independence relationship is probabilistic conditional independence (Lauritzen, 1996; Cowell et al., 2007) and by taking second-order belief as the generalised conditional independence, we obtain a Bayes linear graphical model. See Goldstein and Wilkinson (2000); Goldstein and Shaw (2004).

Bayes linear Bayes graphical models are similar to Bayesian networks, where the nodes represent the parameters in the model or the random variables, and the arrows represent the relationships or the association between the parameters. In Bayesian networks, we make use of the property of conditional independence between variables given other random variables. Goldstein and Wilkinson (2000) proposed the idea of using belief separation. To demonstrate this idea, let us take the following simple example.

Suppose we have unknown quantities X_1, X_2, I_1, I_2. We can present our beliefs about (X_1, X_2) as a Bayes linear belief structure and present our beliefs about each of (X_1, I_1) and (X_2, I_2) as a full-Bayes specification. Figure 6.2 shows a simple way to represent the relationships between all the unknown quantities in a Bayes linear Bayes graphical model.

We notice from Figure 6.2 that the undirected edge between the pair (X_1, X_2) with magenta colour represents a Bayes linear structure and the directed blue arrows refer to the conditional distributions which have a full-Bayes specification.

Suppose that we have multiple unknowns $X_1, X_2, \ldots, X_J$. Let $\underline{X} = (X_1, X_2, \ldots, X_J)'$. First of all, we need to give the second order prior specification in terms of the expectations and the variances of each X_i and the covariance matrix between the variables. In addition, we have variables $I_1, I_2, \ldots, I_J$ that shall be observed where the conditional distribution of $(I_j \mid \underline{X})$ is specified probabilistically. Each I_i is associated with an element X_i of $\underline{X}$ and is conditionally independent of the rest given X_i. Then, a Bayes linear Bayes graphical model can propagate the information using full Bayes and Bayes linear kinematics. The mechanism is that, if we observe I_i, then Bayes theorem is used to calculate the $\mathrm{E}(X_i \mid I_i)$ and $\mathrm{Var}(X_i \mid I_i)$. These changes are passed through the rest of the network using the Bayes linear kinematic equations in (6.5).

We can see from Figure 6.3, that we have three main conditions here.

- If we have a set of unknown quantities, say $K = \{Z, X_1, X_2, ..., X_J, I_1, I_2, ..., I_J\}$, where $Z = X_{J+1}$, then the collection of quantities $(Z, X_1, X_2, ..., X_J)$ has a Bayes linear belief structure.
- We specify a full-Bayesian analysis for each (X_i, I_i), for $i = 1, 2, ..., J$.
- Each I_i is conditionally independent of $K \setminus \{X_i, I_i\}$ given X_i.

In the case where we have $m = 4$, a Bayes linear Bayes graphical model might be represented by Figure 6.3.

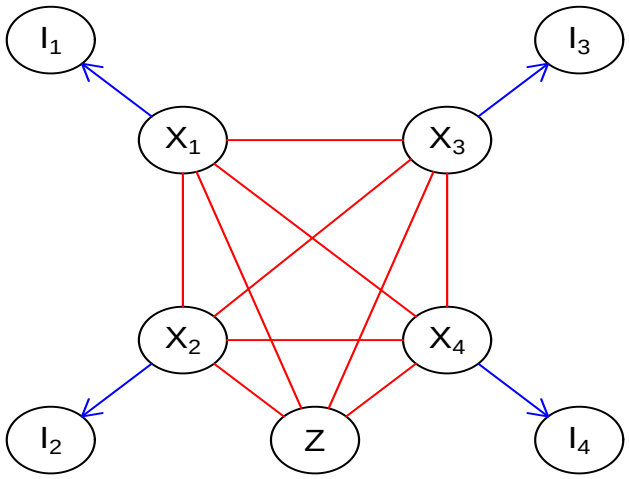


Figure 6.3: Bayes linear Bayes graphical model

6.5 Transformation of the parameters

6.5.1 Introduction

Suppose we have some distribution with a bounded parameter such as a binomial distribution, $0 < \theta_i < 1$, $(i = 1, 2, ..., n)$ or Poisson with $\theta_i > 0$. We can link the parameters $\theta_1, ..., \theta_n$ using a Bayes linear structure exactly as in Section 6.4 where $\theta_1, ..., \theta_n$ correspond to $X_1, ..., X_n$. However, there are some benefits in using a transformation for these parameters. For example we might use the logit function in the binomial case or log in the Poisson case. Then the transformed parameter for a Poisson distribution will be

$$\eta_i = \log(\theta_i)$$

or in the binomial case,

$$\eta_i = \log\left(\frac{\theta_i}{1-\theta_i}\right)$$

The transformed parameters $\eta_1, ..., \eta_n$ are then linked in a Bayes linear structure. After observing I_i that changes our prior mean and variance for η_i, these changes are propagated by using Bayes linear kinematics.

One of the reasons why we should use the transformation is that θ_i has a bounded scale and this boundary makes the use of the Bayes linear method less attractive. Therefore, doing the transformation can guarantee that we can use the linear updates without worrying about crossing the boundary. Secondly, If the parameter is bounded between 0 and 1, that means the variance will depend on the mean and in a Bayes linear analysis the variance should not depend on the mean. Thirdly, when we do not transform, the variance for θ_i might increase when we observe the data. Transforming the parameter η_i can avoid this problem and lead to a reduction in the variance when the data have been observed. For instance, suppose $I|\theta \sim \text{Poisson}(\theta)$, $\theta \sim \text{Gamma}(1, 5)$. Hence, the variance is $1/25$. Now suppose that we observe $I = 2$. Therefore the posterior density will be $\text{Gamma}(3, 6)$. As a result, the posterior variance is $(1/12) > (1/25)$. For more information, see Wilson and Farrow (2010); Wilson et al. (2013); Wilson and Farrow (2017).

6.5.2 Guide relationship

Following West et al. (1985), Wilson and Farrow (2010, 2017) use a guide relationship to suggest how we should relate the moments of η_i to those of λ_i. For example, when λ_i is the mean of Poisson distribution, we can use, as a guide, $\eta_i \approx \log(\lambda_i)$. In this case, Wilson and Farrow (2017) discuss three methods to determine the moments of η_i given those of λ_i, and vice versa. They refer to these as the log-mode method, the log-moment method and the lognormal method. Wilson and Farrow (2017) show that, in each case, if $\lambda_i \sim \text{Ga}(\alpha_i, \theta_i)$, then, for some functions h_1 and h_2, we have $\text{E}(\eta_i) = h_1(\alpha_i) - \log(\theta_i)$ and $\text{Var}(\eta_i) = h_2(\alpha_i)$.

6.5.3 Mode and log-curvature method

Once the parameters have been transformed from a bounded scale to the whole real line we hope that the posterior distribution will be close to symmetric. However, the distribution may not be straightforward. We might choose to use a conjugate prior distribution for the untransformed parameter, for example, a gamma distribution for a Poisson parameter. We then need to relate the hyperparameters of this distribution to moments on the transformed scale. Wilson and Farrow (2017) discuss three methods in the Poisson case. In this section we explain the principal idea of using one of these, the mode and log-curvature method ("log mode") for the transformed parameters. Suppose we obtain the mode of the posterior distribution which can be done by finding the first derivative of the log posterior density with respect to our parameter of interest. We will consider the case when we have a single mode, say $\hat{\eta}$ and we fit the normal distribution based on the second derivatives of the log posterior density at $\hat{\eta}$.

Let the transformed parameter be $\eta = g(\theta)$. Suppose

$$\pi(\eta \mid y) \sim N(\hat{\eta}, \text{Var}(\eta))$$

where $\text{Var}(\eta)$ is the inverse of the curvature of the log posterior density at the mode $\hat{\eta}$

$$\text{Var}(\eta) = \left[- \frac{d^2 \log f(\eta|y)}{d\eta^2} \bigg|_{\eta=\hat{\eta}} \right]^{-1}.$$

This second derivative can be calculated numerically using Newton's method as we

shall illustrate in Section 6.7.3. See Gelman et al. (2014).

We give an example explaining the method using binomial observations. Suppose we have a random variable, $Y_i \sim \text{Bin}(\text{n}, \theta)$, so θ is restricted, $0 < \theta < 1$. For Bayes linear fitting, it is better to have an unrestricted range for the parameter. Therefore we use a transformation to make sure the range of the parameter will be, $-\infty < \eta < \infty$ with

$$\eta_i = \log\Big(\frac{\theta_i}{1-\theta_i}\Big).$$

A conjugate prior for θ is a beta density, i.e. $\theta_i \sim \text{Beta}(\text{a}_\text{i}, \text{b}_\text{i})$. The idea is to find the mean of η_i which is equal to the mode of $\log[\theta_i/(1-\theta_i)]$ and the variance of η_i is the inverse of the curvature of the log density at the mode $\hat{\eta}$.

We have

$$\pi(\theta_i) = \frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}\ \theta_i^{a_i-1}\ (1-\theta_i)^{b_i-1}$$

and because $\theta_i = \Big(\frac{\text{e}^{\eta_i}}{1+\text{e}^{\eta_i}}\Big)$, $\frac{d\theta_i}{d\eta_i} = \theta_i(1-\theta_i)$.

By using the last derivative, we can find the Jacobian. As a result, the density of η_i will be

$$\begin{aligned} g(\eta_i) &= \pi(\theta_i).\Big|\frac{d\theta_i}{d\eta_i}\Big| \\ &= \frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}\ \Big(\frac{\text{e}^{\eta_i}}{1+\text{e}^{\eta_i}}\Big)^{a_i-1}\Big(\frac{1}{1+\text{e}^{\eta_i}}\Big)^{b_i-1} \times \Big(\frac{\text{e}^{\eta_i}}{1+\text{e}^{\eta_i}}\Big)\Big(\frac{1}{1+\text{e}^{\eta_i}}\Big) \\ &= \frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}\Big(\frac{\text{e}^{\eta_i}}{1+\text{e}^{\eta_i}}\Big)^{a_i}\Big(\frac{1}{1+\text{e}^{\eta_i}}\Big)^{b_i}. \end{aligned} \tag{6.16}$$

Taking log of (6.16), we have

$$\log\left[g(\eta_i)\right] = \text{Constant} + a_i \log\Big(\frac{\text{e}^{\eta_i}}{1+\text{e}^{\eta_i}}\Big) + b_i \log\Big(\frac{1}{1+\text{e}^{\eta_i}}\Big).$$

The first derivative of $\log[g(\eta_i)]$ with respect to η_i is

$$\frac{d\log\left[g(\eta_i)\right]}{d\eta_i} = \frac{a_i}{(1+\text{e}^{\eta_i})} - \frac{b_i\text{e}^{\eta_i}}{(1+\text{e}^{\eta_i})}$$

Putting the first derivative equal to 0, we obtain $\text{E}_0(\eta_i) = \log(a_i/b_i)$. In order to find

the variance of η_i, we should find the second derivative of the log density of $g(\eta_i)$ with respect to η_i which is

$$\frac{d^2 \log\left[g(\eta_i)\right]}{d\eta_i^2} = -\left[\frac{a_i \mathrm{e}^{\eta_i}}{(1+\mathrm{e}^{\eta i})^2} + \frac{b_i \mathrm{e}^{\eta_i}}{(1+\mathrm{e}^{\eta i})^2}\right].$$

By substituting $\mathrm{e}^{\eta_i} = a_i/b_i$ and doing some algebra, the variance of η_i will be

$$\mathrm{Var}_0(\eta_i) = -\left[\frac{d^2 \log\left[g(\eta_i)\right]}{d\eta_i^2}\right]^{-1} = \frac{b_i + a_i}{a_i b_i} = \frac{1}{a_i} + \frac{1}{b_i}.$$

See Wilson and Farrow (2010).

6.5.4 Log-moment method

Suppose that we have an exponential survival time with probability density function

$$f_i(t) = \lambda_i \exp\{-\lambda_i t\}, \tag{6.17}$$

and survival function

$$S_i(t) = \exp\{-\lambda_i t\}. \tag{6.18}$$

Now to make inference about the unknown parameter λ, we could give λ a gamma prior distribution and that is conjugate to the density and the survival function in (6.17) and (6.18).

The likelihood function for the individual i is

$$L_i = \lambda_i^{\delta_i} \exp\{-\lambda_i t_i\}.$$

where $\delta_i = 1$ if individual i dies and $\delta_i = 0$ if individual i is censored. The prior distribution of λ_i is gamma distribution with shape parameter α_i and scale parameter θ_i, so $\lambda_i \sim \mathrm{Ga}(\alpha_i, \theta_i)$. Therefore, the posterior density for λ_i will be $\mathrm{Ga}(\alpha_i + \delta_i, \theta_i + t_i)$.

As in the case where we have Poisson observations, $X|\lambda \sim \mathrm{Po}(\lambda)$, a suitable conjugate prior will be a gamma distribution. Wilson and Farrow (2017) proposed $\eta = g(\lambda) = \log(\lambda)$ in order to make η unbounded rather than use the bounded parameter λ. See Section 6.5.1.

If we use the guide relationship with the conjugate update of a gamma distribution, then the mean and the variance of η_i as those of $\log \lambda_i$ gives

$$\begin{aligned} \mathrm{E}_0(\eta_i) &= g_1(\alpha_i, \theta_i) = f_i = h_1(\alpha_i) - \log(\theta_i) \\ \mathrm{Var}_0(\eta_i) &= g_2(\alpha_i, \theta_i) = q_i = h_2(\alpha_i). \end{aligned} \tag{6.19}$$

In the log-moment method, we can calculate the mean and the variance of η_i and that gives us

$$h_1(\alpha_i) = \psi(\alpha_i) \qquad \text{and} \qquad h_2(\alpha_i) = \psi_1(\alpha_i)$$

where $\psi(\cdot)$ is the digamma function and $\psi_1(\cdot)$ is the trigamma function. See Wilson and Farrow (2017). These are the exact moments of η_i if $\eta_i = \log(\lambda_i)$ and $\lambda_i \sim \mathrm{Ga}(\alpha_i, \theta_i)$.

6.5.5 Lognormal method

In this method we equate the first and second moments of the gamma prior distribution for λ to those of a lognormal distribution and use the mean and variance of the corresponding normal distribution for η. We put $\alpha_i/\theta_i = \exp(f_i + q_i/2)$ and $\alpha_i/\theta_i^2 = \exp(2f_i + q_i)[\exp(q_i) - 1]$, that is giving

$$h_1(\alpha_i) = \log\left[\alpha_i \sqrt{\alpha_i/(\alpha_i + 1)}\right] \quad \text{and} \quad h_2(\alpha_i) = \log(1 + \alpha_i^{-1})$$

So,

$$\alpha_i = [\exp(q_i) - 1]^{-1} \quad \text{and} \quad \theta_i = [\exp(q_i) - 1]^{-1} \exp(-q_i/2) \exp(-f_i).$$

See Wilson and Farrow (2017).

In this thesis, we use the lognormal method in the leukaemia example to compare the three methods, full-Bayes method, Bayes linear kinematic methods using conjugate prior update and BLK using non-conjugate prior update.

6.6 Example: Sulfinpyrazone

The Anturane Reinfarction Trial Research Group (1980) reported a clinical trial on the use of the drug sulfinpyrazone in patients who had suffered myocardial infarctions ("heart

Groups	Death	Total
1	44	560
2	62	540

Table 6.1: Sulfinpyrazone example

attacks"). The idea was to see whether the drug had an effect on the number dying. Patients in one group were given the drug while patients in another group were given a "placebo", that is an inactive substitute. Table 6.1 gives the number of all "analysable" deaths up to 24 months after the myocardial infarction and the total number of eligible patients who were not withdrawn and did not suffer a "non-analysable" death during the study. We present the data in Table 6.1 as a bar graph in Figure 6.4.

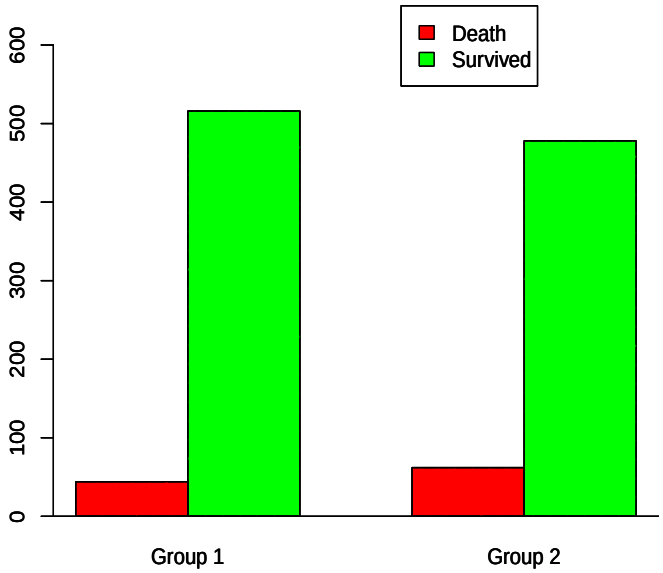


Figure 6.4: Bar plot for the two groups in sulfinpyrazone example.

We have X_i out of n_i die in group i. Hence, $X_i \mid \theta_i \sim \text{Bin}(n_i, \theta_i)$ and in our beliefs we do not regard θ_1, θ_2 as independent.

Let $\underline{\theta} = (\theta_1, \theta_2)'$. For illustrative comparison with the Bayes linear kinematic approach we first show a full-Bayes analysis. Let $\underline{\eta} = (\eta_1, \eta_2)'$ where $\eta_1 = \log[\theta_1/(1-\theta_1)]$ and $\eta_2 = \log[\theta_2/(1-\theta_2)]$. We give η_1, η_2 a bivariate normal prior distribution with $\text{E}(\eta_1) = \text{E}(\eta_2)$.

Suppose that θ_2 has probability 0.95 of being in the interval $0.05 < \theta_2 < 0.20$. The

corresponding interval for η_2 is $-2.944 < \eta_2 < -1.386$. Therefore, $\mathrm{E}_0(\eta_2) = -2.165$ and the prior variance of η_2, $\mathrm{Var}_0(\eta_2) = \Big[(-1.386 + 2.944)/(2 \times 1.96)\Big]^2 = 0.397^2 = 0.158$. In order not to prejudice the analysis, we set $\mathrm{E}(\eta_1) = \mathrm{E}(\eta_2)$. The prior variance of η_1, $\mathrm{Var}_0(\eta_1) = 4 \times \mathrm{Var}_0(\eta_2) = 0.794^2 = 0.630$ and the reason behind increasing the prior standard deviation $\sigma_{\eta_1} = 2\sigma_{\eta_2}$ is because we have less uncertainty about the death rate when the placebo is given. As a result, we have a 95% symmetric prior interval for η_1, $-3.721 < \eta_1 < -0.609$ and thus, $0.02 < \theta_1 < 0.35$ which is a wider interval than that for θ_2.

As we mentioned earlier, in this example θ_1 and θ_2 are dependent in our prior beliefs. We suppose that in our prior beliefs, η_1 and η_2 are correlated with $\rho = 0.5$. Then the prior covariance value can be calculated as follows, $\mathrm{Cov}_0(\eta_1, \eta_2) = \rho\sqrt{\mathrm{Var}_0(\eta_1)\mathrm{Var}_0(\eta_2)} = \mathrm{Var}_0(\eta_2) = 0.158$.

Now, in order to use a full-Bayes analysis, we need to transform the parameters in the model θ_i to the new parameters η_i for $i = 1, 2$. It is more convenient to work with an unbounded scale than work with a (0,1) scale. See Section 6.5 for more detail. We give η_1,η_2 a bivariate normal distribution with density

$$\pi(\eta_1, \eta_2) \propto \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{\eta_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{\eta_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{(\eta_1-\mu_1)(\eta_2-\mu_2)}{\sigma_1\sigma_2}\right)\right]\right\}. \quad (6.20)$$

We observe x_i, $i = 1, 2$. The likelihood function is binomial, $X_i \sim \mathrm{Bin}(n_i, \theta_i)$

$$\mathrm{L}(\underline{\theta}; \underline{x}) = \prod_{i=1}^{n} f(x_i|\theta_i) = \prod_{i=1}^{2} \binom{n_i}{x_i} \theta_i^{x_i}(1-\theta_i)^{n_i-x_i}.$$

Then we transform the parameter θ_i to $\eta_i = \log[\theta_i/(1-\theta_i)]$ where $\eta_i \in (-\infty, \infty)$. The likelihood function for the transformed parameters will be

$$\mathrm{L}(\eta_1, \eta_2; x_i) = \binom{n_i}{x_i}\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{x_i}\left(\frac{1}{1+e^{\eta_i}}\right)^{n_i-x_i} \qquad i = 1, 2. \quad (6.21)$$

Therefore, by applying Bayes' theorem, the marginal posterior density of η_1 is expressed by combining (6.20) and (6.21). We have

$$\pi(\eta_1 \mid x_i) = \frac{\int \pi(\eta_1, \eta_2)\mathrm{L}(\eta_1, \eta_2; x_i)d\eta_2}{\int\int \pi(\eta_1, \eta_2)\mathrm{L}(\eta_1, \eta_2; x_i)d\eta_1 d\eta_2} \qquad i = 1, 2. \quad (6.22)$$

From (6.22) , we find the posterior moments for η_1 respectively using numerical integration as $\mathrm{E}(\eta_1 \mid x_1, x_2) = -2.452$ and $\mathrm{Var}(\eta_1 \mid x_1, x_2) = 0.0231$. Likewise, the posterior moments for η_2 are $\mathrm{E}(\eta_2 \mid x_1, x_2) = -2.074$ and $\mathrm{Var}(\eta_2 \mid x_1, x_2) = 0.0162$.

Figure 6.5 and 6.6 show the posterior densities for $\underline{\theta}$ and $\underline{\eta}$ respectively. See the `R` functions in Appendices A.6.2 and A.6.3.

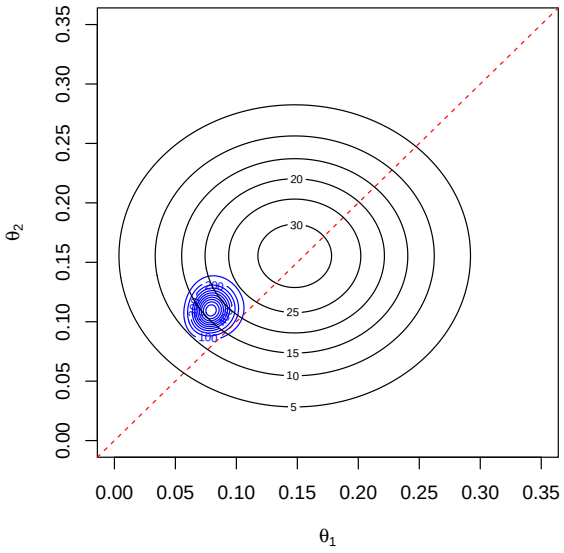


Figure 6.5: The prior (black) and posterior (blue) density of θ_1 and θ_2. The dashed line is when $\theta_1 = \theta_2$.

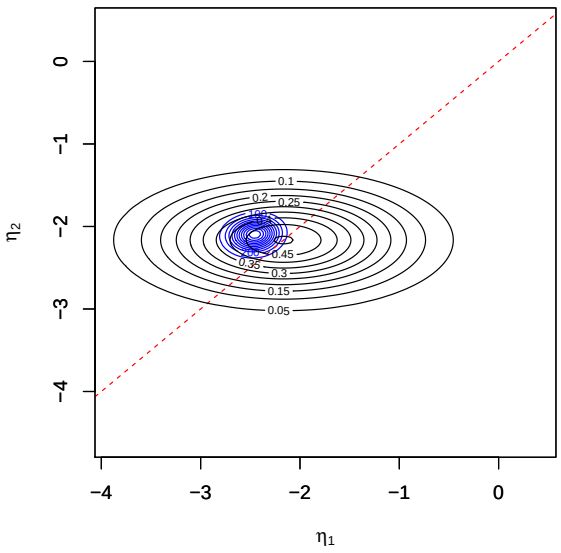


Figure 6.6: The prior (black) and posterior (blue) density of η_1 and η_2. The dashed line is when $\eta_1 = \eta_2$.

We can see from the contour plot in Figure 6.5 which represents the posterior density, that most of the probability lies in the side where $\theta_2 > \theta_1$ and that means that the death rate is probably higher with the placebo than with sulfinpyrazone. As a result we conclude that using sulfinpyrazone has a good effect on the patients. Likewise, we notice also from Figure 6.6 that most of the probability for η_1, η_2 lies in the side where $\eta_2 > \eta_1$ which corresponds with the results from Figure 6.5.

Now we describe a Bayes linear kinematic analysis of the sulfinpyrazone example. We give θ_i a beta prior distribution; $\theta_i \sim \mathrm{Beta}(a_i, b_i)$. We need to specify the parameters a_1, a_2, b_1 and b_2. In order to find a_i and b_i to specify the prior mean and variance for θ_i, we will use the prior moments given to η_1, η_2 in the full-Bayes analysis and work backward using the mode and curvature method as demonstrated in Section 6.5.3.

We have

$$\mathrm{E}_0(\eta_i) = \log\left(\frac{a_i}{b_i}\right).$$

So

$$\left(\frac{a_i}{b_i}\right) = e^{\mathrm{E}_0(\eta_i)}.$$

Hence,

$$b_i = \frac{a_i}{e^{\mathrm{E}_0(\eta_i)}}.$$

We have

$$\mathrm{Var}_0(\eta_i) = \left(\frac{1}{a_i}\right) + \left(\frac{1}{b_i}\right).$$

So

$$\mathrm{Var}_0(\eta_i) = \left(\frac{1}{a_i}\right) + \left(\frac{e^{\mathrm{E}_0(\eta_i)}}{a_i}\right)$$

and

$$a_i = \left(\frac{1 + e^{\mathrm{E}_0(\eta_i)}}{\mathrm{Var}_0(\eta_i)}\right).$$

Hence, $a_1 = 1.77$, $b_1 = 15.42$, $a_2 = 7.07$ and $b_2 = 61.61$.

By applying Bayes' theorem, our posterior density of θ_i is Beta$(a_i + x_i, n_i + b_i - x_i)$, where (x_1, x_2) =(44,62) and (n_1, n_2) =(560,540) as shown in Table 6.1. The summary of the results is given in Table 6.2.

θ_i	$\mathrm{E}_0(\theta)$	$\mathrm{E}_1(\theta)$	$\mathrm{Var}_0(\theta)$	$\mathrm{Var}_1(\theta)$
1	0.103	0.079	0.0051	0.00013
2	0.103	0.113	0.0013	0.00017

Table 6.2: The prior means and variances for θ and the posterior means and variance using the conjugate prior

After observing x_i, we need to update our prior beliefs about η_i. Therefore, we calculate the posterior mean and variance for η_1 given x_1 as

$\mathrm{E}_1(\eta_1) = \log(A_1/B_1)$, where $A_1 = a_1 + x_1 = 45.77$ and $B_1 = b_1 + n_1 - x_1 = 531.42$. So, $\mathrm{E}_1(\eta_1) = -2.452$. Also, $\mathrm{Var}_1(\eta_1) = (\frac{1}{A_1} + \frac{1}{B_1}) = 0.0237$

Similarly, the posterior mean and variance for η_2 given x_2 are, $\mathrm{E}_1(\eta_2) = \log(A_2/B_2)$, where $A_2 = a_2 + x_2$ and $B_2 = b_2 + n_2 - x_2$. So, $\mathrm{E}_1(\eta_2) = -2.056$. Also, $\mathrm{Var}_1(\eta_2) =$

η_i	$\mathrm{E}_0(\eta)$	$\mathrm{E}_1(\eta)$	$\mathrm{Var}_0(\eta)$	$\mathrm{Var}_1(\eta)$
1	-2.165	-2.452	0.630	0.0237
2	-2.165	-2.056	0.158	0.0163

Table 6.3: The prior means and variances for η and the posterior means and variance based on using the conjugate prior update by the corresponding observations.

$(\frac{1}{A_2} + \frac{1}{B_2}) = 0.0163$. Table 6.3 summarises the above calculations.

We now carry on and apply the Bayes linear kinematic equations to the results that we have made. So, the Bayes linear kinematic equations for η_2 after observing x_1 are

$$\begin{aligned} \mathrm{E}_1(\eta_2; x_1) &= \mathrm{E}_0(\eta_2) + \frac{\mathrm{Cov}_0(\eta_1, \eta_2)}{\mathrm{Var}_0(\eta_1)}\Big[\mathrm{E}_1(\eta_1) - \mathrm{E}_0(\eta_1)\Big] \\ \mathrm{Var}_1(\eta_2; x_1) &= \mathrm{Var}_0(\eta_2) - \frac{\mathrm{Cov}_0(\eta_1, \eta_2)^2}{\mathrm{Var}_0(\eta_1)}\left[1 - \frac{\mathrm{Var}_1(\eta_1)}{\mathrm{Var}_0(\eta_1)}\right] \end{aligned} \tag{6.23}$$

and also, the Bayes linear kinematic equations for η_1 after observing x_2 are

$$\begin{aligned} \mathrm{E}_1(\eta_1; x_2) &= \mathrm{E}_0(\eta_1) + \frac{\mathrm{Cov}_0(\eta_1, \eta_2)}{\mathrm{Var}_0(\eta_2)}\Big[\mathrm{E}_1(\eta_2) - \mathrm{E}_0(\eta_2)\Big] \\ \mathrm{Var}_1(\eta_1; x_2) &= \mathrm{Var}_0(\eta_1) - \frac{\mathrm{Cov}_0(\eta_1, \eta_2)^2}{\mathrm{Var}_0(\eta_2)}\left[1 - \frac{\mathrm{Var}_1(\eta_2)}{\mathrm{Var}_0(\eta_2)}\right] \end{aligned} \tag{6.24}$$

So,

$$\begin{aligned} \mathrm{E}_1(\eta_2; x_1) &= -2.237, \qquad \mathrm{E}_1(\eta_1; x_2) = -2.056, \\ \mathrm{Var}_1(\eta_2; x_1) &= 0.1199, \qquad \mathrm{Var}_1(\eta_1; x_2) = 0.4883. \end{aligned}$$

So, this information is propagated through η_1 and η_2 using Bayes linear kinematics and we evaluate the unique commutative Bayes linear kinematic solution, as

$$\begin{aligned} \mathrm{E}_{(2)}(\eta_1; x_1, x_2) &= \mathrm{Var}_{(2)}(\eta_1; x_1, x_2)\Big[\mathrm{Var}_1^{-1}(\eta_1)\mathrm{E}_1(\eta_1) + \\ &\qquad \mathrm{Var}_1^{-1}(\eta_1; x_2)\mathrm{E}_1(\eta_1; x_2) - \mathrm{Var}_0^{-1}(\eta_1)\mathrm{E}_0(\eta_1)\Big] \\ \mathrm{Var}_{(2)}(\eta_1; x_1, x_2) &= \Big[\mathrm{Var}_1^{-1}(\eta_1) + \mathrm{Var}_1^{-1}(\eta_1; x_2) - \mathrm{Var}_0^{-1}(\eta_1)\Big]^{-1} \end{aligned} \tag{6.25}$$

and

$$\begin{aligned} \mathrm{E}_{(2)}(\eta_2; x_1, x_2) &= \mathrm{Var}_{(2)}(\eta_2; x_1, x_2)\Big[\mathrm{Var}_1^{-1}(\eta_2)\mathrm{E}_1(\eta_2) + \\ &\qquad \mathrm{Var}_1^{-1}(\eta_2; x_1)\mathrm{E}_1(\eta_2; x_1) - \mathrm{Var}_0^{-1}(\eta_2)\mathrm{E}_0(\eta_2)\Big] \\ \mathrm{Var}_{(2)}(\eta_2; x_1, x_2) &= \Big[\mathrm{Var}_1^{-1}(\eta_2) + \mathrm{Var}_1^{-1}(\eta_2; x_1) - \mathrm{Var}_0^{-1}(\eta_2)\Big]^{-1} \end{aligned} \tag{6.26}$$

That gives us,

$$\begin{aligned} \mathrm{E}_{(2)}(\eta_1; x_1, x_2) &= -2.439, \qquad \mathrm{E}_{(2)}(\eta_2; x_1, x_2) \quad = -2.071, \\ \mathrm{Var}_{(2)}(\eta_1; x_1, x_2) &= 0.0234, \qquad \mathrm{Var}_{(2)}(\eta_2; x_1, x_2) \quad = 0.0158. \end{aligned}$$

These results differ only slightly from those obtained by the full-Bayes analysis. Alternatively, we can write (6.25) and (6.26) as a vector $\underline{\eta} = (\eta_1, \eta_2)'$ in the following expression

$$\mathrm{E}_2(\eta|x) = \mathrm{Var}_2(\eta|x)\left[\sum_{i=1}^{2} \mathrm{Var}_1^{-1}(\eta|x_i)\mathrm{E}_1(\eta|x_i) - \mathrm{Var}_0^{-1}(\eta)\mathrm{E}_0(\eta)\right] \tag{6.27}$$

$$\mathrm{Var}_2(\eta|x) = \left[\sum_{i=1}^{2} \mathrm{Var}_1^{-1}(\eta|x_i) - \mathrm{Var}_0^{-1}(\eta)\right]^{-1} \tag{6.28}$$

where $\underline{x} = (x_1, x_2)'$.

Therefore, the following *algorithm* is to find the Bayes linear kinematic adjusted moments of η_1, η_2.

1. Find a_1, a_2, b_1, b_2.
2. Find A_1 and B_1 and A_2 and B_2 using the data x_1 and n_1 and x_2 and n_2.
3. Find $\mathrm{E}_1(\eta_1 \mid x_1)$ and $\mathrm{Var}_1(\eta_1 \mid x_1)$ from A_1 and B_1 and $\mathrm{E}_1(\eta_2 \mid x_2)$, $\mathrm{Var}_1(\eta_2 \mid x_2)$ from A_2 and B_2.
4. Use BLK formulae in (6.27) and (6.28).

So, the information is propagated through η_1 and η_2 to update our beliefs having observed x_1 and x_2 , as we can see from Figure 6.7.

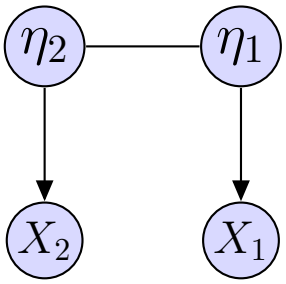


Figure 6.7: Bayes linear Bayes graphical model to update our belief about η_1 and η_2.

6.7 Bayes linear Bayes models with non-conjugate marginal priors

6.7.1 Introduction

In Chapter 7, we will describe a novel Bayes linear Bayes prognostic network. This will be able to use data from covariates of many types, including those where there is no convenient conjugate prior. Furthermore, even when there is a convenient conjugate prior, we will see that there can be advantages in using a prior of a different, non-conjugate, form. Therefore, in this section we introduce a new extension to the theory of Bayes linear kinematics to allow the use of non-conjugate marginal priors.

It is well known that, in the Bayesian framework, if the posterior density and the prior density have the same family of distribution such as normal or beta or gamma, then we denote this prior as a conjugate prior. However, in many applications, the prior and the posterior do not belong to the same family. In other words, the prior has a different type of distribution from the posterior, so this prior is called a non-conjugate prior. This thesis aims to use non-conjugate priors in order to apply Bayes linear kinematics and that can be done using one-dimensional numerical integration and we compare this approach with one using conjugate priors and with a Bayesian analysis using MCMC and a fully specified joint prior distribution.

Wilson (2011), in his thesis, explained in detail how to use Bayes linear Bayes analysis with conjugate marginal priors and he tackled two distributions of observations, binomial and Poisson. However, sometimes this conjugacy condition does not hold. For example, in the case where we have a binomial likelihood, we might use a logit-normal prior. This kind of marginal posterior distribution does not have a closed form. Therefore, we need

to use some numerical integration method such as quadrature or Laplace approximation method to evaluate the posterior mean and the posterior variance. Those posterior means and variances are very important to calculate Bayes linear kinematics.

The use of non-conjugate marginal updates in Bayes linear kinematics is a new feature which is introduced for the first time in this thesis. As well as extending the range of types of variable which can be accommodated in a Bayes linear Bayes model, it avoids the need to pass updated hyperparameters through a link function, or guide relation, by instead updating moments of unknown quantities directly on the Bayes linear scale. Although numerical integration is required, this is typically one-dimensional.

In this thesis we aim to compare the posterior means, variances and covariances using three methods, full-Bayes analysis, Bayes linear kinematics with conjugate prior and Bayes linear kinematics with non-conjugate marginal prior distributions. We will demonstrate in Subsection 6.7.3 how to calculate the posterior using a Laplace approximation method.

6.7.2 Non-conjugate marginal priors

While there are advantages in using a link function, as in Wilson and Farrow (2010, 2017), the use of a conjugate prior and then calculation of the change in mean and variance of a transformed parameter is a somewhat restrictive arrangement. Removing the requirement for a conjugate prior allows many different kinds of observational distributions. The price to be paid for this is that we need to use a numerical integration to find the adjusted mean and variance of X_j given $Y_j = y_j$. However this is typically a one-dimensional integration and suitable fast approximation methods can often be used. Suppose that we give X_j a prior distribution with density $f_j(x)$ and that the likelihood from observing $Y_j = y_j$ is $L_j(x;\ y_j)$. Then the posterior density of X_j is proportional to $g_j(x) = f_j(x)L_j(x;\ y_j)$. For example, in the Poisson case, we might give a normal prior distribution to $\eta = \log \lambda$ and the likelihood is proportional to $\exp(-\lambda)\lambda^{y_j}$.

In suitable cases, particularly where the support is unbounded, we might use a simple normal approximation. Let $G_j(x) = \log g_j(x)$. Then we can finding the maximum m of $G_j(x)$ as an approximation to the posterior mean and use $v = -[\partial^2 G_j(x)/\partial x^2]^{-1}$ evaluated at $x = m$ as an approximation to the posterior variance. Alternatively, we write the posterior mean as

$$\mathrm{E}_1(X_j \mid Y_j = y_j) = \frac{\int_{-\infty}^{\infty} x g_j(x)\ dx}{\int_{-\infty}^{\infty} g_j(x)\ dx} = \frac{\int_{-\infty}^{\infty} \exp\{\log(x) G_j(x)\}\ dx}{\int_{-\infty}^{\infty} \exp\{G_j(x)\}\ dx}$$

and then use Laplace approximations (Tierney and Kadane, 1986) for the integrals in the numerator and denominator. Another possibility would be to use Gauss-Hermite quadrature (e.g. Naylor and Smith, 1982).

6.7.3 Finding the marginal posterior by Laplace approximation

As a first step in summarising a posterior density, we might seek the posterior mode. In practice, we look for one single posterior mode and, if the posterior density is symmetric and unimodal, this locates the centre of the distribution. However, if the posterior density is not symmetric and unimodal then the posterior mode is a poor point summary. In order to make sure that our posterior mode is unique, we should run a mode-finding algorithm with different initial values.

We can use the Newton-Raphson method to find the posterior mode. See Gelman et al. (2014). It is an iterative method which depends on a quadratic Taylor series approximation of the log posterior density,

$$\mathrm{q}(\theta) = \log\left[f(\theta \mid \underline{y})\right],$$

where $f(\theta|\underline{y}) = k^{-1}\pi(\theta)f(\underline{y}|\theta)$ is the posterior density and $k = \int \pi(\theta)f(\underline{y}|\theta)d\theta$ and k does not depend on θ. Then,

$$\mathrm{q}(\theta) = \log(k^{-1}) + \log\left[\pi(\theta)\right] + \log\left[f(\underline{y}|\theta)\right].$$

In order to to find the posterior mode of θ, we should find the first derivative of $\mathrm{q}(\theta)$ with respect to θ and set the first derivative equal to 0.

$$\mathrm{q}^{'}(\theta) = \frac{d\mathrm{q}(\theta)}{d\theta} = \frac{d\log\left[\pi(\theta)\right]}{d\theta} + \frac{d\log\left[f(\underline{y}|\theta)\right]}{d\theta} = 0 \tag{6.29}$$

Sometimes (6.29) does not have a closed form solution for the posterior mode which means that there is no analytical solution for this equation. Therefore, we have to use a numerical optimisation method such as Newton's method as follows.

We start with an initial value $\theta = \theta_0$ and carry on iterating:

$$\theta_{i+1} = \theta_i - \frac{\mathrm{q}^{'}(\theta_{\mathrm{i}})}{\mathrm{q}^{''}(\theta_{\mathrm{i}})}, \qquad i = 1, 2, ... \tag{6.30}$$

where $\mathrm{q}''(\theta)$ is the second derivative of $\mathrm{q}(\theta)$. The sequence $\theta_0, \theta_1, \theta_2, ...$ finally converges to the optimal solution $\hat{\theta}$.

As we mentioned in Section 3.3, there are different techniques to compute the posterior mean $\mathrm{E}(\eta|x)$ in the one-dimensional case. For example, we could use a Laplace approximation to approximate the posterior mean. Using (3.4) we obtain

$$-nk^*(\eta) = \log(\eta) + \log\left[\mathrm{L}(\eta|\underline{x})\right] + \log\left[\pi(\eta)\right].$$

by using the steps that we explained in Subsection 3.3.2. See Tierney and Kadane (1986).

6.7.4 Binomial observations

Suppose we have a random variable X which has a binomial distribution with parameter θ which represents the probability of success, so $X \sim \mathrm{Bin}(n, \theta)$. It is better to transform $\theta \in (0, 1)$ to $\eta \in (-\infty, \infty)$, for the reasons that we mentioned in Section 6.5. The likelihood function will be

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

So, the likelihood function for the transformed parameter η using the logit link function is

$$f(x \mid \eta) = \binom{n}{x} \left(\frac{e^\eta}{1+e^\eta}\right)^x \left(\frac{1}{1+e^\eta}\right)^{n-x}.$$

Suppose that, the prior density for η has a normal distribution with mean μ and precision $\tau = 1/\sigma^2$. Then its density is

$$\pi(\eta) \propto e^{-\frac{\tau}{2}[(\eta-\mu)^2]}.$$

Using Bayes theorem, the posterior density of η is

$$f(\eta|x) \propto \left(\frac{1}{1+e^\eta}\right)^n e^{-\frac{\tau}{2}[(\eta-\mu)^2]} e^{\eta x}.$$

The log posterior density of η can be written as

$$\log\left[f(\eta|x)\right] = k + n\log\left(\frac{1}{1+e^{\eta}}\right) - \frac{\tau}{2}[(\eta-\mu)^2] + \eta x$$

where k is constant. Then, the first derivative w.r.t. η is

$$\frac{d\log\left[f(\eta|x)\right]}{d\eta} = -n\left(\frac{e^{\eta}}{1+e^{\eta}}\right) - \tau(\eta-\mu) + x.$$

Putting the first derivative equal to 0, we have, $x - ne^{\eta}/(1+e^{\eta}) - \tau(\eta-\mu) = 0$. This equation does not have a closed form solution for the posterior mode for η, so we should use the Newton-Raphson method to obtain the mode $\hat{\eta}_{mode}$. After obtaining the posterior mode, we have to substitute it in $\log\left[f(\eta|x)\right]$ because we need it to evaluate the denominator of (3.4).

The second derivative for the log posterior density will be

$$\frac{d^2\log\left[f(\eta|x)\right]}{d\eta^2} = -\frac{ne^{\eta}}{(1+e^{\eta})^2} - \tau = -\left(\frac{\tau(1+e^{\eta})^2 + ne^{\eta}}{(1+e^{\eta})^2}\right).$$

Then, we can calculate the variance of η using the second derivative

$$\mathrm{Var}(\eta) = -\left[\frac{d^2\log\left[f(\eta|x)\right]}{d\eta^2}\right]^{-1} = \frac{(1+e^{\hat{\eta}_{mode}})^2}{\tau(1+e^{\hat{\eta}_{mode}})^2 + ne^{\hat{\eta}_{mode}}}.$$

Likewise, we repeat the calculations above to approximate the numerator of (3.4). We have,

$$f^*(\eta|x) \propto \eta\left(\frac{1}{1+e^{\eta}}\right)^n e^{-\frac{\tau}{2}[(\eta-\mu)^2]} e^{\eta x}$$

where, $f^*(\eta|x) = \eta f(\eta|x)$. Hence, $\log\left[f^*(\eta|x)\right] = \log(\eta) + \log\left[f(\eta|x)\right]$.

Finally, finding the first derivative w.r.t. η and equating it to 0, we have

$$\frac{d\log\left[f^*(\eta|x)\right]}{d\eta} = \frac{1}{\eta} - n\left(\frac{e^{\eta}}{1+e^{\eta}}\right) - \tau(\eta-\mu) + x = 0$$

Again, we have to solve this numerically to find $\hat{\eta}_m$, the posterior mode for $f^*(\eta|x)$.

The second derivative for $\log[f^*(\eta|x)]$ is

$$\frac{d^2 \log[f(\eta|x)]}{d\eta^2} = -\frac{1}{\eta^2} - \frac{ne^\eta}{(1+e^\eta)^2} - \tau.$$

To calculate the variance of η we use the second derivative. So

$$\mathrm{Var}^*(\eta) = -\left[\frac{d^2 \log[f^*(\eta|x)]}{d\eta^2}\right]^{-1} = -\left[-\frac{1}{\hat{\eta}_m^2} - \frac{ne^{\hat{\eta}_m}}{(1+e^{\hat{\eta}_m})^2} - \tau\right]^{-1}.$$

Now, the posterior mean for η using the Laplace approximation is

$$\mathrm{E}(\eta|x) \approx \frac{\sqrt{\mathrm{Var}^*(\eta)}}{\sqrt{\mathrm{Var}(\eta)}} \frac{\exp\{\log[f^*(\hat{\eta}_m|x)]\}}{\exp\{\log[f(\hat{\eta}_{mode}|x)]\}}. \tag{6.31}$$

So, we can rewrite (6.31) as

$$\mathrm{E}(\eta|x) \approx \frac{\sqrt{\mathrm{Var}^*(\eta)}}{\sqrt{\mathrm{Var}(\eta)}} \frac{f^*(\hat{\eta}_m|x)}{f(\hat{\eta}_{mode}|x)}$$

because $\exp\{\log[f(\hat{\eta}_m|x)]\} = f(\hat{\eta}_m|x)$.

Similarly, we can find the posterior expectation for η^2, $\mathrm{E}(\eta^2|x)$ and the posterior variance, $\mathrm{Var}(\eta|x) = \mathrm{E}(\eta^2|x) - [\mathrm{E}(\eta|x)]^2$.

Alternatively, we could use a simple method based on finding the mode and the curvature of $\log[f(\eta|x)]$ and use these to approximate the mean and variance, *i.e.*, the mean is the mode $\hat{\eta}_{mode}$, and the variance is

$$-\left(\frac{d^2 f(\eta|x)}{d\eta^2}\right)^{-1}_{\hat{\eta}_{mode}}.$$

6.7.5 Poisson observations

Suppose we have a random variable X which has a Poisson distribution with the parameter θ, $X \sim \mathrm{Poisson}(\theta)$. The probability density function will be

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x\,!} \qquad \theta > 0.$$

It is more convenient to transform the parameter of the Poisson distribution θ using $\eta_i = \log(\theta_i)$.

As in the case of binomial data, we explain the method of applying the non-conjugate prior to obtain the marginal posterior mean using a Laplace approximation. We have data which is given by Davies and Goldsmith (1972) and reproduced in Hand et al. (1994). The data represent the number of failures of piston rings in two compressors and are shown in Table 6.4.

Compressor	Failures
1	46
2	33

Table 6.4: Piston ring Failures in two compressors

We specify our prior mean and variance for η_1 to be $\mathrm{E}_0(\eta_1) = 3.384$ and $\mathrm{Var}_0(\eta_1) = 0.0340$. So the Poisson likelihood function for compressor 1 is

$$f(x_1|\eta_1) \propto e^{46\eta_1} e^{-e^{\eta_1}}$$

and the prior for η_1 is

$$\eta_1 \sim N(3.384, 0.0340).$$

We also suppose that $\mathrm{E}_0(\eta_2) = \mathrm{E}_0(\eta_1) = 3.384$ and $\mathrm{Var}_0(\eta_2) = \mathrm{Var}_0(\eta_1) = 0.0340$ and the covariance between η_1 and η_2 is $\mathrm{Cov}_0(\eta_1, \eta_2) = \rho\sqrt{\mathrm{Var}_0(\eta_1)\mathrm{Var}_0(\eta_2)} = 0.008475$ where $\rho = 0.25$.

So the prior mean vector and the prior variance covariance matrix for η are respectively $\mathrm{E}_0(\eta) = (3.384, 3.384)'$ and

$$\mathrm{Var}_0(\eta) = \begin{bmatrix} 0.0340 & 0.008475 \\ 0.008475 & 0.0340 \end{bmatrix}.$$

The marginal posterior mean for η_1 using the Laplace approximation is

$$\mathrm{E}(\eta_1|x_1) = \frac{\int \eta_1 f(x_1|\eta_1)\pi(\eta_1)d\eta_1}{\int f(x_1|\eta_1)\pi(\eta_1)d\eta_1}$$

Consider the log of the numerator of the posterior mean,

$$\mathrm{L}^*(\eta_1) = \log(\eta_1) - 14.75(\eta_1 - 3.384)^2 + 46\eta_1 - e^{\eta_1},$$

where $\mathrm{L}^*(\eta_1)$ is the log of the numerator for the expectation. We need to differentiate $\mathrm{L}^*(\eta_1)$ with respect to η_1 and put that derivative equal to zero in order to find the mode $\hat{\eta}_1^*$. So, we obtain $\hat{\eta}_1^* = 3.6493$ numerically, and substituting $\hat{\eta}_1^*$ in $\mathrm{L}^*(\eta_1)$ to evaluate this log likelihood function, we have $\mathrm{L}^*(\hat{\eta}_1^*) = 129.6765$.

Now to find the variance, we should obtain the second derivative of $\mathrm{L}^*(\eta_1)$ with respect to η_1 and substitute $\hat{\eta}_1^* = 3.6493$. So the variance is

$$\mathrm{Var}(\eta_1^*) = -\left[\frac{d^2\mathrm{L}^*(\eta_1)}{d\eta_1^2}\Big|_{\eta_1=\hat{\eta}_1^*}\right]^{-1} = 0.147.$$

So the approximate value for the numerator for the expectation is

$$\sigma^* e^{\mathrm{L}^*(\hat{\eta}_1^*)} = 2.520404 \times 10^{55}$$

where $\sigma^* = \sqrt{\mathrm{Var}(\eta_1^*)}$. Now we can repeat the similar steps to approximate the integral in the denominator. As a result, we have

$$\mathrm{E}(\eta_1|x_1) \simeq \frac{2.520404 \times 10^{55}}{6.921637 \times 10^{54}} \simeq 3.6413$$

and

$$\mathrm{Var}(\eta_1|x_1) \simeq \mathrm{E}(\eta_1^2|x_1) - [\mathrm{E}(\eta_1|x_1)]^2 = 0.0144$$

Table 6.5 shows the posterior means and the posterior variances for $\eta = (\eta_1, \eta_2)$ using the Laplace approximation method.

Compressor	$i = 1$	$i = 2$
$\mathrm{E}(\eta_i\|x_i)$	3.641	3.439
$\mathrm{Var}(\eta_i\|x_i)$	0.0144	0.0160

Table 6.5: Posterior means and posterior variance using Laplace approximation

Now these changes are propagated through to the other group using Bayes linear

kinematics and that gives us

$$\mathrm{E}_1(\eta|x_i) = \mathrm{E}_0(\eta) + \mathrm{Cov}_0(\eta, \eta_i)\mathrm{Var}_0^{-1}(\eta_i)\Big[\mathrm{E}_1(\eta_i|x_i) - \mathrm{E}_0(\eta_i)\Big]$$

and

$$\begin{aligned}\mathrm{Var}_1(\eta|x_i) =& \mathrm{Var}_0(\eta) - \mathrm{Cov}_0(\eta, \eta_i)\mathrm{Var}_0^{-1}(\eta_i)\mathrm{Cov}_0(\eta_i, \eta) \\ &+ \mathrm{Cov}_0(\eta, \eta_i)\mathrm{Var}_0^{-1}(\eta_i)\mathrm{Var}_1(\eta_i|x_i)\mathrm{Var}_0^{-1}(\eta_i)\mathrm{Cov}_0(\eta_i, \eta).\end{aligned}$$

So that our solution is a unique commutative solution, the condition $\mathrm{Var}_0^{-1}(\eta_i)\mathrm{Var}(\eta_i|x_i) < 1$ should hold. Clearly it does.

So, having observed $x = (x_1, x_2)'$, the Bayes linear kinematic commutative solution

$$\mathrm{E}_2(\eta|x) = \mathrm{Var}_2(\eta|x)\left[\sum_{i=1}^{2}\mathrm{Var}_1^{-1}(\eta|x_i)\mathrm{E}_1(\eta|x_i) - \mathrm{Var}_0^{-1}(\eta)\mathrm{E}_0(\eta)\right]$$

$$\mathrm{Var}_2(\eta|x) = \left[\sum_{i=1}^{2}\mathrm{Var}_1^{-1}(\eta|x_i) - \mathrm{Var}_0^{-1}(\eta)\right]^{-1}$$

So, we have

$$\mathrm{E}_2(\eta|x) = (3.643, 3.468)'$$

and

$$\mathrm{Var}_2(\eta|x) = \begin{bmatrix}0.0142 & 0.0017\\ 0.0017 & 0.0157\end{bmatrix}.$$

6.8 Example 1: Sulfinpyrazone with non-conjugate marginal priors

We return to the sulfinpyrazone example which we described in Section 6.6. We have the prior means and prior variances-covariance, $\mathrm{E}_0(\eta_1) = \mathrm{E}_0(\eta_2) = -2.165$, $\mathrm{Var}_0(\eta_1) = 0.630, \mathrm{Var}_0(\eta_2) = 0.158$ and $\mathrm{Cov}_0(\eta_1, \eta_2) = 0.158$. This time we simply give η_1 and η_2 normal prior distributions.

Now, in order to do BLK, we need to obtain $\mathrm{E}_1(\eta_1|x_1), \mathrm{E}_1(\eta_2|x_2), \mathrm{Var}_1(\eta_1|x_1)$ and $\mathrm{Var}_1(\eta_2|x_2)$. To obtain these posterior means and variances, we use a non-conjugate

marginal prior with one-dimensional numerical integration. Therefore, we can use a Laplace approximation. The posterior means and variances for η_1 and η_2 are shown in Table 6.6.

Group	$i = 1$	$i = 2$
$\mathrm{E}(\eta_i \vert x_i)$	-2.4527	-2.0561
$\mathrm{Var}(\eta_i \vert x_i)$	0.0234	0.0164

Table 6.6: Posterior means and variances for η.

Now, we use Bayes linear kinematics to propagate these changes to the other group which gives us

$$\mathrm{E}_1(\eta|x_1) = (-2.4527, -2.2372),$$
$$\mathrm{E}_1(\eta|x_2) = (-2.0561, -2.0561).$$

and

$$\mathrm{Var}_1(\eta|x_1) = \begin{bmatrix} 0.0234 & 0.0059 \\ 0.0059 & 0.1198 \end{bmatrix},$$
$$\mathrm{Var}_1(\eta|x_2) = \begin{bmatrix} 0.4884 & 0.0164 \\ 0.0164 & 0.0164 \end{bmatrix}.$$

For a unique commutative solution, we have

$$\mathrm{E}_2(\eta|x) = \mathrm{Var}_2(\eta|x)\left[\sum_{i=1}^{2} \mathrm{Var}_1^{-1}(\eta|x_i)\mathrm{E}_1(\eta|x_i) - \mathrm{Var}_0^{-1}(\eta)\mathrm{E}_0(\eta)\right]$$

and

$$\mathrm{Var}_2(\eta|x) = \left[\sum_{i=1}^{2} \mathrm{Var}_1^{-1}(\eta|x_i) - \mathrm{Var}_0^{-1}(\eta)\right]^{-1}.$$

So, we have

$$\mathrm{E}_2(\eta|x) = (-2.4445, -2.0691)'$$

and

$$\mathrm{Var}_2(\eta|x) = \begin{bmatrix} 0.0232 & 0.0008 \\ 0.0008 & 0.0159 \end{bmatrix}.$$

We notice that $\mathrm{E}_2(\eta|x)$ and $\mathrm{Var}_2(\eta|x)$ are very close to those obtained by the full-Bayes analysis.

6.9 Example 2: Surgical deaths

6.9.1 Data and model

The data in this example were introduced by Mosteller and Tukey (1977) and reproduced in Hand et al. (1994). The data are collected from two areas in the United States and describe the number of patients classified by sex and age. Table 6.7 shows the number of patients under the surgical operations and the patients who die in one area. The plot of the data is given in Figure 6.8.

In this model, we denote the number of deaths in age group g, area a and sex s as $Y_{g,a,s}$.

$$\text{So} \qquad Y_{g,a,s} \sim \mathrm{Bin}(n_{g,a,s}, \mathrm{P}_{g,a,s,})$$

and

$$\log\left(\frac{\mathrm{P}_{g,a,s}}{1-\mathrm{P}_{g,a,s}}\right) = \eta_{g,a,s}.$$

So we can write $\hat{\eta}$ as an estimate, $\hat{\eta} = \log\left[\hat{\mathrm{P}}/(1-\hat{\mathrm{P}})\right]$ where

$$\hat{\mathrm{P}} = \frac{\text{number of deaths}}{\text{number of patients}},$$

i.e. $\hat{\mathrm{P}} = y_{g,a,s}/n_{g,a,s}$. We also define the covariate x_g to be $\bar{x}_g - 40$ where $\bar{x}_g$ is the age-group midpoint.

We need to specify the prior mean vector and prior variances and covariances matrix for η, $S_0(\eta) = [\mathrm{E}_0(\eta), \mathrm{Var}_0(\eta)]$.

We follow Farrow (2003) to construct a more structured model.

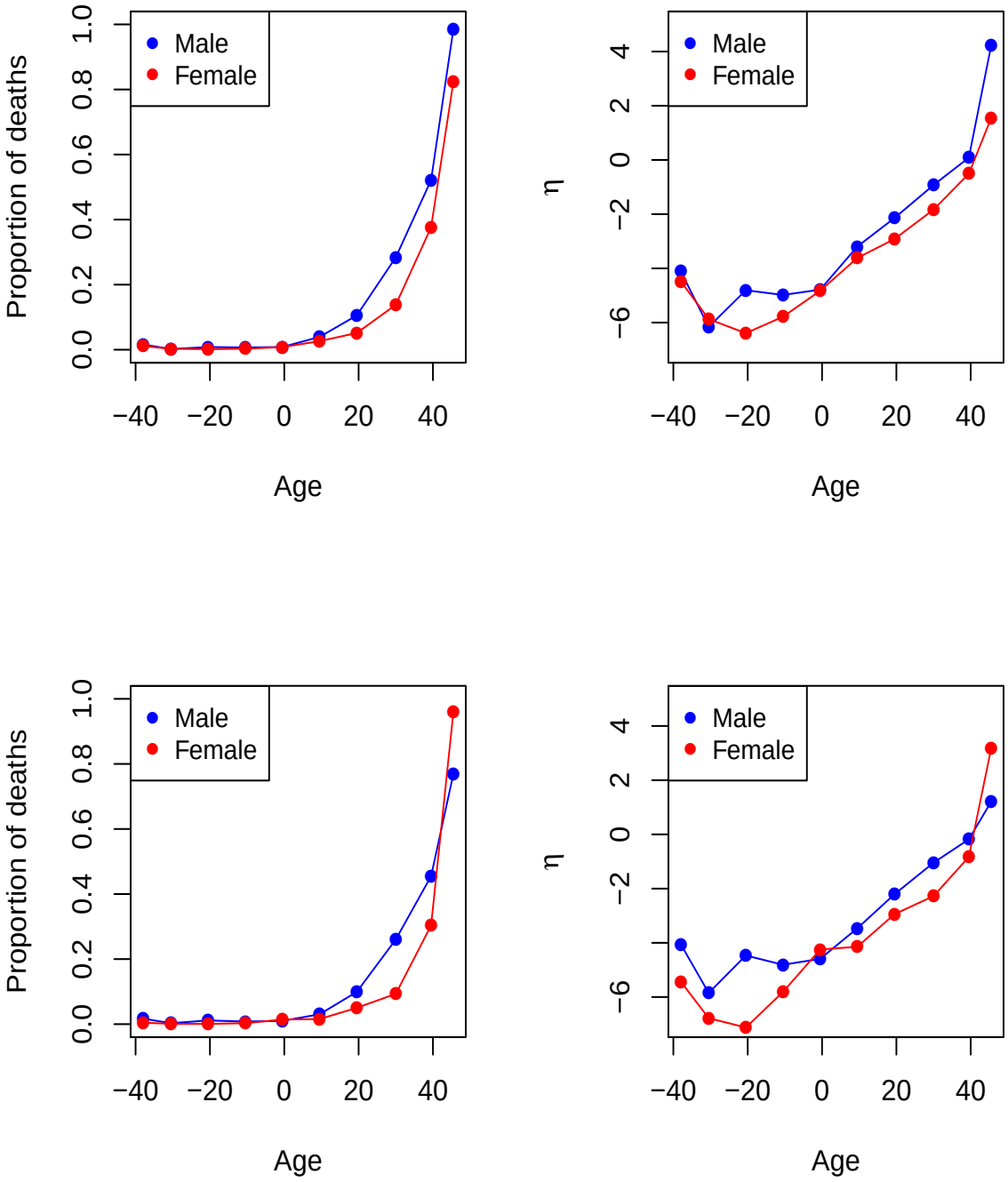


Figure 6.8: Proportion of death for males and females in area 1 against age [Top Left]. A plot of $\hat{\eta} = \log\left[\hat{\mathrm{P}}/(1-\hat{\mathrm{P}})\right]$ for both males and females in area 1 against age [Top Right]. Proportion of death for males and females in area 2 against age [Bottom Left]. A plot of $\hat{\eta} = \log\left[\hat{\mathrm{P}}/(1-\hat{\mathrm{P}})\right]$ for both males and females in area 2 against age [Bottom Right].

Area 1

	Total undergoing surgery		Number dying	
Age	Males	Females	Males	Females
0-4	2104	1952	34	22
5-14	4272	3911	9	11
15-24	2835	2989	23	5
25-34	2785	2606	19	8
35-44	1930	1886	16	15
45-54	1497	1524	59	40
55-64	960	1013	101	52
65-75	652	855	185	118
76-83	186	287	97	108
>83	69	125	68	103

Table 6.7: Death rates amongst subjects classified by age and sex

Therefore,

$$\eta_{g,a,s} = \beta_{a,s} + \gamma_{a,s} x_g + w_{g,a,s} \tag{6.32}$$

where $w_{g,a,s}$ is a specific uncertainty factor for group g, a, s.

We construct the coefficients β and γ respectively as follows:

$$\beta_{1,1} = \beta_0 + \beta_a + \beta_s + \delta_{1,1},$$

$$\beta_{2,1} = \beta_0 - \beta_a + \beta_s + \delta_{2,1},$$

$$\beta_{1,2} = \beta_0 + \beta_a - \beta_s + \delta_{1,2},$$

$$\beta_{2,2} = \beta_0 - \beta_a - \beta_s + \delta_{2,2},$$

$$\gamma_{1,1} = \gamma_0 + \gamma_a + \gamma_s + \kappa_{1,1},$$

$$\gamma_{2,1} = \gamma_0 - \gamma_a + \gamma_s + \kappa_{2,1},$$

$$\gamma_{1,2} = \gamma_0 + \gamma_a - \gamma_s + \kappa_{1,2},$$

$$\gamma_{2,2} = \gamma_0 - \gamma_a - \gamma_s + \kappa_{2,2}.$$

We give prior means and variances to $\beta_0, \beta_a, \beta_s, \delta, \gamma_0, \gamma_a, \gamma_s, \delta_{1,1}, ..., \delta_{2,2}$, and $\kappa_{1,1}, ..., \kappa_{2,2}$,

all of which are mutually independent. We also define w in the following expression

$$w_{g,a,s} = w_{g,a}^{(a)} + w_{g,s}^{(s)} + w_{g,a,s}^{(a,s)} \tag{6.33}$$

where, for $g = 2, ..., 10$,

$$\begin{aligned} w_{g,a}^{(a)} &= \phi\ w_{g-1,a}^{(a)} + \epsilon_{g,a}^{(a)} \\ w_{g,s}^{(s)} &= \phi\ w_{g-1,s}^{(s)} + \epsilon_{g,s}^{(s)} \\ w_{g,a,s}^{(a,s)} &= \phi\ w_{g-1,a,s}^{(a,s)} + \epsilon_{g,a,s}^{(a,s)} \end{aligned} \tag{$\star$}$$

The autoregressive structure in ($\star$) describes our prior beliefs about deviations of the regression lines from the straight-line model in (6.32). The random variables $\epsilon_{g,a}^{(a)}, \epsilon_{g,s}^{(s)}$ and $\epsilon_{g,a,s}^{(a,s)}$ are all given zero means. For $g = 1, ..., G$, where $G = 10$ is the number of groups, $\mathrm{Var}(\epsilon_{g,a}^{(a)}) = \tau_a^{-1}, \mathrm{Var}(\epsilon_{g,s}^{(s)}) = \tau_s^{-1}$ and $\mathrm{Var}(\epsilon_{g,a,s}^{(a,s)}) = \tau_{a,s}^{-1}$. Furthermore, $\epsilon_{g,d}^{(d)}$ is independent of $\epsilon_{g',d'}^{(d')}$ unless $g = g'$ and $d = d'$ and d, and d' are each one of a, s and a, s. The autoregressive parameter ϕ is chosen to reflect the degree of smoothness we expect to see in deviations from the straight-line. We choose $|\phi| < 1$ and set $\mathrm{Var}(w_{g,d}^{(d)}) = \tau_d^{-1}/(1 - \phi^2)$, so that the process is stationary, we then choose the variance τ_a^{-1}, τ_s^{-1} and $\tau_{a,s}^{-1}$ to give a suitable marginal variance to $w_{g,a,s}$ and suitable covariances between areas and sexes. This leads to the prior means, variances and covariances for the elements of η.

We regard areas as exchangeable and sexes as exchangeable so the marginal distribution for each subvector $\eta_{a,s}$ is the same for each a, s.

The prior mean for $\eta_{a,s}$ is

$$\mathrm{E}_0(\eta_{a,s}) = (-9.312, -8.382, -7.142, -5.902, -4.662, -3.422, -2.182, -0.880, 0.298, 1.042).$$

The marginal prior variance for $\eta_{g,a,s}$ is 1.006 for $g = 1, ..., 10$ and the prior autocorrelations are as follows

Lag	1	2	3	4	5	6	7	8	9
Correlation	0.912	0.833	0.762	0.697	0.640	0.587	0.540	0.498	0.460

6.9.2 Bayes linear kinematic analysis

We combine full-Bayes marginal updates, using the Laplace approximation and Bayes linear kinematics in order to obtain $S_1(\eta) = [\mathrm{E}_1(\eta|x), \mathrm{Var}_1(\eta|x)]$. Then, we propagate these changes in beliefs to other age groups using the Bayes linear kinematics equations (6.12) and (6.13).

6.9.3 Results

For comparison with the BLK analysis we also present the results of a full-Bayes analysis in which posterior summaries are computed using MCMC.

First we do an analysis for each area-sex group separately. We use Bayes linear kinematics on $\underline{\eta}$. We use `R` to do all these updates and produce some graphs which show how BLK is close to the full-Bayes analysis. Figure 6.9 represents the adjusted means for $\eta_{g,1,s}$ (red) using BLK with the non-conjugate prior and the corresponding posterior means from the full Bayes analysis. We notice that the posterior means are close to each other. In addition, we can see that our posterior expectations using BLKs are closer to the corresponding data for age group g, area 1 and sex s than the prior expectations. That is an indicator that our prior variance and covariance structure that we used allowed the inference to reflect the relationships between the variables very well. Figure 6.10 shows the posterior expectations, using BLK with $\pm$ 2 standard deviation limits which we can calculate, using

$$\left(\mathrm{E}_1(\eta; x) - 2 \times \sqrt{\mathrm{Var}_1(\eta; x)}, \mathrm{E}_1(\eta; x) + 2 \times \sqrt{\mathrm{Var}_1(\eta; x)}\right).$$

Tables 6.9, 6.11, 6.13 and 6.15, show the posterior means for η from BLK, full-Bayes analysis and the values of $\hat{\eta}$ for the four combinations of area and sex. We can see that, almost all of these posterior means are very close to each other in both methods full-Bayes and BLK. Therefore, the non-conjugate method produces similar results to MCMC but much more quickly. This is also shown in Figure 6.11.

The posterior variances using full-Bayes analysis and the marginal posterior variances using BLK are shown in Tables 6.8, 6.10, 6.12 and 6.14 for the four combinations of sex and area. We can see that most of these marginal posterior variances from BLK and full-Bayes are very close to each other. For example, the posterior variances using BLK

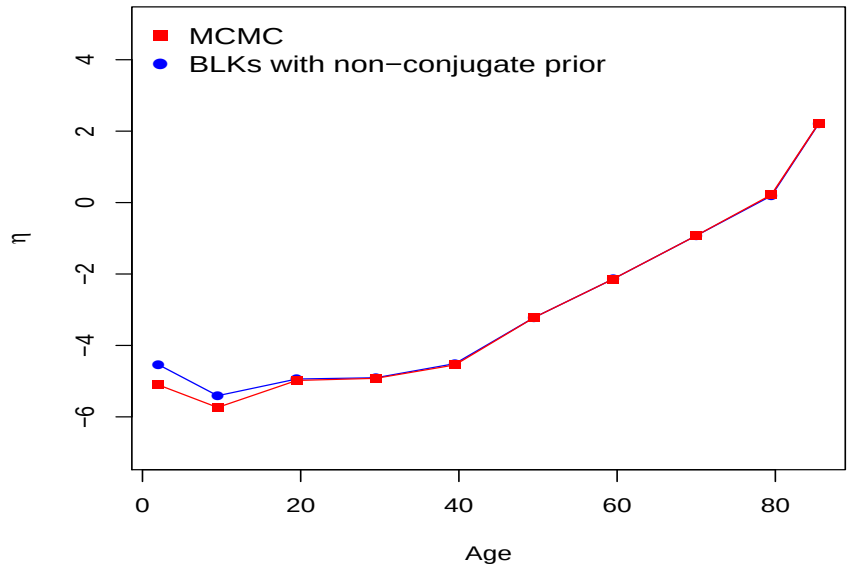


Figure 6.9: Adjusted means for $\eta_{g,1,s}$ using Bayes linear kinematics with the non-conjugate prior and the posterior means using full Bayes analysis (MCMC).

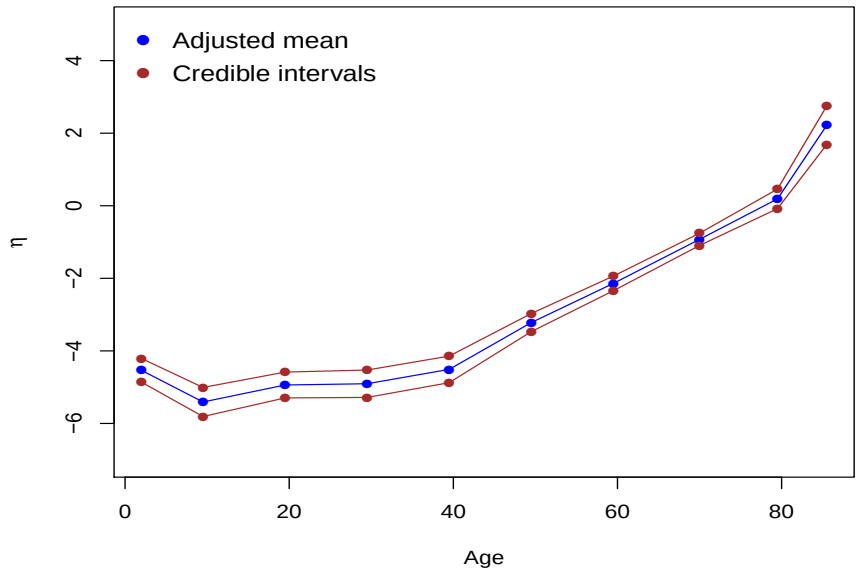


Figure 6.10: Adjusted means, ±2 standard deviation limits for males in area 1.

g, s	$\mathrm{Var}_1(\eta)$ (BLK)	$\mathrm{Var}_1(\eta)$ (full-Bayes)
(1,1)	0.0266	0.0372
(2,1)	0.0401	0.0456
(3,1)	0.0317	0.0375
(4,1)	0.0356	0.0371
(5,1)	0.0335	0.0372
(6,1)	0.0156	0.0156
(7,1)	0.0102	0.0102
(8,1)	0.0071	0.0071
(9,1)	0.0182	0.0192
(10,1)	0.0705	0.0978

Table 6.8: Posterior variances of $\underline{\eta}$ from BLK and full-Bayes analysis for the males in Area 1.

g, s	$\mathrm{E}_1(\eta)$ (BLK)	$\mathrm{E}_1(\eta)$ (full-Bayes)	$\hat{\eta}$ (observations)
(1,1)	-4.5358	-5.1039	-4.1089
(2,1)	-5.4083	-5.7371	-6.1605
(3,1)	-4.9401	-4.9809	-4.8062
(4,1)	-4.9048	-4.9211	-4.9807
(5,1)	-4.5093	-4.5458	-4.7844
(6,1)	-3.2192	-3.2200	-3.1935
(7,1)	-2.1401	-2.1401	-2.1406
(8,1)	-0.9262	-0.9252	-0.9260
(9,1)	0.1902	0.2195	0.0861
(10,1)	2.2184	2.2268	4.2195

Table 6.9: Posterior means of $\underline{\eta}$ from BLK, full-Bayes analysis and the values of $\hat{\eta}$ for the males in Area 1.

for males and females in Area 1 are slightly less than the posterior variances using full-Bayes analysis. However, some of the posterior variances using BLK for both males and females in Area 2 are slightly bigger than the marginal posterior variances using full-Bayes analysis.

Now, we expand the analysis so that the data from both areas and both sexes are analysed together. In fact, we have 4 groups in this example (2 areas $\times$ 2 sexes). So, to do that we give a vector of 10 prior means (for the 10 age-groups for males in Area 1). Then the mean vector for the whole data set is just this vector repeated 4 times.

We also give a 10×10 variance matrix for the males in area 1, V_0. Now we should have a 40×40 variance matrix

g, s	$\mathrm{Var}_1(\eta)$ (BLK)	$\mathrm{Var}_1(\eta)$ (full-Bayes)
(1,2)	0.0433	0.0605
(2,2)	0.0490	0.0563
(3,2)	0.0637	0.0733
(4,2)	0.0666	0.0638
(5,2)	0.0434	0.0444
(6,2)	0.0210	0.0227
(7,2)	0.0169	0.0173
(8,2)	0.0090	0.0092
(9,2)	0.0131	0.0133
(10,2)	0.0389	0.0400

Table 6.10: Posterior variances of $\underline{\eta}$ from BLK and full-Bayes analysis for the females in Area 1.

g, s	$\mathrm{E}_1(\eta)$ (BLK)	$\mathrm{E}_1(\eta)$ (full-Bayes)	$\hat{\eta}$ (observations)
(1,2)	-5.0447	-5.0211	-4.4742
(2,2)	-6.0085	-5.8132	-5.8708
(3,2)	-5.7111	-6.0258	-6.3916
(4,2)	-5.5102	-5.6181	-5.7831
(5,2)	-4.7927	-4.7759	-4.8262
(6,2)	-3.6784	-3.6621	-3.6136
(7,2)	-2.9699	-2.9022	-2.9167
(8,2)	-1.9175	-1.8190	-1.8319
(9,2)	-0.4554	-0.4544	-0.5053
(10,2)	1.3058	1.3335	1.5437

Table 6.11: Posterior means of $\underline{\eta}$ from BLK, full-Bayes analysis and the values of $\hat{\eta}$ for the females in Area 1.

g, s	$\mathrm{Var}_1(\eta)$ (BLK)	$\mathrm{Var}_1(\eta)$ (full-Bayes)
(1,1)	0.1167	0.1251
(2,1)	0.0685	0.0751
(3,1)	0.0596	0.0593
(4,1)	0.0698	0.0658
(5,1)	0.0599	0.0561
(6,1)	0.0282	0.0286
(7,1)	0.0142	0.0146
(8,1)	0.0079	0.0080
(9,1)	0.0180	0.0187
(10,1)	0.0540	0.0560

Table 6.12: Posterior variances of $\underline{\eta}$ from BLK and full-Bayes analysis for the males in Area 2.

g, s	$\mathrm{E}_1(\eta)$ (BLK)	$\mathrm{E}_1(\eta)$ (full-Bayes)	$\hat{\eta}$ (observations)
(1,1)	-5.1069	-5.0346	-4.0532
(2,1)	-5.5059	-5.4175	-5.8487
(3,1)	-4.4720	-4.7241	-4.4667
(4,1)	-4.5038	-4.7009	-4.8091
(5,1)	-4.3459	-4.3733	-4.5917
(6,1)	-3.4048	-3.4342	-3.4624
(7,1)	-2.2176	-2.2194	-2.2070
(8,1)	-1.0685	-1.0503	-1.0448
(9,1)	-0.1349	-0.1276	-0.1804
(10,1)	0.9413	0.9757	1.0609

Table 6.13: Posterior means of $\underline{\eta}$ from BLK, full-Bayes analysis and the values of $\hat{\eta}$ for the males in Area 2.

g, s	$\mathrm{Var}_1(\eta)$ (BLK)	$\mathrm{Var}_1(\eta)$ (full-Bayes)
(1,2)	0.3022	0.3684
(2,2)	0.2260	0.2324
(3,2)	0.1768	0.1688
(4,2)	0.1126	0.1158
(5,2)	0.0516	0.0628
(6,2)	0.0409	0.0426
(7,2)	0.0220	0.0247
(8,2)	0.0129	0.0130
(9,2)	0.0143	0.0141
(10,2)	0.0621	0.0604

Table 6.14: Posterior variances of $\underline{\eta}$ from BLK and full-Bayes analysis for the females in Area 2.

g, s	$\mathrm{E}_1(\eta)$ (BLK)	$\mathrm{E}_1(\eta)$ (full-Bayes)	$\mathrm{E}_1(\eta)$ (observations)
(1,2)	-7.8966	-7.6528	-5.4323
(2,2)	-7.2931	-7.1886	-6.7776
(3,2)	-6.4877	-6.5292	-7.1253
(4,2)	-5.5623	-5.6222	-5.8101
(5,2)	-4.4997	-4.5046	-4.2474
(6,2)	-3.9025	-4.0223	-4.1442
(7,2)	-2.9437	-2.9917	-2.9430
(8,2)	-2.0936	-2.2280	-2.2762
(9,2)	-0.6615	-0.7188	-0.8309
(10,2)	1.9224	2.0961	3.1697

Table 6.15: Posterior means of $\underline{\eta}$ from BLK, full-Bayes analysis and the values of $\hat{\eta}$ for the females in Area 2.

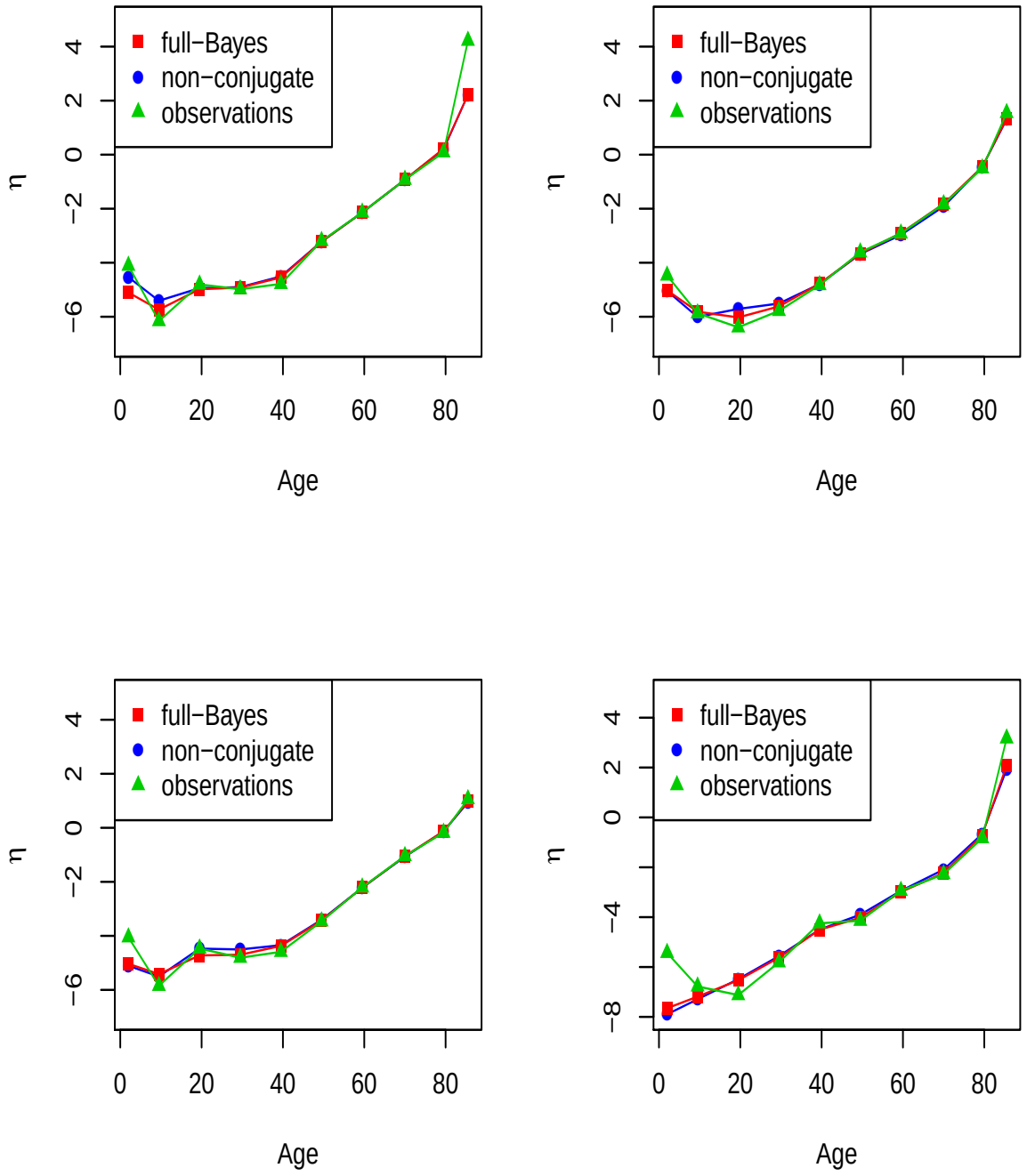


Figure 6.11: Posterior means for η using full-Bayes analysis, BLK with non-conjugate prior and the empirical data. Top left: Posterior means for η for males in Area 1. Top right: Posterior means for η for females in Area 1. Bottom left: Posterior means for η for males in Area 2. Bottom right: Posterior means for η for females in Area 2.

$$\mathrm{V}_0(\eta) = \begin{bmatrix} V_0 & \rho_A V_0 & \rho_S V_0 & \rho_0 V_0 \\ \rho_A V_0 & V_0 & \rho_0 V_0 & \rho_S V_0 \\ \rho_S V_0 & \rho_0 V_0 & V_0 & \rho_A V_0 \\ \rho_0 V_0 & \rho_S V_0 & \rho_A V_0 & V_0 \end{bmatrix}$$

where the first row and the first column of $V_0(\eta)$ represent the males in Area 1, the second row and the second column of $V_0(\eta)$ represent the females in Area 1, the third row and the third column of $V_0(\eta)$ represent the males in Area 2 and the fourth row and the fourth column of $V_0(\eta)$ represent the females in Area 2 and ρ_0, ρ_A and ρ_S are correlation coefficients, where

ρ_0^2 is the proportion of uncertainty shared by all 4 groups,

ρ_S^2 is the proportion of uncertainty shared by 2 groups which are the same sex, and

ρ_A^2 is the proportion of uncertainty shared by 2 groups which are the same area.

So,

$$0 < \rho_0^2 + (\rho_S^2 - \rho_0^2) + (\rho_A^2 - \rho_0^2) < 1$$

In this example, we give $\rho_0^2 = 0.64$, $\rho_A^2 = 0.81$ and $\rho_S^2 = 0.68$. We also have assigned values to $\mathrm{Var}(\beta_0), \mathrm{Var}(\beta_A)$ and $\mathrm{Var}(\beta_S)$, where $\mathrm{Var}(\beta_0) = \rho_0^2 \mathrm{Var}(\beta_{11})$, $\mathrm{Var}(\beta_A) = (\rho_A^2 - \rho_0^2)\mathrm{Var}(\beta_{11})$, $\mathrm{Var}(\beta_S) = (\rho_S^2 - \rho_0^2)\mathrm{Var}(\beta_{11})$, and $\mathrm{Var}(\delta_{11}) = (\rho_A^2 + \rho_S^2 - \rho_0^2)\mathrm{Var}(\beta_{11})$ and so on. So, the posterior mean using BLK in (6.27) for η is $\mathrm{E}_1(\eta)$ as follows

Area	Sex	$g = 1$	2	3	4	5	6	7	8	9	10
1	1	-4.603	-5.445	-5.009	-4.919	-4.487	-3.222	-2.137	-0.921	0.282	2.691
1	2	-4.754	-5.493	-5.018	-4.921	-4.488	-3.222	-2.137	-0.921	0.282	2.691
2	1	-4.754	-5.493	-5.018	-4.921	-4.488	-3.222	-2.137	-0.920	0.282	2.691
2	2	-4.754	-5.493	-5.018	-4.921	-4.488	-3.222	-2.138	-0.924	0.190	1.727

The posterior correlation matrix for η using Bayes linear kinematics is shown in Appendix (A.6.4).

6.10 Categorical and censored variables

6.10.1 Introduction

In a Bayes linear analysis we might observe variables which, typically, we might suppose have approximately normal distributions, perhaps after transformation. In some problems, such as the survival applications in Chapter 7, we observe variables which could not be given a normal distribution even after transformation, for example discrete or categorical variables. In a Bayes linear Bayes model we relate such an observable variable to an underlying latent variable Z. Thus we typically have a latent vector $\underline{Z} = (Z_1, ..., Z_T)^T$ where in an analogous full-Bayes analysis, $\underline{Z}$ might be given a multivariate normal distribution. In this section we consider some particular types of observable variables.

6.10.2 Binary variables

A categorical variable which can take only two values is a binary variable. Typically the values may be labelled 0 or 1.

Two possibilities arise in the case of a binary variable X. The first method, called the *direct method*, is to let $X = 1$ if the corresponding $Z \geq 0$ and $X = 0$ if $Z < 0$ and, for this reason, we assign Z a normal prior distribution. In this case the support of the posterior distribution is bounded at zero and we need to use a quadrature method rather than the normal or Laplace approximation.

A second possibility, which we call the *indirect method*, is often appropriate in applications such as a prognostic index. Suppose that $\Pr(X = 1|Z) = \theta = h^{-1}(Z)$ where $h^{-1}()$ is the inverse of a suitable link function $h()$, for instance, logit, $Z = \log\{\theta/(1-\theta)\}$, or probit, $Z = \Phi^{-1}(\theta)$, where Φ is the standard normal cumulative distribution function. In this case the likelihood function is $\theta^X(1-\theta)^{1-X}$ and the posterior support is not bounded so we can use methods such as a Laplace approximation.

6.10.3 Ordinal variables

Ordinal variables are categorical variables where the values are ordered. A suitable model to use with an ordinal variable is an *ordinal logistic regression*.

Let X be an ordinal variable with K categories. One of the simplest ways to model that is to obtain

$$\Pr(X = 1) = \theta_1,$$

$$\Pr(X = 2|X \neq 1) = \theta_2,$$

$$\Pr(X = 3|X \neq 1, X \neq 2) = \theta_3,$$

and so on. These conditional probabilities are not constrained to sum to one.

We consider the case of ordinal variables as a generalisation of the case of binary variables. Suppose that we have ordered categories labelled $\{1, ..., K\}$. Then we need $K - 1$ cut points $\{c_1, c_2, ..., c_{K-1}\}$ for Z. For example, in the non-Hodgkin lymphoma data set, we have an ordinal variable called Stage. This variable can take the values $(0, 1, 2, 3)$. We relabel these $(1, 2, 3, 4)$. Then we should have in this case three cut points. To avoid non-identifiability, we need to fix two cut points. This is because the distribution of Z has two parameters, the mean and the variance.

Again, as in the case of binary variables, using the direct method, $X = k$ if and only if $c_{k-1} \leq Z < c_k$ for a set of thresholds $\{c_1, ..., c_{K-1}\}$ where $c_0 \to -\infty$ and $c_K \to \infty$. In this case the posterior distribution support is bounded, often both below and above, so we might use a quadrature method to find the posterior moments.

We can also apply the indirect method. Suppose that $\Pr(X \leq k) = h^{-1}(c_k - Z)$ for a suitable link function $h()$. It is convenient to use $h() = \Phi^{-1}()$. In this case we might suppose that there is a latent variable Z^* which has a normal distribution $N(Z, 1)$, given Z. Therefore, the likelihood is $h^{-1}(c_k - Z) - h^{-1}(c_{k-1} - Z)$, for example $\Phi(c_k - Z) - \Phi(c_{k-1} - Z)$. In this case the posterior support for Z is unbounded.

6.10.4 Unordered categorical variables

Dealing with a categorical variable with $K > 2$ unordered categories labelled $1, ..., K$, other than by conditioning the whole model on the categories, requires the use of an underlying vector variable with $K - 1$ elements. This can be handled within the general Bayes linear kinematic and Bayes linear Bayes framework since the elements Z can be vectors. For example in an indirect approach we can set

$$\Pr(X_j = k) = \frac{\exp(Z_{j,k})}{\sum_{i=1}^{K} \exp(Z_{j,i})}$$

and this provides the likelihood. We can then give each $Z_{j,k}$ a normal distribution with variance 1. A constraint, such as $Z_{j,1} = 0$ or $\sum_{k=1}^{K} Z_{j,k} = 0$ is applied for identifiability.

In a direct approach, observing $X_j = k$ corresponds to observing $Z_{j,k} \geq 0$ and $Z_{j,i} < 0$ for all $i \neq k$.

6.10.5 Interval-censored variables

A variable subject to interval censoring may be handled by methods similar to the direct and indirect methods in the case of ordinal variables. In the direct case, if X is not censored, then $Z = X$ and we make a direct observation. If the observation is censored, then we observe that $c_{k-1} < Z < c_k$ for lower and upper bounds c_{k-1} and c_k and the posterior support is bounded. In the indirect method we suppose that, when X is not censored, we observe $Z^* = X$ and, when X is censored, we observe that $c_{k-1} < Z^* < c_k$, where, $Z^* \sim N(Z, 1)$. Hence, in the latter case, the likelihood is $\Phi\{(c_k - Z)\} - \Phi\{(c_k - Z)\}$.

6.10.6 Marginal update calculations for ordinal observations

Consider an ordinal variable X, using the direct method. Suppose that Z has prior mean E_0 and prior variance V_0. Let $\mathrm{S}_0 = \sqrt{\mathrm{V}_0}$ be the prior standard deviation. Suppose that we use probits. Let $\Phi()$ be the standard normal distribution function and $\Phi^{-1}()$ be its inverse. Let $W = \Phi\{(Z - \mathrm{E}_0)/\mathrm{S}_0\}$. Then W has a prior uniform distribution on (0,1). We can also transform the cut points in the same way: $C_x = \Phi\{(C - \mathrm{E}_0)/\mathrm{S}_0\}$. The observation X selects the observed interval. The posterior distribution of W is then simply a uniform distribution over that interval, $U(C_l, C_u)$, where C_l represents the lower cut point and C_u refers to the upper cut point. We use a trapezium rule to integrate over that interval by setting up a grid of W values over the interval and then calculate the values of $Z = \mathrm{E}_0 + \mathrm{S}_0\Phi^{-1}(W)$. Finally we find the average value of Z and the average value of Z^2 over the interval. We used a `R` function to make the adjustment for both binary and ordinal variables in the non-Hodgkin lymphoma example, since a binary variable is equivalent to an ordinal variable with just two categories. See Appendix A.6.5.

When we use the indirect method, we can obtain unbounded support for the posterior rather than obtain bounded support in the direct method.

In the case of binary variables in the model, such as albumin in the non-Hodgkin lymphoma data, our likelihood in this case will be $\theta^X(1-\theta)^{1-X}$. We transform the

parameter θ here, so that we can use, for example, logit, $Z = \log[\theta/(1-\theta)]$. Then the likelihood can be written as

$$f(X, Z) = \left(\frac{e^z}{1+e^z}\right)^x \left(\frac{1}{1+e^z}\right)^{1-x},$$

and the prior for Z in this case is $Z \sim N(m, 1)$ where m is the prior mean of Z and we fix the variance of Z to be 1.

Therefore, the posterior density of Z is

$$\begin{aligned}\pi(Z|x) &\propto f(Z)f(X, Z) \\ &\propto \frac{1}{2}\exp(Z-m)^2 \exp(Zx)[1+\exp(Z)]^{-1}.\end{aligned}$$

Then the support of the posterior density $\pi(Z|x)$ in this case is unbounded so we can use a Laplace approximation to calculate the posterior means and variances for Z when we update these moments using BLK.

6.11 Summary

In this chapter, we have investigated Bayes linear methods with some theoretical aspects of this approach. Bayes linear analysis is different from full-Bayes analysis as Bayes linear methods specify just the first and second order moments and then calculate the posterior moments. So we do not need to specify the prior in a probabilistic way as in a full-Bayes analysis. We explained the idea of Bayes linear analysis using a motivational example. We explained Bayes linear kinematics and mentioned the concept "commutativity" and how to do multiple updates using BLK.

We also described Bayes linear Bayes graphical models as a combination of Bayesian networks and Bayes linear structure. We use the idea of transformation of the parameters for different reasons. However, the most important one is that θ may have a bounded range and this boundary makes the use of Bayes linear methods less attractive.

After transforming the parameters, we can use the mode and log-curvature method that we explained in Section 6.5.3. We apply the mode and curvature method to construct the prior mean and variance for the transformed parameters. We use an example concerning the use of sulfinpyrazone.

This chapter has some proposed ideas including the use of non-conjugate updates in order to calculate Bayes linear kinematics. An example is when we have a binomial likelihood and logit-normal prior. In the case of a non-conjugate prior, we need to use some numerical integration methods such as a Laplace approximation or the trapezoidal rule. However, these integrations are low-dimensional, often one-dimensional, in contrast to the high-dimensional integrations required by a full-Bayes analysis.

We have done two examples in this chapter and the results show that using a non-conjugate prior distribution gives posterior moments closer to those obtained with a full Bayes analysis. The ability to use non-conjugate marginal updates also widens the range of types of observations which can be incorporated in a Bayes linear Bayes model.

We finish this chapter by considering how different types of variables can be handled in a Bayes linear Bayes model. In addition, we have proposed two methods (the direct and the indirect method) which are suitable in some applications such as a prognostic index.

Chapter 7

Application to survival data

7.1 Introduction

This chapter deals with the application of Bayes linear Bayes models and Bayes linear kinematics to survival data. We divide this chapter into two parts. The first part deals with the leukaemia example with investigation and comparison of the full-Bayes method, the BLK method with conjugate prior and our proposed method which is BLK with non-conjugate prior update. This example applies BLK to a data set to make inferences about the values of model parameters. In this part, we start with the introduction of piecewise constant hazard (PCH) models in Section 7.2.2. Then we give a brief description of the leukaemia data set in Section 7.3. In Section 7.3.3, we illustrate in detail the Wilson and Farrow approach which depends on using BLK in a PCH model. Section 7.3.4 uses the idea of non-conjugate prior updates in order to produce posterior means and the posterior variance-covariance matrix for the parameters of interest $\underline{\beta}$ using BLK. We compare the results from using the non-conjugate method with full Bayes methods and BLK with a conjugate prior in Section 7.3.6. In Section 7.3.7, we do some diagnostic plots for the leukaemia example to check the validity of the assumptions.

The second part tackles another type of application with data on patients with non-Hodgkin lymphoma. The idea here is to construct a Bayes linear kinematic network to compute a prognostic index value for a patient given observations on some or all of a set of covariates. The calculation will be fast and relatively simple. The use of Bayes linear kinematics eliminates the problem of non-commutativity which has been experienced in some related work. We start the second part by outlining the novelty of using a latent

prognostic index and how it relates to survival time for patients. In Section 7.4.3, we explain how we can construct a Bayes linear kinematic prognostic network with various sorts of covariates. In Section 7.5.1, we give an explanation of the general strategy for constructing the prognostic network, including the use of the offline learning model.

We propose the BLK method with non-conjugate prior update and investigate its application to a rapid computation of prognostic index values in survival analysis when a patient's values may only be available for a subset of covariates. We explain the offline learning in Section 7.6.3 which simply uses a full-Bayes analysis and MCMC to learn about the values of the parameters. As in the leukaemia example, in Section 7.6.6, we give diagnostic plots for the non-Hodgkin lymphoma in order to assess the validity of our assumptions and to support model selection. In Section 7.6.8, we show the summary of the results for our proposed method using the non-conjugate prior updates in order to evaluate the Bayes linear kinematic approach and compare it with the full-Bayes analysis using MCMC methods. We identify important differences between the direct and the indirect methods and compare the results of using these methods in Section 7.7. In Section 7.8, we explain the prototype calculator for the prognostic index for patients and how it works. Finally, we give a summary of the chapter in Section 7.9.

7.2 Bayes linear Bayes retrospective analysis

7.2.1 Introduction

As our first illustrative example, we will show a retrospective analysis. Here we have a data set with data on a collection of patients and we wish to use Bayes linear kinematics to learn about the values of model parameters. Specifically, we will use a piecewise constant hazard model. Such models will be described in more detail in Section 7.2.2. Briefly, time is divided into a number of intervals and the log hazard for patient i in interval k has the form

$$\eta_{i,k} = \beta_{k,0} + \sum_{j=1}^{J} \beta_{j,k} x_{i,j} \tag{7.1}$$

where $x_{i,j}$ is the value of covariate j for patient i. We wish to learn about the values of the coefficients $\beta_{k,0}, ..., \beta_{k,J}$ for each interval k and we use Bayes linear kinematics to do this.

7.2.2 Piecewise constant hazard models

In a piecewise constant hazard (PCH) model, we do not assume a particular form for the baseline hazard function $h_0(t)$. Instead we relax the parametric assumption about the baseline hazard. Time is divided into different time intervals. Then we assume that the hazard is constant within each interval. However, the hazards are allowed to vary from interval to interval. See Section 5.5.2.

Suppose that we have patients $i = 1, ..., n$ and patient i has the covariates $x_i = (1, x_{i,1}, ..., x_{i,J})'$. A hazard function $h_i(t)$ is associated with patient i at time t. The hazard functions of patients when we assume a proportional hazards model (Cox, 1972) are related as $h_i(t) = \phi_i h_0(t)$, where ϕ_i is a constant and $h_0(t)$ is the baseline hazard function. We might wish to relate the hazard function for patient i to the patient's covariates as follows

$$\phi_i = \exp(\underline{x}_i' \underline{\beta}) \tag{7.2}$$

for some parameters $\underline{\beta} = (\beta_0, \beta_1, ..., \beta_J)'$.

In this scenario, we assume that the parameter values $\underline{\beta}$ are constant over time, which means that the effect of the covariates for the patient is constant over time. Sometimes this is not the case, so if we wish to allow for this possibility we need to use a dynamic model which allows changes in the effects of the covariates over time. This dynamic model can be written as

$$\phi_i(t) = \exp(\underline{x}_i' \underline{\beta}(t)). \tag{7.3}$$

Now the piecewise constant hazards model, (see, for example, Ibrahim et al., 2001; Wilson and Farrow, 2017) uses some fixed cut-points $s_0, s_1, ..., s_k$ such that $s_0 = 0$ and $s_k \to \infty$ is greater than the largest death time. Therefore, the time can be divided into intervals. We define the interval I_k as $[s_{k-1}, s_k)$. So the baseline hazard will be

$$h_0(t) = \lambda_{0,k}, \qquad \text{for} \quad s_{k-1} \leq t < s_k$$

and the hazard function for patient i is

$$h_i(t) = \lambda_{i,k}.$$

More detail about the integrated hazard function $H_i(t)$, the survival function $S_i(t)$ and the probability density function for patient i is given in Section 5.5.2.

In order to apply Bayesian analysis for the dynamic model, first we need to introduce the likelihood function allowing for right censored observations. The likelihood for patients $i = 1, ..., n$ is

$$L = \prod_{i=1}^{n} \prod_{k=1}^{K} L_{i,k}$$

where

$$L_{i,k} = (\lambda_{i,k})^{\delta_{i,k}} \exp\{-\lambda_{i,k}(t_{i,k} - s_{k-1})\} \tag{7.4}$$

where $\delta_{i,k} = 1$ if the patient i dies in the interval I_k, $\delta_{i,k} = 0$ if the patient i is censored or survives in I_k and $t_{i,k} = t_i$ if patient i dies in interval I_k, $t_{i,k} = t_k$ if patient i survives in interval I_k or $t_{i,k} = t_i^*$ if patient i is censored at time t_i^* in interval I_k.

7.2.3 Full Bayes analysis for piecewise constant hazard model

In this section, we give a brief description of the full-Bayes analysis to compute the posterior means and posterior variances for all the parameter values $\underline{\beta}$ in a piecewise constant hazards model.

Given that the patient survives to the beginning of the interval I_k, the conditional lifetime distribution is an exponential distribution with parameter $\lambda_{i,k}$ for patient i in interval I_k. In each interval, our beliefs about $\lambda_{i,k}$ are updated when we observe the data in that interval. These changes in belief are propagated to the quantity $\eta_{i,k}$ (the linear predictor) as follows

$$\log(\lambda_{i,k}) = \eta_{i,k} = \underline{x}_i^T \underline{\beta}_k.$$

To compute the posterior distribution of $\eta_{i,k}$, we need to combine the likelihood function in (7.4) with a suitable prior distribution, for example, a normal prior distribution for $\underline{\beta} = (\underline{\beta}_1^T, ..., \underline{\beta}_k^T)^T$. Since the prior is not conjugate to the likelihood, we need numerical methods to compute posterior summaries and Markov chain Monte Carlo (MCMC) methods are usually used.

7.3 Example: Leukaemia

7.3.1 Introduction

This example refers to the data of Henderson et al. (2002) which are described in Section 2.3. In this example, we have 1043 patients. The dependent variable in this study is the time T measured in days until the event (death) occurs. Of the 1043 patients, 879 died and 164 were right censored. See Section 2.3 for more details.

We have a number of covariates that were thought to have an effect on survival with this disease. We manipulate these covariates as follows:

- The age A_i represents the age of patients in years. So we use $x_{i,1} = A_i - 60$.
- The sex of the patients. We give $x_{i,2} = 1$ if the patient is male and $x_{i,2} = -1$ if the patient is female.
- White blood cell count W_i at the time of diagnosis. We use $x_{i,3} = W_i - 8$.
- Deprivation score (Depscore): This variable measures the deprivation for the residential area of the patient. We use the Townsend deprivation index (TDI) (Townsend et al., 1988). The scale of the variable is from -7 to 10 with lower values indicating more severe deprivation.

We wish to use a piecewise constant hazard model for leukemia survival. Therefore, the hazard function for patient i in interval I_k is $\exp\{\beta_{k,0} + \sum_{j=1}^{4} \beta_{k,j} x_{i,j}\}$. For instance, if we have a male patient with age 60, his white blood cell count is 8 and deprivation score 0, then the hazard function in this case will be $\exp\{\beta_{k,0} + \beta_{k,2}\}$ and so forth.

7.3.2 Exploratory plots in the leukaemia example

Figure 7.1 shows scatter plots of pairs of variables. When we compare Age and $\log(T)$ in Figure 7.1, we see that most of the people are in the age group [50-80] and the younger people are tending to live longer than the older ones.

When we plot one covariate against another, that is plotting Age against log(WBC), Age against Deprivation, log(WBC) against Deprivation, we can see that we have a random scatter for points for both males and females.

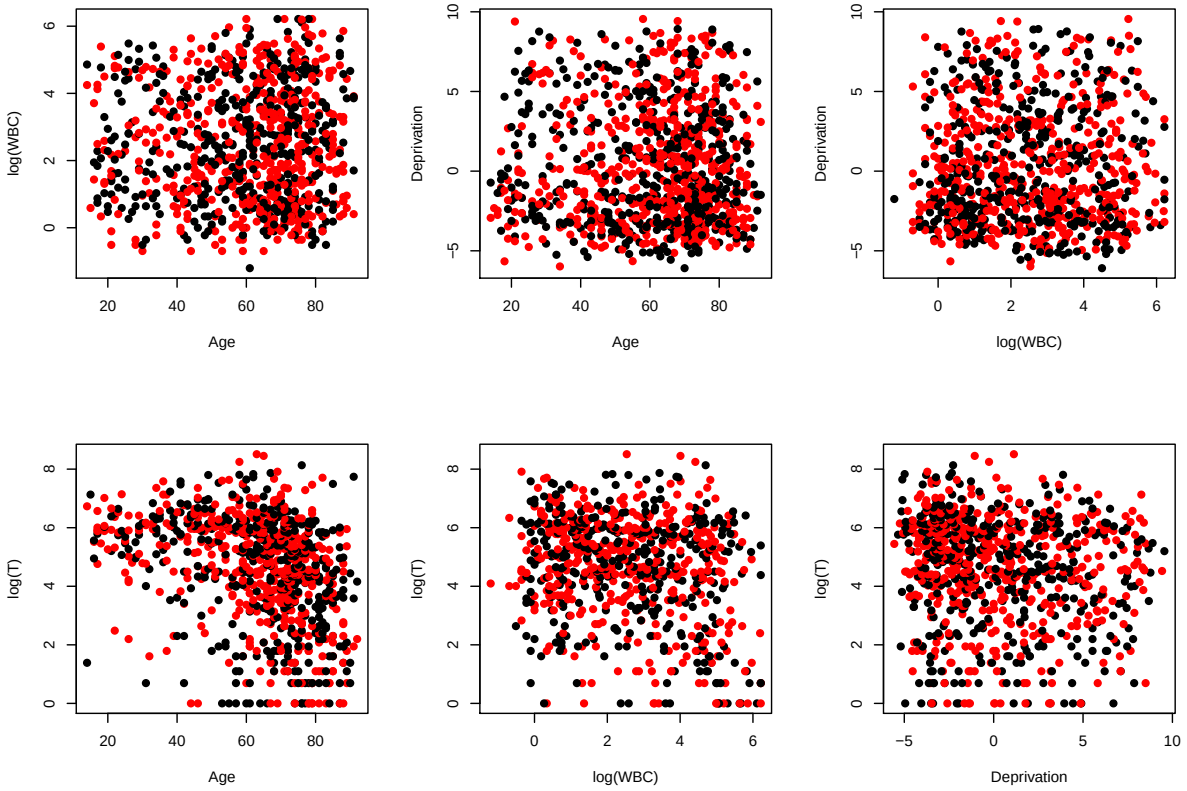


Figure 7.1: Exploratory plots for the covariates in leukaemia example. Black dots for males and red dots for females.

When we plot log(WBC) against $\log(T)$ and Deprivation against $\log(T)$, there is no suggestion that the variance changes. However, we will need to examine diagnostic plots after fitting the model.

7.3.3 Wilson and Farrow approach

Wilson and Farrow (2017) applied Bayes linear kinematics to these data using a piecewise constant hazards model with 10 time intervals. So, they define $s_k = -\nu \log(1 - uk)$, for $u = 0.1$ and $\nu = 500$ in order to give us these 9 cut-points which are (52.7, 111.6, 178.3, 255.4, 346.6, 458.1, 602.0, 804.7, 1151.3).

In order to make inference (i.e. posterior means, posterior variances and covariances) about the collection of parameter values $\underline{\beta}_j = (\beta_{0,j}, \beta_{1,j}, \beta_{2,j}, \beta_{3,j}, \beta_{4,j})^T$ in the model, we

need to specify our prior beliefs for these parameters.

Wilson and Farrow (2017) specified the prior moments for each of the parameters.

This prior specification is developed as follows. To obtain suitable prior moments for $\beta_{0,1}$, we assume a constant hazard with a wide range for the mean lifetime for "baseline" patients. Therefore, we use the mean of -6 and standard deviation of 0.8 for $\beta_{0,1}$. That can give us ± 2 standard deviations range which corresponds to a range for the mean lifetime from 81 to 1998. See Wilson and Farrow (2017).

Now consider $\beta_{1,1}$ which represents the coefficient of Age. The range of the Ages of the patients is between 14 and 92. We expect that the hazard is increasing as the age is increased. Wilson and Farrow (2017) elicited the mean and the variance for $\beta_{1,1}$ by assuming that the patient i is 10 years older than patient $i^{'}$. So, the ratio of the hazard functions for these two patients is

$$\frac{h_i(t)}{h_{i'}(t)} = \exp(10\beta_{1,1}),$$

if patient i and $i^{'}$ have the same values for the other covariates. We can suggest a range $0.8 < h_i(t)/h_{i'}(t) < 1.8$. Suppose that these values give us an approximate 95% interval for a normal prior distribution for $\beta_{1,1}$. Therefore, we obtain $\mathrm{E}_0(\beta_{1,1}) = 0.02$ and $\mathrm{Var}_0(\beta_{1,1}) = 0.0004$.

We can apply this process for the rest of the coefficients in the first time interval, $\beta_{2,1}, \beta_{3,1}$ and $\beta_{4,1}$. We apply this prior elicitation process to the other time intervals. Table 7.1 gives the prior means and variances for $\underline{\beta}$ for the first time interval. The values for the other time intervals are the same.

To complete the prior specification we need to specify prior correlations between the parameters. Following Wilson and Farrow (2017) we make $\beta_{r,j}$ independent of $\beta_{r',j'}$, unless $r = r^{'}$. That is, the coefficients of different covariates are independent of each other. Again following Wilson and Farrow (2017), we construct the covariances between $\beta_{r,1}, ..., \beta_{r,k}$ using a stationary first order autoregressive process. We write, for $j = 2, ..., k$,

$$\beta_{r,j} = \mathrm{B}_r + \phi(\beta_{r,j-1} - \mathrm{B}_r) + \epsilon_{r,j}$$

where $\epsilon_{r,j}$ is a zero-mean random variable with $\epsilon_{r,j}$ independent of $\epsilon_{r',j'}$ unless $r = r^{'}$ and $j = j^{'}$ and $\mathrm{Var}(\epsilon_{r,j}) = V_{\epsilon,r}$.

For stationarity, we choose $|\phi| < 1$ and write

$$\beta_{r,1} = \mathrm{B}_r + \epsilon_r^\star$$

where $\epsilon_r^\star$ is a zero-mean random variable with $\epsilon_r^\star$ independent of $\epsilon_{r'}^\star$ unless $r = r^{'}$ and $\epsilon_r^\star$ independent of $\epsilon_{r^{'},j}$ for all $r^{'}$ and all $j > 1$, and set $\mathrm{Var}(\epsilon_r^\star) = V_{\epsilon,r}(1-\phi^2)^{-1}$. The value of ϕ is chosen to determine the temporal prior correlation in $\beta_{r,1}, ..., \beta_{r,k}$ in conjunction with the choice of $\mathrm{Var}(\mathrm{B}_r)$ (see Wilson and Farrow, 2017). Having chosen the values of ϕ, $\mathrm{Var}(\mathrm{B}_r)$ and $\mathrm{Var}(\beta_{r,j})$, the value of $V_{\epsilon,r}$ is determined since

$$\mathrm{Var}(\beta_{r,j}) = \mathrm{Var}(\mathrm{B}_r) + V_{\epsilon,r}(\phi^2)^{-1}.$$

The values chosen were those used by Wilson and Farrow (2017). Thus $\phi = 0.92, \mathrm{Var}(\mathrm{B}_r) = 0$ for $r = 0, ..., 4$ and the other choices are given in Table 7.1.

Effect		Mean	Variance
Baseline	$\beta_{0,1}$	-6.000	0.64
Age	$\beta_{1,1}$	0.020	0.0004
Sex	$\beta_{2,1}$	0.000	0.1225
WBC	$\beta_{3,1}$	0.005	0.000025
Deprivation score	$\beta_{4,1}$	0.000	0.01

Table 7.1: Prior means and prior variances for each of the effects. Adapted from Wilson and Farrow (2017).

Wilson and Farrow (2017) used conjugate gamma prior distributions for the hazards $\lambda_{i,j} = \exp(\eta_{i,j})$. They considered three different methods for linking the moments of $\lambda_{i,j}$ to those of $\eta_{i,j}$, the log-mode, log-moment and lognormal methods. See Section 6.5.

In this thesis, we will use the same prior specifications as Wilson and Farrow (2017). Then we will use our proposed method which depends upon using the non-conjugate marginal prior (i.e. the non-conjugate method) to update our prior beliefs about $\underline{\beta}$.

7.3.4 Use of non-conjugate updates in the leukaemia example

Our method of using the non-conjugate update might be regarded as an approximation to a full-Bayes analysis. Wilson and Farrow (2017) compared the behaviour of Bayes linear kinematic belief adjustments with full-Bayes posterior inferences in the case of a piecewise constant hazard survival model and found that the results were generally close.

Let us explain how our method works in this example. Suppose we have

$$\eta_{i,k} = \log(\lambda_{i,k}) = \underline{x}_i^T \underline{\beta}_k.$$

We are interested in finding the moments of $\underline{\beta}$, (i.e. $\mathrm{E}_1(\beta)$ and $\mathrm{Var}_1(\beta)$) using Bayes linear kinematics. To do that, we need to specify prior means and prior variance-covariance for $\underline{\beta}$ which are $\mathrm{E}_0(\beta)$ and $\mathrm{Var}_0(\beta)$ respectively. Then, when we have observed patient i in interval I_k, we can use the Laplace approximation method to gain a new mean and new variance for $\eta_{i,k}$ and propagate that through $\underline{\beta}$ and apply the proper BLK in the following way.

We revise the mean and variance using

$$\begin{aligned} \mathrm{E}_1(\underline{\beta} \mid D_\eta) &= \mathrm{E}_{0,\beta} + V_{0,\beta,\eta} V_{0,\eta,\eta}^{-1} (\mathrm{E}_{1,\eta} - \mathrm{E}_{0,\eta}), && (7.5) \\ \mathrm{Var}_1(\underline{\beta} \mid D_\eta) &= V_{0,\beta,\eta} V_{0,\eta,\eta}^{-1} V_{1,\eta,\eta} V_{0,\eta,\eta}^{-1} V_{0,\eta,\beta} \\ &\quad + V_{0,\beta,\beta} - V_{0,\beta,\eta} V_{0,\eta,\eta}^{-1} V_{0,\eta,\beta}. && (7.6) \end{aligned}$$

where D_η is a single observation, $\mathrm{E}_0(\beta)$ is a vector of prior mean of β, $V_{0,\beta,\eta}$ is a vector of the covariance between β and η, $V_{0,\eta,\eta}^{-1}$ is a scalar prior variance of η, $V_{0,\beta,\beta}$ is the prior variance for β, $\mathrm{E}_1(\underline{\beta} \mid D_\eta)$ is the posterior mean of β updated by the single observation D_η and $\mathrm{Var}_1(\underline{\beta} \mid D_\eta)$ is the posterior variance of β updated by a single observation D_η.

That is, we let $\eta = \log(\lambda)$ then we give η a normal prior distribution with mean $\mathrm{E}_0(\eta)$ and variance $\mathrm{Var}_0(\eta)$. Then we obtain a Bayesian update using the numerical methods such as the Laplace approximation method. See Section 6.5.3. Therefore, the update of η will be non-conjugate. Hence, we obtain $\mathrm{E}_1(\eta)$ and $\mathrm{Var}_1(\eta)$. Then we use these $\mathrm{E}_1(\eta)$ and $\mathrm{Var}_1(\eta)$ in (7.5) and (7.6) to find the posterior means and the posterior variances for $\underline{\beta}$.

To combine the updates from all observations, we use Bayes linear kinematics. When a unique commutative update exists, it can be written as

$$\begin{aligned} \mathrm{P}(\underline{\beta} \mid D) &= \sum_{j=1}^{J} \mathrm{P}(\underline{\beta} \mid D_j) - (J-1)\mathrm{P}(\underline{\beta}), && (7.7) \\ \mathrm{P}(\underline{\beta} \mid D)\mathrm{E}(\underline{\beta} \mid D) &= \sum_{j=1}^{J} \mathrm{P}(\underline{\beta} \mid D_j)\mathrm{E}(\underline{\beta} \mid D_j) - (J-1)\mathrm{P}(\underline{\beta})\mathrm{E}(\underline{\beta}), && (7.8) \end{aligned}$$

where $D = (D_1, \ldots, D_J)'$ and $\mathrm{P}(\underline{\beta}) = \mathrm{Var}(\underline{\beta})^{-1}$ is the prior precision matrix. See Section

6.3.4 and equations (6.14) and (6.15).

Our use of non-conjugate updates allows our model to be closer to the corresponding full-Bayes model. We can see from Figure 7.3 that the Bayes linear kinematic adjusted expectations are even closer to the full-Bayes posterior means.

7.3.5 Full Bayes analysis for the leukaemia example

For comparison with the Bayes linear kinematic analysis we also carry out a conventional full-Bayes analysis. In terms of the prior specification, we assume that all $\underline{\beta}_j = (\beta_{0,j}, \beta_{1,j}, \beta_{2,j}, \beta_{3,j}, \beta_{4,j})'$ have a multivariate normal distribution. There are 50 parameters in this case (5 in each time interval). We represent the prior means and prior variances in the first interval in Table 7.1.

To apply a full Bayes analysis, we need to specify our likelihood function which is written in (7.4) in addition to the prior specification.

A Bayesian posterior update is done using the fully specified prior. We give a (non-conjugate) normal prior for $\eta_{i,j}$. Therefore, the posterior update for $\eta_{i,j}$ is non-conjugate. The computations are done using MCMC. Specifically we used `rjags` (Plummer, 2017). The `rjags` model specification is given in Appendix A.7.1.

7.3.6 Results in the leukaemia example

A preliminary analysis gave results which suggested that the hazard was not constant over the first time interval. This was indicted by an excess of residuals close to the left hand end of the histogram, in a residual plot made in the same way as Figure 7.4. To deal with this problem, an extra cut-point was introduced at $s_1 = 10.5$, giving eleven intervals in total. The value of 11.5 was chosen by examination of histograms and Kaplan-Meier plots (See Figure 7.2) of lifetimes in the affected region. Because the hazard appeared to change rapidly in this region, we did not reduce the incremental variance $\mathrm{Var}_0(\beta)$ in the prior specification for β but kept it the same as for the other transitions.

We calculated posterior moments of the parameters using the Bayes linear Bayes approach, with conjugate updates using both the log-mode and the log-normal methods and with non-conjugate updates. We also calculated posterior distributions using a full-Bayes analysis. The log-mode and log-normal methods were chosen because the log-mode method was used by Wilson and Farrow (2017) and the lognormal method is closer to the

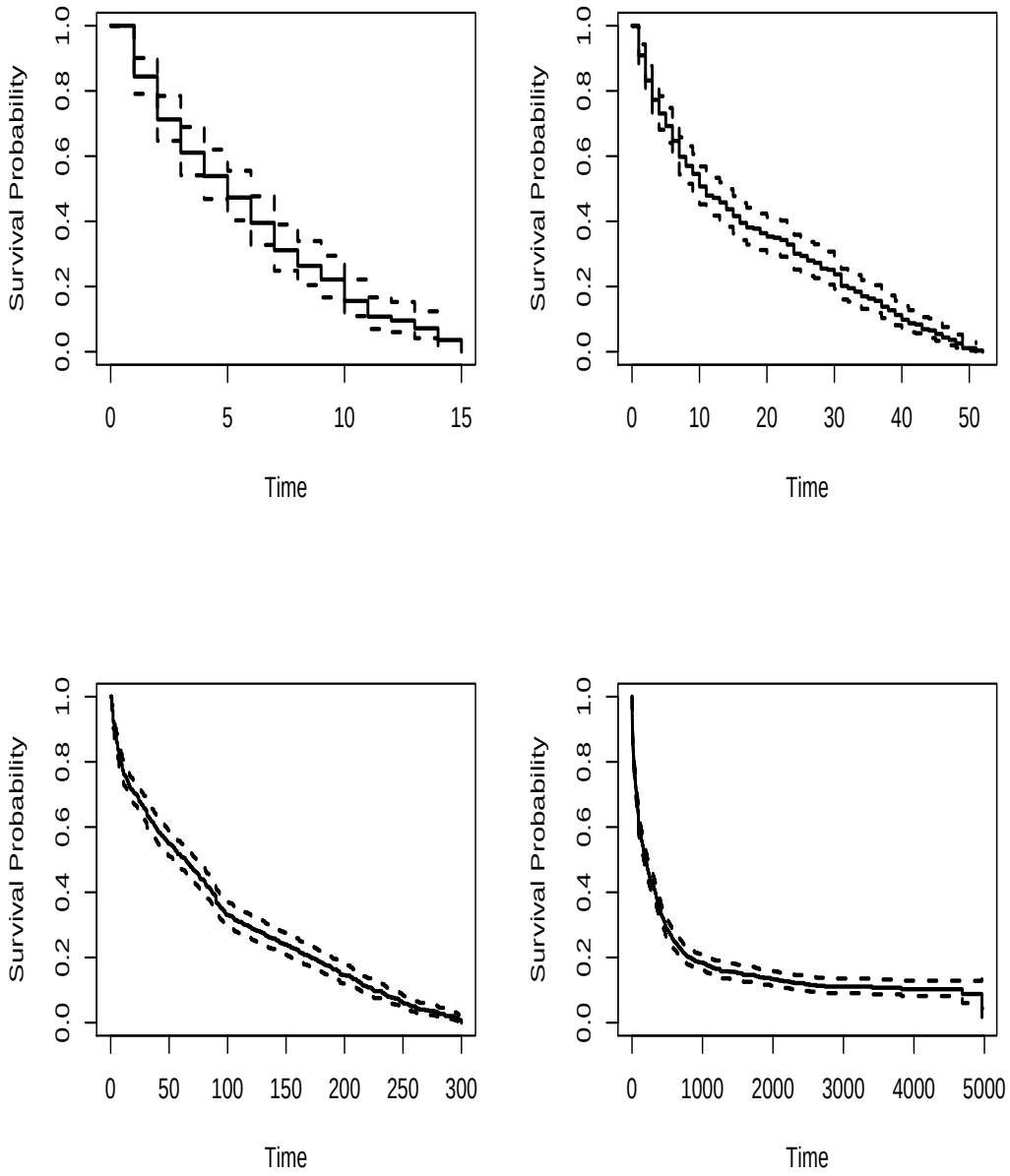


Figure 7.2: Kaplan-Meier estimates $\hat{S}(t)$ with confidence intervals with several lifetime such as 15, 52, 300 and all the observations.

j	τ_j	$\beta_{j1} \times 10^2$	$\beta_{j2} \times 10$	$\beta_{j3} \times 10^3$	$\beta_{j4} \times 10^2$
1	10.5	4.352 (0.331)	-0.199 (0.665)	4.625 (0.526)	3.275 (1.342)
2	52.7	3.692 (0.356)	0.252 (0.703)	3.072 (0.800)	4.262 (1.515)
3	111.6	3.126 (0.386)	0.263 (0.793)	2.882 (0.919)	2.540 (1.774)
4	178.3	2.502 (0.403)	0.310 (0.858)	3.465 (1.007)	1.266 (1.959)
5	255.4	2.345 (0.416)	0.589 (0.896)	3.564 (1.106)	1.147 (2.013)
6	346.6	1.823 (0.423)	0.420 (0.912)	2.589 (1.265)	0.747 (2.069)
7	458.1	1.557 (0.440)	0.090 (0.961)	2.595 (1.364)	1.137 (2.170)
8	602.0	1.590 (0.493)	0.425 (1.047)	3.449 (1.499)	0.033 (2.373)
9	804.7	1.495 (0.532)	1.106 (1.149)	1.762 (1.712)	-2.322 (2.720)
10	1151.3	2.329 (0.580)	1.429 (1.254)	-0.709 (2.063)	-4.815 (2.983)
11	∞	2.887 (0.788)	1.523 (1.647)	-1.179 (2.539)	-4.548 (4.045)

Table 7.2: Posterior means and standard deviations for each of the parameters in each interval using the non-conjugate method.

non-conjugate method.

We compare the results from four methods. Table 7.2 gives us the posterior means for the effects of all the covariates in all the intervals and the posterior standard deviations are given in brackets, using the non-conjugate method. When we look at Table 7.2, we notice that all the posterior means of the effect of age are positive and most of those for sex as well. That means increasing the age and the sex being male both increase the hazard of death from leukaemia.

Furthermore, the age effect is decreasing over time and then eventually increasing while the sex effect increases over the time. So, in general, all the covariates in the model have an effect on the survival time for the patients with leukaemia.

Figure 7.3 shows the comparison between Bayes linear Bayes analysis using the conjugate prior based on the log-normal and log-mode method, full-Bayes analysis and the non-conjugate method. We see that, the posterior means for the Bayes linear kinematic method using the non-conjugate prior gives very similar results to the full-Bayes approach. However, the computations for the non-conjugate BLK were much faster than those for the full Bayes approach.

We have also noticed that the posterior means using the log-normal method are also very close to our posterior means using the non-conjugate method.

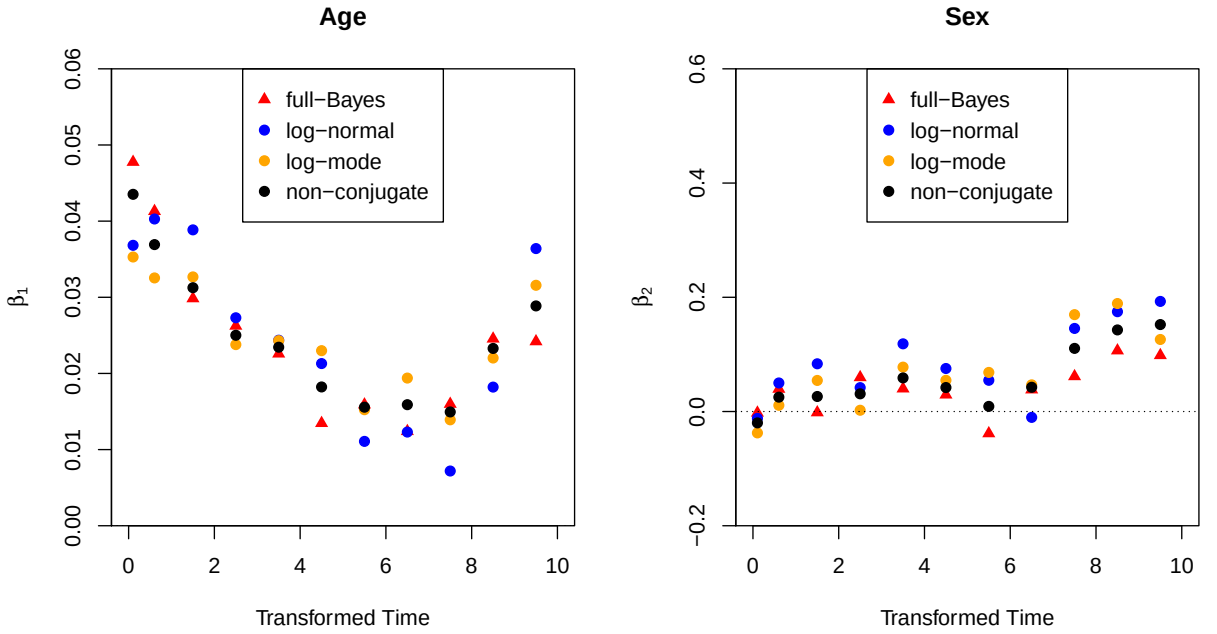


Figure 7.3: The effect of age and sex on the hazard functions of individuals with leukaemia. Triangles represent the posterior means for the full Bayesian method, circles represent different types of Bayes linear Bayes methods such as the black colour represents the posterior means using the non-conjugate prior update method. The transformed time is $[1 - \exp(-t/\nu)]/u$ with $u = 0.1$ and $\nu = 500$. The posterior means are plotted at the mid-points of the time intervals on the transformed scale.

7.3.7 Diagnostic checking in the leukaemia example

In this section, we produce some residual plots to check the validity of the assumptions made in the leukaemia example.

7.3.7.1 Residuals in survival analysis

For a general, non-Bayesian, discussion of residuals in survival analysis see, for example, Collett (1994). Several types of residuals have been defined for survival models. These include Cox-Snell residuals (Cox and Snell, 1968).

In non-Bayesian analyses, the Cox-Snell residual for patient i is an estimate of $-\log S_i(t_i)$, where $S_i(t)$ is the survival function for patient i and t_i is that patient's survival time. In the case of a censored observation, the Cox-Snell residual is also censored.

Clearly there is a 1−1 correspondence between $-\log S_i(t_i)$ and $F_i(t_i)$, where $F_i(t) = 1 - S_i(t)$ is the cumulative distribution function (cdf) of the lifetime distribution for patient i. Given $F_i(t)$ and that the lifetime T_i is a random value from this distribution, then $F_i(T_i)$ has a uniform $U(0,1)$ distribution. Furthermore $\Phi^{-1}[F_i(T_i)]$ has a standard normal distribution, where $\Phi()$ is the standard normal cumulative distribution function.

In the context of a model fitted to data, an observed lifetime t_i is known but the parameters of $F_i(t)$ have a posterior distribution. Therefore $F_i(t_i)$ also has a posterior distribution and we can compute summaries, such as quantiles, of this distribution. In the case of a censored observation, the actual lifetime T_i is also unknown but we can compute summaries of the posterior predictive distribution of $F_i(T_i)$.

The two examples in this chapter, leukaemia and non-Hodgkin lymphoma, are problems of different kinds. The details of how residuals are calculated and used differ between them.

7.3.7.2 Computing the residuals in the leukaemia example

The leukaemia example is a purely Bayes linear Bayes example in which the objective is to make inferences about the values of the model parameters. Because the model is a piecewise constant hazard model, all of the parameters are the coefficients $\underline{\beta}$ in a linear model (including the baseline). The inferences are calculated using Bayes linear kinematics (BLK). As no Markov chain Monte Carlo (MCMC) computations are involved, we can not compute the residuals as a byproduct of the MCMC.

The Bayes linear Bayes model does not assign a distribution (either prior or posterior) to $\underline{\beta}$. However, for the purpose of computing residuals, as an approximation, we use a multivariate normal distribution for the posterior distribution of $\underline{\beta}$. We use the posterior mean vector and variance-covariance matrix calculated by BLK. We draw a large number, M, eg $M = 1000$, of random samples of $\underline{\beta}$ from this multivariate normal distribution. Let these randomly sampled vectors be $\underline{\beta}_1, ..., \underline{\beta}_M$. For each sampled vector $\underline{\beta}_m$ and each patient i, we calculate a residual $R_{m,i}$ as follows.

If the lifetime t_i for patient i is observed (not censored) then we simply calculate $R_{m,i} = F_i(t_i; \underline{\beta}_m)$ which is the cdf evaluated at time t_i with the covariate values for patient i and the parameter values $\underline{\beta}_m$.

If the observation for patient i is right-censored at time c_i then we calculate $F_i(c_i; \underline{\beta}_m)$. If the actual lifetime for patient i is $T_i > c_i$, then $R_{m,i} = F_i(T_i; \underline{\beta}_m)$ is uniformly distributed

on the interval $(F_i(c_i;\underline{\beta}_m), 1)$. So we draw a random sample $U_{m,i}$ from the uniform distribution $U(0,1)$ and set $R_{m,i} = F_i(c_i;\underline{\beta}_m) + U_{m,i}[1 - F_i(c_i;\underline{\beta}_m)]$.

Having done this, for each patient i, we have a random sample of M draws from the distribution of the patient's residual. We can calculate summaries, such as the three quartiles, from these values.

`R` code is shown in Appendix A.7.2 to compute the residuals and `R` a function to find the cdf of a piecewise constant hazard model is shown in Appendix A.7.3. The number of patients is n and the number of saved sets of parameter values is M. As a result, we can compute the three quartiles of each residual.

For the purpose of the graphs which follow, the medians of the residuals are used.

7.3.7.3 Results

Figure 7.4 shows a histogram of the residuals in the leukemia example. This suggests that these residuals have approximately a uniform distribution.

We also need to plot the residuals against the covariates. For example, we plot Age against the residuals in Figure 7.5. It is clearly "random scatter" of points. There is no particular pattern shown in this graph. So, our assumptions are plausible as the residuals are distributed with constant variance for both males and females.

Figure 7.6 shows the scatter plot of log(WBC) and residuals. Again we do two plots, one for males and one for females. Also, there is no concern about any particular changes in the variance.

In Figure 7.7, we plot the Deprivation score against residuals for males and females, again we need to assess the validity of the model assumptions. These residuals are in the range (0,1) and show no dependence on the Deprivation score.

Figure 7.8 shows the scatter plot of the posterior means of η against residuals for males and females. We can see a random pattern indicating a good fit for the model. As a result, we conclude that there is no reason to reject the model assumptions.

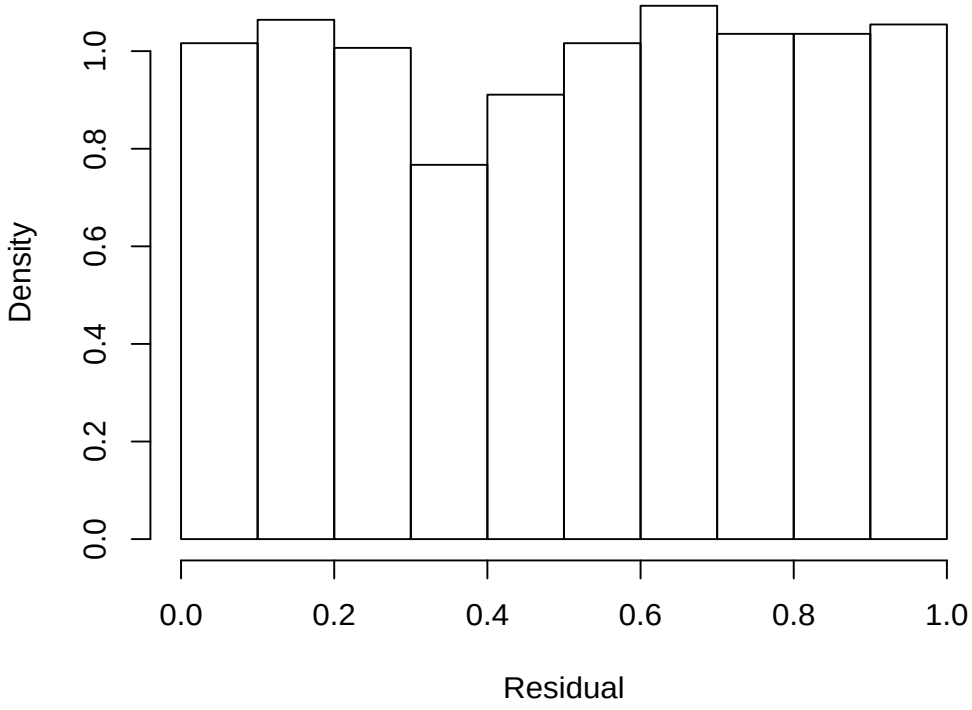


Figure 7.4: Histogram of the posterior medians of the residuals in leukaemia example.

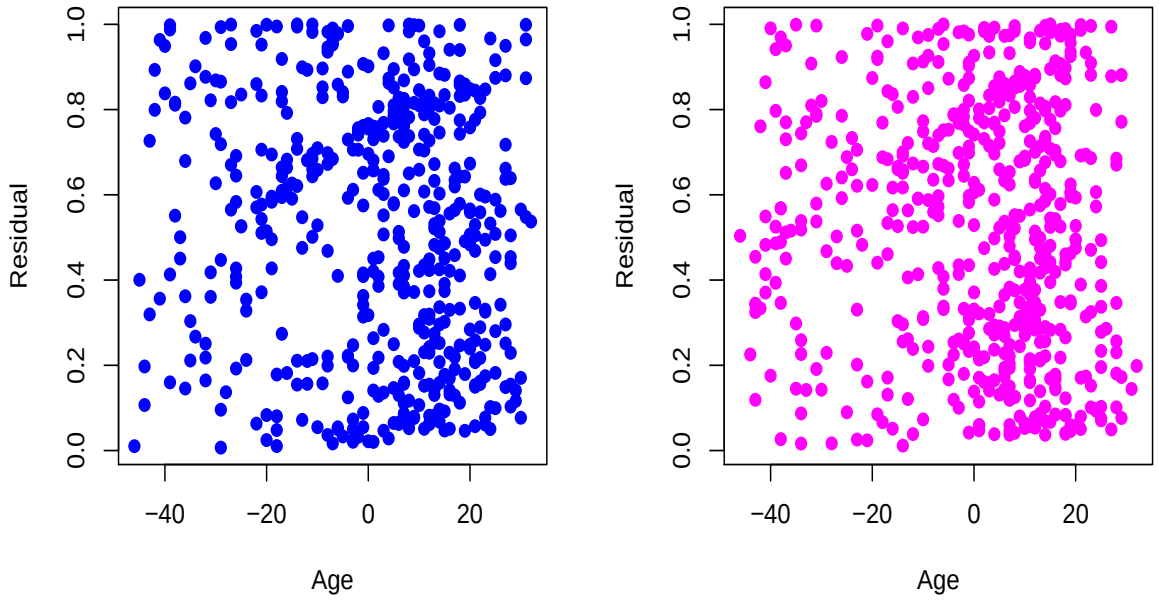


Figure 7.5: Scatter plots for Age against residuals for both sexes. The blue dots for males and pink dots for females.

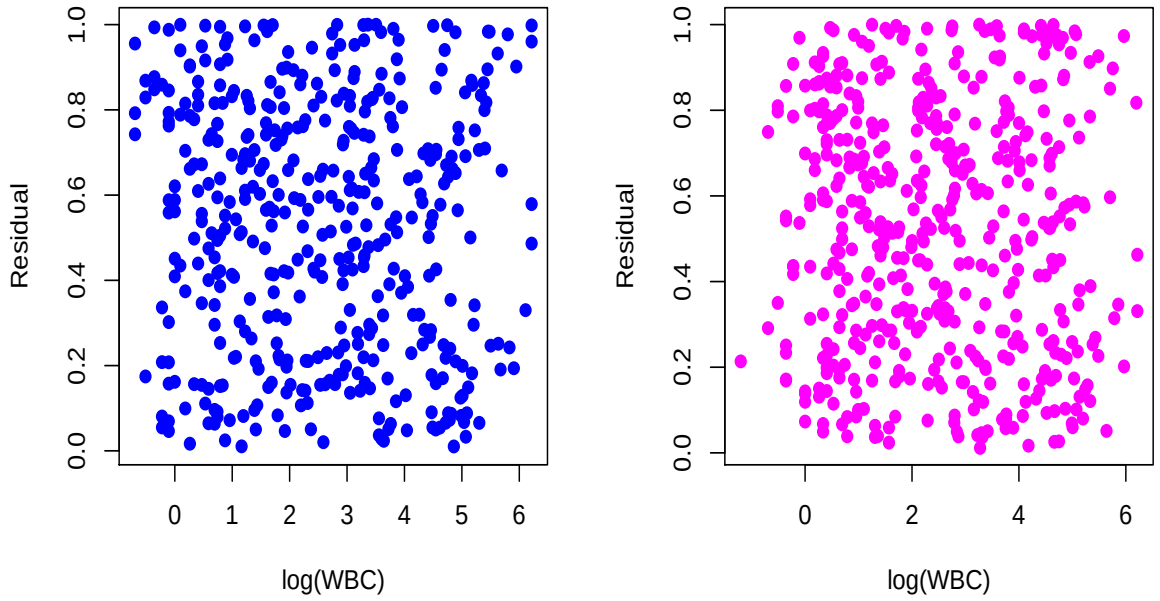


Figure 7.6: Scatter plots for log(WBC) against residuals for both sexes. The blue dots for males and pink dots for females.

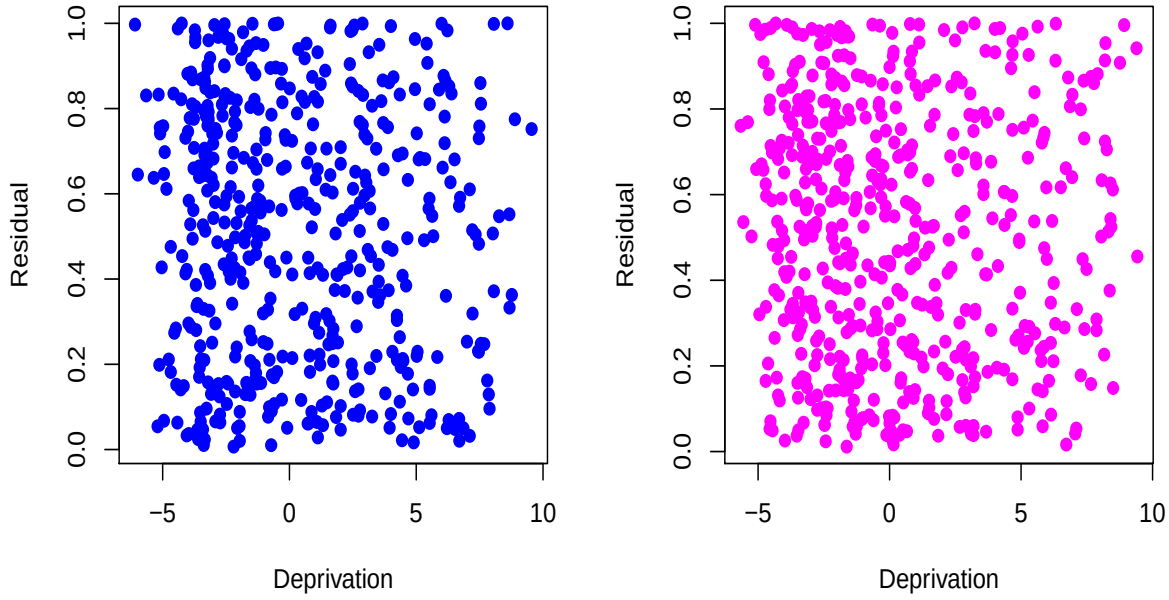


Figure 7.7: Scatter plots for Deprivation score against residuals for both sexes. The blue dots for males and pink dots for females.

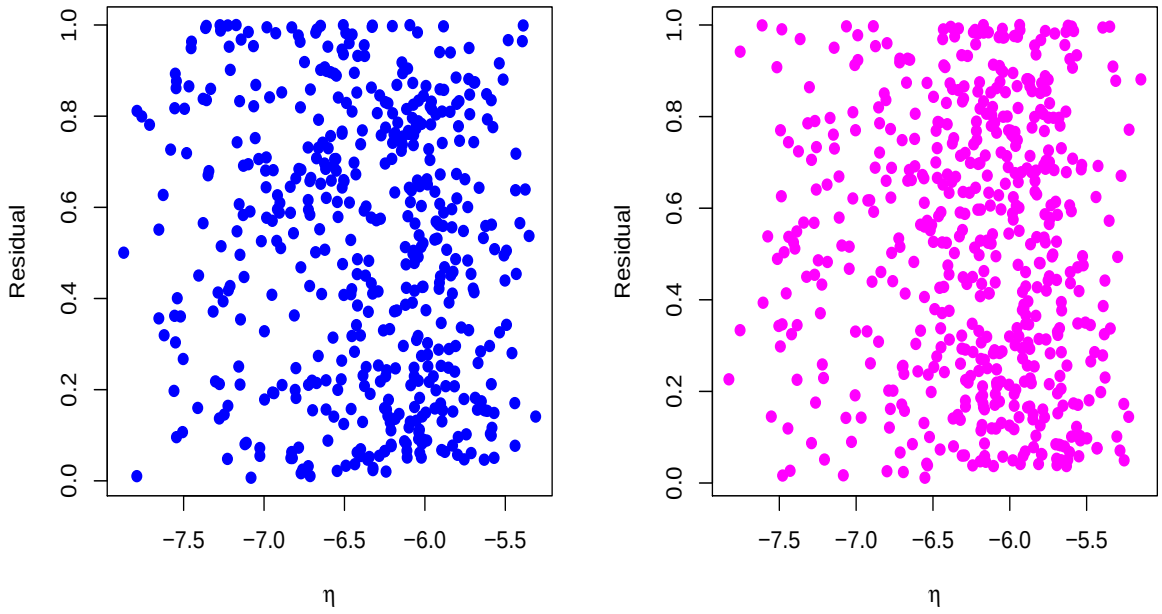


Figure 7.8: Scatter plots for the posterior mean of η against residuals for both sexes. The blue dots for males and pink dots for females.

7.4 Bayes linear Bayes prognostic networks

7.4.1 Introduction

We now consider applications of a different type. We wish to build a system to calculate prognostic index values for individual new patients. This calculation will be done using Bayes linear kinematics. This makes the calculations very fast and, as we shall see, we will be able to compute a value even when observations on some covariates are missing. In Section 7.4.2, we will explain the use of a latent prognostic index and its advantages. In Section 7.4.3, we will describe the general structure of a Bayes linear Bayes prognostic network. In Section 7.5, we will describe how such networks are built, including the use of historical data. In Section 7.6, we will apply these ideas to an example refering to patients with non-Hodgkin lymphoma.

7.4.2 The use of a latent prognostic index

A traditional prognostic index measures the hazard of an individual relative to the baseline in a proportional hazards model. Typically, it is the logarithm of the relative hazard. See

Section 5.6. If we have a fixed list of covariates $S = \{X_1, ..., X_J\}$, then the index must be a function of the values $x_1, ..., x_J$ taken by these covariates. Conventionally, when constructing a prognostic index, we try to choose a suitable function $g(x_1, ..., x_J)$. Suppose now that we have a list $S_{\max} = \{X_1, ..., X_J\}$ of covariates which can be observed but that we might not always observe all of $X_1, ..., X_J$ but rather we observe some subset of $S_{\max}$. Suppose that the possible observed subsets are $S_1, ..., S_M$. We need a different function g_m for each possible subset S_m. To do this in a coherent and principled way, we introduce the idea of a latent variable Z_T. We will refer to this as a prognostic index but we will not observe it. Instead, when we supply an index value to a user, we will give our current expectation of Z_T, given the information available to us. As in a traditional prognostic index, Z_T is a quantity on which the lifetime distribution depends. For example, in a Weibull model with survival function $\exp\{-\lambda_i t^\alpha\}$ for subject i, we can use $Z_T = \eta_i = \log(\lambda_i)$ as the prognostic index. If

$$\eta_i = \beta_0 + \sum_{j=1}^{J} \beta_j X_{i,j},$$

where $X_{i,j}$ is the value of covariate j for subject i, but not all of $X_{i,1}, ..., X_{i,k}$ are observed for subject i, then we use $I = \mathrm{E}(\eta_i \mid S_i)$ where S_i is the subset of observations made for patient i. We refer to I as the *predicted* prognostic index value.

This allows us to compute a (predicted) prognostic index value given observations of any subset of the possible covariates, for example when some values are missing or when some variables are only measured in certain cases. Furthermore, the use of Bayes linear kinematics and a Bayes linear Bayes model allows us to do this quickly and efficiently.

Additional flexibility is provided by modelling the joint distribution of Z_T and the covariates, often through latent variables associated with the covariates, so that Z_T is not known precisely even when all of the covariates are observed. In this way we always use an expectation of Z_T as our declared index value.

7.4.3 Prognostic networks

Suppose that the nodes $X_1, ..., X_J$ can represent covariates. With each covariate X_k we associate a variable Z_k which may be unobserved. The value, or more generally, the distribution of X_k depends on the value of Z_k. Just as X_k depends on Z_k, suppose that we have a lifetime T which depends on Z_T. In general we can have covariates

$X_1, ..., X_J$ depending on $Z_1, ..., Z_J$ respectively and then an additional dependent node T depending on an element Z_{J+1}. The variables $Z_1, ..., Z_{J+1}$ are related in a Bayes linear structure. Then Z_{J+1} represents our prognostic index. When we observe some or all of the covariates this changes our expectation of Z_{J+1} and therefore the index value which we would report. See Figure 7.9. Notice that, in Figure 7.9, the red, undirected, edges represent a Bayes linear structure and that the blue directed edges represent fully-specified conditional probability distributions.

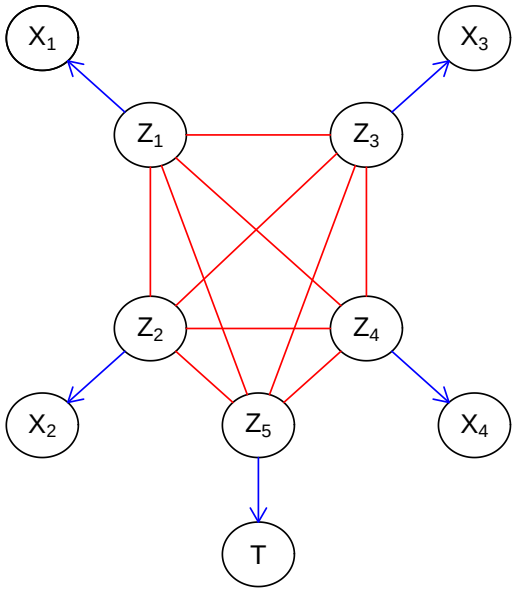


Figure 7.9: Bayes linear Bayes graphical model

The introduction of a latent variable Z_k associated with the covariate X_k allows covariates of different types to be used such as ordinal variables and censored variables. However, we are free to set $X_k = Z_k$ if this is appropriate.

We need to specify a mean vector and a variance-covariance matrix for the elements of the Bayes linear structure, $Z_1, ..., Z_J, Z_{J+1}$. The variance-covariance matrix might be developed in a general, unstructured way as suggested by Figure 7.9. Alternatively, we might impose some structure and exploit conditional independences, perhaps by introducing mediating nodes which induce correlation between related covariates. This might be done by expert judgement. A subjective covariance structure might be developed using an approach similar to methods described in Farrow (2003). On the other hand we might use an automatic method, using an algorithmic approach to determine a suitable network structure. Methods for structure learning for Bayesian networks are discussed

in, for example, Heckerman and Chickering (1995); Neapolitan (2003); Margaritis (2003) and Wang et al. (2015). See Chapter 4.

We might also select certain important variables which are always observed and condition the rest of the model on these.

However we determine the structure, we need to quantify it by specifying means, variances and covariances. Again these might be chosen subjectively. More likely we will use historical data and use an offline learning phase in which we fit an analogous model, with a fully specified prior distribution, using, for example, Markov chain Monte Carlo (MCMC) methods to compute posterior summaries. This latter approach is described in our example in Section 7.6.3.

Once we have a fully specified model, in routine use with new patients, we compute adjusted expectations of the prognostic index given observations of some or all of the covariates. Because we can do the calculation when only a subset of the covariates is observed, we can include a greater number of potential covariates in our model and therefore use more information when it is available.

7.5 Construction of Bayes linear Bayes networks

7.5.1 General strategy

While, in some cases, we might construct our network using only the subjective judgements of experts, more typically we might use historical data to learn values of parameters in our model.

We have an offline learning phase and we use values that we infer from that model in the network. In practice, we use the posterior expectation that we obtained from the offline learning model to construct the Bayes linear Bayes prognostic network. At this stage in our research, we use the posterior means of model parameters as values in our Bayes linear Bayes model. The historical data are, however, independent of future patients, given the model parameters. This raises the possibility, which we will pursue in future research, that we can avoid any such compromise and obtain exactly the expectations which we need. For example, in (6.14) and (6.15), clearly we can obtain the posterior expectations of $\mathrm{P}(\underline{X})$ and $\mathrm{P}(\underline{X})\mathrm{E}(\underline{X})$ directly from the MCMC computations in the learning phase. However, further work is required to address the problem of parameter uncertainty in the

adjusted expectations and precisions. Nevertheless, with a large historical data set, such effects are likely to be small.

7.5.2 Specifying the covariance structure

First of all, we need to construct a variance-covariance matrix of the possibly latent covariates Z. As an alternative to using an inverse-Wishart prior, we can use the following approach.

Suppose that

$$\underline{Z} = G\underline{\varepsilon} + \underline{\mu}.$$

where $\varepsilon \sim N(0, V_\varepsilon)$, and, for example, when $\underline{Z}$ has five elements, V_ε can be written as

$$V_\varepsilon = \begin{bmatrix} \tau_1^{-1} & 0 & 0 & 0 & 0 \\ 0 & \tau_2^{-1} & 0 & 0 & 0 \\ 0 & 0 & \tau_3^{-1} & 0 & 0 \\ 0 & 0 & 0 & \tau_4^{-1} & 0 \\ 0 & 0 & 0 & 0 & \tau_5^{-1} \end{bmatrix}.$$

So the matrix G will be

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \gamma_{21} & 1 & 0 & 0 & 0 \\ \gamma_{31} & \gamma_{32} & 1 & 0 & 0 \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & 1 & 0 \\ \gamma_{51} & \gamma_{52} & \gamma_{53} & \gamma_{54} & 1 \end{bmatrix}$$

Therefore, the variance-covariance matrix $\mathrm{Var}(\underline{Z}) = GV_\varepsilon G' = \Sigma$.

There are four reasons to use this approach rather than just use, for example, an inverse Wishart prior.

1. This structure lends itself to using a more structured network, with some arcs missing.

2. This structure allows us to use a collection of univariate normal distributions rather than a multivariate normal distribution. This avoids problems with standard MCMC software such as JAGS when a multivariate normal vector is sometimes only partially observed.

3. The generalised autoregressive stucture automatically creates a missing-data model. Ibrahim et al. (2001), Section 8.3, suggest a sequence of conditional distributions for covariates. See also Zhao (2010).

4. The generalised autoregressive structure provides greater flexibility in specifying a prior distribution for the variance-covariance matrix than an inverse-Wishart prior. In an inverse-Wishart prior, once $\mathrm{E}(\Sigma)$ is specified, only one parameter is left to specify the uncertainty. In contrast, in the generalised autoregressive structure, we can give the regression coefficients $\gamma_{21}, \gamma_{31}, \gamma_{32}, \ldots$ a multivariate normal prior and also give, for example, a multivariate normal prior to the logarithms of the precisions $\tau_1, \tau_2, \ldots$.

Therefore, we use this structure, adapted from Pourahmadi (1999); Daniels and Pourahmadi (2002). This structure uses a square-root-free Cholesky decompsition of Σ^{-1} as follows.

The Cholesky decomposition of a symmetric positive definite matrix Q with dimension $p \times p$ can be expressed in the following way

$$Q = \widetilde{S}\widetilde{S}'$$

where $\widetilde{S}$ is non-singular and lower-triangular with elements $\widetilde{\zeta}_{j,k}$. Suppose that for each k we divide column k of $\widetilde{S}$ by $\widetilde{\zeta}_{k,k}$ so that

$$Q = SDS' \tag{7.9}$$

where D is diagonal with elements $\widetilde{\zeta}_{1,1}, ..., \widetilde{\zeta}_{p,p}$ and S is lower triangular with elements $\zeta_{j,k} = \widetilde{\zeta}_{j,k}/\widetilde{\zeta}_{k,k}$ for $j > k$ and unit diagonal. This is a square-root-free Cholesky decomposition of Q. See Watkins (2004).

The idea of using a "generalised autoregressive" representation was introduced by Pourahmadi (1999), to provide a non-restrictive parameterisation of the covariance matrix. Suppose that we can write $Z_1 = \varepsilon_1$ and, for $j = 2, ..., J$,

$$Z_j = \sum_{i=1}^{j-1} \phi_{j,i} Z_i + \varepsilon_j,$$

where $\varepsilon_j \sim N(0, \tau_j^{-1})$ and ε_j is independent of $\varepsilon_{j'}$ unless $j = j'$. The coefficients $\phi_{j,i}$ are generalised autoregressive parameters. Therefore, $\underline{\varepsilon} = (\varepsilon_1, ..., \varepsilon_j)' = PZ$ where P is

the lower triangular matrix with unit diagonal and, for $j > k$, elements $-\phi_{j,k}$. So the diagonal covariance matrix of ε is $D^{-1} = P\Sigma P'$, where Σ is the covariance matrix of Z. Rearranging this we obtain

$$\Sigma^{-1} = P'DP.$$

where P' is upper triangular. By reversing the order of the elements of the vectors we obtain (7.9).

7.5.3 Offline learning

Once we have determined a structure for our model we assume that $\underline{Z}$ has a multivariate normal distribution and, with suitable prior distributions assigned, use MCMC to compute posterior means for model parameters, as discussed in Section 7.5.1.

The assumption of a multivariate normal distribution for $\underline{Z}$ is convenient but, given that we can choose a link function between the element of $\underline{Z}$ and the covariates $X_1, ..., X_J$ and T, it is not restrictive.

Specifically we use a generalised autoregressive structure for $\underline{Z}$ in the model fitting. So we set

$$Z_1 = \mu_1 + \varepsilon_1$$

where $\varepsilon_1 \sim N(0, \tau_1^{-1})$. Then, for $j > 1$, we set

$$Z_j = \mu_j + \sum_{k=1}^{j-1} \gamma_{j,k}(Z_k - \mu_k) + \varepsilon_j$$

where $\varepsilon_j \sim N(0, \tau_j^{-1})$ and $\varepsilon_1, \ldots, \varepsilon_J$ are independent.

We then give multivariate normal priors to $\mu_1, \ldots, \mu_{J+1}$ and to $\gamma_{2,1}, \ldots, \gamma_{J+1,J}$. We also give priors to the conditional precisions $\tau_1, \ldots, \tau_{J+1}$. It is simple to give these parameters independent gamma priors but we could also, for example, give a multivariate normal prior to $(\log \tau_1, \ldots, \log \tau_{J+1})$.

The mean vector for $\underline{Z}$ is just $\underline{\mu} = (\mu_1, \ldots, \mu_{J+1})'$ and the variance-covariance matrix of $\underline{Z}$ is given by

$$\Sigma = GD^{-1}G' \tag{7.10}$$

where D is the diagonal matrix with diagonal elements $\tau_1, \ldots, \tau_{J+1}$ and G is the lower triangular matrix with diagonal elements $g_{j,j} = 1$ and off-diagonal elements $g_{j,k} = \gamma_{jk}$ for $j > 1$ and $k < j$.

7.6 Example: Non-Hodgkin lymphoma

7.6.1 Introduction

As an example we use patients with non-Hodgkin lymphoma. The historical data were collected by the Scotland and Newcastle Lymphoma Group from patients in Scotland and the North of England, UK, (Proctor and Taylor, 2000), See Section 2.2. Apart from survival time, which is subject to right censoring, the variables include **Age, Sex, Stage** (Ann Arbor Stage, Carbone et al., 1971), **ECOG** (Eastern Cooperative Oncology Group performance status, Oken et al., 1982), the last two of which are both ordinal variables, and a large number of other covariates, some of which may or may not be observed. Some of these are binary and some are interval-censored, since the results were either recorded as "normal", if the measurement was inside the normal range, or as an actual value if it was not. For further details, see, for example, Zhao (2010). See also Chapter 2. In our example, for illustration, we use a subset of these covariates. We use Age, Sex, Haemoglobin (HB), White Blood Cell (WBC), Stage and Albumin. Of these, HB and WBC are continuous variables, Stage is ordinal and Albumin is binary.

We chose to separate **Age** and **Sex**, which are always observed, and to condition the rest of the model on these. Thus the means of $Z_1, ..., Z_m, Z_{m+1}$, but not the variance and covariances, depend via a linear model on age and sex.

We adopted a general covariance structure for the Bayes linear network and we impose the order of the covariates in the following expression {hb, wbc, stage, albumin, T}, since we always observed {hb, wbc} to form a generalised autoregression. In future work we plan to investigate the use of more structure.

Thus our generalised autoregressive structure becomes the following. We set

$$Z_1 = \mu_{0,1} + \mu_{\text{age},1} x_{\text{age}} + \mu_{\text{sex},1} x_{\text{sex}} + \varepsilon_1$$

where $\varepsilon_1 \sim N(0, \tau_1^{-1})$. Then, for $j > 1$, we set

$$Z_j = \mu_{0,j} + \mu_{\text{age},j} x_{\text{age}} + \mu_{\text{sex},j} x_{\text{sex}} + \sum_{k=1}^{j-1} \gamma_{j,k}(Z_k - \mu_k) + \varepsilon_j$$

where $\varepsilon_j \sim N(0, \tau_j^{-1})$ and $\varepsilon_1, \ldots, \varepsilon_J$ are independent. Here x_{age} is the patient's age in years minus 60 and x_{sex} is 1 for a male patient and -1 for a female patient.

In our offline learning model, we suppose that the distribution of T_i, the lifetime of patient i is a Weibull distribution with two parameters, α and λ_i, where $\log(\lambda_i) = Z_{t,i}$ and $Z_{t,i}$ is the prognostic index value for patient i.

Then we use the generalised autoregressive structure to relate each variable with others in the model. All the variables that we use in the example have means which are conditional on age and sex since we always observed age and sex.

As a result, we obtain the posterior distribution for all the parameters in the model and then use the posterior means to produce a Bayes linear kinematic network.

We obtain the variance-covariance matrix for $\underline{Z}$ from the coefficients $\gamma_{j,k}$ using (7.10).

7.6.2 Exploratory plots in the non-Hodgkin lymphoma example

Before constructing our model, we should look at some plots. So, we plot the covariates used in the NHL example. These covariates are Age, Sex, HB, WBC, Stage and Albumin. In Figure 7.10, we plot Age against Stage which has 4 levels and for both sexes. From these boxplots, we notice that there is no change in the difference between male and female in each stage. Similarly, the boxplots of log(HB) and Stage show no indication that the difference in log(HB) between males and females depends on Stage.

Figure 7.11 shows the boxplots of plotting log(WBC) against Stage and for males and females. This again suggests that the difference between log(WBC) does not depend on the Stage. The same conclusion can be drawn from plotting $\log(T)$ against Stage in Figure 7.11.

Albumin is classified into two categories, Albumin 1 and Albumin 2. In Figure 7.12, we plot Albumin against Age, log(HB), log(WBC) and for males and females. These plots show no concern about the validity of the model assumptions.

Figure 7.13 shows scatter plots for the continuous variables in the data set, Age, HB

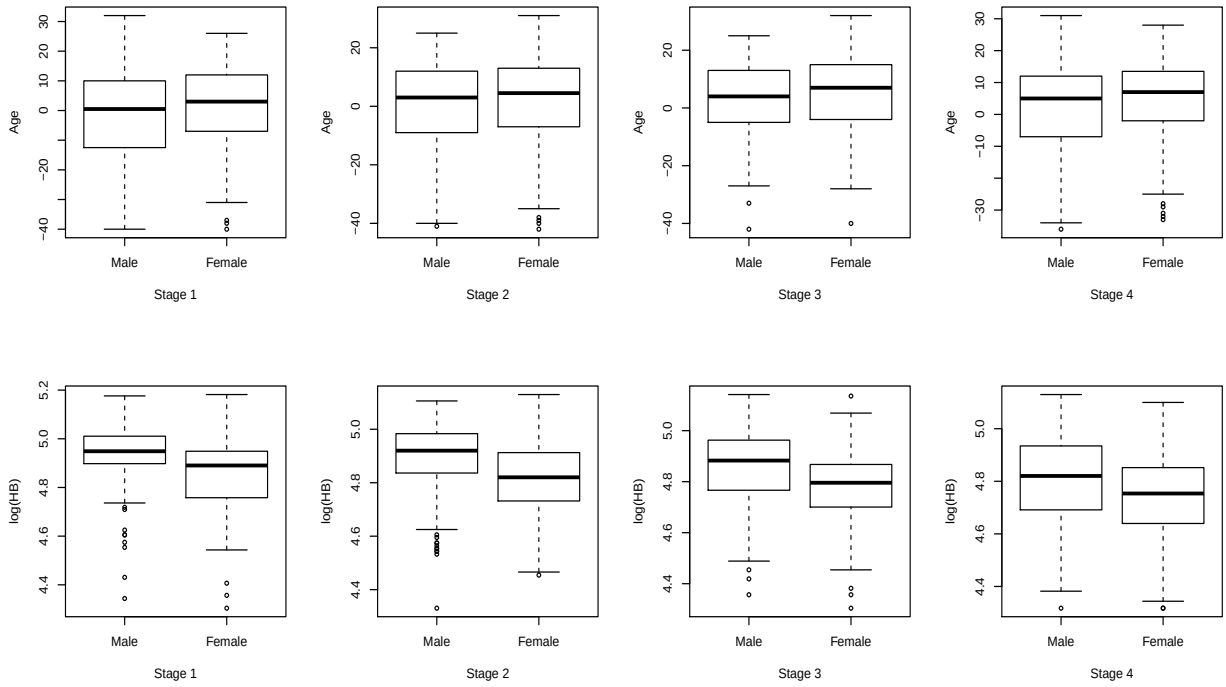


Figure 7.10: Box plots for Stage and Age and box plots for Stage and log(HB).

and WBC. From these scatter plots, we can not see any pattern in the observations which suggests a violation of the model assumptions, so we carry on the analysis.

7.6.3 Offline learning: Introduction

We use a full-Bayes specification and MCMC to learn the values of the parameters. As an example, we use a data set containing 1391 patients with non-Hodgkin lymphoma. See Chapter 2. A large proportion of these patients had at least some missing covariate values. Therefore, it was necessary to include a missing-data model in our model specification. So we construct a model for the joint distribution of all of the variables, including the covariates. This is done using a generalised autoregressive structure. The main model was the same as the Bayes linear Bayes model except that we specified a prior distribution for the unknown model parameters, including the thresholds for ordinal covariates and the means and variance-covariance matrix in the Bayes linear structure. The prior for the parameters of the Bayes linear structure was specified using the generalised autoregressive approach described in Section 7.6.1. This structure implies the missing-data model. The coefficients and conditional precisions in the generalised autoregression were converted to

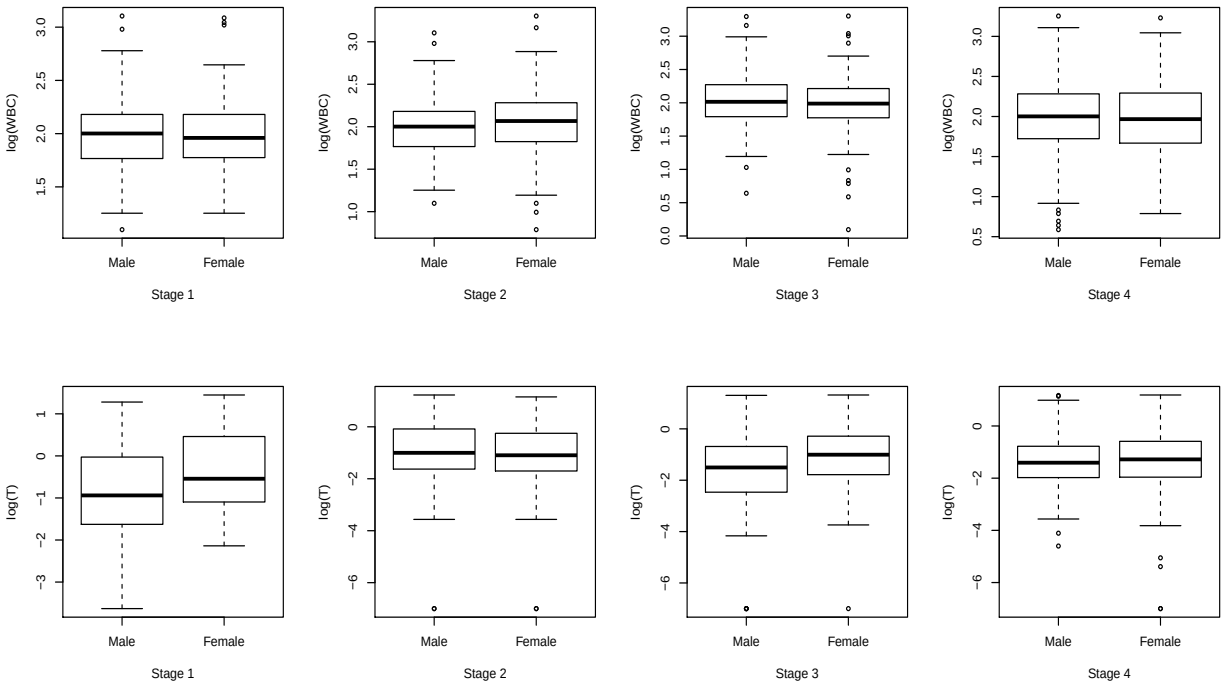


Figure 7.11: Box plots for Stage and log(WBC) and box plots for Stage and log(T).

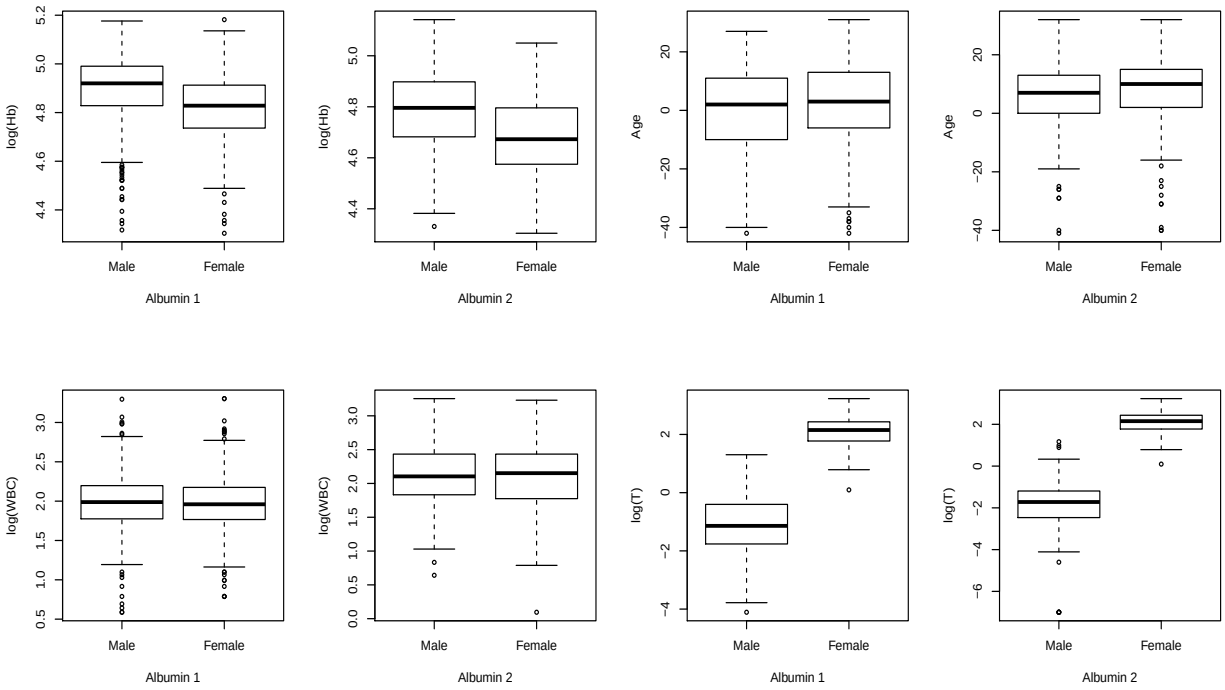


Figure 7.12: Box plots for Age and Albumin, log(HB) and Albumin, log(WBC) and Albumin, and Albumin and log(T).

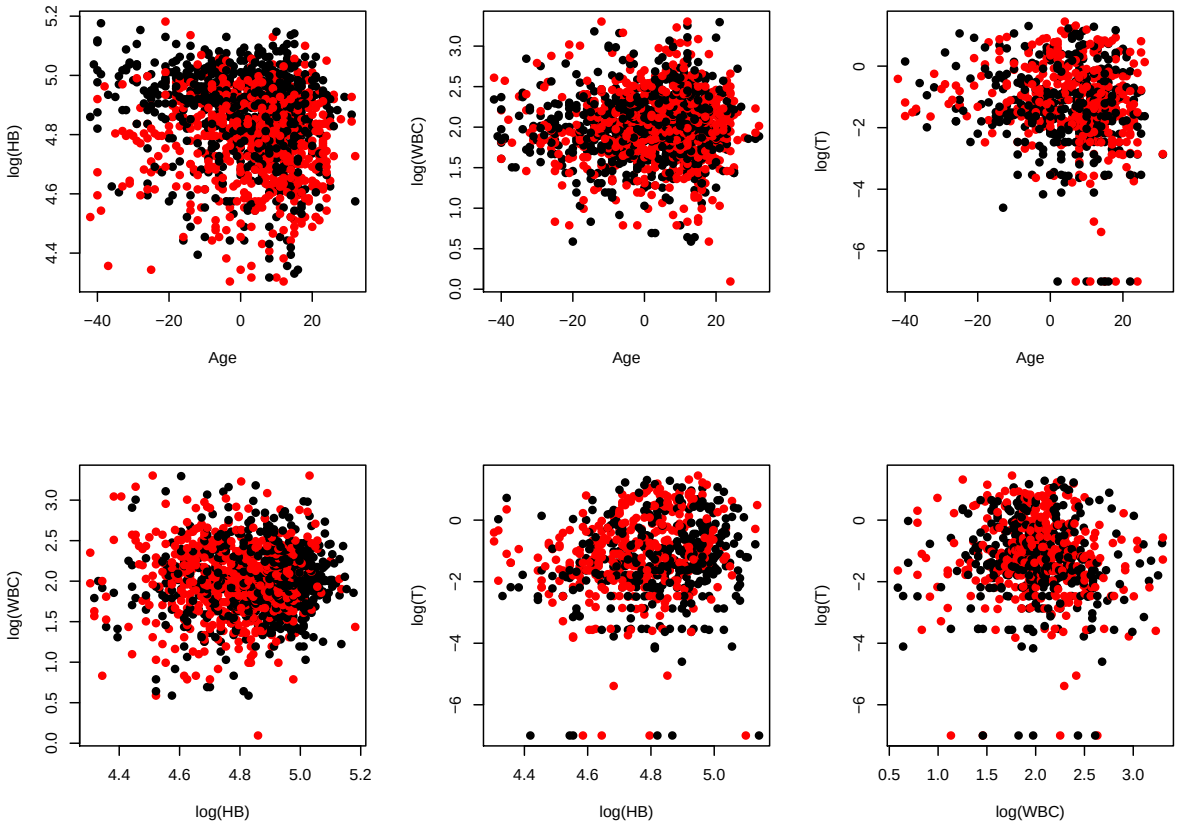


Figure 7.13: Scatter plots of covariates in non-Hodgkin lymphoma example against each other (i.e. Age, HB, WBC) and the lifetime T. Black dots for males and red dots for females.

give the variance-covariance matrix Σ for $\underline{Z}$ using (7.10).

7.6.4 Offline learning model with the direct method

In our example the computations were done using JAGS using the `R` package `rjags` (Plummer 2017; R Core Team, 2018).

Suppose we use the direct method with the case of the ordinal variables. Then, it is necessary to learn about the values of the thresholds, $c_1, ..., c_{K-1}$. However, since the mean and variance of Z are both unknown, for identifiability we fix two thresholds. Thus if $K = 3$, we can fix $c_1 = 0$ and $c_2 = 1$. If $K = 4$, then we can fix $c_1 = 0$ and $c_3 = 2$ and then give $c_2/2$ a scaled beta prior distribution. For $K > 4$, we can fix $c_1 = 0$ and $c_{K-1} = 1$ and then give $\{u_1, ..., u_{k-3}\}$ a Dirichlet prior distribution and let $c_{j+1} = \sum_{i=1}^{j} u_i$. For example, for **ECOG**, $K = 5$ for living patients.

In the case of Stage, $K = 4$. Without loss of generality we can fix $c_1 = 0$ and $c_3 = 2$. We make inference about the second cut point. We give c_2 a scaled beta prior distribution. So $c_2/2 = c^*$ where $c^* \sim \text{Beta}(a_c, b_c)$. We also assume that the underlying latent variable which is associated with this variable has a normal distribution with some mean and unknown variance σ_z^2, so $Z \sim N(\mu_z, \sigma_z^2)$ where the mean μ_z is also unknown. See Appendix A.7.4 for the `rjags` model specification using the direct method.

In the offline-learning phase, we use full Bayes analysis with MCMC computations. As a result, we learn about all the parameters that we need to produce the prognostic index values. Then, we use these parameter values from the offline learning as the prior in the BLK network. We use the generalised autoregressive structure as described in Section 7.6.1 and obtain posterior means for the mean vector and variance-covariance matrix of $\underline{Z}$ to use in the BLK network.

Let us explain how we use the parameters of our generalised autoregression in Table 7.3 to obtain these mean parameters and variance covariance matrix. For the direct model, we obtain the posterior means and posterior variances as described in Appendix A.7.4. Then, we calculate $\text{E}_0(Z)$ from the following

$$\underline{\mu}_0 = (\gamma_{0.hb}, \gamma_{0.wbc}, \gamma_{0.stage}, \gamma_{0.albumin}, \gamma_{0.t})',$$
$$\underline{\mu}_{age} = (\gamma_{age.hb}, \gamma_{age.wbc}, \gamma_{age.stage}, \gamma_{age.albumin}, \gamma_{age.t})',$$

Parameter	Mean	Standard deviation
$\gamma_{0.t}$	0.00	10.00
$\gamma_{sex.t}$	0.00	31.62
$\gamma_{age.t}$	0.00	31.62
$\gamma_{wbc.t}$	0.00	31.62
$\gamma_{albumin.t}$	1.00	100.0
$\gamma_{stage.t}$	0.00	31.62
$\gamma_{hb.t}$	0.00	31.62
$\gamma_{0.hb}$	100	100.0
$\gamma_{hb.age}$	0.00	31.62
$\gamma_{hb.sex}$	0.00	31.62
$\gamma_{0.wbc}$	10.0	31.62
$\gamma_{wbc.age}$	0.00	31.62
$\gamma_{wbc.sex}$	0.00	31.62
$\gamma_{wbc.hb}$	0.00	31.62
$\gamma_{0.stage}$	00.0	31.62
$\gamma_{stage.age}$	0.00	31.62
$\gamma_{stage.sex}$	0.00	31.62
$\gamma_{stage.hb}$	0.00	31.62
$\gamma_{stage.wbc}$	0.00	31.62
$\gamma_{0.albumin}$	00.0	31.62
$\gamma_{albumin.age}$	0.00	31.62
$\gamma_{albumin.sex}$	0.00	31.62
$\gamma_{albumin.hb}$	0.00	31.62
$\gamma_{albumin.wbc}$	0.00	31.62
$\gamma_{albumin.stage}$	0.00	31.62

Table 7.3: Prior means and prior standard deviations for each of the parameters in the NHL example.

and

$$\underline{\mu}_{sex} = (\gamma_{sex.hb}, \gamma_{sex.wbc}, \gamma_{sex.stage}, \gamma_{sex.albumin}, \gamma_{sex.t})^{'}.$$

Now, to obtain $V_0(Z)$, we have $\tau_z = (\tau_{z.hb}, \tau_{z.wbc}, \tau_{z.stage}, 1, \tau_{z.t})^{'}$ as we fixed $\tau_{z.albumin}$ to be 1 and we have $\tau_{z.hb} \sim \text{Gamma}(2, 300), \tau_{z.wbc} \sim \text{Gamma}(2, 30), \tau_{z.stage} \sim \text{Gamma}(2, 3)$ and $\tau_{z.t} \sim \text{Gamma}(1.5, 0.5)$.

We define the matrix V_ε exactly as in Section 7.5.2. We also define the matrix G as follows.

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \gamma_{21} & 1 & 0 & 0 & 0 \\ \gamma_{31} & \gamma_{32} & 1 & 0 & 0 \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & 1 & 0 \\ \gamma_{51} & \gamma_{52} & \gamma_{53} & \gamma_{54} & 1 \end{bmatrix}$$

where $\gamma_{21} = \gamma_{hb.wbc}, \gamma_{31} = \gamma_{hb.stage}, \gamma_{32} = \gamma_{wbc.stage}, \gamma_{41} = \gamma_{hb.albumin}, \gamma_{42} = \gamma_{wbc.albumin}, \gamma_{43} = \gamma_{stage.albumin}, \gamma_{51} = \gamma_{hb.t}, \gamma_{52} = \gamma_{wbc.t}, \gamma_{53} = \gamma_{stage.t}$ and $\gamma_{54} = \gamma_{albumin.t}$.

Therefore, the variance-covariance matrix will be $V_0(Z) = GV_\varepsilon G^T$.

As a result, in the BLK network, we use the prior mean $\text{E}_0(\underline{Z})$ and prior variance-covariance matrix $V_0(Z)$ as follows.

For the prior means we use

$$\underline{\mu}_0^{'} = \begin{matrix} \mu_0(Z_{HB}) & \mu_0(Z_{WBC}) & \mu_0(Z_{Stage}) & \mu_0(Z_{Albumin}) & \mu_0(Z_T) \\ \big(126.6473 & 8.0231 & 1.2037 & -0.8868 & 0.5150\big) \end{matrix}$$

$$\underline{\mu}_{age}^{'} = \begin{matrix} \mu_{age}(Z_{HB}) & \mu_{age}(Z_{WBC}) & \mu_{age}(Z_{Stage}) & \mu_{age}(Z_{Albumin}) & \mu_{age}(Z_T) \\ \big(-0.1777 & 0.0087 & 0.0121 & 0.0189 & 0.0098 \big) \end{matrix}$$

and

$$\underline{\mu}_{sex}^{'} = \begin{matrix} \mu_{sex}(Z_{HB}) & \mu_{sex}(Z_{WBC}) & \mu_{sex}(Z_{Stage}) & \mu_{sex}(Z_{Albumin}) & \mu_{sex}(Z_T) \\ \big(-4.9052 & -0.0084 & 0.0500 & 0.0165 & -0.0628\big) \end{matrix}$$

Therefore,

$$\mathrm{E}_0(\underline{Z}_i) = \mu_0 + \mu_{\text{age}} x_{age,i} + \mu_{\text{sex}} x_{sex,i}.$$

where $x_{age,i}$ is the age in years of patient i, minus 60 and $x_{sex,i}$ is 1 for a male patient or -1 for a female patient, and

$$V_0(Z) = \begin{array}{c} \\ Z_{HB} \\ Z_{WBC} \\ Z_{Stage} \\ Z_{Albumin} \\ Z_T \end{array} \begin{array}{c} \begin{array}{ccccc} Z_{HB} & Z_{WBC} & Z_{Stage} & Z_{Albumin} & Z_T \end{array} \\ \begin{pmatrix} 323.2214 & -2.1272 & -6.9139 & -8.6734 & -3.6989 \\ -2.1272 & 12.1289 & 0.0969 & 0.8100 & 0.3861 \\ -6.9139 & 0.0969 & 1.8214 & 0.4274 & 0.2169 \\ -8.6734 & 0.8100 & 0.4274 & 1.3136 & 0.3980 \\ -3.6989 & 0.3861 & 0.2169 & 0.3980 & 0.4619 \end{pmatrix} \end{array}$$

7.6.5 Offline learning model with the indirect method

In addition to the direct method, we introduce a novel method which is called the indirect method. In this case we relate ordinal variables to the latent variables using ordinal logistic regression. To specify this model (see Appendix A.7.11), suppose we have the variable X which is an ordinal variable with K categories, say $K = 4$ with $i = 1, ..., n$. Then we have probabilities that relate to each category as follows

$$\begin{aligned} p_{i,1} &= 1 - q_{i,1} \\ p_{i,2} &= q_{i,1} - q_{i,2} \\ p_{i,3} &= q_{i,2} - q_{i,3} \\ p_{i,4} &= q_{i,3}. \end{aligned}$$

This ensures that the sum of these probabilities is 1. We can represent q as

$$\text{logit}(q_{i,r}) = Z_i - c_r.$$

where the cut-points are $c_r = (c_1, c_2, c_3)$ since we have four categories.

We can deal with different sorts of covariates in this model exactly as we do with the direct method. So, for example, in the case of a binary variable, since the binary distribution has one parameter, we fix the variance of Z to be 1 and the cut-point to be 0. To include an ordinal variable with $K = 4$ in this model, we should have three cut-points

$(0, c_2, 1)$ where $c_2 \sim \text{Beta}(a, b)$.

In the non-Hodgkin lymphoma example, after obtaining the posterior distribution for all the parameters from the historical data, using a generalised autoregression, we then convert the parameters to the form of $\text{E}_0(Z)$ and $V_0(Z)$. As a result, our prior means and variances for use in the BLK network can be represented as

$$\mu_0 = (126.5844\ ,\ 8.0292\ ,\ 0.1596\ ,\ -0.8988\ ,\ 0.6034)'$$
$$\mu_{\text{age}} = (-0.1766\ ,\ 0.0087\ ,\ 0.0182\ ,\ 0.01910\ ,\ 0.0134)'$$

and

$$\mu_{\text{sex}} = (-4.9086\ ,\ -0.0080\ ,\ 0.0789\ ,\ 0.0157\ ,\ -0.0846)'$$

Therefore,

$$\text{E}_0(Z_i) = \mu_0 + \mu_{\text{age}} x_{age,i} + \mu_{\text{sex}} x_{sex,i},$$

and

$$V_0(Z) = \begin{bmatrix} 323.2659 & -2.1548 & -10.5563 & -8.9206 & -4.3637 \\ -2.1548 & 12.1275 & 0.1440 & 0.8379 & 0.4687 \\ -10.5563 & 0.1440 & 1.3452 & 0.6220 & 0.7403 \\ -8.9206 & 0.8379 & 0.6220 & 1.4024 & 0.4692 \\ -4.3637 & 0.4687 & 0.7403 & 0.4692 & 0.7790 \end{bmatrix}.$$

7.6.6 Offline learning: Diagnostic checking in the direct method

7.6.6.1 Introduction

The non-Hodgkin lymphoma (NHL) example concerns the construction and use of a Bayes linear kinematic prognostic index calculator. There are two phases: the offline-learning stage, which is a full-Bayes analysis with MCMC computations, and the Bayes linear kinematic calculation of prognostic index values for new patients, based on a Bayes linear Bayes network. The appropriate place to calculate and examine residuals in this case is in the offline-learning stage as this is where the model is developed. Some covariate values in this example are missing.

Since we use MCMC to do the model-fitting calculations, it is possible to calculate residuals within a `rjags` run. It is simpler to compute the residuals within `rjags`, especially as the lifetime distribution is Weibull and so the cdf has a simple form. At each MCMC iteration the parameters are sampled and so are the lifetimes for censored observations. We thus obtain samples from the posterior distribution of the residuals as discussed in Section 7.3.7.1. In particular for a censored observation, we obtain samples from the posterior predictive distribution of $F_i(T_i)$. We could use a separate `R` program to deal with the missing covariate values, but there are some difficulties in doing that. On the other hand, since there will be a large number of residuals, sampling the residuals directly within `rjags` might cause difficulties with storage and with processing the results. In fact, no such difficulties were encountered in this example. If they were, we could use thinning to reduce the number of stored samples.

7.6.6.2 Results

In this section, we produce some residual plots in the non-Hodgkin lymphoma example in order to support the selection of our model. In Figure 7.14, we notice that the distribution in the histogram of these residuals is approximately uniform as required. In addition, to check the model assumptions, we plot the residuals against the covariates in the example. For instance, we plot Age against the residuals in Figure 7.15. It is clearly a "random scatter" of points and there is no evidence of any pattern here. So, our assumption which stated that the residuals are distributed with constant variance is plausible for both male and female.

In Figure 7.16, we plot log(WBC) against the residuals. We can see that, there is no concern about the relationship between log(WBC) and residuals. The points are randomly scattered in this graph.

Figure 7.17 shows the scatter plot of HB and residuals. Again we do two plots, one for males and one for females. Also, we can see random scatter with no evidence of changes in the variance.

In Figure 7.18, we plot the Age against residuals but this time we ignore Sex. We do a separate plot for each value of Albumin. Again, everything appears to be in order. There is a random scatter of points. We also notice that there are more observations for the value of Albumin 1 than the value of Albumin 2.

Figure 7.19 shows the scatter plots for residuals against Age. As we have 4 stages in

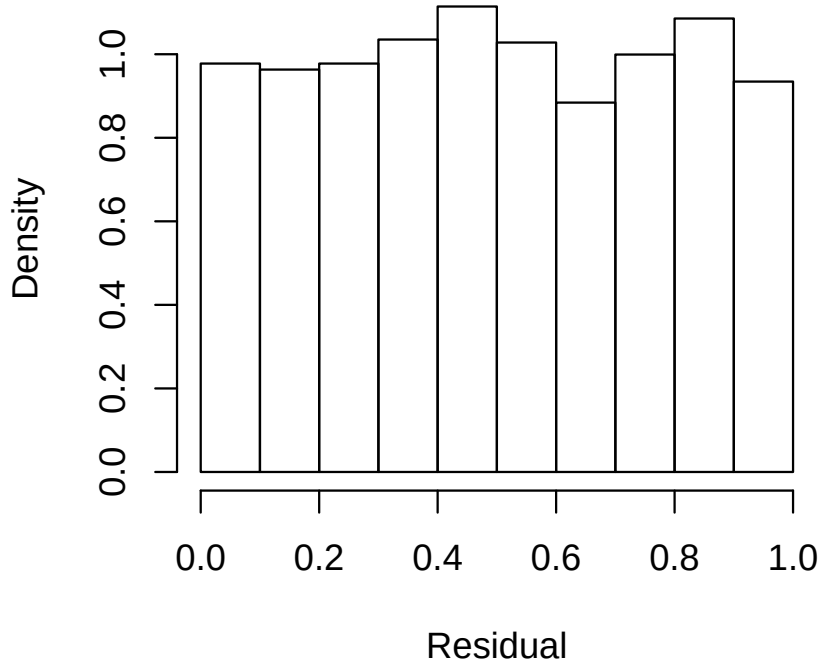


Figure 7.14: Histogram of the posterior means of the residuals.

the variable Stage, we have 4 plots, one for each stage. These graphs raise no concerns about the variability in the variance, which means we have a constant variance in all the stages.

In Figure 7.20, we plot the posterior mean for η using two methods, full Bayes and Bayes linear kinematics, against the residuals and for male and female using the direct method. Again, we have scatter random observations for both methods and for male and female. There is no clear pattern in these plots. As a result, there is no evidence against the validity of our assumptions.

7.6.7 Diagnostic checking in the indirect method

Now we examine the residuals in the case of the indirect model.

Again, the histogram in Figure 7.21 shows approximately a uniform distribution. We

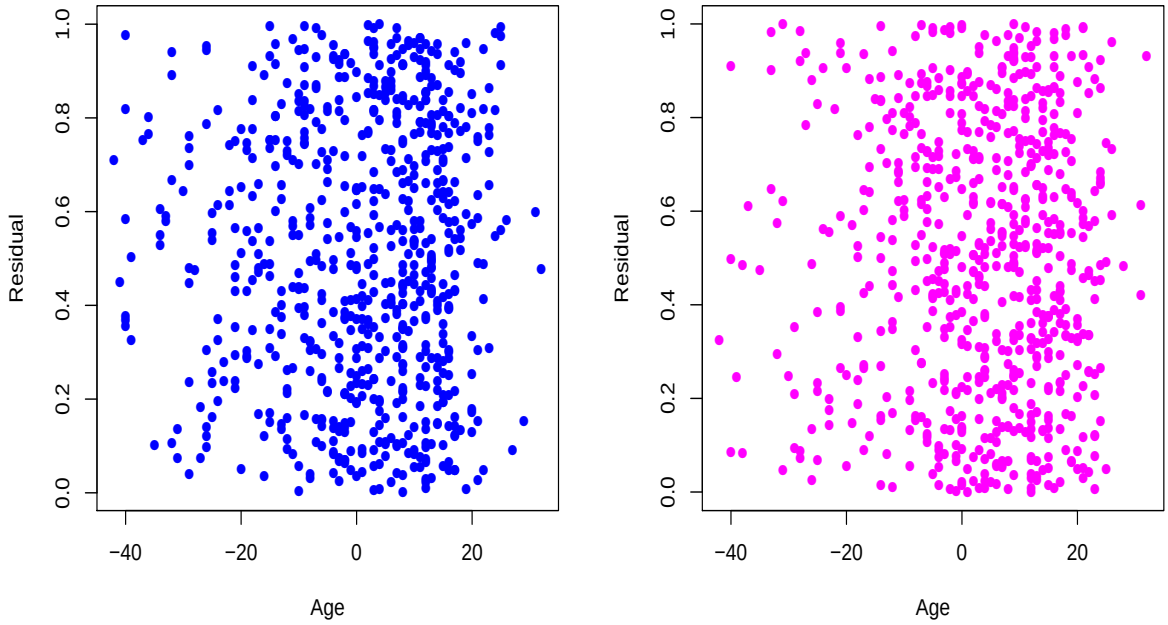


Figure 7.15: Scatter plots for Age against residuals for both sexes. The blue dots for male and the pink ones for female.

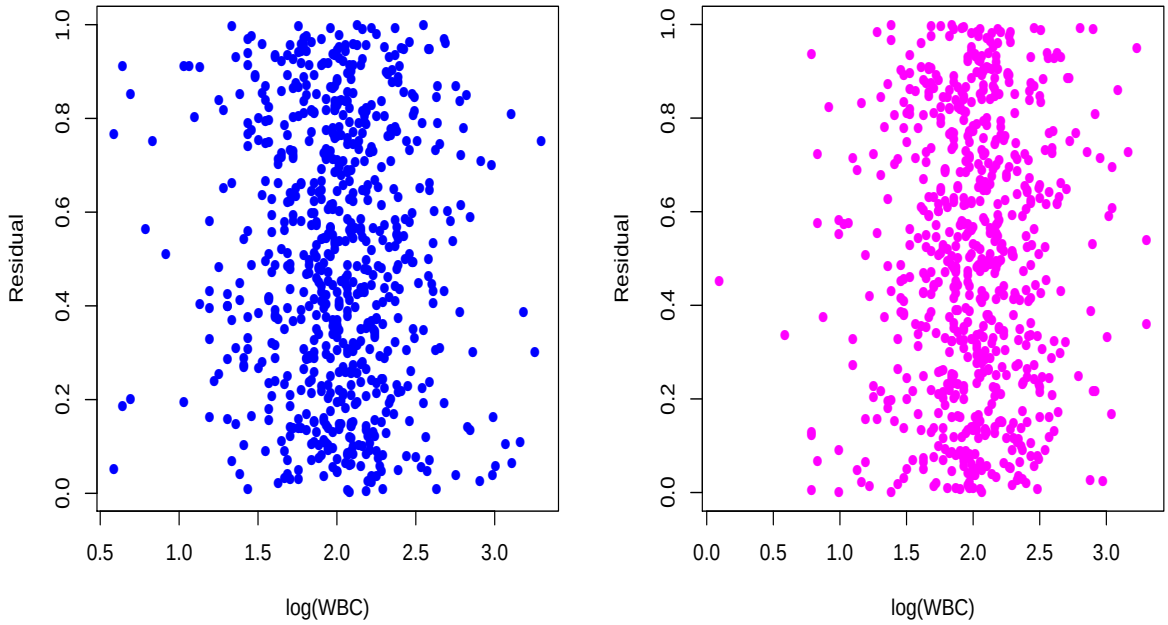


Figure 7.16: Scatter plots for log(WBC) against residuals for both sexes. The blue dots for male and the pink ones for female.

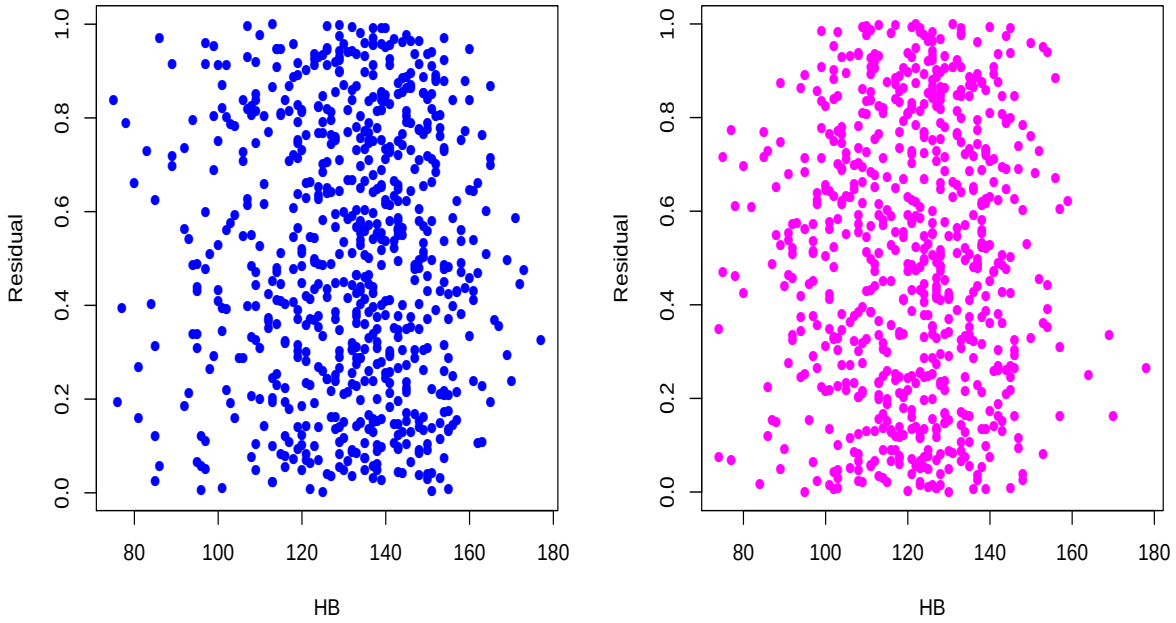


Figure 7.17: Scatter plots for HB against residuals for both sexes. The blue dots for male and the pink ones for female.

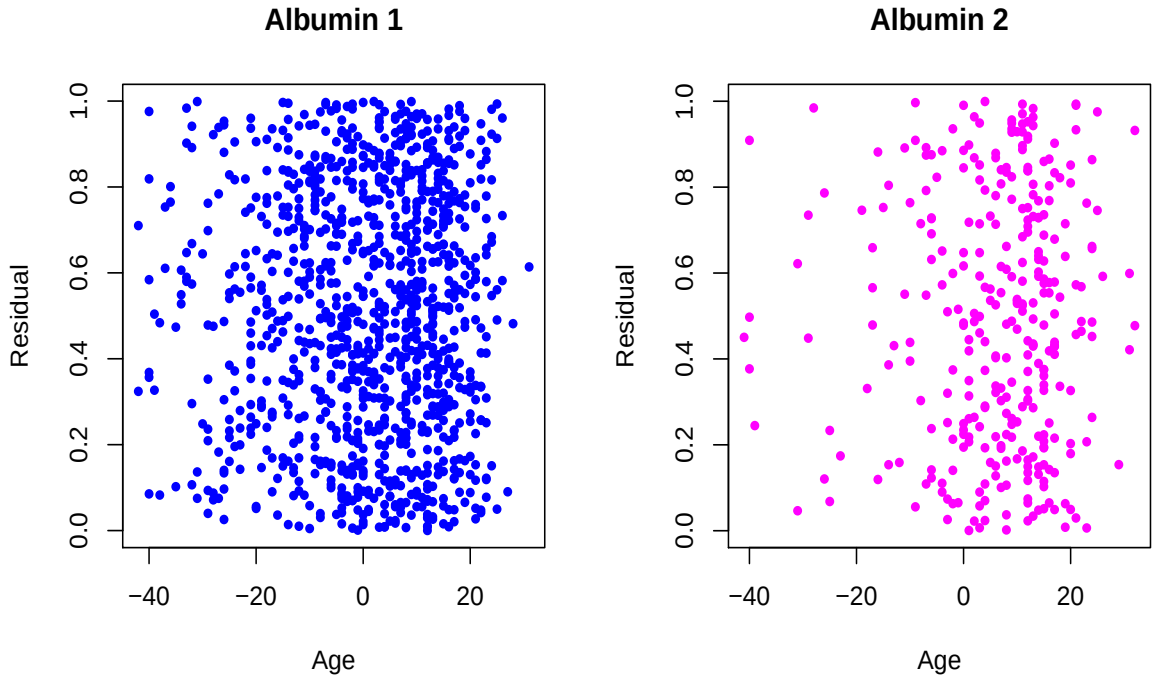


Figure 7.18: Scatter plots for Age against residuals for Albumin 1 and Albumin 2. The blue dots for Albumin 1 and the pink ones for Albumin 2.

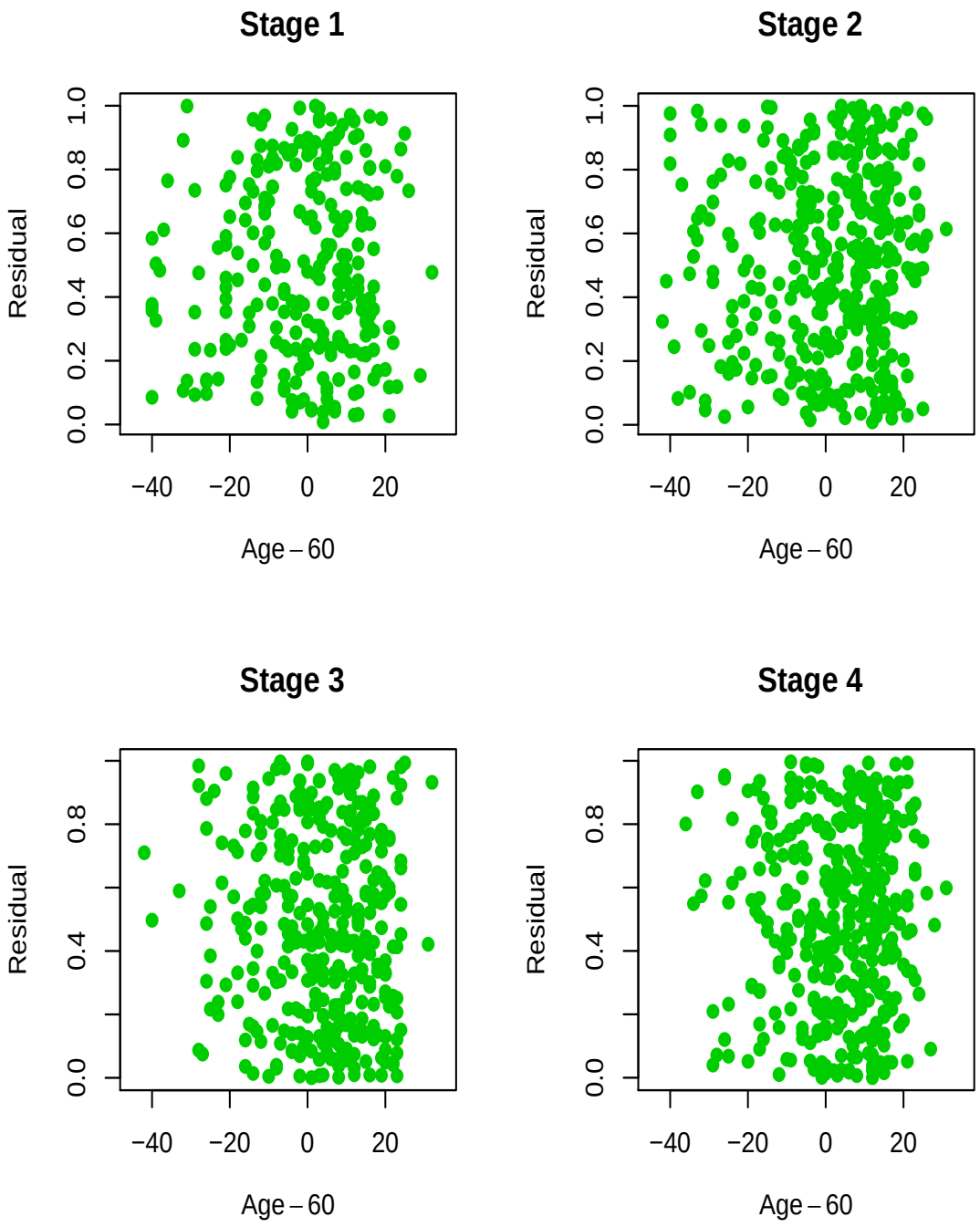


Figure 7.19: Scatter plots for Age against residuals for 4 stages in the covariate Stage.

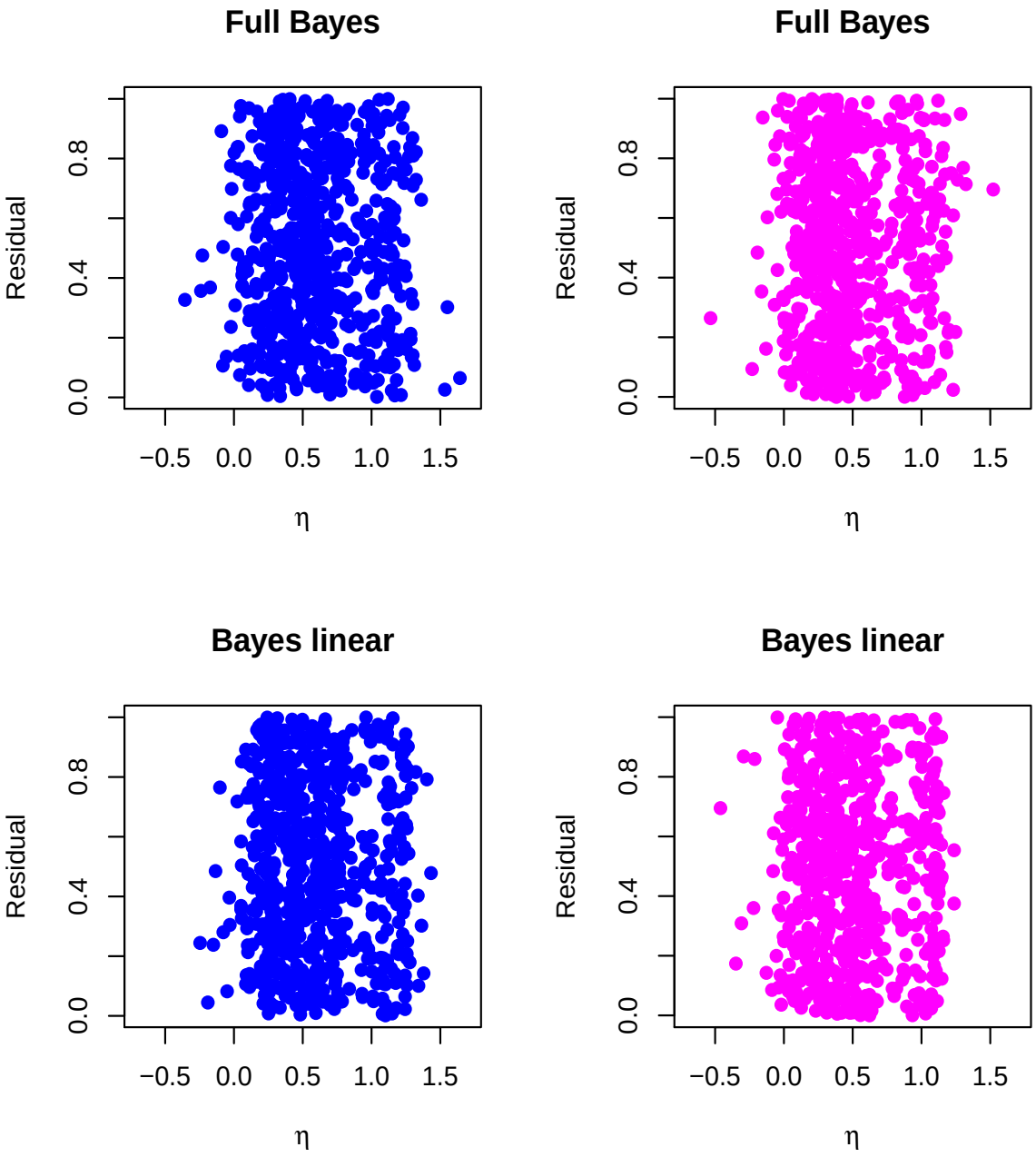


Figure 7.20: Scatter plots for posterior mean of η using full Bayes and Bayes linear kinematic against residuals for both sexes. The blue dots for male and the pink ones for female.

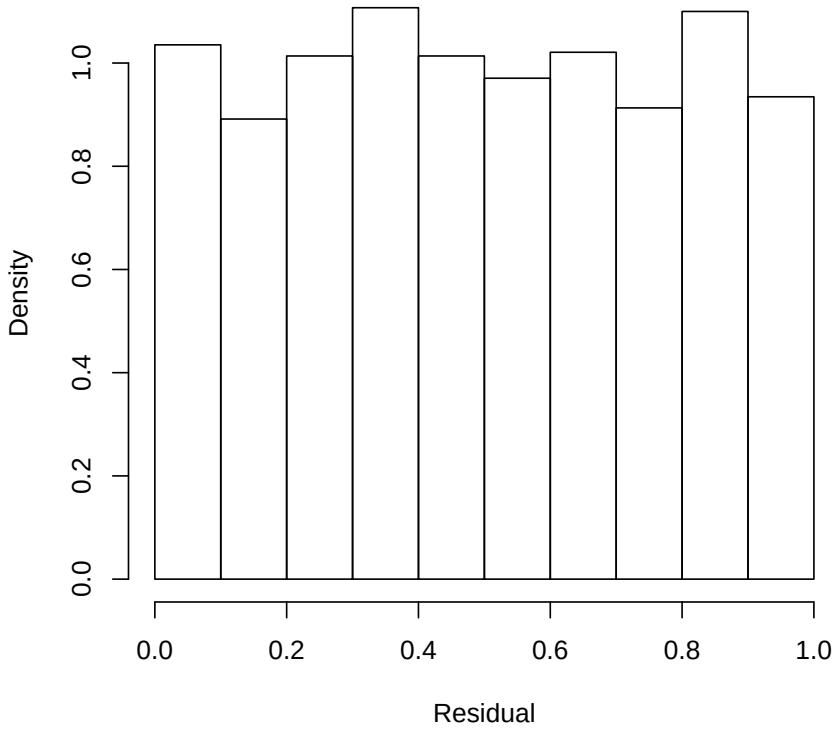


Figure 7.21: Histogram of the posterior means of the residuals in NHL example using the indirect method.

also notice that all our Figures 7.22, 7.23, 7.24, 7.25, 7.26 and 7.27 are very similar graphs to those produced for the direct method. In conclusion, there is no concern about the validity of the model assumptions.

7.6.8 Prognostic index: Comparison with full Bayes analysis

We calculated the BLK prognostic index values for all of the patients in the dataset using the direct method with parameter values obtained from the offline learning.

For comparison with the BLK prognostic index values, we used MCMC to calculate "full Bayes" values. The missing-data ability can be achieved in a full-Bayes model by modelling the joint distribution of all of the variables, rather than just the conditional distribution of the lifetime given the covariates. To make the results comparable and

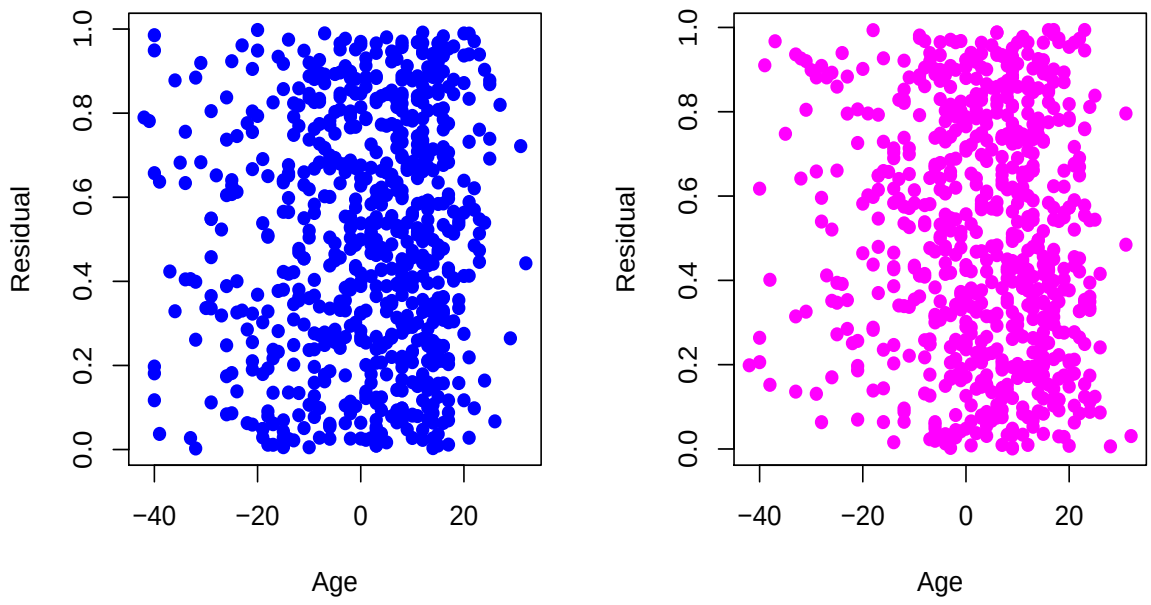


Figure 7.22: Scatter plots for Age against residuals for both sexes in the indirect method. The blue dots for male and the pink ones for female.

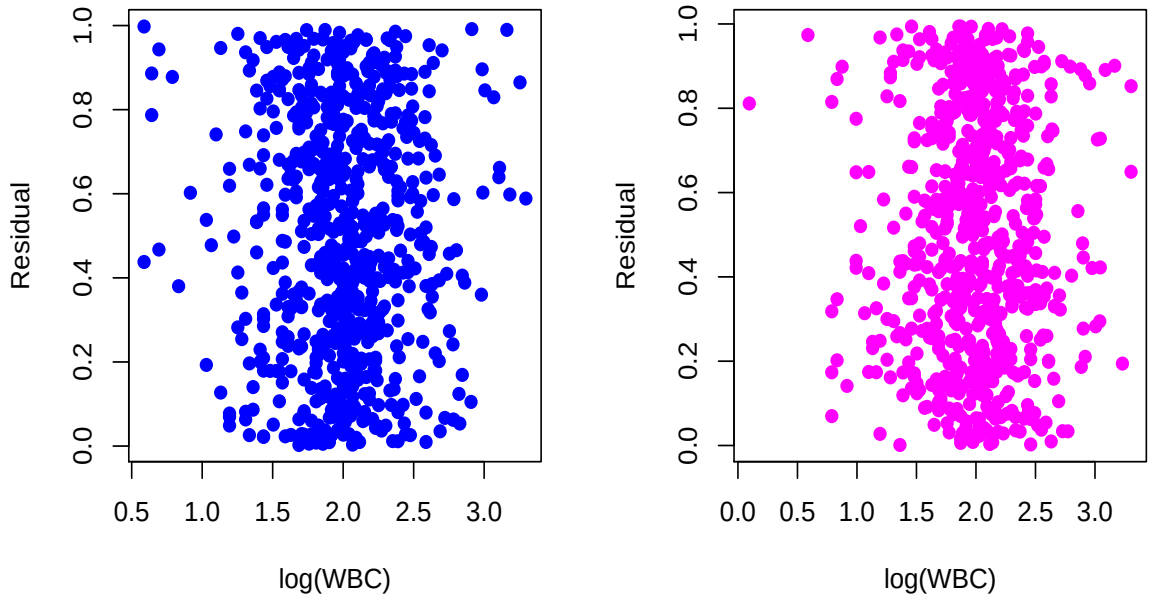


Figure 7.23: Scatter plots for log(WBC) against residuals for both sexes in the indirect method. The blue dots for male and the pink ones for female.

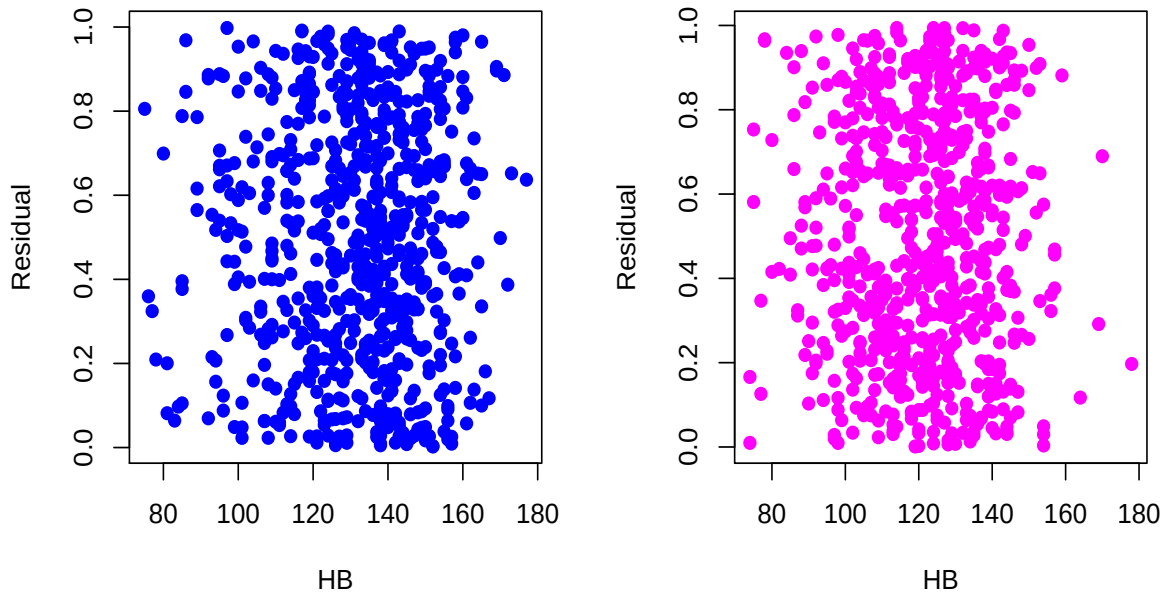


Figure 7.24: Scatter plots for HB against residuals for both sexes in the indirect method. The blue dots for male and the pink ones for female.

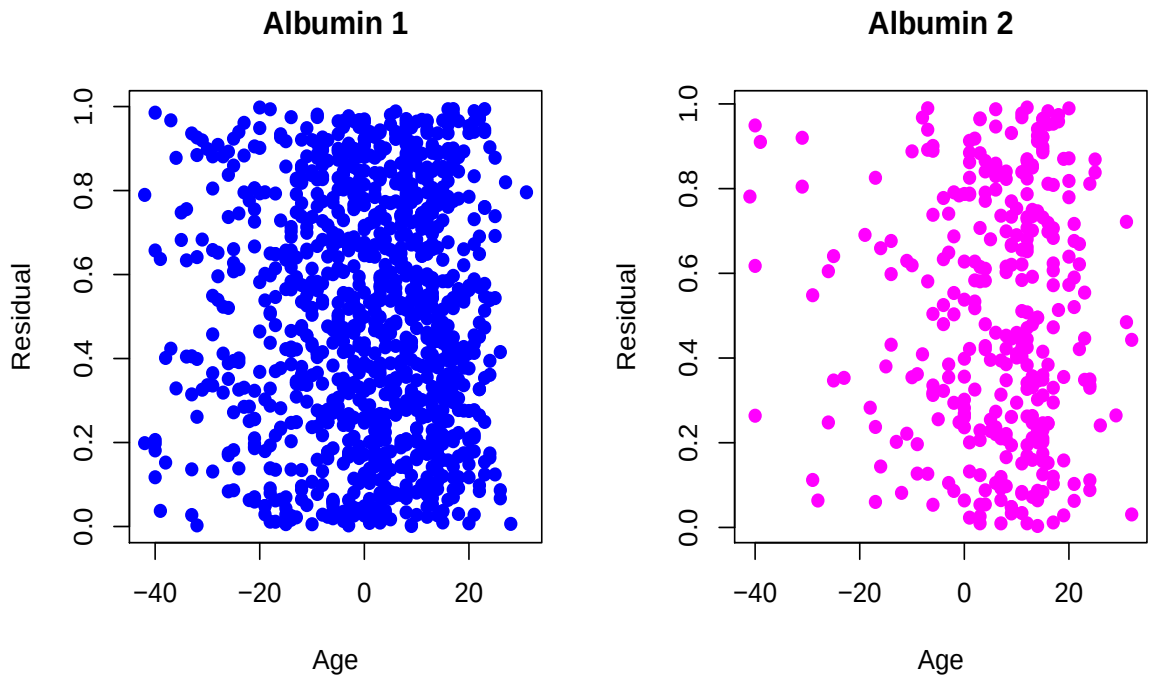


Figure 7.25: Scatter plots for Age against residuals for Albumin 1 and Albumin 2 in the indirect method. The blue dots for Albumin 1 and the pink ones for Albumin 2.

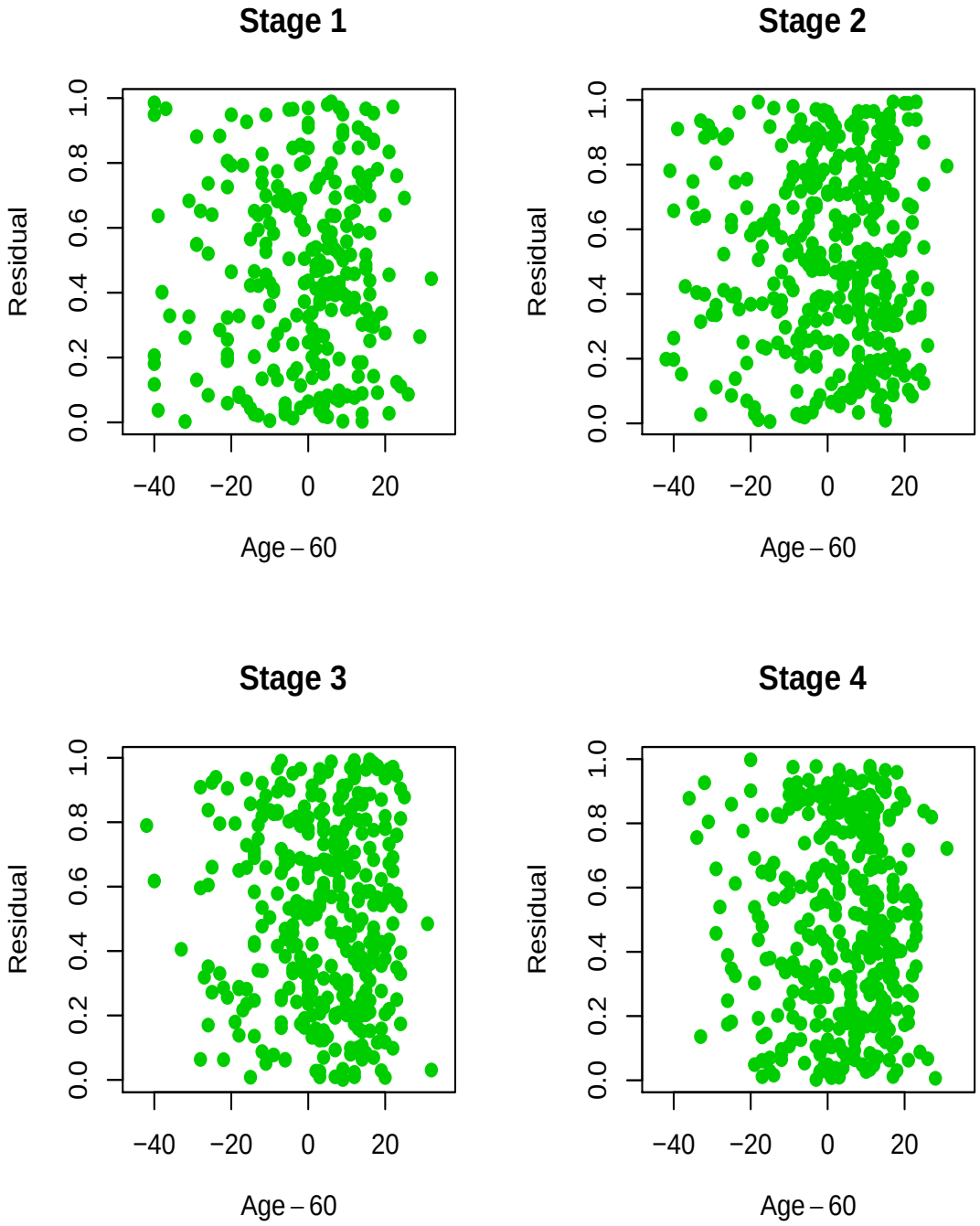


Figure 7.26: Scatter plots for Age against residuals for 4 stages in the covariate Stage using the indirect method.

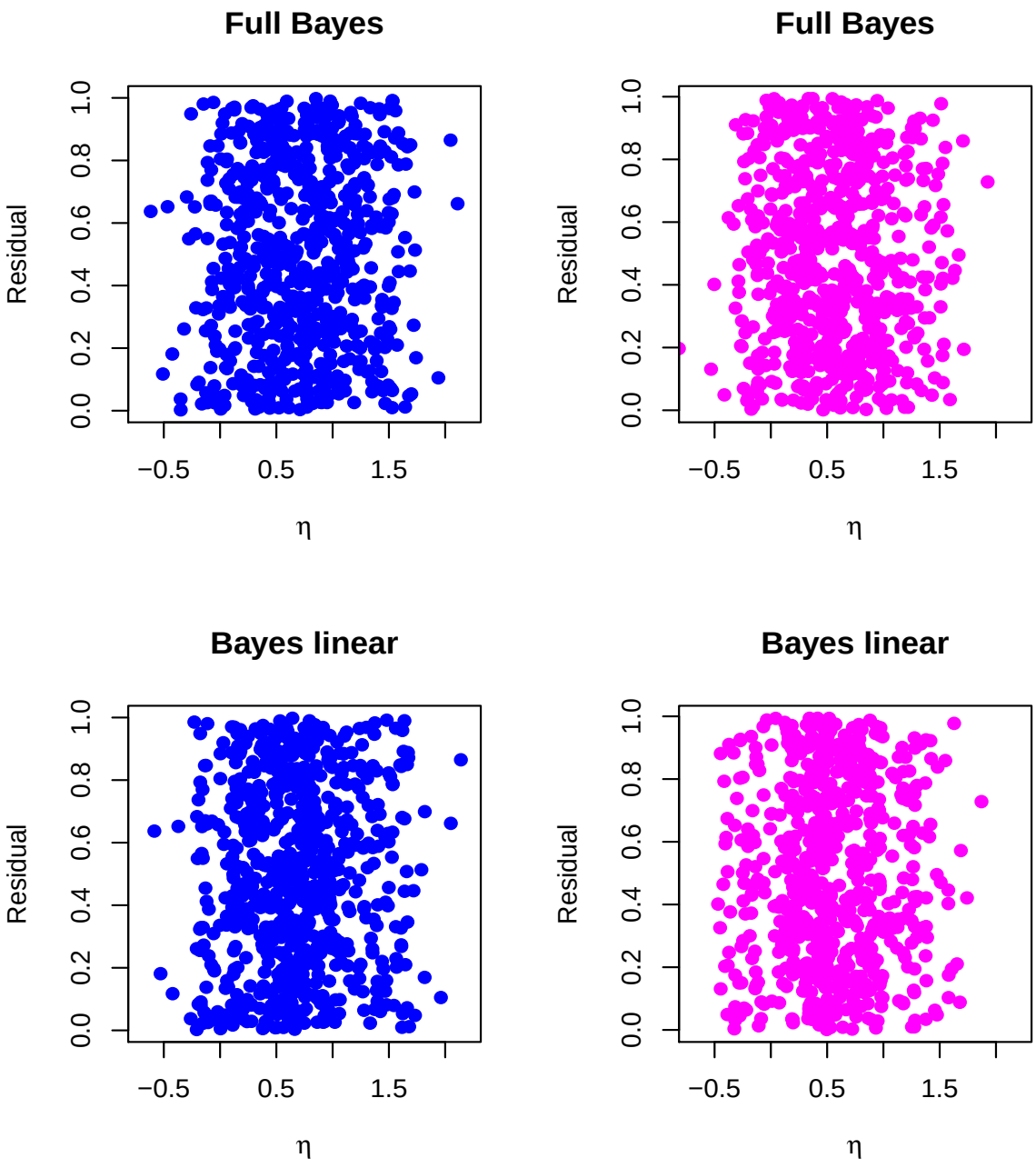


Figure 7.27: Scatter plots for posterior mean of η using full Bayes and Bayes linear kinematic against residuals for both sexes in the indirect method. The blue dots for male and the pink ones for female.

represent routine use in practice, we fixed parameters in the "full Bayes" calculation at the values obtained from the offline learning.

Figure 7.28 shows histograms of the "full Bayes" and BLK prognostic index values for all the patients using the direct method. We notice from Figure 7.28 that both histograms for the prognostic index values using MCMC and BLK are almost the same. The mean "full Bayes" value is 0.5352, with standard deviation 0.3348 and the mean prognostic index value using BLK is 0.5345 with standard deviation 0.3345. Now Figure 7.29 shows that our adjusted means for the prognostic values are close to the posterior mean from the full-Bayes analysis as we can see the straight line of equality is passing through the points which means our BLK method fit the straight line very well. We also use another graph to compare the two methods. This graph is called a "Bland and Altman plot", or agreement plot. This shows the agreement between two methods. Bland and Altman (1986) described such plots which can be done by calculating the mean difference between the two methods and the standard deviations for the differences. In other words, we have, for n cases

$$m_d = \frac{1}{n}\sum_{i=1}^{n}(\hat{Z}_{BLK,i} - \hat{Z}_{MCMC,i}).$$

where $\hat{Z}_{BLK,i}$ is the prognostic index value calculated using the BLK network and $\hat{Z}_{MCMC,i}$ is the "full-Bayes" value for patient i. The lower and the upper limit are $m_d \pm 2S_d$, where S_d is the sample standard deviation of the differences. The sample standard deviation is $S_d = 0.167$.

As we can see from Figure 7.30, there are 96% of the data points within the limits. That indicates that our proposed method to construct Bayes linear kinematic network in Figure 7.9 gives a reasonable result.

In Figure 7.31 the "full Bayes" values are again plotted against the BLK values. However, this time cases where a particular covariate is missing are shown in red. There is one plot for each of the covariates (other than Age and Sex). We see that in three cases, the red dots appear to be fairly evenly distributed among the black dots. However, in the case of the covariate Albumin, there is a distinct group of red dots where the BLK index value is smaller than the corresponding "full Bayes" value. It is not clear why, for example, any model misspecification with respect to Albumin would affect the BLK and "full Bayes" results in distinctly different ways. This feature should be the subject of further research.

The purpose of our Bayes linear Bayes prognostic network is the quick and easy routine

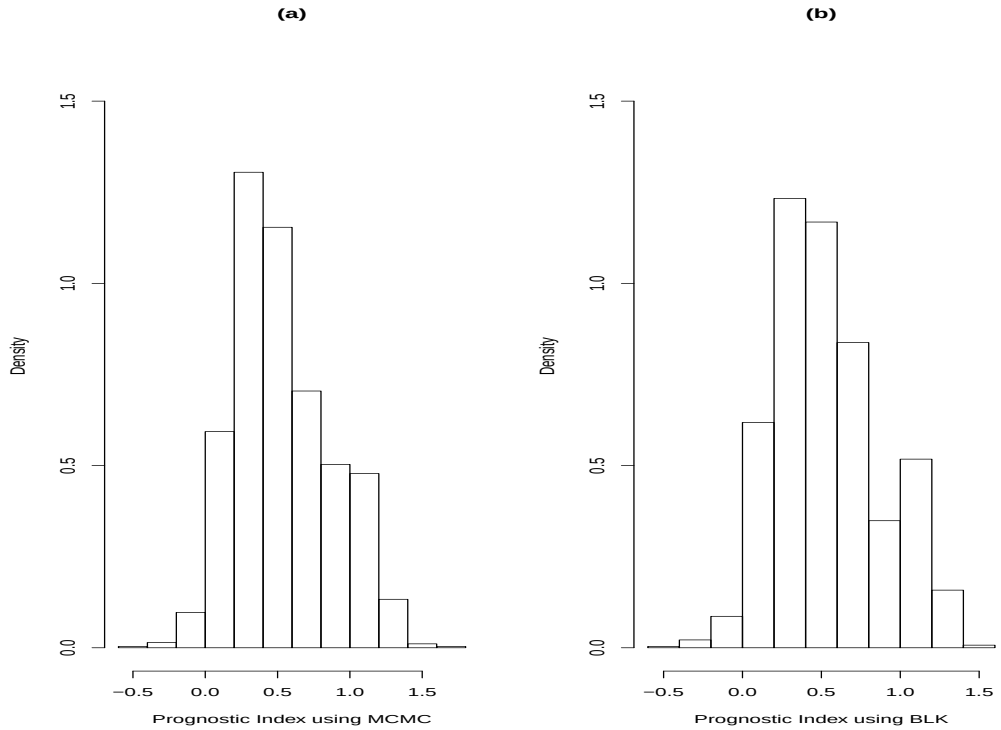


Figure 7.28: Histogram of prognostic index values from MCMC (a), Histogram of prognostic index values from BLK using the direct method (b).

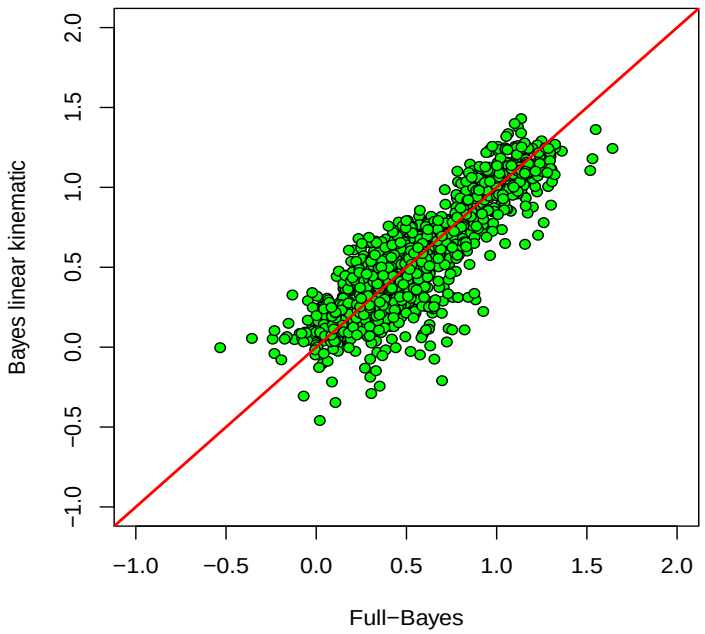


Figure 7.29: Adjusted mean using full-Bayes and BLK in direct method.

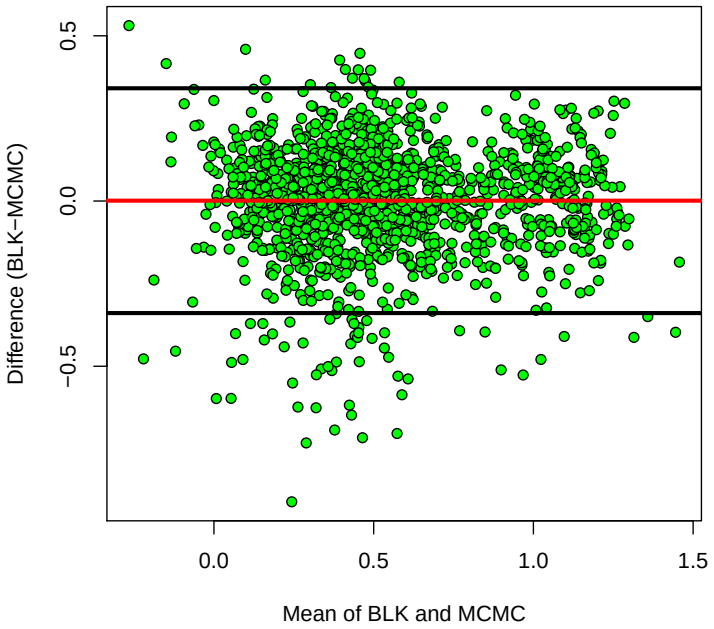


Figure 7.30: Bland and Altman agreement plot for the direct method. The difference $\hat{Z}_{BLK} - \hat{Z}_{MCMC}$ is plotted against the mean $(\hat{Z}_{BLK} + \hat{Z}_{MCMC})/2$ where $\hat{Z}_{BLK}$ and $\hat{Z}_{MCMC}$ are the BLK and full Bayes posterior means of Z_T respectively.

calculation of prognostic index values for new patients, potentially using a large number of covariates but able to work when only some of these are observed. The missing-data ability can be achieved in a full-Bayes model by modelling the joint distribution of all the variables, rather than just the conditional distribution of the lifetime given the covariates. However, in a full-Bayes model, we need to integrate over the joint distribution of the missing covariates, conditional on the observed values, which may be computationally demanding. In our Bayes linear Bayes network, even with non-conjugate marginal updates, we need, at most, a series of one-dimensional integrations which can usually be done very quickly.

Furthermore, a full-Bayes analysis typically requires a lot of decisions to be made about the forms of relationships between variables and these choices may have little basis either in expert judgement or the analysis of historical data. In contrast, the Bayes linear Bayes approach requires a more limited specification of relationships in terms of first and second moments and focussing on these more limited judgements might lead to sounder choices.

Our method might be regarded as an approximation to a full-Bayes analysis. Wilson

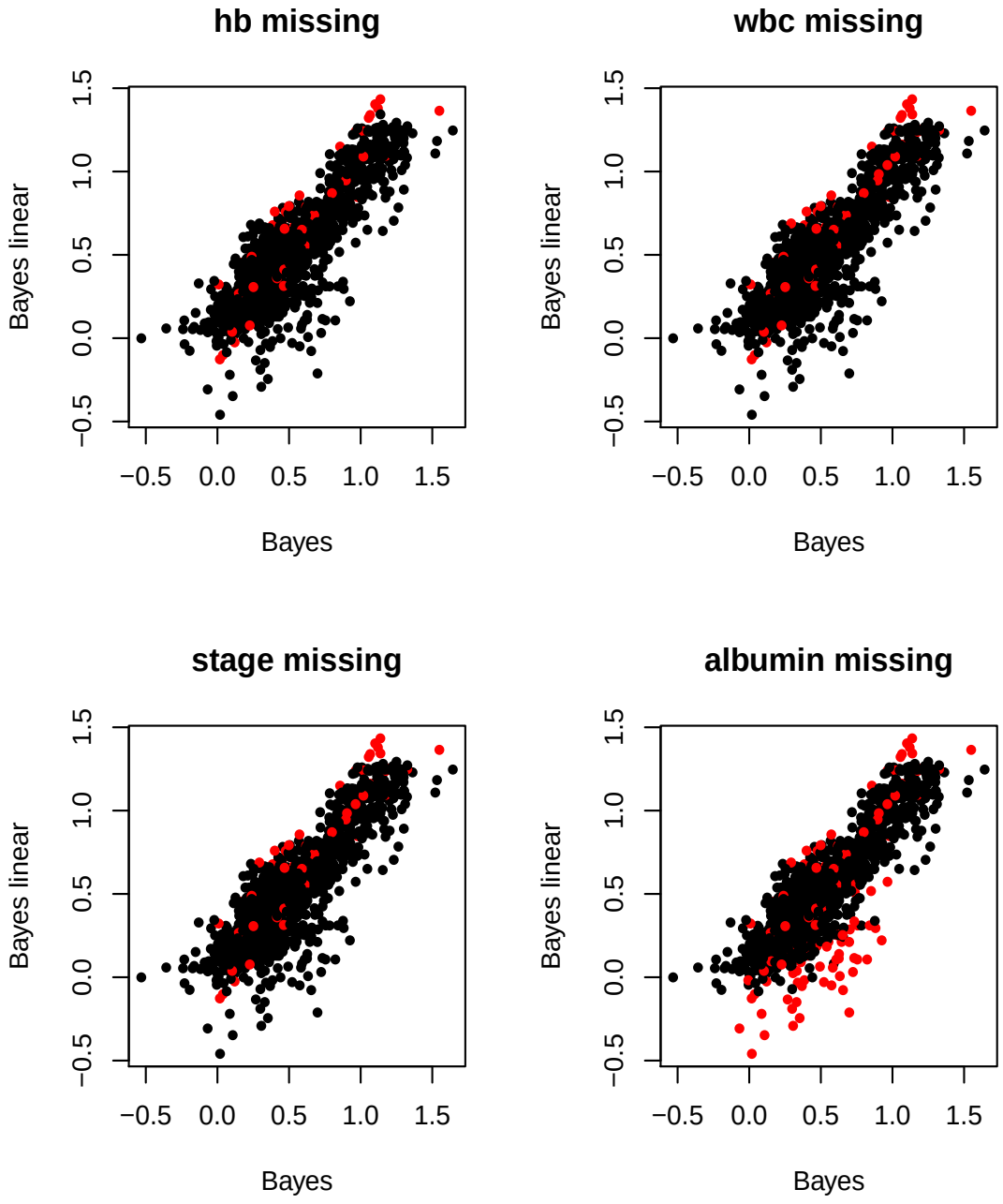


Figure 7.31: Predicted prognostic index values, Bayes linear against full Bayes in the non-Hodgkin lymphoma example using the direct method. In each plot, cases where a particular covariate is missing are shown in red.

and Farrow (2017) compared the behaviour of Bayes linear kinematic belief adjustments with full-Bayes posterior inference in the case of a piecewise constant hazard survival model and found that the results were generally close. Our use of non-conjugate updates allows our model to be closer to the corresponding full-Bayes model and we hope that this will bring Bayes linear kinematic adjusted expectations even closer to full-Bayes posterior means.

7.7 Comparison between the "direct" and the "indirect" methods

We repeated the calculations using the indirect method and the corresponding indirect model in the offline learning.

There are some differences between the direct and the indirect methods here.

In the indirect method, the support of the posterior is unbounded because the likelihood is a function of all the values of the probability and for all the values of Z. As a result, we obtain an unbounded posterior distribution rather than the bounded support for the posterior in the direct method.

In the indirect method, in order to construct a Bayes linear kinematic network, we need to specify three things, $\underline{Z} = (Z_1, ..., Z_n)'$ which represent a Bayes linear structure, $\underline{X} = (X_1, ..., X_n)'$ which represent what we observe and something between them which is the probability P that depends deterministically on $\underline{Z}$. For instance, in the case of an ordinal variable, P is a vector of probabilities and those probabilities depend on the values of Z. Notice that X is not determined by Z, but the probabilities of X depend on Z. So, if we observe x, then the likelihood we obtain is the probability for the category that X is in, and that probability is a function for all the values of Z.

Figure 7.32 shows a comparison between the full Bayes and Bayes linear kinematic posterior means using the indirect method. We see that almost all of the observations lie close to the line of equality which indicates that using the indirect method give us values very close to these given by the full-Bayes method. Figure 7.33 shows that the histograms of full-Bayes and BLK values look normally distributed which is slightly different from the BLK in the direct method. In addition, the agreement plot in Figure 7.34 shows no clear pattern of the trends of these observations.

As for the direct method, in Figure 7.35 the "full Bayes" values are again plotted against the BLK values of the indirect method. However, this time cases where a particular covariate is missing are shown in red. There is one plot for each of the covariates (other than Age and Sex). We see that, in all the cases, the red dots appear to be fairly evenly distributed among the black dots.

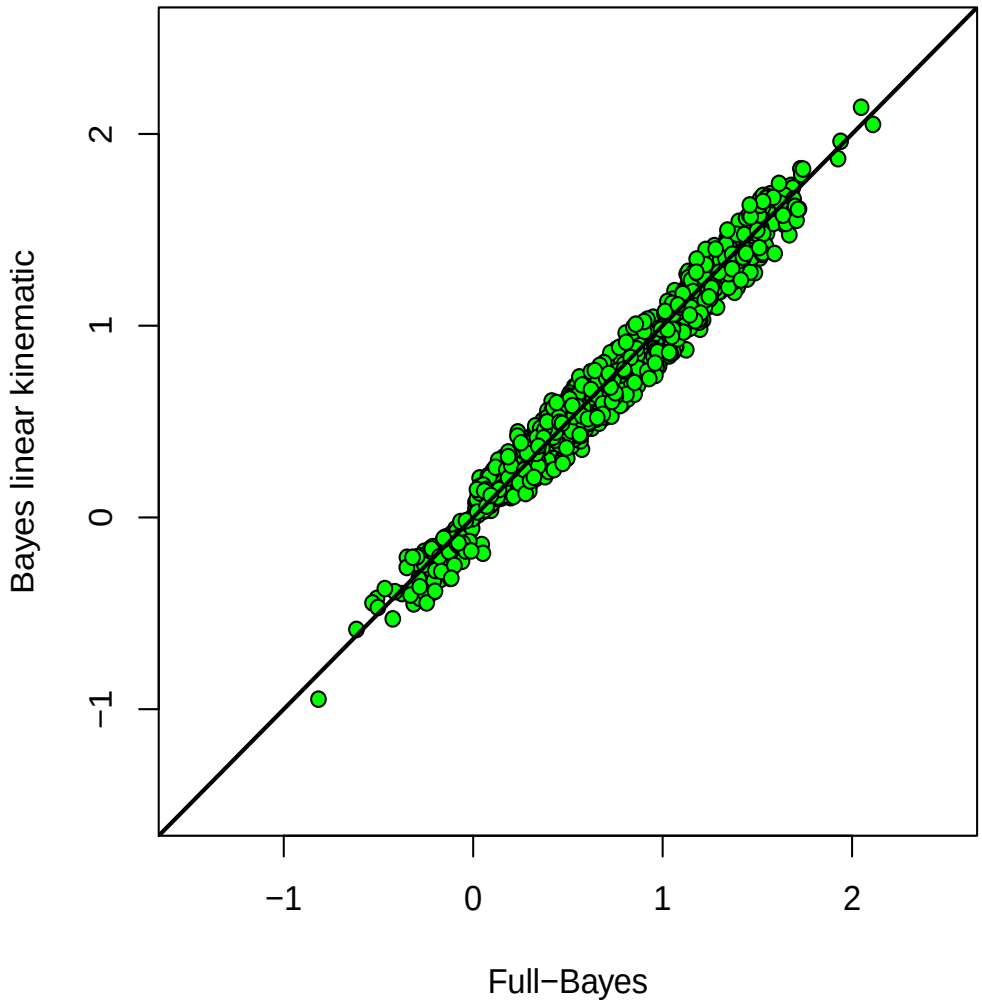


Figure 7.32: Adjusted mean using full-Bayes and BLK in the indirect method.

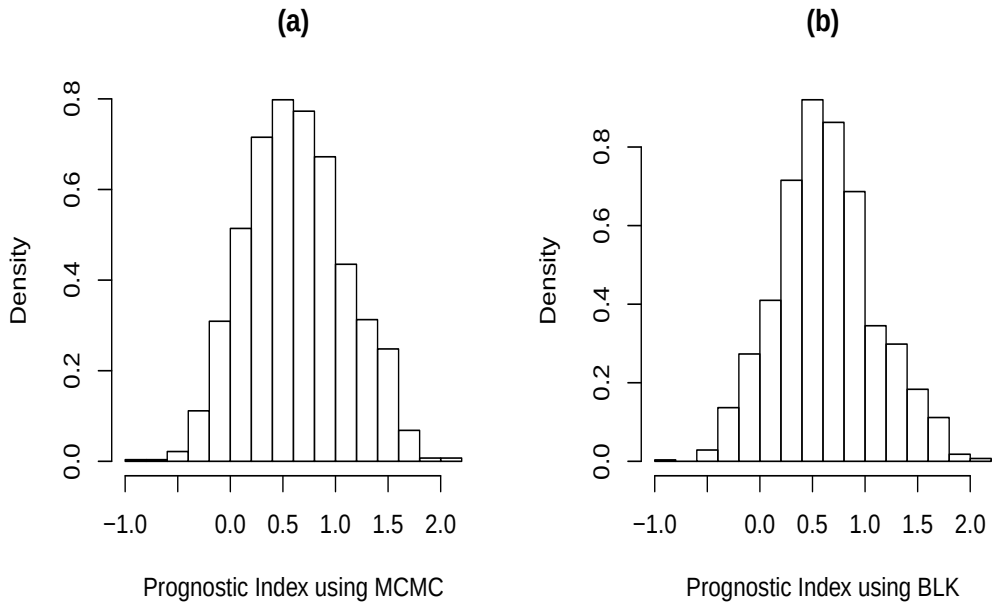


Figure 7.33: Histogram of prognostic index values from MCMC (a), Histogram of prognostic index values from BLK using the indirect method (b).

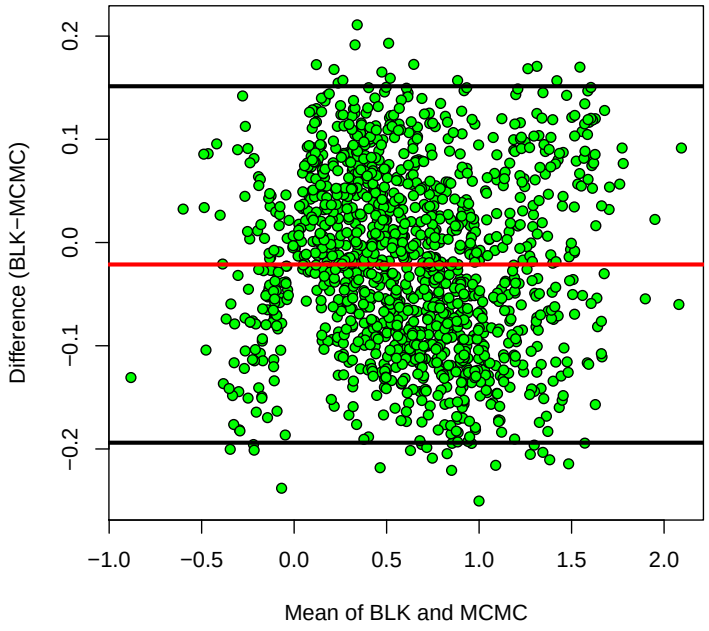


Figure 7.34: Bland and Altman agreement plot for the indirect method. The difference $\hat{Z}_{BLK} - \hat{Z}_{MCMC}$ is plotted against the mean $(\hat{Z}_{BLK} + \hat{Z}_{MCMC})/2$ where $\hat{Z}_{BLK}$ and $\hat{Z}_{MCMC}$ are the BLK and full Bayes posterior means of Z_T respectively.

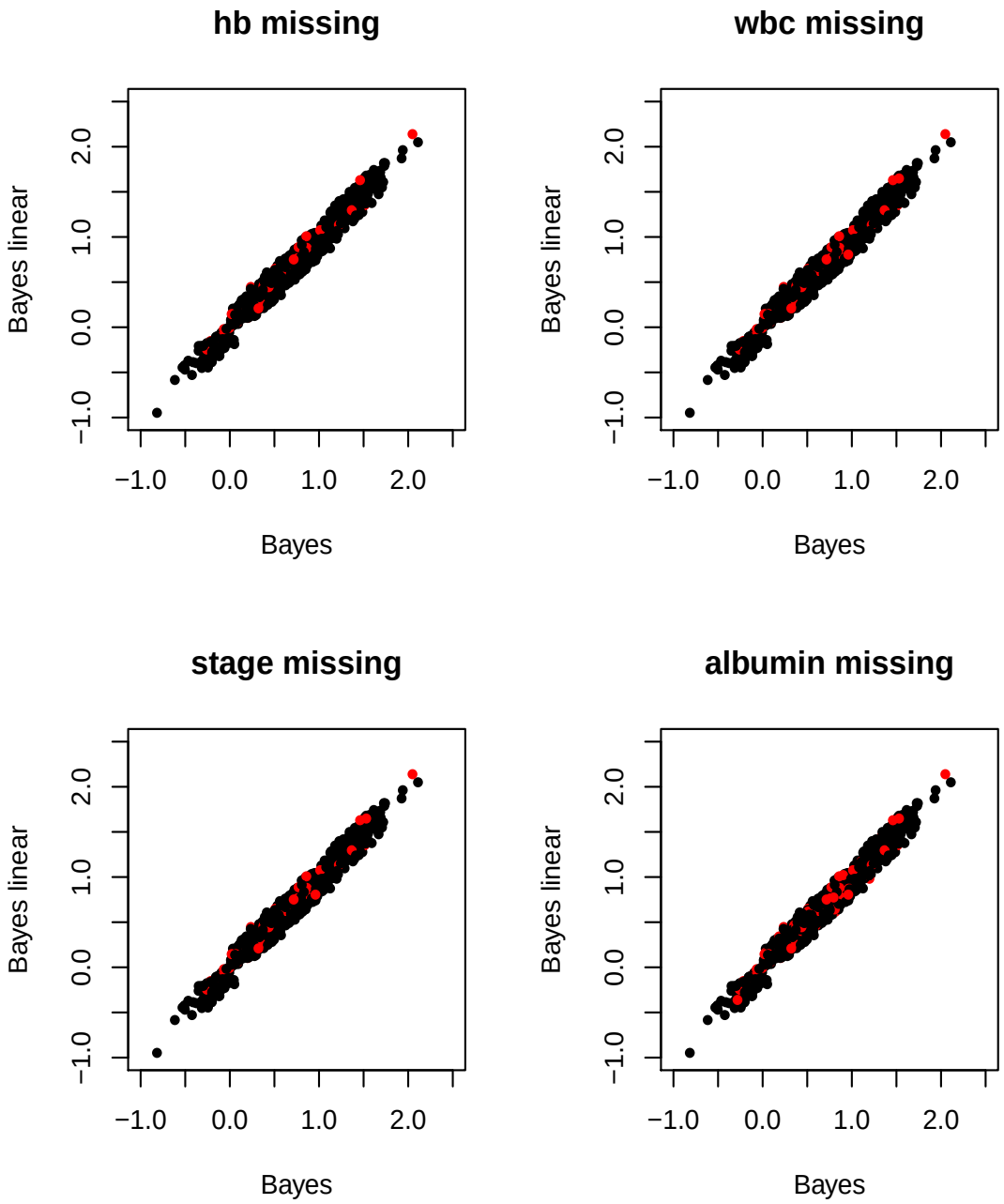


Figure 7.35: Predicted prognostic index values, Bayes linear against full Bayes in the non-Hodgkin lymphoma example using the indirect method. In each plot, cases where a particular covariate is missing are shown in red.

7.8 Prototype prognostic index calculator

In this section, we consider transformation of the adjusted means for the prognostic index values that we calculate from BLK. We wrote a R function which prompts the user to type in the information about the covariates and then gives the user the prognostic value for that particular patient. This can help doctors to know about the current situation of patients. Let the adjusted expectation of Z_T for patient i be $\hat{Z}_{T,i}$. We use transformed index values in the range of (0,100) by using the percentiles of a normal distribution fitted to the values of $\hat{Z}_{T,i}$ for all patients in the data base. That can be done by computing

$$100 \times \Phi\left(\frac{\hat{Z}_{T,i} - m}{S}\right),$$

where $\Phi()$ is the standard normal cumulative distribution function and m and S are the sample mean and sample standard deviation of the values of $\hat{Z}_T$ for patients in the data base. So, for example, if a patient has an index value of 80, this means that this patient has a high risk value. A R function is shown in Appendix A.7.10.

7.9 Summary

In this chapter, we have explained in detail two examples concerning leukaemia and non-Hodgkin lymphoma. For the leukaemia example, we have reviewed the Wilson and Farrow (2017) approach and extended that to our non-conjugate method and we compared our method with three different methods.

We also have reviewed some aspects that relate to Bayes linear kinematics and Bayes linear Bayes graphical models and then described the application of these ideas to the routine calculation of prognostic index values in medical survival.

Initially, we have explained the general strategy of a novel method to construct a Bayes linear Bayes prognostic network. In particular we introduced the idea of using a latent prognostic index and presenting its expectation as a value to be used. We also made some comparisons between full-Bayes and Bayes linear kinematics using the direct method and between the direct and indirect methods. We found that the prognostic index values from MCMC (full-Bayes) and BLK look similar. Our prototype prognostic network produces prognostic index values using all, or some, of the possible covariates almost instantly and

has the potential to be used, for example, as a Web-based calculator.

The general procedure has two phases. The first is to do some inference using past data in an offline learning model. That can be done using MCMC and we obtain the posterior distributions for all the parameters in the model. Secondly, we use these posterior distributions to help us construct a Bayes linear kinematic network. These parameter values are treated as fixed in the BLK network. In practice we might re-run the offline learning from time to time, as new data become available, to obtain new parameter values. This might also be necessary, for example, if a new treatment is introduced.

In conclusion, when using existing methods in more complicated networks, we might have to use computationally intensive methods or some kind of approximations. However, our proposed methods can do fast computations in networks which are not Gaussian, even when not all covariate values are observed.

Chapter 8

Simulation experiment in survival analysis

8.1 Introduction

In this chapter, we investigate further the Bayes linear Bayes methods. We use some simulation experiments to examine the behaviour of the direct and indirect methods. In the following section, we will give more explanation about the way that we generated our simulated data.

The main idea of this chapter is to predict the prognostic index Z_T using three methods, BLK using the direct method, BLK using the indirect method and full-Bayes analysis and to compare the results.

In Section 8.2, we will describe simulation experiments in which the data are generated according to a model corresponding to the direct method. The `R` code for generating the simulated data is given in Appendix A.8.1. These simulated data are then used to predict the prognostic index using both the direct and indirect Bayes linear Bayes methods and also the full-Bayes method assuming both a direct model and an indirect model. The results are then compared. The `R` code for computing the predictions using BLK is given in Appendix A.8.3 for the direct method and Appendix A.8.5 for the indirect method. In the "full Bayes" method used for these comparisons, the population model parameters are treated as known, as in the BLK methods but, unlike the BLK methods, a fully probabilistic prediction, with a multivariate normal distribution for $\underline{Z}$, is computed, using MCMC, with `rjags`. The `rjags` code is given in Appendix A.8.2 for the direct method

and Appendix A.8.4 for the indirect method.

In Section 8.3, the experiments are repeated except that, this time, the data are generated according to a model corresponding to the indirect method. In Section 8.4, conclusions are drawn.

8.2 Data simulated according to the direct model

8.2.1 Simulation method

The simulated data sets are based on the non-Hodgkin lymphoma example. We used the same covariates as in Chapter 7, that is Age, Sex, Hb, Wbc, Stage and Albumin. We have chosen the non-Hodgkin lymphoma as an example to apply our methods because it contains an example of a continuous covariate, a binary covariate and an ordinal covariate.

To do the simulation, we generate the values of $\underline{Z} = (Z_{Hb}, Z_{Wbc}, Z_{Stage}, Z_{Albumin}, Z_T)'$ randomly from a multivariate normal distribution over $\underline{Z}$. Then, given these $\underline{Z}$, we generate randomly all the covariate values X. In other words, we draw samples from the conditional distribution of $X|\underline{Z}$. We use the vector of means $\mathrm{E}_0(\underline{Z})$ and variance-covariance matrix $\mathrm{Var}_0(\underline{Z})$ as we described in Chapter 7, to do the generation. Therefore, the calculations to specify the prior mean vector for $\underline{Z}$, $\mathrm{E}_0(\underline{Z})$ for non-Hodgkin lymphoma will be as follows:

$$\begin{aligned}
\mu_0 &= (126.6473\ ,\ 8.0231\ ,\ 1.2037\ ,\ -0.8868\ ,\ 0.5150)' \\
\mu_{\text{age}} &= (-0.1777\ ,\ 0.0087\ ,\ 0.0121\ ,\ 0.0189\ ,\ 0.0098)' \\
\mu_{\text{sex}} &= (-4.9052\ ,\ -0.0084\ ,\ 0.0500\ ,\ 0.0165\ ,\ -0.0628)'
\end{aligned}$$

Therefore,

$$\mathrm{E}_0(\underline{Z}_i) = \mu_0 + \mu_{\text{age}} x_{age,i} + \mu_{\text{sex}} x_{sex,i}.$$

where $x_{age,i}$ is the age in years of patient i, minus 60 and $x_{sex,i}$ is 1 for a male patient or -1 for a female patient.

The prior variance-covariance matrix for $\underline{Z}$, $\text{Var}_0(\underline{Z})$ is

$$\text{Var}_0(\underline{Z}) = \begin{bmatrix} 323.3027 & -2.1272 & -6.9139 & -8.6734 & -3.6989 \\ -2.1272 & 12.1289 & 0.0969 & 0.8100 & 0.3861 \\ -6.9139 & 0.0969 & 1.8214 & 0.4274 & 0.2169 \\ -8.6734 & 0.8100 & 0.4274 & 1.3136 & 0.3980 \\ -3.6989 & 0.3861 & 0.2169 & 0.3980 & 0.4619 \end{bmatrix}.$$

When we do the simulation in the direct method (see Section 7.6.3), we simulate (randomly) values for $\underline{Z}$. These values can be positive or negative. If we have a binary variable and the value of Z is positive, then $X = 1$ and if Z is negative, then $X = 0$.

In the case of ordinal variables, if the number of categories is K, we have $K - 1$ cut-points. For instance, if we have 4 categories as in the case of the covariate Stage in the non-Hodgkin lymphoma example, then we have three cut-points. In general, if c represents the cut-point, then $X = k$ if and only if $c_{k-1} \leq Z < c_k$ for a set of thresholds $\{c_1, ..., c_{K-1}\}$ where $c_0 \to -\infty$ and $c_K \to \infty$.

We also treated the covariates Hb and Wbc as normal random variables in the direct model. We modelled $Z_{hb} = X_{hb}$ and $Z_{wbc} = X_{wbc}$ in the offline learning model. The actual prediction values of Z_T are generated normally with the prior mean and prior variance that were obtained from offline learning with the real data.

Then we calculated predictions of the prognostic index using both the direct method and the indirect method and also using full Bayes.

We have done three examples in this section. We generate 1200 simulated cases for each of the three examples in this chapter.

In the first example, we are going to do all the simulations in terms of male patients aged 60 to avoid unnecessary extra complications in these simulations.

We also applied our approach to a second example. In this example, we have different ages and sexes. In particular, we have 1/3 of patients with age 50, 1/3 with age 60 and 1/3 with age 70. Therefore, we have 1/6 male aged 50, 1/6 female aged 50, 1/6 male aged 60, 1/6 female aged 60, 1/6 male aged 70 and 1/6 female aged 70. So, we have 1/2 of the patients male and 1/2 female.

In the third example, we artificially increase the variance of Z_T and we use exactly the same age-sex groups from example 2. This has the effect of increasing the range of actual

values somewhat. In order to increase the variance of Z_T, we need to adjust $\mathrm{Var}_0(\underline{Z})$.

Therefore, suppose that we have

$$\mathrm{Var}_0(\underline{Z}) = \left[\begin{array}{c:c} V_c & \underline{c} \\ \hdashline \underline{c}^T & V_{TT} \end{array}\right]$$

where V_c is the variance matrix for the covariate $\underline{Z}_c$, $c = (V_{1T}, V_{2T}, V_{3T}, V_{4T})^T$ and $V_{TT} = \mathrm{Var}(Z_T)$.

Then,

$$\mathrm{Var}(\underline{Z}_c|Z_T) = V_c - \underline{c}\mathrm{Var}(Z_T)^{-1}\underline{c}^T, \tag{8.1}$$

and

$$\mathrm{E}(\underline{Z}_c|Z_T) = \mathrm{E}(\underline{Z}_c) + \underline{c}\mathrm{Var}(Z_T)^{-1}[Z_T - \mathrm{E}(Z_T)]. \tag{8.2}$$

Alternatively, we can write (8.1) and (8.2) as follows

$$\underline{Z}_c = \mathrm{E}(\underline{Z}_c) + \underline{c}\mathrm{Var}(Z_T)^{-1}[Z_T - \mathrm{E}(Z_T)] + U_c$$

where $U_c \sim (\underline{0}, \mathrm{Var}[\underline{Z}_c|Z_T])$.

Now, suppose that the new variance of Z_T is $\mathrm{Var}^*(Z_T)$, then we can write the new variance of Z_c as

$$\begin{aligned} V_c^* &= \underline{c}\mathrm{Var}(Z_T)^{-1}\mathrm{Var}^*(Z_T)\mathrm{Var}(Z_T)^{-1}\underline{c}^T + V_c - \underline{c}\mathrm{Var}(Z_T)^{-1}\underline{c}^T, \\ &= V_c + \underline{c}\underline{c}^T\left\{\frac{\mathrm{Var}^*(Z_T)}{\mathrm{Var}(Z_T)^2} - \frac{1}{\mathrm{Var}(Z_T)}\right\} \\ &= V_c + \underline{c}\underline{c}^T\left\{\frac{\mathrm{Var}^*(Z_T) - \mathrm{Var}(Z_T)}{\mathrm{Var}(Z_T)^2}\right\}. \end{aligned}$$

and the new covariance is

$$\underline{c}^* = \underline{c}\mathrm{Var}(Z_T)^{-1}\mathrm{Var}^*(Z_T).$$

In general, if $\mathrm{Var}^*(Z_T) = b\mathrm{Var}(Z_T)$ then,

$$V_c^* = V_c + \frac{(b-1)}{\mathrm{Var}(Z_T)}\ \underline{c}\underline{c}^T$$

and

$$\underline{c}^* = b\underline{c}.$$

In summary, we used three sets of simulated data as follows. In each case 1200 cases were generated

Example 1: all cases male, aged 60.

Example 2: 200 cases in each of 6 groups: male aged 50, male aged 60, male aged 70, female aged 50, female aged 60, female aged 70.

Example 3: variance of Z_T artificially increased, age-sex groups as in Example 2, b=3.

8.2.2 Results

Figure 8.1 shows scatter plots of predictions, i.e. posterior means, of Z_T given the data, against the true values of Z_T for Example 1. Four methods are used: full Bayes assuming a direct model, full Bayes assuming an indirect model, BLK using the direct method ("BLK direct") and BLK using the indirect method ("BLK indirect").

When the direct method was used for prediction, the parameter values used were those learned by fitting the direct model to the real data.

When the indirect method was used for prediction, the parameter values used were those learned by fitting the indirect model to the real data.

In Figure 8.1, we have the predictions of the prognostic index values for the actual values of Z_T, direct and the indirect method and full-Bayes. These graphs show how successful the methods are at predicting the actual values. The behaviour of the predictions by all the methods is similar as all the points are spreading near the line of equality.

We notice from Figure 8.1 that the predictions are not going as far away from the mean as the actual values do. This is because of "regression to the mean". In addition, we can see that the predictions from Bayes linear kinematics using the direct model are very similar to the predictions from the full Bayes analysis.

Similarly, we can see the same pattern when we use the indirect method with parameter values that we extracted from the offline learning model with the indirect model as the observations lie close to the line of equality.

Figure 8.2 shows the comparison between the various full-Bayes and BLK methods.

These figures show that the two BLK methods look very similar. In addition, the two full-Bayes methods look similar as well.

We should also mention an important point here. The standard deviation for the actual values is bigger than for the predictions (full Bayes and BLK). This is what we would expect to happen because the data do not provide perfect information. If the data were completely non-informative then the prediction would always be the prior mean and the standard deviation of the predictions would be zero. If the data are more informative then the predictions will be closer to the true values and the standard deviation of the predictions will be closer to that of the true values.

From Figure 8.3 and Figure 8.4 in Example 2, Figure 8.5 and Figure 8.6 in Example 3, we can see that increasing the variance makes the relationship between actual and predicted and, more so, between different prediction methods appear stronger.

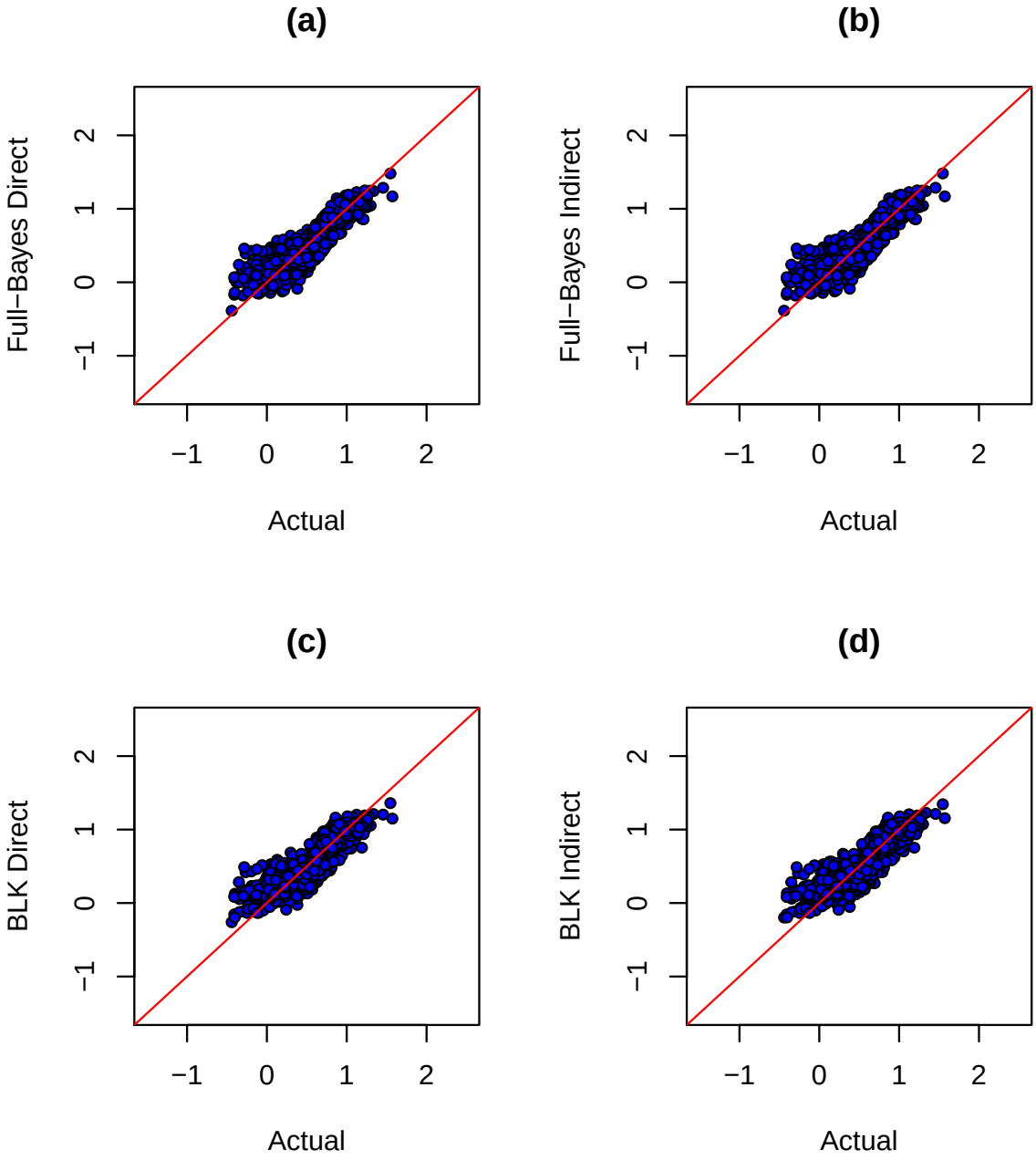


Figure 8.1: Different comparisons between predictions of the prognostic index and the actual values of Z_T, direct method and the indirect method and also using full Bayes, for data simulated using the direct model. (a): actual Z_T vs full Bayes direct. (b): actual Z_T vs full Bayes indirect. (c): actual Z_T vs BLK direct. (d): actual Z_T vs BLK indirect. (Example 1)

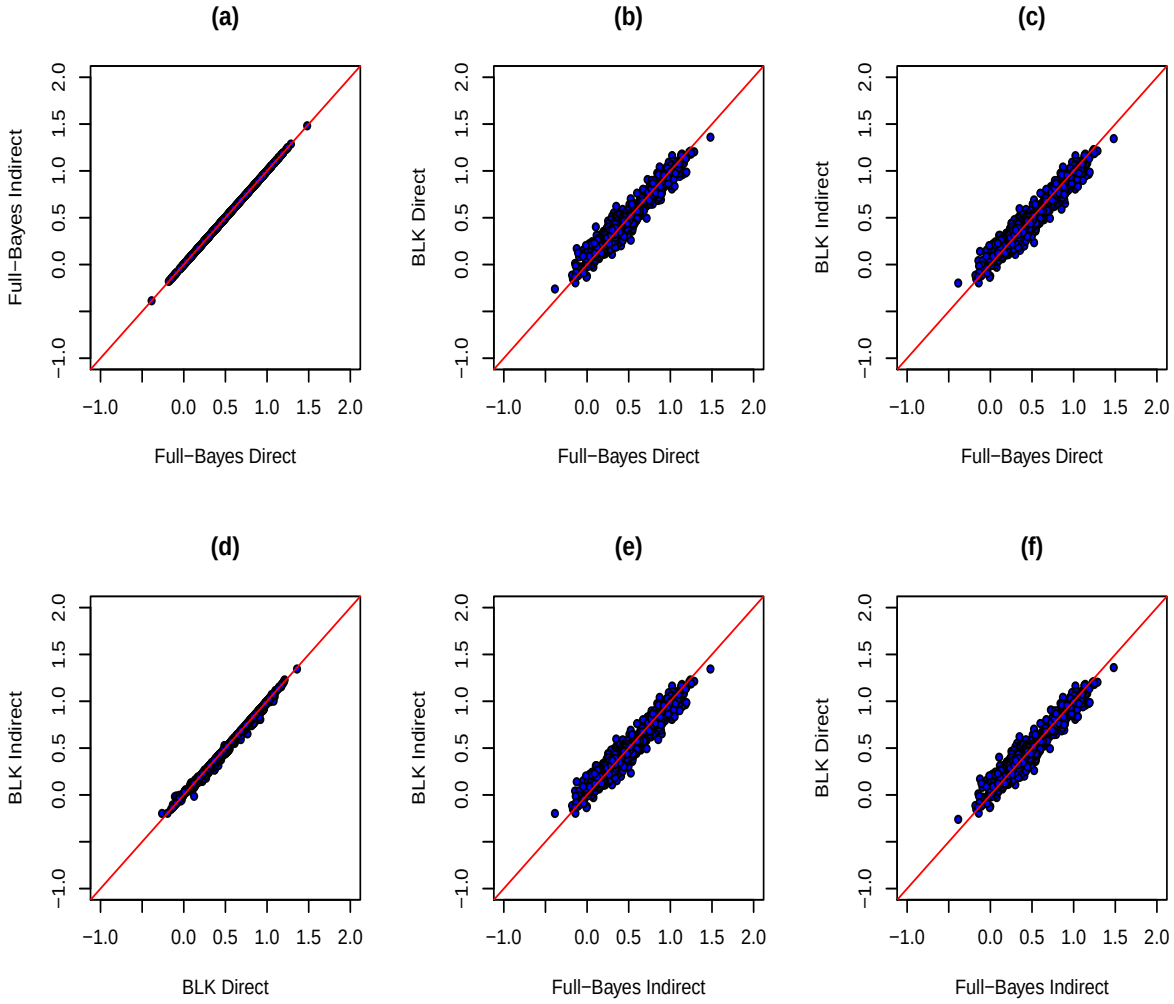


Figure 8.2: Comparing one method with another method to predict the prognostic index using both direct method and the indirect method for data simulated using the direct model. (a): full Bayes direct vs full Bayes indirect. (b): full Bayes direct vs BLK direct. (c): full Bayes direct vs BLK indirect. (d): BLK direct vs BLK indirect. (e): full Bayes indirect vs BLK indirect. (f): full Bayes indirect vs BLK direct. (Example 1)

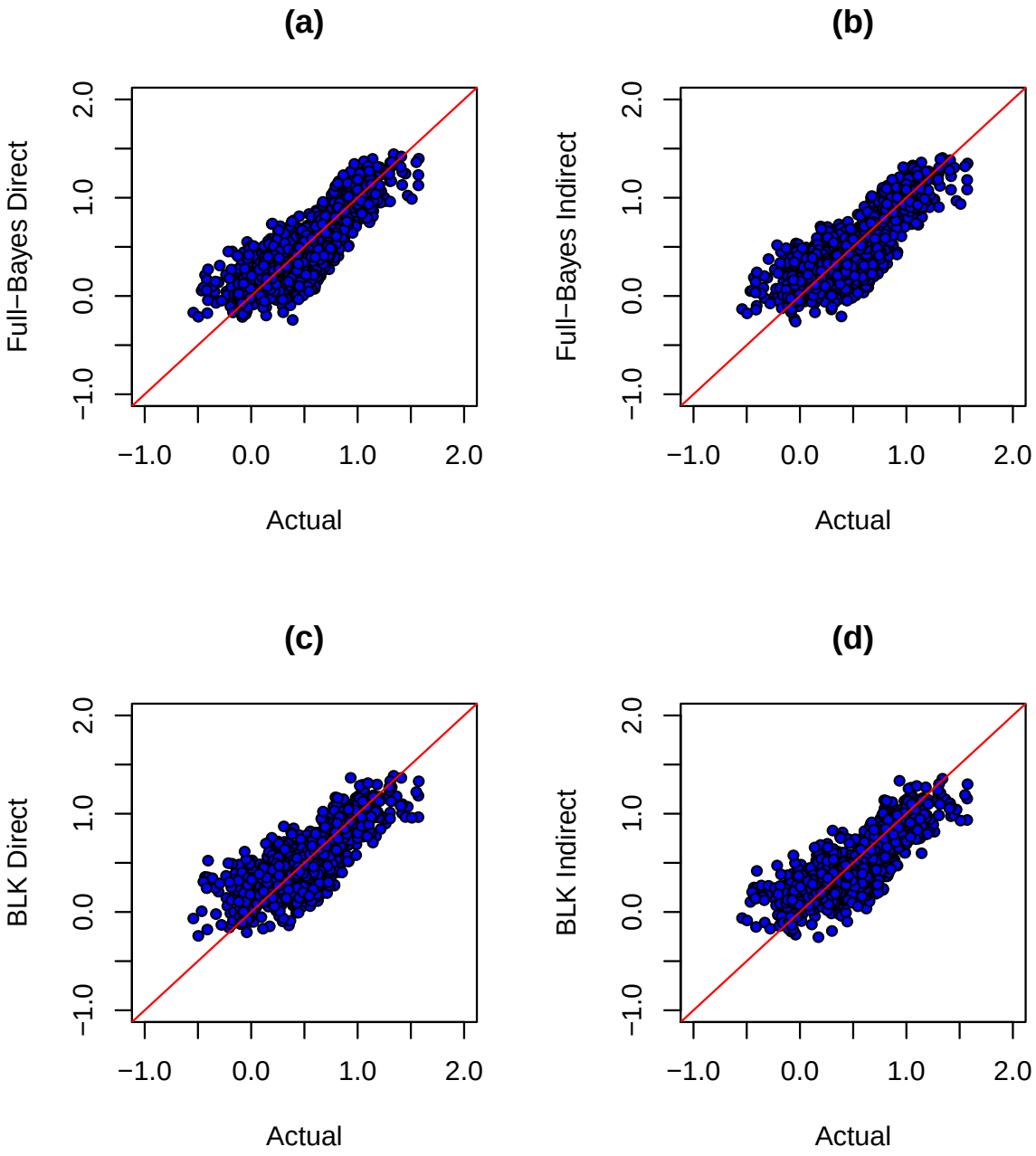


Figure 8.3: Different comparisons between predictions of the prognostic index and the actual values of Z_T, direct method and the indirect method and also using full Bayes, for data simulated using the direct model. (a): actual Z_T vs full Bayes direct. (b): actual Z_T vs full Bayes indirect. (c): actual Z_T vs BLK direct. (d): actual Z_T vs BLK indirect. (Example 2)

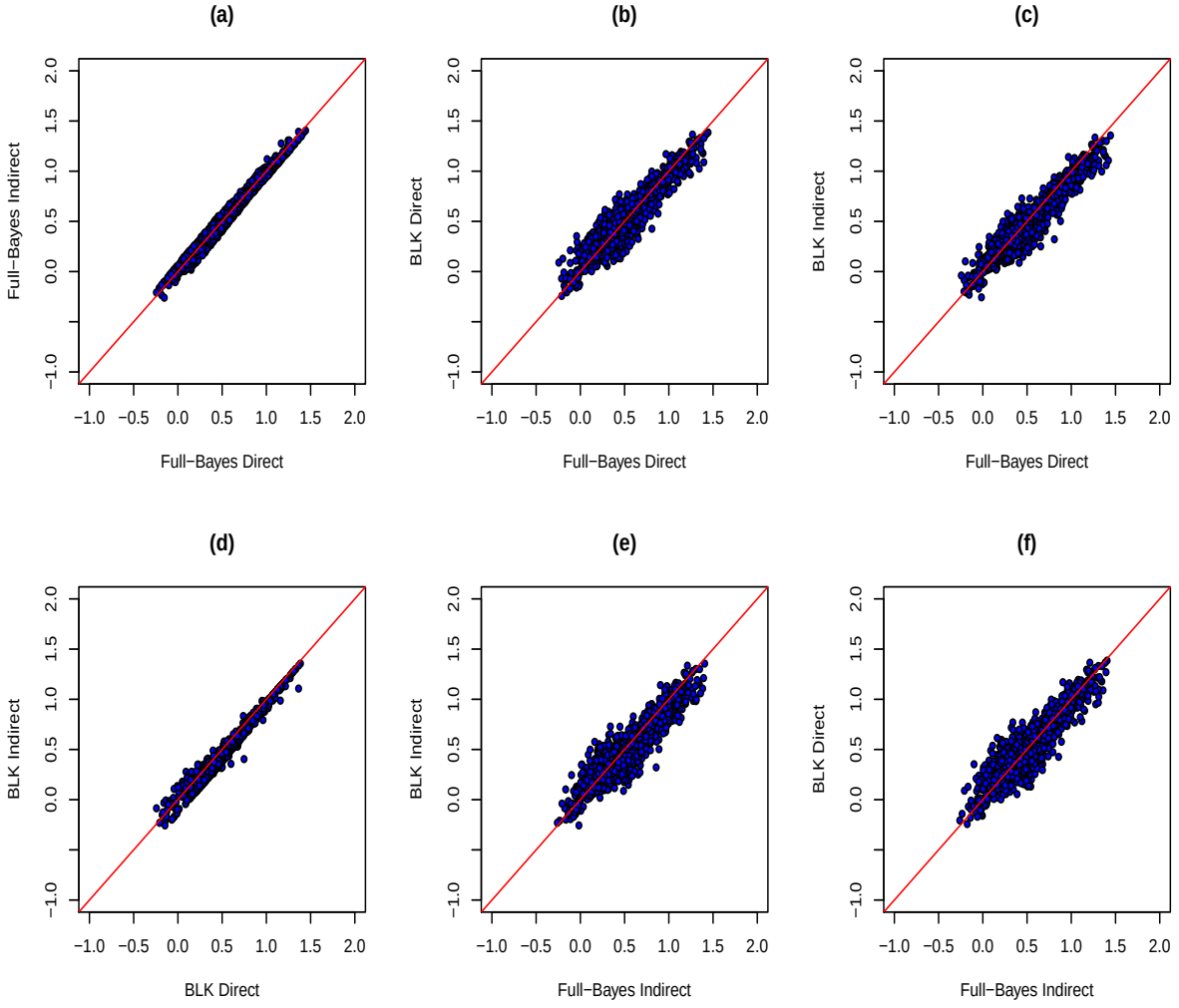


Figure 8.4: Comparing one method with another method to predict the prognostic index using both direct method and the indirect method for data simulated using the direct model. (a): full Bayes direct vs full Bayes indirect. (b): full Bayes direct vs BLK direct. (c): full Bayes direct vs BLK indirect. (d): BLK direct vs BLK indirect. (e): full Bayes indirect vs BLK indirect. (f): full Bayes indirect vs BLK direct. (Example 2)

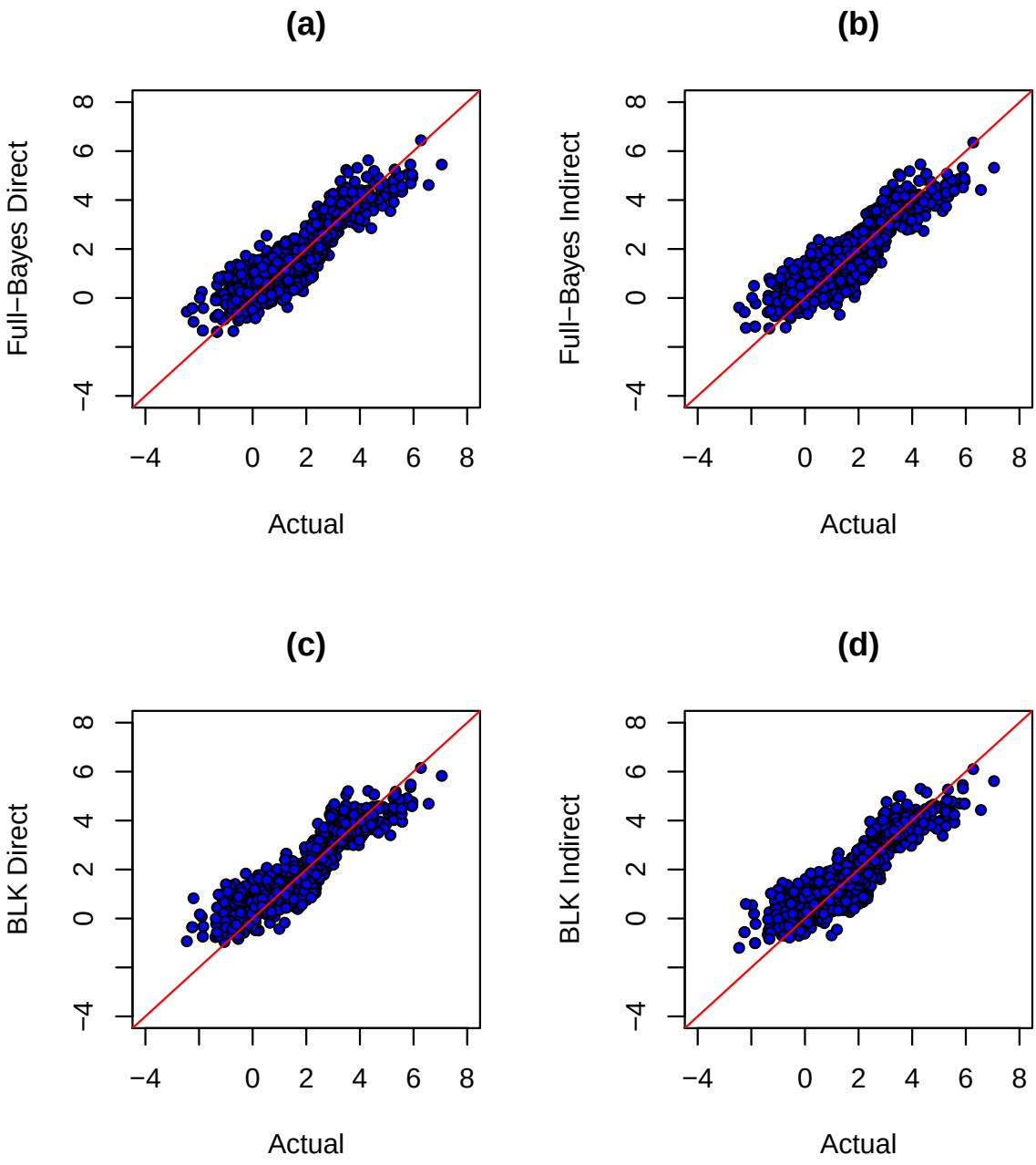


Figure 8.5: Different comparisons between predictions of the prognostic index and the actual values of Z_T, direct method and the indirect method and also using full Bayes, for data simulated using the direct model with increasing the variance of Z_T. (a): actual Z_T vs full Bayes direct. (b): actual Z_T vs full Bayes indirect. (c): actual Z_T vs BLK direct. (d): actual Z_T vs BLK indirect. (Example 3)

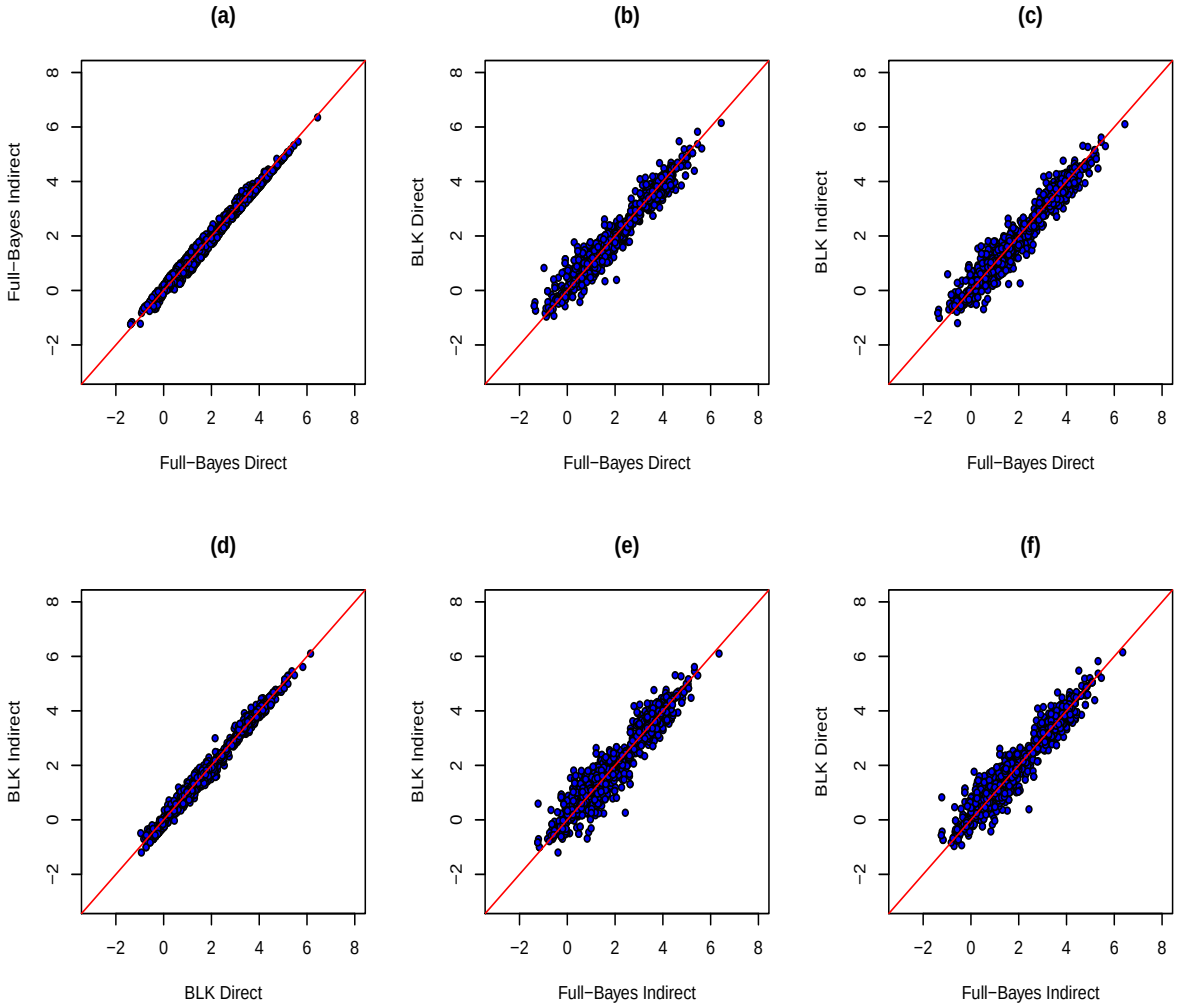


Figure 8.6: Comparing one method with another method to predict the prognostic index using both direct method and the indirect method for data simulated using the direct model with increasing the variance of Z_T. (a): full Bayes direct vs full Bayes indirect. (b): full Bayes direct vs BLK direct. (c): full Bayes direct vs BLK indirect. (d): BLK direct vs BLK indirect. (e): full Bayes indirect vs BLK indirect. (f): full Bayes indirect vs BLK direct. (Example 3)

8.3 Data simulated according to the indirect model

8.3.1 Simulation method

In this section, we simulate data according to a model corresponding to the indirect method (see Section 7.6.3). As with the direct simulations, we first generate a vector $\underline{Z}$ using a multivariate normal distribution. The mean vector and variance matrix of this distribution are those determined following the offline-learning phase with the non-Hodgkin's lymphoma data in Chapter 7. Thus

$$\begin{aligned}
\mu_0 &= (126.5844\ ,\ 8.0292\ ,\ 0.1596\ ,\ -0.8988\ ,\ 0.6034)' \\
\mu_{\text{age}} &= (-0.1766\ ,\ 0.0087\ ,\ 0.0182\ ,\ 0.0191\ ,\ 0.0134)' \\
\mu_{\text{sex}} &= (-4.9086\ ,\ -0.0080\ ,\ 0.0789\ ,\ 0.0157\ ,\ -0.0846)'
\end{aligned}$$

Therefore,

$$\mathrm{E}_0(Z_i) = \mu_0 + \mu_{\text{age}} x_{age,i} + \mu_{\text{sex}} x_{sex,i}.$$

Also

$$\mathrm{Var}_0(Z) = \begin{bmatrix}
323.2659 & -2.1548 & -10.5563 & -8.9206 & -4.3637 \\
-2.1548 & 12.1275 & 0.1440 & 0.8379 & 0.4687 \\
-10.5563 & 0.1440 & 1.3452 & 0.6220 & 0.7403 \\
-8.9206 & 0.8379 & 0.6220 & 1.4024 & 0.4692 \\
-4.3637 & 0.4687 & 0.7403 & 0.4692 & 0.7790
\end{bmatrix}.$$

As in the direct model, we generate all the values of $\underline{Z}$ randomly from a multivariate normal distribution over $\underline{Z}$. From these Z values, we compute the covariates X, this time according to the indirect model. Then, we use the different methods, full-Bayes and BLK in order to calculate the predictions of the prognostic index values using the simulated covariate values. We use three examples, as in Section 8.2. The details of these examples are given respectively as follows

- Example 1: all cases male, aged 60.
- Example 2: 200 cases in each of 6 groups: male aged 50, male aged 60, male aged

70, female aged 50, female aged 60, female aged 70.

- Example 3: variance of Z_t artificially increased, age-sex groups as in Example 2, b=3.

We generate 1200 simulated cases. In the BLK methods, we use the non-conjugate prior update to obtain the predictions for the prognostic index values. The results are shown in Figures 8.7-8.12 which correspond to Figures 8.1-8.6 for the direct data.

8.3.2 Results

Figure 8.7 shows scatter plots of predictions, i.e. posterior means, of Z_T given the data, against the true values of Z_T for Example 1. Four methods are used: full Bayes assuming a direct model, full Bayes assuming an indirect model, BLK direct and BLK indirect.

As in Section 8.2, parameter values obtained from the offline learning with the real data were used. When the direct method was used for prediction, the values used were those from fitting the direct model. When the indirect method was used for prediction, the values used were those from fitting the indirect model.

We notice from Figure 8.7 that the predictions are reasonably good and the predictions from Bayes linear kinematics using the indirect model follow a similar pattern to the predictions from the full Bayes analysis.

Figure 8.8 represents the comparison between the various full-Bayes and BLK methods. These figures show that the two BLK methods look very similar. In addition, the two full-Bayes methods look similar as well. The relationship between the full-Bayes predictions and BLK predictions is not quite as strong.

We notice also from Figure 8.9 and Figure 8.10 in Example 2, having different age groups can make the relationship fairly stronger between the actual and different predicted values.

For Example 3, Figures 8.11 and 8.12 (for the indirect model) show that when we increased the variance of Z_T, we have obtained results which appear better than when we use a small variance of Z_T. We set up the new variance of Z_t to be $\mathrm{Var}^*(Z_t) = b\mathrm{Var}(Z_t)$, where $b = 3$ in this example.

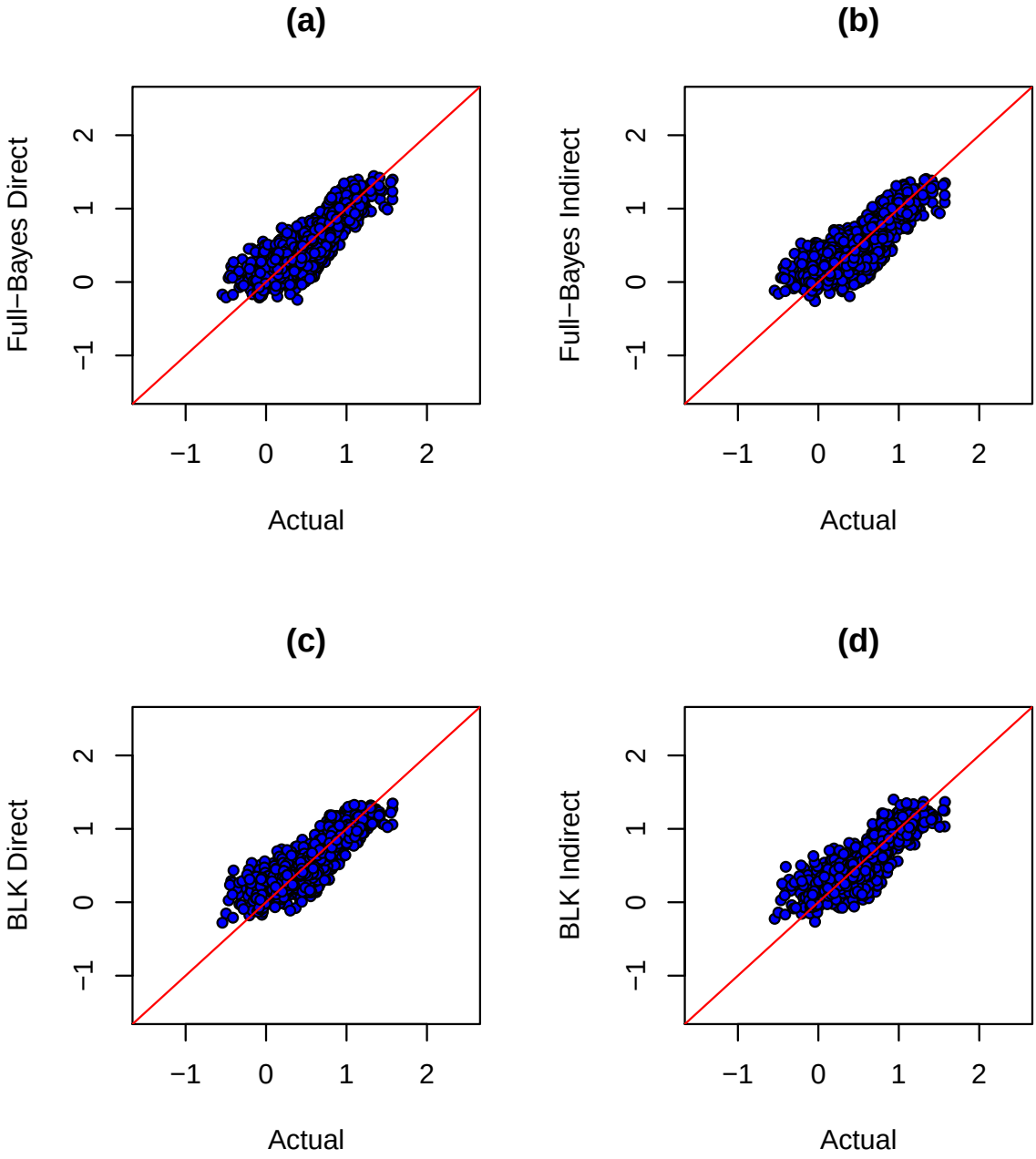


Figure 8.7: Different comparisons between predictions of the prognostic index and the actual values of Z_T, direct method and the indirect method and also using full Bayes, for data simulated using the indirect model. (a): actual Z_T vs full Bayes direct. (b): actual Z_T vs full Bayes indirect. (c): actual Z_T vs BLK direct. (d): actual Z_T vs BLK indirect. (Example 1)

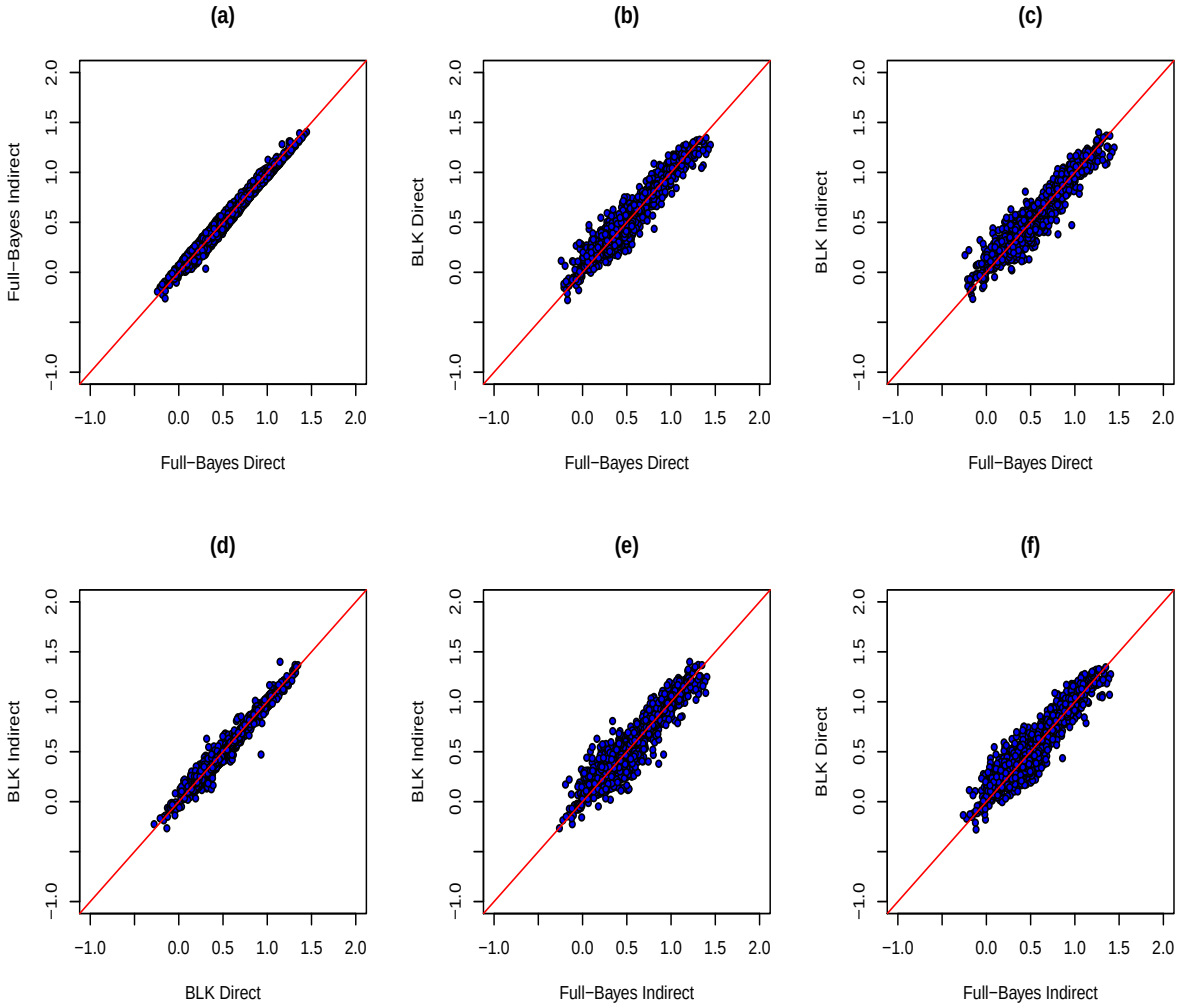


Figure 8.8: Comparing one method with another method to predict the prognostic index using both direct method and the indirect method for data simulated using the indirect model. (a): full Bayes direct vs full Bayes indirect. (b): full Bayes direct vs BLK direct. (c): full Bayes direct vs BLK indirect. (d): BLK direct vs BLK indirect. (e): full Bayes indirect vs BLK indirect. (f): full Bayes indirect vs BLK direct. (Example 1)

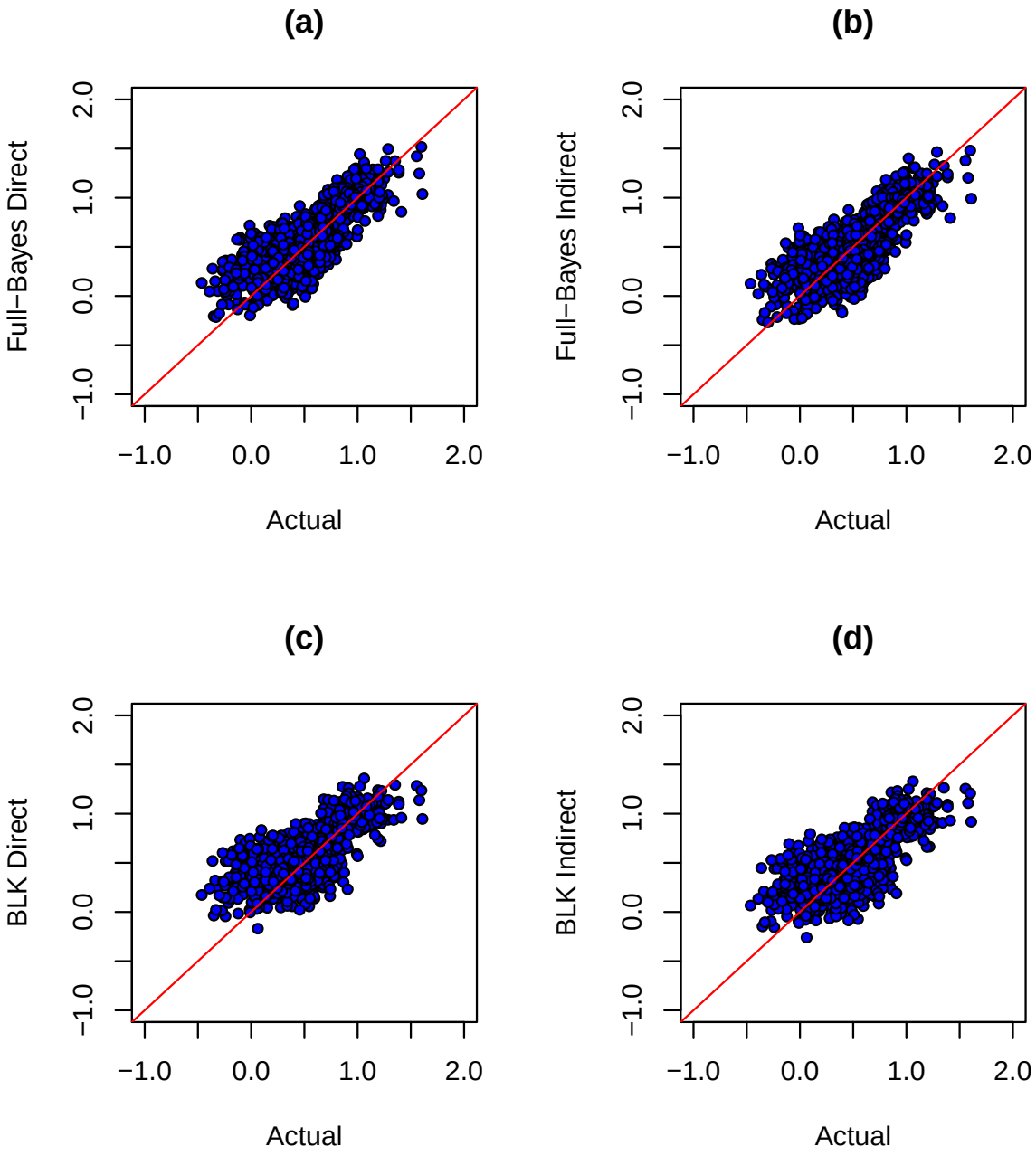


Figure 8.9: Different comparisons between predictions of the prognostic index and the actual values of Z_T, direct method and the indirect method and also using full Bayes, for data simulated using the indirect model. (a): actual Z_T vs full Bayes direct. (b): actual Z_T vs full Bayes indirect. (c): actual Z_T vs BLK direct. (d): actual Z_T vs BLK indirect. (Example 2)

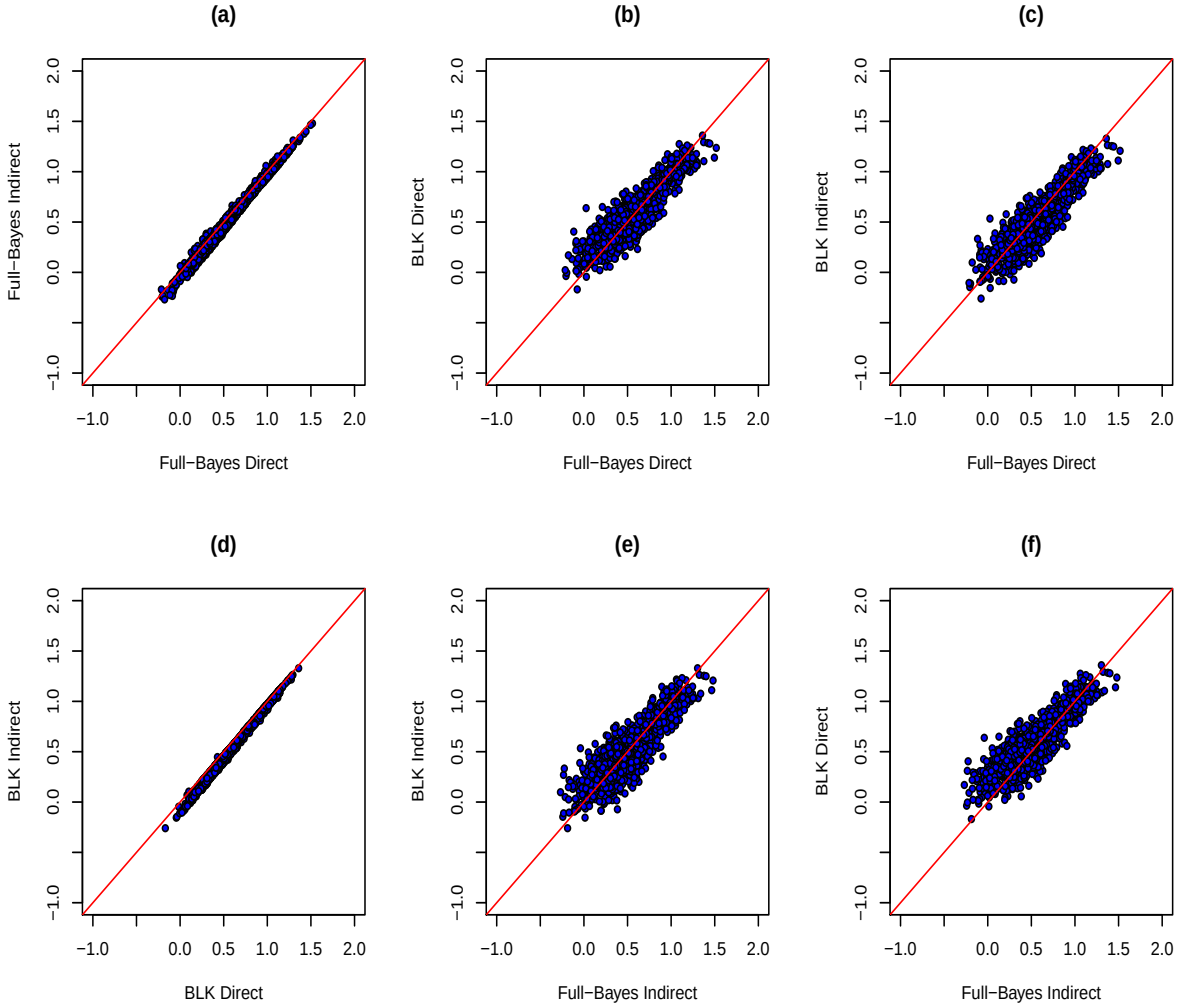


Figure 8.10: Comparing one method with another method to predict the prognostic index using both direct method and the indirect method for data simulated using the indirect model. (a): full Bayes direct vs full Bayes indirect. (b): full Bayes direct vs BLK direct. (c): full Bayes direct vs BLK indirect. (d): BLK direct vs BLK indirect. (e): full Bayes indirect vs BLK indirect. (f): full Bayes indirect vs BLK direct. (Example 2)

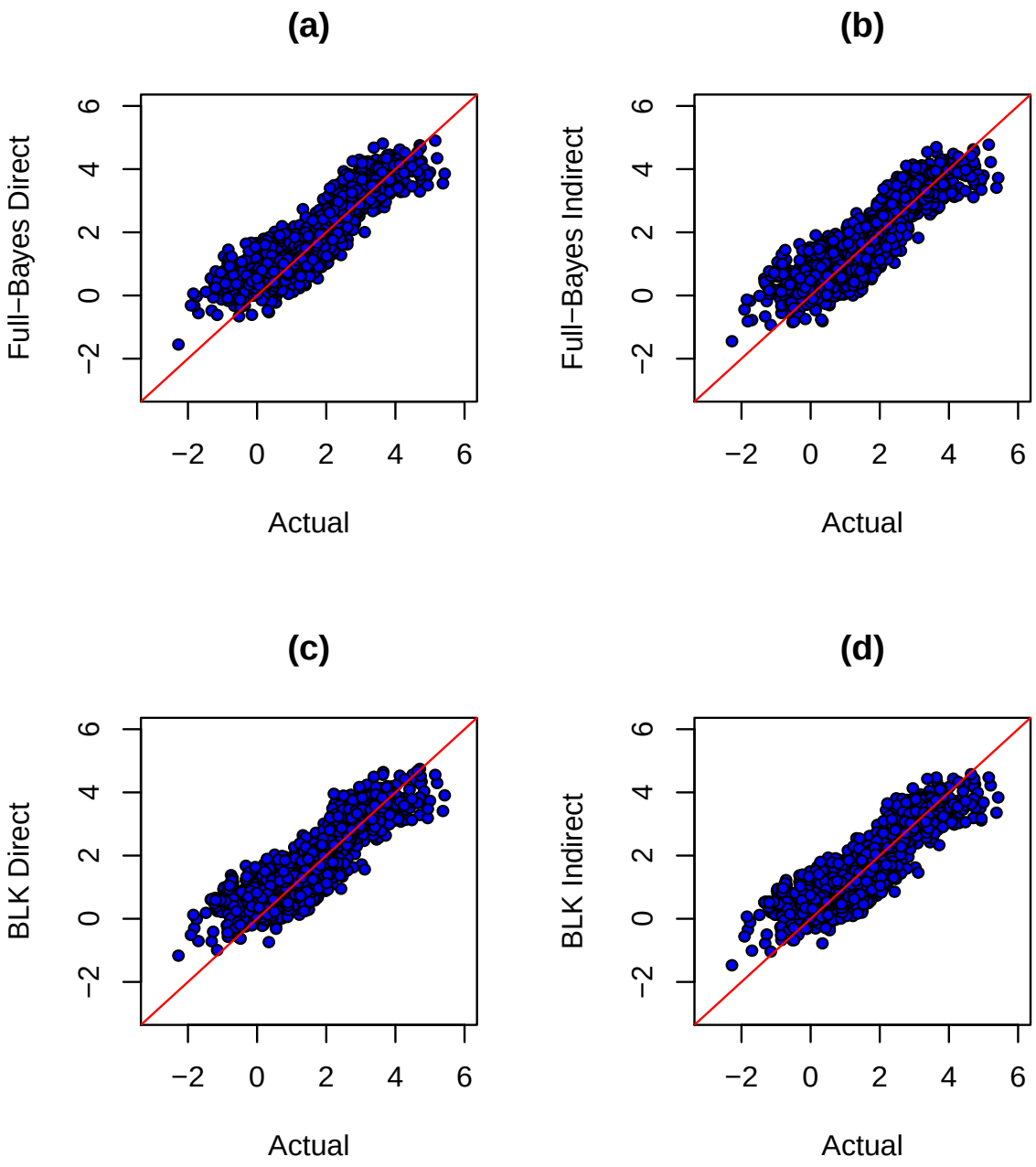


Figure 8.11: Different comparisons between predictions of the prognostic index and the actual values of Z_T, direct method and the indirect method and also using full Bayes, for data simulated using the indirect model with increasing the variance of Z_T. (a): actual Z_T vs full Bayes direct. (b): actual Z_T vs full Bayes indirect. (c): actual Z_T vs BLK direct. (d): actual Z_T vs BLK indirect. (Example 3)

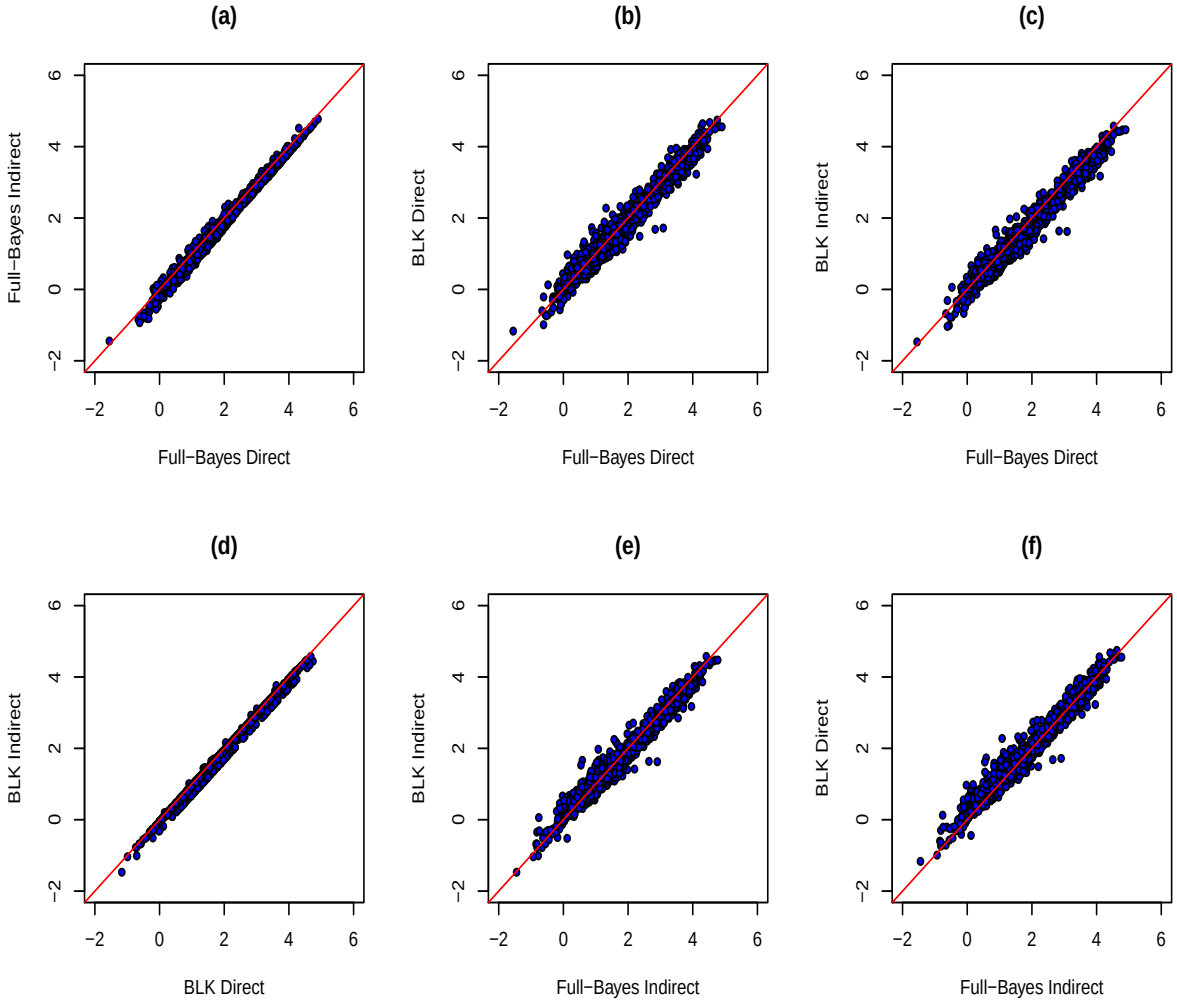


Figure 8.12: Comparing one method with another method to predict the prognostic index using both direct method and the indirect method for data simulated using the indirect model with increasing the variance of Z_T. (a): full Bayes direct vs full Bayes indirect. (b): full Bayes direct vs BLK direct. (c): full Bayes direct vs BLK indirect. (d): BLK direct vs BLK indirect. (e): full Bayes indirect vs BLK indirect. (f): full Bayes indirect vs BLK direct. (Example 3)

8.4 Conclusion

In conclusion, the Bayes linear Bayes approach produced solutions more quickly than the full Bayes method. For example, the time that we need to obtain the results for 1200 simulated patients on a desk-top computer using full-Bayes analysis is about 544 seconds, while the time we need to produce the results using the Bayes linear kinematic method is about 3.3 seconds, which is very much faster than the full Bayes method.

The Bayes linear kinematic predictions appeared to be only slightly less close to the true values than the full-Bayes predictions. Whether the simulated data were generated using the direct or the indirect model, it seemed to make little difference whether the predictions were calculated according to the direct or indirect method.

Importantly, the Bayes linear kinematic methods can easily be used if only some of the covariates are observed.

Following the results that we have obtained so far from the simulations, we may recommend the use of either method, the direct or the indirect, as they both give very close results to each other. Both methods are faster and easier to use once we have obtained the parameter values that we extract from the offline learning. The indirect method may be preferred as it gives unbounded posterior support rather than the bounded posterior in the direct methods.

Chapter 9

Conclusion and Future Work

9.1 Summary of the project

In this thesis we have been concerned with methods that use Bayesian inference and particularly Bayes linear Bayes methods for different types of distribution, such as binomial, Poisson, etc. Firstly, in Chapter 1 we reviewed some literature related to our work such as using structure learning in Bayesian networks and Bayesian networks in survival analysis.

Chapter 2 described two data sets that we used in this thesis which are non-Hodgkin lymphoma (NHL) and leukemia data. We have explained in detail the covariates in each data set.

We consider full-Bayes analysis methods such as MCMC techniques in Chapter 3. Initially, we gave an introduction to Bayesian inference. We have discussed some numerical integration methods that we need in this thesis such as Laplace approximation and the trapezoidal rule. Various types of MCMC algorithms have been illustrated in this chapter, for example, Monte Carlo integration, importance sampling and, of course, the three most important algorithms in Bayesian analysis, Gibbs sampling, Metropolis-Hastings algorithm and Metropolis within Gibbs. We also mentioned how we can obtain samples from the posterior distributions using MCMC and check the convergence of the chains. The second part of this chapter dealt with generalised linear models (GLMs) with common link functions such as logit and probit. We then gave the theory of some variable selection methods and how we can use Bayesian analysis to select the most important variables in the model. We used an illustrative example of a Bayesian logistic regression model. We also explained some variable selection methods which depend on various forms of

prior distribution, such as spike and slab prior and Zellner's g prior. In many data sets, some values are missing. This chapter has described the problem of missing data and the related method of data augmentation. We use the lung transplant example with many covariates and a logistic regression model to illustrate the idea of variable selection from a Bayesian perspective with some results showing that the posterior predictive probabilities are close to the observed values of the dependent variable.

In Chapter 4, we reviewed a special type of probabilistic graphical models called a Bayesian network (BN). Then we gave some important terms and definitions in Bayesian networks. We explained the main points of difference between Bayesian networks and regression models. In particular, in regression model we do not specify a probability distribution for the covariates. However, in a BN, we specify the joint probability distribution for all of the variables so we can use it even when we observe only some of these variables. We may wish to learn about the parameter of a Bayesian network from data. This is called parameter learning in BN. We explained, in brief, learning the parameters in both cases, with complete data and incomplete data. We gave different examples of learning from data in different cases, such as a categorical network and a Gaussian network. We used an `R` package called "`bnlearn`", which stands for Bayesian network learning, to learn about the structure of BNs. We focused on using two algorithms to construct the network which are the "Grow-Shrink" (GS) algorithm and the "Hill-Climbing" algorithm. These algorithms depend on a score function such as the Bayesian information criterion (BIC) to specify the optimal Bayesian network. We explained both algorithms in a motivational example. Finally, we used a proposed method to construct a Bayesian network. The method is called "arc-deletion" method. This method requires that we impose the order of the nodes in the network. Then we used MCMC to select the most likely configuration and that depends on the posterior probability of the coefficients being non-zero in the model. We applied this method to the non-Hodgkin lymphoma data set. We found that the most likely structure from this method had fewer arrows, i.e. we dropped some of the arcs as the posterior probability for them was close to zero. That is beneficial as we do not need as many calculations to compute the joint probability distributions as in a fully-connected network.

We gave general background on survival analysis in Chapter 5. We illustrated some useful models that relate the survival lifetime distribution to some covariates in the model. These models are proportional hazard models, piecewise constant hazard models and accelerated failure time models. We described a prognostic index which is used to predict

the outcome in patients with a certain disease. We showed for example, how to calculate this index by fitting the Weibull lifetime distribution. In this chapter, we demonstrated how we can calculate the posterior distribution for the parameters in survival analysis for exponential and Weibull distributions. We used the software called "`rjags`" to compute all the posterior means and variances for the parameters of interest in the model. We showed some results and graphs showing that our sampler mixed well and the chains converged.

The main contribution of the thesis is in Chapters 6 and 7. In Chapter 6 we investigated Bayes linear methods with some theoretical aspects of this approach. In Bayes linear methods, we do not need to specify the prior in a probabilistic way as in full-Bayes analysis. We explained the idea of Bayes linear analysis using a motivational example. We explained Bayes linear kinematics (BLK) and mentioned the concept of "commutativity" and how to do multiple updates using BLK.

We explained Bayes linear Bayes graphical models as a combination of Bayesian networks and Bayes linear structure. We use the idea of transformation of the parameters for different reasons. However, an important reason is that, when a quantity θ has a bounded range, this makes the use of Bayes linear methods less attractive.

After transforming the parameters, we can use the mode and log-curvature method that we explained in Section 6.5.3 to relate the distribution of the parameters to moments on the transformed scale. We apply the mode and curvature method to construct the mean and variance for the transformed parameters. We use an example for illustration. In this chapter we introduced a new method for updating our means and variances which depends on non-conjugate prior updates in order to calculate Bayes linear kinematics. In the case of a non-conjugate prior, we need to use some numerical integration methods such as Laplace approximation, Gauss-Hermite quadrature and the trapezoidal rule. However, the integrations are typically one-dimensional in contrast to the multi-dimensional integrations often required in a full-Bayes analysis. The use of non-conjugate priors also extends the range of types of variable which we can use in a Bayes linear Bayes model. We used two examples to demonstrate the idea of using non-conjugate prior updates. Finally, we illustrated briefly various sorts of variables that we can deal with, such as binary, ordinal, unordered categorical and interval censored variables.

Chapter 7 reviewed two examples, using data on patients with leukaemia and non-Hodgkin lymphoma. For the leukaemia example, we reviewed the Wilson and Farrow approach using a piecewise constant hazards model. Then we used our non-conjugate

updates in this example in order to compare the results with Wilson and Farrow (2017). We did some diagnostic checking in the leukaemia example to assess the validity of the assumptions in our model.

For the non-Hodgkin lymphoma example, firstly, we introduced Bayes linear Bayes prognostic networks and then applied this idea to the NHL example. Secondly, we introduced the novelty of the model relating T to a latent prognostic index.

Initially, we have explained the general strategy of offline learning to construct a Bayes linear Bayes network and then Bayes linear kinematics in the routine use of a Bayes linear Bayes prognostic network. We described the offline learning model using `rjags` and how we can use the posterior means of the parameters in the model. We introduced two different structures for the Bayes linear Bayes model, called the direct method and the indirect method. Then we used BLK and did some comparisons between full-Bayes and Bayes linear kinematics. We found that the prognostic index values from MCMC (full-Bayes) and BLK look similar. A Bland and Altman agreement plot showed that only 4% of the prognostic index values were outside the limits. Our prototype prognostic network produces prognostic index values using all, or some, of the possible covariates almost instantly and has the potential to be used, for example, as a Web-based calculator.

We also set out some advantages and disadvantages of using the direct and the indirect method.

In summary, the indirect method is preferable to the direct method as the latter has bounded support for the posterior and the indirect method give us posterior means which are closer to the MCMC values than those which the direct method gives.

Since, at present, we use a separate, offline, full-Bayes model to choose parameter values for the Bayes linear Bayes model, in practice, from time to time, after we have observed new data, we might run the offline learning again with the addition of new cases and update the Bayes linear kinematic network.

Finally, in Chapter 8, we implemented a simulation experiment to compare the direct and the indirect methods based on the results of these methods in Chapter 7. So we generated repeated simulations when the direct model is correct but both direct and indirect methods are used to generate the prognostic index distributions. We used full Bayes and BLK methods to compute predictions. We repeated this for the indirect model as well.

9.2 A review of the objectives of the project

In Chapter 1 of this thesis, we stated the aim of the project with some listed points. These were as follows

1. Develop Bayesian methods for selecting, fitting and using models with appropriate conditional independence structures, *i.e.* graphical models, in the context of medical diagnosis and prognosis problems. In addition, we are looking for improvements to some existing methods.

2. Investigate methods for a wider class of conditional distributions.

3. Build probabilistic models for diagnosis and prognosis with various Bayesian network learning algorithms to help the physicians and others to make decisions about their patients more accurately and efficiently.

4. Propose the novelty of using the non-conjugate prior update in order to obtain the posterior moments using Bayes linear kinematics.

5. Construct a Bayes linear kinematic network which can be used when we observe only some of the covariates. Develop methods for incorporating different kind of covariates in such a network.

6. Propose two new methods, the direct and the indirect methods to compute the prototype prognostic network and has the potential to be used, for example, as a Web-based calculator.

We have achieved some of these objectives of the study. For instance, in Chapter 4 we have constructed a Bayesian network with a Weibull lifetime distribution for the leukemia data. We made imposed an order on the covariates in the model and we put the covariates that we always observed first in the network. We also introduced a method to construct a Bayesian network which depends on selecting the most likely configuration and that depends on the posterior probability of the coefficients being non-zero.

The main way in which we achieved the objectives was by the extension of methods for Bayes linear Bayes graphical models to allow non-conjugate marginal updates and the application of this to a Bayes linear Bayes prognostic index.

In Chapter 6, we gave the general theory and some examples where we showed that the adjusted moments from Bayes linear kinematics and the posterior moments from full-Bayes analysis were close to each other. We discussed examples such as binomial observations and Poisson observations. We developed methods for incorporating variables of different types into a Bayes linear Bayes network. All these examples used non-conjugate prior updates which is an extension of the work done by Wilson and Farrow (2010, 2017).

Chapter 7 showed problems of different kinds using the leukaemia example and the non-Hodgkin lymphoma example. In the leukaemia example we demonstrated the use of Bayes linear kinematics to fit a survival model and make inferences about the values of parameters, using our novel non-conjugate marginal updates. In the non-Hodgkin lymphoma example, we showed the idea of constructing a Bayes linear Bayes prognostic index which depends on fitting an offline learning model for a Weibull survival time and making inferences about the parameters in the model and then using Bayes linear kinematics to update our beliefs about $\underline{Z} = (Z_1, ..., Z_J, Z_{J+1})'$ in routine use with new patients. We are particularly interested to predict Z_{J+1} which is linked with the survival time. We compared the posterior moments with full Bayes analysis which gave us reasonable results. We also compared the direct and the indirect method and we found that the indirect method is more accurate than the direct method and the reason is that, in the indirect method, we have an unbounded support for the posterior distribution with ordinal variables rather than the bounded posterior support in the direct method.

In Chapter 8, we did some simulation experiments in survival analysis in order to compare the direct and the indirect methods motivated by the results in the non-Hodgkin lymphoma example in Chapter 7.

9.3 Future work

There is further work to do on selecting Bayesian network structure. Our idea is to use a score criterion the expectation of a suitable utility function. Similarly work can be done which applies this idea to variable selection methods. The idea is to use the more important covariates in the model again depending upon the maximisation of the expected utility.

We have investigated an "arc-deletion" method, which can calculate the posterior mean of the indicators, in other words the posterior probability that those coefficients are

non-zero. This method can drop some of the arrows from the network and that leads to simplifying the calculation of the joint probability distribution. We can do more work to improve this method by changing the directions of the arrows to obtain the optimal structure. That is, we can apply a utility function as a score function in algorithms such as "Hill-Climbing" and "Grow-Shrink". Further work is also required on the choice of suitable utility functions.

In our Bayes linear Bayes graphical models we used a general, that is fully-connected, covariance structure. There is a need for further research on using structures which are not fully-connected and which exploit conditional independence.

In our Bayes linear kinematic prognostic index calculation we used the posterior means of model parameters produced in the offline learning phase. Ideally we would use posterior expectations of the functions of model parameters which are needed in the Bayes linear kinematic calculation. Further work is needed to develop this.

Appendix A

Appendix

A.1 General Appendix

A.1.1 Software

In this thesis, we use the software called jags, which stands for "Just Another Gibbs Sampler" (Plummer, 2017). It uses MCMC to fit many different kinds of models specified using a model specification language based on the `BUGS` language (Spiegelhalter et al., 1996). To run jags we use the `rjags` package (Plummer, 2013) within `R` (R Core Team, 2018).

A.2 Appendix to Chapter 2

A.2.1 Few observations of SNLG data

t	died	age	sex	albumin	ap	bsy ...
0.016438356	1	72	1	2	2	2 ...
0.250228311	1	63	1	1	1	2 ...
0.084018265	1	70	1	1	2	1 ...
0.168949772	0	58	2	1	1	2 ...

Table A.1: Few observations of SNLG data.

A.2.2 Few observations of leukaemia data

t	t.cen	age	sex	wbc	depsc
1	0	61	1	13.3	-1.96
1	0	76	1	450.0	-3.39
1	0	74	1	154.0	-4.95
1	0	79	2	500.0	-1.40
1	0	83	2	160.0	-2.59

Table A.2: Few observations of leukaemia data.

A.3 Appendix to Chapter 3

A.3.1 R function to generate samples from the posterior distribution of μ and τ.

```
function(y,prior,n.iter,start=list(mu=0,tau=1))
{
ybar<-mean(y)
n<-length(y)
Sy<-sum((y-ybar)^2)
musamples<-numeric(n)
tausamples<-numeric(n)
iteration<-1:n.iter
tau<-start$tau
a<-prior$a
b<-prior$b
m<-prior$m
p<-prior$p
ind<-a+n/2
for (i in 1:n.iter)
{d<-p+n*tau
mean<-(p*m+n*tau*ybar)/d
sd<-sqrt(1/d)
```

```
mu<-rnorm(1,mean,sd)
musamples[i]<-mu
scale<-b+0.5*(Sy+n*((ybar-mu)^2))
tau<-rgamma(1,ind,scale)
tausamples[i]<-tau
}
results<-list(iter=iteration,mu=musamples,tau=tausamples)
return(results)
}
```

A.3.2 Rjags specification for the logistic regression model of lung transplant example with missing covariates data

```
model{
for(i in 1:41){
y[i]~dbern(p[i])
logit(p[i])<- beta0+beta1*(new.x1[i]-mean(new.x1[]))+
beta2*(new.x2[i]-mean(new.x2[]))+
beta3*(new.x3[i]-mean(new.x3[]))+beta4*(new.x4[i]-mean(new.x4[]))+
beta5*(new.x5[i]-mean(new.x5[]))+beta6*(new.x6[i]-mean(new.x6[]))+
beta7*(new.x7[i]-mean(new.x7[]))+beta8*(new.x8[i]-mean(new.x8[]))

new.x1[i]~dnorm(m39,m40)
new.x2[i]~dnorm(m41[i],m42)
new.x3[i]~dnorm(m43[i],m44)
new.x4[i]~dnorm(m45[i],m46)
new.x5[i]~dnorm(m47[i],m48)
new.x6[i]~dnorm(m49[i],m50)
new.x7[i]~dnorm(m51[i],m52)
new.x8[i]~dnorm(m53[i],m54)
m41[i]<-beta0.x2+beta1.x1x2*(new.x1[i]-mean(new.x1[]))
m43[i]<-beta0.x3+beta1.x3x2*(new.x2[i]-mean(new.x2[]))+
beta2.x3x1*(new.x1[i]-mean(new.x1[]))
m45[i]<-beta0.x4+beta1.x4x1*(new.x1[i]-mean(new.x1[]))+
beta2.x4x2*(new.x2[i]-mean(new.x2[]))+beta2.x4x3*(new.x3[i]-mean(new.x3[]))
```

```
m47[i]<-beta0.x5+beta1.x5x1*(new.x1[i]-mean(new.x1[]))+
beta2.x5x2*(new.x2[i]-mean(new.x2[]))+
beta2.x5x3*(new.x3[i]-mean(new.x3[]))+beta2.x5x4*(new.x4[i]-mean(new.x4[]))

m49[i]<-beta0.x6+beta1.x6x1*(new.x1[i]-mean(new.x1[]))+
beta2.x6x2*(new.x2[i]-mean(new.x2[]))+
beta2.x6x3*(new.x3[i]-mean(new.x3[]))+beta2.x6x4*(new.x4[i]-mean(new.x4[]))+
beta2.x6x5*(new.x5[i]-mean(new.x5[]))

m51[i]<-beta0.x7+beta1.x7x1*(new.x1[i]-mean(new.x1[]))+
beta2.x7x2*(new.x2[i]-mean(new.x2[]))+
beta2.x7x3*(new.x3[i]-mean(new.x3[]))+
beta2.x7x4*(new.x4[i]-mean(new.x4[]))+
beta2.x7x5*(new.x5[i]-mean(new.x5[]))+
beta2.x7x6*(new.x6[i]-mean(new.x6[]))

m53[i]<-beta0.x8+beta1.x8x1*(new.x1[i]-mean(new.x1[]))+
beta2.x8x2*(new.x2[i]-mean(new.x2[]))+
beta2.x8x3*(new.x3[i]-mean(new.x3[]))+beta2.x8x4*(new.x4[i]-mean(new.x4[]))+
beta2.x8x5*(new.x5[i]-mean(new.x5[]))+beta2.x8x6*(new.x6[i]-mean(new.x6[]))+
beta2.x8x7*(new.x7[i]-mean(new.x7[]))
}
m39~dnorm(0,0.03)
m40~dgamma(2,3)
m42~dgamma(2,3)
m44~dgamma(2,3)
m46~dgamma(2,3)
m48~dgamma(2,3)
m50~dgamma(2,3)
m52~dgamma(2,3)
m54~dgamma(2,3)

beta0.x2~dnorm(0,0.03)
beta1.x1x2~dnorm(0,0.03)
```

```
beta0.x3~dnorm(0,0.03)
beta1.x3x2~dnorm(0,0.03)
beta2.x3x1~dnorm(0,0.03)
beta0.x4~dnorm(0,0.03)
beta1.x4x1~dnorm(0,0.03)
beta2.x4x2~dnorm(0,0.03)
beta2.x4x3~dnorm(0,0.03)

beta0.x5~dnorm(0,0.03)
beta1.x5x1~dnorm(0,0.03)
beta2.x5x2~dnorm(0,0.03)
beta2.x5x3~dnorm(0,0.03)
beta2.x5x4~dnorm(0,0.03)

beta0.x6~dnorm(0,0.03)
beta1.x6x1~dnorm(0,0.03)
beta2.x6x2~dnorm(0,0.03)
beta2.x6x3~dnorm(0,0.03)
beta2.x6x4~dnorm(0,0.03)
beta2.x6x5~dnorm(0,0.03)

beta0.x7~dnorm(0,0.03)
beta1.x7x1~dnorm(0,0.03)
beta2.x7x2~dnorm(0,0.03)
beta2.x7x3~dnorm(0,0.03)
beta2.x7x4~dnorm(0,0.03)
beta2.x7x5~dnorm(0,0.03)
beta2.x7x6~dnorm(0,0.03)

beta0.x8~dnorm(0,0.03)
beta1.x8x1~dnorm(0,0.03)
beta2.x8x2~dnorm(0,0.03)
beta2.x8x3~dnorm(0,0.03)
beta2.x8x4~dnorm(0,0.03)
beta2.x8x5~dnorm(0,0.03)
```

```
beta2.x8x6~dnorm(0,0.03)
beta2.x8x7~dnorm(0,0.03)

beta0~dnorm(-0.5,0.6)
beta1~dnorm(0,0.1)
beta2~dnorm(0,0.1)
beta3~dnorm(0,0.1)
beta4~dnorm(0,0.1)
beta5~dnorm(0,0.1)
beta6~dnorm(0,0.1)
beta7~dnorm(0,0.1)
beta8~dnorm(0,0.1)
}
```

A.4 Appendix to Chapter 4

A.4.1 Rjags specification to compute the posterior probabilities for the coefficients which are non-zero for non-Hodgkin lymphoma data

```
model{
for(i in 1:636){
is.cen[i]~dinterval(t[i],t.cen[i])
t[i]~dnorm(mu.t[i],tau)
mu.t[i]<-beta0t+beta.tage[zbeta.tage]*age[i]+beta.tsex[zbeta.tsex]*sex[i]

age[i]~dnorm(mu.age[i],tau1)
mu.age[i]~dnorm(60,0.005)

sex[i]~dcat(p[])
hb[i]~dnorm(mu.hb[i],tau2)
wbc[i]~dnorm(mu.wbc[i],tau3)
albumin[i]~dcat(p1[])
```

```
mu.wbc[i]<-beta0.wbc+
beta.wbcage[zbeta.wbcage]*age[i]+
beta.wbcsex[zbeta.wbcsex]*sex[i]+
beta.wbchb[zbeta.wbchb]*hb[i]+
beta.wbct[zbeta.wbct]*t[i]
mu.hb[i]<-beta0.hb+beta.hbt[zbeta.hbt]*t[i]+
beta.hbage[zbeta.hbage]*age[i]+beta.hbsex[zbeta.hbsex]*sex[i]

}

pi~dbeta(2,3)
p[1]<-pi
p[2]<-1-pi

pi1~dbeta(1,2)
p1[1]<-pi1
p1[2]<-1-pi1

tau~dgamma(2,3)
tau1~dgamma(1,3)
tau2~dgamma(1,3)
tau3~dgamma(1,3)

beta0t~dnorm(0,3)
beta0.hb~dnorm(0,1)
beta0.wbc~dnorm(1,2)

beta.wbcage[1]~dnorm(0,0.1)
beta.wbcage[2]<-0
zbeta.wbcage~dcat(pi.wbcage[])
pi.wbcage[1]<-0.5
pi.wbcage[2]<-1-pi.wbcage[1]
zzbeta.wbcage<-2-zbeta.wbcage
```

```
beta.wbchb[1]~dnorm(0,0.2)
beta.wbchb[2]<-0
zbeta.wbchb~dcat(pi.wbchb[])
pi.wbchb[1]<-0.5
pi.wbchb[2]<-1-pi.wbchb[1]
zzbeta.wbchb<-2-zbeta.wbchb


beta.wbct[1]~dnorm(0,1)
beta.wbct[2]<-0
zbeta.wbct~dcat(pi.wbct[])
pi.wbct[1]<-0.5
pi.wbct[2]<-1-pi.wbct[1]
zzbeta.wbct<-2-zbeta.wbct

beta.hbt[1]~dnorm(0,0.3)
beta.hbt[2]<-0
zbeta.hbt~dcat(pi.hbt[])
pi.hbt[1]<-0.5
pi.hbt[2]<-1-pi.hbt[1]
zzbeta.hbt<-2-zbeta.hbt

beta.hbage[1]~dnorm(0,0.5)
beta.hbage[2]<-0
zbeta.hbage~dcat(pi.hbage[])
pi.hbage[1]<-0.5
pi.hbage[2]<-1-pi.hbage[1]
zzbeta.hbage<-2-zbeta.hbage

beta.hbsex[1]~dnorm(0,0.2)
beta.hbsex[2]<-0
zbeta.hbsex~dcat(pi.hbsex[])
pi.hbsex[1]<-0.5
pi.hbsex[2]<-1-pi.hbsex[1]
```

```
zzbeta.hbsex<-2-zbeta.hbsex

beta.tage[1]~dnorm(0,0.2)
beta.tage[2]<-0
zbeta.tage~dcat(pi.tage[])
pi.tage[1]<-0.5
pi.tage[2]<-1-pi.tage[1]
zzbeta.tage<-2-zbeta.tage

beta.tsex[1]~dnorm(0,1)
beta.tsex[2]<-0
zbeta.tsex~dcat(pi.tsex[])
pi.tsex[1]<-0.5
pi.tsex[2]<-1-pi.tsex[1]
zzbeta.tsex<-2-zbeta.tsex

beta.wbcsex[1]~dnorm(0,0.2)
beta.wbcsex[2]<-0
zbeta.wbcsex~dcat(pi.wbcsex[])
pi.wbcsex[1]<-0.5
pi.wbcsex[2]<-1-pi.wbcsex[1]
zzbeta.wbcsex<-2-zbeta.wbcsex

}
```

A.4.2 R code to select the most likely configuration using arc deletion method

```
narcs<-9
N<-2^narcs

pick<-matrix(nrow=N,ncol=narcs)
m<-N
for(arc in 1:narcs)
{
```

```
  m<-m/2
  pick[,arc]<-rep(rep(1:2,c(m,m)),2^(arc-1))
}

n.iters<-40000
z1<-matrixout
z2<-1-z1
z<-c(z1,z2)
dim(z)<-c(n.iters,narcs,2)
pmat<-matrix(nrow=n.iters,ncol=N)
for( c in 1:N)
{pmat[,c]<-1
for( arc in 1:narcs){
  pmat[,c]<-pmat[,c]*z[,arc,pick[c,arc]]
}
}
q<-colMeans(pmat)
sum(q)
```

A.5 Appendix to Chapter 5

A.5.1 R function to generate samples from the posterior distribution of α and λ using Metropolis-Hastings algorithm

```
function(t,N,n,prior,sigma,show=TRUE,start=list(alpha=2,lambda=1))
{
  n<-length(t)
  result<-matrix(ncol=2,nrow=N)
  colnames(result)<-c("alpha","lambda")
  a<-prior$a;b<-prior$b;r<-prior$r;s<-prior$s
  d<-sum(delta)
  alpha<-start$alpha
  lambda<-start$lambda
  naccept<-0
```

```
  for(i in 1:N){
    lambda<-rgamma(1,d+r,s+sum(t^2))
    proposal<-rnorm(1,alpha,sigma)
    if(proposal>0){
      logprob<-(a+d-1)*log(proposal/alpha)-b*(proposal-alpha)-
        lambda*(sum(t^proposal)-(sum(t^alpha)))+(proposal-alpha)*sum(delta*log(t))
      u<-runif(1)
      if(log(u)<logprob){
        alpha<-proposal
        naccept<-naccept+1
      }
    }
    result[i,]<-c(alpha,lambda)
  }
  if(show==T)
    message(paste("Acceptance rate=",naccept/N))
  return(result)
}
```

A.5.2 Rjags model specification to fit the exponential survival time with the leukemia data

Setting initial values

```
is.na(t)<-leuk$status==0
is.censored<-1-leuk$status
t.cen<-leuk$t.cen

tinits1<-leuk$t.cen+5
is.na(tinits1)<-leuk$status==1
tinits2<-tinits1+5
```

Model specification

```
model{
for(i in 1:1043)
```

```
{
  is.censored[i]~dinterval(t[i],t.cen[i])
  t[i]~dexp(lambda[i])
  lambda[i]<-exp(beta0+beta.age*(age[i]-mean(age[]))+beta.sex*sex[i]
+beta.wbc*(wbc[i]-8)+beta.depscore*(depscore[i]-mean(depscore[])))
}
beta0~dnorm(-6.90,69.44)
beta.age~dnorm(0.04,1111.111)
beta.sex~dnorm(0.05,44.44)
beta.wbc~dnorm(0.08,33.41)
beta.depscore~dnorm(0.12,82.64)
}
```

A.5.3 Rjags model specification to calculate the survival probability for a new patient

```
model{
lambda.star<-exp(beta0+beta.age*63+beta.sex+beta.wbc*6.8+beta.depscore*2.02)
for(i in 1:1043)
{
is.censored[i]~dinterval(t[i],t.cen[i])
t[i]~dexp(lambda[i])
lambda[i]<-exp(beta0+beta.age*age[i]+beta.sex*sex[i]+beta.wbc*wbc[i]+
beta.depscore*depscore[i])
}
beta0~dnorm(0.0,10)
beta.age~dnorm(0.0,0.1)
beta.sex~dnorm(0.1,10)
beta.wbc~dnorm(0.1,20)
beta.depscore~dnorm(0.0,20)
}
```

Explanation of calculation of the survival probability

Suppose we consider a new patient. That means we are interested in plotting the survival function for one particular patient, say, for example, male and age 63, etc. Therefore,

we have λ for that patient denoted λ^*. If the posterior median for λ^* is λ^*_m and the lower and upper limits of the 95% interval are $\lambda^*_{0.025}$ and $\lambda^*_{0.975}$ respectively, then the corresponding quantities for the survival probability at time t are $\exp(-\lambda^*_m t), \exp(-\lambda^*_{0.025} t)$ and $\exp(-\lambda^*_{0.975} t)$.

As a result, we have

```
u<-exp(-0.002429*t)
l<-exp(-0.001996*t)
m<-(l+u)/2

plot(t,m,type="l",col=4,ylab="Survival probability",xlab="Time")
lines(t,l,type="l",col=2,lty=2)
lines(t,u,type="l",col=2,lty=2)
```

A.6 Appendix to Chapter 6

A.6.1 R functions to use Bayes linear approach

```
function(E_y,E_beta,V_yy,V_beta,C_betay,y){
  dif<-y-E_y
  C_ybeta<-t(C_betay)
  g<-solve(V_yy,C_ybeta)
  e1<-E_beta+t(g)%*%dif
  d<-C_betay%*%g
  v1<-V_beta-d
  return(list(e1=e1,v1=v1))}



first<-function(E_beta,V_beta,y,sigma){
  n<-length(y)
  E_y<-X%*%E_beta
  C_betay<-V_beta%*%t(X)
  V_yy<-X%*%C_betay+diag(sigma,n)
  result<-simplefun(E_y,E_beta,V_yy,V_beta,C_betay,y)
```

```
  return(result)}
```

A.6.2 R function for sulfinpyrazone example using logits

```
function(theta1,theta2,n,x,prior)
{# Evaluates posterior density for logit example.
  # prior is mean1, mean2, sd1, sd2, correlation
  n1<-length(theta1)
  n2<-length(theta2)
  step1<-theta1[2]-theta1[1]
  step2<-theta2[2]-theta2[1]
  theta1<-matrix(theta1,nrow=n1,ncol=n2)
  theta2<-matrix(theta2,nrow=n1,ncol=n2,byrow=T)
  eta1<-log(theta1/(1-theta1))
  eta2<-log(theta2/(1-theta2))
  delta1<-(eta1-prior[1])/prior[3]
  delta2<-(eta2-prior[2])/prior[4]
  r<-prior[5]
  d<-1-r^2
  logprior<- -(delta1^2 + delta2^2 - 2*r*delta1*delta2)/(2*d)
  J<-theta1*(1-theta1)*theta2*(1-theta2)
  logprior<-logprior-log(J)
  loglik<-x[1]*log(theta1)+(n[1]-x[1])*log(1-theta1)
  +x[2]*log(theta2)+(n[2]-x[2])*log(1-theta2)
  logpos<-logprior+loglik
  logpos<-logpos-max(logpos)
  posterior<-exp(logpos)
  int<-sum(posterior)*step1*step2
  posterior<-posterior/int
  e2<-sum(eta1^2*posterior)*step1*step2
  ans<-list(density=posterior,int=int,e2=e2)
  ans
}
```

A.6.3 R function to find the posterior mean and variance for η_1 and η_2 in sulfinpyrazone example

```
function(eta1,eta2,n,x,prior)
{
  n1<-length(eta1)
  n2<-length(eta2)
  step1<-eta1[2]-eta1[1]
  step2<-eta2[2]-eta2[1]
  eta1<-matrix(eta1,nrow=n1,ncol=n2)
  eta2<-matrix(eta2,nrow=n1,ncol=n2,byrow=T)
  theta1<-exp(eta1/(1+eta1))
  theta2<-exp(eta2/(1+theta2))
  delta1<-(eta1-prior[1])/prior[3]
  delta2<-(eta2-prior[2])/prior[4]
  r<-prior[5]
  d<-1-r^2
  logprior<- -(delta1^2 + delta2^2 - 2*r*delta1*delta2)/(2*d)
  J<-exp(eta1)/(1+exp(eta1))^2*exp(eta2)/(1+exp(eta2))^2
  logprior<-logprior-log(J)
  loglik<-x[1]*log(exp(eta1)/(1+exp(eta1)))+(x[1]-
  n[1])*log(1+exp(eta1))
  logpos<-logprior+loglik
  logpos<-logpos-max(logpos)
  posterior<-exp(logpos)
  int<-sum(posterior)*step1*step2
  posterior<-posterior/int
  e2<-sum(eta1*posterior)*step1*step2
  ans<-list(density=posterior,int=int,e2=e2)
  ans
}
```

A.6.4 Posterior correlation matrix for η for both areas and sexes in surgical death example using Bayes linear kinematic

See Table A.3–A.7.

A.6.5 R function to make adjustment for both binary and ordinal variables in the direct method

```
function(y,E0,V0,cuts,nstep=100){
  S0<-sqrt(V0)
  k<-length(cuts)
  if(y==0){
    lower<-0 }
  else
  {lower<-pnorm(cuts[y],E0,S0)}
  if(y==k)
  {upper<-1}
  else {
    upper<-pnorm(cuts[y+1],E0,S0)}
  u<-seq(lower,upper,length.out=(nstep+1))
  Z<-qnorm(u,E0,S0)
  Z2=Z*Z
  E1<-(sum(Z)-(Z[1]+Z[nstep+1])/2)/nstep
  E1Z2<-(sum(Z2)-(Z2[1]+Z2[nstep+1])/2)/nstep
  V1<-E1Z2-E1*E1
  ans<-list(E=E1,V=V1)
  return(ans)
}
```

	1	2	3	4	5	6	7	8
1	1.0000	0.2246	0.0742	0.0332	0.0197	0.0082	0.0024	0.0008
2	0.2246	1.0000	0.2088	0.0662	0.0352	0.0140	0.0039	0.0010
3	0.0742	0.2088	1.0000	0.1726	0.0653	0.0244	0.0070	0.0016
4	0.0332	0.0662	0.1726	1.0000	0.1917	0.0564	0.0152	0.0034
5	0.0197	0.0352	0.0653	0.1917	1.0000	0.1701	0.0408	0.0090
6	0.0082	0.0140	0.0244	0.0564	0.1701	1.0000	0.1204	0.0236
7	0.0024	0.0039	0.0070	0.0152	0.0408	0.1204	1.0000	0.1023
8	0.0008	0.0010	0.0016	0.0034	0.0090	0.0236	0.1023	1.0000
9	0.0010	0.0007	0.0007	0.0011	0.0025	0.0057	0.0224	0.1334
10	0.0017	0.0010	0.0007	0.0008	0.0012	0.0023	0.0081	0.0463
11	0.1330	-0.0840	-0.0377	-0.0264	-0.0175	-0.0071	-0.0015	-0.0005
12	-0.0779	0.3824	-0.0001	-0.0403	-0.0305	-0.0127	-0.0026	-0.0007
13	-0.0735	-0.0643	0.2788	-0.0442	-0.0564	-0.0243	-0.0051	-0.0012
14	-0.0366	-0.0517	-0.0137	0.2466	-0.0548	-0.0366	-0.0083	-0.0018
15	-0.0157	-0.0231	-0.0174	-0.0173	0.2417	-0.0434	-0.0128	-0.0027
16	-0.0068	-0.0102	-0.0087	-0.0260	-0.0533	0.1307	-0.0433	-0.0117
17	-0.0019	-0.0027	-0.0016	-0.0066	-0.0173	-0.0484	0.1057	-0.0388
18	-0.0006	-0.0006	-0.0002	-0.0012	-0.0030	-0.0112	-0.0317	0.0660
19	-0.0005	-0.0003	-0.0001	-0.0002	-0.0004	-0.0019	-0.0059	-0.0321
20	-0.0005	-0.0004	-0.0001	-0.0001	-0.0002	-0.0007	-0.0020	-0.0118
21	0.1264	-0.0321	-0.0621	-0.0255	-0.0119	-0.0060	-0.0022	-0.0006
22	-0.0312	0.3860	-0.0562	-0.0415	-0.0220	-0.0113	-0.0040	-0.0009
23	-0.0465	-0.0225	0.2157	-0.0420	-0.0423	-0.0216	-0.0070	-0.0015
24	-0.0218	-0.0297	-0.0530	0.2459	-0.0349	-0.0333	-0.0114	-0.0022
25	-0.0075	-0.0110	-0.0387	-0.0148	0.2371	-0.0331	-0.0170	-0.0032
26	-0.0035	-0.0055	-0.0214	-0.0258	-0.0361	0.1340	-0.0493	-0.0114
27	-0.0011	-0.0015	-0.0065	-0.0076	-0.0135	-0.0400	0.1001	-0.0400
28	-0.0003	-0.0003	-0.0013	-0.0013	-0.0017	-0.0080	-0.0345	0.0625
29	-0.0003	-0.0002	-0.0005	-0.0002	0.0000	-0.0012	-0.0065	-0.0364
30	-0.0002	-0.0002	-0.0005	-0.0002	-0.0001	-0.0005	-0.0024	-0.0145
31	0.1123	-0.0143	-0.0196	-0.0233	-0.0317	-0.0100	-0.0045	-0.0015
32	0.0058	0.2251	0.0041	-0.0285	-0.0416	-0.0134	-0.0058	-0.0019
33	-0.0166	-0.0076	0.1837	-0.0257	-0.0603	-0.0204	-0.0083	-0.0025
34	-0.0105	-0.0211	0.0002	0.2095	-0.0657	-0.0304	-0.0127	-0.0037
35	-0.0029	-0.0098	-0.0099	-0.0053	0.2079	-0.0287	-0.0202	-0.0060
36	-0.0010	-0.0045	-0.0057	-0.0208	-0.0606	0.1324	-0.0427	-0.0139
37	-0.0002	-0.0015	-0.0015	-0.0078	-0.0279	-0.0330	0.1008	-0.0381
38	0.0000	-0.0004	-0.0003	-0.0019	-0.0074	-0.0093	-0.0322	0.0696
39	0.0000	-0.0002	-0.0001	-0.0005	-0.0019	-0.0020	-0.0078	-0.0304
40	0.0002	-0.0001	-0.0001	-0.0004	-0.0011	-0.0009	-0.0029	-0.0123

Table A.3: Posterior correlation matrix for η for both areas and sexes in surgical death example using Bayes linear kinematic and for the columns $1-8$

	9	10	11	12	13	14	15	16
1	0.0010	0.0017	0.1330	-0.0779	-0.0735	-0.0366	-0.0157	-0.0068
2	0.0007	0.0010	-0.0840	0.3824	-0.0643	-0.0517	-0.0231	-0.0102
3	0.0007	0.0007	-0.0377	-0.0001	0.2788	-0.0137	-0.0174	-0.0087
4	0.0011	0.0008	-0.0264	-0.0403	-0.0442	0.2466	-0.0173	-0.0260
5	0.0025	0.0012	-0.0175	-0.0305	-0.0564	-0.0548	0.2417	-0.0533
6	0.0057	0.0023	-0.0071	-0.0127	-0.0243	-0.0366	-0.0434	0.1307
7	0.0224	0.0081	-0.0015	-0.0026	-0.0051	-0.0083	-0.0128	-0.0433
8	0.1334	0.0463	-0.0005	-0.0007	-0.0012	-0.0018	-0.0027	-0.0117
9	1.0000	0.2145	-0.0007	-0.0004	-0.0005	-0.0005	-0.0005	-0.0027
10	0.2145	1.0000	-0.0010	-0.0006	-0.0005	-0.0004	-0.0004	-0.0012
11	-0.0007	-0.0010	1.0000	0.3101	0.1676	0.0790	0.0386	0.0176
12	-0.0004	-0.0006	0.3101	1.0000	0.3477	0.1474	0.0700	0.0312
13	-0.0005	-0.0005	0.1676	0.3477	1.0000	0.3009	0.1326	0.0581
14	-0.0005	-0.0004	0.0790	0.1474	0.3009	1.0000	0.2249	0.0867
15	-0.0005	-0.0004	0.0386	0.0700	0.1326	0.2249	1.0000	0.1858
16	-0.0027	-0.0012	0.0176	0.0312	0.0581	0.0867	0.1858	1.0000
17	-0.0098	-0.0040	0.0059	0.0103	0.0189	0.0274	0.0547	0.1586
18	-0.0450	-0.0192	0.0017	0.0027	0.0047	0.0068	0.0135	0.0362
19	0.0939	-0.0549	0.0012	0.0011	0.0015	0.0018	0.0033	0.0078
20	-0.0429	0.2034	0.0020	0.0014	0.0014	0.0012	0.0014	0.0027
21	-0.0006	-0.0009	0.1272	-0.0296	-0.0655	-0.0347	-0.0200	-0.0106
22	-0.0006	-0.0008	-0.0299	0.3519	-0.0598	-0.0486	-0.0305	-0.0169
23	-0.0005	-0.0006	-0.0178	0.0285	0.2380	0.0067	-0.0186	-0.0141
24	-0.0005	-0.0005	-0.0200	-0.0236	-0.0260	0.2264	-0.0075	-0.0243
25	-0.0004	-0.0004	-0.0162	-0.0245	-0.0475	-0.0228	0.2045	-0.0370
26	-0.0019	-0.0010	-0.0089	-0.0139	-0.0278	-0.0280	-0.0322	0.1321
27	-0.0079	-0.0036	-0.0028	-0.0043	-0.0088	-0.0090	-0.0135	-0.0335
28	-0.0386	-0.0184	-0.0008	-0.0011	-0.0023	-0.0023	-0.0034	-0.0112
29	0.0992	-0.0465	-0.0005	-0.0005	-0.0008	-0.0006	-0.0007	-0.0026
30	-0.0337	0.1927	-0.0006	-0.0006	-0.0008	-0.0005	-0.0005	-0.0012
31	-0.0013	-0.0008	0.1198	-0.0020	-0.0166	-0.0230	-0.0324	-0.0126
32	-0.0013	-0.0008	0.0155	0.2199	0.0093	-0.0229	-0.0394	-0.0154
33	-0.0015	-0.0008	0.0116	0.0408	0.2069	0.0167	-0.0358	-0.0137
34	-0.0018	-0.0007	0.0000	-0.0013	0.0313	0.2038	-0.0329	-0.0208
35	-0.0025	-0.0008	-0.0055	-0.0130	-0.0122	-0.0065	0.1836	-0.0306
36	-0.0051	-0.0014	-0.0030	-0.0074	-0.0086	-0.0194	-0.0514	0.1321
37	-0.0144	-0.0038	-0.0010	-0.0027	-0.0029	-0.0079	-0.0254	-0.0259
38	-0.0512	-0.0139	-0.0004	-0.0010	-0.0011	-0.0028	-0.0084	-0.0122
39	0.0850	-0.0241	-0.0002	-0.0004	-0.0004	-0.0009	-0.0024	-0.0036
40	-0.0467	0.2231	-0.0002	-0.0004	-0.0004	-0.0006	-0.0014	-0.0017

Table A.4: Posterior correlation matrix for η for both areas and sexes in surgical death example using Bayes linear kinematic and for the columns $9-16$

	17	18	19	20	21	22	23	24
1	-0.0019	-0.0006	-0.0005	-0.0005	0.1264	-0.0312	-0.0465	-0.0218
2	-0.0027	-0.0006	-0.0003	-0.0004	-0.0321	0.3860	-0.0225	-0.0297
3	-0.0016	-0.0002	-0.0001	-0.0001	-0.0621	-0.0562	0.2157	-0.0530
4	-0.0066	-0.0012	-0.0002	-0.0001	-0.0255	-0.0415	-0.0420	0.2459
5	-0.0173	-0.0030	-0.0004	-0.0002	-0.0119	-0.0220	-0.0423	-0.0349
6	-0.0484	-0.0112	-0.0019	-0.0007	-0.0060	-0.0113	-0.0216	-0.0333
7	0.1057	-0.0317	-0.0059	-0.0020	-0.0022	-0.0040	-0.0070	-0.0114
8	-0.0388	0.0660	-0.0321	-0.0118	-0.0006	-0.0009	-0.0015	-0.0022
9	-0.0098	-0.0450	0.0939	-0.0429	-0.0006	-0.0006	-0.0005	-0.0005
10	-0.0040	-0.0192	-0.0549	0.2034	-0.0009	-0.0008	-0.0006	-0.0005
11	0.0059	0.0017	0.0012	0.0020	0.1272	-0.0299	-0.0178	-0.0200
12	0.0103	0.0027	0.0011	0.0014	-0.0296	0.3519	0.0285	-0.0236
13	0.0189	0.0047	0.0015	0.0014	-0.0655	-0.0598	0.2380	-0.0260
14	0.0274	0.0068	0.0018	0.0012	-0.0347	-0.0486	0.0067	0.2264
15	0.0547	0.0135	0.0033	0.0014	-0.0200	-0.0305	-0.0186	-0.0075
16	0.1586	0.0362	0.0078	0.0027	-0.0106	-0.0169	-0.0141	-0.0243
17	1.0000	0.1306	0.0264	0.0084	-0.0041	-0.0065	-0.0062	-0.0110
18	0.1306	1.0000	0.1277	0.0386	-0.0012	-0.0018	-0.0017	-0.0028
19	0.0264	0.1277	1.0000	0.1790	-0.0006	-0.0007	-0.0004	-0.0006
20	0.0084	0.0386	0.1790	1.0000	-0.0009	-0.0009	-0.0005	-0.0005
21	-0.0041	-0.0012	-0.0006	-0.0009	1.0000	0.3729	0.1715	0.0907
22	-0.0065	-0.0018	-0.0007	-0.0009	0.3729	1.0000	0.3524	0.1684
23	-0.0062	-0.0017	-0.0004	-0.0005	0.1715	0.3524	1.0000	0.2607
24	-0.0110	-0.0028	-0.0006	-0.0005	0.0907	0.1684	0.2607	1.0000
25	-0.0210	-0.0052	-0.0008	-0.0006	0.0493	0.0891	0.1215	0.2390
26	-0.0481	-0.0144	-0.0026	-0.0012	0.0250	0.0446	0.0591	0.1064
27	0.1064	-0.0345	-0.0069	-0.0031	0.0082	0.0145	0.0189	0.0336
28	-0.0358	0.0660	-0.0324	-0.0157	0.0023	0.0037	0.0044	0.0077
29	-0.0094	-0.0432	0.0974	-0.0570	0.0017	0.0019	0.0017	0.0023
30	-0.0040	-0.0190	-0.0554	0.2044	0.0025	0.0024	0.0016	0.0015
31	-0.0060	-0.0021	-0.0013	-0.0007	0.1172	0.0353	0.0053	-0.0130
32	-0.0075	-0.0026	-0.0014	-0.0008	0.0391	0.2332	0.0341	-0.0123
33	-0.0077	-0.0029	-0.0014	-0.0006	-0.0029	0.0129	0.1712	-0.0081
34	-0.0120	-0.0043	-0.0018	-0.0006	-0.0023	-0.0052	0.0265	0.1977
35	-0.0237	-0.0081	-0.0029	-0.0009	-0.0026	-0.0078	-0.0040	0.0068
36	-0.0406	-0.0162	-0.0055	-0.0014	-0.0028	-0.0070	-0.0078	-0.0178
37	0.1078	-0.0319	-0.0129	-0.0027	-0.0020	-0.0044	-0.0054	-0.0116
38	-0.0329	0.0743	-0.0444	-0.0099	-0.0007	-0.0014	-0.0018	-0.0037
39	-0.0107	-0.0369	0.0816	-0.0324	-0.0002	-0.0005	-0.0005	-0.0010
40	-0.0045	-0.0168	-0.0713	0.2374	-0.0001	-0.0004	-0.0004	-0.0007

Table A.5: Posterior correlation matrix for η for both areas and sexes in surgical death example using Bayes linear kinematic and for the columns $17-24$

	25	26	27	28	29	30	31	32
1	-0.0075	-0.0035	-0.0011	-0.0003	-0.0003	-0.0002	0.1123	0.0058
2	-0.0110	-0.0055	-0.0015	-0.0003	-0.0002	-0.0002	-0.0143	0.2251
3	-0.0387	-0.0214	-0.0065	-0.0013	-0.0005	-0.0005	-0.0196	0.0041
4	-0.0148	-0.0258	-0.0076	-0.0013	-0.0002	-0.0002	-0.0233	-0.0285
5	0.2371	-0.0361	-0.0135	-0.0017	0.0000	-0.0001	-0.0317	-0.0416
6	-0.0331	0.1340	-0.0400	-0.0080	-0.0012	-0.0005	-0.0100	-0.0134
7	-0.0170	-0.0493	0.1001	-0.0345	-0.0065	-0.0024	-0.0045	-0.0058
8	-0.0032	-0.0114	-0.0400	0.0625	-0.0364	-0.0145	-0.0015	-0.0019
9	-0.0004	-0.0019	-0.0079	-0.0386	0.0992	-0.0337	-0.0013	-0.0013
10	-0.0004	-0.0010	-0.0036	-0.0184	-0.0465	0.1927	-0.0008	-0.0008
11	-0.0162	-0.0089	-0.0028	-0.0008	-0.0005	-0.0006	0.1198	0.0155
12	-0.0245	-0.0139	-0.0043	-0.0011	-0.0005	-0.0006	-0.0020	0.2199
13	-0.0475	-0.0278	-0.0088	-0.0023	-0.0008	-0.0008	-0.0166	0.0093
14	-0.0228	-0.0280	-0.0090	-0.0023	-0.0006	-0.0005	-0.0230	-0.0229
15	0.2045	-0.0322	-0.0135	-0.0034	-0.0007	-0.0005	-0.0324	-0.0394
16	-0.0370	0.1321	-0.0335	-0.0112	-0.0026	-0.0012	-0.0126	-0.0154
17	-0.0210	-0.0481	0.1064	-0.0358	-0.0094	-0.0040	-0.0060	-0.0075
18	-0.0052	-0.0144	-0.0345	0.0660	-0.0432	-0.0190	-0.0021	-0.0026
19	-0.0008	-0.0026	-0.0069	-0.0324	0.0974	-0.0554	-0.0013	-0.0014
20	-0.0006	-0.0012	-0.0031	-0.0157	-0.0570	0.2044	-0.0007	-0.0008
21	0.0493	0.0250	0.0082	0.0023	0.0017	0.0025	0.1172	0.0391
22	0.0891	0.0446	0.0145	0.0037	0.0019	0.0024	0.0353	0.2332
23	0.1215	0.0591	0.0189	0.0044	0.0017	0.0016	0.0053	0.0341
24	0.2390	0.1064	0.0336	0.0077	0.0023	0.0015	-0.0130	-0.0123
25	1.0000	0.2124	0.0645	0.0148	0.0039	0.0019	-0.0260	-0.0325
26	0.2124	1.0000	0.1661	0.0354	0.0084	0.0033	-0.0116	-0.0146
27	0.0645	0.1661	1.0000	0.1206	0.0267	0.0092	-0.0061	-0.0077
28	0.0148	0.0354	0.1206	1.0000	0.1368	0.0451	-0.0020	-0.0025
29	0.0039	0.0084	0.0267	0.1368	1.0000	0.2041	-0.0012	-0.0013
30	0.0019	0.0033	0.0092	0.0451	0.2041	1.0000	-0.0005	-0.0006
31	-0.0260	-0.0116	-0.0061	-0.0020	-0.0012	-0.0005	1.0000	0.3297
32	-0.0325	-0.0146	-0.0077	-0.0025	-0.0013	-0.0006	0.3297	1.0000
33	-0.0478	-0.0227	-0.0113	-0.0035	-0.0017	-0.0008	0.2494	0.3341
34	-0.0328	-0.0230	-0.0135	-0.0044	-0.0019	-0.0008	0.1801	0.2334
35	0.1801	-0.0172	-0.0199	-0.0067	-0.0026	-0.0009	0.1129	0.1447
36	-0.0414	0.1344	-0.0326	-0.0133	-0.0050	-0.0014	0.0591	0.0753
37	-0.0285	-0.0325	0.1027	-0.0350	-0.0140	-0.0037	0.0245	0.0311
38	-0.0090	-0.0127	-0.0349	0.0697	-0.0497	-0.0137	0.0079	0.0097
39	-0.0023	-0.0031	-0.0091	-0.0313	0.0864	-0.0245	0.0037	0.0040
40	-0.0014	-0.0016	-0.0042	-0.0166	-0.0621	0.2242	0.0048	0.0046

Table A.6: Posterior correlation matrix for η for both areas and sexes in surgical death example using Bayes linear kinematic and for the columns $25-32$

	33	34	35	36	37	38	39	40
1	-0.0166	-0.0105	-0.0029	-0.0010	-0.0002	0.0000	0.0000	0.0002
2	-0.0076	-0.0211	-0.0098	-0.0045	-0.0015	-0.0004	-0.0002	-0.0001
3	0.1837	0.0002	-0.0099	-0.0057	-0.0015	-0.0003	-0.0001	-0.0001
4	-0.0257	0.2095	-0.0053	-0.0208	-0.0078	-0.0019	-0.0005	-0.0004
5	-0.0603	-0.0657	0.2079	-0.0606	-0.0279	-0.0074	-0.0019	-0.0011
6	-0.0204	-0.0304	-0.0287	0.1324	-0.0330	-0.0093	-0.0020	-0.0009
7	-0.0083	-0.0127	-0.0202	-0.0427	0.1008	-0.0322	-0.0078	-0.0029
8	-0.0025	-0.0037	-0.0060	-0.0139	-0.0381	0.0696	-0.0304	-0.0123
9	-0.0015	-0.0018	-0.0025	-0.0051	-0.0144	-0.0512	0.0850	-0.0467
10	-0.0008	-0.0007	-0.0008	-0.0014	-0.0038	-0.0139	-0.0241	0.2231
11	0.0116	0.0000	-0.0055	-0.0030	-0.0010	-0.0004	-0.0002	-0.0002
12	0.0408	-0.0013	-0.0130	-0.0074	-0.0027	-0.0010	-0.0004	-0.0004
13	0.2069	0.0313	-0.0122	-0.0086	-0.0029	-0.0011	-0.0004	-0.0004
14	0.0167	0.2038	-0.0065	-0.0194	-0.0079	-0.0028	-0.0009	-0.0006
15	-0.0358	-0.0329	0.1836	-0.0514	-0.0254	-0.0084	-0.0024	-0.0014
16	-0.0137	-0.0208	-0.0306	0.1321	-0.0259	-0.0122	-0.0036	-0.0017
17	-0.0077	-0.0120	-0.0237	-0.0406	0.1078	-0.0329	-0.0107	-0.0045
18	-0.0029	-0.0043	-0.0081	-0.0162	-0.0319	0.0743	-0.0369	-0.0168
19	-0.0014	-0.0018	-0.0029	-0.0055	-0.0129	-0.0444	0.0816	-0.0713
20	-0.0006	-0.0006	-0.0009	-0.0014	-0.0027	-0.0099	-0.0324	0.2374
21	-0.0029	-0.0023	-0.0026	-0.0028	-0.0020	-0.0007	-0.0002	-0.0001
22	0.0129	-0.0052	-0.0078	-0.0070	-0.0044	-0.0014	-0.0005	-0.0004
23	0.1712	0.0265	-0.0040	-0.0078	-0.0054	-0.0018	-0.0005	-0.0004
24	-0.0081	0.1977	0.0068	-0.0178	-0.0116	-0.0037	-0.0010	-0.0007
25	-0.0478	-0.0328	0.1801	-0.0414	-0.0285	-0.0090	-0.0023	-0.0014
26	-0.0227	-0.0230	-0.0172	0.1344	-0.0325	-0.0127	-0.0031	-0.0016
27	-0.0113	-0.0135	-0.0199	-0.0326	0.1027	-0.0349	-0.0091	-0.0042
28	-0.0035	-0.0044	-0.0067	-0.0133	-0.0350	0.0697	-0.0313	-0.0166
29	-0.0017	-0.0019	-0.0026	-0.0050	-0.0140	-0.0497	0.0864	-0.0621
30	-0.0008	-0.0008	-0.0009	-0.0014	-0.0037	-0.0137	-0.0245	0.2242
31	0.2494	0.1801	0.1129	0.0591	0.0245	0.0079	0.0037	0.0048
32	0.3341	0.2334	0.1447	0.0753	0.0311	0.0097	0.0040	0.0046
33	1.0000	0.3250	0.1931	0.0995	0.0409	0.0125	0.0044	0.0043
34	0.3250	1.0000	0.2965	0.1426	0.0578	0.0173	0.0053	0.0042
35	0.1931	0.2965	1.0000	0.2554	0.0988	0.0290	0.0078	0.0046
36	0.0995	0.1426	0.2554	1.0000	0.2152	0.0613	0.0153	0.0071
37	0.0409	0.0578	0.0988	0.2152	1.0000	0.1702	0.0405	0.0167
38	0.0125	0.0173	0.0290	0.0613	0.1702	1.0000	0.1586	0.0621
39	0.0044	0.0053	0.0078	0.0153	0.0405	0.1586	1.0000	0.2440
40	0.0043	0.0042	0.0046	0.0071	0.0167	0.0621	0.2440	1.0000

Table A.7: Posterior correlation matrix for η for both areas and sexes in surgical death example using Bayes linear kinematic and for the columns $33-40$

A.6.6 R function to make adjustment for both binary and ordinal variables in the indirect method

```
function(y,E0,V0,cuts,tol=1E-5,n=20)
{#
  # y is an ordinal variable with possible values 1,2,...,m.
  # If m=2 then we must set V0=1.
  m<-length(cuts)+1
  # Initial approximation
  z0<-E0
  d<-1
  for (i in 1:n)
  {if (abs(d)>tol)
  {if (y==m)
  {puz<-1}
    else
    {euz<-exp(cuts[y]-z0)
    puz<-euz/(1+euz)
    }
    if (y==1)
    {plz<-0}
    else
    {elz<-exp(cuts[y-1]-z0)
    plz<-elz/(1+elz)
    }
    f<-(puz+plz)-1-(z0-E0)/V0
    fd<--(1/V0+puz*(1-puz)+plz*(1-plz))
    d<-f/fd
    z0<-z0-d
  }
  }
  vpost0<- -1/fd
  # Tierney and Kadane normalising constant (denominator)
  if (y==m)
  {puz<-1}
```

```
else
{euz<-exp(cuts[y]-z0)
puz<-euz/(1+euz)
}
if (y==1)
{plz<-0}
else
{elz<-exp(cuts[y-1]-z0)
plz<-elz/(1+elz)
}
g0<-log(puz-plz)-((z0-E0)^2)/(2*V0)
eta0<-exp(g0)*sqrt(2*pi*vpost0)
# Mean shift
C<-5*sqrt(vpost0)-z0
# Tierney and Kadane numerator for E(Z+C)
z1<-z0
d<-1
for (i in 1:n)
{if (abs(d)>tol)
{f<-(puz+plz)-1-(z1-E0)/V0+1/(z1+C)
fd<--(1/V0+puz*(1-puz)+plz*(1-plz)+(z1+C)^(-2))
d<-f/fd
z1<-z1-d
if (y==m)
{puz<-1}
else
{euz<-exp(cuts[y]-z1)
puz<-euz/(1+euz)
}
if (y==1)
{plz<-0}
else
{elz<-exp(cuts[y-1]-z1)
plz<-elz/(1+elz)
}
```

```
}
}
v1<- -1/fd
g1<-log(z1+C)+log(puz-plz)-((z1-E0)^2)/(2*V0)
eta1<-exp(g1)*sqrt(2*pi*v1)
# Tierney and Kadane approximation to E(Z+C) and E(Z)
E1<-eta1/eta0
postmean<-E1-C
# Tierney and Kadane numerator for E([Z+C]^2)
z2<-z1
d<-1
for (i in 1:n)
{if (abs(d)>tol)
{f<-(puz+plz)-1-(z2-E0)/V0+2/(z2+C)
fd<--(1/V0+puz*(1-puz)+plz*(1-plz)+2*(z2+C)^(-2))
d<-f/fd
z2<-z2-d
if (y==m)
{puz<-1}
else
{euz<-exp(cuts[y]-z2)
puz<-euz/(1+euz)
}
if (y==1)
{plz<-0}
else
{elz<-exp(cuts[y-1]-z2)
plz<-elz/(1+elz)
}
}
}
v2<- -1/fd
g2<-2*log(z2+C)+log(puz-plz)-((z2-E0)^2)/(2*V0)
eta2<-exp(g2)*sqrt(2*pi*v2)
# Tierney and Kadane approximation to E([Z+C]^2) and Var(Z)
```

```
  E2<-eta2/eta0
  postvar<-E2-E1^2
  return(list(E1=postmean,V1=postvar))
}
```

A.7 Appendix to Chapter 7

A.7.1 Rjags model specification for the leukaemia data using a piecewise constant hazards model

```
model
{for (i in 1:n)
{is.censored[i]~dinterval(t[i],t.cen[i])
t[i]~dexp(lambda[i])
log(lambda[i])<-beta0[period[i]]+beta.age[period[i]]*(age[i]-60)
+beta.sex[period[i]]*sex[i]+beta.wbc[period[i]]*(wbc[i]-8)
+beta.depsc[period[i]]*depscore[i]
}

#Priors:

beta0[1]~dnorm(-6,p0)
beta.age[1]~dnorm(0.02,p.age)
beta.sex[1]~dnorm(0,p.sex)
beta.wbc[1]~dnorm(0.005,p.wbc)
beta.depsc[1]~dnorm(0,p.depsc)

p0<-1.5625
p.age<-2500
p.sex<-8.16
p.wbc<-40000
p.depsc<-100

rho<-0.92
```

```
rf<-1-rho*rho

p0.e<-p0/rf
p.age.e<-p.age/rf
p.sex.e<-p.sex/rf
p.wbc.e<-p.wbc/rf
p.depsc.e<-p.depsc/rf

for (j in 2:11)
{beta0m[j]<- -6 + rho*(beta0[j-1]+6)
beta0[j]~dnorm(beta0m[j],p0.e)
beta.agem[j]<-0.02 + rho*(beta.age[j-1]-0.02)
beta.age[j]~dnorm(beta.agem[j],p.age.e)
beta.sexm[j]<-rho*beta.sex[j-1]
beta.sex[j]~dnorm(beta.sexm[j],p.sex.e)
beta.wbcm[j]<-0.005+rho*(beta.wbc[j-1]-0.005)
beta.wbc[j]~dnorm(beta.wbcm[j],p.wbc.e)
beta.depscm[j]<-rho*beta.depsc[j-1]
beta.depsc[j]~dnorm(beta.depscm[j],p.depsc.e)
}

for (i in 1:n)
{a[i]~dnorm(0,1)
b[i]~dnorm(0,1)
}
}
```

A.7.2 R code to compute the posterior medians for residuals in leukaemia example

```
probs<-c(0.25,0.5,0.75)
residual<-matrix(nrow=n,ncol=3)
for (i in 1:n)
    {samples<-numeric(m)
     etamat<-matrix(nrow=m,ncol=10)
```

```
    u<-runif(1,0,1)
    for (j in 1:m)
        {eta<-etacalc(x[i,],beta[j,])
         etamat[j,]<-eta
         lambda<-exp(eta)
         if (censored[i]==0)
            {Fc<-ppch(t[i],lambda,cuts)
             samples[j]<-Fc+u*(1-Fc)
            }
         else
            {samples[j]<-ppch(t[i],lambda,cuts)
            }
         }
    residual[i,]<-quantile(samples,probs)
    }
```

A.7.3 R function ppch for finding the cdf of a piecewise constant hazard model

```
ppch<-function(t,lambda,cuts)
{nint<-length(cuts)
 tmax<-cuts[nint]+t
 cuts<-c(0,cuts,tmax)
 S<-1
 for (k in 1:nint)
     {if (t>cuts[k])
         {tk<-min(t,cuts[k+1])
          S<-S*exp(-lambda[k]*(tk-cuts[k]))
         }
      }
 F<-1-S
 return(F)
 }
```

A.7.4 Offline learning model for non-Hodgkin lymphoma data in the direct method

```
#############################################################################
#### Offline learning model for non-Hodgkin lymphoma data in the ####
############### direct method ##############
#############################################################################

model
{
  for (i in 1:1391){
  is.censored[i] ~ dinterval (t[i], t.cen[i])
  t[i]~dweib(alpha,lambda[i])
  log(lambda[i]) <- mean.Z.t[i]

  Z.t[i]~dnorm(mean.Z.t[i],tau.Z.t)
  mean.Z.t[i]<- gamma0.t+gamma.t.age*age[i]+gamma.t.wbc*(Z.wbc[i]-mean.Z.wbc[i])
  +gamma.t.sex*sex[i]+gamma.t.albumin*(Z.albumin[i]-mean.Z.albumin[i])
  +gamma.t.stage*(Z.stage[i]-mean.Z.stage[i])+gamma.t.hb*(Z.hb[i]-mean.Z.hb[i])

  hb[i]~dnorm(mean.Z.hb[i],tau.Z.hb)
  Z.hb[i]<-hb[i]
  mean.Z.hb[i]<-gamma0.hb+gamma.hbage*age[i]+gamma.hbsex*sex[i]

  wbc[i]~dnorm(mean.Z.wbc[i],tau.Z.wbc)
  Z.wbc[i]<-wbc[i]
  mean.Z.wbc[i]<-gamma0.wbc+gamma.wbcage*age[i]+gamma.wbcsex*sex[i]
  +gamma.wbchb*(Z.hb[i]-mean.Z.hb[i])

  stage[i]~dinterval(Z.stage[i],cut.stage[1:3])
  Z.stage[i] ~ dnorm(mean.Z.stage[i], tau.Z.stage)
  mean.Z.stage[i]<-gamma0.stage+gamma.stageage*age[i]+gamma.stagesex*sex[i]
  +gamma.stagehb*(Z.hb[i]-mean.Z.hb[i])
  +gamma.stagewbc*(Z.wbc[i]-mean.Z.wbc[i])
```

```
albumin[i]~dinterval(Z.albumin[i],0)
Z.albumin[i]~dnorm(mean.Z.albumin[i],1)
mean.Z.albumin[i]<-gamma0.albumin
+gamma.albuminage*age[i]+gamma.albuminsex*sex[i]
+gamma.albuminwbc*(Z.wbc[i]-mean.Z.wbc[i])
+gamma.albuminstage*(Z.stage[i]-mean.Z.stage[i])
+gamma.albuminhb*(Z.hb[i]-mean.Z.hb[i])

}

alpha ~ dgamma(4, 4)# prior for alpha
gamma0.t ~ dnorm(0, 0.01)
gamma.t.sex~dnorm(0,0.001)
gamma.t.age~dnorm(0.0,0.001)
gamma.t.wbc~dnorm(0.0,0.001)
gamma.t.albumin~dnorm(1,0.0001)
gamma.t.stage~dnorm(0.0,0.001)
gamma.t.hb~dnorm(0.0,0.001)

gamma0.hb~dnorm(100.0,0.0001)
gamma.hbage~dnorm(0.0,0.001)
gamma.hbsex~dnorm(0.0,0.001)

gamma0.wbc~dnorm(10.0,0.001)
gamma.wbcage~dnorm(0.0,0.001)
gamma.wbcsex~dnorm(0.0,0.001)
gamma.wbchb~dnorm(0.0,0.001)

gamma0.stage~dnorm(0.0,0.001)
gamma.stageage~dnorm(0.0,0.001)
gamma.stagesex~dnorm(0.0,0.001)
gamma.stagehb~dnorm(0.0,0.001)
gamma.stagewbc~dnorm(0.0,0.001)
```

```
  gamma0.albumin~dnorm(0,0.001)
  gamma.albuminage~dnorm(0,0.001)
  gamma.albuminsex~dnorm(0,0.001)
  gamma.albuminwbc~dnorm(0,0.001)
  gamma.albuminstage~dnorm(0,0.001)
  gamma.albuminhb~dnorm(0,0.001)

  tau.Z.wbc~dgamma(2,30)
  tau.Z.stage~dgamma(2,3)
  tau.Z.hb~dgamma(2,300)
  tau.Z.t~dgamma(1.5,0.5)

  cut.stage[1]<-0
  cut.stage[2]<-c2
  cut.stage[3]<-2
  c2<-cc*2
  cc~dbeta(1,1)

}
```

A.7.5 R function to adjust the mean and the variance of the Gaussian random variables in non-Hodgkin lymphoma data

```
##### posterior mean and variance for hb ######
function(y,E0,V0)
{# observed variables (hb, wbc)
  Ey<-E0[1]
  Ez<-E0[3:5]
  Vy<-V0[1,1]
  Vz<-V0[3:5,3:5]
  C<-V0[3:5,1]
  E1<-Ez+C*(y1-Ey)/Vy
```

```
  V1<-Vz-C%*%t(C)/Vy
  P1<-solve(V1)
  return(list(E=E1,P=P1))
}

######## posterior mean and variance for wbc #######

function(y,E0,V0)
{# observed variables (hb, wbc)
  Ey<-E0[2]
  Ez<-E0[3:5]
  Vy<-V0[2,2]
  Vz<-V0[3:5,3:5]
  C<-V0[3:5,2]
  E1<-Ez+C*(y2-Ey)/Vy
  V1<-Vz-C%*%t(C)/Vy
  P1<-solve(V1)
  return(list(E=E1,P=P1))
}
```

A.7.6 R function to update the mean and the variance of the ordinal and the categorical random variables

```
######## posterior mean and variance for Z.stage ####### ##
function(y,E0,V0,cuts,nstep=100){
  S0<-sqrt(V0)
  k<-length(cuts)
  if(y==0){
    lower<-0 }
  else
  {lower<-pnorm(cuts[y],E0,S0)}
  if(y==k)
  {upper<-1}
  else {
```

```
    upper<-pnorm(cuts[y+1],E0,S0)}
  u<-seq(lower,upper,length.out=(nstep+1))
  Z<-qnorm(u,E0,S0)
  Z2=Z*Z
  E1<-(sum(Z)-(Z[1]+Z[nstep+1])/2)/nstep
  E1Z2<-(sum(Z2)-(Z2[1]+Z2[nstep+1])/2)/nstep
  V1<-E1Z2-E1*E1
  ans<-list(E=E1,V=V1)
  return(ans)
}


######## posterior mean and variance for Z.albumin #########
function(y,E0,V0,cuts,nstep=100)
{# y is an integer in [0,k]
  S0<-sqrt(V0)
  k<-length(cuts)
  if (y==0){
    lower<-0
  }
  else
  {lower<-pnorm(cuts[y],E0,S0)}
  if(y==k){
    upper<-1
  }
  else
  {upper<-pnorm(cuts[y+1],E0,S0)}
  u<-seq(lower,upper,length.out=(nstep+1))
  Z<-qnorm(u,E0,S0)
  Z2<-Z*Z
  E1<-(sum(Z)-(Z[1]+Z[nstep+1])/2)/nstep
  E1Z2<-(sum(Z2)-(Z2[1]+Z2[nstep+1])/2)/nstep
  V1<-E1Z2-(E1*E1)
  out<-list(E=E1,V=V1)
  return(out)
```

```
}
```

A.7.7 R function to adjusted the mean and the variance for stage and albumin using BLK with non-conjugate prior update

```
function(y,E0,V0,var,cuts)
{#var=3:stage, var=4:albumin
  if (var==3)
  {adjusted<-adjstage(y,E0[3],V0[3,3],cuts=c(-3.2,-1.5,1,4,5),nstep=100)
  }
  if (var==4)
  {adjusted<-adjalbumin(y,E0[4],V0[4,4],cuts=c(-1.325,1,2.5),nstep=100)
  }
  Vxadj<-adjusted$V
  Exadj<-adjusted$E
  Ez<-E0[3:5]
  Vz<-V0[3:5,3:5]
  C<-V0[3:5,var]
  E1<-Ez+C*(Exadj-E0[var])/V0[var,var]
  V1<-Vz-C%*%t(C)/V0[var,var]+C%*%t(C)*Vxadj/(V0[var,var]*V0[var,var])
  P1<-solve(V1)
  return(list(E=E1,P=P1))
}
```

A.7.8 R function to compute the posterior mean using BLK in order to obtain the prognostic index value for one patient

```
function(patdata,V0,mu0,musex,muage,cuts)
{
  E0<-mu0+muage*patdata[5]+musex*patdata[6]
  P0<-solve(V0[3:5,3:5])
```

```
n<-3
J<-4
d<-n*n*J
E1<-matrix(nrow=n,ncol=J)
P1<-numeric(d)
dim(P1)<-c(n,n,J)

# covariate (hb)

if (is.na(patdata[1])==TRUE)
{P1[,,1]<-P0
E1[,1]<-E0[3:5]
}

else

{ adjust1<-adjbynorm(patdata[1],E0,V0)
P1[,,1]<-adjust1$P
E1[,1]<-adjust1$E
}

# covariate (wbc)

if (is.na(patdata[2])==TRUE)
{P1[,,2]<-P0
E1[,2]<-E0[3:5]
}

else

{ adjust2<-adjbynorm(patdata[2],E0,V0)
P1[,,2]<-adjust2$P
E1[,2]<-adjust2$E
}
```

```
## covariate (stage)


if (is.na(patdata[3])==TRUE)
{P1[,,3]<-P0
E1[,3]<-E0[3:5]
}

else

{ adjust3<-adjbynoncon(patdata[3],E0,V0,3,cuts)
P1[,,3]<-adjust3$P
E1[,3]<-adjust3$E
}

#### covariate (albumin)

if (is.na(patdata[4])==TRUE)
{P1[,,4]<-P0
E1[,4]<-E0[3:5]
}

else

{ adjust4<-adjbynoncon(patdata[4],E0,V0,4,cuts)
P1[,,4]<-adjust4$P
E1[,4]<-adjust4$E
}

V0<-V0[3:5,3:5]
E0<-E0[3:5]
PP<-matrix(0,nrow=3,ncol=3)
PPEE<-rep(0,3)
for(j in 1:4)
{PP<-PP+P1[,,j]
```

```
  PPEE<-PPEE+P1[,,j]%*%E1[,j]
  }
  PP<-PP-3*P0
  PPEE<-PPEE-3*P0%*%E0
  EE<-solve(PP,PPEE)
  return(EE[3])

}
```

A.7.9 R function to obtain the adjusted expectation of Z_T for patient i

```
function(data,V0,mu0,musex,muage,cuts)
{
  indexvalues<-numeric(1391)
  for (i in 1:1391)
  {
    indexvalues[i]<-BLKindex(data[i,],V0,mu0,musex,muage,cuts)
  }
  return(indexvalues)
}
```

A.7.10 R function for prototype prognostic index calculator

```
function(params)
{
mean <- mean(out) ### out: is the posterior expectation of prognostic index values.
std.dev <- sd(out)
############### AGE
write(file="","Please enter the Age in years of the patient at time of diagnosis.")
age<-scan(n=1)
############### SEX
write(file="","Please enter the Sex of the patient.
```

```
Enter 1 for male or 2 for female.")
sex<-scan(n=1)
############### STAGE
write(file="","Please enter the Clinical Stage of the patient (1, 2, 3 or 4).")
stage<-scan(n=1)
############### HB
write(file="","Please enter the Haemoglobin (g/l) measurement for the patient.")
hb<-scan(n=1)
################ WBC
write(file="","Please enter the White Blood Cell count for the patient.")
wbc<-scan(n=1)
################ ALBUMIN
write(file="","Please enter 1 if the Serum Albumin
measurement for the patient is normal")
write(file="","or 2 if it is abnormal")
albumin<-scan(n=1)

eta<-BLKindex(data[1,],params$V0,params$mu0,params$musex,params$muage,params$cuts)
ind<-100*pnorm(eta,mean,std.dev)
index.patient<-round(ind)
write(file="","Index value is")
write(index.patient,file="")
write(file="","The index is on a scale from 0 to 100,
Greater index values indicate greater risk.")

}
```

A.7.11 Offline learning model for non-Hodgkin lymphoma data in the indirect method

```
#########################################################################
#### Offline learning model for non-Hodgkin lymphoma data in the ####
############### indirect method ##############
```

```
#############################################################################

model{
for(i in 1:1391){
stage[i] ~ dcat(p[i,1:4])

p[i,1] <- 1-q[i,1]
for(r in 2:3){
p[i,r] <- q[i,r-1] - q[i,r]
}
p[i,4] <- q[i,3]


for(r in 1:3){
logit(q[i,r]) <- Z.stage[i]- cuts[r]
}


is.censored[i]~dinterval (t[i], t.cen[i])
t[i]~dweib(alpha,lambda[i])
log(lambda[i]) <- mean.Z.t[i]

hb[i]~dnorm(mean.Z.hb[i],tau.Z.hb)
Z.hb[i]<-hb[i]
mean.Z.hb[i]<-gamma0.hb+gamma.hbage*age[i]+gamma.hbsex*sex[i]

wbc[i]~dnorm(mean.Z.wbc[i],tau.Z.wbc)
Z.wbc[i]<-wbc[i]
mean.Z.wbc[i]<-gamma0.wbc+gamma.wbcage*age[i]+gamma.wbcsex*sex[i]
+gamma.wbchb*(Z.hb[i]-mean.Z.hb[i])



Z.stage[i] ~ dnorm(mean.Z.stage[i], tau.Z.stage)
mean.Z.stage[i]<-gamma0.stage+gamma.stageage*age[i]+gamma.stagesex*sex[i]
```

```
+gamma.stagehb*(Z.hb[i]-mean.Z.hb[i])+gamma.stagewbc*(Z.wbc[i]-mean.Z.wbc[i])


albumin[i]~dinterval(Z.albumin[i],0)
Z.albumin[i]~dnorm(mean.Z.albumin[i],1)
mean.Z.albumin[i]<-gamma0.albumin+gamma.albuminage*age[i]
+gamma.albuminsex*sex[i]+gamma.albuminwbc*(Z.wbc[i]-mean.Z.wbc[i])
+gamma.albuminhb*(Z.hb[i]-mean.Z.hb[i])
+gamma.albuminstage*(Z.stage[i]-mean.Z.stage[i])

Z.t[i]~dnorm(mean.Z.t[i],tau.Z.t)
mean.Z.t[i]<- gamma0.t+gamma.t.age*age[i]+gamma.t.wbc*(Z.wbc[i]-mean.Z.wbc[i])
+gamma.t.sex*sex[i]+gamma.t.albumin*(Z.albumin[i]-mean.Z.albumin[i])
+gamma.t.stage*(Z.stage[i]-mean.Z.stage[i])+gamma.t.hb*(Z.hb[i]-mean.Z.hb[i])


}

## priors over thresholds


cuts[1] <- 0
cuts[2] <- c2
cuts[3] <- 1

c2<-cc
cc~dbeta(1,1)

alpha ~ dgamma(4, 4)# prior for alpha
gamma0.t ~ dnorm(0, 0.01)
gamma.t.sex~dnorm(0,0.001)
gamma.t.age~dnorm(0.0,0.001)
gamma.t.wbc~dnorm(0.0,0.001)
gamma.t.albumin~dnorm(1,0.0001)
gamma.t.stage~dnorm(0.0,0.001)
```

```
gamma.t.hb~dnorm(0.0,0.001)

gamma0.hb~dnorm(100.0,0.0001)
gamma.hbage~dnorm(0.0,0.001)
gamma.hbsex~dnorm(0.0,0.001)

gamma0.wbc~dnorm(10.0,0.001)
gamma.wbcage~dnorm(0.0,0.001)
gamma.wbcsex~dnorm(0.0,0.001)
gamma.wbchb~dnorm(0.0,0.001)

gamma0.stage~dnorm(0.0,0.1)
gamma.stageage~dnorm(0.0,0.01)
gamma.stagesex~dnorm(0.0,0.01)
gamma.stagehb~dnorm(0.0,0.01)
gamma.stagewbc~dnorm(0.0,0.01)


gamma0.albumin~dnorm(0,10)
gamma.albuminage~dnorm(0,10)
gamma.albuminsex~dnorm(0,10)
gamma.albuminwbc~dnorm(0,10)
gamma.albuminstage~dnorm(0,10)
gamma.albuminhb~dnorm(0,10)

tau.Z.wbc~dgamma(2,30)
tau.Z.stage~dgamma(2,3)
tau.Z.hb~dgamma(2,300)
tau.Z.t~dgamma(1.5,0.5)


}
```

A.7.12 R functions to do the adjustment by the categorical random variables in the indirect method

```
function(y,E0,tol=1E-5,n=20)
{#
 # Initial approximation
 z0<-E0
 d<-1
 for (i in 1:n)
     {if (abs(d)>tol)
         {ez<-exp(z0)
          pz<-ez/(1+ez)
          f<-y-(z0-E0)-pz
          fd<--(1+pz*(1-pz))
          d<-f/fd
          z0<-z0-d
          }
       }
 ez<-exp(z0)
 v0<- -1/fd
 # Tierney and Kadane normalising constant (denominator)
 g0<-z0*y -((z0-E0)^2)/2-log(1+ez)
 eta0<-exp(g0)*sqrt(2*pi*v0)
 # Mean shift
 C<-5*sqrt(v0)-z0
 # Tierney and Kadane numerator for E(Z+C)
 z1<-z0
 d<-1
 for (i in 1:n)
     {if (abs(d)>tol)
         {ez<-exp(z1)
          pz<-ez/(1+ez)
          f<-y-(z1-E0)-pz+1/(z1+C)
          fd<--(1+pz*(1-pz)+(z1+C)^(-2))
          d<-f/fd
```

```
         z1<-z1-d
         }
      }
ez<-exp(z1)
v1<- -1/fd
g1<-log(z1+C)+z1*y -((z1-E0)^2)/2-log(1+ez)
eta1<-exp(g1)*sqrt(2*pi*v1)
# Tierney and Kadane approximation to E(Z+C) and E(Z)
E1<-eta1/eta0
postmean<-E1-C
# Tierney and Kadane numerator for E([Z+C]^2)
z2<-z1
d<-1
for (i in 1:n)
    {if (abs(d)>tol)
        {ez<-exp(z2)
         pz<-ez/(1+ez)
         f<-y-(z2-E0)-pz+2/(z2+C)
         fd<--(1+pz*(1-pz)+2*(z2+C)^(-2))
         d<-f/fd
         z2<-z2-d
         }
      }
ez<-exp(z2)
v2<- -1/fd
g2<-2*log(z2+C)+z2*y -((z2-E0)^2)/2-log(1+ez)
eta2<-exp(g2)*sqrt(2*pi*v2)
# Tierney and Kadane approximation to E([Z+C]^2) and Var(Z)
E2<-eta2/eta0
postvar<-E2-E1^2
return(list(E1=postmean,V1=postvar))
}
```

A.8 Appendix to Chapter 8

A.8.1 R code for simulation from the direct model with direct parameter values in the NHL example

```
library(MASS)
library(rjags)
# library(tmvtnorm)
# library(matrixcalc)
# library(corpcor)



n<-1200
sex<-rep(1,n)    # all male
age<-rep(0,n)   # because each case aged 60 and we centred
                        # them as 60-60=0

# We use E0(Z) and V0(Z) that we obtained from the
# offline learning model to generate n=1200 samples
# from a multivariate normal distribution for Z.
Z<- mvrnorm(n, E0, V0)

hb<-Z[,1]
wbc<-Z[,2]

# Ordinal variables such as stage
x_3_stage<-ifelse(Z[,3]>cuts[3],3,2)
x_3_stage<-ifelse(Z[,3]<cuts[2],1,x_3_stage)
x_3_stage<-ifelse(Z[,3]<cuts[1],0,x_3_stage)
stage<-x_3_stage



# Binary variables such as albumin
x_4_albumin<-ifelse(Z[,4]>0,1,0)
```

```
albumin<-x_4_albumin


# Actual values of Z_T
actual_Z<-Z[,5]

# Then we have the matrix Z with dimension 1200X5 where
# for example Z[,1] is the generated values for the covariate
# Hb and Z[,2] is the generated values for the covariate Wbc
# and so on. The last column of the matrix Z, Z[,5] represents
# the actual values for Z_T. Afterwards, we use the model
# comparison (in rjags) in Appendix  A.8.1. to compute the posterior
# means of Z_T using MCMC. Then we use Bayes linear
# kinematic to compute the posterior means for Z_T and
# compare the results between the two methods.
```

A.8.2 Rjags model specification for model comparison for non-Hodgkin lymphoma data in the direct method

```
#############################################################################
##### model comparison for non-Hodgkin lymphoma data in the #####
############## direct method #############
#############################################################################

model
{
  for (i in 1:1391){

  mean.Z.t[i]<- gamma0.t+gamma.t.age*age[i]+gamma.t.wbc*(Z.wbc[i]-mean.Z.wbc[i])
  +gamma.t.sex*sex[i]+gamma.t.albumin*(Z.albumin[i]-mean.Z.albumin[i])
  +gamma.t.stage*(Z.stage[i]-mean.Z.stage[i])+gamma.t.hb*(Z.hb[i]-mean.Z.hb[i])

  hb[i]~dnorm(mean.Z.hb[i],tau.Z.hb)
  Z.hb[i]<-hb[i]
  mean.Z.hb[i]<-gamma0.hb+gamma.hbage*age[i]+gamma.hbsex*sex[i]
```

```
  wbc[i]~dnorm(mean.Z.wbc[i],tau.Z.wbc)
  Z.wbc[i]<-wbc[i]
  mean.Z.wbc[i]<-gamma0.wbc+gamma.wbcage*age[i]+gamma.wbcsex*sex[i]
  +gamma.wbchb*(Z.hb[i]-mean.Z.hb[i])

  stage[i]~dinterval(Z.stage[i],cut.stage[1:3])
  Z.stage[i] ~ dnorm(mean.Z.stage[i], tau.Z.stage)
  mean.Z.stage[i]<-gamma0.stage+gamma.stageage*age[i]+gamma.stagesex*sex[i]
  +gamma.stagehb*(Z.hb[i]-mean.Z.hb[i])+gamma.stagewbc*(Z.wbc[i]-mean.Z.wbc[i])

  albumin[i]~dinterval(Z.albumin[i],0)
  Z.albumin[i]~dnorm(mean.Z.albumin[i],1)
  mean.Z.albumin[i]<-gamma0.albumin+gamma.albuminage*age[i]
  +gamma.albuminsex*sex[i]+gamma.albuminwbc*(Z.wbc[i]-mean.Z.wbc[i])
  +gamma.albuminstage*(Z.stage[i]-mean.Z.stage[i])
  +gamma.albuminhb*(Z.hb[i]-mean.Z.hb[i])
  }

  cut.stage=c(0,c2,2)
  c2<-2*cc
}
```

A.8.3 Computing the predictions of Z_T using Bayes linear kinematic for the direct method

```
albumin<-albumin+1
stage<-stage+1
data<-data.frame(hb,wbc,stage,albumin,age,sex)
data<-as.matrix(data)

patdata<-data[1,]
E0<-mu0+muage*patdata[5]+musex*patdata[6]

########## posterior mean and variance for hb ##########
```

```
adjbynorm<-function(y,E0,V0)
{# observed variables (hb, wbc)
  Ey<-E0[1]
  Ez<-E0[3:5]
  Vy<-V0[1,1]
  Vz<-V0[3:5,3:5]
  C<-V0[3:5,1]
  E1<-Ez+C*(y1-Ey)/Vy
  V1<-Vz-C%*%t(C)/Vy
  P1<-solve(V1)
  return(list(E=E1,P=P1))
}


########## posterior mean and variance for wbc ##########
adjbynorm<-function(y,E0,V0)
{# observed variables (hb, wbc)
  Ey<-E0[2]
  Ez<-E0[3:5]
  Vy<-V0[2,2]
  Vz<-V0[3:5,3:5]
  C<-V0[3:5,2]
  E1<-Ez+C*(y2-Ey)/Vy
  V1<-Vz-C%*%t(C)/Vy
  P1<-solve(V1)
  return(list(E=E1,P=P1))
}


adjstage<-function(y,E0,V0,cuts,nstep=100){
  S0<-sqrt(V0)
  k<-length(cuts)
  if(y==0){
    lower<-0 }
  else
```

```
  {lower<-pnorm(cuts[y],E0,S0)}
  if(y==k)
  {upper<-1}
  else {
    upper<-pnorm(cuts[y+1],E0,S0)}
  u<-seq(lower,upper,length.out=(nstep+1))
  Z<-qnorm(u,E0,S0)
  Z2=Z*Z
  E1<-(sum(Z)-(Z[1]+Z[nstep+1])/2)/nstep
  E1Z2<-(sum(Z2)-(Z2[1]+Z2[nstep+1])/2)/nstep
  V1<-E1Z2-E1*E1
  ans<-list(E=E1,V=V1)
  return(ans)
}

########## posterior mean and variance for Z.albumin ##########  adjalbumin
adjalbumin<-function(y,E0,V0,cuts,nstep=100)
{# y is an integer in [0,k]
  S0<-sqrt(V0)
  k<-length(cuts)
  if (y==0){
    lower<-0
  }
  else
  {lower<-pnorm(cuts[y],E0,S0)}
  if(y==k){
    upper<-1
  }
  else
  {upper<-pnorm(cuts[y+1],E0,S0)}
  u<-seq(lower,upper,length.out=(nstep+1))
  Z<-qnorm(u,E0,S0)
  Z2<-Z*Z
  E1<-(sum(Z)-(Z[1]+Z[nstep+1])/2)/nstep
  E1Z2<-(sum(Z2)-(Z2[1]+Z2[nstep+1])/2)/nstep
```

```
  V1<-E1Z2-(E1*E1)
  out<-list(E=E1,V=V1)
  return(out)

}



adjbynoncon<-function(y,E0,V0,var,cuts)
{#var=3:stage, var=4:albumin
  if (var==3)
  {adjusted<-adjstage(y,E0[3],V0[3,3],cuts=c(-4.5,0,2*cc.m, 2,4.5),nstep=100)
  }
  if (var==4)
  {adjusted<-adjalbumin(y,E0[4],V0[4,4],cuts=c(-4.5,0,4.5),nstep=100)
    }
  Vxadj<-adjusted$V
  Exadj<-adjusted$E
  Ez<-E0[3:5]
  Vz<-V0[3:5,3:5]
  C<-V0[3:5,var]
  E1<-Ez+C*(Exadj-E0[var])/V0[var,var]
  V1<-Vz-C%*%t(C)/V0[var,var]+C%*%t(C)*Vxadj/(V0[var,var]*V0[var,var])
  P1<-solve(V1)
  return(list(E=E1,P=P1))
}

##############################
BLKindex<-function(patdata,V0,mu0,musex,muage,cuts)
{
  E0<-mu0+muage*patdata[5]+musex*patdata[6]
  P0<-solve(V0[3:5,3:5])
  n<-3
  J<-4
  d<-n*n*J
```

```
E1<-matrix(nrow=n,ncol=J)
P1<-numeric(d)
dim(P1)<-c(n,n,J)

# covariate (hb)

if (is.na(patdata[1])==TRUE)
{P1[,,1]<-P0
E1[,1]<-E0[3:5]
}

else

{ adjust1<-adjbynorm(patdata[1],E0,V0)
P1[,,1]<-adjust1$P
E1[,1]<-adjust1$E
}

# covariate (wbc)

if (is.na(patdata[2])==TRUE)
{P1[,,2]<-P0
E1[,2]<-E0[3:5]
}

else

{ adjust2<-adjbynorm(patdata[2],E0,V0)
P1[,,2]<-adjust2$P
E1[,2]<-adjust2$E
}

# covariate (stage)
```

```
if (is.na(patdata[3])==TRUE)
{P1[,,3]<-P0
E1[,3]<-E0[3:5]
}

else

{ adjust3<-adjbynoncon(patdata[3],E0,V0,3,cuts)
P1[,,3]<-adjust3$P
E1[,3]<-adjust3$E
}

# covariate (albumin)

if (is.na(patdata[4])==TRUE)
{P1[,,4]<-P0
E1[,4]<-E0[3:5]
}

else

{ adjust4<-adjbynoncon(patdata[4],E0,V0,4,cuts)
P1[,,4]<-adjust4$P
E1[,4]<-adjust4$E
}

V0<-V0[3:5,3:5]
E0<-E0[3:5]
PP<-matrix(0,nrow=3,ncol=3)
PPEE<-rep(0,3)
for(j in 1:4)
{PP<-PP+P1[,,j]
PPEE<-PPEE+P1[,,j]%*%E1[,j]
}
PP<-PP-3*P0
```

```
  PPEE<-PPEE-3*P0%*%E0
  EE<-solve(PP,PPEE)
  return(EE[3])
  }

# BLKindex(patdata,V0,mu0,musex,muage,cuts)


indexBLKall<-function(data,V0,mu0,musex,muage,cuts)
{
  indexvalues<-numeric(1200)
  for (i in 1:1200)
  {
    indexvalues[i]<-BLKindex(data[i,],V0,mu0,musex,muage,cuts)
  }
  return(indexvalues)
}

# out<-indexBLKall(data,V0,mu0,musex,muage,cuts)
```

A.8.4 Rjags model specification for model comparison for non-Hodgkin lymphoma data in the indirect method

```
#############################################################################
##### model comparison for non-Hodgkin lymphoma data in the #####
############# indirect method ############
#############################################################################
model{
for(i in 1:1391){
stage[i] ~ dcat(p[i,1:4])

p[i,1] <- 1-q[i,1]
for(r in 2:3){
p[i,r] <- q[i,r-1] - q[i,r]
}
```

```
p[i,4] <- q[i,3]


for(r in 1:3){
logit(q[i,r]) <- Z.stage[i]- cuts[r]
}


mean.Z.t[i]<- gamma0.t+gamma.t.age*age[i]+gamma.t.wbc*(Z.wbc[i]-mean.Z.wbc[i])
+gamma.t.sex*sex[i]+gamma.t.albumin*(Z.albumin[i]-mean.Z.albumin[i])
+gamma.t.stage*(Z.stage[i]-mean.Z.stage[i])+gamma.t.hb*(Z.hb[i]-mean.Z.hb[i])

hb[i]~dnorm(mean.Z.hb[i],tau.Z.hb)
Z.hb[i]<-hb[i]
mean.Z.hb[i]<-gamma0.hb+gamma.hbage*age[i]+gamma.hbsex*sex[i]

wbc[i]~dnorm(mean.Z.wbc[i],tau.Z.wbc)
Z.wbc[i]<-wbc[i]
mean.Z.wbc[i]<-gamma0.wbc+gamma.wbcage*age[i]+gamma.wbcsex*sex[i]
+gamma.wbchb*(Z.hb[i]-mean.Z.hb[i])

Z.stage[i] ~ dnorm(mean.Z.stage[i], tau.Z.stage)
mean.Z.stage[i]<-gamma0.stage+gamma.stageage*age[i]+gamma.stagesex*sex[i]
+gamma.stagehb*(Z.hb[i]-mean.Z.hb[i])+gamma.stagewbc*(Z.wbc[i]-mean.Z.wbc[i])

albumin[i]~dinterval(Z.albumin[i],0)
Z.albumin[i]~dnorm(mean.Z.albumin[i],1)
mean.Z.albumin[i]<-gamma0.albumin+gamma.albuminage*age[i]+gamma.albuminsex*sex[i]
+gamma.albuminwbc*(Z.wbc[i]-mean.Z.wbc[i])
+gamma.albuminstage*(Z.stage[i]-mean.Z.stage[i])
+gamma.albuminhb*(Z.hb[i]-mean.Z.hb[i])
}


cuts=c(0,c2,1)
```

```
c2<-cc

}
```

A.8.5 Computing the predictions of Z_T using Bayes linear kinematic for the indirect method

```
albumin<-albumin+1
stage<-stage+1
data<-data.frame(hb,wbc,stage,albumin,age,sex)
data<-as.matrix(data)

patdata<-data[1,]
E0<-mu0+muage*patdata[5]+musex*patdata[6]

########## posterior mean and variance for hb ##########
adjbynorm<-function(y,E0,V0)
{# observed variables (hb, wbc)
  Ey<-E0[1]
  Ez<-E0[3:5]
  Vy<-V0[1,1]
  Vz<-V0[3:5,3:5]
  C<-V0[3:5,1]
  E1<-Ez+C*(y1-Ey)/Vy
  V1<-Vz-C%*%t(C)/Vy
  P1<-solve(V1)
  return(list(E=E1,P=P1))
}

########## posterior mean and variance for wbc ##########
adjbynorm<-function(y,E0,V0)
{# observed variables (hb, wbc)
  Ey<-E0[2]
  Ez<-E0[3:5]
  Vy<-V0[2,2]
```

```
  Vz<-V0[3:5,3:5]
  C<-V0[3:5,2]
  E1<-Ez+C*(y2-Ey)/Vy
  V1<-Vz-C%*%t(C)/Vy
  P1<-solve(V1)
  return(list(E=E1,P=P1))
}


### function to find posterior mean and variance for
# Z.albumin and Z.stage.
function(y,E0,tol=1E-5,n=20)
{#
 # Initial approximation
 z0<-E0
 d<-1
 for (i in 1:n)
     {if (abs(d)>tol)
         {ez<-exp(z0)
          pz<-ez/(1+ez)
          f<-y-(z0-E0)-pz
          fd<--(1+pz*(1-pz))
          d<-f/fd
          z0<-z0-d
          }
       }
 ez<-exp(z0)
 v0<- -1/fd
 # Tierney and Kadane normalising constant (denominator)
 g0<-z0*y -((z0-E0)^2)/2-log(1+ez)
 eta0<-exp(g0)*sqrt(2*pi*v0)
 # Mean shift
 C<-5*sqrt(v0)-z0
 # Tierney and Kadane numerator for E(Z+C)
 z1<-z0
```

```
d<-1
for (i in 1:n)
    {if (abs(d)>tol)
        {ez<-exp(z1)
         pz<-ez/(1+ez)
         f<-y-(z1-E0)-pz+1/(z1+C)
         fd<--(1+pz*(1-pz)+(z1+C)^(-2))
         d<-f/fd
         z1<-z1-d
         }
      }
ez<-exp(z1)
v1<- -1/fd
g1<-log(z1+C)+z1*y -((z1-E0)^2)/2-log(1+ez)
eta1<-exp(g1)*sqrt(2*pi*v1)
# Tierney and Kadane approximation to E(Z+C) and E(Z)
E1<-eta1/eta0
postmean<-E1-C
# Tierney and Kadane numerator for E([Z+C]^2)
z2<-z1
d<-1
for (i in 1:n)
    {if (abs(d)>tol)
        {ez<-exp(z2)
         pz<-ez/(1+ez)
         f<-y-(z2-E0)-pz+2/(z2+C)
         fd<--(1+pz*(1-pz)+2*(z2+C)^(-2))
         d<-f/fd
         z2<-z2-d
         }
      }
ez<-exp(z2)
v2<- -1/fd
g2<-2*log(z2+C)+z2*y -((z2-E0)^2)/2-log(1+ez)
eta2<-exp(g2)*sqrt(2*pi*v2)
```

```
 # Tierney and Kadane approximation to E([Z+C]^2) and Var(Z)
 E2<-eta2/eta0
 postvar<-E2-E1^2
 return(list(E1=postmean,V1=postvar))
 }


 adjbynoncon<-function(y,E0,V0,var,cuts)
{#var=3:stage, var=4:albumin
  if (var==3)
  {adjusted<-adjstage(y,E0[3],V0[3,3],cuts=c(-4.5,0,cc.m, 1,4.5),nstep=100)
  }
  if (var==4)
  {adjusted<-adjalbumin(y,E0[4],V0[4,4],cuts=c(-4.5,0,4.5),nstep=100)
  }
  Vxadj<-adjusted$V
  Exadj<-adjusted$E
  Ez<-E0[3:5]
  Vz<-V0[3:5,3:5]
  C<-V0[3:5,var]
  E1<-Ez+C*(Exadj-E0[var])/V0[var,var]
  V1<-Vz-C%*%t(C)/V0[var,var]+C%*%t(C)*Vxadj/(V0[var,var]*V0[var,var])
  P1<-solve(V1)
  return(list(E=E1,P=P1))
}

##############################
BLKindex<-function(patdata,V0,mu0,musex,muage,cuts)
{
  E0<-mu0+muage*patdata[5]+musex*patdata[6]
  P0<-solve(V0[3:5,3:5])
  n<-3
  J<-4
  d<-n*n*J
  E1<-matrix(nrow=n,ncol=J)
```

```
P1<-numeric(d)
dim(P1)<-c(n,n,J)

# covariate (hb)

if (is.na(patdata[1])==TRUE)
{P1[,,1]<-P0
E1[,1]<-E0[3:5]
}

else

{ adjust1<-adjbynorm(patdata[1],E0,V0)
P1[,,1]<-adjust1$P
E1[,1]<-adjust1$E
}

# covariate (wbc)

if (is.na(patdata[2])==TRUE)
{P1[,,2]<-P0
E1[,2]<-E0[3:5]
}

else

{ adjust2<-adjbynorm(patdata[2],E0,V0)
P1[,,2]<-adjust2$P
E1[,2]<-adjust2$E
}

# covariate (stage)


if (is.na(patdata[3])==TRUE)
```

```
{P1[,,3]<-P0
E1[,3]<-E0[3:5]
}

else

{ adjust3<-adjbynoncon(patdata[3],E0,V0,3,cuts)
P1[,,3]<-adjust3$P
E1[,3]<-adjust3$E
}

# covariate (albumin)

if (is.na(patdata[4])==TRUE)
{P1[,,4]<-P0
E1[,4]<-E0[3:5]
}

else

{ adjust4<-adjbynoncon(patdata[4],E0,V0,4,cuts)
P1[,,4]<-adjust4$P
E1[,4]<-adjust4$E
}

V0<-V0[3:5,3:5]
E0<-E0[3:5]
PP<-matrix(0,nrow=3,ncol=3)
PPEE<-rep(0,3)
for(j in 1:4)
{PP<-PP+P1[,,j]
PPEE<-PPEE+P1[,,j]%*%E1[,j]
}
PP<-PP-3*P0
PPEE<-PPEE-3*P0%*%E0
```

```
  EE<-solve(PP,PPEE)
  return(EE[3])

}

# BLKindex(patdata,V0,mu0,musex,muage,cuts)


indexBLKall<-function(data,V0,mu0,musex,muage,cuts)
{
  indexvalues<-numeric(1200)
  for (i in 1:1200)
  {
    indexvalues[i]<-BLKindex(data[i,],V0,mu0,musex,muage,cuts)
  }
  return(indexvalues)
}

# out<-indexBLKall(data,V0,mu0,musex,muage,cuts)
```

A.9 List of abbreviations and notations

Table A.8: Glossary of abbreviation

Symbols	Meaning
AP	Alkaline phosphatase.
BLK	Bayes linear kinematic.
BN	Bayesian network.
bnlearn	Bayesian network learning.
Bsy	B-symptoms.
BUGS	Bayesian inference using Gibbs sampler.
BVS	Bayesian variable selection.
CPT	Conditional probability table.
DA	Data augmentation.
DAG	Directed acyclic graph.
DBN	Dynamic Bayesian network.
Depscore	Deprivation score.
DLBCL	Diffuse large B-cell lymphoma.
DLBCL-NOS	Diffuse large B-cell lymphoma- Not Otherwise Specified.
ECOG	Eastern Co-operative Oncology Group.
Extranod	Extranodal without bone marrow.
EVLP	Ex vivo lung perfusion.
FCD	Full conditional distribution.
GLM	Generalised Linear Model.
GVS	Gibbs variable selection.
HB	Haemoglobin.
i.i.d.	Independent and identically distributed.
INLA	Integrated nested Laplace approximation.
JAGS	Just another Gibbs sampler.
LDH	Serum lactate dehydrogenase.
MAR	Missing at random.
Marrow	Bone marrow involvement.
MB	Markov blanket.

Table A.9: Glossary of abbreviation

Symbols	Meaning
MCAR	Missing completely at random.
MCMC	Markov chain Monte Carlo.
MNAR	Missing not at random.
NHL	Non-Hodgkins Lymphoma.
NRHG	Northern Regional Haematology Group.
PACE	Population Adjusted Clinical Epidemiology.
PCH	Piecewise constant hazard.
PGM	Probabilistic graphical model.
SNLG	Scotland Newcastle Lymphoma Group.
SSVS	Stochastic search variable selection.
TDI	Townsend deprivation index.
Urea	Blood urea nitrogen.
WBC	White blood cell.

Table A.10: Glossary of notations

Symbols	Meaning
$\pi(\theta)$	Prior distribution of parameter θ
$L(\theta\|y)$	Likelihood function.
$\pi(\theta\|y)$	Posterior distribution of parameter θ given data y.
$g(\theta)$	Arbitrary function of θ.
$g(\cdot)$	Link function.
$\theta(x_i)$	The probability of the event for subject i.
Y_{obs}	The observed data.
Y_{miss}	The missing values.
ψ	Unknown quantity for the missing data mechanism.
$F(t)$	Lifetime distribution function.
$S(t)$	Survival function.
$f(t)$	Lifetime probability density function.
$h(t)$	Hazard function.
$H(t)$	Cumulative hazard function.
t_f	Posterior predictive density of future observation.
A	Acceptance probability.
Φ	Standard normal cumulative distribution function.
P_0	Probability that the lung be used.
$\mathrm{P}_{01}, \mathrm{P}_{03}$	Lower and upper quartiles for the proportion.
$f_0(y \mid x)$	Prior predictive distribution.
$f_1(y \mid x)$	Posterior predictive distribution.
$D = \{x_1, ..., x_j\}$	Observed data.
$\theta_1, ..., \theta_P$	Multinomial probability distribution.
$\pi(\theta \mid D)$	Posterior distribution of θ.
$S^\star$	Best structure in Bayesian network.
$\epsilon_1, ..., \epsilon_p$	Autoregression innovations for transformed parameters.

Table A.11: Glossary of notations

Symbols	Meaning
$\mathrm{E}_0, \mathrm{Var}_0, \mathrm{Cov}_0$	Prior expectation, variance, covariance.
$\mathrm{E}_1, \mathrm{Var}_1, \mathrm{Cov}_1$	Posterior expectation, variance, covariance.
γ	Regression coefficients in autoregression model.
$\{X_1, ..., X_J\}$	Covariates in the model.
$\mathrm{R}(Y\|D)$	Error terms from the Bayes linear fit.
$\psi(\cdot)$	Digamma function.
$\psi_1(\cdot)$	trigamma function.
I_k	Interval k in a piecewise constant hazard model where $k = 1, ..., K$.
$s_1, ..., s_{k-1}$	Cut-points in a piecewise constant hazard model.
$\underline{\eta} = (\eta_1, ..., \eta_n)'$	Transformed parameters.
$Z_{m+1} = Z_T$	Prognostic index.
$\hat{Z}_{BLK,i}$	Prognostic index value calculated using BLK.
$\hat{Z}_{MCMC,i}$	Prognostic index value calculated using full Bayes.

References

Aalen, O. (2008). *Survival and Event History Analysis*. New York, NY: Springer.

Alston, R. D., S. Rowan, T. O. Eden, A. Moran, and J. M. Birch (2007). Cancer incidence patterns by region and socioeconomic deprivation in teenagers and young adults in England. *British Journal of Cancer 96*(11), 1760–1766.

Andreas, W. (2011). *Frailty Models in Survival Analysis*. Boca Raton, Florida: CRC Press.

Andreasson, A. S., L. A. Borthwick, C. Gillespie, K. Jiwa, J. Scott, P. Henderson, J. Mayes, R. Romano, M. Roman, S. Ali, et al. (2017). The role of interleukin-1β as a predictive biomarker and potential therapeutic target during clinical ex vivo lung perfusion. *The Journal of Heart and Lung Transplantation 36*(9), 985–995.

Andreasson, A. S., J. H. Dark, and A. J. Fisher (2014). Ex vivo lung perfusion in clinical lung transplantation—state of the art. *European Journal of Cardio-Thoracic Surgery 46*(5), 779–788.

Andreasson, A. S., D. M. Karamanou, C. S. Gillespie, F. Özalp, T. Butt, P. Hill, K. Jiwa, H. R. Walden, N. J. Green, L. A. Borthwick, et al. (2016). Profiling inflammation and tissue injury markers in perfusate and bronchoalveolar lavage fluid during human ex vivo lung perfusion. *European Journal of Cardio-Thoracic Surgery 51*(3), 577–586.

Bland, J. M. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet 327*(8476), 307–310.

Brooks, M. L. (2008). *Exploring Medical Language: A Student-Directed Approach* (7th ed.). Elsevier Health Sciences, USA.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*(4), 434–455.

Buntine, W. (1991). Theory refinement on Bayesian networks. In *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*, pp. 52–60. Morgan Kaufmann Publishers.

Buntine, W. L. (1995). Chain graphs for learning. In *Uncertainty in Artificial Intelligence*, pp. 46–54. Morgan Kaufmann.

Cancer Research UK (2018a). `https://about-cancer.cancerresearchuk.org/about-cancer/non-hodgkin-lymphoma/stages`. Accessed 15-7-2018.

Cancer Research UK (2018b). `https://www.cancerresearchuk.org/about-cancer/hodgkin-lymphoma/symptoms`. Accessed 22-7-2018.

Carbone, P. P., H. S. Kaplan, K. Musshoff, D. W. Smithers, and M. Tubiana (1971). Report of the committee on Hodgkin's disease staging classification. *Cancer Research 31*(11), 1860–1861.

Carlin, B. P. and T. A. Louis (2008). *Bayesian Methods for Data Analysis*. CRC Press.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician 49*(4), 327–335.

Chickering, D. M. and D. Heckerman (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning 29*(2-3), 181–212.

Chipman, H., E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine (2001). The practical implementation of Bayesian model selection. *IMS Lecture Notes-Monograph Series 38*, 124–130.

Clark, T., M. Bradburn, S. Love, and D. Altman (2003). Survival analysis part I: Basic concepts and first analyses. *British Journal of Cancer 89*(2), 232–238.

Collett, D. (2015). *Modelling Survival Data in Medical Research, Third Edition*. CRC Press.

Consul, J. I. (2016). *Flexible Bayesian Modelling of Covariate Effects on Survival*. Ph. D. thesis, University of Newcastle upon Tyne, Newcastle upon Tyne, NE1 7RU, UK.

Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (2007). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks* (1st ed.). Springer.

Cox, D. and D. Oakes (1984). *Analysis of Survival Data*. Chapman and Hall.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society B 34*, 187–220.

Daniels, M. and M. Pourahmadi (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika 89*, 553–566.

Daniels, M. J. and J. W. Hogan (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall/CRC.

Davies, O. L. and P. L. Goldsmith (1972). *Statistical Methods in Research and Production*. Longman Group Ltd., London.

Dellaportas, P., J. J. Forster, and I. Ntzoufras (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing 12*(1), 27–36.

Diaconis, P. and S. L. Zabell (1982). Updating subjective probability. *Journal of The American Statistical Association 77*, 822–830.

Dobson, A. J. and A. Barnett (2008). *An Introduction to Generalized Linear Models*. CRC press.

Edera, A., Y. Strappa, and F. Bromberg (2014). The Grow-Shrink strategy for learning Markov network structures constrained by context-specific independences. In *Ibero-American Conference on Artificial Intelligence*, pp. 283–294. Springer.

El-Galaly, T. C., F. d'Amore, K. J. Mylam, P. de Nully Brown, M. Bøgsted, A. Bukh, L. Specht, A. Loft, V. Iyer, K. Hjorthaug, et al. (2012). Routine bone marrow biopsy has little or no therapeutic consequence for positron emission tomography/computed tomography–staged treatment-naive patients with Hodgkin lymphoma. *Journal of Clinical Oncology 30*(36), 4508–4514.

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC press.

Farrow, M. (2003). Practical building of subjective covariance structures for large complicated systems. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(4), 553–573.

Farrow, M. and M. Goldstein (1993). Bayes linear methods for grouped multivariate repeated measures studies with application to crossover trials. *Biometrika 80*, 39–59.

Farrow, M. and M. Goldstein (2006). Trade-off sensitive experimental design: a multicriterion, decision theoretic, Bayes linear approach. *Journal of Statistical Planning and Inference 136*, 498–526.

Farrow, M. and M. Goldstein (2010). Sensitivity of decisions with imprecise utility tradeoff parameters using boundary linear utility. *International Journal of Approximate Reasoning 51*, 1100–1113.

Ferraris, A., P. Giuntini, and G. Gaetani (1979). Serum lactic dehydrogenase as a prognostic tool for non-Hodgkin lymphomas. *Blood 54*(4), 928–932.

Field, H. (1978). A note on Jeffrey conditionalization. *Philosophy of Science 45*, 361–367.

Freedman, A. S., J. W. Friedberg, J. C. Aster, and A. Lister (2012). Clinical presentation and diagnosis of non-Hodgkin lymphoma. *Waltham (MA): UpToDate Inc.*

Gamerman, D. and H. Lopes (2006). *Markov Chain Monte Carlo*. Boca Raton: Taylor & Francis.

Geiger, D. and D. Heckerman (1994). Learning Gaussian networks. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, pp. 235–243. Morgan Kaufmann Publishers.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2014). *Bayesian Data Analysis* (Third ed.). Chapman & Hall/CRC.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., A. Jakulin, M. G. Pittau, Y.-S. Su, et al. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics 2*(4), 1360–1383.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6*(6), 721–741.

George, E. and R. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

George, E. I., R. E. McCulloch, and R. Tsay (1996). Two approaches to Bayesian model selection with applications. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner 309*, 339–347.

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

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society 57*(6), 1317–1339.

Geyer, C. (1992). Practical Markov Chain Monte Carlo. *Statistical Science 7*(4), 473–511.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Goldstein, M. and S. C. Shaw (2004). Bayes linear kinematics and Bayes linear Bayes graphical models. *Biometrika 91*, 425–446.

Goldstein, M. and D. J. Wilkinson (2000). Bayes linear analysis for graphical models: The geometric approach to local computation and interpretive graphics. *Statistics and Computing 10*(4), 311–324.

Goldstein, M. and D. A. Wooff (2007). *Bayes Linear Statistics: Theory and Methods*. Wiley.

Golub, E. E. and K. Boesze-Battaglia (2007). The role of alkaline phosphatase in mineralization. *Current Opinion in Orthopaedics 18*(5), 444–448.

Gosling, J. P. (2014). Methods for eliciting expert opinion to inform health technology assessment. Technical report, Medical Research Council.

Gosling, J. P., A. Hart, H. Owen, M. Davies, J. Li, and C. MacKay (2013). A Bayes linear approach to weight-of-evidence risk assessment for skin allergy. *Bayesian Analysis 8*, 169–186.

Green, P. J. (1995). Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Hand, D. J., F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski (1994). *A Handbook of Small Datasets*. Chapman and Hall.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97.

Heckerman, D. and D. M. Chickering (1995). Learning Bayesian networks: The combination of knowledge and statistical data. In *Machine Learning*, pp. 20–197.

Henderson, R., M. Jones, and J. Stare (2001). Accuracy of point predictions in survival analysis. *Statistics in Medicine 20*(20), 3083–3096.

Henderson, R., S. Shimakura, and D. Gorst (2002). Modeling spatial variation in leukemia survival data. *Journal of the American Statistical Association 97*(460), 965–972.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science 14*(4), 382–417.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied Logistic Regression*. John Wiley & Sons.

Hosten, A. O. (1990). *BUN and Creatinine in Clinical Methods*. In: Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. Boston: Butterworths.

Howard, R. A. and J. E. Matheson (2005). Influence diagrams. *Decision Analysis 2*(3), 127–143.

Husmeier, D., R. Dybowski, and S. Roberts (2005). *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Springer.

Ibrahim, J. G., M. H. Chen, and D. Sinha (2001). *Bayesian Survival Analysis*. Springer.

Imai, K. and D. A. Van Dyk (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics 124*(2), 311–334.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Wiley.

Jeffrey, R. C. (1965). *The Logic of Decision*. New York: McGrawHill.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks*. London: UCL Press.

Jensen, F. V. and T. V. Nielsen (2007). *Bayesian Networks and Decision Graphs* (Second ed.). Springer.

Ji, Z., Q. Xia, and G. Meng (2015). A review of parameter learning methods in Bayesian network. In *International Conference on Intelligent Computing Theories and Applications*, pp. 3–12. Springer.

Jiang, X., D. Xue, A. Brufsky, S. Khan, and R. Neapolitan (2014). A new method for predicting patient survivorship using efficient Bayesian network learning. *Cancer Informatics 13*, CIN–S13053.

Jones, O., R. Maillardet, and A. Robinson (2014). *Introduction to Scientific Programming and Simulation Using R*. CRC Press.

Kalbfleisch, J. D. and R. L. Prentice (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Kjaerulff, U. B. and A. L. Madsen (2005). Probabilistic networks−An introduction to Bayesian networks and influence diagrams.

Kjaerulff, U. B. and A. L. Madsen (2013). *Bayesian Networks and Influence Diagrams*. Springer.

Klein, J. P. and M. L. Moeschberger (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.

Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Korb, K. B. and A. E. Nicholson (2004). *Bayesian Artificial Intelligence*. London: Chapman and Hall/CRC.

Kraisangka, J. and M. J. Druzdzel (2014). Discrete Bayesian network interpretation of the Cox's proportional hazards model. In *European Workshop on Probabilistic Graphical Models*, pp. 238–253. Springer.

Kułaga, T. (2006). The Markov Blanket Concept in Bayesian Networks and Dynamic Bayesian Networks and Convergence Assessment in Graphical Model Selection Problems. Master's thesis, Jagiellonian University.

Kuo, L. and B. Mallick (1998). Variable selection for regression models. *The Indian Journal of Statistics: Series B* (60), 65–81.

Langseth, H. (1998). Analysis of survival times using Bayesian networks. In *Proceeding of the ninth European Conference on Safety and Reliability (ESREL)*, pp. 647–654.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford Science Publications.

Lauritzen, S. L. and D. J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 157–224.

Lauritzen, S. L. and N. Wermuth (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 31–57.

Lepar, V. and P. P. Shenoy (1998). A comparison of Lauritzen-Spiegelhalter, Hugin, and Shenoy-Shafer architectures for computing marginals of probability distributions. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 328–337. Morgan Kaufmann Publishers.

Lesaffre, E. and A. Lawson (2012). *Bayesian Biostatistics*. Chichester, West Sussex: Wiley.

Link, W. A. and M. J. Eaton (2011). On thinning of chains in MCMC. *Methods in Ecology and Evolution 3*(1), 112–115.

Little, R. J. and D. B. Rubin (2014). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Longmore, M., I. Wilkinson, A. Baldwin, and E. Wallin (2014). *Oxford Handbook of Clinical Medicine*. Oxford University Press.

Lucas, P. J. F., H. Boot, and B. G. Taal (1998). Computer-based decision-support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine 37*, 206–219.

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer Science & Business Media.

Margaritis, D. (2003). *Learning Bayesian Network Model Structure from Data*. Ph. D. thesis, Carnegie Mellon University, School of Computer Science.

Martelli, M., A. J. Ferreri, C. Agostinelli, A. Di Rocco, M. Pfreundschuh, and S. A. Pileri (2013). Diffuse large b-cell lymphoma. *Critical Reviews in Oncology/Hematology 87*(2), 146–171.

Mateo Cerdán, J. L. (2010). *On the Use of Probabilistic Graphical Models: Bayesian and Dependency Networks*. Ph. D. thesis, Universidad de Castilla-La Mancha, Computing Systems Department.

Maton, A. (1997). *Human Biology and Health*. Prentice Hall.

Metropolis, N., W. A. Rosenbluth, M. N. Rosenbluth, and A. H. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Miranda, R. N., J. D. Khoury, and L. J. Medeiros (2013). Diffuse large b-cell lymphoma, not otherwise specified. In *Atlas of Lymph Node Pathology*, pp. 237–245. Springer.

Mitchell, T. and J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), 1023–1032.

Molenberghs, G. and M. Kenward (2007). *Missing Data in Clinical Studies*. John Wiley & Sons.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer.

Moore, D. (2016). *Applied Survival Analysis Using R*. Use R! Springer International Publishing.

Moore, E. H. and R. W. Barnard (1935). General analysis. *Memoirs of the American Philosophical Society 1*, 147–209.

Mosteller, F. and J. W. Tukey (1977). Data Analysis and Regression: a Second Course in Statistics. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*.

Murphy, K. P. and S. Russell (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph. D. thesis, University of California, Berkeley, California.

Naylor, J. C. and A. F. M. Smith (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics 31*, 214–225.

Neapolitan, R. E. (2003). *Learning Bayesian Networks*. Morgan Kaufmann.

Needham, C. J., J. R. Bradford, A. J. Bulpitt, and D. R. Westhead (2007). A primer on learning in Bayesian networks for computational biology. *PLoS Computational Biology 3*(8), e129.

NHS (2016). Lung transplant. `https://www.nhs.uk/conditions/lung-transplant/`. Accessed 20-05-2016.

NHS (2018). Non-Hodgkin lymphoma. `https://www.nhs.uk/conditions/non-hodgkin-lymphoma/`. Accessed 28-6-2018.

Oken, M. M., R. H. Creech, D. C. Tormey, J. Horton, T. E. Davis, E. T. Mcfadden, and P. P. Carbone (1982). Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology 5*(6), 649–656.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, California: Morgan Kaufmann Publishers.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge University Press.

Pearl, J. and T. S. Verma (1995). A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, Volume 134, pp. 789–811. Elsevier.

Penrose, R. (1955). A generalized inverse for matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 51, pp. 406–413. Cambridge University Press.

Pfreundschuh, M., A. D. Ho, E. Cavallin-Stahl, M. Wolf, R. Pettengell, I. Vasova, A. Belch, J. Walewski, P.-L. Zinzani, W. Mingrone, et al. (2008). Prognostic significance of maximum tumour (bulk) diameter in young patients with good-prognosis diffuse large-b-cell lymphoma treated with chop-like chemotherapy with or without rituximab: an exploratory analysis of the mabthera international trial group (mint) study. *The Lancet Oncology 9*(5), 435–444.

Plummer, M. (2013). Package "rjags". `http://cran.r-project.org/web/packages/rjags/rjags.pdf`.

Plummer, M. (2017). *JAGS: Just Another Gibbs Sampler, Version 4.3.0*. `http://mcmc-jags.sourceforge.net/`.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). CODA: convergence diagnosis and output analysis for MCMC. *R news 6*(1), 7–11.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika 86*(3), 677–690.

Press, S. J. (2009). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. John Wiley & Sons.

Proctor and Taylor (2000). A practical guide to continuous population-based data collection (pace): a process facilitating uniformity of care and research into practice. *QJM: An International Journal of Medicine 93*(2), 67–73.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Raghunathan, T. (2015). *Missing Data Analysis in Practice*. CRC Press.

Riggelsen, C. (2006). *Approximation Methods for Efficient Learning of Bayesian Networks*. Ph. D. thesis, Universiteit Utrecht, the Dutch Research School for Information and Knowledge Systems.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Russell, S. J. and P. Norvig (2016). *Artificial Intelligence: A Modern Approach*. Malaysia; Pearson Education Limited.

Schott, J. R. (2016). *Matrix Analysis for Statistics*. John Wiley & Sons.

Scutari, M. (2010). Learning Bayesian networks with the bnlearn R Package. *Journal of Statistical Software 35*(3), 1–22.

Scutari, M. and J. Denis (2014). *Bayesian Networks with Examples in R*. CRC/Chapman and Hall.

Scutari, M. and R. Ness (2012). Package "bnlearn". `https://cran.r-project.org/web/packages/bnlearn/bnlearn.pdf`.

Sesen, M. B., A. E. Nicholson, R. Banares-Alcantara, T. Kadir, and M. Brady (2013). Bayesian networks for clinical decision support in lung cancer care. *PLOS one 8*(12), 1–13.

Sieniawski, M., M. Farrow, X. Zhao, J. Wilkinson, T. Mainou-Fowler, J. White, J. MacIntyre, and S. J. Proctor (2009). A novel Bayesian prognostic index for Diffuse Large B-Cell lymphoma: A new powerful tool for prediction of outcome. *Blood 114*, 1666.

Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks (1996). Bugs 0.5: Bayesian inference using Gibbs sampling manual (version ii). *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*, 1–59.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.

Studenỳ, M. (1998). Bayesian networks from the point of view of chain graphs. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 496–503.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*(398), 528–540.

Tian, G.-L., M. T. Tan, and K. W. Ng (2009). *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation*. Chapman and Hall/CRC.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*(393), 82–86.

Townsend, P., P. Phillimore, and A. Beattie (1988). *Health and Deprivation: Inequality and the North*. Routledge.

Verdurmen, N. (2003). A model for Credit Scoring: Combining Bayesian Networks with Survival Analysis Techniques. Master's thesis, Utrecht University.

Walli, G. M. (2010). Bayesian Variable Selection in Normal Regression Models. Master's thesis, Johannes Kepler Universitat Linz.

Wang, H. et al. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis 10*(2), 351–377.

Watkins, D. S. (2004). *Fundamentals of Matrix Computations*. John Wiley & Sons.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* (14-15), 1871–1879.

West, M., P. J. Harrison, and H. S. Migon (1985). Dynamic generalized linear models and Bayesian forecasting. *80*, 73–83.

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

Wilson, K. J. (2011). *Belief Representation for Counts in Bayesian Inference and Experimental Design*. Ph. D. thesis, University of Newcastle upon Tyne, Newcastle upon Tyne, NE1 7RU, UK.

Wilson, K. J. and M. Farrow (2010). Bayes linear kinematics in the analysis of failure rates and failure time distributions. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 224*, 309–321.

Wilson, K. J. and M. Farrow (2017). Bayes linear kinematics in a dynamic survival model. *International Journal of Approximate Reasoning 80*, 239–256.

Wilson, K. J., J. Quigley, T. Bedford, and L. Walls (2013). Bayes linear Bayes graphical models in the design of optimal test strategies. *International Journal of Performability Engineering 9*, 715–728.

Witteveen, A., G. F. Nane, I. M. Vliegen, S. Siesling, and M. J. IJzerman (2018). Comparison of logistic regression and Bayesian networks for risk prediction of breast cancer recurrence. *Medical Decision Making*, 1–12.

Yadav, C., A. Ahmad, B. D'Souza, A. Agarwal, M. Nandini, K. A. Prabhu, and V. D'Souza (2016). Serum lactate dehydrogenase in non-Hodgkin's lymphoma: A prognostic indicator. *Indian Journal of Clinical Biochemistry 31*(2), 240–242.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Goel, P. and Zellner, A., Eds., Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. Elsevier, New York.

Zhang, Z. and J. Sun (2010). Interval censoring. *Statistical Methods in Medical Research 19*(1), 53–70.

Zhao, X. (2010). *Bayesian Survival Analysis for Prognostic Index Development with Many Covariates and Missing Data*. Ph. D. thesis, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK.

Zhou, M. (2015). *Empirical Likelihood Method in Survival Analysis*. CRC Press.