BAYESIAN METHODS AND APPLICATIONS USING WINBUGS

by

Saman Muthukumarana

B.Sc., University of Sri Jayewardenepura, Sri Lanka, 2002

M.Sc., Simon Fraser University, 2007

A Thesis submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

in the Department

of

Statistics and Actuarial Science

SIMON FRASER UNIVERSITY

Summer 2010

APPROVAL

Name: Saman Muthukumarana

Degree: Doctor of Philosophy

Title of Thesis: Bayesian Methods and Applications using WinBUGS

Examining Committee: Dr. Derek Bingham
Chair

Dr. Tim Swartz, Senior Supervisor

Dr. Paramjit Gill, Supervisor

Dr. Carl Schwarz, Supervisor

Dr. Jiguo Cao, SFU Examiner

Dr. Paul Gustafson, External Examiner
University of British Columbia

Date Approved: _____

Abstract

In Bayesian statistics we are interested in the posterior distribution of parameters. In simple cases we can derive analytical expressions for the posterior. However in most situations, the posterior expectations cannot be calculated analytically due to the complexity of the integrals. This thesis develops some new methodologies for applied problems which deal with multidimensional parameters, complex model structures and complex likelihood functions.

The first project is concerned with the simulation of one-day cricket matches. Given that only a finite number of outcomes can occur on each ball, a discrete generator on a finite set is developed where the outcome probabilities are estimated from historical data. The probabilities depend on the batsman, the bowler, the number of wickets lost, the number of balls bowled and the innings. The proposed simulator appears to do a reasonable job at producing realistic results. The simulator allows investigators to address complex questions involving one-day cricket matches.

The second project investigates the suitability of Dirichlet process priors in the Bayesian analysis of network data. Dirichlet process priors allow the researcher to weaken prior assumptions by going from a parametric to a semiparametric framework. This is important in the analysis of network data where complex nodal relationships rarely allow a researcher the confidence in assigning parametric priors. The Dirichlet process also provides a clustering mechanism which is often suitable for network data where groups of individuals in a network can be thought of as arising from the same cohort. The approach is highlighted on two network models and implemented using WinBUGS.

The third project develops a Bayesian latent variable model to analyze ordinal survey data. The data are viewed as multivariate responses arising from a class of continuous

latent variables with known cut-points. Each respondent is characterized by two parameters that have a Dirichlet process as their joint prior distribution. The proposed mechanism adjusts for classes of personality traits. As the resulting posterior distribution is complex and high-dimensional, posterior expectations are approximated by MCMC methods. The methodology is tested through simulation studies and illustrated using student feedback data from course evaluations at Simon Fraser University.

Keywords: Bayesian latent variable models, Clustering, Dirichlet process, Markov chain Monte Carlo, Simulation, WinBUGS

Acknowledgements

There are many people who helped me to be successful in my academic career over last few years at SFU. It is impossible to name and thank all of them in a page.

I'm deeply indebted to my senior supervisor Dr. Tim Swartz for mentoring me in limitless ways. I'm grateful for the freedom he gave me to explore new directions and understanding and nurturing my research interests. The full and continuous financial support provided during the last five years allowed me to entirely focus on my studies. I was very lucky to have such an amazing mentor in my life.

A big thank-you goes to Dr. Paramjit Gill and Dr. Pulak Ghosh for having stimulating discussions which helped me accomplish my goals. I also want to thank Dr. CJS for giving me the opening to work on the CJS model which yielded my first paper to appear in CJS. There are many others in the Department of Statistics and Actuarial Science who helped in many ways during my life at SFU. Special thanks to Robin Insley, Dr. Richard Lockhart, Dr. Joan Hu, Dr. Jinko Graham, Dr. Brad McNeney, Dr. Charmaine Dean, Dr. Derek Bingham, Dr. Tom Loughin and Ian Bercovitz. I also thank Dr. Jiguo Cao and Dr. Paul Gustafson for their comments and suggestions on the thesis. I must also thank Dr. Sarath Banneheka for initiating my interest in Statistics.

I offer my sincere gratitude to Sadika, Kelly and Charlene for your kindness, help and promptness on all the matters. I would also like to thank all of the graduate students for their friendship. Especially: Ryan, Matt, Crystal, Pritam, Kyle, Dean, Wei, Kelly, Jean, Carolyn, Simon, Chunfang, Elizabeth, Lihui, Wendell, Cindy, Vivien, Joslin, Rianka, Jervyn and many more .... Finally, and most importantly, I appreciate Aruni for her patient support and sacrifices. This would not have been finished without her understanding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Bayesian paradigm

The Bayesian framework was introduced by the Reverand Thomas Bayes (1702-1761) and Bayesian estimation was first used by LaPlace in 1786. Today, Bayesian statistics is widely used by researchers in diverse fields due to significant computational advancements including MCMC, BUGS and WinBUGS software. Researchers in many fields have embraced the Bayesian approach due to its capacity to handle complexity in real world problems. The Bayesian approach has many attractive features over frequentist statistics. In particular, missing data and latent variables often pose no difficulties in Bayesian analyses. The Bayesian approach also provides a way to include expert prior knowledge concerning parameters of interest.

In a frequentist approach, the data are taken as random while parameters are considered fixed. In a Bayesian approach, parameters themselves follow a probability distribution. Furthermore, parameters may be model parameters, missing data or events that are not observed (latent). Frequentist methods also typically rely upon approximations and asymptotic results and these cases are only valid for large samples. Moreover, frequentist methods often replace missing data with guesses and then analyze the data as though the guesses were known or else delete records for subjects that have even one missing value.

This thesis develops some new methodologies for applied problems which deal with multidimensional parameters, complex model structures, latent variables, missing data and

complex likelihood functions. The following components are required in order to carry out a Bayesian analysis:

- the prior distribution of the parameters
- the likelihood corresponding to the data

A prior distribution of a parameter quantifies the uncertainty about the parameter before the data are observed. It is important that priors are selected such that they represent the best knowledge about parameters. If it is not possible, we may be able to use non-informative priors which often produce useful results provided that there is sufficient information in the likelihood. Recall that Bayes formula gives the posterior distribution

$$\pi(\theta \mid y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}$$

where $f(y \mid \theta)$ is the likelihood, $\pi(\theta)$ is the prior density or probability mass function and $f(y)$ is the inverse normalizing constant given by

$$f(y) = \int f(y \mid \theta)\pi(\theta)d\theta.$$

Here θ can be a scalar or a vector of parameters and y is the vector of observed data. In many Bayesian analyses, it is not necessary to calculate the inverse normalizing constant.

In order to perform inferences about components of θ, one is typically faced with high-dimensional integrals. For example, the posterior mean of θ_1 is given by

$$E(\theta_1 \mid y) = \int \theta_1 \pi(\theta_1 \mid y)d\theta_1 = \int \theta_1 \left[\int \pi(\theta \mid y)d\theta_{(-1)}\right] d\theta_1$$

where $\theta_{(-1)}$ is the vector θ with θ_1 removed. In simple models, integration problems can sometimes be avoided by choosing particular types of priors. If the prior and likelihood are natural conjugate distributions, then the posterior is in the same family as the prior and integrals may be tractable. For more complex models, the calculation of integrals is often difficult and sometimes impossible. Sometimes numerical approaches such as quadrature and Laplaces method can be used to approximate the expectations. Evans and Swartz

(1995) provide a discussion of the major techniques available for the approximation of integrals in statistics.

In theory, the functional form of the posterior density provides a complete description of the uncertainty in the parameters. However, to gain insight with respect to the posterior, posterior expectations in the form of integrals are typically desired. In the case of complex posteriors, the integrals can not be evaluated analytically. Instead, simulation procedures are often used to sample variates from the posterior. In a simulation context, sampling directly from the posterior may not be easy in complex problems and there are some alternative sampling strategies which may be useful. The most widely used sampling methods are

- importance sampling
- Markov chain Monte Carlo (MCMC)

In Evans and Swartz (1995), these two methods are discussed where Markov chain methods are recommended for high-dimensional problems such as in the problems considered in this thesis. In MCMC, variates are drawn from a distribution which has the posterior distribution as its equilibrium distribution. In MCMC, output may be averaged to obtain approximations to posterior expectations. A Markov chain is a random process where the variate at iteration i depends only on the variate at iteration $i-1$. Various algorithms have been developed to implement MCMC. The most popular algorithms are

- Metropolis-Hastings
- Gibbs sampling

Given previously generated variates $\theta^{(1)}, \ldots, \theta^{(k-1)}$, the Metropolis-Hastings algorithm proceeds by generating θ^* from a proposal density $q(\theta, \theta^{(k-1)})$. Note that the proposal may depend on the previous variate $\theta^{(k-1)}$. This generated value θ^* is accepted (ie $\theta^k = \theta^*$) with probability

$$min\left(1, \frac{q(\theta^{(k-1)}, \theta^*)\pi(\theta^* \mid y)}{q(\theta^*, \theta^{(k-1)})\pi(\theta^{(k-1)} \mid y)}\right). \tag{1.1}$$

If θ^* is not accepted, then $\theta^{(k)} = \theta^{(k-1)}$. The rate at which the new values are accepted is called the acceptance rate. The process is repeated to obtain a sequence $\theta^{(1)}, \theta^{(2)}, \ldots$ where $\theta^{(k)}$ is approximately a realization from the posterior for sufficiently large k. The Metropolis-Hastings algorithm requires an initial value $\theta^{(0)}$ in order to start the simulation. The choice of initial value may effect the rate of convergence of the algorithm. Initial values which are far away from the range covered by the posterior distribution often lead to chains that take more iterations to attain convergence.

The Gibbs sampling algorithm is a special case of the Metropolis-Hastings algorithm in which samples are drawn by turning the multivariate problem into a sequence of lower-dimensional problems. In Gibbs sampling, parameters are generated from distributions with a 100% acceptance rate.

Fortunately, the software package WinBUGS implements MCMC methods using the Metropolis-Hastings or Gibbs algorithm. The default option in WinBUGS for well behaved models with log concave densities is the Gibbs sampling algorithm. However, Metropolis-Hastings is invoked for nonstandard models. In WinBUGS, we need only specify the likelihood, the prior, the observed data and the initial values. WinBUGS then produces an appropriate Markov chain. Clearly, WinBUGS requires much less MCMC programming than if one was to program the MCMC simulations.

However, we need to make sure that a sequence has converged before inferences are obtained. The number of iterations taken for the practical convergence to the stationary distribution depends on various factors including

- the complexity of the model (models with few parameters generally converge faster)
- whether the prior and likelihood are conjugate
- the closeness of initial values to their respective posterior means
- the parameterization of the problem
- the sampling scheme adopted

The number of iterations prior to convergence is called the burn-in, and we typically discard these variates for the purpose of inference. WinBUGS provides several statistics

and graphical tools to check the convergence of Markov chains. Brooks and Gelman (1997) discuss some these methods.

1.2 MCMC methods using WinBUGS

WinBUGS is a product of the BUGS (Bayesian Inference Using Gibbs Sampling) project which is a joint program of the Medical Research Council of Biostatistics Unit at Cambridge University and the Department of Epidemiology and Public Health of Imperial College at St. Mary's Hospital in London. The software is freely distributed from their web page at (www.mrc-bsu.cam.ac.uk/bugs). Models can be implemented in two ways:

- using the script language
- using the graphical feature, DoodleBUGS which allows the specification of models in terms of a directed graph

The range of model types that can be fitted in WinBUGS is very large. A wide variety of linear and nonlinear models, including the standard set of generalised linear models, spatial models and latent variable models can be fitted. In addition, a range of prior distributions is available with standard defaults. We believe that WinBUGS is a very handy tool in fitting complex models although it is a difficult and frustrating package to master. One of the main issues of WinBUGS is that there is a great amount of embedded statistical theory associated with MCMC. Lack of knowledge of relevant theory can sometimes lead inexperienced users into a state of false security. There are other issues associated with Bayesian modelling, such as choice of priors and initial values that are also relevant. Bayesian analysis using WinBUGS requires three major tasks as follows:

- model specification
- running the model
- Bayesian inference using MCMC output

In WinBUGS, there are three types of nodes referred to as constant, stochastic and deterministic. Constant nodes are used to declare constant terms. Stochastic nodes represent data or parameters that are assigned a distribution. Currently WinBUGS provides 23 familiar distributions. Deterministic nodes are logical expressions of other nodes. Logical expressions can be built using the operators +, -, *, / and various WinBUGS functions. Note that WinBUGS has some special syntax which differs from other languages such as Splus and C++. As an example, WinBUGS requires that each node appear exactly once on the left hand side of an equation. During and after MCMC simulation, WinBUGS provides several numerical and graphical summaries for the parameters. The dynamic trace plots for each parameter, Brooks-Gelman-Rubin convergence statistics (Brooks and Gelman 1997) and the autocorrelation plots of the sequences are handy tools for investigating convergence to the equilibrium distribution. The Brooks-Gelman-Rubin convergence statistics are graphical tools for assessing convergence of multiple chains to the same distribution. Since the simulations from multiple chains are independent, evidence of convergence is based on the equality of within chain variability and between chain variability. CODA (Convergence Output and Diagnostic Analysis) software is also easily accessible from other software platforms for further analysis.

1.3 Organization of the thesis

This thesis involves model development, computation and inferences associated with three applied problems which deal with multidimensional parameters, complex model structures and complex likelihood functions. The commonality of the three problems is that they each involve a Bayesian analysis implemented via WinBUGS. Chapters 2, 3 and 4 individually correspond to one of the three problems. Each chapter is written in a stand-alone fashion and corresponds closely to the technical paper upon which it is based.

1.3.1 One-day international cricket

Chapter 2 is concerned with the simulation of one-day cricket matches. In one day international cricket matches, there are an endless number of questions that are not amenable to

experimentation or direct analysis but could be easily addressed via simulation. A good simulator for ODI cricket is required to provide reliable answers to such questions. Given that only a finite number of outcomes can occur on each ball that is bowled, a discrete generator on a finite set is developed via a Bayesian latent variable model where the outcome probabilities are estimated from historical data involving one-day international cricket matches. The probabilities depend on the batsman, the bowler, the number of wickets lost, the number of balls bowled and the innings. The proposed simulator appears to do a reasonable job at producing realistic results. The simulator allows investigators to address complex questions involving one-day cricket matches. This work was published in the Canadian Journal of Statistics in 2009.

1.3.2 Social network models

Chapter 3 considers the use of Dirichlet process priors in the statistical analysis of social network data. A social network is a data structure which considers relationships (ties) between nodes. The nodes can be people, groups, organizations, cities, countries, etc. The ties can be transfer of goods, money, information, political support, friendship, etc. There are many applications of social network analysis such as citation analysis which identifies influential papers in a research area, dynamics of the spread of disease in epidemiology, identifying most effective areas for product/service distributions in business and telecommunications and coalition formation dynamics in political science. In social networks, it is possible that there are partitions of the data such that data within classes are similar. There may also be some nodes that appear to play network roles or some may be isolated from groups. These types of complex nodal relationships rarely allow a researcher confidence in assigning parametric priors. Dirichlet process priors allow the researcher to weaken prior assumptions by going from a parametric to a semiparametric framework. In addition, the standard normality assumption does not allow outlying subjects and also conclusions may be sensitive to violations from normality. The normality assumption is relaxed by the use of Dirichlet process priors. The Dirichlet process has a secondary benefit due to the fact that its support is restricted to discrete distributions. This provides a clustering mechanism which is often suitable for network data where groups of individuals in a network can be thought

of as arising from the same cohort. Most importantly, in the proposed model, clustering takes place as a part of the model and data decide the clustering structure. The approach is highlighted on two network models. This work has been accepted for publication in the Australian and New Zealand Journal of Statistics (2010).

1.3.3 Ordinal survey data

Chapter 4 presents a Bayesian latent variable model used to analyze ordinal response survey data. The ordinal response data are viewed as multivariate responses arising from a class of continuous latent variables with known cut-points. Each respondent is characterized by two parameters that have a Dirichlet process as their joint prior distribution. The key aspect of the Dirichlet process in this application is that the personality trait parameters of the survey respondents have support on a discrete space and this enables the clustering of personality types. The proposed mechanism adjusts for classes of personality traits. As the resulting posterior distribution is complex and high-dimensional, posterior expectations are approximated by MCMC methods. As a by-product of the proposed methodology, one can identify survey questions where the corresponding performance has been poor or exceptional. The methodology is tested through simulation studies and illustrated using student feedback data in teaching and course evaluations at Simon Fraser University (SFU). Goodness-of fit is investigated using prior predictive methods. This work has been submitted and is currently under review.

The thesis concludes with a discussion in chapter 5. The WinBUGS code for the three problems is provided in Appendix A, B and C respectively. In Appendix D, R code is provided which is used in the goodness-of-fit approach from chapter 4.

Chapter 2

Modelling and Simulation for One-Day Cricket

2.1 Introduction

Simulation is a practical and powerful tool that has been used in a wide range of disciplines to investigate complex systems. When a simulation model is available, it is typically straightforward to address questions concerning the occurrence of various phenomena. One simply carries out repeated simulations and observes the frequency which the phenomena occur.

In one-day international (ODI) cricket, there are an endless number of questions that are not amenable to experimentation or direct analysis but could be easily addressed via simulation. For example, on average, would England benefit from increasing the number of runs scored by changing the batting order of their third and sixth batsmen? As another example, what percentage of time would India be expected to score more than 350 runs versus Australia in the first innings?

To provide reliable answers to questions such as these, a good simulator for one-day cricket matches is required. Surprisingly, the development of simulators for one-day cricket is a topic that has not been vigorously pursued by academics. In the pre-computer days, Elderton (1945) and Wood (1945) fit the geometric distribution to individual runs scored based on results from test cricket. Kimber and Hansford (1993) argue against the geometric

distribution and obtain probabilities for selected ranges of individual scores in test cricket using product-limit estimators. More recently, Dyte (1998) simulates batting outcomes between a specified test batsman and bowler using career batting and bowling averages as the key inputs without regard to the state of the match (e.g. the score, the number of wickets lost, the number of overs completed). Bailey and Clarke (2004, 2006) investigate the impact of various factors on the outcome of ODI cricket matches. Some of the more prominent factors include home ground advantage, team quality (class) and current form. Their analysis is based on the modelling of runs using the normal distribution.

In a non-academic setting, there are currently more than 100 cricket games that have been developed for use on personal computers and gaming stations where some of the games rely on simulation techniques (see www.cricketgames.com/games/index.htm for a comprehensive survey of games and reviews). However, in all of the games that we have inspected, the details of the simulation procedures have not been revealed. Moreover, the games that we have inspected suffer in various ways from a lack of realism.

One-day cricket was introduced in the 1960's as an alternative to traditional forms of cricket that may take up to five days to complete. With more aggressive batting, colourful uniforms and fewer matches ending in draws, one-day cricket has become extremely popular. The ultimate event in ODI cricket takes place every four years where the World Cup of Cricket is contested.

One-day cricket has some similarities with the sport of baseball. The game is played between two teams of 11 players where a batting team attempts to score runs against a fielding team until the first *innings* terminate. At this stage, the fielding team goes "to bat" and attempts to score more runs before the second innings terminate. A bowler on the fielding team "bowls" balls in groups of six referred to as *overs*. Bowling takes place only from one end of the *pitch* during an over, and the bowling end changes on the subsequent over. A batsman on the batting team faces a bowled ball with a bat and attempts to score *runs*. Runs are scored by running between two *stumps* located at opposite ends of the pitch. The running aspect involves two batsmen known as a *partnership* where each partner is running to the opposite stump. They may score $0, \ldots, 6$ runs according to the number of traversals between the stumps. Therefore, the batsman in the partnership who

is located at the batting end faces the next ball. An innings terminates when either 50 overs are completed or 10 *wickets* are lost. A loss of a wicket can occur in various ways with the three most common being (i) a ball is caught in midair after having been batted, (ii) a bowled ball hits the stump located behind the batsman, and (iii) a batsman is *run-out* before reaching the nearest stump. When a wicket is lost, a new batsman is introduced in the batting order. This is a very brief introduction to the rules of one-day cricket, and more information is provided in the paper as required. More detail on the rules (laws) of cricket is available from the Lord's website (www.lords.org).

In this paper, we develop a simulator for one-day cricket matches. The approach extends the work of Swartz, Gill, Beaudoin and de Silva (2006) who investigate the problem of optimal batting orders in one-day cricket for the first innings. Whereas the model used in Swartz et al. (2006) ignores the effect of bowlers, the model used in this paper is more realistic in that specific batsman/bowler combinations are considered. In addition, we now provide a method of generating runs in the second innings. Given that only a finite number of outcomes can occur on each ball that is bowled, a discrete generator on a finite set is developed where the outcome probabilities are estimated from historical data involving ODI cricket matches. The probabilities depend on the batsman, the bowler, the number of wickets lost, the number of balls bowled and the current score of the match. The probabilities are obtained using a Bayesian latent variable model which is fitted using WinBUGS software (Spiegelhalter, Thomas, Best and Lunn 2004).

In Section 2.2, we develop a simulator for ODI cricket which is based upon a Bayesian latent variable model. Particular attention is given to second innings batting where the state of the match (e.g. score, wickets, overs) affects the aggressiveness of batsmen. In Section 2.3, the simulator is constructed using data from recent ODI matches. The methodology is developed for generating runs in the second innings in Section 2.4. In Section 2.5, we consider the adequacy of the approach by comparing simulated results against actual data. In Section 2.6, we demonstrate how the simulator can be used to address questions of interest. We conclude with a short discussion in Section 2.7.

2.2 Simulation

We consider the simulation of runs in the first innings for predetermined batting and bowling orders. We initially investigate the first innings runs since second innings strategies are affected by the number of runs scored in the first innings. By a predetermined batting and bowling order, we mean that a set of rules has been put in place which dictates the batsman and bowler at any given point in the match. These rules could be simple such as maintaining a fixed batting and bowling order. The rules for determining batting and bowling orders could also be very complex. For example, the rules could be Markovian in nature where a specified bowler may be substituted at a state in the match dependent upon the number of wickets lost, the number of overs, the number of runs and the current batsmen. The key point is that they need to be specified in advance for the purpose of simulation.

In one-day cricket, there are a finite number of outcomes arising from each ball bowled. Suppose that the first innings terminate on the m-th ball bowled where $m \leq 300$. Ignoring certain rare events (such as scoring 5 runs), and temporarily ignoring wide-balls and no-balls, let X_b denote the outcome of the b-th ball bowled, $b = 1, \ldots, m$ where

$$X_b = \begin{cases} 1 & \text{if a wicket is taken} \\ 2 & \text{if the batsman scores 0 runs} \\ 3 & \text{if the batsman scores 1 run} \\ 4 & \text{if the batsman scores 2 runs} \\ 5 & \text{if the batsman scores 3 runs} \\ 6 & \text{if the batsman scores 4 runs} \\ 7 & \text{if the batsman scores 6 runs} \end{cases} \tag{2.1}$$

and set $X_{m+1} = \cdots = X_{300} = 0$. Note that the coding in (2.1) includes the possibility of scoring due to byes and and leg byes. Byes and leg byes occur when the batsman has not hit the ball with his bat but decides to run.

Using square brackets to generically denote probability mass functions, the joint distribution of $X_1, \ldots, X_{300}$ can be written as

$$[X_1, \ldots, X_{300}] = [X_{300} \mid X_0, \ldots, X_{299}]\,[X_{299} \mid X_0, \ldots, X_{298}] \cdots [X_2 \mid X_0, X_1]\,[X_1 \mid X_0] \tag{2.2}$$

where we define $X_0 = 0$ for notational convenience. Note that the conditional probabilities in (2.2) suggest that the scoring distribution for a given ball depends on the scoring up to that point in time in the first innings. The proposed simulation algorithm is facilitated by the structure in (2.2) whereby the first outcome X_1 is generated for the first ball bowled, then the second outcome X_2 is generated for the second ball bowled conditional on X_1. The first innings continue until either the overs are completed ($b = 300$) or all wickets are lost (wickets=10). Let v denote the probability of a wide-ball or a no-ball. The simulation algorithm generates a uniform(0,1) random variable u, and if $u < v$, this signals a wide-ball or no-ball condition. In this case, a single run is added to the batting team but the ball is not counted. In addition, further runs may be scored on the no-ball or wide-ball according to the probability ϕ_k for the k-th outcome, $k = 1, \ldots, 7$. The following simulation algorithm generates the number of runs R scored in the first innings:

wickets $= 0$

$R = 0$

for $b = 1, \ldots, 300$

  if wickets $= 10$

  then

    $X_b = 0$

  else

    generate $u \sim \text{uniform}(0, 1)$ $\star$

    if $u < v$

    then

      generate $Y \sim \text{multinomial}(1, \phi_1, \ldots, \phi_7)$

      $R \leftarrow R + 1 + I(Y = 3) + 2I(Y = 4) + 3I(Y = 5) + 4I(Y = 6) + 6I(Y = 7)$

      goto step $\star$

    else

      generate $X_b \sim [X_b \mid X_0, \ldots, X_{b-1}]$

      $R \leftarrow R + I(X_b = 3) + 2I(X_b = 4) + 3I(X_b = 5) + 4I(X_b = 6) + 6I(X_b = 7)$

      wickets $\leftarrow$ wickets $+ I(X_b = 1)$

For the sake of simplicity, the above algorithm does not distinguish a run-out from other forms of wickets. Runs may still be accumulated on a ball prior to a run-out, and the dismissed batsman may be either of the two batsmen in the partnership. We remark that we have estimated the probability of run-outs and have accounted for run-outs in our Fortran implementation of the above algorithm. The proposed simulation algorithm is simple and requires only that we be able to generate from the multinomial$(1, \phi_1, \ldots, \phi_7)$ distribution and the conditional finite discrete distributions given by $[X_b \mid X_0, \ldots, X_{b-1}]$. In the following subsection, we describe a method to compute the conditional distributions by modelling ball by ball outcomes in one-day cricket matches.

2.3 Modelling

The conditional distributions $[X_b \mid X_0, \ldots, X_{b-1}]$ depend on many factors including

- the batsman
- the bowler
- the number of wickets lost
- the number of balls bowled
- the current score of the match
- the opposing team
- the location of the match
- the coach's advice
- the condition of the pitch, etc.

For the first innings, we consider the first four factors and define p_{ijwbk} as the probability corresponding to outcome $k = 1, \ldots, 7$ as described in (2.1), where the the i-th batsman, $i = 1, \ldots, I$ faces the j-th bowler, $j = 1, \ldots, J$ when $w = 0, \ldots, 9$ wickets have been lost and the b-th ball is about to be bowled, $b = 1, \ldots, 300$. Since wide-balls and no-balls have been excluded as possible outcomes of X_b, we have $\sum_k p_{ijwbk} = 1$ for all i, j, w, b. With estimates $\hat{p}_{ijwbk}$, the simulation algorithm generates outcomes according to

$$\text{Prob}(X_b = k \mid X_0, \ldots, X_{b-1}) = \hat{p}_{ijwbk}. \tag{2.3}$$

Our data are based on 472 ODI matches from January 2001 until July 2006 amongst the 10 full member nations of the International Cricket Council (ICC). These matches are those for which ball by ball commentary is available on the Cricinfo website (www.cricinfo.com) and include almost all matches amongst the 10 nations during the specified time period. We note that a Powerplay rule was introduced for a 10 month trial period beginning August 2005 and the Supersub rule was introduced for a portion of the trial period. Although we believe that these temporary rules had some effect on scoring, we do not account for the presence and absence of the temporary rules in our modelling. In the 472 matches, 257922 balls were bowled involving $I = 435$ batsman and $J = 360$ bowlers. In the first innings,

Table 2.1: The number of ODI matches for which data were collected on the ICC teams.

ICC Nation	Number of Matches
Australia	106
Bangladesh	48
England	89
India	125
New Zealand	94
Pakistan	108
South Africa	101
Sri Lanka	118
West Indies	77
Zimbabwe	78

138439 balls were bowled. In Table 2.1, we record the number of matches for which data were collected on the ICC teams. Excepting Bangladesh, we observe that there is reasonable balance in the number of matches played.

Over these matches, we calculate $\hat{v} = 8289/257922 = 0.032$ as the total number of wide-balls and no-balls divided by the total number of balls bowled. The conditional probabilities ϕ_k are similarly estimated by frequencies.

Before describing the model used in the estimation of p_{ijwbk}, we offer some preliminary comments based on our experience in fitting numerous models over several years. Most importantly, our goal is to obtain a model which fits well and provides a realistic simulator. With regard to estimation and prediction, there are many situations for which data do not exist. For example, a given batsman i may never have faced a given bowler j in the third over when two wickets are lost. To predict what may happen in these situations it is necessary to borrow information from similar situations. For example, it is relevant in the above problem to consider how other batsman/bowler combinations fared in the third over when two wickets are lost, and it is also relevant to consider how batsman i fared against bowler j in other stages of a match. With nearly 1000 batsmen and bowlers, our

experience suggests that it is not a good idea to have multiple parameters for each batsman and bowler since this leads to excessive parameterization. When the number of parameters is excessive, computation may be overwhelming and unreliable estimates may arise. We strive for parsimonious model building, assigning only a single parameter to each batsman and assigning only a single parameter to each bowler.

To obtain the estimates $\hat{p}_{ijwbk}$ in (2.3), we develop a Bayesian latent variable model which is related to the classical cumulative-logit models for ordinal responses (Agresti 2002). We initially assume that batting is taking place in the first innings and we temporarily ignore the state of the first innings (e.g. the number of overs completed and the number of wickets lost). We imagine that there is a latent continuous variable U which describes the "quality" of the batting outcome. For example, one might imagine that a ball batted near a fielder has a lower rating than a similar ball batted slightly further away from the fielder. Although U is unobserved (latent), there is an observed data variable X which is related to U as follows:

$$\begin{aligned}
X = 1 \text{ (dismissal)} &\leftrightarrow a_0 < U \leq a_1 \\
X = 2 \text{ (0 runs)} &\leftrightarrow a_1 < U \leq a_2 \\
X = 3 \text{ (1 runs)} &\leftrightarrow a_2 < U \leq a_3 \\
X = 4 \text{ (2 runs)} &\leftrightarrow a_3 < U \leq a_4 \\
X = 5 \text{ (3 runs)} &\leftrightarrow a_4 < U \leq a_5 \\
X = 6 \text{ (4 runs)} &\leftrightarrow a_5 < U \leq a_6 \\
X = 7 \text{ (6 runs)} &\leftrightarrow a_6 < U \leq a_7
\end{aligned}$$

where $-\infty = a_0 < a_1 < \cdots < a_6 < a_7 = \infty$. Note that increasing values of X correspond to higher quality of batting outcomes and that $a_1, \ldots, a_6$ are unknown parameters.

We now define a single batsman characteristic $\mu_i^{(1)}$ for the i-th batsman and a single bowler characteristic $\mu_j^{(2)}$ for the j-th bowler. We assume that the quality of the batting outcome U is expressed as

$$U = \mu_i^{(1)} - \mu_j^{(2)} + \epsilon \tag{2.4}$$

where $\mu^{(1)} = 0$ represents an average batsmen, $\mu^{(2)} = 0$ represents an average bowler and ϵ is a random variable whose distribution determines the quality of the batting outcome for an average batsman and an average bowler. From (2.4), we observe that a good batsman

($\mu_i^{(1)} > 0$) increases the quality of the batting outcome. Similarly, a good bowler ($\mu_j^{(2)} > 0$) decreases the quality of the batting outcome. Letting F denote the distribution function of ϵ, we write

$$\begin{aligned} \text{Prob}(X \leq k) &= \text{Prob}(U \leq a_k) \\ &= \text{Prob}(\mu_i^{(1)} - \mu_j^{(2)} + \epsilon \leq a_k) \\ &= \text{Prob}(\epsilon \leq a_k - \mu_i^{(1)} + \mu_j^{(2)}) \\ &= F(a_k - \mu_i^{(1)} + \mu_j^{(2)}) \end{aligned} \tag{2.5}$$

for the batting outcome $k = 1, \ldots, 7$.

To gain a better appreciation of the model given by (2.5), refer to Figure 2.1 where the density function of the logistic distribution (i.e. $F(\epsilon) = 1/(1 + \exp(-\epsilon))$) has been chosen using typical values of $a_1, \ldots, a_6$. The logistic distribution is symmetric about 0 and has longer tails than the normal distribution. We observe that the area under the density function between a_{k-1} and a_k corresponds to the probability that $X = k$. The effect of better batsmen ($\mu_i^{(1)} > 0$) and weaker bowlers ($\mu_j^{(2)} < 0$) causes a simultaneous leftward shift in the vertical lines and hence decreases the probability of dismissal and increases the probability of scoring 6 runs.

A difficulty with the model given by (2.5) is that it does not account for the variability in aggressiveness which batsmen display during the first innings. For example, it is well-known that batsmen become more aggressive in the final overs if few wickets have been lost. Increasing aggressiveness corresponds to an increase in the probability of dismissal ($k = 1$), a decrease in the probability of scoring 0 runs ($k = 2$) and an increase in the probability of scoring 6 runs ($k = 7$). The effect of increasing aggressiveness on the remaining four outcomes $k = 3, 4, 5, 6$ is not as obvious and is situation dependent. Note that aggressiveness is a characteristic that affects the parameters a_k and is different from the quality of batsmen $\mu_i^{(1)}$ and the quality of bowlers $\mu_j^{(2)}$. In the spirit of parsimonious model building, in Table 2.2 we propose 9 situations where aggressiveness is assumed constant. Situations 1, 2 and 3 are motivated by the fielding restriction which is in place during the first 15 overs. We view overs 16-35 as the middle period of the first innings where any change in aggressiveness is due to wickets lost. The final period of the first innings (overs 36-50) can lead to very aggressive batting if few wickets are lost. We note that situations 4, 7 and 8 are particularly rare and

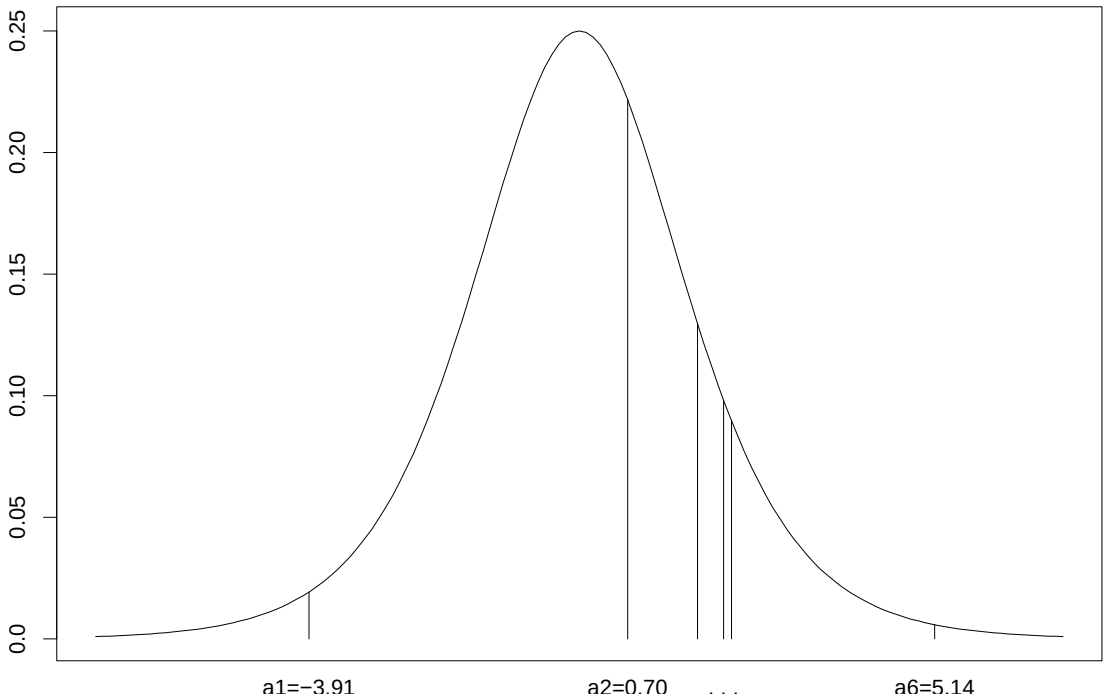

Figure 2.1: Logistic density function and typical parameters for the model in (5).

the parameters corresponding to these situations may not be well estimated. However, this is not a great concern as simulation rarely takes place during these periods.

Having proposed the 9 situations in Table 2.2, we modify the model given by (2.5) to take into account the varying levels of aggressiveness. We introduce a new subscript $l = 1, \ldots, 9$ to denote the 9 situations and we note that l is really a function of w (wickets lost) and b (ball currently facing) as indicated in (2.3). Consider then a batting outcome where batsman i faces bowler j in the l-th situation and outcome k is recorded. The likelihood contribution due to the event is

$$F(a_{lk} - \mu_i^{(1)} - \Delta_l + \mu_j^{(2)}) - F(a_{l,k-1} - \mu_i^{(1)} - \Delta_l + \mu_j^{(2)}) \tag{2.6}$$

where $\Delta_l = 0$ for $l = 1, 2, 3$, $\Delta_l = \Delta^{(1)}$ for $l = 4, 5, 6$ and $\Delta_l = \Delta^{(1)} + \Delta^{(2)}$ for $l = 7, 8, 9$. The complete likelihood is therefore the product of terms of the form (2.6) over all balls bowled in the dataset. The additional parameters $\Delta^{(1)}$ and $\Delta^{(2)}$ provide a link to situations where batsmen do not typically bat. For example, batsmen who bat in positions 7, 8 and 9 in the batting order are usually bowlers and are usually not very good batsmen. If the model given by (2.6) were fit without the Δ terms, then the $\mu_i^{(1)}$ values for these batsmen

Table 2.2: The nine first innings situations where aggressiveness is assumed constant. The fourth column provides the percentage of balls in the dataset that correspond to the given situation.

Situation	Over	Wickets Lost	Percentage of Dataset
1	1-15	0-3	30.3%
2	16-35	0-3	25.3%
3	36-50	0-3	5.0%
4	1-15	4-6	1.2%
5	16-35	4-6	14.0%
6	36-50	4-6	15.2%
7	1-15	7-9	0.0%
8	16-35	7-9	1.7%
9	36-50	7-9	7.3%

would be relative only to batsmen who batted in situations 7, 8 and 9. We therefore adjust situational skill levels using the Δs. Recall that our intention is to develop a simulator that allows experimentation whereby batsmen may bat in atypical positions in the batting order. The estimation of $\Delta^{(1)}$ and $\Delta^{(2)}$ is feasible due to data corresponding to batsmen who crossover. For example, $\Delta^{(1)}$ is estimable since there are some batsmen who have data in at least one of situations 1, 2 or 3 and who have data in at least one of situations 4, 5 or 6. We remark that our approach to handling situational skill levels and aggressiveness in the first innings might be better handled in some sort of continuous fashion rather than imposing 9 states where homogeneity is assumed.

From (2.6) the primary parameters of interest are $\Delta^{(1)}$, $\Delta^{(2)}$, the a_{lk}s, the $\mu_i^{(1)}$s and the $\mu_j^{(2)}$s. This corresponds to $1 + 1 + 9(6) + 435 + 360 = 851$ unknown primary parameters. In a Bayesian formulation, parameters have prior distributions and we assign the following

prior distributions

$$\begin{aligned}
\Delta^{(1)} &\sim \text{uniform}(0,1) \\
\Delta^{(2)} &\sim \text{uniform}(0,1) \\
a_{lk} &\sim \text{normal}(0,\sigma^2) \quad \text{where} \quad \sigma^{-2} \sim \text{gamma}(1.0,1.0) \\
\mu_i^{(1)} &\sim \text{normal}(0,\tau^2) \\
\mu_j^{(2)} &\sim \text{normal}(0,\tau^2) \quad \text{where} \quad \tau^{-2} \sim \text{gamma}(1.0,1.0)
\end{aligned}$$

for batting outcomes $k = 1, \ldots, 6$, for situations $l = 1, \ldots, 9$, for batsmen $i = 1, \ldots, I = 435$ and for bowlers $j = 1, \ldots, J = 360$. The prior distributions are assumed independent except for the order restriction $a_{l1} < a_{l2} < \cdots < a_{l6}$ for $l = 1, \ldots, 9$. The notation $Y \sim \text{gamma}(a, b)$ implies $\text{E}(Y) = a/b$ and $\text{Var}(Y) = a/b^2$. Note that although the prior distributions are somewhat diffuse, prior knowledge is used in the prior specification. For example, it is known that batsmen who bat in positions 1, 2 and 3 in the batting order are generally better than batsmen who bat in positions 4, 5 and 6 in the batting order who are in turn generally better than batsmen who bat in positions 7, 8 and 9 in the batting order. This knowledge implies that $\Delta^{(1)} > 0$ and $\Delta^{(2)} > 0$. The prior distributions for σ^{-2} and τ^{-2} make use of the knowledge that the bulk of the probability for the logistic distribution F lies in the interval $(-5, 5)$. Also, the modelling of common prior means for the $\mu_i^{(1)}$ and the modelling of common prior means for the $\mu_j^{(2)}$ is sensible in that it produces average characteristics for batsmen and bowlers for whom little data has been collected. Finally, note that our prior specification introduces only two extra hyperparameters (σ and τ).

Our model can be specified within the WinBUGS platform, and upon providing the data, WinBUGS constructs a Markov chain whose output consists of generated parameters. Since the Markov chain converges to the posterior distribution, the generated parameters can be averaged to obtain approximations of the posterior means. When the parameters are estimated, the probabilities $\hat{p}_{ijwbk}$ for the simulator (2.3) are obtained by substituting the relevant primary parameter estimates into (2.6). One of the features of the model is that it permits inference on various batsman/bowler combinations at different stages of a match even when they have not faced one another in actual competition.

In our WinBUGS implementation, we used a burn-in of 10000 iterations and a further 10000 iterations for parameter estimation. This required approximately one day of computation. Standard diagnostics such as trace plots, autocorrelation plots and varied starting values were used to assess convergence. We experimented with changes to the vague prior distributions (particularly the gammas) and found that our results were not sensitive to the prior specification.

There is an important remaining point that needs to be made concerning the fitting of the Bayesian latent variable model. In order to fit the model, we require ball by ball data to specify the terms in (2.6). To obtain the data, the ball by ball commentary log from the Cricinfo website was parsed into a convenient format. For example, codes were created to index batsmen and bowlers, and outcomes were categorized according to (2.1).

2.4 Generating runs in the second innings

Up until now, we have considered only the generation of first innings runs. It is evident that the conditional distributions $[X_b \mid X_0, \ldots, X_{b-1}]$ for the second innings also depend on the current score of the match. For example, it is well known that when the team batting first scores an unusually high number of runs, the team batting second becomes more aggressive in its batting style.

One idea is to modify the model given by (2.6) so that batting outcome probabilities also depend on the score of the match. One could imagine introducing additional subscripts to the a_{lk} terms to denote the score of the match. The problem with such an approach is that there would be many more parameters and very few replicate observations for the purposes of estimation.

Our approach is to leave the model given by (2.6) as it stands, and view the p_{ijwbk} terms in (2.3) as the first innings probabilities which share some characteristics with the second innings probabilities. We account for the current score in the second innings by modifying the conditional distributions $[X_b \mid X_0, \ldots, X_{b-1}]$ used in the algorithm for simulation. Consider then the stage of the second innings where batsman i faces bowler j, w wickets have been lost and the b-th ball is about to be bowled. For notational convenience, we suppress

the subscripted notation $ijwb$. Then, referring to (2.1) and ignoring wide-balls and no-balls, the expected number of runs that the batsman scores on the current ball is

$$\mathrm{E}_1(p) = p_3 + 2p_4 + 3p_5 + 4p_6 + 6p_7$$

and the expected proportion of resources that the batsman consumes on the current ball is

$$\begin{aligned} \mathrm{E}_2(p) &= (x+y)p_1 + xp_2 + xp_3 + xp_4 + xp_5 + xp_6 + xp_7 \\ &= x + yp_1 \end{aligned}$$

where the proportion of resources lost x due to the current ball and the proportion of resources lost y due to a wicket are known quantities that are available from the Duckworth/Lewis resource table (Duckworth and Lewis 1998, 2004). For the batsman to become more aggressive during the match, this implies a change in the probabilities $p = (p_1, \ldots, p_7)$. We make the assumption that the batsman modifies his overall batting behaviour from p to $p' = (p'_1, \ldots, p'_7)$ according to the current score in the match.

Consider now the situation where f runs have been scored in the first innings, s runs have been scored in the second innings and the proportion of resources remaining in the second innings is r. To win, the team batting second needs to score $f - s + 1$ runs in the remainder of the match relative to the r resources that are available. This suggests that the run to resource ratio for the batsman should be at least $(f-s+1)/r$. If $(f-s+1)/r > \mathrm{E}_1(p)/\mathrm{E}_2(p)$, then the team batting second is on the verge of losing/tying, and we assume that the batsman becomes more aggressive in his batting style. We therefore propose that the batsman modifies his style from p to p' where

$$p'_2 = cp_2, \quad c \in (0, 1). \tag{2.7}$$

The idea is that a more aggressive batsman swings the bat more often and is less likely to score 0 runs (i.e. $p'_2 < p_2$). When $c = 1$, the batsman is behaving in a neutral fashion (i.e. not extra aggressive), and when $c = 0$, the batsman has reached his limit of aggressive behaviour where scoring 0 runs is impossible. Accordingly, when $(f - s + 1)/r > \mathrm{E}_1(p)/\mathrm{E}_2(p)$, we set

$$c = \frac{r\mathrm{E}_1(p)}{(f - s + 1)\mathrm{E}_2(p)}. \tag{2.8}$$

We propose that when $c \in (0,1)$ as in (2.8), the decrease in probability from p_2 to p_2' results in an increase in the probability of dismissal

$$p_1' = p_1 + \delta p_2(1-c), \quad \delta \in [0,1] \tag{2.9}$$

and a proportional increase in the run-scoring probabilities

$$p_i' = \left(\frac{1-p_1-(c+\delta(1-c))p_2}{1-p_1-p_2}\right) p_i, \quad i = 3,\ldots,7. \tag{2.10}$$

It is easy to establish that $p_1', \ldots, p_7'$ form a simplex, and hence constitute a probability distribution. Observe that the free parameter $\delta \in [0,1]$ determines the degree to which the aggressive behaviour affects dismissals (2.9) and run-scoring (2.10). When $\delta = 1$, all of the aggressive behaviour increases the dismissal probability, and when $\delta = 0$, all of the aggressive behaviour increases the run-scoring probabilities.

When a batsman is aggressive in the second innings (i.e. $c \in (0,1)$), it is easy to establish that

$$\mathrm{E}_1(p') = \left(\frac{1-p_1-(c+\delta(1-c))p_2}{1-p_1-p_2}\right) E_1(p) \geq \mathrm{E}_1(p) \tag{2.11}$$

and

$$\mathrm{E}_2(p') = E_2(p) + \delta p_2(1-c)y \geq \mathrm{E}_2(p). \tag{2.12}$$

In modifying his behaviour from p to p', the inequalities (2.11) and (2.12) imply that the batsman simultaneously increases the expected number of runs scored and the expected number of resources consumed on the current ball. Moreover, it is straightforward to show that both $\mathrm{E}_1(p')$ and $\mathrm{E}_2(p')$ are decreasing functions of $c \in (0,1)$. These consequences correspond to our intuition of more aggressive batting.

The remaining detail in modelling aggressive batting in the second innings is the determination of the parameter $\delta \in [0,1]$. Although an aggressive batsman is attempting to increase his run production $\mathrm{E}_1(p')$, the quantity which really determines the quality of batting is $\mathrm{E}_1(p')/\mathrm{E}_2(p')$. We argue that a batsman is unable to modify his batting style from p to p' so as to make $\mathrm{E}_1(p')/\mathrm{E}_2(p') > \mathrm{E}_1(p)/\mathrm{E}_2(p)$; for if he were able to do this, he would do it all the time, in both the first and second innings. However, when a team is on the verge

of losing (i.e. $(f-s+1)/r > \mathrm{E}_1(p)/\mathrm{E}_2(p)$), we suggest that a batsman may be willing to sacrifice $\mathrm{E}_1(p')/\mathrm{E}_2(p')$ with the benefit of increased run production $\mathrm{E}_1(p')$. In other words, we require $\mathrm{E}_1(p')/\mathrm{E}_2(p') \leq \mathrm{E}_1(p)/\mathrm{E}_2(p)$ and that $\mathrm{E}_1(p')/\mathrm{E}_2(p')$ be an increasing function of $c \in (0,1)$. Using the expressions in (2.11) and (2.12), it is possible to show that $\mathrm{E}_1(p')/\mathrm{E}_2(p')$ is an increasing function of $c \in (0,1)$ provided $\delta > \mathrm{E}_2(p)/(\mathrm{E}_2(p) + y(1-p_1-p_2)) \in (0,1)$. Since a batsman would naturally desire $\mathrm{E}_1(p')/\mathrm{E}_2(p')$ to be as large as possible, we therefore set

$$\delta = \frac{\mathrm{E}_2(p)}{\mathrm{E}_2(p) + y(1-p_1-p_2)}. \tag{2.13}$$

If $(f-s+1)/r < \mathrm{E}_1(p)/\mathrm{E}_2(p)$, the team batting second is on the verge of winning, and although it may not be optimal, batsmen become more cautious. The tendency to become more cautious when protecting a lead is widely acknowledged in many sports including American football, basketball and ice hockey. Following the above development for aggressive batting, a similar modification in a batsman's style from p to p' can be obtained. The idea is that a more cautious batsman provides greater protection of the wicket and is more likely to score 0 runs. Specifically, we set c and δ according to (2.8) and (2.13) respectively, and we determine the probabilities $p'_1, \ldots, p'_7$ according to (2.7), (2.9) and (2.10). In this case, increasing c corresponds to increasing cautiousness. We note that $p'_2 > p_2$ and $p'_i < p_i$ for the remaining probabilities $i = 1, 3, 4, 5, 6, 7$. We keep in mind that it may be necessary to reduce c to an upper bound to ensure that the probability vector p' forms a simplex. We note that the only way that the limit of cautious behaviour is reached (i.e. $c = 1/p_2$) is when $(f-s+1)/r = 0$ which means that the batting side has already won the match and batting has terminated.

There is a final modification in our approach to second innings batting which we now consider. In many sporting activities it is an advantage to have the final offensive opportunity in a game. For example, this is the case in baseball where it is widely viewed as advantageous to bat in the bottom innings since strategy varies according to the number of runs required. Similarly, it is generally beneficial in golf to play in the last foursome of a tournament since the score to beat is known. It therefore seems reasonable that a second innings batting advantage should also exist in one-day cricket. An advantage in second innings batting seems to go hand in hand with the quotation from Sir Francis Bacon that

"knowledge is power". However, upon looking at empirical data, de Silva and Swartz (1997) found no such advantage in second innings batting for ODI cricket. A possible explanation is that the strategic advantage in second innings batting is offset by the deterioration of the pitch during the second innings. As a match progresses, the pitch often becomes worn down, and batting becomes more difficult as bowling is subject to more erratic bounces. Although our procedure for generating runs in the second innings takes strategy into account through modified aggression levels in batting, our procedure fails to account for the deterioration in the pitch. We therefore introduce the "pitch variable" η in (2.6) which is activated in second innings batting but not in first innings batting. We define η as an offset to the parameters a_{l1}, modifying a_{l1} to $a_{l1} + \eta$. This has the effect of increasing the probability of dismissal in second innings batting. Since our estimation procedure is based on first innings data only, we do not treat the pitch variable η as a standard parameter which is estimated but rather we treat η as a tuning parameter. We have set $\eta = 0.15$ (a very small adjustment relative to the size of a_{l1}) to coincide with the findings of de Silva and Swartz (1997). The treatment of η as a tuning parameter may be more appropriate than viewing η as a pitch variable. As pointed out by a referee, it is not always the case that batting conditions deteriorate during the course of a match.

2.5 Testing model adequacy

As a type of cross-validation procedure, we fit the Bayesian latent variable model using only first innings data. Although this reduces the size of the dataset (by roughly 50%), it permits us to compare simulated results for the second innings with actual second innings results that were not used in determining the parameter estimates.

The model was fit using WinBUGS software where posterior means were calculated for the 853 model parameters. Although the WinBUGS program requires two hours of computation, once the parameter estimates are obtained, they can be used over and over again as inputs to the simulation program. In the Appendix A, we provide the WinBUGS code to emphasize the simplicity in which WinBUGS software facilitates the implementation of latent variable models.

Another advantage of the Bayesian formulation concerns the use of parameter estimates in the simulation program. It is a widely held belief that the performances of batsmen and bowlers are not constant. For example, batsmen have good days and bad days, and this can be related to their health or any number of reasons. In a Bayesian formulation, we need not use the same parameter estimates $\mu^{(1)}$ and $\mu^{(2)}$ (posterior means) for batsmen and bowlers over all matches. Alternatively, at the beginning of a match, the $\mu^{(1)}$ and $\mu^{(2)}$ values can be generated from their respective posterior distributions to reflect match by match variation in performance.

Now, there are countless ways that one might test the adequacy of the model. In Table 2.3, we provide the estimated probabilities p_{ijwbk} for some batsmen/bowler combinations at different states of a match. We have also included the expected number of runs per over for each combination. We have presented batting outcome probabilities when Alistair Cook of England is batting against Glenn McGrath of Australia, and against Nazmul Hossain of Bangladesh. At the beginning of a match (i.e. ball 1, 0 wickets), we observe that with probabilities 0.681 and 0.078, Cook scores 0 runs and 4 runs respectively against McGrath. At the beginning of a match, these probabilities change to 0.626 and 0.100 respectively when Hossain is bowling. These changes are consistent with the general belief that McGrath is a better bowler than Hossain. We then investigate a situation where batsmen ought to become more aggressive (ball 271 when 2 wickets are lost). Indeed, the probability that Cook scores 0 runs decreases substantially to 0.338 and 0.285 depending on whether McGrath or Hossain is the bowler. We also note a curious result concerning the probability of scoring 4 runs. Even though batsmen are more aggressive on ball 271 with 2 wickets than at the beginning of a match (ball 1 with 0 wickets), the fielding restriction that is in place at the beginning of a match enables batsmen to score 4's at a higher rate. This batting behaviour is observed in Table 2.3 and has been verified by looking at empirical data. In Table 2.3, we also investigate the case of ball 271 when 4 wickets are lost which according to common knowledge should be a less aggressive batting situation than ball 271 when 2 wickets are lost. Accordingly, we observe that the probability of 0s increase and the probability of 1s decrease in the less aggressive situation.

In Table 2.4, we investigate the adjustment from p to p' in the second innings. Consider

Table 2.3: Batting probabilities p for various states and the expected number of runs per over $E(R)$ where CM denotes the Cook/McGrath matchup and CH denotes the Cook/Hossain matchup.

State of the Match	Dismissal	Zero	One	Two	Three	Four	Six	$E(R)$
CM (ball 1, 0 wickets)	0.024	0.681	0.165	0.038	0.010	0.078	0.004	3.7
CM (ball 271, 2 wickets)	0.039	0.338	0.452	0.077	0.006	0.069	0.018	6.1
CM (ball 271, 4 wickets)	0.033	0.352	0.435	0.085	0.007	0.071	0.017	6.1
CH (ball 1, 0 wickets)	0.018	0.626	0.191	0.047	0.012	0.100	0.006	4.5
CH (ball 271, 2 wickets)	0.030	0.285	0.472	0.094	0.008	0.088	0.024	7.1
CH (balls 271, 4 wickets)	0.025	0.297	0.454	0.103	0.008	0.091	0.022	7.1

again the matchup between the batsman Cook and the bowler Hossain. Suppose that Bangladesh has scored $f = 250$ runs in the first innings. Suppose further that ball $b = 183$ is about to be bowled and $w = 3$ wickets have been lost in the second innings (situation 2). From the Duckworth/Lewis table, we therefore have the proportion of resources lost $x = 0.0027$ due to the current ball and the proportion of resources lost $y = 0.044$ due to a wicket. This might be considered as a "middle point" of the second innings since the proportion of resources used is $R(w, b) = 0.4993$, and therefore, the proportion of resources remaining is $r = 0.5007$. In this case, the estimated parameters give $\mathrm{E}_1(p) = 0.9723$ and $\mathrm{E}_2(p) = 0.00393$. We now investigate the outcome probabilities p'_{ijwbk} when England has scored $s = 127, 90, 60$ runs. When $s = 127$, then $(f-s+1)/r = 247.7 \approx \mathrm{E}_1(p)/\mathrm{E}_2(p) = 247.5$ and England is on pace to draw the match, and $p' = p$ (i.e. no adjustment). When $s = 90$, Cook should become more aggressive ($c = 0.768$), and when $s = 60$, Cook should become even more aggressive ($c = 0.648$). The entries in Table 2.4 appear reasonable and support these tendencies.

We now compare actual runs versus simulated runs. For this, we consider the 23 matches between Sri Lanka and India from November 1998 through March 2007 in which Sri Lanka batted first. The 23 matches consist of 15 matches from the original dataset (2001-2006)

Table 2.4: Second innings batting probabilities p' and the expected number of runs per over for the Cook/Hossain matchup when Bangladesh has scored $f = 250$ runs in the first innings. In the second innings, $w = 3$ wickets have been lost, ball $b = 183$ is about to be bowled and England has scored s runs.

England Runs (s)	Dismissal	Zero	One	Two	Three	Four	Six	Runs/Over
127	0.016	0.452	0.397	0.058	0.008	0.062	0.008	5.0
90	0.029	0.347	0.466	0.068	0.010	0.072	0.009	5.8
60	0.036	0.293	0.501	0.073	0.010	0.078	0.010	6.3

used for model fitting and 8 matches outside of the training period. We simulate 1000 first innings results for Sri Lanka based on representative batting and bowling orders employed during the time period. The resultant QQ plot comparing the actual runs and the simulated runs is given in Figure 2.2. We observe excellent agreement and we remark that satisfactory plots are also observed for other pairs of teams that we investigated. In comparing wickets taken, the actual results also compare favourably with the simulated results.

We also investigate the effect of the second innings adjustment. The difficulty in this exercise is that given the number of first innings runs between two teams, replicate observations tend not to occur. Therefore, to address goodness-of-fit, we provide some evidence that the second innings adjustment p' is an improvement over having the second innings team bat in a neutral fashion (i.e. $p' = p$). Consider then simulated matches between Australia and the other 9 ICC teams where Australia is batting in the second innings and where the batting and bowling lineups resemble those used in the 2007 World Cup matches. We generate first innings runs for the other teams, and then second innings runs for Australia with the proposed batting adjustment p' based on the target scores. The simulation is repeated for 1000 hypothetical matches for each of the 9 teams. We observe that Australia uses their full 50 overs in 8.5% of the simulated matches. The small percentage seems sensible since Australia rarely uses all 50 overs in matches that they win. In matches that Australia

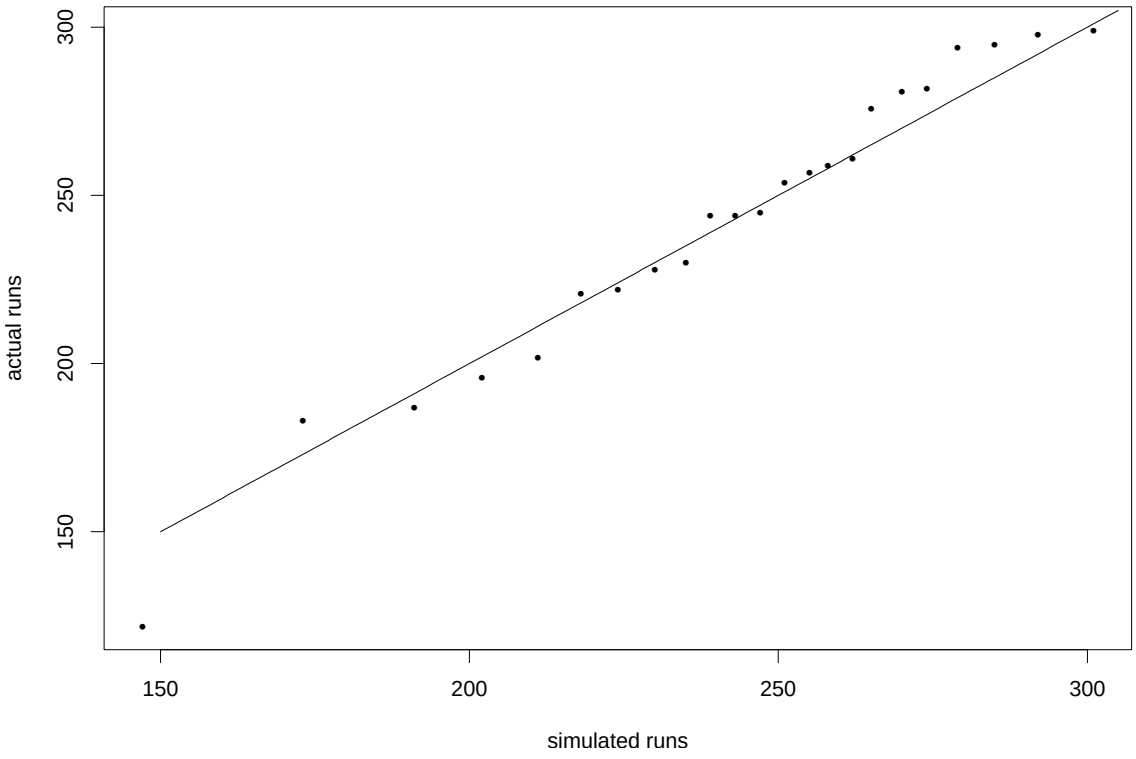

Figure 2.2: QQ plot corresponding to first innings runs for Sri Lanka batting against India.

loses, at some point when they are falling behind, they become desperate (aggressive), and typically consume all of their wickets before using the allotted 50 overs. When we repeat the simulation with neutral batting in the second innings (i.e. Australia behave as they would in the first innings), Australia uses all 50 overs 13% of the time. To get a sense of the percentages using actual data, we look at all 83 matches from 2000 to 2006 where Australia batted second, and observe that Australia used the full 50 overs only 7% of the time. This suggests that there is merit in our modification of aggressiveness in second innings batting.

2.6 Addressing questions via simulation

Having developed a simulator for ODI cricket matches, there is no limit to the number and type of questions that may be posed. The greatest utility of the simulator occurs for circumstances in which there is limited empirical data. In these cases, without a simulator, the best that one can do is to rely on hunches with respect to the questions of interest. In this section, we give a flavour for the types of questions that might be posed. We see these

types of applications as being of value not only to cricket devotees but also to selection committees and team strategists. We note that each of the simulations described below requires less than one minute of computation.

2.6.1 Question 1.

Adam Gilchrist is often an opening batsman for Australia, and Australia has not played the West Indies often in recent history. We are interested in the probability of Gilchrist hitting a century as an opening batsman against the West Indies when Australia is batting in the first innings and the West Indies are using a bowlng lineup from the 2007 World Cup. Based on 1000 first innings simulations, Gilchrist reaches a century 5.1% of the time. The result appears consistent with Gilchrist's actual batting performances where Gilchrist made a century 8 times as an opening batsman in 138 first innings ODI matches (5.8%) throughout his career (1996-2008).

2.6.2 Question 2.

England has occasionally sent Alastair Cook and Matt Prior as opening batsmen. In other matches, they used Ian Bell and Michael Vaughan as opening batsmen. We are interested in the performance in the untested opening partnership of Alastair Cook and Ian Bell. More specifically, we consider the length of the partnership (i.e. the number of overs prior to losing the first wicket) for Cook and Bell when they are batting in the first innings against Pakistan where Pakistan uses a bowling lineup comparable to the lineup used in their December 15/2005 match against England. In Figure 2.3, we provide a histogram of the number of overs in the length of their partnership based on 1000 simulations. We observe that the median and the mean length of the partnership is 5 overs and 7.1 overs respectively. It appears very unlikely for Cook and Bell to have a partnership exceeding 20 overs.

2.6.3 Question 3.

Consider a match between New Zealand and Sri Lanka where Sri Lanka is batting in the second innings and New Zealand has scored an impressive 300 runs in the first innings.

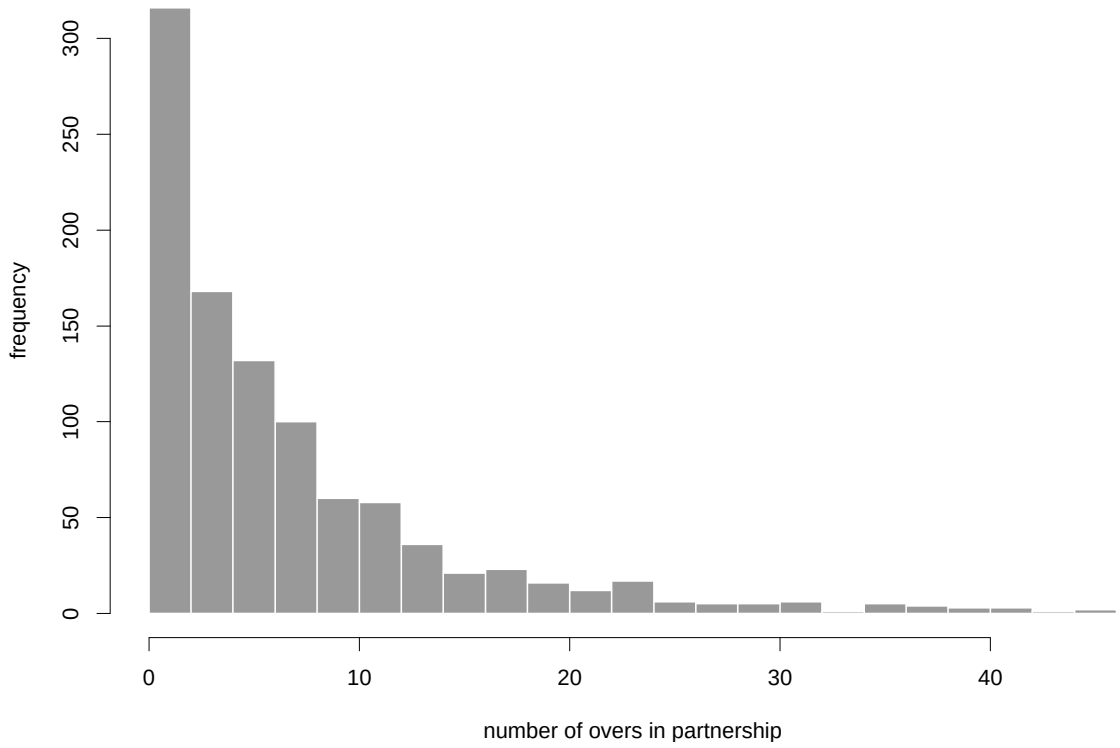

Figure 2.3: Histogram of the length of the partnership (in overs) of the untested opening partnership of Alastair Cook and Ian Bell.

We are interested in the probability that Sri Lanka can overcome the barrier and win the match. Based on 1000 simulations, and taking batting/bowling lineups used in the 2007 World Cup matches, we observe that Sri Lanka wins 10.0% of the simulated matches. The result is consistent with Sri Lanka's first innings performance in the current decade where Sri Lanka has scored over 300 runs in 9 out of 121 matches (i.e. 7.4% of the time).

2.6.4 Question 4.

Muttiah Muralitharan of Sri Lanka is a spin bowler and is widely regarded as one of the best bowlers in cricket. The question arises as to his value to the Sri Lankan team. Consider a match between Sri Lanka and India where India is batting in the first innings and India's batting lineup is based on their 2007 World Cup team. We use a bowling lineup comparable to Sri Lanka's bowling lineup in the 2007 World Cup where Muralitharan was a prominent bowler. Based on 1000 simulations, we observe that India scores 247 runs on average. When we make Muralitharan the only Sri Lankan bowler (which is of course against the rules),

then India scores only 185 runs on average. Clearly, if every bowler on Sri Lanka were as good as Muralitharan, Sri Lanka would have a much better team. When we instead replace Muralitharan with Upul Chandana (a more typical bowler), then India scores 267 runs on average.

2.6.5 Question 5.

Here is a crazy question that surely few people have contemplated. What would happen if we reverse a team's batting order? We consider a batting order used by Australia in the 2007 World Cup matches. Based on 1000 simulations against each of the other 9 teams using their 2007 World Cup bowling lineups, Australia produces on average 272 runs, losing 6.3 wickets in 48.8 overs during the first innings. This compares favourably with empirical data over the data collection period where Australia produces on average 273 runs, losing 6.8 wickets in 49.8 overs during the first innings. When we reverse the Australian batting lineup, and simulate 1000 matches against each of the other 9 teams, Australia produces on average 234 runs, losing 7.3 wickets in 48.4 overs during the first innings. A simple explanation for the difference in expected runs is that higher-scoring batsmen tend to be placed at the beginning of the lineup. When the batting order is reversed, they often do not get an opportunity to bat or they bat for shorter periods of time.

2.6.6 Question 6.

We now determine the probability that one team defeats another team. Using typical batting and bowling lineups taken from matches during the 2007 World Cup of Cricket, Table 2.5 provides estimated probabilities based on 10000 simulations between each pair of ODI teams. Accordingly, Australia is clearly the best team, and Bangladesh and Zimbabwe are the weakest teams. The probabilities for the other teams roughly agree with the authors' beliefs although we note that the probabilities are sensitive to the choice of the batting and bowling lineups. Referring to the row and column averages, we observe that the probability of winning is nearly the same for first and second innings batting. This corresponds to the observations of de Silva and Swartz (1997) and provides a justification for the tuning parameter η introduced at the end of Section 2.4.

Table 2.5: Estimated probabilities of the row team defeating the column team where the row team corresponds to the team batting first. The final column are the row averages and correspond to the average probabilities of winning when batting in the first innings. The final row are the average probabilities of winning when batting in the second innings.

Batting First	Batting Second										
	Aus	Bang	Eng	Ind	NZ	Pak	SA	SL	WI	Zimb	Average
Aus		0.88	0.69	0.65	0.72	0.81	0.58	0.64	0.74	0.96	0.74
Bang	0.14		0.29	0.23	0.34	0.42	0.19	0.25	0.35	0.77	0.33
Eng	0.29	0.75		0.43	0.54	0.61	0.36	0.43	0.54	0.89	0.54
Ind	0.33	0.75	0.54		0.57	0.65	0.40	0.46	0.57	0.89	0.57
NZ	0.30	0.81	0.50	0.44		0.64	0.40	0.46	0.59	0.92	0.56
Pak	0.16	0.58	0.31	0.27	0.37		0.22	0.28	0.37	0.78	0.37
SA	0.44	0.89	0.63	0.57	0.71	0.77		0.60	0.72	0.96	0.70
SL	0.37	0.81	0.58	0.52	0.61	0.70	0.46		0.63	0.92	0.62
WI	0.23	0.67	0.41	0.34	0.47	0.55	0.30	0.36		0.85	0.46
Zimb	0.05	0.33	0.12	0.10	0.15	0.21	0.08	0.12	0.16		0.15
Average	0.74	0.28	0.55	0.61	0.50	0.40	0.67	0.60	0.48	0.12	

2.7 Discussion

In this paper, a simulator for ODI cricket is developed. One of the virtues of the approach is that the characteristics of individual batsmen and bowlers are used to generate ball by ball outcomes. As time progresses and more matches are played, the database may be updated to reflect changes in player characteristics.

Beyond obvious uses in betting, the simulator may be used to help teams determine optimal strategies. In ODI cricket, it is not so easy to test ideas as a team may only play 20 matches per year and a team does not typically play all ICC nations. For example, it is not clear how changes in the batting and bowling orders affect a match. The simulator allows a team to easily investigate the results of making changes to the batting and bowling orders.

Our simulator was developed with an attempt to realistically model one-day cricket, and it appears to do a reasonable job of reflecting major tendencies. Nevertheless, there are improvements that might be made in future implementations of the simulator. For example, rather than treat wide-balls and no-balls as aggregate characteristics, it may be possible to incorporate wide-balls and no-balls as individual characteristics of bowlers. It is also possible to include a home-field advantage term in the Bayesian latent variable model. Home field advantage has been estimated to be worth roughly 16 runs for the home team (de Silva, Pond and Swartz 2001).

There are other modelling issues that we have not considered yet may have an impact on scoring. For example, in many sports there tends to be an aging effect where younger players improve, reach a plateau and then experience a decline in performance. Modelling the aging effect is a challenge (Berry, Reese and Larkey 1999) and appears to be sport specific. We hope that by discarding old data and regularly updating our database, we might mitigate the aging effect by retaining data reflective of current performance. We might also consider the possibility that batsmen are more vulnerable to dismissal when they first come to the crease. Another modelling challenge involves the recognition that some batsmen struggle with certain types of bowlers (e.g. fast bowlers, spin bowlers).

Immediately prior to the World Cup of Cricket held in the West Indies in March 2007, a significant rule change was made with respect to fielding restrictions. This rule is known

as the Powerplay rule which differs slightly from the temporary Powerplay rule considered during the 10 month period beginning August 2005. The new Powerplay rule has the potential of affecting the periods of constant aggressiveness as assumed in Table 2.2. In a few years time, when more data have been collected subject to the new rule, we plan on altering the definition of the situations in Table 2.2 and refitting our model. Of course, our simulations are only as good as the data which have been collected. Therefore it is advisable to regularly update our database and eliminate data that is deemed to have occurred too far in the past.

Chapter 3

A Bayesian Approach to Network Models

3.1 Introduction

The analysis of network data is an active research topic. The range of applications is vast and includes such diverse areas as the detection of fraud in the telecommunications industry (Cortes, Pregibon and Volinsky 2003), the development of adaptive sampling schemes for populations at risk of HIV/AIDS infection (Thompson 2006), the study of conflicts between nations (Ward and Hoff 2007, Hoff 2008), the quantification of social structure in elephant herds (Vance 2008) and the investigation of the cooperative structure between lawyers (Lazega and Pattison 1999).

Not only are the areas of application varied, the statistical approaches to the analysis of network data are also varied. The approaches depend on many factors including the inferential goal of the analysis whether it be description, testing or prediction, the size of the data set and the nature of the data. Data may be continuous or discrete, there may be complex dependencies amongst nodes, relationships may be directed or non-directed, data may be dynamic, multivariate, have missing values, include covariates, lack balance, etc. Network analyses have been considered under both classical and Bayesian paradigms.

Although a complete review of the network literature strikes us as daunting task, we remark on some of the prominent approaches to the statistical analysis of network data.

With continuous observations between network nodes, Warner, Kenny and Stoto (1979) introduced the social relations model whose structure considers dependencies in the measurements between nodes. In the social relations model, nodes (e.g. subjects) have dual roles as both actors and partners where measurements between nodes are dependent on both actor and partner effects. Social relations models (also referred to as round robin models) were originally studied using analysis of variance methodology. Other inferential approaches have since been explored including maximum likelihood (Wong 1982), multilevel methods (Snijders and Kenny 1999) and Bayesian methods (Hoff 2005, Gill and Swartz 2007).

More research effort has taken place in the context of binary network data where a greater amount of mathematics and graph theory have come into play (Besag 1974, Frank and Strauss 1986). In the context of binary network data, a seminal contribution is due to Holland and Leinhardt (1981) who broke away from the often unrealistic assumption of independence between pairs of nodes and proposed the p_1-model for directed graphs. The original p_1-model has been expanded upon in many ways including empirical Bayes approaches (Wong 1987), fully Bayesian approaches (Gill and Swartz 2004) and the consideration of more complex dependencies (Wasserman and Pattison 1996). All of these models fall under the general framework of exponential random graph models whose various limitations have been discussed by Besag (2001) and Handcock (2003). A main feature of exponential random graph models is that the entire graph is modelled. A distinct approach to the analysis of binary network data involves modelling the individual nodal relationships; these models have been generalized in various ways and are referred to as latent factor models (Hoff, Raftery and Handcock 2002, Handcock, Raftery and Tantrum 2007). Finally, a recent approach which is related to the latent factor methodology provides a greater emphasis on the socio-spatial structure typically inherent in networks (Linkletter 2007). The approach requires the existence of meaningful spatial covariates and appears well suited for prediction.

This paper investigates the suitablity of Dirichlet process priors in the Bayesian analysis of network data. The Dirichlet process (Ferguson 1974) which was once a mathematical curiousity is becoming a popular applied tool (Dey, Müller and Sinha 1998). Dirichlet process priors allow the researcher to weaken prior assumptions by going from a parametric

framework to a semiparametric framework. This is important in the analysis of network data where complex nodal relationships rarely allow a researcher the confidence in assigning parametric priors. The Dirichlet process has a secondary benefit due to the fact that its support is restricted to discrete distributions. This results in a clustering effect which is often suitable for network data where groups of individuals in a network can be thought of as arising from the same cohort. Importantly, we demonstrate how Dirichlet process priors can be easily implemented in network models using WinBUGS software (Spiegelhalter, Thomas and Best 2003). The ease in which this can be done increases the potential of the methodology for widespread usage.

In section 3.2, we provide an overview of the Dirichlet process with an emphasis on issues that are most relevant to the implementation of the network models that are considered in this paper. In section 3.3, we begin with a simple network model where the observations between nodes are measured on a continuous scale. For this model, we demonstrate how the Dirichlet process can be easily implemented using WinBUGS software. This example involves a social relations model previously studied by Gill and Swartz (2007) where the observations between nodes are measured on a continuous scale. In section 3.4, we implement the Dirichlet process on a second model where the presence or absence of a tie between a pair of nodes implies a binary response. This example concerns an enhanced binary network model that studies the working relationships between lawyers. This is a variation of the p1-model of Holland and Leinhardt (1981) and stratifies the lawyers according to their professional rank. In section 3.5, a simulation study involving a simple but popular binary network model is presented. Some concluding remarks are provided in section 3.6.

3.2 The Dirichlet process

In a Bayesian framework, parameters are not viewed as fixed quantities whose values are unknown to us. Rather, parameters are thought of as random quantities that arise from probability distributions. For the sake of discussion, consider parameters $\theta_1, \ldots, \theta_n \in \mathcal{R}$ from a parametric Bayesian model where

$$\theta_i \overset{\text{iid}}{\sim} G_0. \tag{3.1}$$

In (3.1), we specify the parametric distribution G_0, and note that sometimes G_0 may depend on additional parameters. For example, G_0 may correspond to a normal distribution whose mean and variance are left unspecified.

With a Dirichlet process (DP) prior, we instead write

$$\begin{aligned} \theta_i &\overset{\text{iid}}{\sim} G \\ \text{where} \quad G &\sim \text{DP}(m, G_0). \end{aligned} \tag{3.2}$$

In (3.2), we are stating that the parameter θ arises from a distribution G but G itself arises from a distribution of distributions known as the Dirichlet process with concentration parameter $m > 0$ and mean $E(G) = G_0$. The Dirichlet process in (3.2) is defined (Ferguson 1974) as follows: For finite k and any measurable partition $(A_1, \ldots, A_k)$ of $\mathcal{R}$, the distribution of $G(A_1), \ldots, G(A_k)$ is Dirichlet$(mG_0(A_1), \ldots, mG_0(A_k))$. It is apparent that the baseline distribution G_0 may serve as an initial guess of the distribution of θ and that the concentration parameter m determines our apriori confidence in G_0 with larger values corresponding to greater degrees of belief. Under (3.2), we think of a distribution G arising from the Dirichlet process followed by a parameter θ arising from G.

There are several ways to implement a DP prior. The Pólya urn characterization of the DP (Blackwell and MacQueen, 1973) within a Markov chain sampling setting was considered by Escobar (1994), and Escobar and West (1995). The collapsed cluster sampling method of MacEachern (1994) and the "no-gaps" algorithm of MacEachern and Müller (1998) for non-conjugate DPM models also exploits the Pólya-urn structure.

An illuminating and alternative definition of the Dirichlet process was given by Sethuraman (1994). His constructive definition of (3.2) which is also known as the stick breaking representation is given as follows: Generate a set of iid atoms $\theta_i^* \sim G_0$ and generate a set of weights $w_i = y_i \prod_{j=1}^{i-1}(1 - y_j)$ where the y_i are iid with $y_i \sim \text{Beta}(1, m)$ for $i = 1, \ldots, \infty$. Then

$$G = \sum_{i=1}^{\infty} w_i I_{\theta_i^*} \tag{3.3}$$

where $I_{\theta_i^*}$ is a point mass at θ_i^*.

For our purposes, the Sethuraman (1994) construction is most useful. First, we see that the stick breaking mechanism creates smaller and smaller weights w_i. This suggests that

at a certain point we can truncate the sum (3.3) and obtain a reasonable approximation to G (Muliere and Tardella 1998). Ishwaran and Zarepour (2002) suggest that the number of truncation points be n when the sample size is small and $\sqrt{n}$ when the sample size is large. Secondly, in WinBUGS modelling, it is required to specify the distributions of parameters. Whereas the Ferguson (1974) definition does not provide an adequate WinBUGS specification, the truncated version of (3.3) can be easily implemented. Finally, the stick breaking construction clearly shows that a generated G is a discrete probability distribution which implies that there is non-negligible probability that θ's generated from the same G have the same value. As later demonstrated in the examples, it is often desirable to facilitate clustering in network modelling.

Although the DP is a highly technical tool, the simple introduction above is all that is required to use Dirichlet process priors in the network models considered in this paper.

3.3 Example 1: A social relations model

We consider a simplification of the social relations model considered by Gill and Swartz (2007). The model involves paired continuous observations y_{ijk} and y_{jik} where y_{ijk} represents the k-th response of subject i as an actor towards subject j as a partner, $k = 1, \ldots, n_{ij}$, $i \neq j$. In y_{jik}, the roles are reversed. We let n denote the number of subjects. The model expresses the paired responses in an additive fashion

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\ y_{jik} &= \mu + \alpha_j + \beta_i + \varepsilon_{jik} \end{aligned}$$

where μ is the overall mean, α_i is the effect of subject i as an actor, β_j is the effect of subject j as a partner and ε_{ijk} is the error term. We refer to μ, the α's and the β's as first-order parameters. The Bayesian model specification then assigns prior distributions

$$\mu \sim \text{Normal}(\theta_\mu, \sigma_\mu^2), \tag{3.4}$$

$$(\alpha_i, \beta_i)' \overset{\text{iid}}{\sim} \text{Normal}_2\left(0, \Sigma_{\alpha\beta}\right), \tag{3.5}$$

$$(\varepsilon_{ijk}, \varepsilon_{jik})' \overset{\text{iid}}{\sim} \text{Normal}_2\left(0, \Sigma_\varepsilon\right) \tag{3.6}$$

where

$$\begin{aligned}\Sigma_{\alpha\beta} &= \begin{pmatrix} \sigma_\alpha^2 & \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta \\ \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}, \\ \Sigma_\varepsilon &= \sigma_\varepsilon^2 \begin{pmatrix} 1 & \rho_{\varepsilon\varepsilon} \\ \rho_{\varepsilon\varepsilon} & 1 \end{pmatrix}.\end{aligned}$$

The parameters $\{\sigma_\alpha, \sigma_\beta, \rho_{\alpha\beta}, \sigma_\varepsilon, \rho_{\varepsilon\varepsilon}\}$ are called the variance-covariance parameters (or variance components). Note that the joint distributions (3.5) and (3.6) induce a dependence structure amongst the observations y_{ijk}. The interpretation of the variance-covariance parameters is naturally problem specific. However, for the sake of illustration, suppose that the response y_{ijk} is the k-th measurement of how much subject i likes subject j. In this case, $\rho_{\alpha\beta}$ represents the correlation between α_i and β_i, and we would typically expect a positive value. That is, an individual's positive (negative) attitude towards others is usually reciprocated. To complete the Bayesian model specification, hyperpriors are assigned as follows:

$$\begin{aligned}\theta_\mu &\sim \text{Normal}(\theta_0, \sigma_{\theta 0}^2), && (3.7)\\ \sigma_\mu^{-2} &\sim \text{Gamma}(a_0, b_0), && (3.8)\\ \Sigma_{\alpha\beta}^{-1} &\sim \text{Wishart}_2((\nu_0 R_0)^{-1}, \nu_0), && (3.9)\\ \sigma_\varepsilon^{-2} &\sim \text{Gamma}(c_0, d_0), && (3.10)\\ \rho_{\varepsilon\varepsilon} &\sim \text{Uniform}(-1.0, 1.0) && (3.11)\end{aligned}$$

where $X \sim \text{Gamma}(a, b)$ implies $\text{E}(X) = a/b$ and hyperparameters subscripted with a 0 are set to give diffuse prior distributions (Gill and Swartz 2007).

We now consider a modification of the above social relations model where the prior assumptions (3.4) through (3.11) are maintained except that (3.5) is modified according to

$$\begin{aligned}(\alpha_i, \beta_i)' &\overset{\text{iid}}{\sim} G \\ G &\sim \text{DP}(m, \text{Normal}_2(0, \Sigma_{\alpha\beta})) \\ m &\sim \text{Uniform}(0.4, 10.0).\end{aligned} \qquad (3.12)$$

Via (3.12), a DP prior has been introduced where the prior for the concentration parameter m is similar to the choices made by Ohlssen, Sharples and Spiegelhalter (2007).

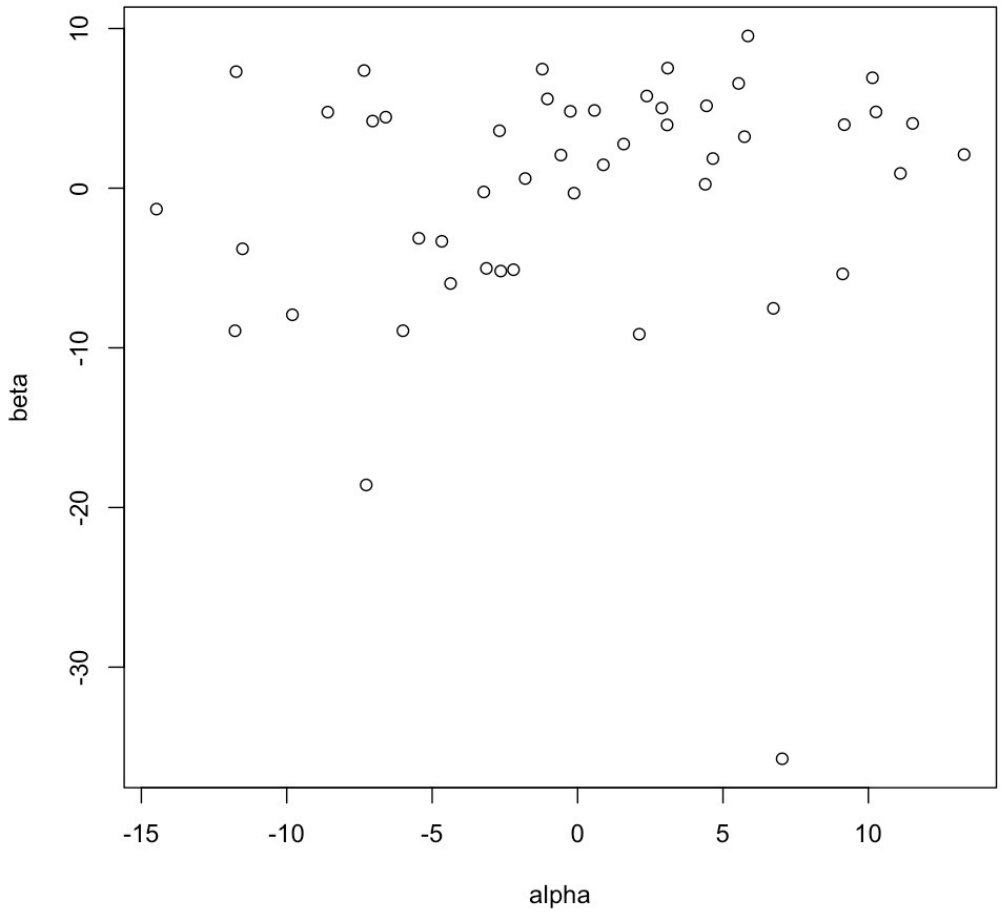

Figure 3.1: Posterior means of the (α_i, β_i) pairs under the normal prior in Example 1.

We have weakened the parametric normality assumption concerning (α_i, β_i) and have also introduced the potential for clustering individuals according to (α_i, β_i). In the context of interpersonal attraction, this is important as one can imagine four broad classifications of individuals:

- those who like others and are also liked
- those who like others and are disliked
- those who dislike others and are liked

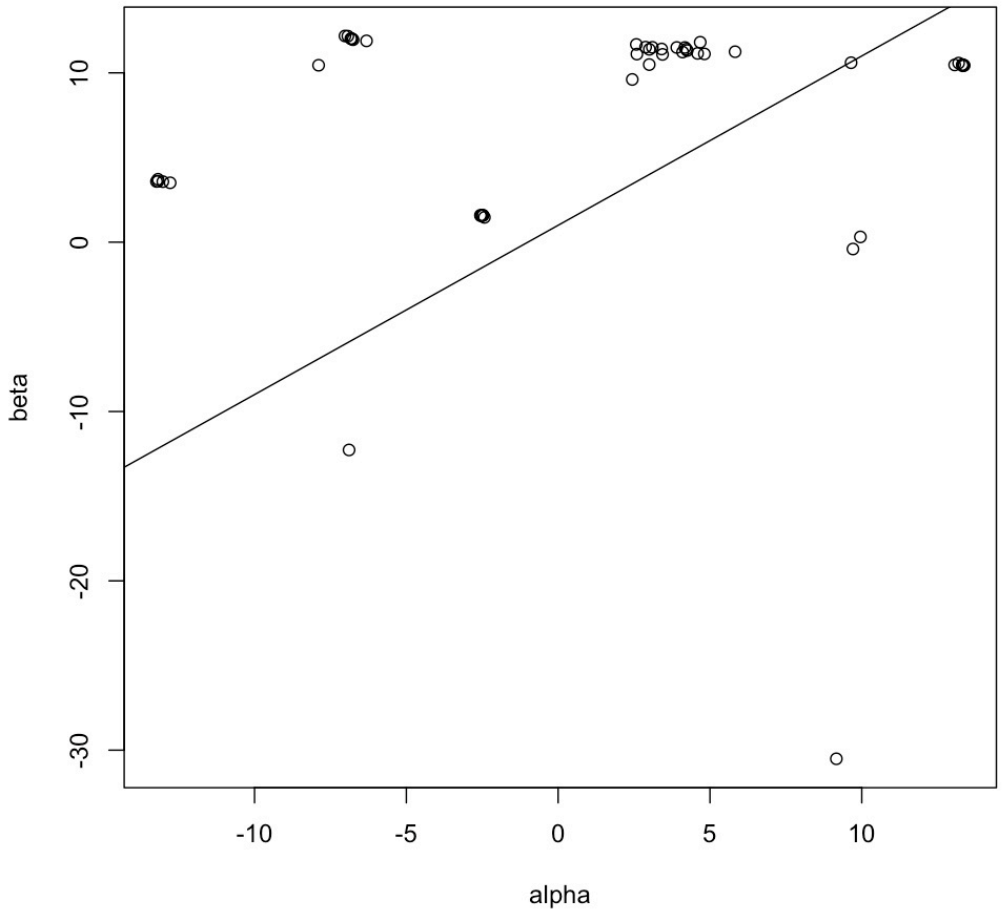

Figure 3.2: Posterior means of the (α_i, β_i) pairs under the DP prior in Example 1.

- those who dislike others and are also disliked .

Whereas social relations models focus on the variance components which are characteristics of the population, the enhanced social relations model using the Dirichlet process also permits the investigation of individuals.

To demonstrate the approach, we consider a study of students who lived together in a residence hall at the University of Washington (Curry and Emerson 1970). Data were collected on $n = 48$ individuals and measured on occasions $k = 1, 2, 3, 4, 5$ according to their pairwise levels of attraction. There is a missing data aspect to the problem as measurements

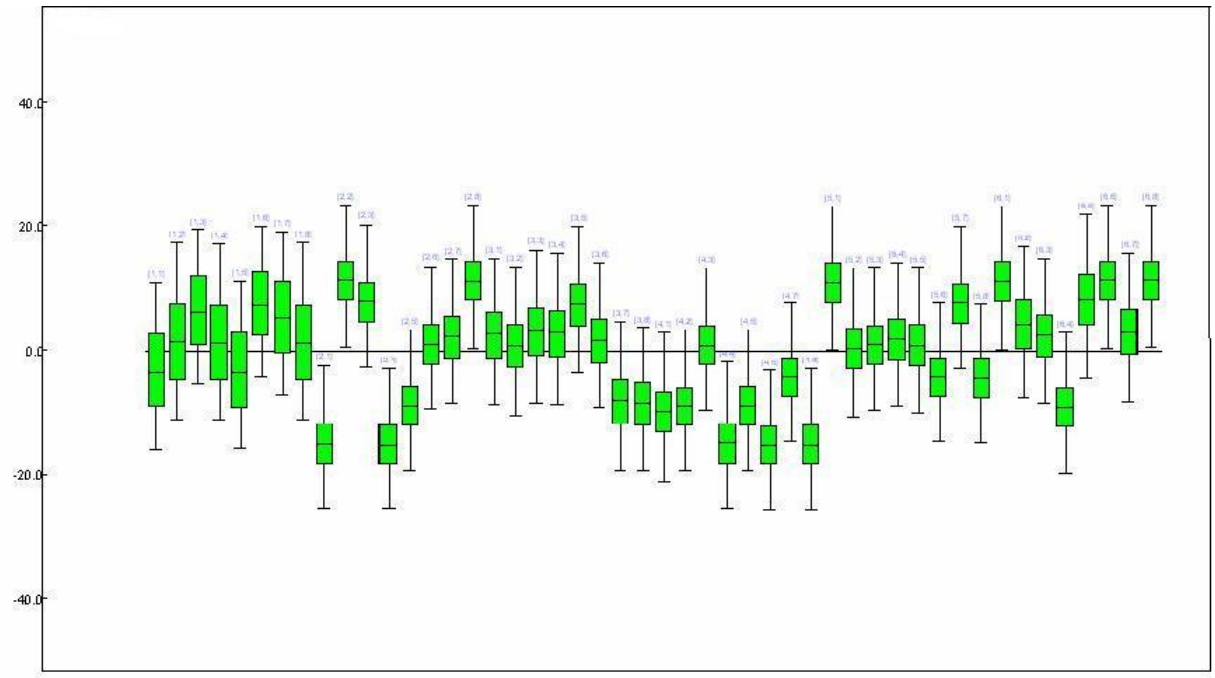

Figure 3.3: Posterior box plots of the α_i's under the DP prior in Example 1.

were only taken between pairs of 8 individuals in each of six dorm groups. Markov chain Monte Carlo simulations were carried out in WinBUGS using the original normal prior and the DP prior. We allow 5000 iterations for the sampler to converge and another 10000 iterations for sampling from the posterior. Convergence is checked visually and by using several starting points. In Figure 3.1, we provide a plot of the posterior means of the 48 (α_i, β_i) pairs using the normal prior. What is evident from Figure 3.1 is that the normal prior (3.5) is inconsistent with the data as the plot does not exhibit an elliptical shape. The outlier in the bottom right corner is also problematic. In Figure 3.2, we provide a plot of the posterior means of the 48 (α_i, β_i) pairs using the DP prior. We have also included the line $y = x$ for comparison purposes. Figure 3.2 suggests a tendency of individuals to cluster together with points scattered about the line $y = x$ corresponding to individuals who extend friendship to a similar extent that friendship is returned. The outlier in the bottom right corner corresponds to an individual who likes others but is disliked. The two clusters of points in the top left corner correspond to individuals who may be regarded as having false personalities; they do not generally like others although they convey signals that in turn

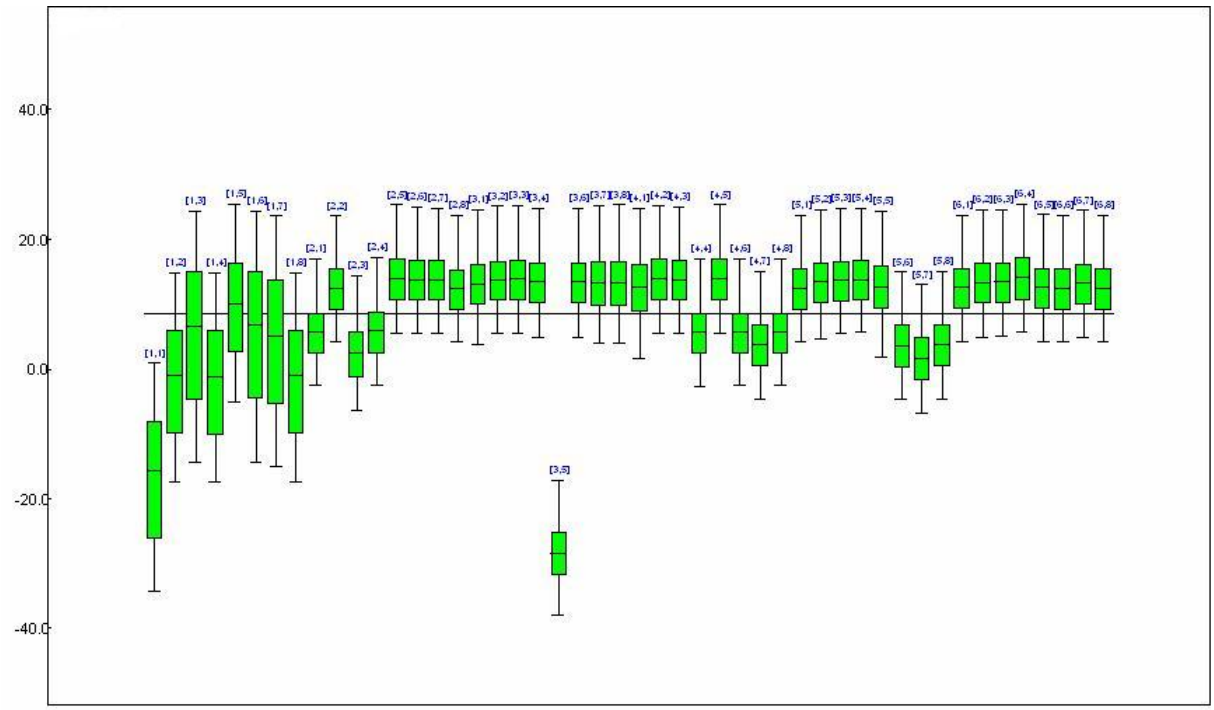

Figure 3.4: Posterior box plots of the β_i's under the DP prior in Example 1.

cause them to be liked. The posterior box plots of α_i's and β_i's are given in Figure 3.3 and 3.4. To further emphasize the superiority of the DP prior in this example, we calculate the log pseudo marginal likelihood (LPML) proposed by Gelfand, Dey and Chang (1992) as a model selection technique. Using the LPML, the DP prior is preferred (LPML $= -5016.9$) over the normal prior (LPML $= -5180.9$).

To investigate the effect of the prior choice involving the concentration parameter m in (3.12), we considered various priors. For instance, we considered $m \sim \text{Gamma}(2.0, 0.1)$ which is greatly different from the Uniform$(0.4, 10.0)$ prior. In comparing these two priors, we found that the posterior distributions of m differs substantially with $\text{E}(m|y) = 9.3$ under the Gamma prior and $\text{E}(m|y) = 6.6$ under the Uniform prior. However, our applied focus does not concern m. When looking at the posterior distributions of the (α_i, β_i) pairs under the two priors, we see very little difference. This is comforting and provides us with a sense of prior robustness with respect to the concentration parameter m.

In the Appendix B, we provide the WinBUGS code to emphasize the simplicity in which WinBUGS software facilitates the implementation of the DP prior in this example.

3.4 Example 2: A binary network model

We now consider an exponential random graph model previously studied by Gill and Swartz (2004). The data is an n by n matrix $Y = (y_{ij})$ describing the relationships between n nodes where $y_{ij} = 1$ denotes a tie from node i to node j and $y_{ij} = 0$ denotes the absence of such a tie. The p_1-model of Holland and Leinhardt (1981) states

$$\text{Prob}(Y) \propto \exp\left(\sum_{i<j} \phi y_{ij} y_{ji} + \sum_{i \neq j} (\theta + \alpha_i + \beta_j) y_{ij}\right) \tag{3.13}$$

where (3.13) implies the independence of the dyads $D_{ij} = (y_{ij}, y_{ji})$, $i < j$. The parameter ϕ measures the average degree of reciprocity or mutuality of ties in the population whereas θ measures the density of ties. The subject specific effects α_i and β_i represent the ability of subject i to extend and attract ties respectively. The Bayesian model specification then assigns prior distributions to the primary parameters of interest

$$\begin{aligned}
\phi &\sim \text{Normal}(\mu_\phi, \sigma_\phi^2), && (3.14)\\
\theta &\sim \text{Normal}(\mu_\theta, \sigma_\theta^2), && (3.15)\\
(\alpha_i, \beta_i)' &\overset{\text{iid}}{\sim} \text{Normal}_2\left(0, \Sigma_{\alpha\beta}\right). && (3.16)
\end{aligned}$$

To complete the Bayesian model specification, hyperpriors are assigned as follows:

$$\begin{aligned}
\mu_\phi &\sim \text{Normal}(\mu_0, \sigma_0^2), && (3.17)\\
\mu_\theta &\sim \text{Normal}(\mu_0, \sigma_0^2), && (3.18)\\
\sigma_\phi^{-2} &\sim \text{Gamma}(a_0, b_0), && (3.19)\\
\sigma_\theta^{-2} &\sim \text{Gamma}(a_0, b_0), && (3.20)\\
\Sigma_{\alpha\beta}^{-1} &\sim \text{Wishart}_2(r_0, \Sigma_0). && (3.21)
\end{aligned}$$

Again, the parameters subscripted with a 0 in expressions (3.17) through (3.21) are set to provide diffuse prior distributions. To implement the Dirichlet process version of this model, priors (3.14) through (3.21) are maintained except that (3.16) is modified as in (3.12). Clearly, there is similarity between these two models and the continuous data models of section 3.3. To investigate the enhanced Dirichlet process model, we consider a subset of the law firm data originally studied by Lazega and Pattison (1999). The directed data

matrix Y specifies whether or not advice was given between lawyers in a law firm consisting of 36 partners and 35 associates. The use of the Dirichlet process provides an approach to modelling the heterogeneity amongst the lawyers with respect to the parameters α and β. In the law firm example, one line of reasoning suggests that:

- senior lawyers are more likely to give advice but are less likely to receive advice (positive α and negative β)
- junior lawyers are more likely to receive advice but are less likely to give advice (negative α and positive β)
- intermediate lawyers are likely to provide advice to the same extent that it is sought (comparable α and β) .

The idea of partitioning the network actors into classes is very similar to the concept of blockmodelling. Wasserman and Faust (1994, chapters 10 and 16) describe in detail apriori and aposteriori blockmodelling. In apriori blockmodelling, exogenous attributes of actors are used for partitioning. For example, as a simple first approach following the previous intuition, we consider the introduction of the term δh_{ij} within the second sum in (3.13) where the covariate $h_{ij} = 1$ if i is a partner and j is an associate, $h_{ij} = -1$ if i is an associate and j is a partner, and $h_{ij} = 0$ otherwise. We then assign an additional prior

$$\delta \sim \text{Normal}(\mu_\delta, \sigma_\delta^2)$$

where

$$\begin{aligned} \mu_\delta &\sim \text{Normal}(\mu_0, \sigma_0^2), \\ \sigma_\delta^{-2} &\sim \text{Gamma}(a_0, b_0). \end{aligned}$$

Although this may appear sensible, there may very well be lawyers who do not fit the mold for apriori blockmodelling. For example, there may be young associates brimming with confidence who rarely ask for advice but readily offer their opinions. We prefer to let the data determine the clusters and this is possible with the proposed Dirichlet process model. Another objection to apriori blockmodelling is that often many models are fit before satisfactory covariates are determined. This suggests the problem of multiple comparisons

where the final model may only include covariates that fit the dataset in question and may not provide adequate fit to the population of interest. In aposteriori blockmodelling, estimates of the subject parameters α_i and β_i are obtained, and then standard clustering methods are applied to the estimates with the intention of grouping individuals. Aposteriori blockmodelling strikes us a somewhat of an ad-hoc procedure. We prefer a principled Bayesian approach where the individuals are clustered as a by-product of the DP model.

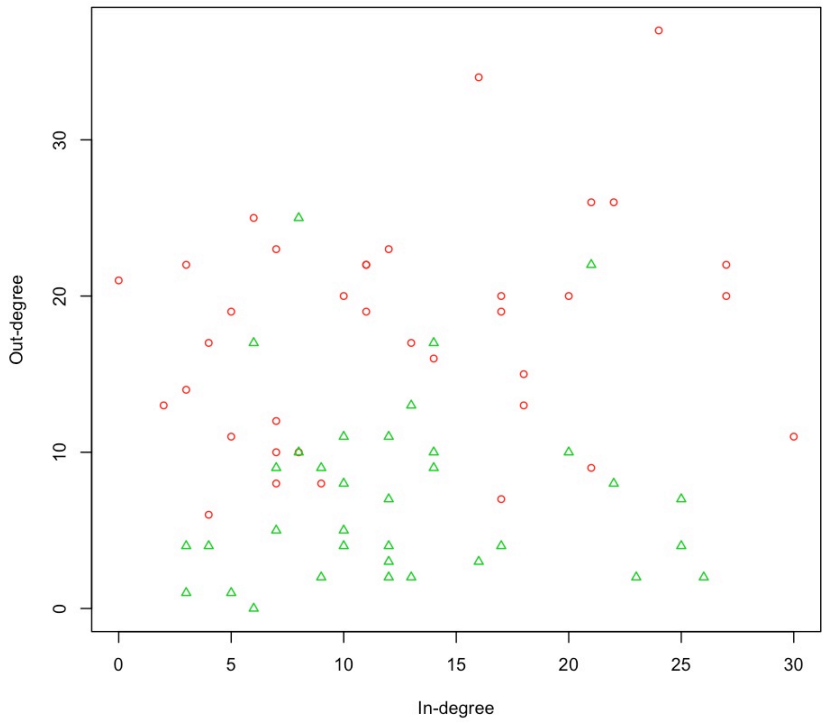

Figure 3.5: Plot of out-degree versus in-degree for the 71 lawyers in Example 2 where the lawyers labelled with triangles are associates and the lawyers labelled with circles are partners.

Figure 3.5 provides a plot depicting the relationship between providing advice and receiving advice. For each of the 71 lawyers, out-degree (number of individuals to whom advice was given) is plotted against in-degree (number of individuals from whom advice

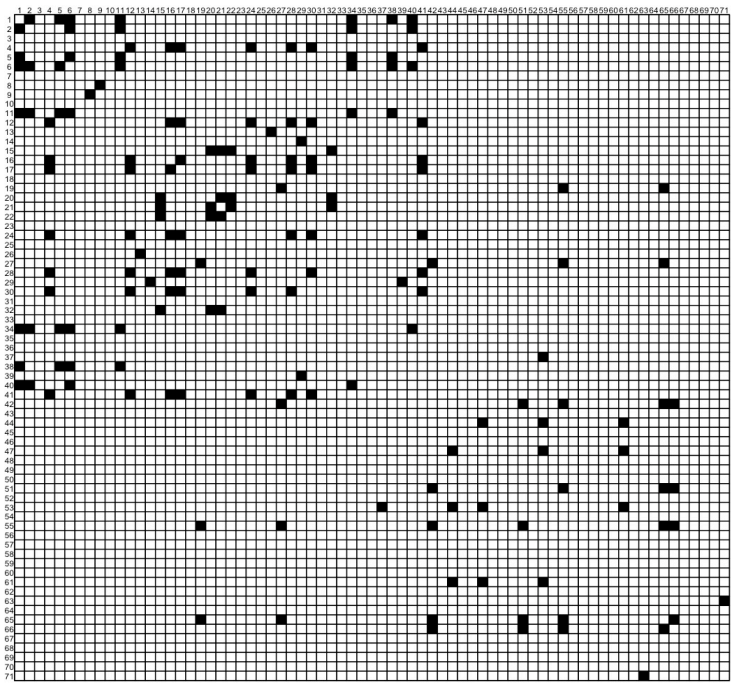

Figure 3.6: Plot of pairwise clustering of the 71 lawyers based on the DP model in Example 2. Black (white) squares indicate posterior probabilities of clustering greater than (less than) 0.5. Labels 1-36 correspond to partners and labels 37-71 correspond to associates.

was received). As expected, we observe that the younger associates generally give less advice than they receive. For example, one associate gave advice to only two colleagues yet received advice from 26 different colleagues. However, we notice that there are exceptions to the general heuristics. For example, there is a partner who gave advice to 11 colleagues yet received advice from 30 colleagues.

We fit the Bayesian DP model to the lawyer data and consider the clustering of (α_i, β_i) amongst the 71 lawyers. In a single iteration of MCMC, lawyers are clustered according to whether their (α_i, β_i) values are the same. In subsequent iterations of MCMC, the cluster membership may differ. With the MCMC output, we are able to calculate the

proportion of iterations that any given pair of lawyers cluster together and this provides an estimate of the posterior pairwise probability of clustering. We contrast this feature with aposteriori blockmodelling where clustering is based on a deterministic algorithm and there is no probability measure associated with resultant clusters. In Figure 3.6, we provide a plot which highlights the pairwise clustering involved in the DP analysis. For every pair of lawyers, a black square represents the posterior probability of clustering using a threshold value of 0.5. An interesting observation from Figure 3.6 is that the grid is roughly divided into four quadrants. It appears that partners (the top left quadrant) tend to cluster together and that associates (the bottom right quadrant) tend to cluster together. In other words, partners tend to behave similarly and associates tend to behave similarly. What this relevation suggests is that the original intuition was not quite right, and this argues again for the DP approach over apriori blockmodelling. In the DP approach, the data determine the clusters whereas in apriori blockmodelling, it is often easy to fail to find the best covariates.

3.5 Example 3: A Simulation Study

We report on a simulation study that investigates the performance of clustering using the Dirichlet process mixture in a simple binary network model. The model is a variation of logistic regression where binary responses describe the presence of ties between nodes. The simulated network data consist of an n by n matrix Y where $y_{ij} = 1$, $i \neq j$ indicates that subject i has a tie towards subject j, and $y_{ij} = 0$ denotes the absence of such a tie. Each $y_{ij} \sim \text{Bernoulli}(p_{ij})$ is assumed independent of other y's and the independence assumption is a common criticism of the simple model. We use a logistic link for $p_{ij} = Pr(y_{ij} = 1)$ whereby

$$\begin{aligned} log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \mu + \alpha_i + \beta_j \\ log\left(\frac{p_{ji}}{1-p_{ji}}\right) &= \mu + \alpha_j + \beta_i. \end{aligned} \tag{3.22}$$

In (3.22), the parameters α_i and β_i quantify the strength with which subject i produces and attracts ties respectively. With the inclusion of the α and β random effects, a type of dependency is introduced among dyads which share a common subject. The parameter μ measures the overall density of ties in the network.

In order to induce clustering amongst the random effects, we divide $n = 100$ subjects into four groups of equal size. This is a substantial network as each subject has 2(99) = 198 observations that describe its associated ties. The large size of the dataset helps demonstrate the utility of the approach. We set $\mu = 0$ and set the random effects according to $(\alpha_i, \beta_i) = (-1,-1), (1,1), (-1,1), (1,-1)$ for the four groups.

A Bayesian model for this network consists of the Bernoulli model description for Y, the logistic link (3.22), and the diffuse prior distributions $\mu \sim \text{Normal}(0, 10000)$, $(\alpha_i, \beta_i)' \overset{\text{iid}}{\sim} \text{Normal}_2(0, \Sigma_{\alpha\beta})$ and $\Sigma_{\alpha\beta}^{-1} \sim \text{Wishart}_2(2, I)$. To implement the Dirichlet process mixture version of the model, these priors are maintained except that the prior distribution for (α_i, β_i) is modified according to (3.2) where the number of truncation points $L = 20$ and the baseline distribution G_0 is the bivariate normal. The prior for the concentration parameter is given by $m \sim \text{Uniform}(0.4, 10)$ which is similar to the choices made by Ohlssen, Sharples and Spiegelhalter (2007).

In testing the adequacy of the model, we note that all of the 100 subjects are correctly clustered into their corresponding groups. The posterior probabilities of pairs of subjects (from the same group) clustering together range from 0.77 to 0.99. For pairs of subjects from different groups, the posterior probabilities of clustering are all 0.00. WinBUGS simulations for this substantial dataset require roughly two hours of computation for 20000 iterations.

We then modify the density parameter from $\mu = 0$ to $\mu = -1$ prior to simulating the data Y. This has the effect of radically decreasing the number of ties between subjects. Again, we find perfect clustering for the 100 subjects.

As a third test of the utility of the model, we introduce some variation in the random effects (α_i, β_i) as might be expected in most networks. We generate the (α_i, β_i) from a bivariate normal distribution having zero correlation and standard deviation 0.1 in both the α and β parameters. This time, the clustering is again perfect in the sense that none of the subjects from a given group cluster with subjects outside of their own group. However, there is a little bit of sub-clustering of subjects within their own groups. In particular,

- the group with mean (-1, -1) has two sub-clusters of sizes 3 and 22
- the group with mean (1, 1) has two sub-clusters of sizes 5 and 20

- the group with mean (-1, 1) is a single cluster
- the group with mean (1, -1) has three sub-clusters of sizes 2, 4, and 19.

Cluster membership is based on posterior probability of pairwise clustering exceeding 0.5. When the threshold level is reduced to 0.25, we again observe perfect clustering with each of the 100 subjects assigned to its original group.

3.6 Discussion

In this paper, we have considered the use of Dirichlet process priors for network problems. The relaxation of parametric assumptions and the ability to facilitate clustering are both seen as advantages in network analyses. Furthermore, the models that we have considered are easily implemented using WinBUGS software.

It is worth asking where DP priors can be reasonably employed in network models. There are many networks where data can be modelled using a random effects specification. When some of the random effects might possibly be the same, then it is good to have methodology to accommodate and identify this type of clustering, and DP mixture modelling accomplishes this goal. For example, in various disease transmission networks, it is useful to identify individuals who have high probabilities of transmission. By clustering these individuals, patterns of behaviour may be deduced and this may be useful in disease prevention. As another example, consider the complex network structures that can be studied between states or nations. These structures may involve trade, information flow, immigration/tourism, military cooperation, etc. Here, it may be useful to cluster the states or nations so that idealogical categorizations can be inferred. For example, it may be interesting to know which eastern countries (if any) are close idealogically to western countries.

There are a number of future directions for this line of research. We are interested in using the Dirichlet process in more complex network problems with more complex dyadic dependencies. We are also interested in the treatment of longitudinal data and dynamic data networks. The development of complementary software to handle the special features of Dirichlet modelling may also be of value.

Chapter 4

Bayesian Analysis of Ordinal Survey Data

4.1 Introduction

For the sake of convenience, many surveys consist of ordinal data, often collected on a five-point scale. For example, in a typical course evaluation survey, a student may express his view concerning an aspect of the course from a set of five alternatives: 1-poor, 2-satisfactory, 3-good, 4-very good, and 5-excellent. Sometimes five-point scales have alternative interpretations. For example, the symmetric Likert scale measures a respondent's level of agreement with a statement according to the correspondence: 1-strongly disagree, 2-disagree, 3-neither agree nor disagree, 4-agree, and 5-strongly agree. Student feed-back on course evaluation surveys represents a modern approach for measuring quality of teaching. Nowadays, a growing number of websites use student feed-back as their main peformance indicator in teaching evaluations. As an example, http://www.ratemyprofessors.com/ rate over one million professors based on student feed-back on a five-point ordinal scale. The scenario is similar in customer satisfaction surveys and social science surveys.

The simplest method of summarizing ordinal response data is to report the means corresponding to the ordinal scores for each survey question. At a slightly higher level of statistical sophistication, standard ANOVA methods may be applied to the ordinal scores by treating the data as continuous. However, the standard models for the analysis of ordinal

data are logistic and loglinear models (Agresti, 2010; McCullagh, 1980; Goodmann, 1979). These models correctly take into account the true measurement scales for ordinal data and permit the use of statistical inference procedures for assessing population characteristics. An overview of the methodologies for ordered categorical data is given by Liu and Agresti (2005).

The approach in this paper is Bayesian and considers an aspect of ordinal survey data that is sometimes overlooked. It is widely recognized that respondents may have differing personalities. For example, consider a company which conducts a customer satisfaction survey where there is a respondent with a negative attitude. The respondent may complete the survey with a preponderance of responses in the 1-2 range. In this case, a response of 1 may not truly represent terrible performance on the part of the company. The response may reflect more on the disposition of the individual than on the performance of the company. As another example of an atypical personality, consider an individual who only provides extreme responses of 1's and 5's. It would be useful if statistical analyses could adjust for personality traits. This is the motivation of the paper, and the tool which we use to account for personality traits is the Dirichlet process, first introduced by Ferguson (1973). As a by-product of the proposed methodology, we attempt to identify areas (survey questions) where performance has been poor or exceptional. In addition, we attempt to identify questions that are highly correlated. Clearly, surveyors desire accurate responses and by identifying highly correlated questions, it allows surveyors to remove redundant questions from the survey which in turn reduces fatigue on the part of the respondents.

Our paper is not the first Bayesian paper to consider this problem. Rossi, Gilula and Allenby (2001) refer to the problem as scale usage heterogeneity and our approach shares many features of their approach. Specifically, both approaches assume an underlying continuous response and the use of cut-points. We comment on the Ross, Gilula and Allenby (2001) paper in greater detail as we introduce our model. Alternative Bayesian approaches include Johnson (1996), Johnson (2003), Dolnicar and Grun (2007), Javaras and Ripley (2007) and Emons (2008). Johnson (2003) uses a hierachical ordinal regression model with heterogenious thresholds structure. Dolnicar and Grun (2007) use a ANOVA approach to assess the inter-cultural differences in responses. One of the main features of this paper

is that there is a mechanism to cluster subjects based on personality traits. Most importantly, in our approach, clustering takes place as a part of the model and data determine the clustering structure. Often, clustering is done in a post hoc fashion, following some fitting procedure. We also relax the bivariate normality assumption made by Rossi, Gilula and Allenby (2001). The normality assumption prevents outlying subjects and also conclusions may be sensitive to violations of normality. We relax this assumption by use of the Dirichlet process as the joint prior distribution.

In section 4.2, we provide a detailed development of the Bayesian latent variable model. The model assumes that ordinal response data arise from continuous latent variables with known cut-points. Furthermore, each respondent is characterized by two parameters that have a Dirichlet process as their joint prior distribution. The mechanism adjusts for classes of personality traits leading to standardized scores for respondents. Prior distributions are defined on the model parameters. We provide details about nonidentiability in our model and we overcome nonidentifiability issues by assigning suitable prior distributions. Computation is discussed in section 4.3. As the resulting posterior distribution is complex and high-dimensional, we approximate posterior summary statistics which describe key features in the model. In particular, posterior expectations are obtained via MCMC methods using WinBUGS software (Spiegelhalter, Thomas and Best 2003). In section 4.4, the model is applied to actual student survey data obtained in course evaluations. We demonstrate the reliability of the approach via simulation. In section 4.5, goodness-of-fit procedures are developed for assessing the validity of the model. The proposed procedures are simple, intuitive and do not seem to be a part of current Bayesian practice. We conclude with a short discussion in section 4.6.

4.2 Model development

Consider a survey where the observed data are described by a matrix $X : (n \times m)$ whose entries X_{ij} are the ordinal responses. The n rows of X correspond to the individuals who are surveyed and the m columns refer to the survey questions. Without loss of generality, we assume that the responses are taken on a five-point scale. For the time being, we also

assume that there are no missing data although the assumption is later relaxed.

When considering the degrees of opinion towards a particular survey question, it is reasonable to assume that X_{ij} arises from an underlying continuous variable Y_{ij}. We consider a cut-point model which converts the latent variable Y_{ij} to the observed X_{ij} as follows:

$$\begin{aligned} \mathrm{X}_{ij} = 1 &\Longleftrightarrow \lambda_0 < Y_{ij} \leq \lambda_1 \\ \mathrm{X}_{ij} = 2 &\Longleftrightarrow \lambda_1 < Y_{ij} \leq \lambda_2 \\ \mathrm{X}_{ij} = 3 &\Longleftrightarrow \lambda_2 < Y_{ij} \leq \lambda_3 \\ \mathrm{X}_{ij} = 4 &\Longleftrightarrow \lambda_3 < Y_{ij} \leq \lambda_4 \\ \mathrm{X}_{ij} = 5 &\Longleftrightarrow \lambda_4 < Y_{ij} \leq \lambda_5 \end{aligned} \tag{4.1}$$

Up until this point, our approach is identical to that of Rossi, Gilula and Allenby (2001). Our approach now deviates slightly as we assume that the cut-points are known and are given by $\lambda_0 = -\infty$, $\lambda_1 = 1.5$, $\lambda_2 = 2.5$, $\lambda_3 = 3.5$, $\lambda_4 = 4.5$ and $\lambda_5 = \infty$. We suggest that the chosen cut-points correspond to the way that respondents actually think. When asked to supply information on a five-point scale, we hypothesize that respondents make assessments on the continuum where the values $1.0, \ldots, 5.0$ have precise meaning. The respondents then implicitly round the continuous score to the nearest of the five integers. Although our methodology can be modified using unknown cut-points, the estimation of cut-points introduces difficulties involving nonidentifiability. Rossi, Gilula and Allenby (2001) address nonidentifiability and parsimony by imposing various complex constraints on the cut-points.

Using the notation $Y_i = (Y_{i1}, \ldots, Y_{im})'$, Rossi, Gilula and Allenby (2001) then model

$$Y_i \sim \text{Normal}(\mu + \tau_i \underline{1},\ \sigma_i^2 \Sigma) \tag{4.2}$$

for $i = 1, \ldots, n$ where τ_i and σ_i are respondent-specific parameters used to address scale usage heterogeneity. Although (4.2) contains many of the features we desire, it does not, for example, adequately describe an individual whose responses are mostly 2's and 4's.

We consider a structure that has similarities to (4.2). We propose

$$Y_i \sim \text{Normal}(b_i(\mu + a_i \underline{1} - 3\underline{1}) + 3\underline{1},\ b_i^2 \Sigma) \tag{4.3}$$

where it is instructive to adjust for personality traits via a "pure" or standardized score

$Z_i = (Z_{i1}, \ldots, Z_{im})' \sim \text{Normal}(\mu,\ \Sigma)$ such that

$$Y_{ij} = b_i(Z_{ij} + a_i - 3) + 3 \tag{4.4}$$

for $i = 1, \ldots, n$.

For an interpretation of the *disposition* parameter $a_i \in \mathcal{R}$ in (4.4), it is initially helpful to consider a_i conditional on $b_i = 1$. In this case, when $a_i = 0$, the ith respondent has a neutral disposition and the latent response Y_{ij} is equal to the standardized score Z_{ij}. When $a_i > 0$ ($a_i < 0$), the ith respondent has a positive (negative) attitude because Z_{ij} is adjusted by a_i to give Y_{ij}. In the Ross, Gilula and Allenby (2001) parameterization, the parameter τ_i in (4.2) is the counterpart to the parameter a_i in (4.4).

For an interpretation of the *extremism* parameter $b_i > 0$ in (4.4), it is helpful to consider b_i conditional on $a_i = 0$. In this case, when $b_i > 1$, the amount by which Z_{ij} exceeds 3.0 is magnified and is added to 3.0 and gives a more extreme result towards the tails on the five-point scale. When $0 \leq b_i < 1$, then the extremism parameter has the effect of pulling the latent response Y_{ij} closer to the middle. A respondent whose $b_i \approx 0$ might be described as moderate and we impose the constraint $b_i > 0$ to avoid nonidentifiability. Note that the parameter σ_i in (4.2) addresses variability which is somewhat different from our concept of extremism.

To provide a little more clarity, when $Z_{ij} + a_i - 3 > 0$, the i-th respondent is positively inclined towards survey question j. When $Z_{ij} + a_i - 3 < 0$, the i-th respondent is negatively inclined towards survey question j. The quantity $Z_{ij} + a_i - 3$ is then scaled by b_i to account for extremism on the part of the i-th respondent. The personality differential $b_i(Z_{ij} + a_i - 3)$ is then added to the non-commital score 3 to yield the latent variable Y_{ij}. Having adjusted for respondent personalities, we are interested in the average response μ for the m questions and the corresponding correlation structure Σ. We recognize that not all individuals share the same temperment. The i-th respondent is characterized by the parameters a_i and b_i where a_i is the disposition parameter and b_i is the extremism parameter.

As the proposed approach is Bayesian, prior distributions are required for the model

parameters in (4.3). Specifically, we assign moderately diffuse priors

$$\begin{aligned} \mu_j &\sim \text{Uniform}(0, 6) \\ \Sigma^{-1} &\sim \text{Wishart}_m(I, m) \end{aligned}$$

where the components of $\mu = (\mu_1, \ldots, \mu_m)'$ are apriori independent. For the personality traits a_i and b_i, the prior assignment is based on the supposition that there are classes of personality traits. We therefore consider the Dirichlet process

$$\begin{aligned} (a_i, b_i)' &\overset{\text{iid}}{\sim} G \\ G &\sim \text{DP}(\alpha, \text{tr-Normal}_2(\mu_G, \Sigma_G)) \end{aligned} \tag{4.5}$$

for $i = 1, \ldots, n$. The specification in (4.5) states that (a_i, b_i) arises from a distribution G but G itself arises from a distribution of distributions known as the Dirichlet process. The Dirichlet process in (4.5) consists of the concentration parameter α and baseline distribution tr-Normal$_2(\mu_G, \Sigma_G)$) where tr-Normal refers to the truncated bivariate Normal whose second component b_i is constrained to be positive. The baseline distribution serves as an initial guess of the distribution of (a_i, b_i) and the concentration parameter determines our confidence in the baseline distribution with large values of $\alpha > 0$ corresponding to greater degrees of belief. Prior distributions can be assigned to the hyperparameters in (4.5). Our analyses involving course evaluation surveys on a five-point give sensible results with $\alpha \sim \text{Uniform}(0.4, 10)$ (Ohlssen, Sharples and Spiegelhalter, 2007), $\mu_G = (0, 1)'$ and $\Sigma_G = (\sigma_{ij})$ where $\sigma_{11} = 1.0$, $\sigma_{22} = 0.5$ and $\sigma_{ij} = 0$ for $i \neq j$. Note that the choice of 1.0 and 0.5 are sufficiently diffuse in the range of parameters a_i and b_i. The key aspect of the Dirichlet process in our application is that the personality trait parameters (a_i, b_i) have support on a discrete space and this enables the clustering of personality types. An advantage of the Dirichlet process approach is that clustering is implicitly carried out in the framework of the model and the number of component clusters need not be specified in advance. Once a theoretical curiousity, the Dirichlet process and its extensions are finding diverse application areas in nonparametric modelling (e.g. Qi, Paisley and Carin 2007, Dunson and Gelfand 2009, Gill and Casella 2009).

4.3 Computation

The model described in section 4.2 is generally referred to as a Dirichlet process mixture model, and various Markov chain methodologies have been developed to facilitate sampling-based analyses (Neal 2000). However, these algorithms require considerable sophistication on the part of the programmer.

A goal in this paper is to simplify the programming aspect of the analysis by carrying out computations in WinBUGS. The basic idea behind WinBUGS is that the programmer need only specify the statistical model, the prior and the data. The Markov chain calculations are done in the background whereby the user is then supplied with Markov chain output. Markov chain output is then conveniently averaged to give approximations of posterior means.

To implement the analysis of our model in WinBUGS, we make use of the constructive definition of the Dirichlet process given by Sethuraman (1994). The definition is known as the stick breaking representation, and in the context of our problem, it is given as follows: Generate a set of iid atoms (a_i^*, b_i^*) from tr-Normal$_2$ (μ_G, Σ_G) and generate a set of weights $w_i = y_i \prod_{j=1}^{i-1}(1 - y_j)$ where the y_i are iid with $y_i \sim \text{Beta}(1, \alpha)$ for $i = 1, \ldots, \infty$. Then

$$G = \sum_{i=1}^{\infty} w_i I_{(a_i^*, b_i^*)} \tag{4.6}$$

where $I_{(a_i^*, b_i^*)}$ is the point mass at (a_i^*, b_i^*).

For programming in WinBUGS, the Sethurman (1994) construction is most useful as it allows us to approximately specify the prior. We see that the stick breaking mechanism creates smaller and smaller weights w_i. This suggests that at a certain point we can truncate the sum (4.6) and obtain a reasonable approximation to G (Muliere and Tardella 1998). Ishwaran and Zarepour (2002) suggest that the number of truncation points be n when the sample size is small and $\sqrt{n}$ when the sample size is large. The stick breaking construction clearly shows that a generated G is a discrete probability distribution which implies that there is non-negligible probability that (a_i, b_i)'s generated from the same G have the same value. This facilitates the clustering of personality traits in ordinal survey data. We note that the original definition of the Dirichlet process (Ferguson 1973) does not provide a WinBUGS-tractable expression for the prior.

4.4 Examples

4.4.1 Course evaluation survey data

The proposed model is fit to data obtained from teaching and course evaluations in the Department of Statistics and Actuarial Science at Simon Fraser University (SFU). The standard questionnaire at SFU contains $m = 15$ questions with responses on a five-point scale ranging from 1 (a very negative response) to 5 (a very positive response) where the specific interpretation of responses are question dependent. The survey questions are given as follows:

1. The course text or supplementary material was
2. I would rate this course as
3. The assignments and lectures were
4. The assignments and exams were on the whole
5. The marking scheme was on the whole
6. How informative were the lectures
7. The Instructor's organization and preparation were
8. The Instructor's ability to communicate material was
9. The Instructor's interest in the course content appeared to be
10. The Instructor's feed back on my work was
11. Questions during class were encouraged
12. Was the Instructor accessible for extra help
13. Was the Instructor responsive to complaints/suggestions
14. Overall, the Instructor's attitude towards students was
15. I would rate the Instructor's teaching ability as

Data were collected from $n = 75$ students pertaining to an introductory Statistics course. Various summary statistics corresponding to the parameter μ are given in Table 4.1. These are based on a MCMC simulation using WinBUGS with a burn-in period of 1000 iterations followed by 4000 iterations, taking roughly two hours of computation on a personal computer. We observe that nearly all of the posterior means of the μ_j exceed the corresponding sample means of the individual questions. This suggests that ratings can be viewed more favourably (i.e. higher scores) once the personality traits of the students have been removed. The highest posterior mean was recorded for the 9-th question which asked about "the Instructor's interest in the course material". The smallest mean was recorded for the 10-th question which asked about "the Instructor's feed back on work". These results are consistent with past surveys taken in the same course with the same Instructor. In particular, the Instructor does not grade assignments and this yields some criticisms from the students. Note that the posterior standard deviations are sufficiently small such that we can sensibly discuss the posterior means.

In Figure 4.1, we provide a plot of the posterior means of the (a_i, b_i) pairs. As expected, roughly 50% of the a_i's are greater than 0.0, and roughly 50% of the b_i's are greater than 1.0. There are some interesting observations that are revealed by the plot. Figure 4.1 indicates that there are four main clusters on the extremism parameter b and less clustering on the disposition parameter a. Note that one of the respondents provided a score of 5.0 for all $m = 15$ questions. It turns out that the corresponding posterior mean of (a_i, b_i) for this student was $(0.36, 1.09)$. Based on an average posterior response $\bar{\mu} = 4.1$, this student's mean latent Y-score is $1.09(\bar{\mu} + 0.36 - 3) + 3 = 4.59$ which rounds to a respondent X-score of 5.0 according to the cut-point model (4.1). This provides some evidence that the (a_i, b_i) parameters are estimated sensibly. As another example, the smallest posterior $a_i = -0.30$ was recorded for a student with a 2.06 average response for $m = 15$ questions.

It is also instructive to look at the posterior mean of the variance-covariance matrix Σ which describes the relationships amongst the $m = 15$ survey questions. The largest correlation value 0.63 occurred between survey questions 14 and 15. This is consistent with our intuition and personal teaching experience whereby students think highly of their instructors when they believe that their instructors care about them. The second highest

Table 4.1: Estimates of posterior means and posterior standard deviations for the actual SFU survey data.

Parameter	Sample Mean	Posterior Mean	Posterior SD
μ_1	3.62	3.69	0.19
μ_2	3.53	3.53	0.15
μ_3	3.89	4.04	0.18
μ_4	3.40	3.45	0.17
μ_5	3.62	3.85	0.17
μ_6	4.24	4.54	0.19
μ_7	4.06	4.33	0.18
μ_8	4.05	4.41	0.17
μ_9	4.40	4.78	0.15
μ_{10}	3.40	3.23	0.17
μ_{11}	4.11	4.51	0.19
μ_{12}	3.89	4.01	0.18
μ_{13}	3.96	4.11	0.17
μ_{14}	4.22	4.57	0.19
μ_{15}	4.05	4.52	0.18
$\bar{\mu}$	3.89	4.10	

correlation 0.57 occurred between survey questions 6 and 7 which is also believable from the view that learning is best achieved when material is clearly presented. However, we emphasize that the elimination of questions on the basis of redundancy should not be done solely on the basis of high correlations. In addition to high correlations, we should also have similar psoterior means. With the estimated posterior means $\mu_{14} = 4.57$ and $\mu_{15} = 4.52$, SFU may feel comfortable in dropping either question 14 or question 15 from the survey. Furthermore, we note that there were no negative posterior correlations and the minimum correlation 0.11 occurred between question 1 and question 13. Our intuition suggests that these two questions are independent.

Figure 4.2 provides an estimate of the posterior density of μ_1 using a kernel smoother from WinBUGS. The plot suggests a nearly symmetric unimodal distribution as might be expected. Similar plots were obtained for the other μ parameters.

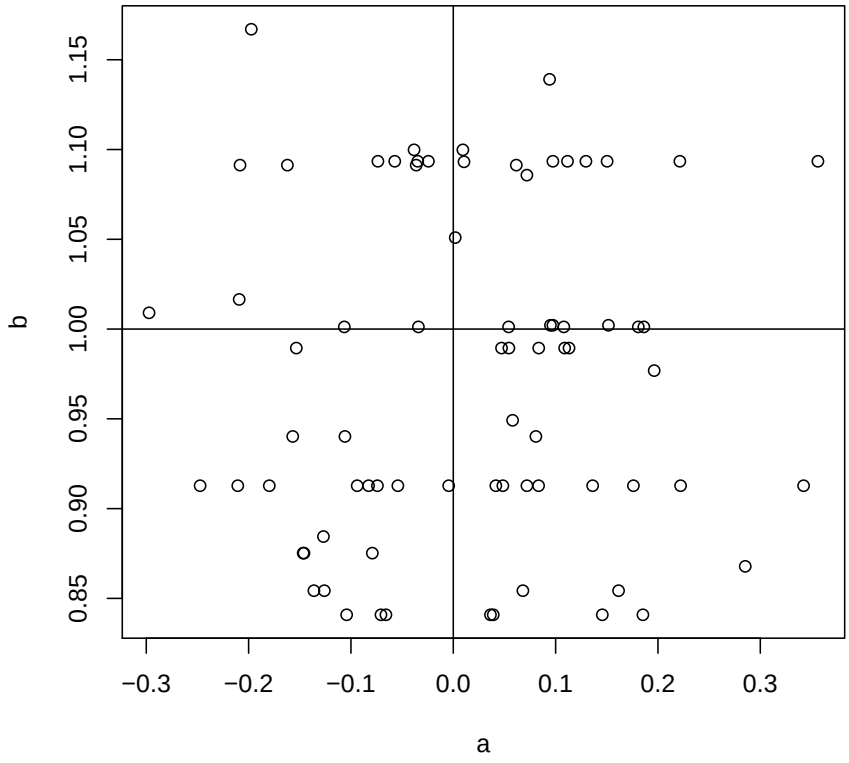

Figure 4.1: Plot of the posterior means of the personality trait parameters (a_i, b_i) for the actual SFU survey data.

It is good statistical practice to look at some plots related to the MCMC simulation. A trace plot for μ_1 is given in Figure 4.3. The trace plot appears to stabilize immediately and hence provides no indication of lack of convergence in the Markov chain. In Figure 4.4, an autocorrelation plot for μ_1 is also provided. The autocorrelations appear to dampen quickly. This provides added evidence of the convergence of the Markov chain and also suggests that it may be appropriate to average Markov chain output as though the variates were independent. Similar plots were obtained for all of the parameters in the model. In addition to the diagnostics described, multiple chains were generated to provide further assurance of the reliability of the methods. For example, the Brooks-Gelman-Rubin statistic

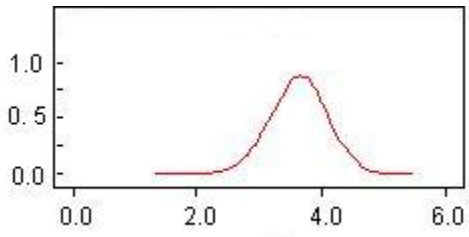

Figure 4.2: Estimate of the posterior density of μ_1 for the actual SFU survey data.

(Brooks and Gelman 1997) gave no indication of lack of convergence.

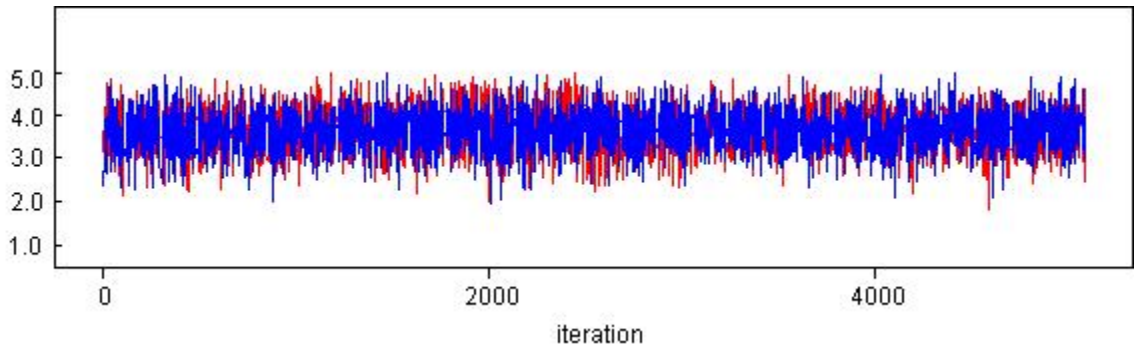

Figure 4.3: Trace plot for μ_1 based on the MCMC simulation for the actual SFU survey data.

4.4.2 Simulated data

Several simulation studies were carried out to investigate the model. We report on one such simulation. A dataset corresponding to $n = 150$ students with $m = 10$ questions was simulated using R code. In this example, the mean vector $\mu = (3, 3, 3, 3, 3, 3, 3, 3, 3, 3)'$ and variance covariance matrix $\Sigma = (\sigma_{ij})$ with $\sigma_{ii} = 4$ and $\sigma_{ij} = 2$ for $i \neq j$ were used in generating the latent variable matrix Y. The personality trait parameters a_i and b_i were generated from the Dirichlet process with $\alpha = 5.0$, $\mu_G = (0, 1)'$ and $\Sigma_G = 0.01I$ using the stick breaking method. Having generated the data as described, we then transformed the

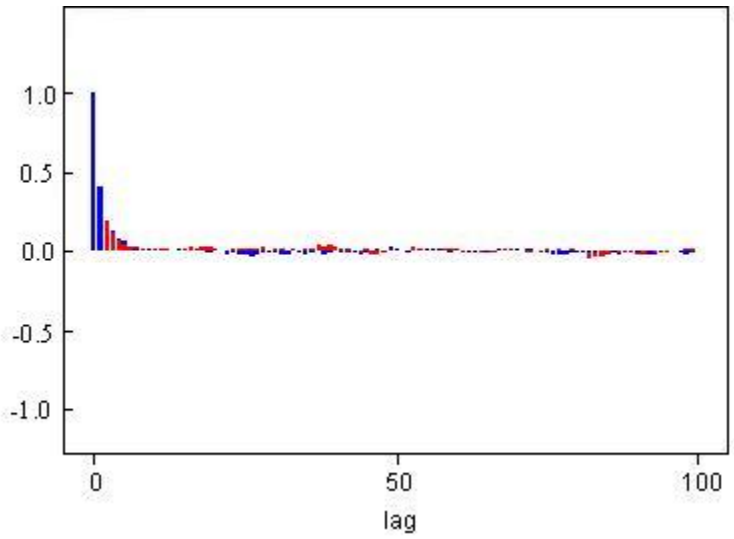

Figure 4.4: Autocorrelation plot for μ_1 based on MCMC simulation for the actual SFU survey data.

latent matrix to the observed data matrix X using the cut-point model (4.1).

The model was fit using WinBUGS software where 1000 iterations were used for the burn-in period. The posterior estimates in Table 4.2 were based on 4000 iterations. We observe that the posterior means of the mean vector are in rough agreement with the simulated values $\mu = 3.0$. The posterior means of Σ also appear consistent with the simulated values. The level of agreement improves as we increase the number of respondents n.

A plot of the posterior means of the personality trait parameters (a_i, b_i) is given in Figure 4.5. We are confident that the model is behaving as intended based on comparisons of the posterior estimates with the simulated values of (a_i, b_i). For example, there are 61, 53, 11 and 25 simulated (a_i, b_i) pairs in the first four quadrants respectively and this is in agreement with Figure 4.5. The level of agreement improves as we increase the number of survey questions m.

Table 4.2: Estimates of posterior means and standard deviations in the simulated data example.

Parameter	Posterior Mean	SD
μ_1	3.06	0.11
μ_2	3.05	0.11
μ_3	3.11	0.12
μ_4	3.06	0.13
μ_5	2.91	0.11
μ_6	3.07	0.12
μ_7	2.97	0.12
μ_8	2.93	0.13
μ_9	3.07	0.11
μ_{10}	3.03	0.13

4.5 Goodness-of-Fit

In Bayesian statistics, the assessment of goodness-of fit in complex models is problematic. A possible explanation for this is that aposteriori testing of model adequacy is not a Bayesian construct and may be seen as violating the Bayesian paradigm. From the point of view of a subjective Bayesian purist, any uncertainty concerning a model ought to be expressed via prior opinion. For example, if an experimenter is unsure whether the sampling distribution of the data is normal or Student, then the uncertainty might be expressed via a mixture. In theory, if we are able to express uncertainty in a model (and this includes both the sampling model and the parameters given the sampling model), then there is no need to assess model adequacy as all possible models have been considered and our inferences are subjective. However, from a practical point of view, it is typically difficult/impossible to determine the space of possible sampling models and parameters, and to assign prior opinion to the space.

Therefore, what does the practical Bayesian do in the context of model assessment? An honest answer may be that the assessment of complex Bayesian models is not a routine activity and there is no consensus regarding the "correct" approach. When Bayesian model assessment is considered, it appears that the prominent modern approaches are based on

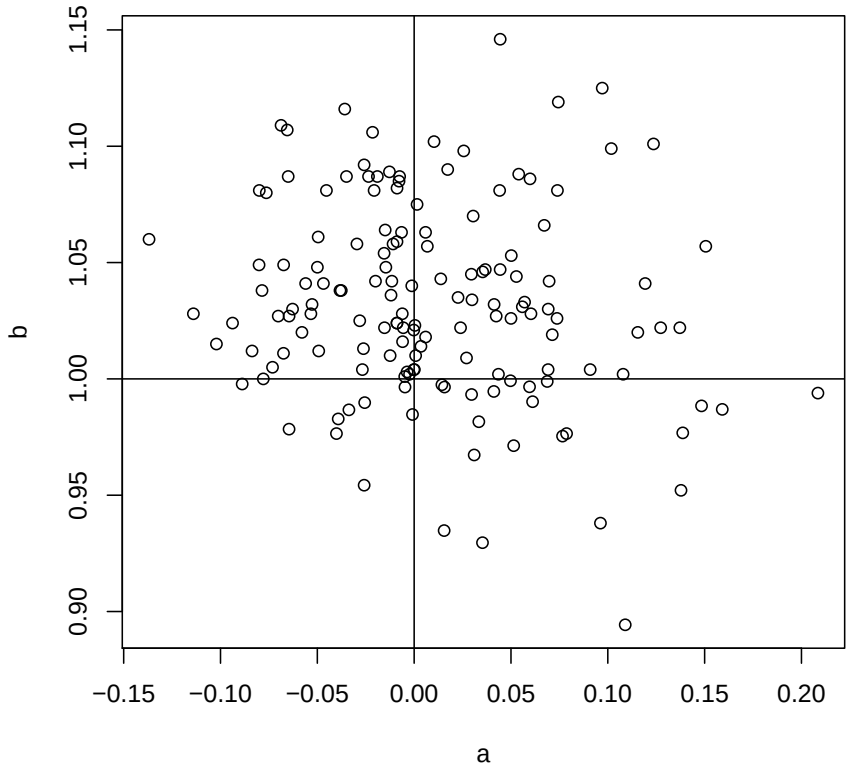

Figure 4.5: Plot of the posterior means of the personality trait parameters (a_i, b_i) in the simulated data example.

the posterior predictive distribution (Gelman, Meng and Stern 1996). These approaches rely on sampling future variates y from the posterior predictive density

$$f(y \mid x) = \int f(y \mid \theta)\ \pi(\theta \mid x)\ d\theta \tag{4.7}$$

where x is the observed data, $f(y \mid \theta)$ is the sampling density and $\pi(\theta \mid x)$ is the posterior density. In MCMC simulations, approximate sampling from (4.7) proceeds by sampling y_i from $f(y \mid \theta^{(i)})$ where $\theta^{(i)}$ is the i-th realization of θ from the Markov chain. Model assessment then involves a comparison of the future values y_i versus the observed data x. One such comparison involves the calculation of posterior predictive p-values (Meng 1994). A major difficulty with posterior predictive methods concerns double use of the

data. Specifically, the observed data x is used both to fit the model giving rise to the posterior density $\pi(\theta \mid x)$ and then is used in the comparison of y_i versus x. For this reason, some authors prefer a cross-validatory approach (Gelfand, Dey and Chang 1992) where the data $x = (x_1, x_2)$ are split such that x_1 is used for fitting and x_2 is used for validation.

We take the view that in assessing a Bayesian model, the entire model ought to be under consideration, and the entire model consists of both the sampling model of the data and the prior. We also want a methodology that does not suffer from double use of the data. For the models proposed here, we recommend an approach that is similar to the posterior predictive methods but instead samples "model variates" y from the prior predictive density

$$f(y) = \int f(y \mid \theta)\ \pi(\theta)\ d\theta \tag{4.8}$$

where $\pi(\theta)$ is a proper prior density. This approach was advocated by Box (1980) before simulation methods were common. It is not difficult to write R code to simulate $y_1, \ldots, y_N$ from the prior predictive density in (4.8). R code for the simulated data is given in Appendix D. It is then a matter of deciding how to compare the y_i's against the observed data matrix X. In our application, the data are high dimensional, and we advocate a comparison of "features" that are of direct interest. This is an intuitive and simple approach which is not part of current statistical practice. For example, one might compare observed subject means $\bar{X}_i = \sum_{j=1}^{m} X_{ij}/m$ with subject means generated from the prior predictive simulation. A simple comparison of these vectors can be easily carried out through the calculation of Euclidean distances. Naturally, as the priors become more diffuse, it becomes less likely to find evidence of model inadequacy. We do not view this as a failing of the methodology. Rather, if you really want to detect departures from a model, it is necessary that you have strong prior opinion concerning your model.

To provide a more stringent test, we consider a modification of our model where subjective priors $\mu_j \sim \text{Uniform}(2, 5)$ and $\Sigma_G = 0.01I$ are introduced. We assess goodness-of fit on the SFU data discussed in section 4.4. With N simulated vectors from the prior predictive distribution, there are $\binom{N+1}{2}$ Euclidean distances of interest; N of these distances are between the observed mean vector and the simulated vectors, and the remaining $\binom{N}{2}$ distances correspond to distances between simulated vectors. These distances are displayed

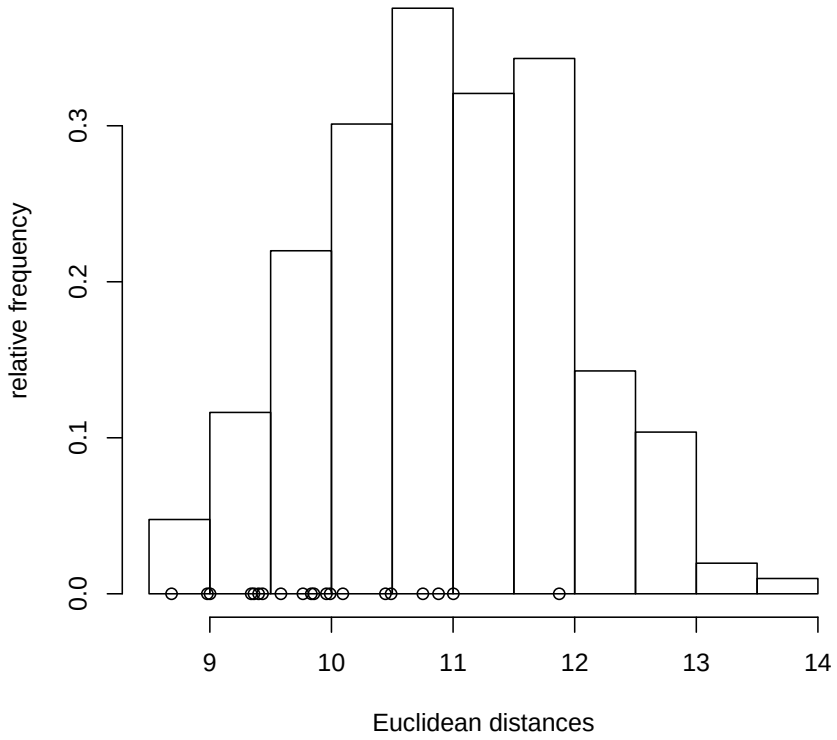

Figure 4.6: Histogram corresponding to the $\binom{N+1}{2}$= 210 Euclidean distances with respect to the prior-predictive check for the actual SFU survey data.

in a histogram with the $N = 20$ distances highlighted in Figure 4.6. Because these distances appear typical, there is no evidence of lack of fit. In fact, we observe that most of the Euclidean distances involving observed data lie on the left side of the histogram. This suggests that the most extreme variates arose from the prior-predictive distribution. Clearly, graphical displays for alternative features can also be produced.

4.6 Discussion

We have developed a Bayesian latent variable model to analyze ordinal response survey data. We have also relaxed the normality assumption made by Rossi, Gilula and Allenby (2001)

where we facilitate a clustering mechanism based on personality traits. Most importantly, clustering takes place as a consequence of the modelling using the Dirichlet process prior for personality trait parameters.

Our model identifies areas where performance has been poor or exceptional in a ordinal survey data by investigating standardized parameters. It also allows us to check whether some questions in a survey are redundant. A goodness-of-fit procedure is advocated that is based on comparing prior-predictive output versus observed data. The approach is intuitive and is flexible in the sense that one can investigate features which are relevant to the particular model. Future enhancements may be considered such as including subject covariates and handling longitudinal data structures.

Chapter 5

Discussion

The goal of this thesis was to show how Bayesian methodologies can be developed and implemented in WinBUGS in applied problems which deal with multidimensional parameters, complex model structures and complex likelihood functions. Three applied problems were considered in the fields of sport, network and survey data.

In the first project, we investigated the sport of cricket and developed a Bayesian latent variable model to simulate the game. A Bayesian framework was highly useful in order to capture the realism of the game. The Bayesian approach not only provided a flexible way to handle the variability of the game, it also allowed us to include prior knowledge about the game. This approach may be extended to T20 cricket which is the latest version of cricket. However, modelling rapid changes due to high variation of aggressiveness in this short form of the game may pose a challenge. Another future research direction is the development of realistic gaming software for cricket.

The second project proposed a semiparametric methodology to model network data using Dirichlet process priors. The approach not only provided a natural clustering mechanism which is highly useful in network data, it also relaxed the traditional normality assumptions. In future work, incorporating longitudinal aspects of the data and enhancing the methodology for triadic data are two challenges.

The third project developed a Bayesian latent variable model used to analyze ordinal survey data. The concept of personality traits that take place in survey data were introduced and modelled using Dirichlet process priors. The model was assessed using a prior predictive

approach. Incorporating covariate information of subjects in the model will be an interesting line of future research.

In all three projects, Bayesian inference was carried out via Markov chain Monte Carlo simulation. MCMC was implemented via the Metropolis-Hastings algorithm using WinBUGS software which is a very useful and powerful tool for Bayesian computation. Standard problems in MCMC simulation are the assessment of convergence and the determination of the length of the burn-in period. We used several graphical tools and the Brooks-Gelman-Rubin convergence statistic from WinBUGS to assess convergence and determine the length of the burn-in period.

Bibliography

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, Second Edition*, Wiley: New York.

Bailey, M.J. and Clarke, S.R. (2004). "Market inefficiencies in player head to head betting on the 2003 cricket world cup", In *Economics, Management and Optimization in Sport*, S. Butenko, J. Gil-Lafuente and P.M. Pardalos (editors), Heidelberg, Springer-Verlag, 185-202.

Bailey, M.J. and Clarke, S.R. (2006). "Predicting the match outcome in one day international cricket matches, while the match is in progress", *Journal of Science and Sports Medicine*, 5, 480-487.

Berry, S.M., Reese, C.S. and Larkey, P.D. (1999). "Bridging different eras in sports", *Journal of the American Statistical Association*, 94, 661-676.

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems", *Journal of the Royal Statistical Society, Series B*, 36, 192-236.

Besag, J. (2001). "Markov chain Monte Carlo for statistical inference", Working Paper No. 9, Center for Statistics and the Social Sciences, University of Washington.

Box, G. E. (1980). "Sampling and Bayes' inference in scientific modelling and robustness" (with discussion), *Journal of the Royal Statistical Society, Series A*, 143, 383–430.

Brooks, S.P. and Gelman, A. (1997). "Alternative methods for monitoring convergence of iterative simulations", *Computational and Graphical Statistics*, 7, 434-455.

Brooks, S.P. and Roberts, G.O. (1998). "Assessing convergence of Markov chain Monte Carlo algorithms", *Statistics and Computing*, 8, 319-335.

Cortes, C., Pregibon, D. and Volinsky, C. (2003). "Computational methods for dynamic graphs", *Journal of Computational and Graphical Statistics*, 12, 950-970.

Curry, T.J. and Emerson, R.M. (1970). "Balance theory: a theory of interpersonal attraction", *Sociometry*, 33, 216-238.

de Silva, B.M. and Swartz, T.B. (1997). "Winning the coin toss and the home team advantage in one-day international cricket matches", *The New Zealand Statistician*, 32, 16-22.

de Silva, B.M., Pond, G.R. and Swartz, T.B. (2001). "Estimation of the magnitude of victory in one-day cricket", *The Australian and New Zealand Journal of Statistics*, 43, 259-268.

Dey, D., Müller, P. and Sinha, D., editors (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lecture Notes in Statistics, Vol. 133, New York: Springer-Verlag.

Dolnicar, S. and Grun, B. (2007). "Cross-cultural differences in survey response patterns", *International Marketing Review*, 24, 127-143.

Duckworth, F.C. and Lewis, A.J. (1998). "A fair method for resetting targets in one-day cricket matches", *Journal of the Operational Research Society*, 49, 220-227.

Duckworth, F.C. and Lewis, A.J. (2004). "A successful operational research intervention in one-day cricket", *Journal of the Operational Research Society*, 55, 749-759.

Dunson, D.B. and Gelfand, A.E. (2009). "Bayesian nonparametric functional data analysis through density estimation", *Biometrika*, 96, 149-162.

Dyte, D. (1998). "Constructing a plausible test cricket simulation using available real world data", In *Mathematics and Computers in Sport*, N. de Mestre and K. Kumar (editors), Bond University, Queensland, Australia, 153-159.

Elderton, W.E. (1945). "Cricket scores and some skew correlation distributions", *Journal of the Royal Statistical Society, Series A*, 108, 1-11.

Emons, W.H.M. (2008). "Nonparametric person-fit analysis of polytomous item scores", *Applied Psychologial Measurement*, 32, 224-247.

Evans, M. and Swartz, T.B. (1995). "Methods for approximating integrals in Statistics with special emphasis on Bayesian integration problems", *Statistical Science*, 10, 254-272.

Evans, M. and Swartz, T.B. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*, Oxford University Press.

Ferguson, T.S. (1973). "A Bayesian analysis of some nonparametric problems", *Annals of Statistics*, 1, 209-230.

Ferguson, T.S. (1974). "Prior distributions on spaces of probability measures", *Annals of Statistics*, 2, 615-629.

Frank, O. and Strauss, D. (1986). "Markov graphs", *Journal of the American Statistical Association*, 81, 832-842.

Geisser, S. (1980). "Discussion on Sampling and Bayes' inference in scientific modelling and robustness by G.E.P. Box", *Journal of the Royal Statistical Society, Series A*, 143, 416-417.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). "Model determination using predictive distributions with implementation via sampling-based methods" (with discussion), In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, editors), Oxford: Oxford University Press, 147–167.

Gelman, A., Meng, X. L. and Stern, H. S. (1996). "Posterior predictive assessment of model fitness via realized discrepancies", *Statistica Sinica*, 6, 733–807.

Ghosh P., Muthukumarana, S., Gill, P.S. and Swartz, T.B. (2010). "A semiparametric Bayesian approach to network modelling using Dirichlet process priors", *The Australian and New Zealand Journal of Statistics*, To appear.

Gill, J. and Casella, G. (2009). "Nonparametric priors for ordinal Bayesian social science models: specification and estimation", *Journal of the American Statistical Association*, 104, 453-464.

Gill, P.S. and Swartz, T.B. (2004). "Bayesian analysis of directed graphs data with applications to social networks", *Applied Statistics*, 53, 249-260.

Gill, P.S. and Swartz, T.B. (2007). "Bayesian analysis of dyadic data", *American Journal of Mathematical and Management Sciences: Special Volume on Modern Advances in Bayesian Theory and Applications*, 27, 73-92.

Goodmann, L.A. (1979). "Simple models for the analysis of association in cross-classifications having ordered categories", *Journal of the American Statistical Association*, 74, 537-552.

Handcock, M.S. (2003). "Assessing degeneracy in statistical models of social networks", Working Paper No. 39, Center for Statistics and the Social Sciences, University of Washington.

Handcock, M.S., Raftery, A.E. and Tantrum, J.M. (2007). "Model-based clustering for social networks", *Journal of the Royal Statistical Society, Series A*, 170, 301-354.

Hoff, P.D. (2005). "Bilinear mixed effects models for dyadic data", *Journal of the American Statistical Association*, 100, 286-295.

Hoff, P.D. (2009). "Multiplicative latent factor models for description and prediction of social networks", *Computational and Mathematical Organization Theory*, 15, 261-272.

Hoff, P.D., Raftery, A.E. and Handcock, M.S. (2002). "Latent space approaches to social network analysis", *Journal of the American Statistical Association*, 97, 1090-1098.

Holland, P.W. and Leinhardt, S. (1981). "An exponential family of probability distributions for directed graphs", *Journal of the American Statistical Association*, 76, 33-65.

Ishwaran, H. and Zarepour, M. (2002). "Dirichlet prior sieves in finite normal mixtures", *Statistica Sinica*, 12, 941-963.

Javaras, K.N. and Ripley, B.D. (2007). " An 'unfolding' latent variable model for Likert attitude data: Drawing inferences adjusted for response style", *Journal of the American Statistical Association*, 102, 454-463.

Johnson, V.E. (1996). "On Bayesian analysis of multirater ordinal data: An application to automated essay grading", *Journal of the American Statistical Association*, 91, 42-51.

Johnson, T.R. (2003). "On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style", *Psychometrika*, 68, 563-583.

Kimber, A.C. and Hansford, A.R. (1993). "A statistical analysis of batting in cricket", *Journal of the Royal Statistical Society, Series A*, 156, 443-455.

Lazega, E. and Pattison, P.E. (1999). "Multiplexity, generalized exchange and cooperation in organizations: a case study", *Social Networks*, 21, 67-90.

Linkletter, C.D. (2007). "Spatial process models for social network analysis", *PhD Thesis, Department of Statistics and Actuarial Science, Simon Fraser University.*

Liu, I. and Agresti, A. (2005). "The analysis of ordered categorical data: An overview and a survey of recent developments", *Test*, 14, 1-73.

McCullagh, P. (1980). "Regression models for ordinal data (with discussion)", *Journal of the Royal Statistical Society*, Series B, 42, 109-142.

Meng, X.L. (1994). "Posterior predictive p-values", *The Annals of Statistics*, 22, 1142–1160.

Muliere, P. and Tardella, L. (1998). "Approximating distributions of random functionals of Ferguson-Dirichlet priors", *Canadian Journal of Statistics*, 26, 283-297.

Neal, R.M. (2000). "Markov chain sampling methods for Dirichlet process mixture models", *Journal of Computational and Graphical Statistics*, 9, 249-265.

Ohlssen, D., Sharples, L.D. and Spiegelhalter, D.J. (2007). "Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons", *Statistics in Medicine*, 26, 2088-2112.

Qi, Y., Paisley, J.W. and Carin, L. (2007). "Music analysis using hidden markov mixture models" *IEEE Transactions in Signal Processing*, 55, 5209-5224.

Rossi, P.E., Gilula, Z. and Allenby, G.M. (2001). "Overcoming scale usage heterogeneity", *Journal of the American Statistical Association*, 96, 20-31.

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors", *Statistica Sinica*, 4, 639-650.

Snijders T.A.B. and Kenny, D.A. (1999). "The social relations model for family data: a multilevel approach", *Personal Relationships*, 6, 471-486.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2004). *WinBUGS User Manual Version 1.4.1*, Medical Research Council Biostatistics Unit, Cambridge.

Swartz, T.B., Gill, P.S., Beaudoin, D. and de Silva, B.M. (2006). "Optimal batting orders in one-day cricket". *Computers and Operations Research*, 33, 1939-1950.

Swartz, T.B., Gill, P.S. and Muthukumarana, S. (2009). " Modelling and simulation for one-day cricket". *The Canadian Journal of Statistics*, 37, 143-160.

Thompson, S.K. (2006). "Adaptive web sampling", *Biometrics*, 62, 1224-1234.

Vance, E.A. (2008). "Statistical methods for dynamic network data", *PhD Thesis, Department of Statistical Science, Duke University.*

Ward, M.D. and Hoff, P.D. (2007). "Persistent patterns of international commerce", *Journal of Peace Research*, 44, 157-175.

Warner, R.M., Kenny, D.A. and Stoto M. (1979). "A new round robin analysis of variance for social interaction data", *Journal of Personality and Social Psychology*, 37, 1742-1757.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*, Cambridge: Cambridge University Press.

Wasserman, S. and Pattison, P. (1996). "Logit models and logistic regression for social networks: an introduction to Markov graphs and p*", *Psychometrika*, 61, 401-425.

Wong, G.Y. (1982). "Round robin analyses of variance via maximum likelihood", *Journal of the American Statistical Association*, 77, 714-724.

Wong, G.Y. (1987). "Bayesian models for directed graphs", *Journal of the American Statistical Association*, 82, 140-148.

Wood, G.H. (1945). "Cricket scores and geometrical progression", *Journal of the Royal Statistical Society, Series A*, 108, 12-22.

Appendix A

WinBUGS Code for the ODI Cricket Model

```
 model
 {
 for(i in 1:9)  # 9 situations
  {
  alpha[i,1]<- -10
  alpha[i,8]<-10
  }
 for(i in 1:9)  {
 for(j in 2:7)  {
u[i,j] <-alpha[i,j+1]
l[i,j] <-alpha[i,j-1]
} }

 for(i in 1:9) {
 for(j in 2:7) {
 alpha[i,j]~dnorm(0,1)I(l[i,j],u[i,j])
 }}

 for (i in 1:Nobowlers)  {
 mu2[i] ~ dnorm(0,tau) }

 for (i in 1:Nobats)  {
 mu1[i] ~ dnorm(0,tau) }

 tau ~ dgamma(0.1,0.1)
 sig <- 1/tau

 del[1] <- 0
 del[2] <- 0
 del[3] <- 0
```

```
 del[4] <- del1
 del[5] <- del1
 del[6] <- del1
 del[7] <- del1+del2
 del[8] <- del1+del2
 del[9] <- del1+del2

 del1 ~ dunif(0,1) ; del2 ~ dunif(0,1)

 for(i in 1:Nobs) {
 F[i,1] <- 0 ; F[i,8] <- 1
for(j in 2:7) {
 logit(F[i,j]) <- alpha[situ[i],j] - mu1[Batsman[i]] + mu2[Bowler[i]]
 - del[situ[i]]
 }}

   for(i in 1:Nobs) {
   for(j in 1:7) {
   p[i,j]<-F[i,j+1]-F[i,j]  }}

  for (i in 1:Nobs)  {
  n[i] <- sum(out[i,])
 out[i,1:7] ~ dmulti(p[i,1:7],n[i]) }}

 Data are in the format:
 Batsman[] Bowler[] situ[] out[,1] out[,2] out[,3] out[,4] out[,5]
 out[,6] out[,7]
   1 103   5     0    5    3    0    0    0    1
   1 103   6     0    3    2    1    1    0    0
   1 111   6     0    2    7    1    0    0    1
   1 113   6     1    2    2    1    0    1    0
   1 117   2     0    1    0    0    0    0    0
   1 117   5     0    1    0    0    0    0    0
   1 117   6     0    0    3    1    0    0    0
   1 123   5     0    2    2    0    0    0    0
   1 125   5     0    4    3    0    0    0    0
   1 125   6     0    1    3    1    0    1    0
   5 101   2     0    7    5    0    0    0    0
   5 101   3     0    1    2    1    0    1    0
   5 102   2     0    9    8    0    0    1    0
   5 102   3     0    2    2    1    0    0    0
   .
   .
   .
   END
```

Appendix B

WinBUGS Code for the Network Model

```
model{

# m = number of subjects ;  g = number of groups ;  k = number of occasions

  for(g in 1:6) { for(i in 1:m-1) { for(j in i+1:m)  { for (k in 1:5)  {

  y[g,(i-1)*m-i*(i-1)/2 +j-i,k,1:2] ~
  dmnorm(Ey[g,(i-1)*m-i*(i-1)/2+j-i,k,1:2],S[k,1:2,1:2])

  Ey[g,(i-1)*m-i*(i-1)/2+j-i,k,1] <- mu+a[g,i]+b[g,j]
  Ey[g,(i-1)*m-i*(i-1)/2+j-i,k,2] <- mu+a[g,j]+b[g,i]

  } } } }

  for (g in 1:6) { for (i in 1:m) {
   a[g,i] <- aa[(g-1)*m+i,1]
   b[g,i] <- aa[(g-1)*m+i,2]
  } }

# Prior for mu

  mu ~ dnorm(theta,tau)

# Hyperpriors
  theta ~ dnorm(0,0.0001); tau ~ dgamma(0.0001,0.0001)

# Hyperprior for var-cov matrix

  Sab[1:2,1:2] ~ dwish(Omega[1:2,1:2],2)
```

```
  Sigma[1:2,1:2] <- inverse(Sab[1:2,1:2])
  siga <- Sigma[1,1] ;  sigb <- Sigma[2,2]
  corrab <- Sigma[1,2]/(sqrt(Sigma[1,1]*Sigma[2,2]))

#  Prior for var-cov matrix of error terms

  for (k in 1:5) {
  taug[k] ~ dgamma(3,A02); sigg[k] <- 1/taug[k]

  corrg[k] ~ dunif(-1,1)

  S[k,1,1] <-  taug[k]/(1-corrg[k]*corrg[k])
  S[k,1,2] <- -corrg[k]*taug[k]/(1-corrg[k]*corrg[k])
  S[k,2,1] <- S[k,1,2]; S[k,2,2] <-  S[k,1,1]

  }
  A02 <- 100

# DP Priors for alpha's and beta's using stick-breaking method

  for (j in 1:48) {  for (kk in 1:2) {
  aa[j,kk] <- theta1[latent[j],kk]
  }  }
  for (i in 1:48) {
  latent[i] ~ dcat(pi[1:L1])
  }
  pi[1] <- r[1]

  for (j in 2:(L1-1)) {
  log(pi[j]) <- log(r[j])+sum(R1[j,1:j-1])

  for (l in 1:j-1) {
  R1[j,l] <- log(1-r[l])
  } }

  pi[L1] <- 1-sum(pi[1:(L1-1)])

  for (j in 1:L1) {
  r[j] ~ dbeta(1,alpha)
  }

# Baseline distribution for DP

 for (i in 1:L1) {
```

```
 theta1[i,1:2] ~ dmnorm(zero[1:2],Sab[1:2,1:2])
 }
 zero[1] <- 0; zero[2] <- 0

# Prior for concentration parameter
  alpha ~ dgamma(2,1)
}
```

Appendix C

WinBUGS Code for the Ordinal Survey Model

```
model
{
# cut-points model as in 4.1

alpha[1]<- -5
alpha[6]<-10
alpha[2]<- 1.5
alpha[3]<- 2.5
alpha[4]<-3.5
alpha[5]<-4.5


 for(i in 1:n)
 {
 for(j in 1:m)
 {
lo[i,j]<-((alpha[x[i,j]]-3)/b[i])+3 -a[i]

up[i,j]<-((alpha[x[i,j]+1]-3)/b[i])+3 -a[i]
}}

# Prior for mu

for (i in 1:m)  {
mu[i] ~ dunif(0,6)
 }
```

```
for(i in 1:n) {
y[i,1:m] ~ dmnorm(mu[] , G[,])I(lo[i,],up[i,])
}

#  Prior for variance-covariance matrix

G[1:m,1:m] ~ dwish(R[,],m)

varcov[1:m,1:m] <- inverse(G[,])

for(j in 1:m)
     {
        cor[j,j] <- varcov[j,j]
     }

for(i in 1:m-1)
    {
       for(j in i+1:m)
       {
           cor[i,j]<- varcov[i,j]/(sqrt(varcov[i,i]*varcov[j,j]))
           cor[j,i]<-cor[i,j]
        }
    }

# DP Priors for a's and b's using stick-breaking method as in 4.5

for ( i in 1:n) {
    a[i]<-aa[i,1]
b[i]<-(aa[i,2])
  }

for ( j in 1:n) {
for ( kk in 1:2) {
aa[j,kk]<- theta1[latent[j],kk]

}}
 for (i in 1:n) {
latent[i]~dcat(pi[1:L1])

  }
pi[1]<-r[1]
  for ( j in 2:(L1-1)) {
  log(pi[j])<-log(r[j])+sum(R1[j,1:j-1])

     for ( l in 1:j-1) {
```

```
        R1[j,l]<-log(1-r[l])
}
}
pi[L1]<-1-sum(pi[1:(L1-1)])


for ( j in 1:L1) {
r[j]~dbeta(1,mm)
}

 for ( i in 1:L1) {
 theta1[i,1:2]~dmnorm(zero[1:2],Sab[1:2,1:2])I(LB[],)

              }
       zero[1] <- 0; zero[2] <- 1

Sab[1:2,1:2] ~ dwish(Omega[1:2,1:2], 2)
varcovab[1:2,1:2] <- inverse(Sab[,])
corab<- varcovab[1,2]/sqrt(varcovab[1,1]*varcovab[2,2])

mm~dunif(0.4,10)

for( i in 1:n) {
   for (j in (i+1):n-1) {
            equalsmatrix[i,j]<-equals(aa[i,1],aa[j,1])*equals(aa[i,2],aa[j,2])
            equalsmatrix[j,i]<-equalsmatrix[i,j]
}}
}
```

Appendix D

R Code for the Prior Predictive Simulation

```
data=matrix(scan("surveydata.txt"),ncol=15,nrow=75,byrow=T)
s=vector(length=75)
mu=vector(length=15)
s=mean(data.frame(t(data)))

for (i in 1:20){

sigma = matrix(c(4,2,2,4),2,2)

n=75  # no. of subjects
m=15  # no.of questions
L1=30 # truncation in stick breaking method

mumu=matrix(ncol=15,nrow=n)
theta1=matrix(ncol=2,nrow=L1)
R1=matrix(ncol=L1-2,nrow=L1)
aa=matrix(ncol=2,nrow=n)
v=c(1:30)

a=vector(length=n)
latent=vector(length=n)

b=a
zero=vector(length=2)
r=vector(length=L1)
mm=runif(1,0.4,10)

for (i in 1:m){
mu[i]=runif(1,2,5)
```

```
}

G = matrix(ncol=15,nrow=15)

for (i in 1:m){
G[i,i]=4
}

for (i in 1:(m-1)){
for (j in (i+1):m){
G[i,j]=2
G[j,i]=2
}}

Sab = matrix(c(.01,0,0,.01),2,2)
zero[1] <- 0; zero[2] <- 1

 for ( i in 1:L1) {
 theta1[i,1:2]<-mvrnorm(1,zero[1:2],Sab[1:2,1:2])
     }

for ( j in 1:L1) {
r[j]<- rbeta(1,1,mm)
}

pi[1]<-r[1]

R1[2,1]<-log(1-r[1])
  for ( j in 2:(L1-1)) {
for ( l in 1:j-1) {
        R1[j,l]<-log(1-r[l])
}
}

for ( j in 2:(L1-1)) {
  pi[j]<-exp(log(r[j])+sum(R1[j,1:j-1]))
}
pi[L1]<-1-sum(pi[1:(L1-1)])

for (i in 1:n) {
latent[i]<-sample(v,1,replace=TRUE,pi)
  }

for ( j in 1:n) {
for ( kk in 1:2) {
```

```
aa[j,kk]<- theta1[latent[j],kk]

}}

for ( i in 1:n) {
   a[i]<-aa[i,1]
 b[i]<-(aa[i,2])
  }

z=matrix(ncol=m,nrow=n)
y=matrix(ncol=m,nrow=n)

z= mvrnorm(n=n,mu[], G)

for (i in 1:n){
for (j in 1:m){
y[i,j] =  b[i]*(z[i,j]+a[i]-3)+3
}}

x=matrix(ncol=m,nrow=n)

alpha=c(-10,1.5,2.5,3.5,4.5,15)

# converting latent matrix to observed data matrix using cut-points model

for (i in 1:n){
for (j in 1:m){

x[i,j] = ifelse(y[i,j]<(alpha[2]),1,x[i,j])
x[i,j] = ifelse(y[i,j]>(alpha[2]) & y[i,j]<(alpha[3]),2,x[i,j])
x[i,j] = ifelse(y[i,j]>(alpha[3]) & y[i,j]<(alpha[4]),3,x[i,j])
x[i,j] = ifelse(y[i,j]>(alpha[4]) & y[i,j]<(alpha[5]),4,x[i,j])
x[i,j] = ifelse(y[i,j]>(alpha[5]),5,x[i,j])

}}

s=rbind(s,mean(data.frame(t(x))))
}
```