Sexual Network Random Effects Models of Migration and Spread of HIV and other STIs in South Africa

A thesis presented to

The University of Waikato

in fulfilment of the requirement for the degree

of

Doctor of Philosophy in Statistics

by

Khangelani Zuma

Department of Statistics

The University of Waikato

Te Whare Wānanga o Waikato

University of Waikato

Abstract

South Africa is experiencing an explosive epidemic of Human Immunodeficiency Virus (HIV) and of Sexually Transmitted Infections (STI)s. Furthermore, South Africa has extraordinarily high rates of migration. The predominant type of migration is the *circular migration* in which young men migrate to work in urban areas leaving their sexual partners behind, to whom they return periodically. Conditions of migration bring men into sexual contact with prostitutes and other women at high risk of HIV/STIs. In this way, migrant men form sexual networks, which become a critical bridge for transmitting HIV/STIs between rural and urban areas.

The thesis investigates the determinants of HIV and those of STIs, taking into account the migration status and sexual network clustering effect in the data. The data investigated is from cohorts of *migrant men* from Hlabisa district working in urban areas, non-migrant *partner(s) of migrant men* residing in Hlabisa district, *non-migrant men* and their *non-migrant partner(s)* residing in Hlabisa district in northern KwaZulu-Natal, South Africa. Initially, the expectation-maximization (EM) algorithm is used to estimate parameters of the logistic-mixed model investigating risk factors of STIs. The interval-censored time until HIV infection is investigated using the Cox proportional hazards model which includes sexual network random effects in addition to the fixed effect. The parameters of this model were initially estimated using the EM algorithm. The main parameter estimation was carried out using the Gibbs sampler, a Bayesian Markov chain Monte Carlo (MCMC) method.

The results show that *migration* is a risk factor of HIV/STI. The results further show that *age*, *marital status*, *age at first sexual intercourse*, *sexual contact partners*, *lifetime partners* and other *biomedical factors* are important determinants of HIV/STIs. The study shows that ignoring sexual network random effects in the analysis of HIV/STIs biases the results. The Gibbs sampler is shown to be a plausible alternative to the EM algorithm in the analysis of correlated interval-censored data. It allows full Bayesian inference, which provides a natural framework with which to integrate the uncertainty about parameters and incorporate heterogeneity between sub-groups, without the need to evaluate high-dimensional integrals.

Notes

A number of papers have been produced from this thesis. The abstracts of the papers appear in the Appendix. The full references of the papers are as follows:

1. Zuma, K, Gouws E, Williams BG, *et al.* (2003). Risk factors for HIV infection among women in Carletonville, South Africa: migration, demography and sexually transmitted diseases. *International Journal of STD & AIDS*, **14**:814–817

2. Lurie MN, Williams BG, Zuma K, *et al.* (2003a). The impact of migration on HIV-1 transmission in South Africa: A study of migrant and nonmigrant men and their partners. *Sexually Transmitted Diseases*, **30**(2):149–156

3. Lurie M, Williams BG, Zuma K, *et al.* (2003b). Who Infects Whom? HIV-1 Concordance and Discordance Among Migrant and Non-Migrant Couples in South Africa. *AIDS*, **17**:2245–2252

4. Zuma K, Lurie M, Williams BG, *et al.* (2004). The risk factors of sexually transmitted infections among migrant and non-migrant sexual networks from rural South Africa. *Submitted for publication*

5. Zuma K, Lurie M, Jorgensen M. (2004). Analysis of interval-censored data from circular migrant and non-migrant sexual partnerships using the EM algorithm. *Submitted for publication*

6. Zuma K and Lurie MN. (2005). Application and comparison of methods for analysing correlated interval-censored data from sexual partnerships. *Journal of Data Science*, **3**(3):000–000

The earlier version of paper 4 was presented at the Australian Statistical Association conference in July 2002, Canberra: Australia and at the International Society for Clinical Biostatistics conference in September 2002, Dijon: France. The earlier version of paper 5 was presented at the Biostatistics seminar series at Limburgs University, September 2002, Diepenbeek: Belgium.

Acknowledgements

This work would not have been accomplished without the assistance of a number of persons and whom I would like to take these few lines to thank them warmly for their help. I owe my deep gratitude to my chief supervisor, Dr. William M. Bolstad for all his supervision and continual support in the course of this work. I also received valuable comments from Dr. Murray Jorgensen, the other member of the supervising panel.

I also wish to acknowledge the Migration Project Team without whom this project would not have been possible and particularly Dr. Mark Lurie for his valuable comments since the conception of this work. The study investigated was part of the Africa Centre for Population Studies and Reproductive Health, South Africa and was supported by the Wellcome Trust and South African Medical Research Council whom I also wish to acknowledge. This project would not have been done without the financial support of the New Zealand Overseas Development Assistance whom I wish to thank most dearly.

I am very thankful to my parents, my brother and sisters for their support and encouragement throughout my life.

Finally, on a personal note, I am very grateful to my family, whose presence and support have always been important to me. I dedicate this work to my wife Zodwa, our daughter Khabelihle Minenhle and our son Amahle Mhlabawethu. They all shared the burden of my limited ability to contribute effectively to our family life during this thesis work.

Contents

List of Figures

List of Tables

List of Abbreviations

STI	Sexually Transmitted Infection
HIV	Human Immunodeficiency Virus
AIDS	Acquired Immunodeficiency Syndrome
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
GEE	Generalized Estimating Equation
ML	Maximum Likelihood
EM	Expectation Maximization
IRLS	Iterative Re-weighted Least Squares
OR	Odds Ratio
RR	Relative Risk
MCMC	Markov Chain Monte Carlo
ARS	Acceptance-Rejection Sampling
SIR	Sampling Importance Resampling
AdRS	Adaptive-Rejection Sampling
DAG	Directed Acyclic Graph

Map of South Africa

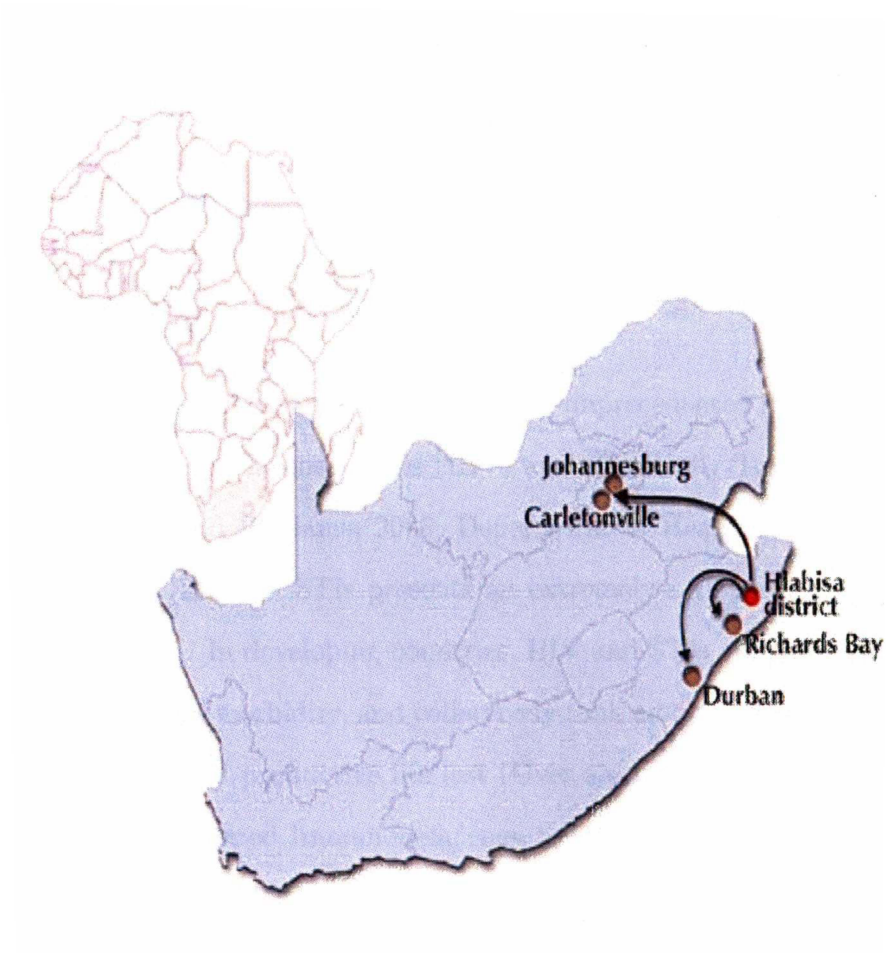


Figure 0.1: *The map of South Africa with the study sites*

Chapter 1

Introduction

1.1 Overview

In the last decade, South Africa has experienced an unprecedented rise in the prevalence of Human Immunodeficiency Virus (HIV) and of Sexually Transmitted Infections (STI)s (Gouws and Williams, 2000; Department of Health, 2002). The rise in the prevalence of HIV and STIs presents an extremely serious threat to public health in South Africa. In developing countries, HIV and STIs are among the leading causes of substantial morbidity, and collectively rank among the most important causes of years of healthy productive life lost (Over and Piot, 1993; Gerbase, *et al* 1998). Long before Acquired Immunodeficiency Syndrome (AIDS) was discovered as a consequent disease of HIV, STIs such as *gonorrhoea, chlamydia, syphilis* and *genital ulcers* ranked among top diseases for which sexually active individuals from developing countries sought health care (Buve, *et al* 1993). However, epidemiological factors determining the geographical spread of HIV/STIs are still not completely understood.

The geographical spread of HIV/STIs is determined by an interaction of factors related to demography, socioeconomic and sexual behaviour. The predominant socioeconomic factor is the rural-urban *labour migration* of young sexually active

men leaving their sexual partners behind (Pison, *et al* 1993; Decosas, *et al* 1995). Conditions of migration bring men into heterosexual contact with commercial sex workers and other women at high risk of HIV/STIs (Jochelson, *et al* 1991). The consequent sexual networking between urban and rural areas determines the diffusion rate of HIV/STIs into local societies (Fleming and Wasserheit, 1999). Furthermore, the women left behind sometimes have to exchange sex for favours as a survival strategy (Evian, 1993). The stark reality of the impact of HIV/STIs on the society requires deeper understanding of factors determining the spread of HIV/STIs and further understanding of the relationship between HIV and STIs.

The literature on the epidemiology and relationships between HIV and STIs is presented in Sections 1.1.1 to 1.1.3. Section 1.2 describes how the spread of HIV/STIs is influenced by the pattern of migration. Section 1.2.3 describes other important determinants of HIV/STIs. The clustered data set which is the main focus of this thesis is presented in Section 1.3. Models commonly used to analyse clustered data are reviewed in Sections 1.4. The model proposed in this thesis is formulated in Section 1.5. Section 1.6 presents the thesis objectives. Finally, in Section 1.7, the organization of subsequent chapters of the thesis is presented.

1.1.1 The epidemic and impact of STIs

Africa and other developing countries bear a heavy burden of STIs. Gonorrhoea, syphilis and chanchroid are the most common STIs (Mann, *et al* 1992). In 1995, Gerbase, *et al* (1998) estimated over 300 million new cases of syphilis, gonorrhoea, chlamydia and trichomoniasis in adults aged between 15 and 49 years worldwide. Gonorrhoea alone accounted for 18.8% of these new cases. The highest number of new cases of these STIs occurred in developing countries with 19.7% in sub-Saharan Africa.

STIs can cause acute symptoms such as genital ulcers and genital discharges.

The health repercussions of STIs affect women disproportionately. For example, STIs increase a woman's risk of ectopic pregnancy, which causes 1 to 15% of maternal deaths in developing countries (Population Reports, 1993). In women, STI pathogens can migrate from the lower reproductive tract causing pelvic inflammatory infection, which accounts for up to 40% of admissions to gynaecological wards in many African hospitals. Without prompt and appropriate treatment, 55 to 85% of women with pelvic inflammatory infection may become infertile (Piot and Tezzo, 1990). Pelvic inflammatory infection can further increase the risk of ectopic pregnancy (Meheus, 1992). In sub-Saharan Africa, 50% of infertility cases are attributed to pelvic inflammatory infection, which are usually caused by gonorrhoea or chlamydia (Adler, *et al* 1998). In men, infertility can follow a venereal infection that spreads from the urethra to the epididymis. The most common cause of epididymitis in men under 35 years is gonorrhoea or chlamydia infection (Piot and Tezzo, 1990).

1.1.2 Epidemiology and burden of HIV

AIDS was first recognized as a new disease in the early 1980s. However, it existed at least since the late 1970s. AIDS was first recognized among homosexual and bisexual men in the United States, and then in heterosexual men and women in Central and East Africa. The HIV virus was identified as the cause of AIDS two years after the identification of AIDS as a disease. Today, HIV affects all countries of the globe, making it and its disease consequences the most significant emerging infection of the late 20th century (Nicoll and Gill, 1999).

In South Africa, the first two cases of AIDS were reported in 1983 (Ras, *et al* 1983). Between 1983 and 1989 the population prevalence of HIV was estimated below 0.5%. HIV infection relentlessly spread out within urban areas and then to the rural areas in the early 1990s. The trend in HIV prevalence continued its unprecedented rise between 1995 and 2000. For a thorough review of literature on HIV epidemiology in South Africa, see for example Gouws and Williams (2000). The

national prevalence of HIV amongst women attending, for the first time, ante-natal clinics reached 23% in 1998 (Department of Health, 1999). KwaZulu-Natal province had the highest prevalence (32%) and Western Cape province had the lowest (5.2%) prevalence. Macro-simulation models predicted that the AIDS epidemic could reach 30% prevalence in sexually active population by 2000 to 2005 (Schall, 1990). The estimate is close to the current national antenatal HIV prevalence of 24.5% (Department of Health, 2001). However, there are some indications that the national prevalence has reached a plateau.

1.1.3 Relationship between HIV and STIs

The relationship between HIV and STIs is complicated. This is because HIV is also sexually transmitted and therefore shares the same behavioural risk factors and common human reservoirs as STIs. Thus, acquisition of an STI could merely be a marker of exposure to a sexual partner at higher risk of HIV infection (Mertens, *et al* 1990) rather than due to causal relationships. However, the epidemiological importance of STIs has acquired greater significance as it became apparent that they promote transmission of HIV and are important co-factors driving the HIV epidemic. The first evidence of possible relationships between STIs and HIV came from epidemiological studies that showed high prevalence of HIV among individuals with history of STIs (Wasserheit, 1992; Grosskurth, *et al* 1995; Fleming and Wasserheit, 1999).

Biological mechanisms facilitating interrelationship between HIV and STIs are well established (Cohen, *et al* 1997; Cohen, 1998). The studies show high shedding of HIV virus into genital fluids in the presence of genital ulcers and other inflammatory infections associated with non-ulcerative STIs. The implications are that people who are infected with HIV and have an STI are more infectious to their sexual partners than those infected with HIV but without an STI. Empirical data indicate that women infected with chlamydia or gonorrhoea are more susceptible to HIV

infection due to disproportionate increase in CD4 cell count in the endocervix and HIV virus targets this cell (Levine, *et al* 1994). Ulcerative STIs disrupt epithelial barriers in the genital tract. Disruption of epithelial barriers permits penetration of viral infections (Laga, *et al* 1993). Current evidence points to the conclusion that correct management of STIs should influence transmission of HIV. The community based randomised trial conducted in Tanzania demonstrated that improved treatment of STIs reduces the incidence of HIV (Grosskurth, *et al* 1995). Furthermore, immunosuppression associated with HIV can reduce resistance to STIs (Wasserheit, 1992).

1.2 Circular migration and HIV/STIs

1.2.1 Circular migration

Southern Africa has extraordinarily high rates of population movement both within and between countries. It is difficult to accurately quantify the extent and nature of population movements. However, Crush (1995) estimated that approximately 2.5 million legal migrants have come to South Africa from neighbouring countries along with an unknown number of illegal migrants. In addition, millions of men migrate within South Africa from rural to urban areas in search of work. In rural Hlabisa district of KwaZulu-Natal province (Figure 0.1) where this study was carried out, 62% of adult men spend most nights away from their homes (Lurie, *et al* 1997).

The roots of migration in South Africa can be traced back to the discovery of gold in the 1880s and the associated labour demands. Various types of migration currently exist in southern Africa. However, the predominant type is the *circular labour migration* in which young men migrate to work in urban areas leaving their rural sexual partners behind, to whom they return periodically. Furthermore, the system of circular labour migration was a cornerstone for apartheid policy, in which movements of South Africa's black population was strictly controlled. However,

patterns of migration have changed dramatically in the last decade. The rapid development of an informal but efficient transport infrastructure means people can now move freely between urban and rural areas.

1.2.2 Migration and spread of HIV/STIs

HIV, like other infectious diseases that spread from person to person, follows the movement of people (Quinn, 1994; Decosas, *et al* 1995; Decosas and Adrien, 1997; Mabey and Mayaud, 1997). Mobile people are at higher risk of HIV/STIs than those in stable living arrangements (Pison, *et al* 1993; Legarde, *et al* 1996). In Uganda, people who had changed residence within the last five years were three times more likely to be infected with HIV than those who had lived in the same place for more than ten years (Nunn, *et al* 1995). In South Africa, similar results were found among people who had recently changed their residence compared to those who had not changed their residence over time (Abdool Karim, *et al* 1992).

The role of migration in the spread of HIV has been described primarily as a result of migrant men becoming infected while away from home and infecting their partners when they return. In a study of seasonal migration in Senegal, Pison *et al* (1993) argued that the virus was mainly transmitted in two steps: first to adult single or married men through sexual contacts with infected women met during their seasonal migration, and second to their female partners when they return. Since this study focuses on seasonal migration, where men spend on average six months a year away from their rural homes, implications for South Africa may be important as migration patterns in the two countries appear to be similar. Kane, *et al* (1993) found higher prevalence of HIV among Senegalese men who had travelled and worked in another African country and among their rural sexual partners compared to men and women who had never travelled to another African country.

Decosas *et al* (1995) argue that it is not so much the movement itself rather

the *conditions and structure of migration* that put people at risk of HIV/STIs. Social and cultural data reveal that in many African countries where men migrate to cities, they engage in high risk sexual behaviour (Jochelson, *et al* 1991; Mbizvo, *et al* 1996; Mabey and Mayaud, 1997). In extreme cases, migrant men establish parallel families in urban areas and rural homes (Lurie, *et al* 1997). In this way, migrant men form sexual networks, which become a critical bridge for transmitting HIV/STIs between rural and urban areas.

Recently, the concept of sexual network core groups has become the integral part in understanding the epidemiology of HIV/STIs within human populations and identification of key populations for intervention programs (Wylie and Jolly, 2001; Koumans, *et al* 2001; Johnson, *et al* 2003). Sexual networks are often derived from contact tracing or asking participants to report on their partner's behaviour (Johnson, *et al* 2003). The leading studies in infectious diseases and sexual networks have demonstrated higher likelihood of HIV infection within core group sexual networks (Friedman, *et al* 1997). Results based on partner-reporting provide valuable information about sexual network sizes but fail, however, to provide sufficient information necessary to estimate the degree of heterogeneity between sexual networks.

1.2.3 Other determinants of HIV/STIs

Various other demographic and behavioural factors are associated with HIV (Celentano, *et al* 1996; Brewer, *et al* 1998; Auvert, *et al* 2001; Gibney, *et al* 2003 and references therein). A large *age difference* between sexual partners is an important risk factor for women (Gregson, *et al* 2002). Most women form partnership with men 5 to 10 years older than themselves. Understanding the effects of *number of lifetime partners*, *age at first sexual intercourse*, *recent sexual contact partners*, *condom use* and *type of sexual relationship* is much more problematic and often confounded by several factors including *respondent's age* and *duration of relationship*. *Alcohol use* has also been considered as a possible risk factor of HIV/STIs. In this discussion,

documented effects of these factors, as well as their interrelationships are highlighted.

Simulation models of sexual network partnerships and HIV/STI transmission identify measures of risk behaviour accumulated over the period parallel to HIV epidemic, such as *age at first sexual intercourse* and *number of lifetime partners* as important risk factors (Ghani and Garnett, 2000). An increase in the *number of lifetime partners* is associated with an exponential increase in the risk of HIV (Eisenberg, 1989; Auvert, *et al* 2001). However, the *number of lifetime partners* is related to the respondent's age since young people who recently started having sexual intercourse will most likely have fewer *lifetime partners* than older respondents.

The study of sexual networks of pregnant women found that *recent sexual contacts* accounted for most of HIV infections in women (Johnson, *et al* 2003). In this study, women reported fewer *lifetime partners* and the conclusion was that their increased risk was due to their male partners who had sexual contacts with commercial sex workers. Predominant risky sexual practices include having *casual sexual relationships*, *increased frequency* and *type of sexual contacts.* Empirical evidence shows less coital frequency in casual relationships than in marital relationships (Gregson, *et al* 2002). However, the risk of HIV is much higher in sexual contacts with *casual partners* than with *wives* or *regular partners* (Celentano, *et al* 1996; Auvert, *et al* 2001). Unmarried men engage in much more high risk sexual behaviour than married men (Gibney, *et al* 2003) and this increases the risk of infection among these men.

The *use of alcohol* is not itself a risk factor but sexual behaviour associated with drinking alcohol is. Gibney, *et al* (2003) reported that *drinking alcohol* was a significant factor associated with having sexual contact with a commercial sex worker. In most societies, very few women ever acknowledge *drinking alcohol* and thus data becomes less variable for any valid statistical analysis. Consistent *condom*

use during sexual intercourse is an effective measures of preventing heterosexual transmission of HIV/STIs (Conant, *et al* 1984). However, the frequency of *condom use* is relatively low and varies with the *type of relationship*. Occasional *use of condoms* is reported in *casual relationships* and, an even lower rate, in *regular* or *marital relationships*. The most disturbing aspect of *condom use* is that condoms are rarely used in casual relationships involving young women and older men because men consider young women to be free of HIV (Gregson, *et al* 2002). This partly explains higher rates of HIV infection among young women compared to men of the same age group.

1.3 The data set

In this thesis, the HIV and STIs data from migrant and non-migrant sexual networks from Hlabisa rural health district in northern KwaZulu-Natal South Africa, Figure 0.1, will be studied. South Africa is the country at the bottom of Africa. The Indian and Atlantic oceans form the eastern and western coastlines respectively, and they meet in the south at the Cape of Good Hope. To the north, South Africa is bordered by Namibia, Botswana, Zimbabwe, Swaziland and Mozambique. Lesotho is a country entirely surrounded by South Africa.

The data investigated is from cohorts of *migrant men* from Hlabisa district, *partner(s) of migrant men* residing in Hlabisa district, *non-migrant men* from Hlabisa district and *partner(s) of non-migrant men* residing in Hlabisa district. The study was designed to test the hypothesis that migrant men and their rural partners are at increased risk of HIV/STIs compared to non-migrant men and their partners. The investigation was intended to determine the extent to which the rural epidemic of HIV and that of STIs is being fuelled by circulation within the rural population as opposed to introduction from outside the home community by returning circular migrant men.

In October 1998, a sample of migrant men from Hlabisa district working in Carletonville gold mines near Johannesburg or Richard's bay factories near Durban (Figure 0.1) were invited from their workplaces to participate in the study. Migrant men were eligible to participate if they had been migrants for at least six months and had at least one regular sexual partner in Hlabisa who was not a migrant herself. Migrant men in this sample provided details of their sexual partners from Hlabisa, who were then located and invited to participate. In the neighbourhood of each migrant man's household, a non-migrant man and his partner(s) were selected and invited to participate. A non-migrant was defined as someone who spends most of the nights at home and has not been a migrant for more than a total of six months in the last five years.

Recruitment and logistical support for migrant men in Carletonville was embedded within a community based study carried out in Carletonville district. This community survey collected data from men and women aged between 13 and 60 years. The main objective of the survey was to investigate the extent of HIV infection in the community. The subsidiary goals were to determine the extent of female migration and investigate the risk factors of HIV among women who self-identified themselves as migrants compared to women who self-identified themselves as non-migrants in the area.

In the period between October 1998 and October 2001, the study participants were visited approximately every four months. During each visit, a detailed survey questionnaire was administered. The survey questionnaire elicited information related to demographic and socioeconomic characteristics, and to sexual behaviour and biomedical factors. In particular, the survey questionnaire collected information on each individual's accumulated sexual behaviour and partnership characteristics such as *age* and *other concurrent partners* within the last four months.

Two millilitres of venous blood were collected from those who consented to participate. The blood was screened for HIV using the Determine Rapid Test (Abbott Diagnostics). Samples that tested positive were re-tested using two additional ELISA tests (HIV 1.2.0 - Abbott/Murieux and Vironosticka HIV uniform 2+0, Omnimed). A random sample of 10% of the specimens that were negative on the Determine Rapid Test was also subjected to the ELISA confirmation to validate the specificity of the testing method. These tests remained negative on ELISA test.

All participants were offered extensive pre- and post-test counselling, condoms at each visit, and free treatment for symptomatic and laboratory-diagnosed STIs. The medical professional physically examined participants for presence of symptomatic STIs. Symptomatic ulcers and genital discharge were treated on enrolment according to the KwaZulu-Natal Province syndromic management guidelines (Department of Health, 1995), and laboratory-diagnosed STIs were treated ten days later.

The study group consists of 631 men and women aged between 18 and 69 years who were interviewed during the first clinical visit. There are 287(45.4%) women and 344(55.6%) men in the group. Of the men, circular migrants from Carletonville and Richard's bay accounted for 27.3% and 37.2% respectively and the rest of the men were non-migrants (35.5%). About 49.8% of women were partners of migrant men whilst 51.0% were partners of non-migrant men. The number of female partners interviewed for each man ranged from 0 to 4 women. Composition of sexual network partnerships consisted of 187 dyads, 40 triads, 4 quadriads and 1 pentad. The study planned to get data on each individual for the initial visit and on six follow-up visits at four monthly intervals. However, many participants dropped out of the study after each visit so the data on most participants covers only a few follow-up visits.

The study was part of the Africa Center for Population Studies and Reproductive Health, South Africa and was supported by the Wellcome Trust and South African Medical Research Council. The study was approved by the ethics committees of the University of Natal, Durban, South Africa and the Johns Hopkins University School of Hygiene and Public Health, USA.

1.4 Clustered data

Traditional statistical methods of data analysis assume that an individual response is the unit of analysis. The fundamental assumption of these statistical methods is that observational units are independent. Often data is collected using designs that gather data in dependent sub-groups or clusters. Familiar examples of clusters are families, schools or communities. In simple terms, a cluster is a collection of subunits on which observations are made. Another, common type of cluster is when observations are collected repeatedly on the same unit over time. The feature of clustered data is that observations within the same cluster tend to be more similar than observations in different clusters. The observations within a cluster are *correlated.* In standard settings, there is only one source of variation between observational units. Heterogeneity between clusters introduces an additional source of variation, which complicates the analysis. Classical methods that do not explicitly correct for clustering are inappropriate. Correlated data often arise in scientific disciplines such as health and social sciences, and require sophisticated statistical methods. In the next section, current statistical approaches to clustered data are described.

1.4.1 Models for correlated data

Scientific interest in clustered data is either in the pattern of change over time when measurements are taken repeatedly within the same unit, or simply the dependence of the *outcome variable* on the *explanatory variable(s).* The methods for

an approximately Gaussian *outcome variable* are well developed (Laird and Ware, 1982, Verbeke and Molenberghs, 1997; 2000). The linear mixed model has played a prominent role in extending the general linear model to handle correlated continuous data. The model relies on the elegant properties of a multivariate normal distribution. If the *outcome variable* is discrete, complete specification of the joint distribution of the response vector becomes problematic and likelihood methods get tedious. Three broad classes of models for clustered data have been proposed and are briefly described.

The first class of models often used to model clustered data is the class of *conditional models.* In conditional models, an outcome is modelled conditional on other outcomes rather than integrating them out. Parameter estimates from conditional models describe a feature of a set of outcomes conditionally on other outcomes. Conditional models are related to the family of *transition models* such as Markov models (Diggle, Liang and Zeger, 1994). Molenberghs and Ryan (1999) gave an example of such models in the case of binary response data. The main criticism of conditional models is their conditional interpretation of parameters on other outcomes and on cluster size. The conditional interpretation of the parameters renders these models less useful for regression analysis.

The second approach is *marginal models.* Marginal models directly model the marginal distribution of the response as a function of explanatory variables (Prentice, 1988; Liang, Zeger and Qaqish, 1992). In marginal models, the regression model is of scientific interest and correlation between observations within the same cluster is considered a nuisance parameter. However, we often do not know the precise details of the probabilistic function from which the data is generated. Liang and Zeger (1986) proposed a method of *generalized estimating equations* (GEE) that does not require assumptions about the complete joint distribution of the response vector. Zeger and Liang (1986) generalized the GEE approach. The GEE approach

provides a natural extension of quasi-likelihood (Wedderburn, 1974) to account for correlation within clusters. The GEE approach only requires correct specification of the univariate marginal probabilities with an adoption of some working assumption about the correlation structure. GEE has received much attention for some time, perhaps due to their relative computational ease and availability of good software (e.g. SAS procedure GENMOD). Prentice (1988) proposed an extension to GEE which allows for modelling of pairwise association using correlation or odds ratio (OR) as the measure of association. The main criticism of the GEE approach is that it does not generally correspond directly to a likelihood which could be used to calculate deviances (Hardin and Hilbe, 2003). Some approximations to the likelihood ratio statistic have been proposed (Rotnitzky and Jewell, 1990).

The model that mimics a linear mixed model for continuous data assumes the existence of an underlying unobserved continuous (*latent*) variable that represents various features shared by elements of a cluster and hence introduces correlation among observations. The latent variable is often called a *random effect.* Random effect models were introduced to account for extra-binomial variation due to larger variability among clustered binary responses than what would have been expected due to binomial variability alone. The families of linear mixed models and generalized linear models (GLM)s (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) are combined and form a class of generalized linear mixed models (GLMM)s if random effects are assumed normally distributed. These models have been extensively studied (see for example Stiratelli, Laird and Ware, 1984; Anderson and Aitken, 1985; Im and Gianola, 1988; Zeger, Liang and Albert, 1988; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993).

There are some critical distinctions between *marginal models* and *random effects models.* In marginal models, parameters are interpreted with respect to the marginal or population-averaged distribution. Such models are referred to as the *population-*

averaged models. In the random effects models, on the other hand, parameters have cluster-specific effects and are consequently called *cluster-specific* models. Cluster-specific models assume that correlation arises because regression parameters vary across clusters. This distinction is irrelevant in normal response variables since parameters have both the population-averaged and the cluster-specific interpretations. The distinction is critical in discrete data. Zeger, Liang and Albert (1988) discuss these two approaches to modelling of longitudinal data using the GEE.

1.4.2 Inference for GLMMs

Inference for GLMMs is a topic that has received much attention recently (Littell, *et al* 1996; Yu and Zelterman, 2002; Mills, *et al* 2002). Mills *et al* (2002) proposed an approach that allows for both population-averaged and individual-specific inference in GLMMs. Yu and and Zelterman (2002) discuss exact methods of inference when data are sparse or alternative hypothesis do not have a parametric form. The importance of each fixed effect in GLMMs is determined by a *Student's t-test* obtained by the ratio of the estimate to its standard error. The estimates of variability for fixed effects underestimate the true variability since they do not take into account the variability introduced when estimating random components. In practice, the *t-test* is used to account for this downward bias. The degrees of freedom are obtained as described in chapters 1 and 2 of Littell, *et al* (1996).

The deviance and the scale deviance are interpreted as goodness-of-fit chi-squared statistics for conditional models given the random effects (Littell, *et al* 1996). Self and Liang (1987) proposed a method for testing the importance of random effects in the model. They suggest fitting models with and without random effects and testing the difference in the deviance between the two nested models as a likelihood ratio chi-squared test. Modifications are needed to test for zero random effect variance since the hypothesized value may lie on the boundary of the parameter space. In

such a situation, Self and Liang (1987) proved that the likelihood ratio test is a mixture of χ_0^2 and χ_1^2 distributions. The correct approach is to use an upper tail critical region of the normal distribution. Thus, the usual ratio z-test, for a random effect variance is compared to a critical value of 1.645 at 0.05 level of significance.

1.5 Model formulation

The two primary outcomes of the thesis are time until HIV infection and infection with any of the curable STIs. The curable STIs considered are *active syphilis, chlamydia, gonorrhoea, genital ulcers* and *genital discharge.* Given the high cure rate of STIs through syndromic management, presence of any curable STI is considered a new infection. In the cross-sectional baseline descriptive analysis of HIV infection, an indicator of HIV status is all that is considered. This is in contrast to incidence analysis where time until HIV infection will be the primary outcome. One important dynamical feature of infection taking place within the constraints of a sexual network is the rapid increase of correlation in the infection status of members of a sexual network (Friedman, *et al* 1997). For example, many infected individuals will have their partners also infected since either they infected their partners, or vice-versa. In this section, a model which includes sexual network dependency through introducing a sexual network random effects term is formulated.

Consider an indicator of HIV or STI status of an individual from a particular sexual network. The infection status is nested within a sexual network. Sexual networks form clusters within which subjects are more alike than those from different sexual networks. Let y_{ij} $(i = 1, \cdots, I; j = 1, \cdots, J_i)$ denote the infection indicator for the jth individual in sexual network i. The infection indicator is 1 if an individual is infected and 0 otherwise. The extension to include series of observations from an individual over time follows immediately with the response becoming y_{ijk} where $(k = 1, \cdots, K_{ij})$. Let X_{ij} represent a known fixed design vector of explana-

tory variables and β be a p-dimensional vector representing covariate effects. The ith sexual network random effect is represented by b_i.

The response variable y_{ij} is premised on the assumption that it represents an underlying continuous variable T, known as 'threshold' or tolerance variable (Anderson and Aitkin, 1985; Im and Gianola, 1988). An individual is diagnosed with an infection if the tolerance level exceeds t which is the threshold or tolerance level. Thus, if π_{ij} is the conditional probability that the jth member of sexual network i is infected, then

$$\pi_{ij} = \Pr(T_{ij} > t|b_i) = \Pr(Y_{ij} = 1|b_i).$$

It is further assumed that the unobservable tolerance variable follows the mixed linear model

$$T_{ij} = \beta' X_{ij} + b_i + e_{ij}.$$

The e_{ij} are assumed symmetric and identically distributed random variables with a continuous unimodal density function $f()$. Without loss of generality, we set $t = 0$. In GLMMs the explanatory variables in X_{ij} and shared random effects b_i influence y_{ijk} through a linear combination $\eta_{ij} = \beta' X_{ij} + b_i$ where η_{ij} is called a *linear predictor.* The linear predictor η_{ij} is related to π_{ij} of y_{ij} through the *link function* g such that $\eta_{ij} = g(\pi_{ij})$.

1.5.1 Parameter estimation for GLMMs

The fundamental distributional assumptions for the GLMMs are first stated here below:

- The distribution of y_{ij} given b_i follows a distribution from the exponential family $f(y_{ij}|b_i; \beta)$,
- Given b_i, the observations $\{y_{i1}, \cdots, y_{iJ_i}\}$ are independent,
- The b_i are independent and identically distributed with density function $f(b_i; D, \Sigma)$ where D and Σ are the mean and variance respectively.

The parameters $\Lambda = \{D, \Sigma\}$ known as *hyperparameters* are also estimated from the data. The method of maximum likelihood (ML) estimation amounts to selecting estimates of those parameters that make the observed data most likely to have occurred. The likelihood function for unknown parameters β and Λ is

$$L(\beta, \Lambda; \mathbf{y}) = \prod_{i=1}^{I} \int \prod_{j=1}^{J_i} f(y_{ij}|b_i; \beta) f(b_i; \Lambda) db_i$$

which is the marginal distribution of the response vector $\mathbf{y}$ obtained after integrating out b_i. Except in normal linear models with normally distributed random effects, the computation of the marginal likelihood presents substantial problems because the marginal distribution of the response variable is usually intractable. In some instances numerical integrations may have to be used. For more complex problems involving high dimensional parameter space, numerical integration can be infeasible. Breslow and Clayton (1993) constructed a Laplace approximation for the marginal quasi-likelihood which is then maximized via linearization methods (Goldstein, 1995). Wolfinger and O'Connell (1993) proposed a pseudo-likelihood approach which circumvents the need for numerical integration. The approach of Wolfinger and O'Connell (1993) is implemented in SAS macro GLIMMIX. Zeger and Karim (1991) avoided the need for numerical integration by casting the GLMM in Bayesian framework and estimate parameters via the Gibbs sampler.

The standard approach in ML estimation is to take the partial derivatives of the log-likelihood with respect to each parameter and set them to zero. The resulting systems of equations are either solved directly or iteratively. The common strategy in mixed models is to use the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). In Chapter 2, we describe and implement the EM algorithm on the logistic mixed model. In the Bayesian paradigm, all unknown quantities are treated as random variables and a joint *hyperprior* probability distribution $p(\beta, \Lambda)$ is specified for the unknown parameters β and Λ that define the GLMM. The joint

posterior distribution $g(\mathbf{b}, \beta, \Lambda|\mathbf{y})$ is

$$g(\mathbf{b}, \beta, \Lambda|\mathbf{y}) = \frac{f(\mathbf{y}|\mathbf{b}; \beta)\, f(\mathbf{b}; \Lambda) p(\beta, \Lambda)}{\int\int\int f(\mathbf{y}|\mathbf{b}; \beta)\, f(\mathbf{b}; \Lambda) p(\beta, \Lambda) d\beta d\mathbf{b}\, d\Lambda}$$

where $\mathbf{b} = \{b_1, \cdots, b_I\}$. Finding $g(\mathbf{b}, \beta, \Lambda|\mathbf{y})$ analytically can be very complicated. Even computing it numerically is extremely problematic when the dimension of the parameter space is high. Instead we generate samples from the joint posterior $g(\mathbf{b}, \beta, \Lambda|\mathbf{y})$ that can be used to estimate any quantity that is of interest to us. In Chapter 4, we describe Bayesian techniques that can be used to generate samples from $g(\mathbf{b}, \beta, \Lambda|\mathbf{y})$. The marginal posteriors for regression parameters β and variance component parameters Λ are computed by marginalization.

1.6 The thesis objectives

The primary objective of this thesis is to investigate the effects of urban-rural circular migration of men on the spread of HIV and other STIs in the rural health district of KwaZulu-Natal, South Africa. The thesis aims to analyse the determinants of STIs and those of HIV infection, with the main focus being on migration status. These goals will be attained by developing, describing and evaluating statistical methods for modelling multivariate binary data and interval-censored time until HIV infection data. Therefore, the thesis is both substantive and methodological.

It is acknowledged that some sexual networks might have higher risk of STIs than others. Also, some sexual networks might have higher risk of HIV infection than other sexual networks (Friedman, *et al* 1997). The thesis aims to quantify the magnitude and the importance of sexual network clustering in the transmission of STIs and of time until HIV infection separately. The quantification of sexual network clustering effect of STIs will be achieved by using random effects models which allow for clustering of STI risk within sexual networks. The magnitude of the effect of sexual network clustering of time until HIV infection will be accomplished by using a random effects model that takes into account the interval-censored nature

of time until HIV infection and sexual network dependencies. The main focus on interval-censored frailty model will be on the full Bayesian inference which has not been previously used for this type of analysis and data.

The estimation of coefficients of the explanatory variables and variance components for sexual network random effects, for both response variables of interest, will also be carried out using the EM algorithm. A subsidiary goal for this thesis is to compare the resulting estimates of the EM algorithm to the full Bayesian estimates for the interval-censored frailty model obtained via the Gibbs sampler.

1.7 Structure of the thesis

The subsequent chapters of this thesis are organized as follows: Chapter 2 describes the basic theory behind the EM algorithm. The EM algorithm is then used to compute ML estimates of parameters associated with risk factors of STIs and the ML estimates of the sexual network variance. The analysis correctly accounts for possible sexual network random effects through casting the logistic mixed model as a *missing data* problem to facilitate the use of the EM algorithm. The chapter compares estimates from the model with and without random effects.

The multiplicative proportional hazards frailty model of time until HIV infection, where infection time is interval-censored and is treated as missing data, is formulated and analysed in Chapter 3 using the EM algorithm. Unobserved frailties and interval-censored infection times further form the missing data to facilitate the EM algorithm. Parameter estimates obtained from the model with and without sexual network frailty term are compared.

The major hindrance to full Bayesian implementation of many models has been the difficulty of evaluating the integrals to obtain posterior densities analytically.

Instead, methods have recently been developed to generate samples from the posteriors. In Chapter 4, we present the theory behind Markov chain Monte Carlo (MCMC) simulation techniques used in Bayesian analysis to generate these samples. The MCMC methods circumvent the complexities involved in dealing with intractable integrals.

Chapter 5 presents the results of the multiplicative frailty model obtained via the Gibbs sampler, a MCMC method. Furthermore, the results from the Gibbs sampler are compared to those obtained from the EM algorithm. Finally, Chapter 6 presents the substantive and methodological conclusions of the thesis. The chapter further highlights issues related to future research.

Chapter 2

Analysing STIs with the EM algorithm

2.1 Introduction

The expectation-maximization (EM) algorithm is a broadly applicable iterative technique for finding maximum likelihood (ML) estimates for parametric models. The ideas behind the EM algorithm appeared in various situations even before it was presented in a general formulation by Dempster, Laird and Rubin (1977), in which they showed its basic properties. The EM algorithm is profitably applied in situations of *incomplete-data* problems, where ML estimation is complicated by the data being *missing*. Data can be missing because of spoiled specimens, non-response from participants, or censored data (Cox, 1972). For example, in panel studies of HIV the exact infection time may occur between two widely spaced clinical examination times and thus it is interval-censored. The exact infection time is missing. In some cases, incompleteness of the data is not trivial. This is the case in random effects models, mixture models and latent variable structures where data is actually *unobservable*.

In this chapter, we will analyse migration data using the EM algorithm. The

outcome of interest is the infection with at least one curable STI. Syndromic management of diagnosed STIs was completed according to standard provincial guidelines (Department of Health, 1995). Therefore, a new infection is considered an incident case, not a carryover from a previous infection. The model formulated in Section 1.5 is extended to incorporate an extra index for clinical examination visit times. Section 2.2 presents the basic theory behind the EM algorithm. Section 2.3 discusses the rate of convergence and possible techniques for improving convergence. Methods of obtaining the information matrix are presented in Section 2.4. Section 2.5 describes parameter estimation of the model and derive equations required to implement the E-step and M-step of the algorithm. Finally, analyses of STIs data is presented in Section 2.6.

2.2 Theory of the EM algorithm

The basic idea behind the EM algorithm is to relate a given incomplete-data problem with *complete-data* problem in which the ML estimates are more tractable. In this framework, the complete-data x is perceived as an augmented form of the observed (incomplete) data y such that the distribution of y can be obtained from that of x as a marginal distribution. Let $f_c(x;\theta)$ and $f_o(y;\theta)$ denote the probability density function of the complete-data x and observed data y respectively, and θ is a vector of unknown parameters. The complete-data is related to the incomplete data through

$$f_o(y;\theta) = \int f_c(x;\theta)dx$$

where the integral is taken over $\{x : y = h(x)\}$ for some known function $h(.)$. The integral is replaced by the summation in discrete data.

Instead of directly maximizing the observed data log likelihood function $\log L_o(\theta) = \log f_o(y;\theta)$, the EM algorithm proceeds iteratively by updating the current estimate of θ by alternating between the E-step and M-step as follows.

- E-step: Compute the expected value of the complete-data log likelihood function given the observed data y and the current estimate θ.

$$Q(\theta;\theta^{(r)}) = E_{\theta^{(r)}}\left[\log f_c\left((x;\theta)|y,\theta^{(r)}\right)\right]$$

  where $\theta^{(r)}$ is the current estimate of θ after r^{th} iteration. In the exponential family case this is achieved by using the conditional expectation of the complete-data sufficient statistics given y and $\theta^{(r)}$. In other cases, the conditional expectations are not available in closed form, and they are computed using numerical methods such as Laplace approximation (Steele, 1996, Skaung, 2002), numerical quadratures (Anderson and Aitkin, 1985; Im and Gianola, 1988) and adaptive Gaussian quadrature (Monahan and Stefanski, 1992). McCulloch (1997) used the Metropolis algorithm that does not require specification of the observed data likelihood.

- M-step: Compute $\theta^{(r+1)}$ as the θ that maximizes the complete-data log likelihood $Q(\theta;\theta^{(r)})$ after replacing unobserved data with their conditional expectations obtained from the E-step. In general, this step will be fairly easy to compute since in most cases it coincides with complete data ML specifications. In most practical problems, estimation will be iterative in nature.

It is a general result from EM methodology that if $\theta^{(r+1)}$ maximized $Q(\theta;\theta^{(r)})$ then $\log L_o(\theta^{(r+1)}) \geq \log L_o(\theta^{(r)})$. That is, $\theta^{(r+1)}$ is better than $\theta^{(r)}$ (Dempster, Laird and Rubin, 1977). Jorgensen (2002) noted that this EM property is a consequence of Kullback-Leibler divergence properties. The E-step and M-step are alternated until the difference between consecutive values $\log L_o(\theta^{(r)})$ and $\log L_o(\theta^{(r+1)})$ becomes smaller than the pre-specified value ϵ. The EM algorithm is guaranteed to converge to at least a local maximum of $\log L_o(\theta)$. However, convergence to a global maxima in the presence of multiple maxima is not guaranteed. This is also the case with other algorithms, including Newton-type algorithms. McLachlan and Krishnan (1997) devoted the whole monograph to the EM algorithm and discussed extensions to the algorithm.

2.3 Rate of EM convergence

The EM algorithm is an attractive tool due to its simplicity. However, it can be extremely slow to converge compared to other methods such as Newton-Raphson. The rate of convergence of the EM algorithm is linear rather than the quadratic rate of convergence achieved in the Newton-Raphson algorithm. The EM algorithm implicitly defines a mapping function M such that

$$\theta^{(r+1)} = M(\theta^{(r)}).$$

If $\theta^{(r)}$ converges to some point θ^* and $M(\theta^{(r)})$ is continuous, then θ^* is a fixed point of the algorithm and $\theta^* = M(\theta^*)$. Using the first term of the Taylor series expansion of $\theta^{(r+1)} = M(\theta^{(r)})$ about the point $\theta^{(r)} = \theta^*$, we have that in the neighbourhood of θ^*

$$\theta^{(r+1)} - \theta^* = J(\theta^*)(\theta^{(r)} - \theta^*)$$

where $J(\theta^*)$ is a matrix of partial derivatives (Jacobian), $J(\theta) = \partial M(\theta)/\partial\theta$ evaluated at $\theta = \theta^*$ (Laird, *et al* 1987). In the neighbourhood of θ^* the algorithm is a linear iteration with convergence rate $J(\theta^*)$. They show that for large enough r, $J^* \approx J^{(\infty)}$ where $J^{(\infty)} = J(\theta^{(\infty)})$ and $\theta^{(\infty)} = \lim_{r\to\infty}\theta^*$. Thus, we can write

$$\theta^{(\infty)} \approx \theta^{(r)} + \left\{\sum_{h=0}^{\infty}\left[J(\theta^{(\infty)})\right]^h\right\}(\theta^{t+1} - \theta^t).$$

The power series $\sum_{h=0}^{\infty}\{J(\theta^{(\infty)})\}^h$ converges to $(I - J^{(\infty)})^{-1}$ if all eigenvalues of $J(\theta^{(\infty)})$ are between 0 and 1. Therefore, convergence can be improved by trying

$$\theta^{(\infty)} \approx \theta^{(r)} + \{I - J^{(\infty)}\}^{-1}(\theta^{(r+1)} - \theta^{(r)})$$

where I is a $p \times p$ identity matrix and θ is a $p-$dimensional vector of unknown parameters. The rate matrix $J^{(\infty)}$ can be estimated by J^*. In ML estimation, explicit formulae for J^* can be obtained by directly differentiating the mapping function. The rate of convergence matrix can also be expressed in terms of information matrices of the *pseudo-complete* data and unobserved data as detailed in the following section.

2.4 Observed information matrix

The common criticism of the EM algorithm is that it does not explicitly provide an estimate of the variance-covariance matrix while other methods such as the Newton-Raphson algorithm do. Methods of obtaining the variance-covariance matrix have been suggested by several authors (Louis, 1982; Meng and Rubin, 1991; Jamshidian and Jennrich, 2000). McLachlan and Krishnan (1997, chapter 4) give a survey of earlier work on the calculation of information matrix in the context of the EM algorithm. Proposed methods have their basis on the decomposition of the observed information matrix. Let $S_c(\theta; x)$ and $S_o(\theta; y)$ be the score vectors of the complete-data x and observed data y respectively. Also, let $I_c(\theta; x)$ and $I_o(\theta; y)$ be the negative square matrices of first partial derivatives of $S_c(\theta; x)$ and $S_o(\theta; y)$ respectively. Let $x = (y, z)$ where z represents the missing part of the complete-data x. The conditional density function of z given y is

$$g(z|y;\theta) = \frac{f_c(x;\theta)}{L_o(\theta;y)}. \tag{2.1}$$

Since $\log L_o(\theta; y) = \log L_c(\theta; x) - \log g(z|y; \theta)$ we then have that

$$\begin{aligned} -\frac{\partial^2 \log L_o(\theta;y)}{\partial\theta\partial\theta'} &= -\frac{\partial^2 \log L_c(\theta;x)}{\partial\theta\partial\theta'} - -\frac{\partial^2 \log g(z|y;\theta)}{\partial\theta\partial\theta'} \\ I_o(\theta;y) &= I_c(\theta;x) - I_{z|y}(\theta;x) \\ I_o(\theta;y) &= \mathcal{I}_c(\theta;y) - \mathcal{I}_{z|y}(\theta;y) \end{aligned} \tag{2.2}$$

where $\mathcal{I}_c(\theta; y)$ is the conditional expectation of the complete-data information matrix. Although $\log g(z|y; \theta)$ cannot be considered to be a log likelihood it seems sensible by analogy to define

$$\begin{aligned} I_{z|y}(\theta;x) &= -\frac{\partial^2 \log g(z|y;\theta)}{\partial\theta\partial\theta'} \\ \mathcal{I}_{z|y}(\theta;y) &= E\left[I_{z|y}(\theta;x)|\theta, y\right] \end{aligned}$$

where $\mathcal{I}_{z|y}(\theta; y)$ is the expected information matrix of the conditional distribution of unobserved data z given the observed data y. In the class of exponential family, $\log f_c(x; \theta) = t(x)\, q(\theta) + \log\, c(\theta) + \log\, h(x)$. Therefore, the elements of $\mathcal{I}_c(\theta; y)$ can

be expressed in terms of the complete-data sufficient statistics $t(x)$ and second partial derivatives of $q(\theta)$ and $c(\theta)$. Consequently, $\mathcal{I}_c(\theta; y)$ can be computed from the conditional expectations of the complete-data sufficient statistics. These conditional expectations are in fact required for implementation of the EM algorithm.

The decomposition in (2.2) is the application of the *missing information principle* as a consequence of observing only y and not z, meaning the observed information is equal to the complete-data information less missing information. Efron and Hinkley (1978) argued in favour of the observed data information $I_o(\theta; y)$ as a more appropriate measure of information than its expectation $\mathcal{I}(\theta) = E[I_o(\theta; y)]$ over y. Louis (1982) used the definition of $S_o(\theta; y)$ to show that $\mathcal{I}_{z|y}(\theta; y)$ can be expressed as

$$\begin{aligned} \mathcal{I}_{z|y}(\theta; y) &= \text{Cov}\{S_c(\theta; x)|y\} \\ &= E\left[S_c(\theta; x)S_c'(\theta; x)|y\right] - S_o(\theta; y)S_o'(\theta; y). \end{aligned} \tag{2.3}$$

where $S_o(\theta; y) = E\left[S_c(\theta; x)|y\right]$ (Louis, 1982; McLachlan and Krishnan, 1997). Therefore $I_o(\theta; y)$ becomes

$$\begin{aligned} I_o(\theta; y) &= \mathcal{I}_c(\theta; y) - \mathcal{I}_{z|y}(\theta; y) \\ &= \mathcal{I}_c(\theta; y) - \text{Cov}\{S_c(\theta; x)|y\} \\ &= \mathcal{I}_c(\theta; y) - E\left[S_c(\theta; x)S_c'(\theta; x)|y\right] + S_o(\theta; y)S_o'(\theta; y). \end{aligned} \tag{2.4}$$

The conditional expectations in (2.4) can be computed in the EM algorithm using $S_c(\theta; x)$ and $\mathcal{I}_c(\theta; y)$ which are the gradient and the curvature of the complete-data problem introduced within the EM algorithm. The conditional expectation needs only be evaluated at the last step of the EM algorithm where $\theta = \theta^*$ is the ML estimate. Thus,

$$I_o(\theta; y) = [\mathcal{I}_c(\theta; y)]_{\theta=\theta^*} - E\left[S_c(\theta; x)S_c'(\theta; x)|y\right]_{\theta=\theta^*}.$$

The last term on the right hand side of (2.4) disappears since $S_o(\theta^*; y) = 0$. The estimates of the variance-covariance matrix obtained by the EM algorithm are based

on the second derivative of the log likelihood and thus guaranteed to be valid asymptotically.

The methods of computing the variance-covariance matrix discussed thus far have their limitations. For example, the method proposed by Louis (1982) requires, in addition to the code for the complete-data variance-covariance matrix and the code for E-step and M-step, calculation of the conditional expectation of the square of the complete-data score function. Calculating such a conditional expectation can be cumbersome. Smith (1977) proposed calculating the asymptotic variance using the rate of convergence and the complete-data asymptotic variance. However, Smith's (1977) approach is inadequate in multi-parameter problems. The method cannot produce the entire rate matrix. Observed component-wise rate of convergence provides only a few eigenvalues. In most cases, it only yields the largest eigenvalue of the rate matrix. Meng and Rubin (1991) extended Smith's (1977) approach to multi-parameter problems. Theorem 4 of Dempster, Laird and Rubin (1977) proves that the convergence rate matrix $J(\theta^*)$ is

$$J(\theta^*) = \mathcal{I}_c^{-1}(\theta^*; y)\mathcal{I}_{z|y}(\theta^*; y).$$

An intuitive interpretation of this equation is that if more information is missing from the complete-data, the slower the convergence. The consequence of this result is that we can express $J(\theta^*)$ in terms of information matrix in (2.2) (McLachlan and Krishnan, 1997) as follows

$$\begin{aligned} J(\theta^*) &= \mathcal{I}_c^{-1}(\theta^*; y)[\mathcal{I}_c(\theta^*; y) - I(\theta^*; y)] \\ &= I - \mathcal{I}_c^{-1}(\theta^*; y)I(\theta^*; y) \end{aligned}$$

Interchanging terms and inverting matrices yields $I^{-1}(\theta^*; y) = \mathcal{I}_c^{-1}(\theta^*; y)[I - J(\theta^*)]^{-1}$. Inverted information matrices are used as estimates of the variance-covariance matrix of θ. Therefore, the observed asymptotic variance can be obtained by inflating the ordinary complete-data asymptotic variance with a factor of $[I - J(\theta^*)]^{-1}$. The factor $[I - J(\theta^*)]^{-1}$ is readily available from the output of the EM algorithm (Meng

and Rubin, 1991) without extra computational code outside the EM context. This only requires a supplemented step used to compute $[I - J(\theta^*)]^{-1}$ and evaluation of $\mathcal{I}_c^{-1}(\theta^*; y)$, hence Meng and Rubin, (1991) gave the name Supplemented EM (SEM). Jamshidian and Jennrich (2000) considered methods similar to SEM. However, their methods are based on numerical differentiation of the mapping function $M(\theta)$ and the observed data score vector.

2.5 Likelihood for the logit model

Much attention to random effects models has been given to problems where the conditional distribution of the response variable is normal, and the marginal distribution of random effects is normal (Harville, 1977; Laird and Ware, 1982; Verbeke and Molenberghs, 1997; 2000). The rationale for normal random effects models carries over to a wide range of probability distributions (Stiratelli, Laird and Ware, 1984; Anderson and Aitkin, 1985). If the distribution of the threshold level T_{ijk} is a standard logistic model as we discussed in Section 1.5, then

$$\pi_{ijk} = g(\beta' X_{ijk} + b_i) = \frac{\exp(\beta' X_{ijk} + b_i)}{1 + \exp(\beta' X_{ijk} + b_i)} \quad k = 1, \cdots, K_{ij} \tag{2.5}$$

where K_{ij} is the total number of clinical examination visits for the jth member of sexual network i and b_i is the random effect for network i. The model for π_{ijk} gives a *logit* link defined as $\log[\pi_{ijk}/(1 - \pi_{ijk})]$. The logistic mixed model has been considered before. Im and Gianola (1988) used a logistic mixed model to analyse lamb mortality with a two-way nested random effects of dam within sire. A similar approach was considered earlier by Anderson and Aitkin (1985) in the study of interviewer effect within areas.

The random effect variable $b_i \sim \mathcal{N}(0, \sigma^2)$. For convenience we consider $s_i = b_i/\sigma$, and thus $s_i \sim \mathcal{N}(0, 1)$, probability density function of a normal distribution, denoted by $\phi(s_i)$. The approach often adopted is to estimate the value of sexual network random effects (s_i) corresponding to ith sexual network response vector

$\mathbf{y}_i = \{y_{i11}, \cdots, y_{i1K_{ij}}, y_{i21}, \cdots, y_{iJ_iK_{ij}}\}$ by $E[s_i|\mathbf{y}_i;\theta]$ and obtain the ML estimates of θ, where $\theta = \{\beta, \sigma\}$. The predicted values of s_i are important in their own right. They provide valuable information about sexual network formations that are associated with high risk of STIs. The elements of $\mathbf{y}_i$ are assumed conditionally independent given s_i. Therefore, under the logit model the complete-data likelihood for sexual network i is given by:

$$L_c(\theta;\mathbf{y}_i,s_i) = \phi(s_i)\left(\prod_{j=1}^{J_i}\prod_{k=1}^{K_{ij}}\pi_{ijk}^{y_{ijk}}(1-\pi_{ijk})^{1-y_{ijk}}\right)$$

and the marginal log likelihood is given by

$$l(\theta) = \sum_{i=1}^{I}\log\left\{\int_{-\infty}^{\infty}\phi(s_i)\left(\prod_{j=1}^{J_i}\prod_{k=1}^{K_{ij}}\pi_{ijk}^{y_{ijk}}(1-\pi_{ijk})^{1-y_{ijk}}\right)ds_i\right\}. \quad (2.6)$$

The central function of the EM algorithm is the complete-data log likelihood given by

$$l_c(\theta) = \sum_{i=1}^{I}\left[\log\phi(s_i) + \sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} y_{ijk}\log\pi_{ijk} + (1-y_{ijk})\log(1-\pi_{ijk})\right]. \quad (2.7)$$

This function depends on unobserved data s_i through π_{ijk} (2.5). To estimate functions of unobserved data, we require their conditional distribution given the observed data and current estimate of θ, apart from a term not depending on θ. The following section describes the calculation for the required conditional expectations.

2.5.1 Expectation step

It is noted from (2.7) that we require $E[\log(\pi_{ijk})|\mathbf{y}_i;\theta]$ and $E[\log(1-\pi_{ijk})|\mathbf{y}_i;\theta]$. Thus, unobserved quantities will be replaced by their conditional expectations. The conditional density $g(s_i|\mathbf{y}_i;\theta)$ for sexual network i is given by

$$g_i(s_i|\mathbf{y}_i;\theta) = \frac{f(\mathbf{y}_i,s_i;\theta)}{L_i(\theta;\mathbf{y}_i)}$$

where

$$f(\mathbf{y}_i,s_i;\theta) = \phi(s_i)\prod_{j=1}^{J_i}\prod_{k=1}^{K_{ij}}\pi_{ijk}^{y_{ijk}}(1-\pi_{ijk})^{1-y_{ijk}}$$

and the $\mathbf{y}_i$ component of the observed data likelihood of θ is

$$L_i(\theta;\mathbf{y}_i) = \int_{-\infty}^{\infty} \phi(s_i) \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \pi_{ijk}^{y_{ijk}} (1-\pi_{ijk})^{1-y_{ijk}} ds_i. \tag{2.8}$$

It is not possible to solve $L_i(\theta;\mathbf{y}_i)$ analytically. Methods of numerical integration are required. Since the integration is over normal densities, Gaussian-Hermite quadratures can be used. Gaussian-Hermite quadratures replace the integral by a summation of the weighted integrand function evaluated at optimal quadrature points. Therefore,

$$\int f(a)\phi(a)da \approx \sum_{h=1}^{M} w_h f(a_h),$$

where $\{a_h : h = 1,\cdots,M\}$ denote M-optimal Gaussian quadrature points and $\{w_h : h = 1,\cdots,M\}$ the corresponding weights. The terms $w_h\sqrt{\pi}$ and $a_h/\sqrt{2}$ are given in Abramowitz and Stegun (1972). Other methods such as automatic differentiation have been used to facilitate ML estimation (Skaung, 2002). Applying Gaussian-Hermite quadratures to (2.8) gives us

$$L_i(\theta;\mathbf{y}_i) \approx \sum_{h=1}^{M} w_h \left(\prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \pi_{ijkh}^{y_{ijk}} (1-\pi_{ijkh})^{1-y_{ijk}} \right)$$

where $\pi_{ijkh} = g(\beta' X_{ijkh} + \sigma a_h)$. The $E[\log(\pi_{ijk})|\mathbf{y}_i;\theta]$ and $E[\log(1-\pi_{ijk})|\mathbf{y}_i;\theta]$ given $\mathbf{y}_i$ are

$$\begin{aligned} E[\log(\pi_{ijk})|\mathbf{y}_i;\theta] &= \int_{-\infty}^{\infty} \log(\pi_{ijk})\; g_i\,(s_i|\mathbf{y}_i;\theta) ds_i \\ &\approx \frac{\sum_{h=1}^{M} w_h \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \log(\pi_{ijkh})\, \pi_{ijkh}^{y_{ijk}} (1-\pi_{ijkh})^{1-y_{ijk}}}{\sum_{h=1}^{M} w_h \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \pi_{ijkh}^{y_{ijk}} (1-\pi_{ijkh})^{1-y_{ijk}}} \end{aligned}$$

and

$$\begin{aligned} E[\log(1-\pi_i)|\mathbf{y}_i;\theta] &= \int_{-\infty}^{\infty} \log(1-\pi_{ijk})\, g_i(s_i|\mathbf{y}_i;\theta)\, ds_i \\ &\approx \frac{\sum_{h=1}^{M} w_h \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \log(1-\pi_{ijkh})\, \pi_{ijkh}^{y_{ijk}} (1-\pi_{ijkh})^{1-y_{ijk}}}{\sum_{h=1}^{M} w_h \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \pi_{ijkh}^{y_{ijk}} (1-\pi_{ijkh})^{1-y_{ijk}}}, \end{aligned}$$

respectively. The $E[s_i|\mathbf{y}_i;\theta]$ is calculated in a similar fashion.

2.5.2 Maximization step

The M-step proceeds by replacing functions of unobserved data with their conditional expectations obtained from the E-step into the complete-data log likeli-

hood (2.7). Let the vector of weights $\mathbf{w}_{hi}$ be

$$\mathbf{w}_{hi} = \frac{w_h \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \pi_{ijkh}^{y_{ijk}} (1-\pi_{ijkh})^{1-y_{ijk}}}{\sum_{h=1}^{M} w_h \prod_{j=1}^{J_i} \prod_{k=1}^{K_{ij}} \pi_{ijkh}^{y_{ijk}} (1-\pi_{ijkh})^{1-y_{ijk}}}.$$

Therefore, the M-step maximizes

$$\begin{aligned} Q(\theta;\theta^{(r)}) &= \sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} y_{ijk}\left(\sum_{h=1}^{M} \mathbf{w}_{hi}\log(\pi_{ijkh})\right) + (1-y_{ijk})\left(\sum_{h=1}^{M}\mathbf{w}_{hi}\log(1-\pi_{ijkh})\right) \\ &= \sum_{h=1}^{M}\sum_{i=1}^{I}\mathbf{w}_{hi}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} y_{ijk}\log(\pi_{ijkh}) + (1-y_{ijk})\log(1-\pi_{ijkh}) \end{aligned} \tag{2.9}$$

with respect to θ. Functions of the data not involving θ are ignored in (2.9). The ML estimators are the solutions to

$$\sum_{h=1}^{M}\sum_{i=1}^{I}\mathbf{w}_{hi}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} x_{ijkhs}\left\{y_{ijk} - \frac{\exp(\beta' X_{ijkh} + \sigma a_h)}{1+\exp(\beta' X_{ijkh} + \sigma a_h)}\right\} = 0, \qquad s = 1,\cdots,p \tag{2.10}$$

and

$$\sum_{h=1}^{M}\sum_{i=1}^{I}\mathbf{w}_{hi}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}} a_h\left\{y_{ijk} - \frac{\exp(\beta' X_{ijkh} + \sigma a_h)}{1+\exp(\beta' X_{ijkh} + \sigma a_h)}\right\} = 0 \tag{2.11}$$

where p is the total number of fixed effects. It turns out that these are the equations used in weighted logistic regression estimation where the weight of $\mathbf{y}_i$ is $\mathbf{w}_{hi}$. The weights $\mathbf{w}_{hi}$ depend on current parameter estimate $\theta^{(r)}$, which is updated at each iteration. The equations are similar to those obtained when differentiating (2.6) with the integral approximated by numerical quadratures. This then suggests estimating parameters using an iterative technique such as iterative re-weighted least squares (IRLS) used for estimation in GLMs. The IRLS will require the second derivatives of (2.10) and (2.11). In this estimation approach, σ is considered a regression coefficient and estimated by concatenating σ to β and a_h into the design matrix. If the design matrix is $N = \sum K_{ij}$ long, then the new design matrix is $N \times M$ long. Therefore, M should be as small as practical for computational ease (Bock and Aitkin, 1981; Brillinger and Preisler, 1983; Im and Gianola, 1988). The effect of M is only noticeable when the variance component is large, but the results do not change much for large M (Brillinger and Preisler, 1983; Anderson and Aitkin,

1985). The advantage of this approach is that any available software that can fit IRLS can be used. In the context of GLMMs several other numerical techniques for maximizing likelihood functions have been suggested. These include, among others, Monte Carlo EM-algorithm (McCulloch, 1994; Booth and Hobert, 1999) and Monte Carlo Newton-Raphson (McCulloch, 1997 and Kuk and Cheng, 1997).

2.6 Application to the data

In this chapter, the outcome variable is considered to be the presence or absence of at least one curable STI in an individual at each of the first five examination visits (including baseline). The first five clinical examination visits constitute most of the data due to high subject dropout rate. Infection with HIV is irreversible whilst other STIs are curable and recurrent, so HIV infection is not part of the outcome. Genital sores and genital discharge could merely be markers of acute symptoms of syphilis and gonorrhoea respectively, therefore confusing the disease effects with its symptoms in the derivation of the outcome. However, such manifestations will not affect the outcome since the outcome indicates presence of an infection or acute symptom due to an infection. The outcome would have been affected had counts of multiple correlated events within an individual been used.

The available data for this analysis is from 628 individuals. Individuals constituting a sexual network range from 1 to 5 with varying number of visits attended. The followup rate was slightly better among migrant men, Table 2.1. Each sexual network had one man. The data consisted of 189 couples, 39 triads, 5 quadriads and 1 pentad. The data exhibit a higher rate of individual dropout than sexual network dropout. Major reasons for dropout were that the subject was lost to follow-up or refused to continue to cooperate with interviewers. There is no apparent consistent pattern of change in the risk of STIs by migration status over time, Table 2.2. However, migrant men are seemingly at higher risk of infection than their female

Table 2.1: The distribution by migration status at each visit

Visit	Migrants	Non-migrants	Migrants' partners.	Non-migrant's partners	**Total**
1	223	122	139	144	631
2	131	68	97	104	408
3	62	29	66	52	212
4	32	19	39	25	115
5	19	13	22	20	75

partners. Female partners of migrant men are, in 4 out of 5 visits, at higher risk of

Table 2.2: Prevalence of STIs at each clinical examination visit

	Migrant networks (%)[a]		Non-migrant networks(%)		TOTAL	
Visit	**Male**	Female	**Male**	Female	N	Percent
1	**33.04**	23.91	**21.31**	21.53	628	26.11
2	**20.00**	12.00	**14.81**	11.11	439	15.03
3	**9.41**	8.96	**10.34**	15.09	234	10.68
4	**12.00**	16.28	**15.79**	8.33	136	13.24
5	**18.52**	13.64	**0.00**	5.00	81	11.11

[a]Defined depending on whether a man is a migrant or non-migrant

infection than partners of non-migrant men. However, partners of non-migrant men are, in 3 out of 5 visits, at higher risk than their male partners. Twenty-two people were diagnosed with the same STI in two consecutive visits, with 4.5% co-infected with both active syphilis and gonorrhoea. Out of 5 people diagnosed with the same STI at a subsequent visit, 3 were diagnosed with active syphilis and 2 with gonorrhoea. None was diagnosed with chlamydia. The results are in agreement with the assumption that a new STI is an incident case.

Preliminary analysis identified *migration status*, *age at recruitment*, *marital status*, *age at first sexual intercourse*, *recent sexual contact partners*, *HIV status* and

number of visits as important determinants of STIs. We first fitted an ordinary logistic regression model ignoring sexual network induced correlation in the data. The results are presented in Table 2.3. The model estimate of the constant is the log odds of being infected with an STI for someone in reference categories of all other variables. Exponentiation of -1.984 gives odds ratio (OR)=0.134 of being infected with an STI. Applying inverse link yields the probability (0.118) of being infected with an STI in that category. Interpretation of other variables is made in relation to a reference category and keeping all other variables fixed. Reference categories are constrained to unit odds. Migrant men and their partners are consistently at higher risk of STIs compared to partners of non-migrant men with OR=1.542 [95%CI: 1.001 - 2.373] and OR=1.196 [95%CI: 0.762 - 1.877] respectively. Non-migrant men were at lesser risk of STIs than their partners, but this was not significant. The risk of infection was significantly higher among migrant sexual networks than non-migrant sexual networks (p-value=0.031). However, the risk of STI did not differ between males and females (p-value=0.652).

Being aged younger than 35 years was associated with increased risk of STIs. But the risk effect of age diminished when migration status was considered. The risk factors considered are somehow interrelated. Individuals who have never been married were at higher risk of infection with an STI, OR=1.442 compared to those who were currently or had been married.

Earlier commencement (16 years or younger) of sexual activity increased the likelihood of infection with an STI, OR=1.426, p-value=0.023. The risk of infection was higher among those reporting recent sexual contact with one partner but not significantly different from those who reported no sexual contact. The risk was significantly higher among those reporting recent sexual contact with at least two partners, OR=2.342 [95%CI: 1.236 - 4.437], compared to those who reported no sexual contact. Larger number of lifetime partners increased the risk of STIs.

Table 2.3: Results for the standard logistic regression model

Parameter	Estimate	Standard Error	Z-statistic
Constant	-1.984	0.404	-4.906
Migration status			
Migrant men	0.433	0.220	1.968
Partners of migrant men	0.179	0.230	0.776
Non-migrant men	-0.096	0.259	-0.371
Partners of non-migrant men[a]			
Age less than 35 years			
(0=no, 1=yes)	0.153	0.175	0.875
Never married			
(0=no, 1=yes)	0.366	0.178	2.054
Age first sexual contact			
16 years or younger	0.355	0.156	2.275
More than 16 years[a]			
Recent sexual contact partners			
None[a]			
One	0.352	0.298	1.178
Two or more	0.851	0.326	2.608
HIV status			
(0=negative, 1=positive)	0.422	0.168	2.518
Visit numbers			
Linear	-0.293	0.072	-4.083

[a]Reference category

However, its effects were completely diminished when the number of recent sexual contact partners was included in the model.

Presence of HIV was associated with a significantly increased risk of contracting an STI (OR=1.525). For every clinical visit attended, the odds of having at least one STI were reduced by OR=0.746 [95%CI: 0.646 - 0.859]. The study design makes it more sensible to include the time factor in the model which makes interpretation of the results time related. The reduction in the risk of STIs showed the importance of continuous treatment of STIs, sexual behavioural education and appropriate health seeking behaviour.

These variables (Table 2.3) were analysed using the EM algorithm to take into account the sexual network correlation. The analysis was implemented in S-plus 2000. However, since S-plus is not optimized for iterative loops, intensive iterative statements were carried out in Microsoft Visual C++ version 6.0 and integrated into S-plus by creating a Dynamic Link Library. The fixed effects estimates from the standard logistic regression model were used as initial estimates in the EM algorithm. The initial estimate of the random effect variance was set to 1. The sampling nodes in the estimation of random effects were modified accordingly to ensure that sampling of the integrand is in a suitable range of values (Liu and Pierce, 1994). Fewer number of quadrature points were considered. In similar models, Anderson and Aitkin (1985) reported that five quadrature points suffice. Thus, only six quadrature points were used in fitting the sexual network random effects model.

The results of the EM analysis are presented in Table 2.4. The results indicate that unobserved sexual network random effects have sizeable impact on the risk of contracting STIs. All estimates of fixed effects, except HIV status, were magnified. The corresponding standard errors were also inflated as a consequence of including random effects in the model. This has important implications for behavioural

Table 2.4: The EM parameter estimates of STI data from migrant and non-migrant sexual networks

Parameter	Estimate	Standard Error	Z statistic
Constant	-2.968	0.448	-6.631
Migration status			
Migrant men	0.554	0.244	2.276
Partners of migrant men	0.071	0.252	0.281
Non-migrant men	-0.103	0.281	-0.365
Partners of non-migrant men[a]			
Age less than 35 years			
(0=no, 1=yes)	0.187	0.193	0.971
Never married			
(0=no, 1=yes)	0.498	0.198	2.522
Age first sexual contact			
16 years or younger	0.483	0.171	2.818
More than 16 years[a]			
Recent sexual contact partners			
None[a]			
One	0.407	0.322	1.264
Two or more	1.055	0.357	2.956
HIV status			
(0=negative, 1=positive)	0.397	0.187	2.131
Visit numbers			
Linear	-0.363	0.079	-4.605
Random effect variance[b]			
Sexual network	1.457	0.111	13.090

[a]Reference category

[b]Z test is equivalent to a one-sided test with a critical value of 1.645 at 5% significance level

epidemiological studies since random effects of sexual networks are rarely taken into account in studies of HIV/STIs. The increased standard errors indicate extra variability taken into account in the model. However, most variables still reached statistical significance level (p-value$\leq$0.05) so the main fixed effects inferences remained unchanged.

The notable change in the risk of other STIs is the effect of HIV status. In the standard logistic model, the risk of an STI was OR=1.525 times higher among those infected with HIV and highly significant (p-value=0.012). The risk estimate of HIV in the EM analysis (Table 2.4) was reduced by about six percent. Transmission of HIV in a sexual network is high if at least one partner is infected. Therefore, inclusion of sexual network effect which accounts for unmeasurable common sexual network behaviour is likely to reduce the magnitude of the effect of HIV status. However, HIV status remained significant (p-value=0.033). This indicates that the reduced immune response due to HIV infection increased the risk of being infected by a STI.

The estimate of sexual network variance is 1.457, which is considerably greater than zero at 5% level of significance based on the one-sided $\mathcal{Z}$-test with critical value of 1.645. The estimate indicates a high degree of heterogeneity between sexual networks, Figure 2.1. The estimated sexual network variance implies a substantial degree of association between members of the same sexual network with respect to the risk of STIs even after adjusting for other individual specific covariates in the model. The corresponding estimated correlation is 0.592, which is quite substantial.

2.7 Conclusion

The results show that circular migration contributes significantly towards transmission dynamics of STIs. Contacts between migrant men and highly sexually active

women during migration bring the epidemic of HIV/STIs into local regions, through their less sexually active rural partners (Lurie, *et al* 1997). However, an ongoing local epidemic of HIV/STIs in rural areas is also responsible for sustaining the epidemic of HIV/STIs (Lurie, Williams, Zuma, *et al* 2003a). Efficient transport system between urban and rural areas provide adequate conditions for the formation of linear sexual network components which are an important reservoir for maintenance of STIs, particularly gonorrhoea (Wylie and Jolly, 2001).

The risk factors considered are interdependent, thus making it difficult to isolate the effect of each factor. Age and marital status, for example, are interrelated in that older people are more likely to be married than younger people. Migrant sexual networks are more likely to be younger than non-migrant sexual networks as migrant men are often at their working age. Certain variables commonly considered in such analysis, for example education and income, were excluded either because of collinearity between variables such as income, job status and migration status or because there was little variation in the data for any valid statistical analysis. For example, all migrant men were recruited at their workplaces, and therefore all employed whilst few non-migrant men were employed.

Sexual mixing between older men and younger women has potentials of introducing infections among younger women (Gregson, *et al* 2002). In the model, age was not a significant risk factor for STIs after adjusting for other factors. However, it was kept in the model due to its epidemiological importance. The risk of STIs increases exponentially with an increase in the number of recent sexual contact partners. This indicates the intensity of STIs in the presence of multiple partners as confirmed by stochastic simulation models (Morris and Kretzschmar, 1997). The risk associated with the number of recent sexual contact partners increased in the analysis that corrects for sexual network correlation in the data. This shows the standard logistic model underestimated the effect of multiple partners in the trans-

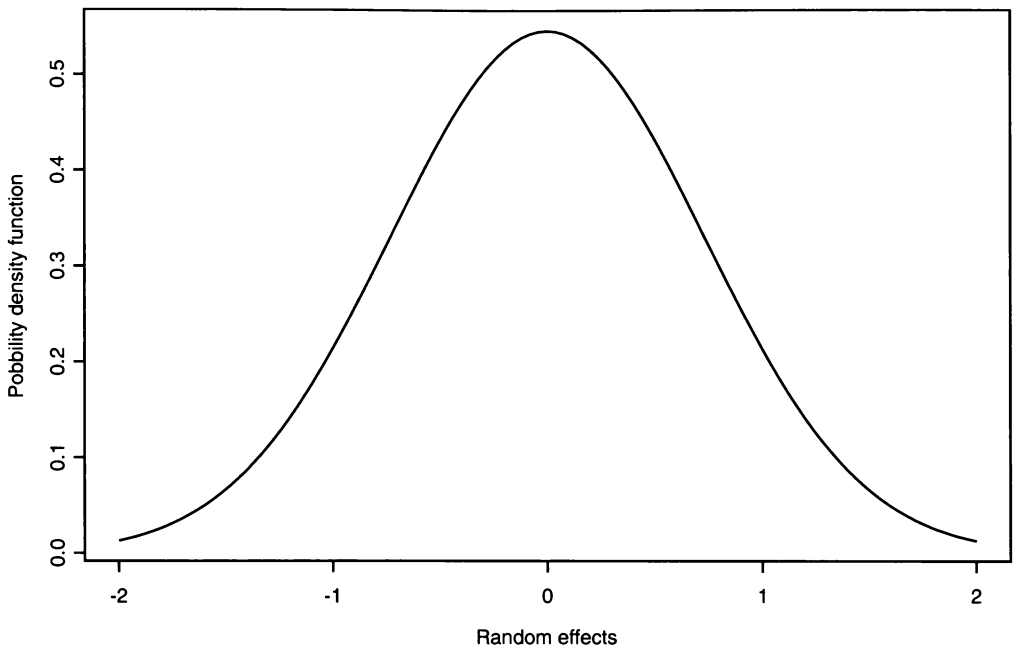


Figure 2.1: *Distribution of sexual network random effects*

mission dynamics of STIs.

The transmission of HIV in this society is mainly heterosexual (O'Farrell, *et al* 1991). Therefore its presence is a good indicator of history of risky behaviour. Inclusion of unobserved random effects variable which accounts for sexual network specific behaviour is likely to reduce the magnitude of HIV effects. However, importance of HIV status was not completely removed. The remaining importance could be that the reduced immune system response caused by HIV infection increases the likelihood of acquiring an STI during unprotected sexual contact with an infected partner. The most unfavourable scenario is that people infected with HIV continue to engage in unprotected sexual contact even though efforts are made to educate them about the importance of safe sex.

Heterogeneity across sexual networks has important implications for transmis-

sion dynamics of HIV/STIs (Ghani, *et al* 1997). The main source of heterogeneity in HIV/STIs studies arises from patterns of sexual mixing and structural composition of sexual networks. Migrant men frequent prostitutes or have sexual contacts with sexually active women during their migration periods (Hunt, 1989; Jochelson, *et al* 1991; Lurie, 2000). The risky behaviour of migrant men is followed by sexual contacts with their less sexually active rural partners. Their rural partners may have other short-term relationships while their partners are away (Lurie, *et al* 1997). If this mixing pattern occurs, a multiply peaked epidemic of HIV/STIs may occur and the epidemic will spread rapidly through the small proportion of sexually active men and women but more slowly and larger through the less sexually active group of men and women (Anderson, *et al* 1991). The less sexually active networks constitute the majority of the population.

There are two possible scenarios leading to the positive correlation within sexual networks. Firstly, one member of a sexual network could get an infection from some other non-regular sexual partner(s) and pass the infection onto the regular sexual partners in a sexual network. Secondly, whilst the partners are away from each other, each can get infected through different outside sexual contacts. These links between networks act as a co-transmitter sexual network. Transmission of an infection is much more efficient in the present network structures since men act as links to all their partners within a sexual network. Therefore, epidemiological implications are much more severe than in sexual networks that are linear in structure (Morris and Kretzschmar, 1997).

The estimates obtained in both the standard logistic and the logistic with sexual network random effects models are subject to some bias. We obtained each man's consent prior to contacting his rural female sexual partner(s), which probably reduced the response rate and contributed to biased participation of rural female partners. In studies involving partner tracing, it is often partners with whom the re-

lationship is more stable that are divulged to the interviewers (Johnson, *et al* 2003). Furthermore, it is possible that not all women's extra relationships were divulged to interviewers. Societal values play an important role in the formation of sexual networks and their structure. In this society, extra marital relationships among women are not as widely accepted as they are among men. Women may underreport their sexual activity whilst men may overstate theirs. The sampling method used started from a random sample of people thus sexual behavioural aspects are expected to be the same as in the true sexual network. However, the refusal to participate or to identify sexual partners might lead to underestimation of sexual mixing. Individuals who were infected and sought treatment for STIs elsewhere between consecutive clinical visits were not counted. Therefore, the risk of STIs might even be higher than observed.

In the estimation phase, Gaussian-Hermite quadratures were used to approximate intractable integrals (Abramowitz and Stegun, 1972). The final results are based on six quadrature points. Brillinger and Preisler (1983) reported that results do not change much for quadrature points greater than eight. Moreover, even five quadrature points have been considered sufficient (Bock and Aitkin, 1981; Anderson and Aitkin, 1985). The inherent disadvantage of using large number of quadrature points is the need for strong assumption of normally distributed random effects. For example, ten quadrature points fit a symmetric distribution thus forcing the tails of random effects distribution to be normal. Methods similar to this have been used in the analysis of binomial data with two levels of nested random effects (Anderson and Aitkin, 1985; Im and Gianola, 1988).

The estimation approach accounted for sexual network heterogeneity in the data by iteratively calculating functions of unobserved data and estimating the fixed effects through weighted logistic using IRLS. Ignoring unobserved sexual network random effects in the analysis leads to spurious associations between the risk of STIs

and some covariates due to underestimation of the standard errors. In the analysis, the standard errors of the model that ignores correlation were underestimated by at most 11% compared to the model that corrected for correlation. Therefore, treating each individual as independent gives false impression that there is more information in the data than there really is. The advantage of using the EM algorithm is the assurance of convergence to (at least a local) ML estimate and the possibility of including more than one random effects term in the model. For example, in community mass treatment of STIs, a random term corresponding to distinct communities, and for distinct sexual networks within communities, could easily be incorporated into the model.

The results have important implications for the control of STIs. Control measures of STIs should extend further from focussing on high-risk individuals to considering high-risk sexual networks. This is more imperative for women who are in weak positions to negotiate safe sex or prevent their partners from having extra relationships. Partner notification should be encouraged and facilitated in order to stop the continuing transmission cycle of STIs and further transmission of HIV within sexual networks. Interventions targeted at local communities will attain only short-term success in the presence of urban-rural migration which creates opportunities for re-entry of STIs. Therefore, interventions should fully incorporate the effects of migration and sexual network structures in their approaches. The results of this chapter have been submitted for publication (Zuma, *et al* 2004).

Chapter 3

Analysing time until HIV infection using the EM algorithm

3.1 Introduction

Epidemiological studies of disease incidence seek to relate the risk of contracting a disease to a set of measurable risk factors. In reality, some important risk factors are neither collected nor measurable. In many cases, unmeasured risk factors vary across sub-groups and thus inducing sub-group correlation. In studies of disease incidence, correlation arises because disease occurrence tends to cluster within families. The common statistical approach to familial data is treating the clustering variable as random effect. In studies of HIV/STIs, the risk of HIV infection within sexual networks depends on common sexual behaviour, susceptibility to an infection and high likelihood of HIV transmission if at least one partner is infected. In this study, a similar approach will be taken by introducing membership to a sexual network variable as a random effect in the investigation of risk factors associated with time until HIV infection.

The last several years has seen significant research regarding the inclusion of random effects in models of *failure time* data. A model becoming popular in modelling

correlated failure times is the *frailty model.* The frailty model is an extension of the Cox proportional hazards model (Cox, 1972). Frailty is the term describing common excess risk of infection among members of the same sub-group. The topic of frailty models has received considerable attention in demography (Vaupel, *et al* 1979) and statistics (Clayton, 1978; Clayton and Cuzick, 1985). Considerable research in statistical literature has focused on estimation techniques of frailty models. Estimation techniques include derivation of a full likelihood function that depends on both the observed data and the unobserved frailties (Klein, 1992; Guo and Rodriguez, 1992; Sastry, 1997).

The models discussed thus far assume that the exact failure time is either precisely known or it is right-censored. This is basically true when failure time is the time of death. For some events other than death, the failure time may not be precisely known but only the examination times to which the exact failure time lies. The resultant data is referred to as *interval-censored* data (see for example Rücker and Messerer, 1988). Interval-censored data assume that the event is *irreversible.* For instance, it is impossible to be cured from HIV infection. In panel studies of AIDS, HIV status is determined by performing laboratory analysis of blood samples at periodic examination times. As these tests are performed infrequently, diagnosis is delayed compared to disease onset (Jewell, *et al* 1994; Farrington and Gay, 1999). Interval-censored data require special statistical methods due to imprecise knowledge of event time.

Methods of analysing interval-censored data stem from the Cox model (Cox, 1972). Finkelstein (1986) generalized the Cox model to correctly account for interval-censored event time. Pan (2000) proposed an approach based on multiple imputation of failure times. Parametric methods for analysing interval-censored data are readily available (Lindsey and Ryan, 1998). A particular drawback of popular methods for analysing interval-censored data is their failure to reduce to standard survival

settings when data are not interval-censored. Goetghebeur and Ryan (2000) proposed an approximate likelihood function for interval-censored data that reduces to standard survival settings when data are right-censored. Huang and Wellner (1997) provide a rigorous theoretical account for methods of analysing interval-censored data. Methods discussed thus far further assume that failure times are independent. The thesis of this chapter is to analyse dependent interval-censored time until HIV infection, until the end of the study or until the subject was lost to follow-up. This dependency is modelled as frailties. The frailties and interval-censored infection times form the missing data for the application of the EM algorithm (Dempster, Laird and Rubin, 1977).

The chapter presents an approach for estimating the parameters when the data is correlated and interval-censored. Section 3.2 presents data notation and related work. The model under consideration is formulated in Section 3.3. Section 3.4 describes the sexual network frailty distribution. The details of parameter estimation are presented in Section 3.5. Computation and inference are presented in Section 3.6. Finally, in Section 3.7 baseline descriptive statistics are presented and HIV data from cohorts of migrant and non-migrant sexual networks from a rural health district of South Africa are analysed using the model formulated in Section 3.3.

3.2 Data notation and related work

The HIV data to be analysed in this chapter is clustered within sexual networks. Sexual networks are considered distinct sub-groups connected by sexual relationships. Let X_{ij} denote a vector of covariates associated with the jth member ($j = 1, \cdots, J_i$) of the ith sexual network ($i = 1, \cdots, I$) and β be a vector of coefficients representing covariate effects. Available data for each person consist of ordered clinical visitation (examination) times $\{0 < v_{ij,1} < v_{ij,2} < \cdots < v_{ij,n_{ij}} < \infty\}$ and corresponding binary indicators $\{\delta_{ij,1}, \delta_{ij,2}, \cdots, \delta_{ij,n_{ij}}\}$ of HIV status. The exact infection time t_{ij}

is only known to be before $v_{ij,1}$ or between $\{v_{ij,k}, v_{ij,k+1}\}$ or after $v_{ij,n_{ij}}$ if the person remained uninfected at the last examination time.

Irreversibility of HIV infection means that HIV test results will be negative at all examination times before the first positive test result and will be positive at all subsequent examination times. It is therefore sufficient to record only the examination time interval $v_{ij} = \{v_{ij,k}; v_{ij,k+1}\}$ encompassing the transition time. Examination time $v_{ij,k} = \max\{v_{ij,k}|\delta_{ij,k} = 0\}$ and $v_{ij,k+1} = \min\{v_{ij,k}|\delta_{ij,k} = 1\}$. The width of v_{ij} possibly varies between individuals. Therefore, techniques for multivariate grouped survival data (Guo and Lin, 1994) are inappropriate. Let $y_{ij} = t_{ij} \in (v_{ij,k}; v_{ij,k+1}]$ if infection occurred and $y_{ij} = v_{ij,k}$ if right-censored. Note that for interval-censored observations y_{ij} is unobserved. Instead, we only know clinical examination endpoints such that v_{ij} contains the observed data. Define a non-censoring indicator $\delta_{ij} = 1$ if infected with HIV and 0 otherwise. The ith sexual network specific frailty is denoted by b_i.

The attempt to include frailties in the interval-censored data likelihoods of Finkelstein (1986) or Huang and Wellner (1997) inevitably results in rather complex intractable likelihood functions. The gamma frailty distribution, which is conjugate to the standard proportional hazards likelihood, is not conjugate to the interval-censored data likelihood. Therefore, implementing the EM algorithm becomes laborious and prohibitive in terms of computing efforts. The alternative approach is to ignore correlation in the data and utilize standard univariate techniques for interval-censored data (Finkelstein, 1986; Huang and Wellner, 1997). However, naive standard errors obtained through this approach can lead to invalid inference (Guo and Lin, 1994; Wei, *et al* 1989). A nonparametric ML estimator for discrete bivariate interval-censored data using techniques for convex optimization has been proposed (Betensky and Finkelstein, 1999). But, this technique does not account for covariate effects. Kim and Xue (2002) proposed a marginal proportional haz-

ards model for discrete survival times previously utilised by Finkelstein (1986). The model is an extension of the marginal approach used by Wei *et al* (1989) and Guo and Lin (1994) for the analysis of correlated continuous survival times and multivariate grouped survival data respectively. Farrington and Gay (1999) proposed a Laplace approximate method that combines individual frailty and interval-censored survival data using empirical Bayes estimates of individual frailty effect. The method underestimates standard errors and is only recommended as an exploratory tool.

In the univariate context, the common *fix-up* approach is to assume that the event occurred at the beginning or end of each examination time to facilitate use of the Cox proportional hazards model. Estimates calculated from these two extreme assumptions roughly enclose the estimates from interval-censored data methods (Finkelstein, 1986; Lindsey and Ryan, 1998). However, making assumptions about infection time can lead to biased estimates of regression parameters for both univariate (Rücker and Messerer, 1988; Odell, *et al* 1992) and multivariate interval-censored data (Kim and Xue, 2002). The bias is severe when data are heavily censored (Lindsey and Ryan, 1998). In particular, the method tends to underestimate standard errors and result in wrong inference.

In this work, we consider an approach where both the exact infection time for interval-censored observations and frailties form the missing data to facilitate the EM algorithm. We then use techniques for correlated survival data which iteratively augment common unobserved sexual partnership frailties (Guo and Rodriguez, 1992; Klein, 1992; Sastry, 1997) to estimate parameters. For a review of methods for analysing correlated standard survival data, see for example Lin (1994) and Kelly and Lim (2000).

3.3 Model formulation

The Cox proportional hazards model (Cox, 1972) is assumed. The model has extensively been used to handle right-censored data. The Cox model assumes that the hazard function $\lambda(y_{ij}|X_{ij})$ is related to the baseline hazard function $\lambda_0(y_{ij})$ as follows:

$$\lambda(y_{ij}|X_{ij}) = \lambda_0(y_{ij})e^{\beta' X_{ij}}$$

and $\Lambda_0(y_{ij}) = \int_0^{y_{ij}} \lambda_0(y_{ij})dy_{ij}$ is the corresponding integrated baseline hazard. The integrated fixed effects hazard is

$$\Lambda(y_{ij}|X_{ij}) = \int_0^{y_{ij}} \lambda_0(y_{ij})e^{\beta' X_{ij}} dy_{ij}.$$

Conditional on sexual network frailties, survival times within a sexual network are assumed mutually independent and their conditional marginal distributions have a hazard function $h(y_{ij}|b_i, X_{ij})$. The hazard function satisfies the multiplicative frailty model

$$h(y_{ij}|b_i, X_{ij}) = b_i\lambda_0(y_{ij})e^{\beta' X_{ij}}.$$

In this formulation, the frailties operate multiplicatively on baseline hazards. If baseline hazard for someone with frailty 1 is $\lambda_0(y_{ij})$, then baseline hazard for someone with frailty b_i is $b_i\lambda_0(y_{ij})$. Frailties are interpreted as relative risks (RR)s. The quantity $\exp(\beta' X_{ij})$ is the RR associated with covariate X_{ij}. Sexual networks with only one member included are permitted. In that case, individuals are affected by their own frailty. Members of the same sexual network are indistinguishable, except for values of X_{ij}, so they have a common baseline hazard rate. The associated integrated hazards are

$$\begin{aligned} H(y_{ij}|b_i, X_{ij}) &= \int h(y_{ij}|b_i, X_{ij})dy_{ij} \\ &= \int b_i\lambda_0(y_{ij})e^{\beta' X_{ij}} dy_{ij} \\ &= b_i\Lambda(y_{ij}|X_{ij}). \end{aligned}$$

The conditional survival function (3.1)

$$S(y_{ij}|b_i, X_{ij}) = \exp\{-H(y_{ij}|b_i, X_{ij})\} \quad (3.1)$$

is the probability of infection time being greater than y_{ij}. The likelihood contribution for someone infected with HIV is the conditional density (3.2) given by

$$f(y_{ij}|b_i, X_{ij}) = S(y_{ij}|b_i, X_{ij}) \times h(y_{ij}|b_i, X_{ij}). \quad (3.2)$$

But, someone uninfected with HIV only contributes the conditional survival function.

A parametric form for the baseline hazards is assumed. Baseline hazards are assumed constant, $\lambda_0(y_{ij}) = \lambda_0$. The corresponding cumulative baseline hazards are $\Lambda_0(y_{ij}) = \lambda_0 y_{ij}$. Therefore, survival times come from exponential distributions. The constant hazards assumption is plausible in the analysis of HIV incidence. The risk of HIV infection likely varies substantially between individuals, depending on individual behaviour for which covariate information is not available. However, modelling behavioural risk factors and unobserved frailties absorb individual variability. Consequently, the remaining risk of HIV infection will be less variable. Furthermore, there are indications that the national HIV prevalence has reached a mature stage (Department of Health, 2001). Constant hazards have previously been assumed in HIV incidence studies (Farrington and Gay, 1999). Farrington and Gay (1999) assumed constant hazards in the model investigating association between HIV incidence and frequency of examination among homosexual men.

3.4 Sexual network frailty

Sexual network frailties are unobservable. The strategy is to adopt some distributional assumptions about frailties. Either parametric or nonparametric frailty distribution can be assumed (Guo and Rodriguez, 1992). Frailties are assumed to be mutually independent random variables. The multiplicative frailty model takes

only positive frailty values. Therefore, we assume a gamma distribution with shape and scale parameters α and α^{-1} respectively. The density function of b_i is

$$f(b_i) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} b_i^{\alpha-1} e^{-\alpha b_i}. \tag{3.3}$$

Hence, the RR for sexual networks has mean 1 and variance $1/\alpha$.

The choice of a gamma frailty distribution arises partly because of its mathematical convenience, flexibility and nonnegativity. The gamma distribution is conjugate to the likelihood. This conjugacy property immensely simplifies mathematical computations. Several authors in both the frequentist (Klein 1992; Sastry, 1997) and the Bayesian literature (Clayton, 1991; Manda, 1998) have used gamma frailty. Other frailty distributions that are of interest and also flexible include log-gamma and log-normal distributions (Duchateau, *et al* 2002; Palmgren and Ripatti, 2002).

The current debatable issue is the effect of using a particular frailty distribution. In the study of factors affecting duration of unemployment, Heckman and Singer (1984) conducted a sensitivity analysis comparing the effects of choosing different frailty distributions on the estimates of covariate parameters. They found many changes in signs and absolute magnitude of parameters. The authors argued in favour of nonparametric frailty distributions. Hougaard (1986) pointed out that the assumption of positive stable frailty distribution preserves the proportionality of hazards to the marginal distributions, which is invalid in the presence of frailties. Estimated parameters are biased towards zero when the frailty with finite mean is ignored (Schumacher, *et al* 1987). Therefore, including frailty would effectively increase the magnitude of covariate coefficients depending on the proximity of the assumed frailty distribution to the true frailty distribution (Pickles and Crouchley, 1995). Guo and Rodriguez (1992) showed that estimates do not markedly differ whether nonparametric or parametric frailty is assumed. Simulation studies also suggest that the choice of a particular frailty distribution is not critical in estimating regression parameters (Pickles and Crouchley, 1995; Sastry, 1997). Therefore, the

fixed effect inferences would not be expected to change by much had we assumed a different frailty distribution.

3.5 Parameter estimation

The theory of the EM algorithm presented in Chapter 2 is implemented in survival data when both the frailties and exact infection time are unobserved. ML estimates of $\theta = \{\alpha, \lambda_0, \beta\}$ cannot be calculated straight away without an estimate of functions of unobserved exact infection times and frailties. The response vector $\mathbf{y}_i$ of sexual network i consists of (possibly sub-vectors) of known clinic visitation times v_i and unobserved true infection time $\mathbf{t}_i$. Using conditional independence between the elements of $\mathbf{y}_i$ given b_i the complete-data likelihood function for sexual network i is

$$\begin{aligned} & L_i(\theta; b_i, v_i, \mathbf{t}_i) \\ = \; & \frac{\alpha^\alpha}{\Gamma(\alpha)} b_i^{\alpha-1} e^{-\alpha b_i} \\ \times \; & \prod_{j=1}^{J_i} \left(e^{-b_i \Lambda_0(t_{ij})\, e^{\beta' X_{ij}}} \; b_i \, \lambda_0 \, e^{\beta' X_{ij}} \right)^{\delta_{ij}} \left(e^{-b_i \Lambda_0(v_{ij,k})\, e^{\beta' X_{ij}}} \right)^{1-\delta_{ij}} . \end{aligned} \tag{3.4}$$

If t_{ij} was known and the frailties b_i observed but variable, one would fit model (3.4) maximising the complete-data log-likelihood given by

$$\begin{aligned} l_i(\theta) \; = \; & \alpha \log \alpha - \log \Gamma(\alpha) + (\alpha - 1) \log b_i - \alpha b_i \\ + \; & \sum_{j=1}^{J_i} \delta_{ij} \left[-b_i \, \lambda_0 \, t_{ij} \, e^{\beta' X_{ij}} + \log(b_i \, \lambda_0) + \beta' X_{ij} \right] \\ - \; & (1 - \delta_{ij}) \, b_i \, \lambda_0 \, v_{ij,k} \, e^{\beta' X_{ij}} . \end{aligned} \tag{3.5}$$

The complete-data log-likelihood (3.5) could be maximized over all sexual networks. However, $\mathbf{t}_i$ and b_i are unobserved. Functions of unobserved data need to be estimated from the data.

Let $f_i(b_i, v_i; \theta)$ be the joint marginal distribution of b_i and v_i given by

$$
\begin{aligned}
& f_i(b_i, v_i; \theta) \\
= & \int_{v_{i\delta_{i+},k}}^{v_{i\delta_{i+},k+1}} \cdots \int_{v_{i2,k}}^{v_{i2,k+1}} \int_{v_{i1,k}}^{v_{i1,k+1}} L_i(\theta; b_i, v_i, \mathbf{t}_i)\, dt_{i1}\, dt_{i2} \cdots dt_{i\delta_{i+}} \\
= & \frac{\alpha^\alpha}{\Gamma(\alpha)} b_i^{\alpha-1} e^{-b_i\left(\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\,\Lambda(v_{ij,k}|X_{ij})\right)} \prod_{j=1}^{\delta_{i+}} \left(e^{-b_i\Lambda(v_{ij,k}|X_{ij})} - e^{-b_i\Lambda(v_{ij,k+1}|X_{ij})}\right)
\end{aligned}
$$

where $\delta_{i+} = \sum_{i=1}^{J_i} \delta_{ij}$ is the total number of HIV infected members of a particular sexual network. In order to facilitate the EM algorithm, conditional expectations of functions of unobserved data given observed data and current parameter estimates need to be computed. The next sections describe computation of these conditional expectations and maximization of parameters.

3.5.1 Expectation step

The E-step of the algorithm computes the conditional expectation of (3.5) which mainly involves computation of the conditional expectations of functions of unobserved data. It can be seen from (3.5) that we need only the conditional expectations of b_i, $\log b_i$ and $b_i t_{ij}$. Conditional expectations of b_i and $\log b_i$ require the marginal conditional distribution of b_i given the observed data. The marginal conditional distribution $g_i(b_i|v_i; \theta)$ can be computed from $f_i(b_i, v_i; \theta)$ as

$$
g_i(b_i|v_i; \theta) = \frac{f_i(b_i, v_i; \theta)}{L_i(\theta; v_i)}
$$

where

$$
\begin{aligned}
& L_i(\theta; v_i) \\
= & \int_0^\infty f_i(b_i, v_i; \theta)\, db_i \\
= & \int_0^\infty \frac{\alpha^\alpha}{\Gamma(\alpha)} b_i^{\alpha-1} e^{-b_i\left(\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\,\Lambda(v_{ij,k}|X_{ij})\right)} \prod_{j=1}^{\delta_{i+}} (e^{-b_i\Lambda(v_{ij,k}|X_{ij})} - e^{-b_i\Lambda(v_{ij,k+1}|X_{ij})})\, db_i
\end{aligned}
$$

is the observed data likelihood.

Therefore, the conditional expectation of b_i is

$$
\begin{aligned}
& E[b_i|v_i;\theta] \\
= & \int_0^\infty b_i \, g_i(b_i|v_i;\theta) db_i \\
= & \frac{\int_0^\infty b_i^\alpha e^{-b_i(\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij}))} \prod_{j=1}^{\delta_{i+}} (e^{-b_i\Lambda(v_{ij,k}|X_{ij})} - e^{-b_i\Lambda(v_{ij,k+1}|X_{ij})}) \, db_i}{\int_0^\infty b_i^{\alpha-1} e^{-b_i(\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij}))} \prod_{j=1}^{\delta_{i+}} (e^{-b_i\Lambda(v_{ij,k}|X_{ij})} - e^{-b_i\Lambda(v_{ij,k+1}|X_{ij})}) \, db_i}.
\end{aligned}
$$

The integrands expand depending on the total number δ_{i+} of those infected with HIV in a sexual network which is known. Therefore, each integrand mimics a gamma density and thus the conditional expectation simplifies. If $\delta_{i+} = 0$ then $E[b_i|v_i;\theta]$ is the expected value of a gamma distribution with parameters α and $\alpha + \sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})$. If $\delta_{i+} = 1$ then

$$
\begin{aligned}
& E[b_i|v_i;\theta] \\
= & \frac{\int_0^\infty b_i^\alpha e^{-b_i[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})]} \times (e^{-b_i\Lambda(v_{ij,k}|X_{ij})} - e^{-b_i\Lambda(v_{ij,k+1}|X_{ij})}) \, db_i}{\int_0^\infty b_i^{\alpha-1} e^{-b_i[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})]} \times (e^{-b_i\Lambda(v_{ij,k}|X_{ij})} - e^{-b_i\Lambda(v_{ij,k+1}|X_{ij})}) \, db_i}
\end{aligned}
$$

$$
\begin{aligned}
= & \{\int_0^\infty b_i^\alpha e^{-b_i[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})+\delta_{ij}\Lambda(v_{ij,k}|X_{ij})]} db_i \\
- & \int_0^\infty b_i^\alpha e^{-b_i[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})+\delta_{ij}\Lambda(v_{ij,k+1}|X_{ij})]} db_i\} \\
\div & \{\int_0^\infty b_i^{\alpha-1} e^{-b_i[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})+\delta_{ij}\Lambda(v_{ij,k}|X_{ij})]} db_i \\
- & \int_0^\infty b_i^{\alpha-1} e^{-b_i[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})+\delta_{ij}\Lambda(v_{ij,k+1}|X_{ij})]} db_i\}
\end{aligned}
$$

$$
\begin{aligned}
= & \{\frac{\Gamma(\alpha+1)}{[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})+\delta_{ij}\Lambda(v_{ij,k}|X_{ij})]^{\alpha+1}} \\
- & \frac{\Gamma(\alpha+1)}{[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})+\delta_{ij}\Lambda(v_{ij,k+1}|X_{ij})]^{\alpha+1}}\} \\
\div & \{\frac{\Gamma(\alpha)}{[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})+\delta_{ij}\Lambda(v_{ij,k}|X_{ij})]^{\alpha}} \\
- & \frac{\Gamma(\alpha)}{[\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})+\delta_{ij}\Lambda(v_{ij,k+1}|X_{ij})]^{\alpha}}\}.
\end{aligned}
$$

Similarly, the $E[b_i|v_i;\theta]$ for $\delta_{i+} > 1$ can be computed.

The conditional expectation of $\log b_i$ is

$$E[\log b_i|v_i;\theta] \;=\; \int_0^\infty \log b_i \; g_i\,(b_i|v_i;\theta)\; db_i$$

and can be approximated numerically. For the product of b_i and particular $t_{ij'}$, we need to integrate out the remaining sexual network unobserved infection times t_{ij}, for $j \neq j'$ from $L_i(\theta; b_i, v_i)$ to obtain the joint marginal distribution of b_i, $t_{ij'}$ and v_i. From the joint marginal distribution $f_{ij'}(b_i, t_{ij'}, v_i)$ we compute the conditional distribution $f_{ij'}(b_i, t_{ij'}|v_i;\theta) = f_i(b_i|v_i;\theta)\, f_{ij'}(t_{ij'}|b_i, v_i;\theta)$. The conditional distribution

$$f_{ij'}(t_{ij'}|b_i, v_i;\theta) \;=\; \frac{e^{-b_i\Lambda(t_{ij'}|X_{ij'})}\; b_i\lambda_0 e^{\beta' X_{ij'}}}{S(v_{ij',k}|b_i;\theta) - S(v_{ij',k+1}|b_i;\theta)}.$$

The conditional expectation $E[b_i t_{ij'}|v_{ij',k} < t_{ij'} \leq v_{ij',k+1};\theta]$ is derived as

$$\begin{aligned} & E[b_i t_{ij'}|v_{ij',k} < t_{ij'} \leq v_{ij',k+1};\theta] \\ = \; & \int_0^\infty b_i\, f_i(b_i|v_i;\theta) \left\{ \frac{\int_{v_{ij',k}}^{v_{ij',k+1}} t_{ij'}\, f_{ij'}(t_{ij'}|b_i, v_i;\theta)\; dt_{ij'}}{\int_{v_{ij',k}}^{v_{ij',k+1}} f_{ij'}(t_{ij'}|b_i, v_i;\theta)\; dt_{ij'}} \right\} db_i. \end{aligned}$$

The integral over b_i expands depending on the number of infected members (δ_{i+}) of a sexual partnership.

Denote $E[b_i|v_i;\theta]$, $E[\log b_i|v_i;\theta]$ and $E[b_i t_{ij}|v_{ij,k} < t_{ij} \leq v_{ij,k+1};\theta]$ by $\overline{b_i}$, $\overline{\log b_i}$ and $\overline{b_i t_{ij}}$ respectively. The quantities $\overline{b_i}$, $\overline{\log b_i}$ and $\overline{b_i t_{ij}}$ are evaluated at current estimate of θ.

3.5.2 Maximization step

The M-step concerns finding ML estimate $\hat{\theta} = \{\hat{\alpha}, \hat{\lambda}_0, \hat{\beta}\}$ after replacing functions of unobserved data in (3.5) by their conditional expectations $\overline{b_i}$, $\overline{\log b_i}$ and $\overline{b_i t_{ij}}$. The consequent complete-data log-likelihood function is denoted by $Q(\theta;\theta^{(r)})$ where $\theta^{(r)}$ is the current estimate of θ. Maximisation process of $Q(\theta;\theta^{(r)})$ separates itself into two distinct parts, one involving a one dimensional parameter α, say $l_1(\alpha)$, and the other involving a one dimensional parameter λ_0 and multi-dimensional parameter

β, say $l_2(\lambda_0, \beta)$.

There is no closed form expression for $\hat{\alpha}$. Maximization with respect to α can be accomplished by Newton-Raphson algorithm which requires the following first and second derivatives:

$$\frac{\partial Q(\theta;\theta^{(r)})}{\partial \alpha} = \sum_{i=1}^{I}(\overline{\log b_i} - \overline{b_i}) + \sum_{i=1}^{I}\left[1 + \log\alpha - \frac{d\log\Gamma(\alpha)}{d\alpha}\right]$$

$$\frac{\partial^2 Q(\theta;\theta^{(r)})}{\partial \alpha^2} = \sum_{i=1}^{I}\left[\frac{1}{\alpha} - \frac{d^2\log\Gamma(\alpha)}{d\alpha^2}\right]$$

where the first and second derivatives of $\log\Gamma(\alpha)$ are computed from the corresponding recursive formulae given in Abramowitz and Stegun (1972). The ML estimates for λ_0 and β are obtained in a similar fashion.

3.6 Computation and inference

Computation of $\{\hat{\lambda}_0, \hat{\beta}\}$ is carried out using the profile likelihood approach. This is achieved by maximising $Q(\theta;\theta^{(r)})$ over λ_0 for all values of β to obtain $\hat{\lambda}_0(.,\beta)$, thereafter maximise $Q(\alpha, \beta, \hat{\lambda}_0(.,\beta))$ over β to find $\hat{\beta}$. The estimate $\hat{\theta}$ is then used to compute $\overline{b_i}$, $\overline{\log b_i}$ and $\overline{b_i t_{ij}}$. The following algorithm details the computation of $\{\overline{b_i}, \overline{\log b_i}, \overline{b_i t_{ij}}\}$ and $\hat{\theta}$. Let $\theta^{(0)} = \{\alpha^{(0)}, \lambda_0^{(0)}, \beta^{(0)}\}$ be the initial parameter estimates and set $r = 1$.

- Step (i) Compute $\overline{b_i}$, $\overline{\log b_i}$ and $\overline{b_i t_{ij}}$ as described in Section 3.5.1 using $\theta^{(r-1)}$.
- Step (ii) Maximize $l_1(\alpha)$ with respect to α to get $\alpha^{(r)}$
- Step (iii) Maximize $l_2(\lambda_0, \beta^{(r-1)})$ with respect to λ_0 to get $\lambda_0^{(r)}$
- Step (iv) Maximize $l_2(\lambda_0^{(r)}, \beta)$ with respect to β to get $\beta^{(r)}$
- Step (v) Let $\theta^{(r)} = \{\alpha^{(r)}, \lambda_0^{(r)}, \beta^{(r)}\}$. Set r=r+1 and repeat steps (i) to (v) until the difference between consecutive values $\log L_o(\theta^{(r)})$ and $\log L_o(\theta^{(r+1)})$ becomes smaller than the prespecified value ϵ.

In principle, the values maximising $l_1(\alpha)$ and $l_2(\lambda_0, \beta)$ also maximise $Q(\theta; \theta^{(r)})$. The log likelihood $l_1(\alpha)$ is a concave function. Proposition 3.1 in Huang and Wellner (1997) states that for fixed β, $l_2(\lambda_0, \beta)$ is concave in λ_0, and for fixed λ_0, $l_2(\lambda_0, \beta)$ is concave in β. Therefore, maximization steps (ii) to (iv) concern maximising well defined concave functions.

The matrix $\partial^2 Q(\theta; \theta^{(r)})/\partial\theta\partial\theta'$ can be used to check if the estimate is a local maximum. The point estimate is a local maximum if $\partial^2 Q(\theta; \theta^{(r)})/\partial\theta\partial\theta'$ is negative definite. Huang (1996) and Murphy (1995) studied asymptotic properties of (λ_0, β) and (α, λ_0) respectively. Inference in these models is based on asymptotic normality of the parameter estimate around the true value. Murphy (1994) proved consistency of α and λ_0 in one-sample frailty model. The asymptotic variance is estimated by *observed information matrix* rather than its expectation. The latter requires knowledge of the probability distribution of censoring pattern. However, the observed information matrix only requires knowledge of censoring times. In frailty models estimated via EM algorithm, the observed information matrix underestimates variability. This is because the method ignores variability introduced by estimating b_i, $\log b_i$ and $b_i t_{ij}$. The SEM algorithm (Meng and Rubin, 1991) as briefly described in Chapter 2 was used to obtain adjusted standard errors.

3.7 Application to the data

3.7.1 Baseline description

The baseline HIV prevalence was 20.1%. The prevalence of HIV did not differ significantly between men (22.7%) and women (19.1%). Migrant men and their partners were (based on a χ^2 test) significantly at higher risk of HIV infection compared to non-migrant men and their partners, 24.0% and 15.0% respectively, p-value=0.02. The prevalence of HIV was significantly higher among migrant men (25.9%) than non-migrant men (12.7%) in all age-groups. The prevalence was also higher among

partners of migrant men (21.1%) compared to partners of non-migrant men (16.5%), but this was not statistically significant. The use of condoms was rare in both men (20%) and women (10%). Ever use of condoms was higher among women reporting many lifetime partners and men reporting many casual partners. The complete results of baseline description of the data with socio-demographic and other biomedical risk factors of HIV infection have been published in a paper by (Lurie, Williams, Zuma, *et al* 2003a). Migration is an important risk factor of HIV among migrant men and their partners in rural areas.

The study of self-identified migrant women and non-migrant women in Carletonville where some migrant men were recruited was subsidiary to the migration project, Section 1.3. Migration was identified as an important risk factors of HIV among these women. Self-identified migrant women were almost twice as likely to be infected with HIV compared to self identified non-migrant women (OR=1.61, 95%CI: 1.11 - 2.31). Most self-identified migrant women were from other rural areas of South Africa. The paper published by Zuma *et al* (2003) reports the risk factors of HIV infection among self-identified migrant and non-migrant women.

In the 168 couples recruited earlier into the study, 58.3% had a migrant male partner and 41.7% had a non-migrant male partner. The overall prevalence of HIV was 19.9% with infection significantly higher among men (24.4%) than women (15.5%), p-value=0.04. Neither partner was infected with HIV in 69.6% of the couples. Migrant couples were as likely as non-migrant couples to have neither partner infected with HIV (65.3% versus 75.7%; χ^2 test p-value=0.148). In 9.5% of the couples, both partners were infected with HIV, and this did not differ significantly by the migration status of the male partner. In 20.8% of the couples, only one partner was infected with HIV (HIV discordant). Migrant couples were 2.5 times more likely than non-migrant couples to be discordant (26.5% versus 12.8%, p-value=0.031). Of the 35 discordant couples, the man was infected in 25 (71%) of the cases and the

woman in the remaining 10 (29%) cases. The proportion of infected men in migrant discordant couples was essentially the same as in non-migrant discordant couples. The full results describing infection patterns in the couples and the associated risk factors of the presence of an infection in a couple have been published in a paper by (Lurie, Williams, Zuma, *et al* 2003b).

3.7.2 Main data analysis

The main focus in this chapter is in the analysis of the time since 1990 until HIV infection, until the end of the study or until the subject was lost to follow-up. Thus the primary outcome variable is the survival time after 1990. It is considered that the epidemic of HIV in South Africa became well established in 1990 (Gouws and Williams, 2000). Anyone infected prior to that time is likely to have died before the study commenced. Since the HIV rate has been high over the whole time since 1990 and fairly constant over the period of clinical examination, a constant baseline hazard seems reasonable. The most common mode of HIV transmission in this society is through heterosexual contacts. An exception to this is the pattern of infection found in the white population of South Africa where homosexual contacts accounted for 87% of infections between 1982 and 1990 (Zwi and Bachmayer, 1990). In the black population, the pattern of HIV infection is similar to the rest of sub-Saharan Africa. More than 75% of AIDS cases among black South Africans between 1982 and 1990 resulted from heterosexual transmission (O'Farrell and Windsor, 1991).

The current analysis is restricted to 339 identifiable distinct sexual networks from 604 individuals. The mean sexual network size is 1.78 individuals. Sexual network size ranges from 1 to 5 with only one man in each sexual network. Table 3.1 provides the distribution of sexual networks and percentage of persons infected with HIV. A considerable number of migrant men gave incorrect information about the location of their partners and some of their identified partners refused to participate. This led to a large number of sexual networks where only the man was included. Migrant

Table 3.1: Distribution of sexual networks and HIV infection

			HIV infection	
Sexual network size	Number of sexual networks	Percentage	N	Percentage
1	122	36.0	48	39.34
2	175	51.6	88	25.14
3	37	10.9	38	34.23
4	4	1.2	2	12.5
5	1	0.3	0	0.0

men contributed considerably to high HIV infection in networks where only a man was included. Fifty-two percent of sexual networks were couples. HIV infection was considerably higher in triads than in the couples. The number of sexual networks of size greater than three is too small to make valid comparisons. The maximum total number of HIV infected members per sexual partnership size was three, and was among triads. The overall mean years since first sexual intercourse is 20.6. The distributions of years of sexual activity are shown in Figure 3.1 for males and females respectively. The mean sexual activity age does not differ significantly between men and women. The mean age at first sexual intercourse was 18 and 17 years for men and women respectively. The mean number of lifetime partners was 15.8 and 2.0 for men and women respectively.

The analysis of the data using the EM algorithm was implemented in Microsoft Visual C++ 6.0 and S-plus 2000. S-plus subroutines were used for numerical approximations (in particular, TRIGAMMA and DIGAMMA). Table 3.2 presents descriptive statistics of variables considered in the analysis. The considerable imbalance between migrant men and their partners is due to large number of partners of migrant men who were not part of the study.

Two models were fitted to the data: a model that does not take into account

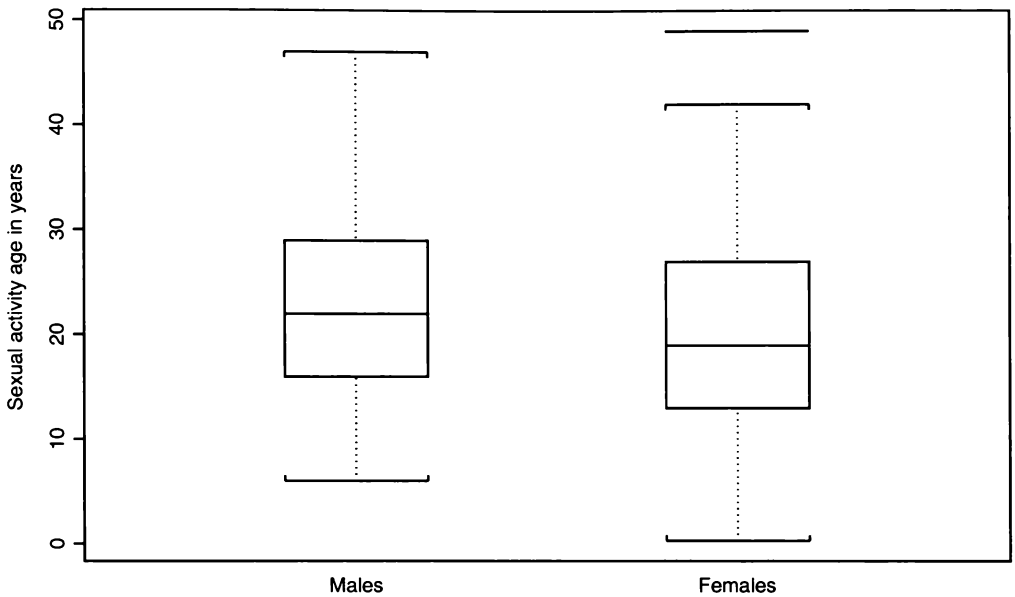


Figure 3.1: *Distribution of sexual activity age by gender*

dependency induced by sexual networks and one that takes into account the dependency through the introduction of a frailty term. The resulting parameter estimates are presented in Table 3.3. Estimated constant baseline hazard of the model without frailty was minimal. Being a migrant men or a partner of a migrant men is associated with increased risk of HIV infection compared to partners of non-migrant men, RR=1.170 and RR=1.167 respectively. Non-migrant men were at reduced risk of HIV infection compared to their partners (RR=0.812). The estimated survival times show that migrant men and their partners were similarly and higher risk of HIV than non-migrant men and their partners, Figure 3.2. Being in the age category 18 to 24 is associated with considerably higher risk of HIV compared to someone aged above 34 years (RR=2.804). The RR of HIV is 1.689 times higher for someone aged between 25 and 34 years compared to someone aged above 34 years. Someone who reported recent sexual contact with more than one sexual partner is almost twice as likely to be HIV positive compared to someone who reported recent sexual

Table 3.2: Descriptive statistics of variables used in HIV infection

Variable	Percent	Variable	Percent
Migration status		*Recent sexual partners*[a]	
Migrant men	31.6	More than one	20.2
Partners of migrant men	24.8	*Lifetime partners*	
Non-migrant men	18.6	More than one	67.8
Partners of non-migrant men	25.0		
Age in years		*Active syphilis*	
18 to 24	5.5	Positive	15.7
25 to 34	29.9	*Status of other STIs*	
35 or above	64.6	Positive	28.3

[a]Partners with sexual contact in the last four months

contact with one or no partner. Reporting more than one lifetime sexual partner was associated with an increased risk of HIV infection, RR=1.418. Infection with syphilis significantly increases the RR=1.548 of HIV. Also, an infection with other STIs is clearly important: their presence is associated with 1.623 times more risk of HIV infection.

The model with frailty term contains the effects of other covariates not specifically included in the model. The Akaike information criteria (AIC) for the mdel without sexual network frailty term and the model with sexual network frailty term is 1445.11 and 1377.61 respectively. The AIC leads us to conclude that the model with sexual network frailty term fits the data better than the model without sexual network frailty term.

The sexual network frailty parameter represents sexual network effect. The parameter is interpreted as the variance of the frailty distribution. Large values indicate greater heterogeneity between sexual networks and stronger association

Table 3.3: Parameter estimates obtained with the EM-algorithm

	Model without frailty		Model with frailty	
Parameter	Estimate	SE	Estimate	SE
Baseline hazard				
Constant	0.016	0.002	0.017	0.003
Migration status				
Migrant men	0.157	0.220	0.343	0.226
Partners of migrant men	0.156	0.204	0.227	0.208
Non-migrant men	-0.205	0.258	-0.132	0.265
Partners of non-migrant men[a]				
Age in years				
18 to 24	1.031	0.292	1.577	0.297
25 to 34	0.524	0.160	0.623	0.164
35 and above[a]				
Recent sexual contact partners				
Only one[a]				
More than one	0.611	0.192	0.498	0.199
Number of lifetime partners				
Only one[a]				
More than one	0.349	0.171	0.417	0.182
Syphilis				
0=Negative,1=Positive	0.437	0.157	0.387	0.167
Status of other STIs				
0=Negative,1=Positive	0.484	0.180	0.503	0.185
Frailty variance[b]				
Sexual network			0.462	0.054

[a]Reference category

[b]The Z test for testing frailty variance equal to zero at 5% significance level is equivalent to a one-sided test with a critical value of 1.645.

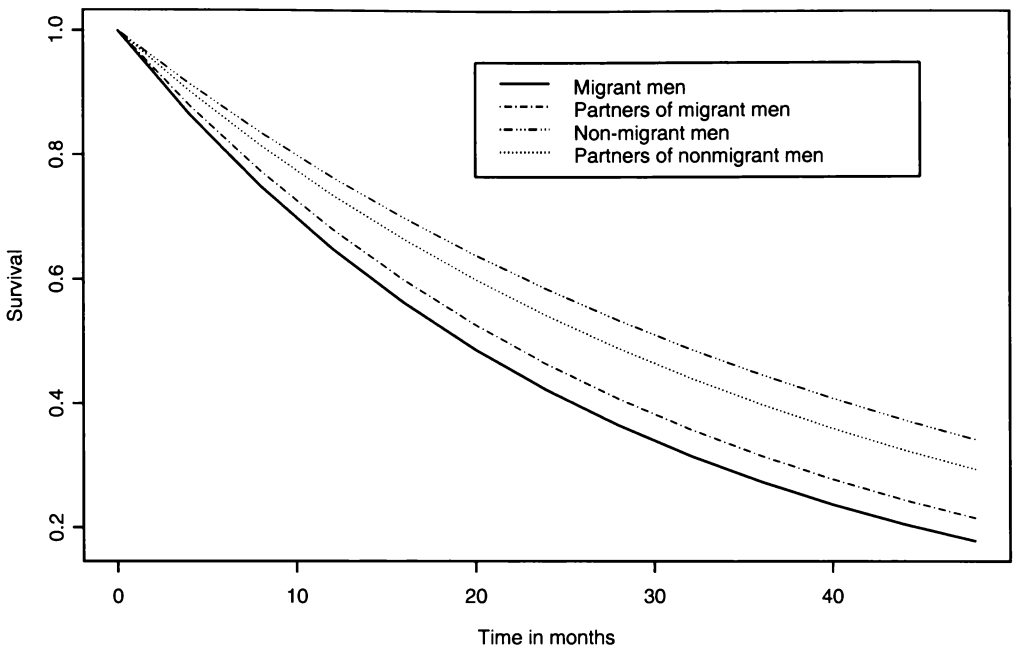


Figure 3.2: *The estimated survival times by migration status*

within sexual networks. The frailty variance is 0.462 with standard error 0.054 and is significantly different from zero. The frailty variance implies moderate degree of correlation (0.188) within sexual networks even after controlling for other covariates. Standard errors are slightly magnified in the model with frailties. This indicates that fixed effect parameters are now estimated with a more realistic, but lower precision.

The estimates of constant baseline hazards did not change very much between the model without frailty and the model with frailty. Results of the model with frailty indicate that unobserved sexual network frailty has considerable impact on the risk of HIV infection. The coefficient estimates of migration status, age at recruitment, number of lifetime partners and other STIs were slightly magnified in the model with frailty compared to the model without frailty. The RR associated with migrant men was considerably inflated. However, the RRs of recent sexual contact partners and syphilis infection were deflated. The RR for someone reporting

recent sexual contact with more than one partner was reduced by about 23%. But, the RR remained statistically significant. Reported number of recent sexual contact partners indicates the risk associated with a particular sexual network. Intuitively, someone from a large sexual network is frailer. Including sexual network random effects corrects for such proneness of particular sexual networks. Therefore, the magnitude of RR associated with recent sexual contact partners is reduced in the model with frailty compared to the model without frailty. This indicates that recent sexual contact partners were acting as a proxy for unobserved sexual network effect in the model without frailty. The effect is now partly captured by sexual network random effects. Similar effects are observed in the RRs of syphilis. This is critical since syphilis is often asymptomatic and left undiagnosed in the general population (Wilkinson, *et al* 1999).

3.8 Conclusion

The clustering of socio-demographic and sexual behavioural risk factors within sexual networks provides an opportunity to investigate factors contributing to the epidemics of HIV/STIs. The results of the model indicate that sexual networks contribute significantly in the spread of STIs, HIV in particular. Even after controlling for some important risk factors, the risk of HIV infection varies considerably across sexual networks. The importance of sexual network frailty term indicates that some sexual networks are at increased risk of HIV infection compared to others. Therefore, interventions should consider sexual networks as social units rather than focussing on individuals. Intervention strategies such as counselling, treatment of STIs and education messages specifically designed to deal with situations where sexual partners are discordant are urgently needed to protect uninfected partners who are at high risk of HIV infection due to their infected partners.

The observed patterns of HIV infection were unexpected but revealing in as

much as they shed light on the role of migration in the spread of HIV to the rural areas (Lurie, Williams, Zuma, *et al* 2003b). It has long been hypothesized that the primary direction of HIV transmission is from returning migrant men, who contract HIV while migrating, and return home to infect their rural partners (Pison, *et al* 1993; Decosas, *et al* 1995). If this was the case, one would expect the men to be the infected partner in most discordant couples. However, in nearly one-third of discordant couples the female was the infected partner. While this confirms the importance of migration as a risk factor for HIV infection in both men and women, it changes our understanding of the way in which migration enhances this risk. We have found that migration is a risk factor not simply because men return home to infect their rural partners, but also because their rural females - both those who are partners of migrant men and those who are partners of non-migrant men - are as likely to become infected from sexual contacts outside of their primary relationship. The fact that the patterns of HIV discordance are similar in non-migrant couples, with the woman being the infected partner in one-third of non-migrant discordant couples, indicates that some partners of non-migrant men become infected not through their husbands.

The data provided by respondents in this study about the age at first sexual contact and the number of lifetime partners were consistent with those of other South African studies (Department of Health, 1998; Williams, *et al* 2000; Eaton, *et al* 2003). A community-based survey in Carletonville (Williams, *et al* 2000) found the age at first sexual intercourse to be slightly younger (a year for girls, a year and a half for boys) than in the migration study, but this may be due to urban composition of the Carletonville study's sample. A review by Eaton *et al* (2003) on sexual behaviour among South African youth concluded that at least 50% of South African youth is sexually active by the age of 16 years. The Carletonville study also found similar high rates of reported STI symptoms and numbers of lifetime partners (Williams, *et al* 2000).

The results indicate modest differences between the models with and without frailty. A number of factors can lead to this observation. Firstly, it is possible that the assumed distribution of random effects is inappropriate. The results will be sensitive to the assumed frailty distribution if the proportion uninfected with HIV within the period of analysis is low, and the frailty variance is large (Sastry, 1997). These data are from South Africa, which is among countries with the world's highest rate of HIV infection. We have also found an estimate of frailty variance much smaller than 2. Secondly, high risk sexual networks will get infected earlier than low risk sexual networks. In the long run, low risk sexual networks will dominate the sample and the risk will decline over time (Guo and Rodriguez, 1992).

The analysis was restricted to the duration of sexual activity since the epidemic of HIV became well established in South Africa. However, it is possible that some people were infected before 1990. Although their chances of surviving for more than ten years in the absence of antiretroviral drugs are minimal, the results of the analysis using time since first sexual intercourse led to similar conclusions and these are shown in Table 3.4. The minor differences were in the reduced baseline hazards and increased fixed effects estimates in the analysis of time since first sexual intercourse. The reduction in baseline hazard estimate was expected since sexual activity prior to the period when the epidemic was well established posed very little risk compared to the period after 1990.

Inference based on correlated observations can lead to substantial bias in estimated parameters, especially when unobserved random effect variance is large (Klein, 1992; Guo and Rodriguez, 1992). The effect of including sexual partnership frailty on fixed effect parameters and standard errors, though very minor, appears to be important because it reveals systematic bias in the same direction. Standard errors are underestimated in the model that ignores the within sexual network cor-

Table 3.4: Estimates for time since first sexual intercourse until HIV infection

	Model with frailty	
Parameter	Estimate	SE
Baseline hazard		
Constant	0.0069	0.0011
Migration status		
Migrant men	0.4598	0.2161
Partners of migrant men	0.2991	0.2104
Non-migrant men	-0.2186	0.2590
Partners of non-migrant men[a]		
Age in years		
18 to 24	2.4547	0.2963
25 to 34	1.0720	0.1628
35 and above[a]		
Recent sexual contact partners		
Only one[a]		
More than one	0.5575	0.1893
Number of lifetime partners		
Only one[a]		
More than one	0.3284	0.1719
Syphilis		
0=Negative, 1=Positive	0.2836	0.1580
Status of other STIs		
0=Negative, 1=Positive	0.5031	0.1807
Frailty variance		
Sexual network	0.4588	0.0691

[a]Reference category

relation in contrast to the model that takes into account that dependency within the data. If this dependency is not taken into account in the analysis, confidence intervals and credible intervals for fixed effects will be too narrow (Wei, *et al* 1989; Guo and Lin, 1994) even in cases where failure times are modestly correlated (Kim and Xue, 2002). This leads to overstated significance levels and coverage probabilities of the corresponding confidence intervals fall below nominal level.

In this chapter, we have considered an approach of finding ML estimates in correlated interval-censored data. Parameters are estimated using the EM algorithm with appreciably simple steps. The idea is to treat both the interval-censored infection times and frailties as unobserved data to facilitate the EM algorithm. Methods of correlated standard survival data were used to estimate parameters. Only the interval-censored observations were estimated, conditional on clinical examination times. There is no practical relevance in augmenting right-censored observations as these observations are censored at the end of the study or lost to follow-up. Thus, their last clinical examination times are known. The approach grossly simplifies the analysis in correlated interval-censored data and yields the tractable marginal likelihoods needed to facilitate the EM algorithm. The results of this chapter have been submitted for publication (Zuma, Lurie, Jorgensen, 2004).

Chapter 4

Bayesian simulation methods

4.1 Introduction

The previous two chapters showed how the parameters of the logistic mixed model and the proportional hazards frailty model could be estimated using the EM algorithm, which is a frequentist approach. The likelihood function could not be obtained analytically and was approximated using numerical methods such as Gaussian-Hertmite quadrature which may be imprecise (Crouch and Spiegelman, 1990; Monahan and Stefanski, 1992). Bayesian methods avoid the need for evaluating complex integrals. A Bayesian paradigm treats all unknown parameters as random variables and assigns a prior distribution $g(\theta)$ for the unknown parameter vector θ. Posterior inference about θ is obtained by using the likelihood function $f(y|\theta)$ to convert prior uncertainty $g(\theta)$ into posterior probability statement $g(\theta|y)$. In this way, the posterior distribution summarizes our knowledge of θ after observing the data y. Instead of obtaining $g(\theta|y)$ analytically, we use Markov chain Monte Carlo (MCMC) methods to generate samples from the joint posterior. In this chapter, we discuss various methods of generating samples from the posterior distribution and the related implementation issues.

Bayes' theorem presents the posterior distribution of θ as

$$g(\theta|y) = \frac{f(y|\theta)\, g(\theta)}{\int f(y|\theta)\, g(\theta)\, d\theta}.$$

The posterior density is usually presented in its proportional form as

$$g(\theta|y) \propto f(y|\theta)g(\theta). \tag{4.1}$$

Therefore, the posterior distribution of θ is proportional to the product of the likelihood function and the prior distribution. The unscaled posterior (4.1) only provides the shape of $g(\theta|y)$. From the unscaled posterior we can find the modes and relative frequencies at any two locations. The exact posterior distribution is found by re-scaling (4.1) so it forms a density. Re-scaling (4.1) requires division by

$$\int f(\theta|y)g(\theta)d\theta.$$

In practise, it may be very difficult to evaluate this integral. A closed form solution of the posterior density can be found in few special cases, such as when $f(y|\theta)$ is a member of the exponential family and the prior density is from the conjugate family of priors. For other cases, the density has to be approximated numerically using complicated asymptotic techniques such as Laplace methods (Tierney and Kadane, 1986; Carlin and Louis, 1996), which maybe inadequate in high dimensional parameter space.

Sampling based methods were developed which allow us to draw samples from the posterior density when we don't know the posterior completely, only its unscaled form. Direct sampling methods such as *Acceptance-Rejection Sampling, Sampling Importance Resampling* and *Adaptive-Rejection Sampling* are discussed in Section 4.2. They generate samples directly from the posterior distribution by drawing random samples from an easily sampled candidate density and reshaping it to accept some of the values into the final sample. The accepted values give us a random sample from the posterior. These methods work very well for low-dimensional parameter space. However, direct sampling methods lose efficiency very quickly as

the dimension of the parameter space increases. The candidate density has heavier tails in each dimension. So, almost all candidate draws will be from the region of extremely low posterior probability and will not be accepted into the final sample.

The MCMC methods (Gilks, *et al* 1996; Carlin and Louis, 1996) for obtaining a random sample from the posterior have been developed and are discussed in Section 4.3. They are much more efficient than the direct sampling methods when the parameter dimension is high. They set up a Markov chain that has the posterior as its long-run distribution. Running the Markov chain for a while moves the chain out of the region of the parameter space that has extremely low posterior probability. Practical issues involved in implementing MCMC sampling methods are addressed in Section 4.3.5.

4.2 Direct sampling methods

Direct non-iterative sampling techniques require more than one step for obtaining a sample from the posterior $g(\theta|y)$ which may only be known in its unscaled form $f(\theta|y)g(\theta)$. Typically, these sampling methods involve two steps. The first step samples random variables from a candidate distribution $g_0(\theta)$. The second step adjusts the sample to approximate $g(\theta|y)$. Direct sampling techniques generate statistically independent samples unless correlation was introduced as a variance reduction tool. In this section, we briefly describe direct sampling methods that are often used: *Acceptance-Rejection Sampling, Sampling Importance Resampling* and *Adaptive-Rejection Sampling.*

4.2.1 Acceptance-Rejection Sampling

Acceptance-Rejection Sampling (ARS) utilises a candidate density $g_0(\theta)$ with heavier tails than the posterior $g(\theta|y)$. The density $g_0(\theta)$ is chosen such that it is easy to draw samples from. By only accepting some candidates, the sample is reshaped

to be a random sample from $g(\theta|y)$. Suppose that there exists a positive constant $M < \infty$ where M is the smallest value such that $f(\theta|y)g(\theta) \leq Mg_0(\theta)$, for every possible value of θ. The density $g_0(\theta)$ is referred to as a *blanketing density* or envelop and M is the *envelop constant.* The ARS algorithm proceeds as follows:

- Draw θ_j from $g_0(\theta)$
- Draw U from a uniform (0,1) distribution
- If $U \leq f(\theta_j|y)g(\theta_j)/Mg_0(\theta_j)$ then accept θ_j, otherwise reject θ_j
- Return to first step until the required sample is attained.

An accepted random variable θ_j comes from $g(\theta_j|y)$. The proof that ARS samples from $g(\theta_j|y)$ only requires us to show that the conditional density of $[\theta_j|U \leq f(\theta_j|y)g(\theta_j)/Mg_0(\theta_j)]$ is $g(\theta_j|y)$. Since θ and U are independent and U is uniform then their joint density is $g_0(\theta)$. In a narrow slice, that is, for a very small Δ

$$\begin{aligned}
& P\left[\left(\theta_j - \frac{\Delta}{2} < \theta_j < \theta_j + \frac{\Delta}{2}\right) \bigcap \text{we accept } \theta\right] \\
\doteq\ & P\left[\left(\theta_j - \frac{\Delta}{2} < \theta_j < \theta_j + \frac{\Delta}{2}\right) \bigcap U \leq \frac{f(\theta_j|y)g(\theta_j)}{Mg_0(\theta_j)}\right] \\
\doteq\ & \int_{\theta_j - \frac{\Delta}{2}}^{\theta_j + \frac{\Delta}{2}} \frac{f(\theta_j|y)g(\theta_j)}{Mg_0(\theta_j)} \times g_0(\theta)\, d\theta \\
\doteq\ & \frac{f(\theta_j|y)g(\theta_j)}{Mg_0(\theta_j)} \times \Delta g_0(\theta_j).
\end{aligned}$$

Dividing this probability by Δ and taking the limit as $\Delta \longrightarrow 0$ we get the derivative which is the density of accepted θ at θ_j proportional to $f(\theta_j|y)g(\theta_j)$. Therefore the density of accepted θ is proportional to $g(\theta|y)$ which is the posterior.

The efficiency of the algorithm depends on similarities between $g_0(\theta)$ and $g(\theta|y)$.

If π is the acceptance probability for θ_j, then

$$\begin{aligned}
\pi &= \Pr[U \leq f(\theta_j|y)g(\theta_j)/Mg_0(\theta_j)] \\
&= \int \Pr[U \leq f(\theta_j|y)g(\theta_j)/Mg_0(\theta_j)]g_0(\theta_j)d\theta_j \\
&= \int [f(\theta_j|y)g(\theta_j)/M]d\theta_j \\
&= \frac{1}{M}\int f(\theta_j|y)g(\theta_j)d\theta_j \\
&= \frac{c}{M}
\end{aligned}$$

where c is the standardizing constant for $g(\theta|y)$. Therefore, the number of iterations required to accept a single θ_j is a geometric random variable with mean $\pi^{-1} = M/c$. The value of M is determined from $g_0(\theta)$. Efficiency of the algorithm can be improved by choosing $g_0(\theta)$ with the same shape as $g(\theta|y)$ but with heavier tails. Heavy tailed distributions such as student's t-distribution with low degrees of freedom are recommended (Chib and Greenberg, 1995). Typically, ARS is characterized with elaborate exploration of different candidate envelope functions. Gilks and Wild (1992) proposed an *Adaptive-Rejection Sampling* technique that constructs and adapts the candidate generating distribution at each non-accepted value until an accepted value is achieved.

4.2.2 Sampling Importance Resampling

The Sampling Importance Resampling (SIR) first proposed by Rubin (1987) is a two stage method of sampling from $g(\theta|y)$. The SIR stems from an idea of sampling from $g_0(\theta)$ without having to determine the value of M. Suppose we want a sample of size n' from $g(\theta|y)$ which is hard to sample from, whilst a sample $\{\theta_1, \cdots, \theta_{N'}\}$, $n' < N'$, is available from $g_0(\theta)$. Then, calculate the *sampling importance weight* w_j for each value of the sample where

$$w_j = \frac{q_j}{\sum_{j=1}^{N'} q_j} \quad \& \quad q_j = \frac{f(\theta_j|y)g(\theta_j)}{g_0(\theta_j)}.$$

The ultimate $\{\theta_1^*, \cdots, \theta_{n'}^*\}$ is obtained by resampling from $\{\theta_1, \cdots, \theta_{N'}\}$ using weights $\{w_1, \cdots, w_{N'}\}$. The resulting sample is approximately distributed as $g(\theta|y)$ with approximation improving as N' increases. Smith and Gelfand (1992) proved that SIR

effectively samples from $g(\theta|y)$. Typically, SIR is a 'bootstrap' resampling with unequal probabilities determined by w_j. However in SIR, the parameters rather than the data are resampled hence it is often called Bayesian bootstrap.

4.2.3 Adaptive-Rejection Sampling

Adaptive-Rejection Sampling (AdRS) is a sampling algorithm applicable to a special class of log-concave univariate densities (Gilks and Wild, 1992). The AdRS approximates $\log g(\theta|y)$ by drawing line segments that blanket $\log g(\theta|y)$. The line segments are formed by constructing tangents to $\log g(\theta|y)$. In many applications, full conditionals are log-concave (Dellaportas and Smith, 1993). The blanketing envelope is piece-wise exponential and easy to sample from. Gilks and Wild, (1992) proposed an alternative squeezing function that does not require evaluation of the derivative of $\log\ g(\theta|y)$. The method is iterative. At each iteration both the envelope and squeezing functions are updated thus improving efficiency of the algorithm. However, the complexities involved in multivariate generalization of the AdRS makes it unattractive for cases beyond univariate densities.

4.3 Markov Chain Monte Carlo methods

Direct sampling methods discussed thus far reshape the sample drawn from the candidate distribution to that drawn from the posterior distribution. These methods can be very inefficient in high dimensional parameter space. The MCMC methods provide an efficient way of sampling from a high dimensional posterior through setting up a Markov chain that has the posterior distribution as its stationary distribution. There are a number of books devoted to the subject of Markov chains. These include, among others (Meyn and Tweedie, 1993; Ross, 1996). For a more statistical oriented approach, see for example (Guttorp, 1995; Gamerman, 1997).

We wish to sample from $g(\theta|y)$ over the parameter space. Suppose $g(\theta|y)$ is

known up to a multiplicative constant. The posterior $g(\theta|y)$ is a *long-run* distribution of an aperiodic ergodic Markov chain with *transition kernel* $P(\theta, A)$ if and only if it satisfies the *steady state equation*

$$\int_A g(\theta|y)d\theta = \int g(\theta|y)P(\theta, A)d\theta$$

where the probability transition kernel $P(\theta, A)$ is a mapping from points in the parameter space into measurable sets in the parameter space

$$P(\theta^{(r)}, A) = P(\theta^{(r+1)} \in A|\theta^{(r)}).$$

Unlike classical Markov chain analysis where we know the transition function and we want to find the long-run distribution, in this case we know the long-run distribution of the Markov chain and the problem involves finding the appropriate transition kernel $P(\theta, A)$. There are many possible transition kernels that have the long-run distribution. Surprisingly it is not too difficult to find one. The MCMC sampling methods such as *Metropolis-Hastings algorithm*, *substitution sampler* and *Gibbs sampler* are all different methods of finding the transition kernel whose long-run distribution is $g(\theta|y)$. These methods are discussed in Section 4.3.1 through to Section 4.3.3. Sample values drawn after running the Markov chain for a reasonably long time, say c iterations, approximate draws from $g(\theta|y)$ and thus $\{\theta^{(r)} : r = c+1, \cdots, N\}$ are dependent samples from the desired posterior density. We discuss methods of determining c in Section 4.3.5. The estimate of the expectation of any function $E[h(\theta^{(r)}]$ can be obtained using the *ergordic average*

$$\bar{h} = \frac{1}{N-c}\sum_{r=c+1}^{N} h(\theta^{(r)}).$$

However, the constructed Markov chain has to satisfy the regularity conditions of *irreducibility* and *aperiodicity* before this can hold (Smith and Roberts, 1993). Suppose that the candidate distribution which generates a candidate value θ' given θ is $q(\theta, \theta')$. If $q(\theta, \theta')$ satisfies reversibility condition

$$g(\theta|y)q(\theta, \theta') = g(\theta'|y)q(\theta', \theta) \tag{4.2}$$

for all values of θ and θ', then $g(\theta|y)$ is the long-run distribution for a Markov chain whose transition kernel is

$$P(\theta, A) = \int_A q(\theta, \theta')d\theta' + s(\theta)\delta_A(\theta)$$

where $s(\theta) = 1 - \int q(\theta, \theta')d\theta'$ is the probability that a chain remains at θ and $\delta_A(\theta) = 1$ if $\theta \in A$ and 0 otherwise. To prove this, we need to evaluate $\int g(\theta|y)P(\theta, A)d\theta$ and show that it is equal to $\int_A g(\theta|y)d\theta$ as follows

$$\int g(\theta|y)P(\theta, A)d\theta = \int \int_A g(\theta|y)q(\theta, \theta')d\theta' d\theta + \int g(\theta|y)s(\theta)\delta_A(\theta)d\theta.$$

Since $\delta_A(\theta) = 1$ if $\theta \in A$ and 0 otherwise, the corresponding integral needs only be evaluated over the region A

$$\begin{aligned}
\int g(\theta|y)P(\theta, A)d\theta &= \int \int_A g(\theta|y)q(\theta, \theta')d\theta' d\theta + \int_A g(\theta|y)s(\theta)d\theta \\
&= \int_A \int g(\theta|y)q(\theta, \theta')d\theta d\theta' + \int_A g(\theta|y)s(\theta)d\theta \\
&= \int_A \int g(\theta'|y)q(\theta', \theta)d\theta d\theta' + \int_A g(\theta|y)s(\theta)d\theta \\
&= \int_A g(\theta'|y)[1 - s(\theta')]d\theta' + \int_A g(\theta|y)s(\theta')d\theta \\
&= \int_A g(\theta|y)d\theta. \qquad (4.3)
\end{aligned}$$

Therefore, the chain is positive recurrent and irreducible with stationary distribution $g(\theta|y)$.

On the other hand, suppose the candidate distribution does not satisfy the reversibility condition. In Section 4.3.1 we show how the candidate generating distribution can be modified so that we can construct a Markov chain having $g(\theta|y)$ as its long run distribution for that case.

4.3.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings (M-H) algorithm is an MCMC sampling method which first appeared in mathematical physics dealing with the calculation of properties of chemical substances (Metropolis, *et al* 1953). Hastings (1970) subsequently generalized the algorithm and hence it is now known as the M-H algorithm. Tierney

(1994) gives a comprehensive theoretical exposition of this algorithm, whilst Chib and Greenberg (1995) provide an excellent tutorial on the topic of the M-H algorithm.

Similar to the ARS, suppose we have a candidate generating density $q(\theta, \theta')$. In the context of Markov chains, the candidate generating density is allowed to depend on the current state of the process. If $q(\theta, \theta')$ satisfies the reversibility condition (4.2) then its long-run distribution is $g(\theta|y)$, see Equation (4.3). In general however, it does not always satisfy the reversibility condition such that for some θ and θ'

$$g(\theta|y)q(\theta, \theta') > g(\theta'|y)q(\theta', \theta).$$

The process moves from θ to θ' more often than from θ' to θ. A convenient way to circumvent this is to reduce the number of moves from θ to θ' while leaving the number of moves from θ' to θ unchanged. This can be achieved by introducing a probability of move from $\alpha(\theta, \theta') \leq 1$. The reversibility condition becomes

$$g(\theta|y)q(\theta, \theta')\alpha(\theta, \theta') = g(\theta'|y)q(\theta', \theta)\alpha(\theta', \theta)$$

In this case $\alpha(\theta', \theta) = 1$ so that

$$g(\theta|y)q(\theta, \theta')\alpha(\theta, \theta') = g(\theta'|y)q(\theta', \theta).$$

Clearly, the probability of move from θ to θ' is given by

$$\alpha(\theta, \theta') = \frac{g(\theta'|y)q(\theta', \theta)}{g(\theta|y)q(\theta, \theta')}.$$

To ensure that $q(\theta', \theta)\alpha(\theta', \theta)$ satisfies the reversibility condition, the probability of a move from θ to θ' should be set to

$$\alpha(\theta, \theta') = \min\left\{\frac{g(\theta'|y)q(\theta', \theta)}{g(\theta|y)q(\theta, \theta')}, 1\right\} \text{ whenever } g(\theta|y)q(\theta, \theta') > 0.$$

It is important to note that calculation of $\alpha(\theta, \theta')$ requires that $g(\theta|y)$ be known only to a multiplicative constant. This means the M-H algorithm can be implemented

even when we only know $g(\theta|y)$ up to a proportionality constant. Consequently, $g(\theta|y)$ is the long-run distribution of a Markov chain with the transition kernel

$$P(\theta, A) = \int_A q(\theta, \theta')\alpha(\theta, \theta')d\theta' + s(\theta)\delta_A(\theta)$$

where $s(\theta) = 1 - \int q(\theta, \theta')\alpha(\theta, \theta')d\theta'$ is the probability that the chain remains at θ, $\delta_A = 1$ if $\theta \in A$ and 0, otherwise.

We now summarize the steps of the M-H algorithm initiated at $\theta^{(0)}$

- Repeat for $r = 1, 2, \cdots, N$
- Draw θ' from $q(\theta^{(r-1)}, \theta)$ and U from uniform $(0, 1)$ distribution independently of each other and previous draws
- Calculate $\alpha(\theta^{(r-1)}, \theta')$
- If $U < \alpha(\theta^{(r-1)}, \theta')$ then set $\theta^{(r)} = \theta'$, else set $\theta^{(r)} = \theta^{(r-1)}$
- Return the values $\{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(N)}\}$.

The chain converges to $g(\theta|y)$ after it has been run for a sufficiently long *burn-in* period such that the effect of the fixed starting point is negligible. Sample values obtained after the *burn-in* period approximate a dependent sample from $g(\theta|y)$.

Successful implementation of M-H algorithm depends on how close $q(\theta, \theta')$ mirrors $g(\theta|y)$. If $q(\theta, \theta')$ is the true posterior density, then $\alpha(\theta, \theta') = 1$ and θ' would always be accepted. Generally, $q(\theta, \theta')$ would be from the family of densities that require specification of parameters such as the location and spread. The spread of the candidate density affects both the acceptance and mixing of the chain (Chib and Greenberg, 1995). Metropolis, *et al* (1953) proposed a random walk chain $q(\theta, \theta') = g_1(\theta' - \theta)$ where g_1 is a multivariate density. The density generates a candidate $\theta' = \theta + \varepsilon$ where ε is called the increment random variable. The increment random variable ε follows the distribution g_1. Since the candidate is equal

to the current value plus noise, Chib and Greenberg (1995) call this case a random walk chain. The possible forms of g_1 are multivariate normal and multivariate-t densities. Roberts *et al* (1997) show that if the target and the proposal densities are normal then the scale of the proposal distribution should be tuned such that the acceptance rate is approximately 0.45 in one-dimension and decrease towards 0.23 as the number of dimensions approach infinity.

Hastings (1970) proposed a candidate density $q(\theta, \theta') = g_2(\theta')$ which is independent of the current value of θ. The independence M-H chain results in

$$\alpha(\theta, \theta') = \min\left\{\frac{w(\theta')}{w(\theta)}, 1\right\}$$

where $w(\theta) = g(\theta/y)/g_2(\theta)$. The function w is the importance weight that would be used in the SIR if observations were sampled from g_2. Thus, a candidate θ' with low weight would rarely be accepted. Chib and Greenberg (1995) suggest that an independent candidate density $g_2(\theta)$ with heavier tails be used. Tierney (1994) presents several classes of possible candidate densities.

4.3.1.1 M-H Acceptance-Rejection chains

It was noted in the ARS algorithm that M and $g_0(\theta)$ should be chosen such that $Mg_0(\theta)$ dominates $g(\theta|y)$ for all values of θ. This can be difficult especially when $g(\theta|y)$ depends on parameters that are updated at every iteration. Tienery (1994) proposed a rejection sampling scheme that drives an independent M-H chain to circumvent the problem of finding a suitable value of M. Define $C = \{\theta : g(\theta|y) < Mg_0(\theta)\}$. The candidate θ' is assumed to come from an ARS algorithm. Since θ and θ' can each be in C or $\bar{C}$, there are four possible cases: $\theta \in C$ and $\theta' \in C$; $\theta \notin C$ and $\theta' \in C$; $\theta \in C$ and $\theta' \notin C$; and $\theta \notin C$ and $\theta' \notin C$. Chib and Greenberg (1995) provided the M-H acceptance probability such that $q(\theta')\alpha(\theta, \theta')$ satisfies the reversibility condition. Finally, the acceptance probability $\alpha(\theta, \theta')$ can be written

as

$$\alpha(\theta,\theta') = \begin{cases} 1 & \text{if } \theta \in C \\ \frac{Mg_0(\theta)}{g(\theta|y)} & \text{if } \theta \notin C \text{ and } \theta' \in C \\ \min\left\{\frac{g(\theta'|y)g_0(\theta)}{g(\theta|y)g_0(\theta')}, 1\right\} & \text{if } \theta \notin C \text{ and } \theta' \notin C. \end{cases}$$

It is clear that in both cases where $\theta \in C$ the probability of move to θ' is 1 irrespective of where θ' lies.

4.3.1.2 Blockwise M-H algorithm

It is often efficient to work with the components of θ instead of the full dimensional parameter. Suppose the parameter vector θ can be partitioned into sub-blocks of $\{\theta_0, \cdots, \theta_b\}$ where θ_k is a block of parameters which may contain sub-vectors of parameters. Let θ_{-k} denote a set comprising all components of θ except θ_k. Hastings (1970) showed how to apply M-H algorithm to subblocks of θ rather than to all components of θ simultaneously. Let $P_k(\theta_k, A_k|\theta_{-k})$ be the transition kernel for the M-H algorithm applied to subblock θ_k of the vector θ keeping all other parameter blocks fixed. Hastings (1970) showed that

$$P(\theta, A) = \prod_{k=1}^{b} P_k(\theta_k, A_k|\theta_{-k})$$

has $g(\theta|y)$ as its long-run distribution. Chib and Greenberg (1995) illustrated the idea of *product of kernels principle* for two blocks of move from $\theta = \{\theta_1, \theta_2\}$ to $\theta' = \{\theta_1', \theta_2'\}$. The principle allows us to draw consecutively from each kernel, instead of running each of the kernels to convergence for every value of the conditioning block. The practical significance of the principle is that it is often much easier to find several conditional kernels that converge to their respective posterior densities than finding one kernel that converges to the joint posterior density. The Gibbs sampler is an example of the component M-H algorithm since it uses fixed sequence of Gibbs transition kernels which update different components of the state vector, Section 4.3.3.

4.3.2 Substitution sampling

Tanner and Wong (1987) proposed a Markov chain algorithm that has $g(\theta|y)$ as its limiting distribution using ideas similar to those of the EM algorithm. The algorithm is sometimes referred to as the *data augmentation* since observed data y is augmented with unobserved data z. The variable y is augmented in such a way that if both z and y were known, it could be easy to evaluate the complete-data posterior $g(\theta|z,y)$. The incomplete data posterior $g(\theta|y)$ is given by

$$g(\theta|y) = \int g(\theta|z,y)p(z|y)dz$$

where the predictive distribution of the unobserved variable $p(z|y)$ is given by

$$p(z|y) = \int p(z|\theta',y)g(\theta'|y)\,d\theta'.$$

Substituting the second equation into the first equation and interchanging the order of integration results in

$$\begin{aligned} g(\theta|y) &= \int g(\theta|z,y)\int p(z|\theta',y)g(\theta'|y)d\theta'dz \\ &= \int h(\theta,\theta')g(\theta'|y)d\theta' \end{aligned} \tag{4.4}$$

where $h(\theta,\theta') = \int g(\theta|z,y)p(z|\theta',y)dz$. Hence $g(\theta|y)$ is a fixed point solution of the integral (4.4). The uniqueness of $g(\theta|y)$ is discussed in Gelfand and Smith (1990). Given an appropriate approximation $g_r(\theta)$ to $g(\theta|y)$, Tanner and Wong (1987) showed that the sequence of distributions found by successive solutions of the integral equations

$$g_r(\theta) = \int h(\theta,\theta^{(r-1)}))g_{(r-1)}(\theta^{(r-1)})d\theta^{(r-1)} \tag{4.5}$$

converge to $g(\theta|y)$ at a uniform convergence rate. Equation (4.5) requires evaluation of the integral that maybe difficult to perform analytically. Therefore, a sampling method is used to generate a sequence of random variables from each of the successive distributions. The algorithm proceeds as follows:

- Draw $\theta^{(0)}$ from $g_0(\theta)$

- Draw $z^{(0)}$ from $p(z|\theta^{(0)}, y)$
- Iteratively for $r = 1, \cdots, n$
- Draw $\theta^{(r)}$ from $g(\theta|z^{(r-1)}, y)$
- Draw $z^{(r)}$ from $p(z|\theta^{(r)}, y)$.

In general, the algorithm draws $\theta^{(r)}$ from $g(\theta|z^{(r-1)}, y)$ then draws $z^{(r)}$ from $g(z|\theta^{(r)}, y)$ successively substituting θ and z in turn at each iteration, hence the name substitution sampler. At each iteration, the algorithm produces a pair $(\theta^{(r)}, z^{(r)})$ with marginal densities $g_r(\theta)$ and $g_r(z)$ respectively (Carlin and Lewis, 1996). The final form of $g(\theta|y)$ based on the final sample can be obtained using the kernel density estimate (Gelfand and Smith, 1990). An extension of the algorithm to more than two components is illustrated in Gelfand and Smith (1990). Implementation of substitution sampler to d variables requires the availability of all $d(d-1)$ conditional distributions. These conditional distributions are sometimes hard to find.

4.3.3 Gibbs sampler

Geman and Geman (1984) proposed an MCMC method which forms the transition kernel using fewer full conditional distributions. The method is named the Gibbs sampler since it originated in image processing where the posterior was the Gibbs distribution. As discussed in Besag and Green (1993), the Gibbs sampler is founded on the ideas of Grenander (1983). Gelfand and Smith (1990) noted its potential, popularized the method to the wider statistical community and pointed out the connections between the Gibbs sampler and Substitution sampler.

The Gibbs sampler assumes that all full conditional distributions

$$g_k(\theta_k|\theta_1, \theta_2, \cdots, \theta_{k-1}, \theta_{k+1}, \cdots, \theta_b, y)$$

are available to sample from in order to approximate $g(\theta|y)$. Based on Besag (1974), the set of conditional posterior distributions are sufficient, after satisfying some

conditions, to determine the joint posterior distribution. These conditions are conditional independence between y given the model parameters and covariates, and independence between the parameters themselves. Let $\theta^{(0)} = \{\theta_1^{(0)}, \cdots, \theta_b^{(0)}\}$ be an arbitrary starting value. The Gibbs sampler moves from $\theta^{(i)}$ to $\theta^{(i+1)}$ by successively updating each variable as follows:

$$\begin{aligned}
&\theta_1^{(i+1)} \quad \text{sampled from } g(\theta_1|\theta_2^{(i)}, \cdots, \theta_b^{(i)}, y) \\
&\theta_2^{(i+1)} \quad \text{sampled from } g(\theta_2|\theta_1^{(i+1)}, \theta_3^{(i)}, \cdots, \theta_b^{(i)}, y) \\
&\\
&\theta_k^{(i+1)} \quad \text{sampled from } g(\theta_k|\theta_1^{(i+1)}, \theta_2^{(i+1)}, \cdots, \theta_{k-1}^{(i+1)}, \theta_{k+1}^{(i)}, \cdots, \theta_b^{(i)}, y) \\
&\\
&\theta_b^{(i+1)} \quad \text{sampled from } g(\theta_b|\theta_1^{(i+1)}, \theta_2^{(i+1)}, \cdots, \theta_{b-1}^{(i+1)}, y).
\end{aligned}$$

This completes a transition from $\theta^{(i)}$ to $\theta^{(i+1)}$. Iteration of the algorithm produces a sequence of $(\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(r)}, \cdots)$ which is a realization from a Markov chain with transition probability of a move from $\theta^{(r-1)}$ to $\theta^{(r)}$ given by

$$P(\theta^{(r-1)}, \theta^{(r)}) = \prod_{k=1}^{b} g_k(\theta_k|\theta_1^{(r-1)}, \cdots, \theta_{k-1}^{(r-1)}, \theta_{k+1}^{(r-1)}, \cdots, \theta_b^{(r)}, y)$$

and the stationary distribution $g(\theta|y)$. The Gibbs sampler is an example of a component-based MCMC method. In fact, the Gibbs sampler is the block-wise M-H algorithm using the exact full conditionals as the candidate generating distribution. Geman and Geman (1984) showed that after running the chain for a sufficiently long time, say c, $(\theta_1^{(c)}, \cdots, \theta_b^{(c)})$ converges to $g(\theta|y)$ at an exponential rate, see also Gelfand and Smith (1990). Thus, by taking c large enough any population characteristic including the density can be obtained with high accuracy. General applications of the Gibbs sampler are illustrated in Smith and Roberts (1993). Brooks (1998) provides a comprehensive review of MCMC methods and their applications.

The degree of disaggregation of the parameter into blocks should take the correlation structure of $g(\theta|y)$ into account. Parameters that are highly correlated should be included in the same block, and sampled together in such a way that takes the

correlation structure into account. Otherwise, successive draws from the chain will move very slowly through the parameter space. The chain will be said to have *poor mixing properties.* This will result in high *burn-in* time required. Gelfand and Smith (1990) suggest a method analogous to the Monte Carlo integration for n' independent Gibbs sequences

$$\bar{g}_k(\theta_k|y) = \frac{1}{n'}\sum_{j=1}^{n'} g_k(\theta_k|\theta_{j1}^{(c)}, \theta_{j2}^{(c)}, \cdots, \theta_{j(k-1)}^{(c)}, \theta_{j(k+1)}^{(c)}, \cdots, \theta_{jb}^{(c)}, y)$$

as a density estimator of $g_k(\theta_k|y)$. Gilks *et al* (1993) reviewed applications of Gibbs sampler in medicine, longitudinal data, disease mapping and survival models. Illustrations of the Gibbs sampler in multinomial and normal models are presented in Gelfand and Smith (1990). Casella and George (1992) gave details of how and why the Gibbs sampler works. Brooks (1998) tackles the applications of both Gibbs sampler and M-H algorithm. Casting of GLMMs in Bayesian framework using Gibbs sampler appeared in Zeger and Karim (1991). Clayton (1991) introduced the Bayesian model for analysis of multivariate survival data using the frailty model and estimated parameters via the Gibbs sampler. Other successful applications of the Gibbs sampler have appeared in the analysis of three-way multilevel survival data by Bolstad and Manda (2001).

4.3.3.1 Hierarchical and graphical modelling

Many statistical applications involve estimating parameters that are somehow interrelated. The joint probability model should reflect the dependency between the parameters. In Bayesian modelling, some parameters can be viewed as samples from a common population distribution. Parameters describing the population distribution are referred to as *hyperparameters.* Hyperparameters are unknown and estimated from the observed data. Priors are specified for hyperparameters and referred to as *hyperpriors.* This setup forms a hierarchical structure of parameters. In most practical problems, such as in random effects models, parameters can be subsumed within the hierarchical framework.

The hierarchical structure of parameters can be represented by a *directed acyclic graph* (DAG) which shows the dependency structure of parameters including *priors*, hyperpriors, hyperparameters and the observed data (Clayton, 1991). Each model parameter appears as a node in the graph with edges representing the direction of the relationship. Some quantities are fixed constants such as priors, hyperpriors and observed data. These are often represented by rectangles around them. Circles are drawn around stochastic parameters that are subject to estimation. Each node is connected, with an arrow, to those nodes that depend on it. Probabilistic dependencies are represented by solid lines and dashed lines represent deterministic dependencies. Parameters having direct influence on the data appear at the bottom of the hierarchy and those with lesser influence placed at the top of the hierarchy.

Let θ_k be the component of interest and θ be the set of all nodes on the graph. Factors that node θ_k directly depends on are called its *parent* nodes. The nodes that directly depend on θ_k are called its *children* nodes. The DAG represents the assumption of conditional independence. That is, for any node θ_k if we know its parents then no other node contains information about θ_k except its children. Therefore, it is possible to factorize the full joint posterior $f(\theta; y)$ as a product of conditional distributions

$$f(\theta; y) = \prod_{\nu \in \{\theta, y\}} g(\nu | \text{parents of } \nu)$$

where ν is any particular node. The distribution of all other nodes θ_{-k} is found by integrating θ_k out of the full joint posterior. All terms not containing θ_k are constant with respect to the integration and can be moved outside of the integral. Therefore, in calculating the conditional distribution of $\theta_k | \theta_{-k}; y$ terms not containing θ_k cancel out and we are left with

$$g_k(\theta_k | \theta_{-k}; y) \quad \propto \quad g(\theta_k | \text{parents of } \theta_k) \prod_{\omega: \phi_\omega \in \{\text{children of } \theta_k\}} f(\phi_\omega | \theta_k \text{ and co-parents}) \cdot$$

The co-parents nodes are other nodes which are also parents of ϕ_ω. The $g_k(\theta_k | \theta_{-k}, y)$ has the prior component $g(\theta_k | \text{parents of } \theta_k)$ and the likelihood component given

by $\prod_{\omega:\phi_\omega \in \{\text{children of } \theta_k\}} f(\phi_\omega|\theta_k \text{ and co-parents})$. Thus, in hierarchical model the conditional posterior of any node given all other nodes, only depends on its parents, its children and other parents of its children. Conjugacy of the prior component to the likelihood component reduces mathematical complexities involved in formulating a Gibbs conditionals. However, it is not necessary to use conjugate priors as the nodes can be sampled directly using AR or AdR sampling.

4.3.3.2 Relationship to M-H and Substitution algorithms

In M-H algorithm, consider a block-wise transition kernel

$$P_k(\theta_k, A_k) = \int_{A_k} g_k(\theta_k|\theta_{-k}; y)d\theta_k.$$

This shows that the Gibbs sampler is a special case of M-H algorithm. For each block θ_k we are sampling from a correct full conditional posterior. Therefore, the acceptance probability of a move from θ to θ' at each step is $\alpha(\theta, \theta') = 1$. See, for example, Brooks (1998) for a formal proof that $\alpha(\theta, \theta') = 1$ in the Gibbs sampler.

In the substitution sampler, suppose θ is partitioned into components. For each θ_k let the missing data $Z_k = \theta_{-k}$. In this case, the substitution sampler is equivalent to Gibbs sampler but with different visitation order (Gelfand and Smith, 1990).

4.3.4 Prior and propriety of posterior distributions

The posterior distribution discussed thus far has been implicitly assumed to be proper. That is, the integral of $g(\theta|y)$ is finite and $g(\theta|y)$ can be written as a known probability density function. Specification of prior information for both fixed effects and variance components partly determines the propriety of the posterior distribution. It is common in hierarchical linear mixed models to specify an improper prior $f(\beta_1, \cdots, \beta_p) = 1$ for fixed effects and Gamma(α, β) or inverse-Gamma(α, β) priors for variance components (Gelfand and Smith, 1990; Hobert and Casella, 1996; Bolstad, 1997; Daniels, 1999; Bolstad and Manda, 2001; Gelman, 2004).

A comprehensive review of methods of eliciting *noninformative* priors are given by Kass and Wasserman (1996), Gelman (2004) and references therein. Gelman (2004) constructed a new folded-noncentral t family of conditionaly conjugate priors for hierarchical standard deviation parameters. Conjugate priors lead to proper posteriors or tractable Gibbs conditionals. Unfortunately, propriety of the Gibbs conditionals does not necessarily imply propriety of the complete posterior distribution (Hobert and Casella, 1996). Improper priors often lead to improper posteriors. Sun, *et al* (2001) investigated conditions leading to the propriety of the posterior in hierarchical linear mixed model when an improper prior was specified for fixed effects and proper prior for variance components, and improper priors for both fixed and variance components. The authors found that the posterior is proper if a vague prior is specified for fixed effects and proper prior for variance components. Bolstad (1997) specified conjugate priors in the hierarchical normal model reducing Gibbs conditionals to simpler forms that can be sampled directly. Hobert and Casella (1996) suggest specifying normal prior with large variances for fixed effects and inverse gamma prior with small parameter values for the variance components. The conditions of the propriety of the posterior distribution discussed thus far are special cases of a unified treatment of necessary and sufficient conditions for propriety of the posterior distribution discussed by Sun, *et al* (2001).

4.3.5 Practical implementation issues

The theoretical formulation of MCMC methods with a *long-run* distribution as the target posterior distribution is well founded (Tierney, 1994). A series of contentious issues arise though in the implementation of the Markov chain. These issues include among others determining the sample size, determination of *burn-in* length, single or multiple chains, and determination of starting points. A roundtable discussion by experts in the field considered some of these contentious issues (Kass, *et al* 1998).

The standard error of the sample mean $h(\theta)$ from an independent and identically distributed $(i.i.d)$ sample of size N from $g(\theta|y)$ is $\sigma/\sqrt{N}$ where σ is the posterior standard deviation of θ. If a reasonable estimate of σ is available, then N can easily be estimated. The iterates $\{\theta^{(r)}\}$ from an MCMC methods are correlated. Because of this correlation one needs larger samples than would be required if samples were independent and computing standard deviation is rather complicated. If the series can be approximated by a *first-order autoregressive* process, then the asymptotic standard deviation of the sample mean is given by

$$\frac{\sigma}{\sqrt{N}}\sqrt{\left(\frac{1+\rho}{1-\rho}\right)}$$

where ρ is the autocorrelation estimate of $\{h(\theta^{(r)})\}$. The estimate of the asymptotic standard deviation can be used to estimate the required sample size if a reasonable estimate of ρ is available. Autocorrelation can be reduced by subsampling the chain. However, this will depend on the cost of sampling from $g(\theta|y)$ (Geyer, 1992). An informal check can be attained by running parallel chains from independent starting points and compare $\bar{h}$. If they do not adequately agree, N must be increased. Raftery and Lewis (1992) proposed a method of estimating N using cumulative normal distribution.

Determining the *burn-in* period involves identifying the length of initial sample to be discarded on the basis that the chain has not reached its stationary distribution. That is, identifying the minimum point after which the sample can be claimed to come from $g(\theta|y)$ and from which the effect of the starting points is negligible. Visual inspection of the iterates $\{\theta^{(r)} : r = 1, \cdots, N\}$ is the most obvious and commonly used method of detecting the *burn-in* period. Other formal techniques have been proposed (Gelman and Rubin, 1992; Raftery and Lewis, 1992). Geyer (1992) suggested identification of *burn-in* using autocovariances. The arguments are that discarding initial sample will have minimal effect on inference since when the chain has been running long enough the *burn-in* iterates will constitute a small percentage ($\approx 1\%$) of the total sample (Geyer, 1992). Formal techniques of determining the

burn-in often make use of Monte Carlo output analysis. Raftery and Lewis (1996) outlined a way of determining *burn-in* for a single run which is also a diagnostic tool. Gelman (1996) suggests using multiple runs with overdispersed starting distribution. Gelman's idea is to monitor the chains to a point where they all stabilize to one value, thus having forgotten their respective starting points.

The other contentious issue discussed is whether to run one chain or multiple chains to achieve valid inference (Geyer, 1992; Gelman and Rubin, 1992). Both approaches have their advantages and disadvantages. Theoretically, running one chain seems more efficient because only one burn-in phase is involved and less values are discarded. Multiple chains have an advantage of transversing a wider parameter space. Gelman and Rubin (1992) recommended sampling the starting points from an overdispersed distribution which is a close approximation of the target distribution. The resulting sample from parallel chains is closer to being $i.i.d$ than a large sample from one chain. However, running multiple chains can pose a heavy computational burden. Geyer (1992) provides theoretical background related to the use of *autocovariance* to ascertain convergence in single run. See Cowles and Carlin (1996) for an extensive expository of diagnostic tools for MCMC output.

4.4 Summary

The advents of MCMC sampling methods make possible the use of flexible Bayesian models that would have previously been almost impossible to fit. Applied statisticians can now formulate more realistic models. Flexibility of Bayesian sampling techniques enables applied statisticians to tailor the statistical model to the problem at hand. Simulation methods result in samples from the posterior distribution of parameters. In this way, exploratory data analysis techniques can be used to critically explore the features of the posterior distribution. However, implementation of MCMC sampling methods is computationally involved in terms of coding the

method, generating samples, and storing and processing the results.

The M-H algorithm appeared in the literature (Metropolis, *et al* 1953; Hastings, 1970) much earlier than the Gibbs sampler (Geman and Geman, 1984). But, the Gibbs sampler has gained much popularity in the statistical community due to Gelfand and Smith (1990) who noted its potentials. Implementing the Gibbs sampler is relatively simpler than M-H algorithm, which possibly contributes to its popularity. Nevertheless, both methods complement each other. Direct sampling methods are easy to implement. However, they require cumbersome exploitation of the posterior density. Usually, they are only used to sample single dimensional nodes in the larger MCMC sampling scheme when those nodes are non-conjugate.

The sample based Bayesian procedures (MCMC) are conceptually attractive. If there is an abundance of data, likelihood inference based on the EM algorithm and the Bayesian inference will give similar results. However, Bayesian inference is valid even when data are sparse. The advantage of the Bayesian approach is the possibility of including informative priors. This allows external information to be added to the model in a coherent way. However, there is still a difficulty in accessing convergence in the Bayesian MCMC sampling methods which is not the issue in the EM algorithm as the likelihood monotonically increases. However, as discussed in Chapter 2 the EM algorithm is also only guaranteed to converge to at least a local maximum and convergence to a global maximum in the presence of multiple maxima becomes an issue.

Chapter 5

A Bayesian analysis of time until HIV infection

5.1 Introduction

The model framework presented in Chapter 3 is already hierarchical and fully specified from the frequentist point of view and the model parameters have been estimated using the EM algorithm. From the Bayesian perspective, we also need to specify priors for the fixed effects vector β, the constant baseline hazard λ_0 and the hyperparameter α before the model is fully specified. The directed graph for the fully specified Bayesian model is shown in Figure 5.1.

The prior for the vector of fixed effects β is assumed multivariate normal with mean vector $\mathbf{d_0} = \mathbf{0}$ and diagonal covariance matrix $\mathbf{\Sigma_0} = \upsilon_0\mathbf{I}$, where υ_0 is a suitably chosen large number. In the assumption of the proportional hazards model, fixed effects represent logarithm of the relative risk and thus will not be far away from zero. The normal priors with large variances are almost identical to the flat priors for all practical purposes in terms of their effect on the marginal posteriors of the regression coefficients. The choice of such proper priors ensures valid approximations to the posterior distribution.

The available knowledge about the extent of HIV infection in South Africa has been gained through an annual HIV surveillance among women presenting for the first time in the antenatal clinics. The province of KwaZulu-Natal, where this study is conducted, has the highest antenatal prevalence of HIV compared to other provinces. In 2000, the antenatal prevalence of HIV infection was 36% rising from 32.5% in 1998 (Department of Health, 2001). These estimates provide information used to determine parameters of the prior for λ_0 used in this analysis. The baseline hazard is a rate and the interval considered is between 8 to 10 years with time measured in months. Therefore, the suitable prior should represent this. A proper prior is assumed for the baseline hazard. We assume a Gamma(ξ_0, ζ_0) prior distribution $\xi_0 = 1$ and $\zeta_0 = 20$. This gives 0.05 and 0.05 as the prior mean and standard deviation for λ_0.

The sexual network frailties b_i act multiplicatively on the baseline hazard and take only positive values. The frailties represent relative risks and thus should have mean 1. We model them as independent random variates from a Gamma(α, α) distribution. Therefore, the relative risk for sexual networks has mean 1 and variance $1/\alpha$. The unit mean constraint on the relative risk ensures that the sexual network effect represents deviations from the population average risk. To accomplish model specification, a prior distribution for the variance component is required.

The standard noninformative prior for variance component σ^2 is the density $f(\sigma^2) \propto 1/\sigma^2$. The equivalent noninformative prior for generic precision component τ is the density $f(\tau) \propto 1/\tau$. This prior is improper and may lead to an improper posterior. To avoid this potential pitfall, we specify a proper prior for the precision component. It seems reasonable to specify a prior density that is positive, finite and decreases monotonically in τ. The prior of this nature favours models with smaller magnitude of frailty effect. We used a Gamma (ν_0, κ_0) prior for α where $\nu_0 = 1$

and $\kappa_0 = 1$. This is equivalent to variance prior having mode just below 1, which leads to a posterior marginal density of frailty variance with mode depending on the observed data.

The modelling framework proposed here is related to the work of Clayton (1991), Gustafson (1997) and Bolstad and Manda (2001). All these authors discuss Bayesian models for hierarchical multivariate survival data with *precisely known* failure times. However, the context and the modelling approach considered in this thesis differs from theirs in several aspects including baseline hazards specification, variance components and the extension of estimation in correlated interval-censored data. Clayton (1991) treated the increment of the cumulative baseline hazards as independent gamma variates. In this way, conditional on β, the cumulative baseline hazards posterior follows an independent gamma process (Kalbfleisch, 1978). Gustafson (1997) used similar approach in the implementation of Cox partial likelihood by integrating out the baseline hazards with respect to the gamma process. Furthermore, Gustafson (1997) assumed a log-normal frailty distribution. Bolstad and Manda (2001) specified piecewise constant baseline hazards in a three-way multilevel model of child mortality. The parameters of the piecewise constant baseline hazards were absorbed into the fixed effects, and estimated with them.

Constant baseline hazards are assumed in this analysis. Baseline hazards and fixed effects are sampled sequentially. Furthermore, we consider an important aspect of sampling interval-censored failure times conditional on clinical examination times, frailties and observed data. Sinha and Dey (1997) reviewed a number of Bayesian methods of analysing survival data. Their review also covers interval-censored survival data. However, extensions of semiparametric Bayesian models for analysis of multivariate survival data using frailties (Clayton, 1991) to interval-censored data are not immediate. Thus, we cast the problem of multivariate interval-censored survival data as a *missing data* problem. We are not aware of any previous im-

plementation of Bayesian hierarchical model for correlated interval-censored data through augmentation of infection times conditional on clinical examination times, frailties and observed data.

5.2 The joint posterior distribution

The full Bayesian model considered is the proportional hazards frailty model:

$$
\begin{aligned}
& h(y_{ij}|\beta, b_i) = b_i \lambda(y_{ij}) e^{\beta' X_{ij}} \qquad \text{where } \lambda(y_{ij}) = \lambda_0 \\
& \beta \sim \text{MVN}(d_0, \Sigma_0) \\
& \lambda_0 \sim \text{Ga}(\xi_0, \zeta_0) \\
& b_i \sim \text{Ga}(\alpha, \alpha) \\
& \alpha \sim \text{Ga}(\nu_0, \kappa_0) \\
& t_{ij} \sim \text{Exp}(b_i \lambda_0 \exp(\beta' X_{ij}))
\end{aligned}
$$

where y_{ij} denotes both the infection time for someone infected with HIV and right-censoring time for someone uninfected with HIV at the end of the study or lost to follow-up, $y_{ij} \in [v_{ij,k}; v_{ij,k+1}]$. The MVN$(d, s)$ generically denotes a multivariate normal distribution with mean vector d and covariance matrix s; Ga(d, s) generically denotes a gamma distribution with mean d/s and variance d/s^2. Exp(d) generically denotes an exponential distribution with parameter d. Here, we assume independence between $\{y_{ij}\}$ given all other parameters of the model; between $\{b_i\}$ given hyperparameter α and between β and α.

The joint posterior distribution of parameters, hyperparameters and the data is given by

$$
\begin{aligned}
& f(\text{data}, \beta, \lambda_0, t_{ij}, b_i, \alpha) \\
= \; & f(\beta) f(\lambda_0) f(\alpha) \times \\
& \left\{ \prod_{i=1}^{I} f(b_i|\alpha) \prod_{j=1}^{J_i} L_i(y_{ij}|\beta, \lambda_0, b_i) \right\}.
\end{aligned}
$$

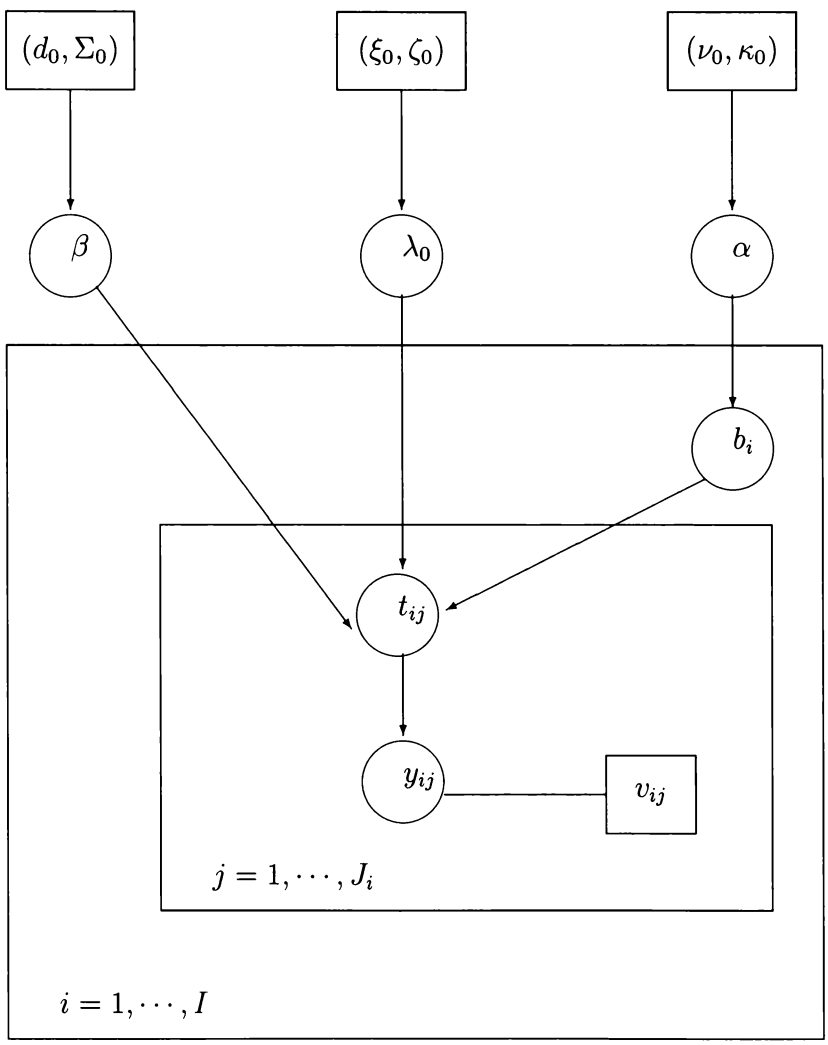


Figure 5.1: *The directed acyclic graphical model representation of migration data*

After inserting the relevant quantities, the joint posterior density function becomes

$$\begin{aligned} f(\text{data}, \beta, \lambda_0, t_{ij}, b_i, \alpha) &= \\ (2\pi)^{\frac{-p}{2}} |\Sigma_0|^{\frac{-1}{2}} \exp\{-\frac{1}{2}(\beta - d_0)'\Sigma_0^{-1}(\beta - d_0)\} &\times \\ \frac{\zeta_0^{\xi_0}}{\Gamma(\xi_0)} \lambda_0^{\xi_0 - 1} e^{-\lambda_0 \zeta_0} \times \frac{\kappa_0^{\nu_0}}{\Gamma(\nu_0)} \alpha^{\nu_0 - 1} e^{-\alpha \kappa_0} &\times \\ \left\{ \prod_{i=1}^{I} \frac{\alpha^\alpha}{\Gamma(\alpha)} b_i^{\alpha-1} e^{-\alpha b_i} \prod_{j=1}^{J_i} \left[e^{-b_i \Lambda(t_{ij}|X_{ij})} \, b_i \, \lambda_0 \, e^{\beta' X_{ij}} \right]^{\delta_{ij}} \times \left[e^{-b_i \, \Lambda(v_{ij,k}|X_{ij})} \right]^{1-\delta_{ij}} \right\} & \end{aligned} \tag{5.1}$$

where p is the number of fixed effects parameters. Bayesian statistical inference requires the joint posterior density of all parameters and hyperparameters given the data. In our model, the posterior density cannot be obtained analytically. The Gibbs sampler can be used to obtain samples of parameters from the posterior density using the conditional distribution of each node given all the other nodes. Most of these conditional distributions are relatively tractable and they can be easily sampled. However, some conditional distributions are intractable, and other methods for sampling from intractable conditional distributions will be used.

5.3 Gibbs conditional distributions

The required Gibbs conditional distributions are for sexual network random effects $f(b_i|\text{data}, \beta, \lambda_0, t_{ij}, \alpha)$, infection time $f(t_{ij}|v_{ijk} < t_{ij} \leq v_{ijk+1}, \text{data}, \beta, \lambda_0, b_i, \alpha)$, sexual network random effect inverse variance $f(\alpha|\text{data}, \beta, \lambda_0, t_{ij}, b_i)$, baseline hazards $f(\lambda_0|\text{data}, \beta, t_{ij}, b_i, \alpha)$ and fixed effects $f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha)$. In a hierarchical model, the conditional distribution of one node given all the other nodes is proportional to the prior distribution of that node times the conditional distribution of all its direct child nodes and co-parent nodes. The following sections present the Gibbs conditional nodes required in the analysis.

5.3.1 Sexual network random effects

The sexual network random effects conditional distribution is calculated as

$$\begin{aligned} f(b_i|\text{data}, \beta, \lambda_0, t_{ij}, \alpha) &\propto \frac{\alpha^\alpha}{\Gamma(\alpha)} b_i^{\sum_{j=1}^{J_i}\delta_{ij}+\alpha-1} e^{-b_i\left[\alpha+\sum_{j=1}^{J_i}\delta_{ij}\Lambda(t_{ij}|X_{ij})+\{1-\delta_{ij}\}\Lambda(v_{ij,k}|X_{ij})\right]} \\ &\propto b_i^{\sum_{j=1}^{J_i}\delta_{ij}+\alpha-1} e^{-b_i\left[\alpha+\sum_{j=1}^{J_i}\delta_{ij}\Lambda(t_{ij}|X_{ij})+\{1-\delta_{ij}\}\Lambda(v_{ij,k}|X_{ij})\right]} \end{aligned}$$

which we recognize as the *kernel* of a gamma distribution with shape $\alpha + \sum_{j=1}^{J_i}\delta_{ij}$ and inverse scale $\alpha + \sum_{j=1}^{J_i}[\delta_{ij}\Lambda(t_{ij}|X_{ij}) + (1-\delta_{ij})\Lambda(v_{ij,k}|X_{ij})]$. Hence this can be sampled directly.

5.3.2 Infection time

The conditional distribution of the HIV infection time is

$$\begin{aligned} f(t_{ij}|v_{ij,k} < t_{ij} \leq v_{ij,k+1}, \text{data}, \beta, \lambda_0, b_i, \alpha) &= \frac{f(t_{ij}|\text{data}, \beta, \lambda_0, b_i)}{\int_{v_{ij,k}}^{v_{ij,k+1}} f(t|\text{data}, \beta, \lambda_0, b_i, \alpha)dt} \\ &= \frac{e^{-H(t_{ij}|b_i,X_{ij})} \times h(t_{ij}|b_i, X_{ij})}{e^{-H(v_{ij,k}|b_i,X_{ij})} - e^{-H(v_{ij,k+1}|b_i,X_{ij})}} \\ &= \frac{\exp(-b_i\lambda_0 t_{ij} e^{\beta' X_{ij}}) \times b_i\lambda_0 \exp(\beta' X_{ij})}{S(v_{ij,k}|b_i, X_{ij}) - S(v_{ij,k+1}|b_i, X_{ij})} \\ &\propto \exp(-t_{ij} b_i \lambda_0 e^{\beta' X_{ij}}) \end{aligned}$$

which we recognize as the *kernel* of a gamma distribution with shape 1 and inverse scale $b_i\lambda_0 e^{\beta' X_{ij}}$. Such a gamma distribution is equivalent to an exponential distribution with parameter $b_i\lambda_0 e^{\beta' X_{ij}}$. Hence this node can also be sampled directly on condition that the sampled value $t_{ij} \in (v_{ij,k}, v_{ij,k+1}]$.

5.3.3 Random effects inverse variance

The conditional distribution of the sexual network random effects inverse variance is

$$\begin{aligned} f(\alpha|\text{data}, \beta, \lambda_0, b_i) &\propto \frac{\kappa_0^{\nu_0}}{\Gamma(\nu_0)} \alpha^{\nu_0-1} e^{-\alpha\kappa_0} \times \prod_{i=1}^{I} \frac{\alpha^\alpha}{\Gamma(\alpha)} b_i^{\alpha-1} e^{-\alpha b_i} \\ &\propto \alpha^{\nu_0-1} \times \left(\frac{\alpha^\alpha}{\Gamma(\alpha)}\right)^I \left(\prod_{i=1}^{I} b_i\right)^{\alpha-1} e^{-\alpha\left[\kappa_0+\sum_{i=1}^{I} b_i\right]}. \end{aligned}$$

The full conditional does not simplify to any standard distribution that can be sampled directly. Thus, we require methods for sampling from an arbitrary conditional distribution. It turns out that the full conditional distribution is a simple *log-concave* distribution in α and can be sampled efficiently using the *adaptive-rejection sampling* scheme (Gilks and Wild, 1992).

5.3.4 Baseline hazard

The baseline hazards conditional distribution is computed as

$$\begin{aligned}
& f(\lambda_0|\text{data}, \beta, t_{ij}, b_i, \alpha) \\
\propto\ & \frac{\zeta_0^{\xi_0}}{\Gamma(\xi_0)} \lambda_0^{\xi_0 - 1} e^{-\lambda_0 \zeta_0} \\
\times\ & \lambda_0^{\sum_{i=1}^{I} \sum_{j=1}^{J_i} \delta_{ij}} e^{-\lambda_0 \left[\sum_{i=1}^{I} \sum_{j=1}^{J_i} \delta_{ij}\, b_i\, t_{ij}\, e^{\beta' X_{ij}} + (1-\delta_{ij})\, b_i\, v_{ij,k}\, e^{\beta' X_{ij}} \right]} \\
\propto\ & \lambda_0^{\xi_0 - 1 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \delta_{ij}} e^{-\lambda_0 \left[\zeta_0 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \delta_{ij}\, b_i\, t_{ij}\, e^{\beta' X_{ij}} + (1-\delta_{ij})\, b_i\, v_{ij,k}\, e^{\beta' X_{ij}} \right]}
\end{aligned}$$

which we recognize as the *kernel* of a gamma density with the shape parameter $\xi_0 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \delta_{ij}$ and scale parameter $\zeta_0 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} [\, \delta_{ij}\, b_i\, t_{ij}\, e^{\beta' X_{ij}} + (1 - \delta_{ij})\, b_i\, v_{ij,k}\, e^{\beta' X_{ij}}]$. Hence this node can also be sampled directly.

5.3.5 Fixed effects

The full conditional distribution of the fixed effects is

$$\begin{aligned}
& f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha) \\
\propto\ & (2\pi)^{\frac{-p}{2}} |\Sigma_0|^{\frac{-1}{2}} \exp\{-\frac{1}{2}(\beta - d_0)' \Sigma_0^{-1} (\beta - d_0)\} \\
\times\ & \prod_{i=1}^{I} \prod_{j=1}^{J_i} \left[e^{-b_i \Lambda_0(t_{ij}|X_{ij})}\, b_i\, \lambda_0\, e^{\beta' X_{ij}} \right]^{\delta_{ij}} \times \left[e^{-b_i \Lambda_0(v_{ij,k}|X_{ij})} \right]^{1-\delta_{ij}}
\end{aligned}$$

which does not simplify to any standard distribution. A Taylor series expansion of $f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha)$ centered at the posterior mode $\tilde{\beta}$ leads to

$$
\begin{aligned}
& \log f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha) \\
= \; & \log f(\tilde{\beta}|\text{data}, \lambda_0, t_{ij}, b_i, \alpha) + (\beta - \tilde{\beta})' \frac{\partial \log f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha)}{\partial \beta}|_{\beta=\tilde{\beta}} \\
+ \; & \frac{1}{2}(\beta - \tilde{\beta})' \frac{\partial^2 \log f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha)}{\partial \beta^2}|_{\beta=\tilde{\beta}}(\beta - \tilde{\beta}) \\
= \; & \log f(\tilde{\beta}|\text{data}, \lambda_0, t_{ij}, b_i, \alpha) + \frac{1}{2}(\beta - \tilde{\beta})' \frac{\partial^2 \log f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha)}{\partial \beta^2}|_{\beta=\tilde{\beta}}(\beta - \tilde{\beta}).
\end{aligned}
\tag{5.2}
$$

The linear term in (5.2) is zero because the log posterior density has zero derivative at its mode. The remainder terms of higher order in the Taylor series expansion fade in importance relative to the quadratic term when β is close to $\tilde{\beta}$ and the sample size is large. If we consider (5.2) as a function of β, the first term is a constant whilst the second term is proportional to the logarithm of a normal density. This yields the following approximation:

$$f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha) \approx N(\tilde{\beta}, [I(\tilde{\beta})]^{-1})$$

where $I(\tilde{\beta})]$ is the observed information

$$I(\tilde{\beta}) = -\frac{\partial^2 \log f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha)}{\partial \beta^2}|_{\beta=\tilde{\beta}}.$$

Asymptotically, $f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha)$ can be approximated by a normal distribution with mean being the posterior mode and covariance matrix equal to minus the inverse of the second derivative of the log posterior evaluated at the posterior mode $\tilde{\beta}$. If a flat prior is assumed for β, the posterior mode can be replaced by the ML estimate $\hat{\beta}$ and log posterior density by log likelihood function. The observed information $I(\tilde{\beta})$ becomes $I(\hat{\beta})$, the Fisher information matrix evaluated at the ML estimator. Therefore, samples from $f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha)$ can be easily generated by calculating $\hat{\beta}$ and $[I(\hat{\beta})]^{-1}$ using the likelihood methods shown in Chapter 3. These values are used for the multivariate normal proposal distribution. To ensure that the samples obtained come from the specified conditional posterior distribution, we inserted a Metropolis step where candidates from the multivariate normal

proposal distribution are either accepted or rejected. The acceptance rate for candidates was about 54%, which was well within 30% and 70%, the recommended acceptance rate (Raftery and Lewis, 1996). The high acceptance rate indicates that the multivariate normal proposal distribution is a good initial approximation to the actual conditional posterior.

The computer code implementing the MCMC simulations was written and implemented in Microsoft Visual C++ Version 6.0. Microsoft Visual C++ Version 6.0 does not, however, have subroutines for generating random samples from the standard distributions. The software only has a Uniform(0,1) generator. Various functions were written to sample from these distributions building from a Uniform(0,1) generator. For example, to sample β from $\text{MVN}_p(\hat{\beta}, [I(\hat{\beta})]^{-1})$ we wrote a function that generates random variables $\mathbf{z}$ from a standard normal distribution. We performed a Cholesky decomposition of covariance matrix $[I(\hat{\beta})]^{-1} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is the lower triangular matrix. Samples from $\text{MVN}_p(\hat{\beta}, [I(\hat{\beta})]^{-1})$ were generated as $\hat{\beta} + \mathbf{L}\mathbf{z}$ (Ripley, 1987).

5.4 Application to the data

The Markov chain Monte Carlo (MCMC) analysis described in the previous section was implemented for the model with and without sexual network frailties, the same as was done in Chapter 3 using the EM algorithm. The frailty model required estimation of a total of 526 parameters: 9 fixed effects, 1 baseline hazard, 339 sexual network specific random effects, 176 infection times and the inverse scale for sexual network frailty distribution.

Five parallel chains were run from independent starting points. When we iterated through the MCMC sampling scheme at an equal rate for all parameters we found that successive values for b_i, t_{ij}, α and λ_0 were highly correlated. The successive fixed

effects values were much less correlated. However, the fixed effect sampling scheme involved an EM estimation of ML estimates and calculation of Fisher information for the proposal density for the M-H step. This was computationally intensive. Because of this we modified the iteration scheme to iterate through b_i, t_{ij}, α and λ_0 five times for each draw of β, which greatly improved efficiency.

Monitoring of all parameters was impractical. We monitored all the fixed effects, the baseline hazard, the sexual network inverse scale, some of the sexual network random effects and some individual infection times from all five chains. We found no evidence from the multiple chains suggesting that the monitored nodes were not converging to the same node. The median and the 97.5% percentile of Gelman and Rubin's (1992) scale reduction factor (GR) for each monitored variable were calculated. GR compares the *between chain* variation to the *within chain* variation and should be close to one if the *burn-in* time has been sufficiently long for the Gibbs sampler to be nearly convergent to the target posterior distribution. From the first chain, we calculated the Z-test of equality based on the arguments of Geweke (1992). In this test Geweke proposed a simple method based on time series ideas stating that if the chains were in equilibrium, the means of the first 10% and the last 50% of the iterates should be nearly equal. Therefore, the diagnostic computes the Z-test of the hypothesis of equality between two means. We first ran five parallel chains from independent starting points for $2n = 2000$. The GR statistics for all other parameters except baseline hazards were reasonably close to one, Figures 5.2 and 5.3. The first-order autocorrelations AR(1) for all parameters were quite substantial. The autocorrelation plots showed high degree of autocorrelation even after lag 30 for some parameters. We increased the number of iterations to $2n = 4000$. Output analysis of $2n = 4000$ simulated observations resulted in GR statistics extremely close to 1 for all variables indicating substantial improvement in the convergence of the estimates (Kass, *et al* 1998). Thus, we took 2 000 iterations as satisfactory *burn-in* time. We simulated a further 38 000 values from each chain and took every

Table 5.1: Geweke convergence diagnostics

	The Z-scores for each chain and parameter				
Parameter	Chain 1	Chain 2	Chain 3	Chain 4	Chain 5
Baseline hazard	-1.280	-0.333	1.490	-0.054	-1.760
Migrant men	1.020	-0.542	0.659	-0.971	0.317
Part. of migrant men	1.260	-0.614	0.196	0.472	1.300
Non-migrant men	1.560	-1.404	1.530	1.250	-0.180
Age:18 to 24	-0.128	1.130	-1.280	-0.806	1.130
Age:25 to 34	0.875	0.510	-1.050	-0.924	-1.130
Current partners	-0.978	0.560	-1.170	0.556	2.600
Lifetime partners	0.914	-0.163	-2.140	0.670	0.217
Syphilis	-0.840	-2.860	-0.877	-1.950	1.950
Other STIs	-0.967	-0.192	2.290	2.280	0.853
Frailty variance	1.140	-0.228	-0.011	0.267	-0.861

100th value after *burn-in* time. Therefore, the result was 2 000 nearly independent simulated observations from the posterior distribution. The autocorrelation estimates for the final simulated values were near zero. The Z-scores from all chains showed reasonable convergence, Table 5.1. Figures 5.2 and 5.3 are the trace plots for simulated parameters from the posterior density developed in Section 5.2. The trace plots show consistent random fluctuations around the convergent value.

The histograms for fixed effects parameters with an overlaid normal curve are presented in Figure 5.4. The histograms are fairly symmetric as would be expected. Figure 5.5(a) shows the marginal posterior distribution for the baseline hazard. Table 5.2 presents the results of the baseline hazards, fixed effects and frailty variance from the Gibbs sampler. The estimates of the fixed effects and baseline hazards are similar for all practical purposes to the respective modes obtained from the EM algorithm. However, the estimates from the Gibbs sampler are more variable than

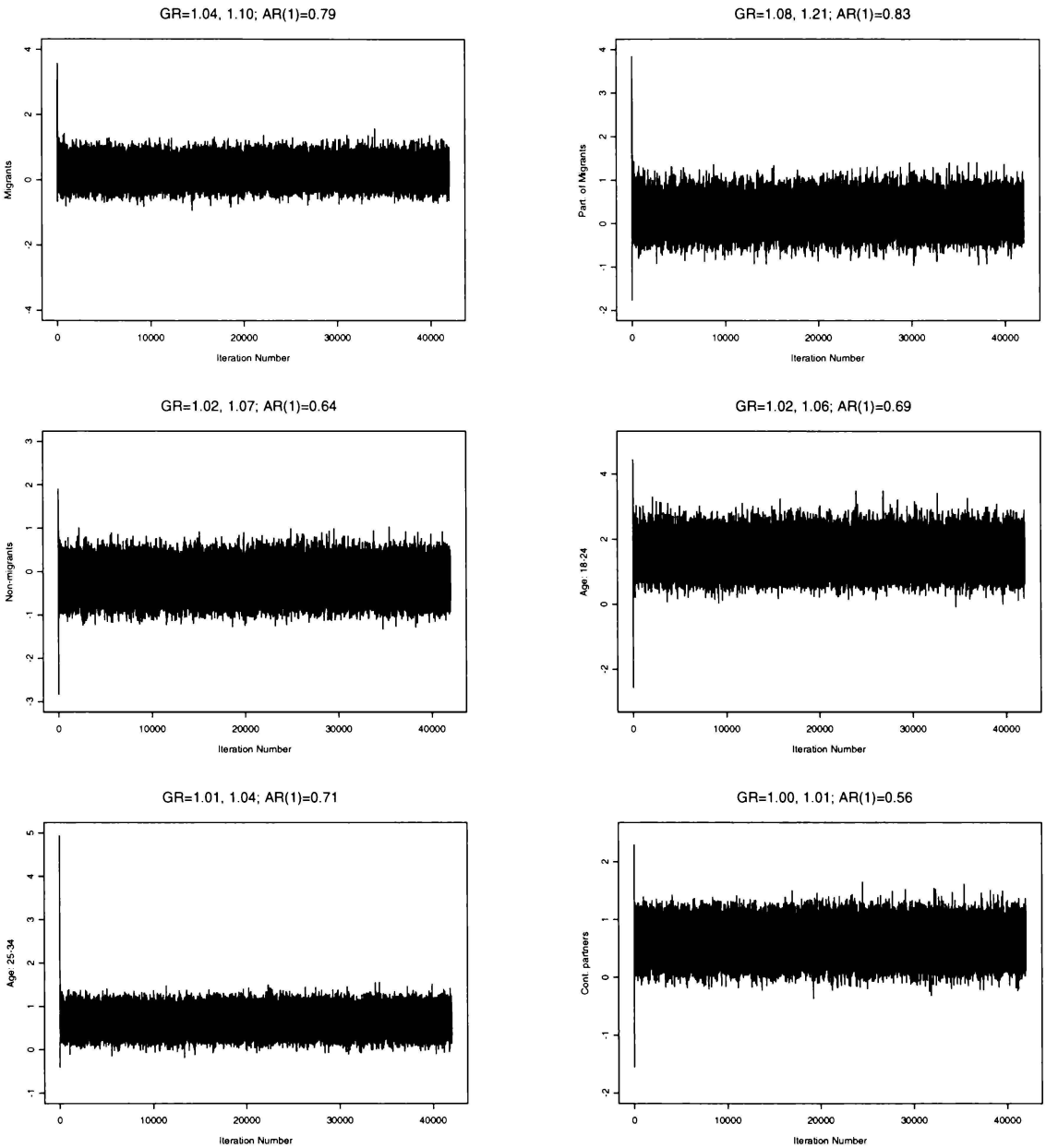


Figure 5.2: *Convergence monitoring trace plots for selected fixed effects. In each panel, all five independent chains are plotted. Included is the mean and 97.5% percentile of GR statistic from the first 1000 observations. Also included is the first-order autocorrelation AR(1) estimated from the first chain.*

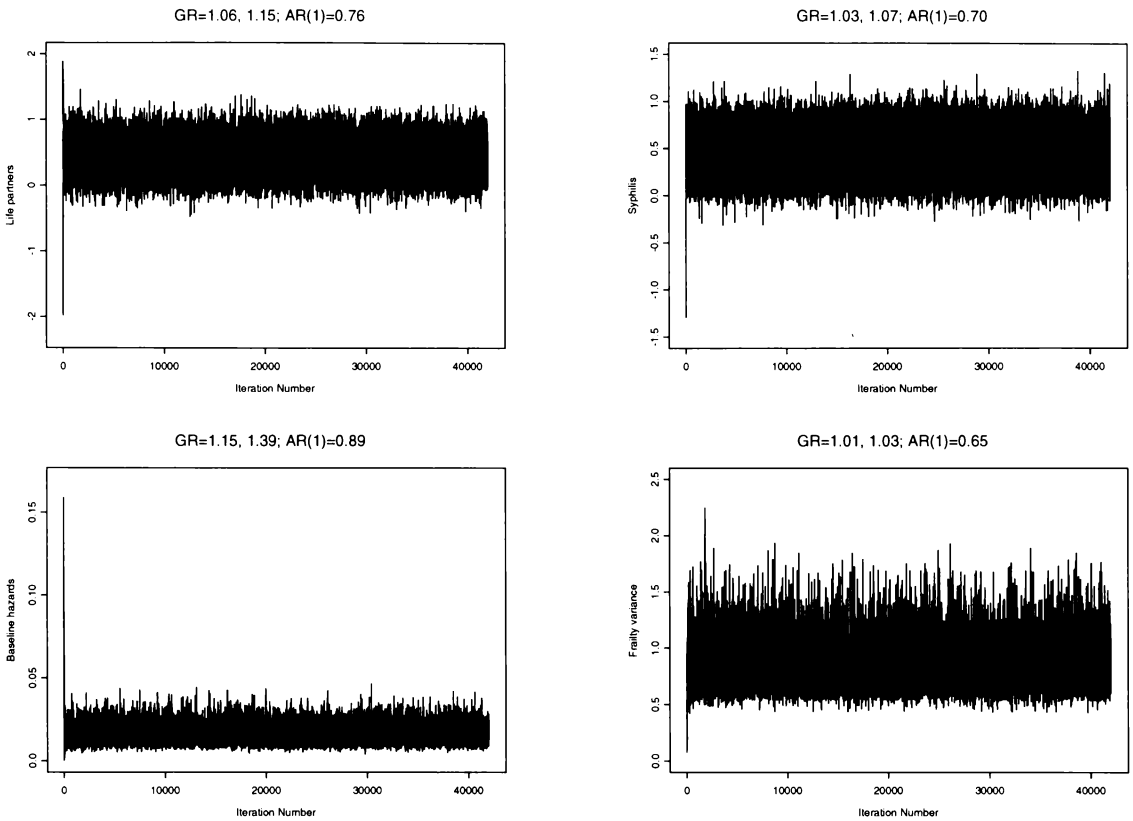


Figure 5.3: *Convergence monitoring trace plots for some fixed effects, baseline hazards and frailty variance. In each panel, all five independent chains are plotted. Included is the mean and 97.5% percentile of GR statistics for the first 1000 observations. Also included is the first-order autocorrelation AR(1) estimated from the first chain.*

those from the EM algorithm. The estimate of sexual network frailty variance from the Gibbs sampler is quite large compared to the estimate obtained from the EM algorithm. The posterior median and mean is 0.788 and 0.812 respectively. The 95% credible interval for sexual network frailty variance is (0.614, 1.120). The distribution for the frailty variance is shown in Figure 5.5(b). In the EM algorithm, the mode of the sexual network frailty variance was estimated to be 0.462.

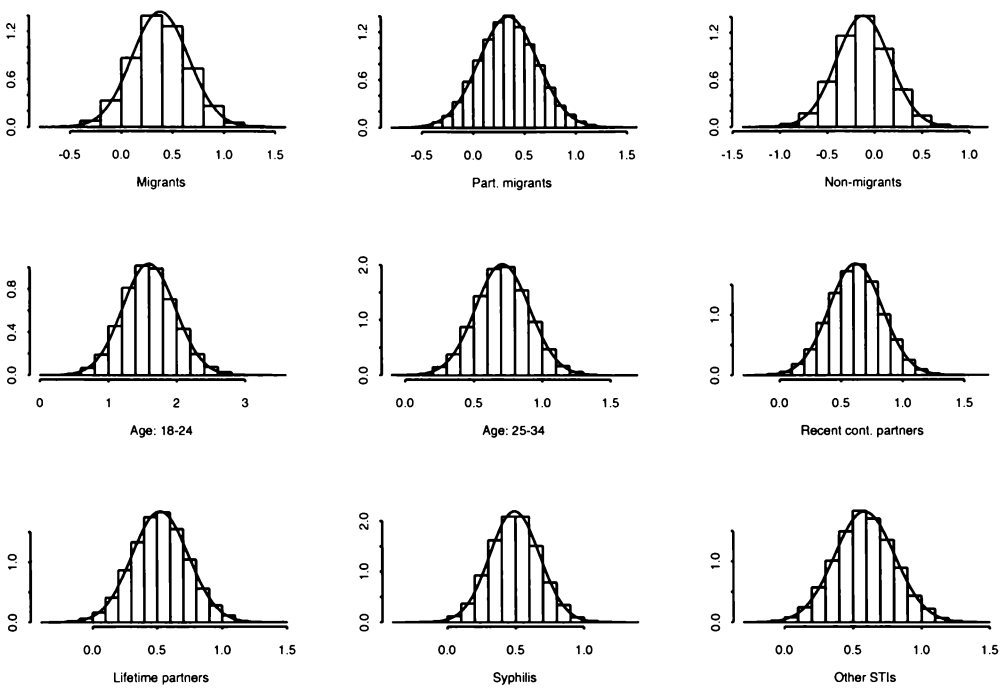
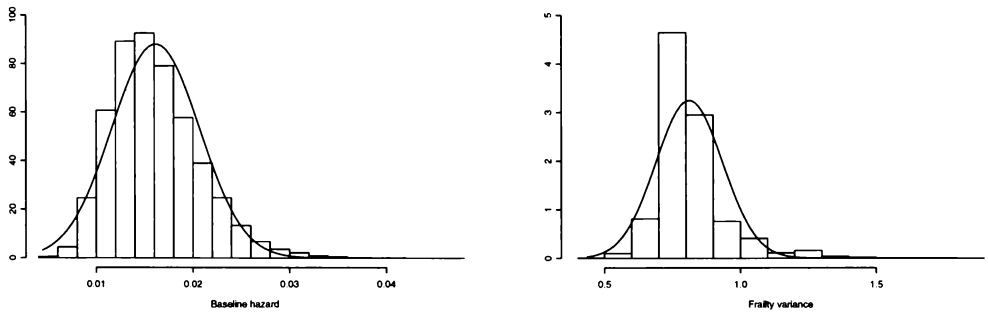


Figure 5.4: *Histograms of the fixed effects parameters.*

5.5 Conclusion

We have successfully implemented the Gibbs sampler to investigate the risk factors associated with HIV infection among people in networks of sexual partnerships involving migrant and non-migrant men and their non-migrant partners from a rural health district of South Africa. The approach focussed on reducing the complex posterior likelihood for correlated interval-censored data, whose direct sampling is not very straight forward (Sinha and Dey, 1997), to a simpler correlated right-censored data problem. Fitting Bayesian frailty models to interval-censored data likelihoods, (for example Finkelstein, 1986; Huang and Wellner, 1997) presents analytical challenges for computing the posterior distribution. The Gibbs sampler implemented here provides full Bayesian inference without requiring evaluation of complex integrals as was the case with the EM algorithm.

The Gibbs sampler yields a sample of parameters and hyperparameter obtained



(a) Baseline hazard (b) Frailty variance

Figure 5.5: *The marginal posterior distributions*

from a well-defined Markov process such that the values are thought of as realisations from the corresponding marginal posterior distribution and can be explored over a range of values. The Gibbs conditionals simplify to two iterative steps involving *imputation step* which draws $b_i^{(r)}$ and $t_{ij}^{(r)}$ from the conditional predictive distribution $f(b_i|\text{data}, \beta, \lambda_0, t_{ij}^{(r-1)}, \alpha)$ and $f(t_{ij}|v_{ijk} < t_{ij} \leq v_{ijk+1}, \text{data}, \beta, \lambda_0, b_i^{(r)}, \alpha)$ respectively, and a *posterior step* which draws $\theta^{(r)} = \{\beta^{(r)}, \lambda_0^{(r)}, \alpha^{(r)}\}$ from conditional posterior distribution $f(\theta^{(r)}|\text{data}, b_i^{(r)}, t_{ij}^{(r)})$. The two iterative steps can be viewed as the stochastic counterparts to the E-step and M-step of the EM algorithm. Under broad regularity conditions, the sequence $(b_i^{(r)}, t_{ij}^{(r)}, r = 1, 2, \cdots)$ converges to the joint posterior $f(b_i^{(r)}, t_{ij}^{(r)}|\text{data})$ and the sequences of their components converge to their marginal posteriors (Gilks, *et al* 1996).

Compared to the traditional ML estimation, Bayesian analysis is capable of not only incorporating information about frailties and infection time, but also uncertainties about available information. For example, the uncertainty about the true values of variance components is formally incorporated into the analysis through the choice of a plausible prior distribution. The fixed effects results from the Gibbs sampler are in good agreement with the corresponding posterior modes from the EM algorithm. The agreement between the modes of marginal posteriors and ML estimates is generally expected due to the specified proper prior for fixed effects which

is nearly flat in the region near zero (Harville, 1974). However, estimated standard deviations from the ML approach are severely biased downwards. The bias reflects the incapability of ML approach to correct for variability of unobserved frailties and infection time. Downward bias in standard deviations is highly undesirable because it provides false sense of security for the estimates. The frailty variance estimate from the EM algorithm also shows similar downward bias compared to the estimate from the Gibbs sampler. However, inference and conclusions from the Gibbs sampler were not markedly different from inference based on the EM algorithm. These comparative results have been published elsewhere (Zuma and Lurie, 2005).

Table 5.2: Frailty model estimates and credible intervals (CI)s from the Gibbs sampler

	Frailty model		95% CI	
Parameter	Mean	SD	2.5%	97.5%
Baseline hazard				
Constant	0.013	0.004	0.007	0.022
Migration status				
Migrant men	0.391	0.276	-0.156	0.911
Partners of migrant men	0.354	0.276	-0.204	0.886
Non-migrant men	-0.103	0.276	-0.616	0.440
Partners of non-migrant men[a]				
Age in years				
18 to 24	1.590	0.383	0.861	2.360
25 to 34	0.709	0.201	0.330	1.110
35 and above[a]				
Recent sexual contact partners				
Only one[a]				
More than one	0.609	0.216	0.174	1.020
Number of lifetime partners				
Only one[a]				
More than one	0.521	0.215	0.113	0.944
Syphilis				
0=Negative, 1=Positive	0.501	0.179	0.147	0.849
Status of other STIs				
0=Negative, 1=Positive	0.588	0.218	0.167	1.020
Frailty variance				
Sexual network	0.812	0.120	0.614	1.120

[a]Reference category

Chapter 6

Conclusion

6.1 Thesis theme

The thesis has introduced the concept of incorporating sub-groups of correlated sexual networks as random effects in the investigation of the effects of circular migration in the spread of HIV and other curable STIs in the rural Hlabisa district from northern KwaZulu-Natal, South Africa. The sexual network random effects formed part of unobserved data as has been done in various other applications involving correlated data. Another development of this thesis is the treatment of interval-censored HIV infection time as unobserved data. Then, the complete-data likelihood functions were developed which are compatible with estimation via the EM algorithm (Dempster, Laird and Rubin, 1977) and the Gibbs sampler (Geman and Geman, 1984). Finally, the thesis compares the frailty model results from the EM algorithm to those obtained from the Gibbs sampler.

The introductory chapter reviewed literature on epidemiology and reciprocal impact of HIV infection and STIs. The introductory chapter discusses migration and other factors as risk determinants of HIV/STIs. The chapter further outlined current statistical methods for analysing correlated data. Chapter 2 presented the theory behind the EM algorithm. The EM algorithm was implemented in the anal-

ysis of curable STIs. In Chapter 3 the EM algorithm, outlined in Chapter 2, was further used to analyse correlated interval-censored data where both sexual network frailties and interval-censored infection time formed the missing data used to facilitate the EM algorithm. This is in contrast to Chapter 2 where only the sexual network random effects constituted missing data. In both preceding chapters, the results of a standard and random effects models were compared.

In Chapter 4 we outlined the basics of Bayesian parameter estimation and of Markov chain Monte Carlo (MCMC) simulation techniques. Full Bayesian analysis of the proportional hazards frailty model with interval-censored HIV infection time was carried out in Chapter 5. The Gibbs sampler, an MCMC simulation technique, was used to attain full Bayesian inference of the model. Some Gibbs conditionals were intractable and required methods of sampling from a non-standard Gibbs conditional distribution.

6.2 Thesis conclusions

6.2.1 Substantive

The logistic mixed model suggests that *migration* of men is a risk factor of acquiring at least one STI (p-value=0.049). Migrant men and their rural female sexual partners are at marginally increased risk of STIs compared to non-migrant men and their rural female sexual partners. Being *never married* or having *first sexual intercourse* before the seventeenth birthday are associated with increased risk of being infected with at least one curable STIs, p-values=0.039 and 0.023 respectively. Recent *sexual contact* with more than one sexual partner increases the risk of STIs. Infection with *HIV* further increases the risk of contracting STIs, p-value=0.012. Provision of syndromic management and sexual behavioural education reduces the risk of subsequent transmission of curable STIs and hence HIV.

In the cross-sectional baseline investigation of the effects of *migration* on the risk of HIV among couples only, *migration* was identified as an important risk factor of HIV (Lurie, Williams, Zuma, *et al* 2003b). Migration is a risk factor not simply because returning migrant men infect their partners, but also because their rural female partners -including those who are partners of non-migrant men - are likely to become infected from outside their primary relationships. However, in the main analysis which included all sexual network sizes and corrected for correlation induced by clustering of sexual networks, migration did not appear to be a significant risk factor, Table 5.2. At this late stage of the epidemic, migration might be becoming less important due to the existing high rates of HIV infection in rural areas and ongoing spread of HIV within the rural areas. The risk of HIV is considerably high in ages between 18 and 24 years and decreases slightly in ages 25 and 34 years, Table 5.2. Recent *sexual contact* with more than one partner or having more than one *lifetime sexual* partners are associated with increased risk of HIV. Infection with *syphilis* or *other curable STIs* greatly increase the risk of HIV infection. The risk of HIV/STIs varies considerably across sexual networks.

The results of this thesis have important policy implications. Interventions aimed at combating the spread of HIV/STIs should extend further from focussing on individuals as social units to treating sexual networks as social units. Interventions have often been aimed at individual-level behavioural changes promoting condom use, fewer concurrent sexual partners and sexual abstinence. However, these approaches are of less benefit to women who are in weaker positions to negotiate safe sex or discourage their partners from having extra marital relationships. The urgently required changes in the policy include formulating specialized educational programs targeting HIV discordant partnerships. Interventions should enforce counselling, educational messages and treatment of STIs within sexual networks rather than only the infected individual members. Health care providers should enforce contact partner tracing to reduce further transmission of an infection within a sex-

ual network.

South Africa should reconsider the system of labour migration and conditions of migration in this post-apartheid era. The mining sector and other industrial areas attracting migrant men should improve social conditions and provide *family friendly* accommodation to curb family separation. Currently, a very small proportion of migrant men live with their families at their workplaces. The majority of migrant men still live in single sex hostels. Rates of circular migration can possibly be reduced by encouraging industrial decentralization and promotion of regional development.

6.2.2 Methodological

The study of migrant and non-migrant sexual networks has shown that ignoring sexual network random effects in the analysis of HIV/STIs biases the results. Inclusion of sexual network random effects leads to slightly magnified fixed effects estimates and standard errors are consistently larger in the random effects models. However, the effect of *HIV infection* was reduced in the logistic mixed model albeit the effects of all other factors inflated. Similar results were seen in the frailty model where the effects of *recent sexual partners* and of *syphilis infection* were slightly reduced. In the standard logistic model, *HIV infection* was acting as a proxy for sexual network effects probably due to high likelihood of HIV transmission if at least one partner is infected. Furthermore, *HIV infection* is a potential indicator of high risk behaviour. Similar arguments hold for *recent sexual partners* and *syphilis infection* in the frailty model. However, the importance of these variables was not completely removed from their respective models and the substantive inference remained unchanged.

The EM analysis shows that inclusion of sexual network random effects has similar effects in both logistic mixed model and frailty model. Fixed effects and baseline hazard estimates from a full Bayesian analysis of the frailty model do not markedly differ from ML estimates, Table 6.1. Since the priors for fixed effects were nearly

flat, the Gibbs sampler should give approximately the EM estimates at the mode of the joint posterior distribution (Harville, 1974). However, the standard errors and variance component estimates from the EM algorithm are biased downwards. This is particularly the feature of ML estimates for variance components as degrees of freedom lost due to estimation of fixed effects are not accounted for. The size

Table 6.1: Frailty model estimates from the EM algorithm and Gibbs sampler

	EM algorithm		Gibbs sampler	
Parameter	Mean	SE	Mean	SD
Baseline hazard				
Constant	0.007	0.001	0.013	0.004
Migration status				
Migrant men	0.460	0.216	0.391	0.276
Partners of migrant men	0.299	0.210	0.354	0.276
Non-migrant men	-0.219	0.259	-0.103	0.276
Age in years				
18 to 24	2.455	0.296	1.590	0.383
25 to 34	1.072	0.163	0.709	0.201
Recent sexual contact partners				
More than one	0.558	0.189	0.609	0.216
Number of lifetime partners				
More than one	0.328	0.172	0.521	0.215
Syphilis				
0=Negative, 1=Positive	0.284	0.158	0.501	0.179
Status of other STIs				
0=Negative, 1=Positive	0.503	0.181	0.588	0.218
Frailty variance				
Sexual network	0.459	0.069	0.812	0.120

and sparseness of the data can also have an effect. A considerable number of sexual

networks with only one partner included had an infection. The ML estimates from such data are biased towards zero. The sparseness of the data is less problematic in Bayesian analysis.

The Bayesian inference provides a natural framework with which to integrate the uncertainty about parameters and incorporate heterogeneity between sub-groups. The models incorporating this heterogeneity and estimated via ML approach become complex and require numerical integrations. Often the stability of numerical integration has to be carefully checked, involving additional computations. In situations where the sample size is small, the asymptotic normality of parameter estimates based on ML estimation is questionable, a problem which does not arise when using MCMC methods. The Gibbs sampler provides a useful and advantageous alternative to the EM algorithm when working with incomplete-data through 'data augmentation' techniques. The idea is to sample the missing data in addition to parameters, as was done in the frailty model. Superiority of Bayesian analysis has also been shown in GLMMs (Tu, Kowalski and Jia, 1999).

6.3 Further research

In this work, we have touched on aspects through which *migration* influences the spread of HIV/STIs. The focal point was on *migrant men* and their *female partners* from rural areas. Future studies and implementation of prevention strategies should also include female partners of migrant men at work places. Recently, there has been an increase of women who become migrants and are at risk of HIV infection (Brewer, *et al* 1998; Zuma, *et al* 2003). Sexual contacts between migrant men and these women not only connects HIV infection between urban and particular rural areas but has a potential of introducing HIV to the other rural areas where these women come from. The conditions and circumstances under which women migrate, and their role in transmitting HIV to other rural areas require further research.

It could be of interest to investigate the efficacy of providing antiretroviral therapies for HIV-positive partner(s) in a sexual network. However, in South Africa this kind of treatment is unlikely to be implemented on a large scale in the immediate future. Presently, antiretroviral treatment is not even routinely provided to HIV-positive pregnant women. Further research on understanding the factors which put women in weaker positions to negotiate safe sex and how they can be empowered to do so is required. Kavinya (2002) investigated factors related to women's empowerment in the context of reproductive decision making processes. Similar, research can be extended to sexual behavioural related issues.

Numerous numerical approximations have been used to accomplish estimation in the logistic mixed model. This has led to a range of statistical methods being used to fit these models. Many of these methods underestimate variance components and fixed effects. A number of corrections for this bias have been suggested, but thus far none of them has proven completely satisfactory. More work is needed to improve estimation in GLMMs. In frailty model estimation we used the gamma frailty distribution. The gamma distribution was chosen on the basis of its conjugacy status. It is worth investigating the performance of other forms of frailty distributions in similar context. However, fitting Bayesian models with this complexity does present analytical challenges for computing the joint posterior distribution. The main challenge is the Gibbs conditionals that are intractable. Intractable univariate Gibbs conditionals are handled by direct sampling methods. However, their multivariate generalizations are often inefficient and difficult to implement.

Appendix A

Abstracts of papers from the thesis

Risk factors for HIV infection among women in Carletonville, South Africa: migration, demography and sexually transmitted diseases

Abstract

We investigate the prevalence of, and risk factors for, HIV infection among women in an urban South African setting. A random sample of 834 women was recruited into a community-based cross-sectional study. HIV prevalence was 37.1% with higher prevalence among migrant women (46.0%) than non migrant women (34.7%), (odds ratio (OR)=1.61, 95%CI:1.11 2.31). The highest HIV prevalence (50.9%) was between ages 26 and 35 years. Having two or more lifetime partners increased the risk of HIV infection (OR=4.88, 95%CI:3.01-7.89). Migration, age, marital status, alcohol use, syphilis and gonorrhoea were independently associated with HIV infection. Migration increases the risk of HIV infection. Provision of services to treat sexually transmitted diseases (STDs) and educational empowerment programmes that will promote safer sex among migrant women are urgently needed.

Who Infects Whom? HIV-1 Concordance and Discordance Among Migrant and Non-Migrant Couples in South Africa

Abstract

Objectives: To measure HIV-1 discordance among migrant and non-migrant men and their rural partners, and to estimate the relative risk of infection from inside versus outside primary relationships.

Design: A cross-sectional behavioural and HIV-1 seroprevalence survey among 98 couples in which the male partner was a migrant and 70 couples in which the male was not a migrant. Methods: Following informed consent, a detailed questionnaire was administered and blood was collected for laboratory analysis. A mathematical model was developed to estimate the relative risk of infection for men and women from inside versus from outside the regular relationship.

Results: 70% (117/168) of couples were negatively concordant for HIV, 9% (16/168) were positively concordant and 21% (35/168) were discordant. Migrant couples were more likely than non-migrant couples to have one or both partners infected (35% versus 19%; p=0.026; OR=2.28) and to be HIV-1 discordant (27% versus 15%; p=0.066; OR=2.06). In 71.4% of discordant couples, the male was the infected partner; this did not differ by migration status. In the mathematical model, migrant men were 26 times more likely to be infected from outside their regular relationships than from inside (RR=26.3; p=0.000); non-migrant men were 10 times more likely to be infected from outside their regular relationships than inside (RR=10.5; p¡0.0001).

Conclusions: Migration continues to play an important role in the spread of HIV-1 in South Africa. The direction of spread of the epidemic is not only from returning migrant men to their rural partners, but also from women to their migrant partners. Prevention efforts will need to target both migrant men and women who remain at home.

The Impact of Migration on HIV-1 Transmission in South Africa A Study of Migrant and Nonmigrant Men and Their Partners

Abstract

Background: To investigate the association between migration and HIV infection among migrant and nonmigrant men and their rural partners. Goal: The goal was to determine risk factors for HIV-1 infection in South Africa.

Study Design: This was a cross-sectional study of 196 migrant men and 130 of their rural partners, as well as 64 nonmigrant men and 98 rural women whose partners are nonmigrant. Male migrants were recruited at work in two urban centers, 100 km and 700 km from their rural homes. Rural partners were traced and invited to participate. Nonmigrant couples were recruited for comparison. The study involved administration of a detailed questionnaire and blood collection for HIV testing.

Results: Testing showed that 25.9% of migrant men and 12.7% of nonmigrant men were infected with HIV (P=0.029; odds ratio (OR)=2.4; 95%CI:1.1-5.3). In multivariate analysis, main risk factors for male HIV infection were being a migrant, ever having used a condom, and having lived in four or more places during a lifetime. Being the partner of a migrant was not a significant risk factor for HIV infection among women; significant risk factors were reporting more than one current regular partner, being younger than 35 years, and having STD symptoms during the previous 4 months.

Conclusion: Migration is an independent risk factor for HIV infection among men. Workplace interventions are urgently needed to prevent further infections. High rates of HIV were found among rural women, and the migration status of the regular partner was not a major risk factor for HIV. Rural women lack access to appropriate prevention interventions, regardless of their partners' migration status.

The risk factors of sexually transmitted infections among migrant and non-migrant sexual networks from rural South Africa

Abstract

Objectives: To identify important risk factors of sexually transmitted infections (STI)s among migrant and non-migrant sexual networks. To estimate the degree of variability across sexual networks, and identify the effects of ignoring correlation on the risk factors of STIs.

Method: Cohorts of circular migrant men and their non-migrant sexual partners; and non-migrant men and their non-migrant sexual partners from rural South Africa were recruited between October 1998 and October 2001. Recruited female partners ranged from 0 to 4 per man, forming a sexual network. About 631 individuals aged between 18 and 69 years were recruited and followed-up every four months for interviews and examination. The main outcome is the presence of at least one curable STI in an individual at each visit.

Results: Prevalence of STI at each follow-up visit was 27.4%, 15.9%, 11.6% and 13.6%, respectively. Migration status, age, marital status, age at first sexual intercourse, recent sexual partners, HIV status were found to be important risk factors of STI. Syndromic management reduced the risk of STIs, odds ratio (OR)=0.75, p-value<0.0001. The risk of STI varies (1.46) considerably across sexual networks, and implies substantial correlation (0.59) between members of the same sexual network. Ignoring correlation underestimates standard errors by at most 11%.

Conclusion: Migration influences the spread of STIs. Different sexual networks are at different risks of STIs. Community interventions of HIV/STIs should target high-risk and co-transmitter sexual networks rather than high-risk individuals. This is more imperative for women who are in weak positions to negotiate safe sex.

Analysis of interval-censored data from circular migrant and non-migrant sexual partnerships using the EM algorithm

Abstract

In epidemiological studies where subjects are seen periodically on follow-up visits, interval-censored data occur naturally. The exact time the change of state (such as HIV seroconversion) occurs is not known exactly, only that it occurred sometime within a specific time interval. Methods of estimation for interval-censored data are readily available when data are independent. However, methods for correlated interval-censored data are not well developed. This paper considers the problem of finding maximum likelihood estimates when survival times are interval-censored and correlated within sexual partnerships. We consider the exact failure times for interval-censored observations as unobserved data, only known to be between two time points. Dependency induced by sexual partnerships is modelled as unobserved frailties assuming a parametric distribution. In this context, both the unobserved failure times and frailties form the missing data for the application of the EM algorithm. Maximization process maximises the standard survival frailty model. Results show high degree of heterogeneity between sexual partnerships. Intervention strategies aimed at combating the spread of HIV/STIs should treat sexual partnerships as social units and fully incorporate the effects of circular migration.

Application and comparison of methods for analysing correlated interval-censored data from sexual partnerships

Summary

In epidemiological studies where subjects are seen periodically on follow-up visits, interval-censored data occur naturally. The exact time the change of state (such as HIV seroconversion) occurs is not known exactly, only that it occurred sometime within a specific time interval. This paper considers estimation of parameters when HIV infection times are interval-censored and correlated. It is assumed that each sexual partnership has a specific unobservable random effect that induces association between infection times. Parameters are estimated using the expectation-maximization algorithm and the Gibbs sampler. The results from the two methods are compared. Both methods yield fixed effects and baseline hazard estimates that are comparable. However, standard errors and frailty variance estimates are underestimated in the expectation-maximization algorithm compared to those from the Gibbs sampler. The Gibbs sampler is considered a plausible alternative to the expectation-maximization algorithm.

Bibliography

[1] ABDOOL KARIM, Q.,ABDOOL KARIM, S.S., SINGH, B., SHORT, R., and NGXONGO, S. (1992). Seroprevalence of HIV infection in rural South Africa. *AIDS*, **6**(12):1535–1539.

[2] ABRAMOWITZ, M. and STEGUN, I. (1972). *Handbook of Mathematical Functions.* New York: Dover Publications.

[3] ADLER, M., FOSTER, S., GROSSKURTH, H., RICHENS, J., and SLAVIN, H. (1998). Sexual health and health care: sexually transmitted infections guidelines for prevention and treatment. *Department for International Development Occasional Paper*: London.

[4] ANDERSON, D.A. and AITKIN, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society, Series B*, **47**(2):203–210.

[5] ANDERSON, R.M., MAY, R.M., BOILY, M.C., GARNETT, G.P., and ROWLEY, J.T. (1991). The spread of HIV-1 in Africa: sexual contact patterns and the predicted demographic impact of AIDS. *NATURE*, **352**:581–589.

[6] AUVERT, B., BALLARD, R., CAMPBELL, C., *et al.* (2001). HIV infection among youth in a South African mining town is associated with herpes simplex virus-2 seropositivity and sexual behaviour. *AIDS*, **15**:885–898.

[7] BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**:192–236.

[8] BESAG, J. and Green, P.J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B*, **55**:25–37.

[9] BETENSKY, R.A. and FINKELSTEIN, D.M. (1999). A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, **18**(22):3089–3100.

[10] BOCK, D. and AITKIN, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, **46**(2):443–459.

[11] BOLSTAD, W.M. (1997). Monte Carlo method in Bayesian statistics. *New Zealand Statistician*, **32**(1):2–17.

[12] BOLSTAD, W.M. and MANDA, S.O. (2001). Investigating child mortality in Malawi using family and community random effects: a Bayesian analysis. *Journal of the American Statistical Association*, **96**(453):12–19.

[13] BOOTH, J.G. and HOBERT, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, **61**):265–285.

[14] BRESLOW, N.E. and CLAYTON, D.G. (1993). Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, **88**:9–25.

[15] BREWER, T.H., HASBUN, J., RYAN, C.A., *et al.* (1998). Migration, ethnicity and environment: HIV risk factors for women on the sugar cane plantations of the Dominican Republic. *AIDS*, **12**(14):1879–1887.

[16] BRILLINGER, D.R. and PREISLER, M.K. (1983). Maximum likelihood estimation in a latent variable problem. In *Studies in Econometrics, Time Series and Multivariate Statistics*, (eds. S. Karlin, T. Ameya and L.A. Goodman), pp. 31–65. New York: Academic Press.

[17] BROOKS, S.P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, **47**(Part 1):69–100.

[18] BUVE, A., LAGA, M., and PIOT, P. (1993). Sexually transmitted diseases: where are we now? *Health Policy and Planning*, **8**(3):277–281.

[19] CARLIN, B.P. and LOUIS, A.T. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman & Hall.

[20] CASELLA, G. and GEORGE, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**(3):167–174.

[21] CELENTANO, D.D., NELSON, K.E., SUPRASERT, S., *et al.* (1996). Risk factors for HIV-1 seroconversion among young men in northern Thailand. *Journal of the American Medical Association*, **275**(2):122–127.

[22] CHIB, S. and GREENBERG, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**(4):327–335.

[23] CLAYTON, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies in familial tendency in chronic disease incidence. *Biometrika*, **61**(1):141–151.

[24] CLAYTON, D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, **47**:467–485.

[25] CLAYTON, D. and CUZICK, J. (1985). Multivariate generalisations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series A*, **148**:82–117.

[26] COHEN, S.M. (1998). Sexually transmitted diseases enhance HIV transmission: no longer a hypothesis. *Lancet*, **351**(suppl III):5–7.

[27] COHEN, M.S., HOFFMAN, I.F., ROYCE, R.A., *et al.* (1997). Reduction of concentration of HIV-1 in semen after treatment of urethritis: implications for prevention of sexual transmission of HIV-1. *Lancet*, **349**:1868-1873.

[28] CONANT, M.A., SPICER, D.W. and SMITH, C.D. (1984). Herpes simplex virus transmission: condom studies. *Sexually Transmitted Diseases*, **11**(2):94–95.

[29] COWLES, K. and CARLIN, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**(434):883–904.

[30] COX, D.R. (1972). Regression models and Life Tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**:187–220.

[31] CROUCH, A.C. and SPIEGELMAN, E. (1990). The evaluation of integrals of the form $\int f(t)\exp(-t^2)dt$: application to logistic-normal models. *Journal of the American Statistical Association*, **85**:464–469.

[32] CRUSH, J. (1995). Mine migrancy in the contemporary era. In *Crossing Boundaries: Mine Migrancy in a Democratic South Africa*, (eds. J. Crush and W. James). Cape Town: IDASA/IDRC.

[33] DANIELS, M. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, **27**:569–580.

[34] DECOSAS, J., KANE, F., ANARFI, J.K., SODJI, K.D., and WAGNER, H.U. (1995). Migration and AIDS. *Lancet*, **346**(8978):826–828.

[35] DECOSAS, J. and ADRIEN, A. (1997). Migration and HIV. *AIDS*, **11**(Suppl A):S77–S84.

[36] DELLAPORTAS, P. and SMITH, A.F.M. (1993). Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, **42**:443–460.

[37] DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**:1–38.

[38] DEPARTMENT OF HEALTH, KwaZulu-Natal Province, South Africa. (1995). *Syndromic Management of STDs.* Durban: Department of Health, Coordinating Committee.

[39] DEPARTMENT OF HEALTH, South Africa. (1998). *South Africa Demographic and Health Survey.* Pretoria: Department of Health.

[40] DEPARTMENT OF HEALTH, South Africa. (1999). *National HIV Seroprevalence Survey of Women Attending Public Antenatal Clinics in South Africa.* Pretoria: Department of Health.

[41] DEPARTMENT OF HEALTH, South Africa. (2001). *National HIV and Syphilis Sero-Prevalence Survey of Women Attending Public Antenatal Clinics in South Africa.* Pretoria: Department of Health.

[42] DIGGLE, P.J., LIANG, K.Y., and ZEGER, S.L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

[43] DUCHATEAU, L., JANSSEN, P., LINDSEY, P., LEGRAND, C., NGUTI, R., and SYLVESTER, R. (2002). The shared frailty model and power for heterogeneity tests in multicenter trials. *Computational Statistics & Data Analysis*, **40**:603–620.

[44] EATON, L., FLISHER, A.J., and AARO, L.E. (2003). Unsafe sexual behaviour in South African youth. *Social Science and Medicine*, **56**:149–165.

[45] EFRON, B. and HINKLEY, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, **65**(3):457–487.

[46] EISENBERG, B. (1989). The number of partners and the probability of HIV infection. *Statistics in Medicine*, **8**:83–92.

[47] EVIAN, C. (1993). The socio-economic determinants of the AIDS epidemic in South Africa - a cycle of poverty. *South Africa Medical Journal*, **83**:653–656.

[48] FARRINGTON, C.P. and GAY, N.J. (1999). Interval-censored survival data with informative examination times: parametric models and approximate inference. *Statistics in Medicine*, **18**:1235–1248.

[49] FINKELSTEIN, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**:845–854.

[50] FLEMING, D.T. and WASSERHEIT, J.N. (1999). From epidemiological synergy to public health policy and practice: the contribution of other sexually transmitted diseases to sexual transmission of HIV infection. *Sexually Transmitted Infections*, **75**(1):3–17.

[51] FRIEDMAN, S.R., NEAIGUS, A., JOSE, B., *et al.* (1997). Sociometric risk networks and risk for HIV infection. *American Journal of Public Health*, **87**:1289–1296.

[52] GAMERMAN, D. (1997). *Markov Chains Monte Carlo.* London: Chapman & Hall.

[53] GELFAND, A.E. and SMITH, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**:398–409.

[54] GELMAN, A. (2004). Prior distributions for variance parameters in hierarchical models. Unpublished.

[55] GELMAN, A. (1996). Inference and monitoring convergence. In *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 131–143. London: Chapman & Hall.

[56] GELMAN, A. and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences (with comment). *Statistical Science*, **7**(4):457–511.

[57] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**:721–741.

[58] GERBASE, A.C., ROWLEY, J.T., HEYMANN, D.H., BERKLEY, S.F., and PIOT, P. (1998). Global prevalence and incidence estimates of selected curable STDs. *Sexually Transmitted Infections*, **74**(Suppl 1):s12–s16.

[59] GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian statistics 4*, (eds. J.M. Bernardo, A.F.M. Smith, A.P. Dawid and J.O. Berger), pp. 169–193. Oxford UK: Oxford University Press.

[60] GEYER, C.J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, **7**(4):473–511.

[61] GHANI, A.C. and GARNETT, G.P. (2000). Risks of acquiring and transmitting sexually transmitted diseases in sexual partner networks. *Sexually Transmitted Diseases*, **27**(10):579–587.

[62] GHANI, A.C., SWINTON, J., and GARNETT, G.P. (1997). The role of sexual partnership networks in the epidemiology of gonorrhea. *Sexually Transmitted Diseases*, **24**(1):45–56.

[63] GIBNEY, L., SAQUIB, N., and METZGER, J. (2003). Behavioral risk factors for STD/HIV transmission in Bangladesh's trucking industry. *Social Science and Medicine*, **56**(7):1411–1424.

[64] GILKS, W.R., CLAYTON, D.G., SPIEGELHALTER, D.J., *et al.* (1993). Modelling complexity: applications of Gibbs sampling in medicine. *Journal of the Royal Statistical Society, Series B*, **55**(1):39–52.

[65] GILKS, W.R., RICHARDSON, S., and SPIEGELHALTER, D.J. (1996). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter), pp.1–20. London: Chapman & Hall.

[66] GILKS, W.R. and WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**:337–348.

[67] GOETGHEBEUR, E. and RYAN, L. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics*, **56**:1139–1144.

[68] GOLDSTEIN, H. (1995). *Multilevel Statistical Models,* 2nd edition. London: Edward Arnold.

[69] GOUWS, E. and WILLIAMS, B.G. (2000). Science and HIV/AIDS in South Africa: a review of literature. *South African Journal of Science*, **96**:274–276.

[70] GREGSON, S., NYAMUKAPA, C.A., GARNETT, G.P., *et al.* (2002). Sexual mixing patterns and sex-differentials in teenage exposure to HIV infection in rural Zimbabwe. *Lancet*, **359**:1896–1903.

[71] GRENANDER, U. (1983). Tutorial in pattern theory. *Technical Report.* Providence, R.I.: Division of Applied Mathematics, Brown University.

[72] GROSSKURTH, H., MOSHA, F., TODD, J., *et al.* (1995). Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomised controlled trial. *Lancet*, **346**:530–536.

[73] GUO, G. and RODRIGUEZ, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala. *Journal of the American Statistical Association*, **87**:969–976.

[74] GUO, S.W. and LIN, D.Y. (1994). Regression analysis of multivariate grouped survival data. *Biometrics*, **50**:632–639.

[75] GUSTAFSON, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics*, **55**:230–242.

[76] GUTTORP, P. (1995). *Stochastic Modeling of Scientific Data.* London: Chapman & Hall.

[77] HARDIN, J.W. and HILBE, J.M. (2003). *Generalized estimating equations.* London: Chapman & Hall.

[78] HARVILLE, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**:383–385.

[79] HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**:320–340.

[80] HASTINGS, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**:97–109.

[81] HECKMAN, J. and SINGER, B. (1984). A method for minimising the impact of distributional assumption in econometric models for duration data. *Econometrica*, **52**:271–320.

[82] HOBERT, J.P. and CASELLA, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, **91**:1461–1473.

[83] HOUGAARD, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**:387–396.

[84] HUANG, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics*, **24**:540–568.

[85] HUANG, J. and WELLNER, J.A. (1997). Interval censored survival data: a review of recent progress. In *Proceedings of the First Seatle Symposium in Biostatistics: Survival Analysis*, (eds. D.Y. Lin and T.R. Fleming), pp. 123–169. New York: Springer-Verlag.

[86] HUNT, C.W. (1989). Migrant labor and sexually transmitted disease: AIDS in Africa. *Journal of Health and Social Behaviour*, **30**:353–373.

[87] IM, S. and GIANOLA, D. (1988). Mixed models for Binomial data with an application to lamb mortality. *Applied Statistics*, **37**:196–204.

[88] JAMSHIDIAN, M. and JENNRICH, R.I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society, Series B*, **62**:257–270.

[89] JEWELL, N.P., MALANI, H.M., and VITTINGHOFF, E. (1994). Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS. *Journal of the American Statistical Association*, **89**:7–18.

[90] JOCHELSON, K., MOTHIBELI, M., and LEGER, J.P. (1991). Human immunodeficiency virus and migrant labor in South Africa. *International Journal of Health Services*, **21**(1):157–173.

[91] JOHNSON, K.M., ALARCÓN, J., WATTS, D.M., *et al.* (2003). Sexual networks of pregnant women with and without HIV infection. *AIDS*, **17**:605–612.

[92] JORGENSEN, M. (2002). EM algorithm. *Encyclopedia of Environmetrics*, **2**:627–653.

[93] KALBFLEISCH, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B*, **40**:214–221.

[94] KANE, F., ALARY, M., NDOYE, I., *et al.* (1993). Temporary expatriation is related to HIV-1 infection in rural Senegal. *AIDS*, **7**:1261–1265.

[95] KASS, R.E., CARLIN, P.B., GELMAN, A., and NEAL, R.M. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician*, **52**:93–100.

[96] KASS, R.E. and WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**(435):1343–1370.

[97] KAVINYA, A.M. (2002). Women's empowerment, spousal communication and reproductive decision-making in Malawi. Unpublished Ph.D. thesis, University of Waikato.

[98] KELLY, P.J. and LIM, L.L-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, **19**:13–33.

[99] KIM, M.Y. and XUE, X. (2002). The analysis of multivariate interval-censored survival data. *Statistics in Medicine*, **21**:3715–3726.

[100] KLEIN, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, **48**:795–806.

[101] KOUMANS, E.H., FARLEY, T.A., GIBSON, J.J., *et al.* (2001). Characteristics of persons with syphilis in areas of persisting syphilis in the United States: sustained transmission associated with concurrent partnerships. *Sexually Transmitted Diseases*, **28**(9):504–507.

[102] KUK, A.Y.C. and CHENG, Y.W. (1997). The Monte Carlo Newton-Raphson algorithm. *Journal of Statistical Computation and Simulation*, **59**:91–99.

[103] LAGA, M., MANOKA, A., KIVUVU, M., *et al.* (1993). Non-ulcerative sexually transmitted diseases as risk factors for HIV-1 transmission in women: results from a cohort study. *AIDS*, **7**:95–102.

[104] LAIRD, N.M. and WARE, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**:963–974.

[105] LAIRD, N., LANGE, N., and STRAM, D. (1987). Maximum likelihood computations with repeated measures: an application of the EM Algorithm. *Journal of the American Statistical Association*, **82**(397):97–105.

[106] LAGARDE, E., PISON, G., and ENEL, C. (1996). A study of sexual behavior change in rural Senegal. *Journal of Acquired Immune Deficiency Syndromes*, **11**:282–287.

[107] LEVINE, W.C., POPE, V., BHOOMKAR, A., *et al.* (1994). Increase in endocervical CD4 lymphocytes in women with non-ulcerative STD. *Abstracts from the Tenth International Conference on AIDS/International Conference on STD.* Yokohama, Japan: Abstract 457C.

[108] LIANG, K.Y., ZEGER, S.L., and QAQISH, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**(1):3–40.

[109] LIANG, K.Y. and ZEGER, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1):13–22.

[110] LIN, D.Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, **13**:2233–2247.

[111] LINDSEY, J.C. and RYAN, L.M. (1998). Tutorial in biostatistics methods for interval-censored data. *Statistics in Medicine*, **17**:219–238.

[112] LITTELL, R.C., MILLIKEN, G.A., STROUP, W.W., and WOLFINGER, R.D. (1996). *SAS System for Mixed Models.* Cary, NC: SAS Institute Inc.

[113] LIU, Q. and PIERCE, D.A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, **81**(3):624–629.

[114] LOUIS, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**:226-223.

[115] LURIE, M., HARRISON, A., WILKINSON, D., and ABDOOL KARIM, S.S. (1997). Circular migration and sexual networking in rural KwaZulu/Natal: implications for the spread of HIV and other sexually transmitted diseases. *Health Transition Review*, **7**(Suppl. 3):15–24.

[116] LURIE, M. (2000). Migration and AIDS in Southern Africa: a review. *South African Journal of Science*, **96**:343–347.

[117] LURIE, M.N., WILLIAMS, B.G., ZUMA, K., *et al.* (2003a). The impact of migration on HIV-1 transmission in South Africa: a study of migrant and non-migrant men and their partners. *Sexually Transmitted Diseases*, **30**(2):149–156.

[118] LURIE, M., WILLIAMS, B.G., ZUMA, K., *et al.* (2003b). Who infects whom? HIV-1 concordance and discordance among migrant and non-migrant couples in South Africa. *AIDS*, **17**:2245–2252.

[119] MABEY, D. and MAYAUD, P. (1997). Sexually transmitted diseases in mobile populations. *Genitourinary Medicine*, **73**(1):18–22.

[120] MANDA, S.O.M. (1998). A nested random effects model analysis of child survival in Malawi. Unpublished Ph.D. thesis, University of Waikato.

[121] MANN, J., TARANTOLA, D.J.M. and NETTER, T.W. (1992). *AIDS In the World.* Cambridge: Harvard University Press.

[122] MBIZVO, M.T., MACHEKANO, R., MCFARLAND, W., *et al.* (1996). HIV seroincidence and correlates of seroconversion in a cohort of male factory workers in Harare, Zimbabwe. *AIDS*, **10**:895–901.

[123] MCCULLAGH, P. and NELDER, J.A. (1989). *Generalized Linear Models.* London: Chapman & Hall.

[124] MCCULLOCH, C.E. (1994). Maximum likelihood variance components-components estimation for binary data. *Journal of the American Statistical Association*, **89**:330–335.

[125] McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**:162–170.

[126] McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.

[127] Meheus, A. (1992). Women's Health: Importance of Reproductive Tract Infections, Pelvic Inflammatory Disease and Cervical Cancer. In *Reproductive Tract Infections: Global Impact and Priorities for Women's Reproductive Health* (eds. G. Adrienne, K.K. Holmes, P. Piot and J.N. Wasserheit). New York: Plenum Press.

[128] Meng, X.L. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, **86**:899–909.

[129] Mertens, T.E., Hayes, R.J., and Smith, P.G. (1990). Epidemiological methods to study the interaction between HIV infection and other sexually transmitted diseases. *AIDS*, **4**(1):57–65.

[130] Metropolis, N.A., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemichal Physics*, **21**:1087–1092.

[131] Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. New York: Springer.

[132] Mills, J.E., Field, C.A., and Dupuis, D.J. (2002). Marginally specified generalized linear mixed models: A robust approach. *Biometrics*, **58**:727–734.

[133] Molenberghs, G. and Ryan, L. (1999). An exponential family model for clustered multivariate binary data. *Environmetrics*, **10**:279–300.

[134] Monahan, J.F. and Stefanski, L.A. (1992). Normal scale mixture approximations to $F^*(x)$ and computation of the logistic-normal integral. In *Handbook of the logistic distribution* (ed. N. Balakrishnan), pp. 529–540, New York:Marcel Dekker.

[135] Morris, M. and Kretzschmar, M. (1997). Concurrent partnerships and the spread of HIV. *AIDS*, **11**:641–648.

[136] Murphy, S.A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics*, **22**:712–731.

[137] MURPHY, S.A. (1995). Asymptotic theory for the frailty model. *Annals of Statistics*, **23**:182–198.

[138] NELDER, J.A. and WEDDERBURN, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**:370–384.

[139] NICOLL, A. and GILL, O.N. (1999). The global impact of HIV infection and disease. *Commun Dis Public Health*, **2**(2):85–95.

[140] NUNN, A.J., WAGNER, H.U., KAMALI, A., KENGEYA-KAYONDO, J.F., and Mulder, D.W. (1995). Migration and HIV-1 seroprevalence in a rural Ugandan population. *AIDS*, **9**:503–506.

[141] ODELL, P.M., ANDERSON, K.M., and D'AGOSTINO, R.B. (1992). Maximum likelihood estimation for interval-censored data using a Weibul-based accelerated failure time model. *Biometrics*, **48**:951–959.

[142] O'FARRELL, N., WINDSOR, I., and BECKER, P. (1991). HIV-1 infection among heterosexual attenders at a sexually transmitted diseases clinic in Durban. *South African Medical Journal*, **80**:17–20.

[143] OVER, M. and PIOT, P. (1993). HIV infection and sexually transmitted diseases. In *Disease Control Priorities in Developing Countries* (eds. D.T. Jameson, W.H. Mosely, A.R. Measham and J.L. Babadilla), pp. 445–529. New York: University Press.

[144] PALMGREN, J. and RIPATTI, S. (2002). Fitting exponential mixed models. *Statistical Modelling*, **2**:23–38.

[145] PAN, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56**:192–203.

[146] PICKLES, A. and CROUCHLEY, R. (1995). A comparison of frailty models for multivariate survival data. *Statistics in Medicine*, **14**:1447–1461.

[147] PIOT, P. and TEZZO, R. (1990). The epidemiology of HIV and other sexually transmitted infections in the developing world. *Scandinavian Journal of Infectious Diseases*, **69**:89–97.

[148] PISON, G., LE GUENNO, B., LAGARDE, E., ENEL, C., and SECK, C. (1993). Seasonal migration: a risk factor for HIV infection in rural Senegal. *Journal of Acquired Immune Deficiency Syndromes*, **6**:196–200.

[149] POPULATION REPORTS. (1993). *Controlling Sexually Transmitted Diseases.* Series L, Number 9. Baltimore: Johns Hopkins Center for Communication Programs.

[150] PRENTICE, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**:1033–1048.

[151] QUINN, T.C. (1994). Population migration and the spread of types 1 and 2 human immunodeficiency viruses. *Proceedings of the National Academy of Sciences*, **91**:2407–2414.

[152] RAFTERY, A.E. and LEWIS, S.M. (1992). One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Statistical Science*, **7**:493–497.

[153] RAFTERY, A.E. and LEWIS, S.M. (1996). Implementing MCMC. In *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson, D.J. Spiegelhalter), pp. 115–130. London: Chapman & Hall.

[154] RAS, G.J., SIMSON, I.W., ANDERSON, R., PROZESKY, O.W., and HAMERSMA, T. (1983). Acquired immunodeficiency syndrome: a report of two South African cases. *South African Medical Journal*, **64**:140–142.

[155] RIPLEY, B.D. (1987). *Stochastic Simulation.* New York: Wiley.

[156] ROBERTS, G.O., GELMAN, A., and GILKS, W.R. (1994). Weak convergence and optimal scaling of random walk Metropolis algorithms. Technical Report, University of Cambridge.

[157] ROSS, S. (1996). *Stochastic Processes*, 2nd edn. New York: Wiley.

[158] ROTNITZKY, A. and JEWELL, P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for clustered correlated data. *Biometrika*, **77**:485–497.

[159] RUBIN, D.B. (1987). Comment of: The calculation of posterior distributions by data augmentation, by M.A. Tanner and W.H. Wong. *Journal of the American Statistical Association*, **82**:543–546.

[160] RÜCKER, G. and MESSERER, D. (1988). Remission duration: an example of interval-censored data observations. *Statistics in Medicine*, **7**:1139–1145.

[161] SASTRY, N. (1997). A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil. *Journal of the American Statistical Association*, **92**(438):426–435.

[162] SCHALL, R. (1990). On the maximum size of the AIDS epidemic among the heterosexual black population in South Africa. *South African Medical Journal*, **78**(9):507–510.

[163] SCHUMACHER, M., OLSCHEUSKI, M., and SCHMOOR, C. (1987). The impact of heterogeneity of comparisons of survival times. *Statistics in Medicine*, **6**:773–784.

[164] SELF, S.G. and LIANG, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**(398):605–610.

[165] SINHA, D. and DEY, D.K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*, **92**(439):1195–1212.

[166] SKAUNG, H.J. (2002). Automatic Differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. *Journal of Computational and Graphical Statistics*, **11**(2):458–470.

[167] SMITH, A.F.M. and GELFAND, A.E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *American Statistician*, **46**:84–88.

[168] SMITH, A.F.M. and ROBERTS, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, **55**:3–23.

[169] SMITH, C.A.B. (1977). Discussion of: Maximum likelihood estimation from incomplete data via the EM algorithm by A.P. Dempster, N.M. Laird and D.B. Rubin. *Journal of the Royal Statistical Society, Series B*, **39**:24–25.

[170] STEELE, B.M. (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrics*, **52**(4):1295–1310.

[171] STIRATELLI, R., LAIRD, N., and WARE, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, **40**:961–971.

[172] SUN, D., TSUTAKAWA, R.K., and HE, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica*, **11**(1):77–95.

[173] TANNER, M.A. and WONG, W.H. (1987). The calculation of posterior distributions by data augmentation (with comment). *Journal of the American Statistical Association*, **82**:528–550.

[174] TIERNEY, L. (1994). Markov chains for exploring the posterior distributions. *Annals of Statistics*, **22**:1701–1762.

[175] TIERNEY, L. and KADANE, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**:82–86.

[176] TU, X.M., KOWALSKI, J., and JIA, G. (1999). Bayesian analysis of prevalence with covariates using simulation-based techniques: applications to HIV screening. *Statistics in Medicine*, **18**:3059–3073.

[177] VAUPEL, J.W., MANTON, K.G., and STALLARD, E. (1987). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**:439–454.

[178] VERBEKE, G. and MOLENBERGHS, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Aproach.* New York: Springer.

[179] VERBEKE, G. and MOLENBERGHS, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

[180] WASSERHEIT, J.N. (1992). Epidemiological synergy. interrelationships between human immunodeficiency virus infection and other sexually transmitted diseases. *Sexually Transmitted Diseases*, **19**:61–77.

[181] WEDDERBURN, R.W.M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**:439–447.

[182] WEI, L.J., LIN, D.Y., and WEISSFELD, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association*, **84**:1065–1073.

[183] WILLIAMS, B., GILGEN, D., CAMPBELL, C., TALJAARD, D., and MACPHAIL, C. (2000). *The natural history of HIV/AIDS in South Africa: A biomedical and social survey in Carletonville.* Johannesburg, South Africa: Centre for Scientific and Industrial Research.

[184] WILKINSON, D., ABDOOL KARIM, S.S., HARISON, A., *et al.* (1999). Unrecognized sexually transmitted infections in rural South African women: a hidden epidemic. *Bulletin of the World Health Organization*, **77**:22–28.

[185] WOLFINGER, R.D. and O'CONNELL, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**:233–243.

[186] WYLIE, J.L. and JOLLY, A.M. (2001). Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada. *Sexually Transmitted Diseases*, **28**:14–24.

[187] YU, C. and ZELTERMAN, D. (2002). Statistical inference for familial disease clusters. *Biometrics*, **58**:7481–491.

[188] ZEGER, S.L. and LIANG, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**:121–130.

[189] ZEGER, S.L., LIANG, K.Y., and ALBERT, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**:1049–1060.

[190] ZEGER, S.L. and KARIM, M.Z. (1991). Generalised linear models with random effects, a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**(413):79–86.

[191] ZUMA, K., GOUWS, E., WILLIAMS, B.G., and LURIE, M. (2003). Risk factors for HIV infection among women in Carletonville, South Africa: migration, demography and sexually transmitted diseases. *International Journal of STD & AIDS*, **14**:814–817.

[192] ZUMA K, and LURIE, M.N. (2005). Application and comparison of methods for analysing correlated interval-censored data from sexual partnerships. *Journal of Data Science*, **3**(3):000–000.

[193] ZUMA, K., LURIE, M., and JORGENSEN, M. (2004). Analysis of interval-censored data from circular migrant and non-migrant sexual partnerships using the EM algorithm. *Submitted.*

[194] ZUMA, K., LURIE, M., WILLIAMS, B.G., *et al.* (2004). Statistical modelling of risk factors of sexually transmitted infections and heterogeneity between sexual networks from a rural district of South Africa. *Submitted.*

[195] ZWI, A. and BACHMAYER, D. (1990). HIV and AIDS in South Africa: What is an Appropriate Public Health Response? *Health Policy and Planning*, **5**(4):316–326.