Journal of Data Science **3**(2005), 241-256

Application and Comparison of Methods for Analysing Correlated Interval-censored Data from Sexual Partnerships

Khangelani Zuma[1] and Mark N. Lurie[2]
[1]*Human Sciences Research Council and* [2]*Brown University*

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. 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.

Key words: EM algorithm, frailty, Gibbs sampler, HIV, sexual partnerships.

1. Introduction

Interval-censored data arise in research settings where the exact time an event occurs is not observed directly, but only the time interval to which the observation belongs is observed. For instance, the exact time HIV seroconversion occurs is not observed exactly (Jewell, *et al.* 1994) but only the clinical examination times between which HIV infection occurred. Estimation methods for interval-censored data are often based on the Cox proportional hazards model (Cox, 1972). Finkelstein (1986) generalized the Cox proportional hazards model to account for interval-censored data. Huang and Wellner (1997) provide a rigorous theoretical account for maximum likelihood methods for interval-censored data.

Dependency of event times further complicates estimation in interval-censored data. Dependency may arise as a result of the sampling method used, such as in the study of HIV seroconversion among cohorts of circular migrant men, non-migrant men and their non-migrant sexual partners as described in Lurie *et al.*

(2003a; 2003b), and is the subject of our analysis, see Section 2. This dependency is often modelled as random effects or frailties. Frailty is the term describing common excess risk of infection among members of the same sub-group. Frailties are considered unobserved mutually independent random variables specified by some parametric distribution. The topic of frailty models has received considerable attention in demography (Vaupel, *et al.* 1987) and statistics (Clayton, 1978; Clayton and Cuzick, 1985). Including frailties in the interval-censored data likelihoods of (Finkelstein, 1986) or (Huang and Wellner, 1997) results in complex intractable likelihood functions. The conjugate gamma frailty distribution often assumed in standard survival frailty model (Klein, 1992) is no longer conjugate in the interval-censored likelihood (Finkelstein, 1986; Huang and Wellner, 1997).

In this paper, both the interval-censored infection time and frailties are treated as missing data. The primary goal of this paper is to apply and compare two statistical methods of analysing correlated interval-censored data. The two statistical methods considered are the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) and Bayesian analysis using Markov chain Monte Carlo (MCMC) methods (Gilks, *et al.* 1996; Carlin and Louis, 1996). The primary outcome variable is the survival time from 1990 until HIV infection, or until the end of the study or until the subject was lost to follow-up. In South Africa, research shows that the epidemic of HIV well established in 1990 (Gouws and Williams, 2000 and references therein) and thus 1990 is used as the initial time. The main focus is on full Bayesian analysis and its comparison to maximum likelihood estimation. Section 3 of this paper presents the migration data to be analyzed. The assumed conditional survival model is presented in Section 3. Section 4 presents the likelihood formulation for the EM estimation. Full Bayesian estimation is presented in Section 5. The application and conclusion are presented in Sections 6 and 7, respectively.

2. The Data

A total of 631 individuals aged between 18 and 60 years were recruited into the study. The study composed of circular migrant men, non-migrant men and their rural based non-migrant sexual partners. Circular migration is the predominant type of migration in South Africa where young men migrate to work in urban areas leaving their rural sexual partners behind, and return home periodically. Circular migrant men were recruited from their workplaces in urban areas. They provided details of their sexual partners residing in rural areas, who were then located and invited to participate. In the neighbourhood of each migrant man's rural household, a non-migrant man and his partner(s) were selected and invited to participate. The study participants were visited approximately every four months to administer a detailed questionnaire eliciting information related to

Table 1: Distribution of sexual partnerships and HIV infection

Sexual partnership size	Number of sexual partnerships	Percentage	HIV infection	
			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

demographic and socioeconomic characteristics, sexual behavioural and biomedical factors. At each visit, blood and urine specimen were collected to test for HIV status and status of other sexually transmitted infections (STI)s, respectively. This resulted in discrete interval-censored infection times due to known clinical visit time before and after the infection time. Further details of the study including details of inclusion criteria and testing of specimen have been reported elsewhere (Lurie, *et al.* 2003a; 2003b).

The current analysis is restricted to 339 identifiable distinct sexual partnerships from 604 individuals. The mean sexual partnership size is 1.78 individuals. Sexual partnership size ranges from 1 to 5 individuals with only one man in each sexual partnership. Table 1 shows the distribution of sexual partnerships and percentage of persons infected with HIV. Considerable numbers of migrant men gave incorrect information about the location of their partners and some of their identified partners refused to participate resulting in cases where only a man was included making up to 36% of five different number of included participants in a partnership. Fifty-two percent of sexual partnerships were couples.

Men (mostly migrants) whose partners were not part of the study, contributed considerably to high HIV infection in sexual partnerships where only one partner was included. HIV infection was considerably higher in triads than couples. There are small proportions of sexual partnerships of size greater than three to make valid comparisons. The maximum total number of HIV infected members per sexual partnership size was three, and was among triads. 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. Men tend to overstate their sexual behaviour whilst women understate theirs.

3. The Conditional Survival Model

The data analyzed is clustered within sexual partnerships. The HIV infection time is interval-censored such that it is unobserved but only the time interval within which HIV infection occurs is observed. Let $v_{ij} = \{L_{ij}; U_{ij}\}$ $(i = 1, \cdots, I; j = 1, \cdots, J_i)$ denote the examination endpoints encompassing the unobserved HIV infection time $t_{ij} \in (L_{ij}, U_{ij}]$ where L_{ij} is the last time a person tested HIV negative and U_{ij} is the first time a person tested HIV positive. For right-censored observation, the observed time is L_{ij}. For notational simplicity, let $Y_{ij} = (L_{ij}, t_{ij})$ denote both observed right-censored time L_{ij} and unobserved infection time t_{ij}. Define a non-censoring indicator $\delta_{ij} = 1$ if HIV positive and 0 otherwise. The *ith* sexual partnership frailty is denoted by b_i.

The multiplicative frailty model is assumed. Baseline hazards are assumed constant $\lambda_0(y_{ij}) = \lambda_0$ and the corresponding integrated baseline hazard is $\Lambda_0(y_{ij}) = \lambda_0 y_{ij}$. Throughout this paper, it is assumed that censoring and infection times are independent. Therefore, the censoring process is non-informative. Conditional on b_i, survival times are independent and their conditional hazards distribution is

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

where X_{ij} is the vector of covariates and β represents the corresponding covariate effect. Those infected with HIV contribute to the likelihood the product of their conditional hazards and conditional survival function whilst those who were right-censored contribute only the conditional survival function. The conditional survival distribution is $S(y_{ij}|b_i, X_{ij}) = \exp[-H(y_{ij}|b_i, X_{ij})]$ where $H(y_{ij}|b_i, X_{ij}) = b_i \Lambda_0(y_{ij}) \exp(\beta' X_{ij})$ is the integrated hazards corresponding to $h(y_{ij}|b_i, X_{ij})$. The integrated fixed effects hazards are $\Lambda(y_{ij}|X_{ij}) = \Lambda_0(y_{ij}) \exp(\beta' X_{ij})$.

The frailties act multiplicatively on the baseline hazard and are interpreted as relative risks (RR)s. They are unobserved and take only positive values. In this work, they are modelled as independent random variates from a Gamma(α, α) distribution. The RR for sexual partnerships has mean 1 and variance $1/\alpha$ in this case. The unit mean constrain ensures that the sexual partnership effects represent deviations from the population average risk. The gamma distribution is a popular choice for frailties, possibly due to its flexible shape and conjugacy property (Guo and Rodriguez, 1992; Klein, 1992; Bolstad and Manda, 2001). Heckman and Singer (1984) and Pickles and Crouchley (1995) discuss problems associated with the choice of the frailty distribution.

4. Maximum Likelihood Estimation

This section examines estimation of fixed effect parameters β, baseline hazard λ_0 and association parameter α from the Gamma frailty distribution using maximum likelihood approach. To do this, we need the joint distribution of the response vector and frailties. The response vector y_i of sexual partnership i consists of (possibly sub-vectors) observed v_i and unobserved t_i. Using conditional independence between y_i given b_i, the complete-data likelihood contribution for sexual partnership i is

$$\begin{aligned} L_i(b_i, v_i, t_i; \theta) &= \frac{\alpha^\alpha}{\Gamma(\alpha)} b_i^{\alpha-1} e^{-\alpha\, b_i} \\ &\times \prod_{j=1}^{J_i} \left(e^{-b_i \lambda_0 t_{ij} e^{\beta' X_{ij}}} b_i \lambda_0 e^{\beta' X_{ij}} \right)^{\delta_{ij}} \left(e^{-b_i \lambda_0 L_{ij} e^{\beta' X_{ij}}} \right)^{1-\delta_{ij}} \qquad (4.1) \end{aligned}$$

where $\theta = \{\alpha, \lambda_0, \beta\}$. The complete-data log-likelihood for sexual partnership i corresponding to (4.1) is

$$\begin{aligned} l_i(\theta) &= \alpha \log \alpha - \log \Gamma(\alpha) + (\alpha - 1) \log b_i - \alpha b_i \\ &+ \sum_{j=1}^{J_i} \delta_{ij} [-b_i \lambda_0 t_{ij} \exp(\beta' X_{ij}) + \log(b_i) + \log(\lambda_0) + \beta' X_{ij}] \\ &- (1 - \delta_{ij}) b_i \lambda_0 L_{ij} \exp(\beta' X_{ij}). \qquad (4.2) \end{aligned}$$

The complete-data log-likelihood (4.2) depends on functions of unobserved infection time (t_{ij}) and sexual partnership specific frailty (b_i). Implementation of the EM algorithm requires calculation of $Q(\theta; \theta^{(r)})$ equal to the conditional expectation of (4.2) over all functions of unobserved data, given the observed data and current estimate $\theta^{(r)}$ of θ. The observed data likelihood $L_i(v_i; \theta)$ is attained by integrating out unobserved data from (4.1) as follows:

$$\begin{aligned} & L_i(v_i; \theta) \\ = & \int_0^\infty \int_{L_{i\delta_{i+}}}^{U_{i\delta_{i+}}} \cdots \int_{L_{i2}}^{U_{i2}} \int_{L_{i1}}^{U_{i1}} L_i(b_i, v_i, t_i; \theta)\, dt_{i1}\, dt_{i2} \cdots dt_{i\delta_{i+}}\, db_i \\ = & \frac{\alpha^\alpha}{\Gamma(\alpha)} \int_0^\infty b_i^{\alpha-1} e^{-b_i \left(\alpha + \sum_{j=1}^{J_i} (1-\delta_{ij})\, \Lambda(L_{ij}|X_{ij}) \right)} \\ & \times \prod_{j=1}^{\delta_{i+}} \left(e^{-b_i \Lambda(L_{ij}|X_{ij})} - e^{-b_i \Lambda(U_{ij}|X_{ij})} \right) db_i \end{aligned}$$

where $\delta_{i+} = \sum_{i=1}^{J_i} \delta_{ij}$ is the total number of HIV positive members of a particular sexual partnership network. The number of people infected with HIV in a particular sexual partnership is known at the analysis stage and thus δ_{i+}.

It can be seen from (4.2) that we need only the conditional expectations of b_i, $\log b_i$ and $b_i t_{ij}$. This is because (4.2) is a linear function of these quantities. The conditional expectations of b_i and $\log b_i$ are computed using the marginal conditional distribution of b_i given the observed data. Firstly, the joint marginal distribution $f(b_i, v_i|\theta)$ is

$$\begin{aligned} f(b_i, v_i; \theta) &= \int_{L_{i\delta_{i+}}}^{U_{i\delta_{i+}}} \cdots \int_{L_{i2}}^{U_{i2}} \int_{L_{i1}}^{U_{i1}} L_i(b_i, v_i, t_i; \theta)\, dt_{i1}\, dt_{i2} \cdots dt_{i\delta_{i+}} \\ &= \frac{\alpha^\alpha}{\Gamma(\alpha)} b_i^{\alpha-1} e^{-b_i(\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(L_{ij}|X_{ij})} \\ &\qquad \times \prod_{j=1}^{\delta_{i+}} (e^{-b_i\Lambda(L_{ij}|X_{ij})} - e^{-b_i\Lambda(U_{ij}|X_{ij})}). \end{aligned}$$

Thus, the marginal conditional distribution of b_i is given by

$$\begin{aligned} g(b_i|v_i; \theta) &= \frac{f(b_i, v_i; \theta)}{L_i(v_i; \theta)} \\ &= \frac{b_i^{\alpha-1} e^{-b_i(\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\Lambda(L_{ij}|X_{ij})} G(b_i)}{\int_0^\infty b_i^{\alpha-1} e^{-b_i\left(\alpha+\sum_{j=1}^{J_i}(1-\delta_{ij})\,\Lambda(L_{ij}|X_{ij})\right)} G(b_i)\, db_i}, \end{aligned}$$

where

$$G(b_i) = \prod_{j=1}^{\delta_{i+}} \left(e^{-b_i\Lambda(L_{ij}|X_{ij})} - e^{-b_i\Lambda(U_{ij}|X_{ij})} \right).$$

The integrand expands depending on δ_{i+}. Therefore, the sums of integrals will disappear by making each integrand mimic a gamma density. The EM algorithm proceeds by iteratively computing the following E- and M-steps.

- *E-step*: The E-step computes the conditional expectations of b_i, $\log b_i$ and $b_i t_{ij}$ given the observed data and current parameter estimates of θ. The conditional expectation of b_i simplifies whilst that of $\log b_i$ can be approximated numerically. The conditional expectation $E[b_i t_{ij'}|L_{ij'} < t_{ij'} \leq U_{ij'}; \theta]$ is calculated using $f(b_i, t_{ij'}; \theta)$ where $f(b_i, t_{ij'}; \theta)$ is the joint conditional distribution obtained by first integrating out the remaining sexual partnership unobserved infection times t_{ij}, for $j \neq j'$ from $L_i(b_i, v_i, t_i; \theta)$.

- *M-step*: The M-step of the algorithm involves maximizing $Q(\theta; \theta^{(r)})$ after replacing b_i, $\log b_i$ and product $b_i t_{ij}$ by their conditional expectations in (4.2). Maximization is accomplished via Newton-Raphson algorithm which requires evaluation of the first and second derivatives of $Q(\theta; \theta^{(r)})$. The parameters (α, λ_0) and the parameter vector (β) are maximized sequentially.

The two steps are iterated until convergence criterion is met. The EM algorithm may show slow convergence if there is large amount of missing data, or if the estimated hyperparameter heavily depends on missing data. The rate of convergence in this study was satisfactory and thus complex computations involved in the acceleration procedures were not used (Laird, Lange and Stram, 1987).

5. The Full Bayesian Estimation

The model framework presented in the preceding section is hierarchical and fully specified from the frequentist point of view and the model parameters have been estimated using the EM algorithm. In addition to this, we need to specify priors for fixed effects β, baseline hazard λ_0 and hyperparameter α before the model is fully specified from a Bayesian perspective. The prior for β is assumed multivariate normal with mean vector $\mathbf{d_0} = \mathbf{0}$ and diagonal covariance matrix $\mathbf{\Sigma_0} = \upsilon_\mathbf{0}\mathbf{I}$, where υ_0 is a suitably chosen large number. The prior mean is set to 0 since the fixed effects represent logarithms of RRs and not expected to be far from 0. The effect of these priors on the marginal posteriors of the regression coefficients is almost identical to the flat priors. A Gamma(ξ_0, ζ_0) prior distribution is specified for the baseline hazard. The specification of priors for precision parameter is more difficult in hierarchical model setting. An improper prior can lead to an improper posterior (Hobert and Casella, 1996). Thus, a Gamma (ν_0, κ_0) prior for precision component is often assumed due to its conjugacy status. All prior distributions are assumed independent of each other.

The modelling framework considered 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 for precisely known failure times. Gustafson (1997) used similar approach in the implementation of Cox partial likelihood. Bolstad and Manda (2001) presented a three-way multilevel model for child mortality. In this work, an important aspect of sampling interval-censored failure times conditional on examination times, frailties and other observed data is considered. Sinha and Dey (1997) reviewed a number of Bayesian methods for analysing survival data and clearly, the extensions of semiparametric Bayesian model for analysis of multivariate survival data using frailty model (Clayton, 1991) to interval-censored data are not immediate.

Figure 1 presents the directed acyclic graph of the model. Each parameter node is circled; the data and prior constants are indicated by rectangles denoting that they are fixed. The joint distribution of all parameters, hyperparameters and the data can be written as the product of all prior and conditional distribution as follows:

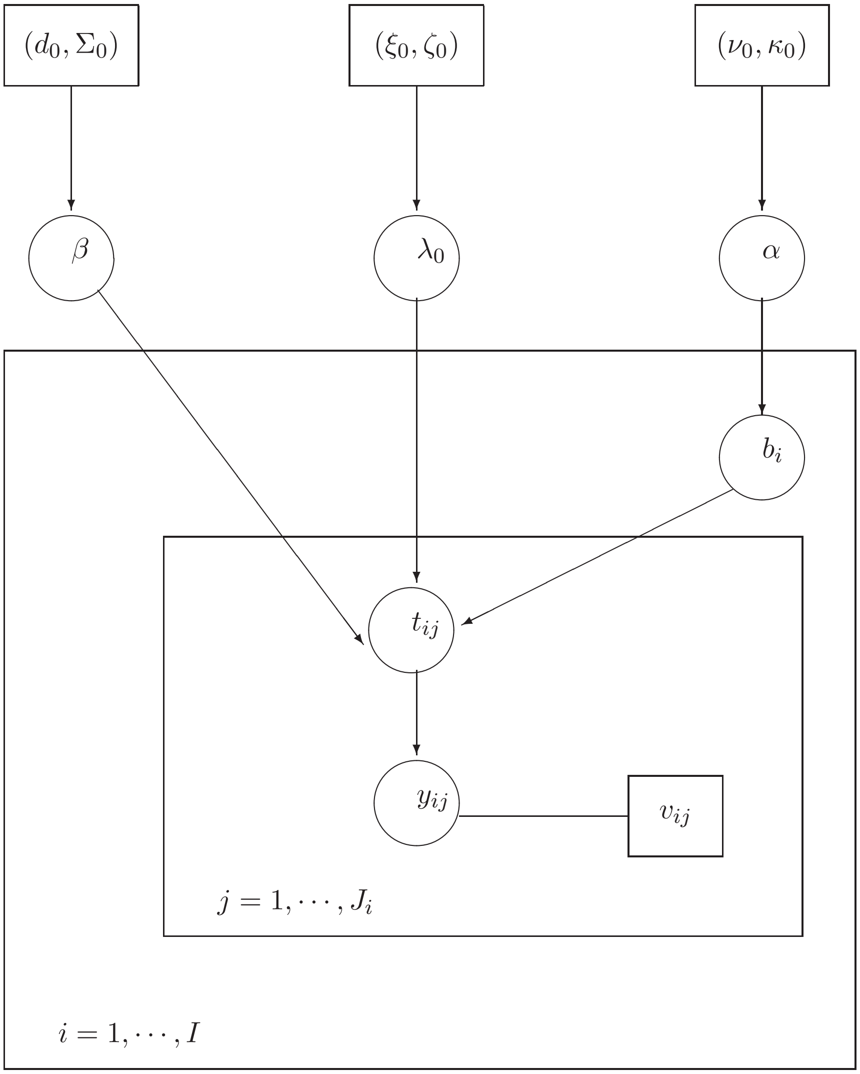

Figure 1: The directed acyclic graphical model representation of migration data

$$f(\text{data}, \beta, \lambda_0, t_{ij}, b_i, \alpha) = f(\beta) f(\lambda_0) f(\alpha) \prod_{i=1}^{I} f(b_i|\alpha) \prod_{j=1}^{J_i} L_i(y_{ij}|\beta, \lambda_0, b_i). \qquad (5.1)$$

In Bayesian analysis, the joint posterior distribution of all parameters given the data is required. In our model, the joint posterior distribution cannot be obtained analytically. Instead, we use the Gibbs sampler (Geman and Geman, 1984) to obtain the required posterior. The Gibbs sampler proceeds by iteratively sampling from the conditional posterior distribution of each parameter using the most recent values of the given parameters. This generates a Markov chain in the process and the chain has the joint posterior as its long-run distribution.

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 relevant Gibbs conditionals can be computed from the joint distribution (5.1) and some of these conditionals are presented here below:

- $f(b_i|\text{data}, \beta, \lambda_0, t_{ij}, \alpha) \propto b_i^{\sum_{j=1}^{J_i} \delta_{ij} + \alpha - 1}$
$$\times\, e^{-b_i\left[\alpha + \sum_{j=1}^{J_i} \delta_{ij} \Lambda(t_{ij}|X_{ij}) + \{1-\delta_{ij}\}\Lambda(L_{ij}|X_{ij})\right]}$$

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

- $f(t_{ij}|L_{ij} < t_{ij} \le U_{ij}, \text{data}, \beta, \lambda_0, b_i, \alpha) = \frac{f(t_{ij}|\text{data}, \beta, \lambda_0, b_i)}{\int_{L_{ij}}^{U_{ij}} f(t|\text{data}, \beta, \lambda_0, b_i, \alpha) dt}$
$$\propto \exp(-t_{ij} b_i \lambda_0 e^{\beta' X_{ij}})$$

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 $b_i \lambda_0 e^{\beta' X_{ij}}$. This node can also be sampled directly on condition that the sampled value $t_{ij} \in (L_{ij}, U_{ij}]$.

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

This full conditional does not simplify to any standard distribution. Methods for sampling from an arbitrary conditional distribution are required. It turns out that the full conditional distribution is a simple log-concave distribution in α and thus can be sampled efficiently using the adaptive-rejection sampling scheme (Gilks and Wild, 1992).

- $f(\beta|\text{data}, \lambda_0, t_{ij}, b_i, \alpha)$

If a flat prior $f(\beta) = 1$ is assumed for β, the posterior mode can be replaced by the maximum likelihood estimate $\hat{\beta}$ and the log posterior density by the log likelihood function. The normal approximation, with mean and covariance matrix equal to the mode and inverse of the information obtained from the maximum likelihood estimation, can be used in the Metropolis step to generate candidates for β.

6. Application and Comparison

The possible risk factors of HIV considered in our analysis include migration status, age at recruitment, number of lifetime partners, and number of recent sexual contact partners, syphilis status and status of other STIs. Other STIs refer to the status of any of the following STIs: chlamydia, gonorrhoea, genital discharge and genital sores. These are typical covariates that are considered important determinants of HIV infection. Circular migration is one of the structural factors associated with HIV infection, but the dynamics and complex role of circular migration as a determinant of HIV infection is still a major issue for social science research. Importance of migration as a risk factor lies in the assumption that circular migrant men, whilst away from their partners, engage in risky sexual behaviour with other female sexual partners (Lurie, *et al.* 1997). During this period of migration, partners of circular migrant men are also as likely to acquire extra sexual partners (Lurie, *et al.* 1997). Risk factors parallel to the epidemic of HIV such as the number of lifetime partners are considered important determinants of HIV infection due to their cumulative effect. Evidence shows that STIs, both ulcerative and non-ulcerative, facilitate transmission of HIV (Wasserheit, 1992).

The EM algorithm and the Gibbs sampler were implemented on Microsoft Visual C++ Version 6.0. For Bayesian inference, five parallel chains were run from independent starting points for $2n$ =4 000. All the fixed effects parameters, some random effects, inverse scale, baseline hazard and some infection times were monitored for convergence. Gelman and Rubin's (1992) scale reduction factor and other convergence checks were computed. These convergence checks were satisfactory. The first 2 000 iterations were discarded. Starting from the 4 000th iteration, a further 38 000 values were simulated. Every 100th value was taken resulting in 2 000 nearly independent samples from the joint posterior distribution.

It is worth mentioning that fixed effect sampling scheme involved an EM estimation for maximum likelihood estimates and the calculation of Fisher information matrix. The resulting estimates were used in the proposal density for the Metropolis step. This was computationally intensive and equal sampling of all parameters led to correlated values of b_i, t_{ij}, α and λ_0 compared to β values.

Table 2: Results for HIV infection among migratory partnerships form South Africa

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

Thus, the iteration scheme was modified to iterate through b_i, t_{ij}, α and λ_0 five times for each draw of β. The modification greatly improved efficiency. The acceptance rate for candidate β 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 estimates from both methods are presented in Table 2. The fixed effects estimates are presented on a log scale where no risk is represented by 0. Standard deviations from the EM algorithm were computed using the SEM methods proposed by Meng and Rubin (1991). The estimates of the fixed effects and baseline hazard are similar for all practical purposes to the respective modes obtained from the EM algorithm. However, variance parameter estimates differ markedly between the two methods. The Gibbs sampler provides variance estimates that are larger than those from the EM algorithm. The estimate of sexual partnership 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). In the EM algorithm, the mode of the sexual partnership frailty variance was estimated to be 0.462. The unfavorable consequence is that if one based inference on the EM algorithm, the resulting confidence intervals would be narrower and differences more significant.

7. Data

The Gibbs sampler and the EM algorithm have been implemented on correlated interval-censored data. The paper showed that Bayesian analysis via MCMC 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. This agreement 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 likelihood approach are severely biased downwards. The bias reflects the incapability of likelihood approach to correct for variability of unobserved frailties and infection time (Ripatti and Palmgren, 1999). Downward bias in standard deviations observed in the likelihood estimation 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.

The Gibbs sampler has been shown to be a plausible alternative to the EM algorithm in this setting. The Gibbs sampler does not require evaluation of high-dimensional integrals as done in the EM algorithm. Estimation via the Gibbs sampler is advantageous in that it is easily extendable to other frailty distributions (Sargent, 1998). Superiority of the Gibbs sampler has also appeared in three-way multilevel hazards model for right-censored data (Manda, 2001).

Acknowledgements

The authors thank all members of the Migration Project Team, without whom this project would not have been possible. The authors also thank Dr. William M. Bolstad for his constructive comments and Ms. Lebogang Letlape for her editorial comments. The study was part of the Africa Centre for Population Studies and Reproductive Health, South Africa and was supported by the Wellcome Trust (Grant 05057/z/97abc) and South African Medical Research Council. This publication was made possible in part through the support of the New Zealand

Overseas Development Assistance. Additional support came from a training grant awarded by the National Institute of Drug Abuse to the Miriam Hospital (Grant 5 T32 DA13911) and National Institute of Mental Health Career Development Award granted to Dr. Lurie at Brown University (Grant 1 K01 MH069113-01A1).

References

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**, 12-19.

Carlin, B. P. and Louis, A. T. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman & Hall.

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**, 141-151.

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467–485.

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.

Cox, D. R. (1972). Regression models and Life Tables (with discussion). *Journal of the Royal Statistical Society* Series B **34**, 187-220.

Dempster, A. P., Laird, N. M., and Rubin, R. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society* Series B **39**, 1-38.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845-854.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with comment). *Statistical Science* **7**, 457-511.

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.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (Edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), 1-20. Chapman & Hall.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337-348.

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.

Guo, G. and Rodriguez, G. (1992). Estimating 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.

Gustafson, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics* **55**, 230-242.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383-385.

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.

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.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: A review of recent progress. In *Preceedings of the First Seatle Symposium in Biostatistics: Survival Analysis*, (Edited by D. Y. Lin and T. R. Fleming), 123-169. Springer-Verlag.

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.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795-806.

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**, 97-105.

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* Suppl. 3 **7**, 15-24.

Lurie, M. N., Williams, B. G., Zuma, K., Mkaya-Mwamburi, D., Garnett, G. P., Sturm, A. W., Sweat, M. D., Gittelsohn, J. and Abdool Karim, S. S. (2003). 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**, 149-156.

Lurie, M., Williams, B. G., Zuma, K., Mkaya-Mwamburi, D., Garnett, G. P., Sweat, M. D., Gittelsohn, J. and Abdool Karim, S. S. (2003). Who infects whom? HIV-1 concordance and discordance among migrant and non-migrant couples in South Africa. *AIDS* **17**, 2245-2252.

Manda, S. O. M. (2001). A comparison of methods for analysing a nested frailty model to child survival in Malawi. *Aust. N. Z. J. Stat.* **43**, 7-16.

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

Pickles, A. and Crouchley, R. (1995). A comparison of frailty models for multivariate survival data. *Statistics in Medicine* **14**, 1447-1461.

Raftery, A. E. and Lewis, S. M. (1996). Implementing MCMC. In *Markov Chain Monte Carlo in Practice* (Edited by. W. R. Gilks, S. Richardson, D. J. Spiegelhalter), pp. 115-130. Chapman & Hall.

Ripatti, S. and Palmgren, J. (1999). Estimation of multivariate frailty models using the penalised partial likelihood. Department of Biostatistics research Report 99/1. University of Copenhagen.

Sargent, D.J. (1998). A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics* **54**, 1486-1497.

Sinha, D. and Dey, D. K. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* **92**, 1195-1212.

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.

Wasserheit, N. J. (1992). Epidemiological synergy, interrelationships between human immunodeficiency virus infection and other sexually transmitted diseases. *Sexually Transmitted Diseases* **19**, 61-77.

Received Marcch 25, 2004; accepted May 4, 2004.

Khangelani Zuma
Human Sciences Research Council
Private Bag 41
PRETORIA, 0001, South Africa
e-mail: kzuma@hsrc.ac.za

Mark N. Lurie
Department of Community Health and the Miriam Hospital
School of Medicine
Brown University
Providence, RI, USA
e-mail: Mark_Lurie@brown.edu