Modeling and Estimation of Patterns of Relationship Formation and Dissolution

Modeling and Estimation of Patterns of Relationship Formation and Dissolution

A dissertation presented

by

Yared Gurmu

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

May 2016

Dissertation Advisor: Professor Victor De Gruttola Yared Gurmu

Modeling and Estimation of Patterns of Relationship Formation and Dissolution

Abstract

This dissertation describes and develops methods for modeling sexual partnership formation and termination using retrospectively collected survey data. Such methods are required to produce information necessary to model propagation of sexually transmitted diseases and the impact of interventions on such processes which are being used both to design and to monitor HIV combination prevention studies. Sexual history data are commonly obtained through surveys that collect information on relationships that are ongoing during a fixed time window. This sampling mechanism leads to incomplete sexual history data and duration data that are left truncated and right censored.

In Chapter 1, we describe a common sampling scheme for collecting sexual partnership data, discuss a key assumption required for unbiased estimation, and provide the conditions under which the nonparametric maximum likelihood estimator of the relationship duration distribution is unique and consistent. We also investigate the conditions required for the consistency of the regression coefficient from a Cox proportional hazards model that apply even when the distribution of duration is not completely identifiable due to restrictions on the support of the truncation distribution. Lastly, we will provide some illustrative examples on estimating distribution of most recent partnerships and present spline regression results based on sexual history data from Botswana.

In Chapter 2, we present a Markov framework for modeling and estimation of partnership transition probabilities for sexual history data collected under a retrospective sampling scheme. We propose a stochastic expectation maximization algorithm (stEM) coupled with rejection-sampling scheme in order to estimate transition probabilities from a state of celibacy to monogamy and to concurrency (or vice versa). This approach ac-

commodates the retrospective sampling scheme from which sexual partnership data is obtained and utilizes all available information from the sexual history data. In particular, this paper will address maximum likelihood estimation via stEM when our observed data includes information on the number of certain types of transitions without specifying the sojourn time in the states. For example, with regards to partnership data, the total life time number of partnerships (or number of partnerships within a fixed window of time) may be known even though the sojourn time of each of the partnerships in the different states may not be known. In the process of estimating transition rates, we incorporate such information by using rejection sampling. Simulation results showing the performance of the stEM will be presented. We also provide an application example based on partnership data collected from South Africa.

In Chapter 3, we extend the Markov model presented in Chapter 2 so that the sexual history process can be fully characterized. This approach combines a Markov model and a logistic regression framework. The Markov model states we consider include celibacy, monogamy and concurrency; the logistic regression model classifies the pattern of concurrency, which can be either transitional (older partnership ends first) or embedded (new ends first). By using both types of models we can fully characterize the processes of interest. Estimation of model parameters is based on a stochastic expectation maximization algorithm (stEM) coupled with rejection-sampling scheme. Strategies based on statistics that arise naturally from the estimation procedure itself stEM are used to validate model assumptions. The method is illustrated using sexual history data collected from South Africa. Simulation results are used to demonstrate properties of the estimation methods.

Contents

Acknowledgments

I am deeply grateful to my advisor, Victor De Gruttola, for guiding, encouraging and supporting me over the years. Victor's energy and enthusiasm for solving high impact public health problems has been quite inspirational to me.

I would like to thank my thesis committee members, Rebecca Betensky, Jing Qian, and Lorenzo Trippa, for all your guidance through this process; your discussion, ideas, and feedback have been absolutely invaluable.

I am thankful to my friends and family who made this journey bearable, meaningful and purposeful. In particular, I want to thank my wife Alana for her love, support, and constant encouragement and sacrifice throughout the past several years. Despite your dislike for anything winter related, you chose to move to New England. I am forever indebted to you.

Sexual partnership duration: characterizing sampling conditions that permit unbiased estimation of survivorship and effect on it of covariates

Yared Gurmu, Jing Qian, and Victor De Gruttola

1.1 Introduction

Estimating the duration of sexual partnerships is important in investigation of the epidemic dynamics of sexually transmitted infections (STI). Duration of such partnerships is a key feature in mathematical models of STIs and has been shown to be an important predictor of STI risk and of concurrency (Matson et al., 2012; Chen et al., 2008). Goodreau et al. (2012) utilize data on duration to model sexual partnership networks in their study of the roles of acute infection and concurrent partnerships in HIV transmission dynamics.Wang et al. (2014) used duration in modeling spread of HIV for the purpose of designing intervention studies. In a different application of duration information, Matson et al. (2012) investigated the association between concurrency and duration of relationships based on a prospective cohort that followed participants every six months. Using a multilevel mixed effect logistic regression, the study found that odds of concurrency (OR = 1.03, 95 % CI: [1.02,1.11]) increased with length of relationship.

Distributions of duration of relationships are often estimated retrospectively from surveys that collect information about the length of partnerships that are ongoing or have ended within a fixed period (typically 6 months or a year) before the date of the survey. This form of sampling yields data that are left truncated (because the relationship had to have endured long enough to be present within the time window before the survey) and potentially right censored, should the relationship be ongoing at the time of the survey. Several authors have considered the problem of right censoring and left truncation (RCLT) in analysis of survival or failure-time data. For right-censored observations, the survival time lies in an interval of the form $[C, \infty]$ where C is the censoring time. In contrast, left truncation arises from sampling of observations conditional on the failure time itself. Denoting T as the left truncation time and X as the survival time, X is observable only if $X > T$. All observations on subjects for whom $X < T$ are excluded from the observation process.

Kaplan and Meier (1958) discussed the problem of right censoring and left truncation ; and Tsai et al. (1987) described the asymptotic properties of the truncation product limit estimator (TPLE), the left truncated version of the Kaplan Meier estimator, in this set-

ting. These estimators are similar except in that the risk set of the TPLE is adjusted at each observed failure time x to include those subjects who have not yet failed and whose truncation time is less than x. Woodroofe (1985) demonstrated the consistency and weak convergence of the Lynden-Bell estimator–a product limit estimator with left truncation but no censoring. Tsai et al. (1987) and Woodroofe Woodroofe (1985) used empirical processes to show asymptotic properties. Later, Lai and Ying (1991) noted that the TPLE severely underestimates the true survival duration if the risk set is small at the start of follow-up. This results from the fact that subjects can only be at risk at failure time x_i if they have truncation time less than x_i. However, for small values of x_i it is possible to have only one subject in the risk set leading to a situation where the estimator takes a value of zero. In order to remedy this problem, the authors proposed a modified version of the TPLE and demonstrated its consistency and weak convergence to a Gaussian process using a counting process approach. Lai and Ying (1991) also relaxed Woodroofe's assumption that the event times need to follow a continuous distribution.

Wang (1989) proposed a more efficient estimator of the survival distribution that required a parametric assumption about the truncation distribution, noting that the TPLE is not a maximum likelihood estimator when the truncation distribution can be parameterized. Several other authors have addressed the problem of estimating the distribution function in the special case that initiation times have uniform distribution; the sampling in this case is referred to as length-biased (Vardi, 1982, 1989; Asgharian et al., 2002, 2005; Qin and Shen, 2010). With this additional assumption, one can estimate the distribution function for the failure time without conditioning on the truncation times as shown for example in Asgharian et al. (2002). Assuming uniformly distributed truncation times is equivalent to assuming that relationship initiation times follow a stationary Poisson process. For the relationship data described in the next section, the relationship initiation times are not likely to follow a Poisson process with constant intensity throughout a lifetime; people tend to initiate more relationships in their youth than in their later years.

One of the approaches that has been used to address length-biased sampling of relationship durations involves the calculation of harmonic mean of observed relationship durations (Goodreau et al., 2012). Since longer relationships are more likely to be observed

at a given point in time, the sample mean will tend to overestimate the true mean. One rationale for using harmonic mean is that it is always less than or equal to the arithmetic mean for positive random variables and therefore will tend to reduce the extent of overestimation of the mean. Sen (1987) has shown that the harmonic mean is unbiased and consistent estimator of the true mean in the setting of length-biased sampling. As this result only applies to uncensored length-biased data, harmonic mean may not provide a consistent estimator of the mean if the data are also censored.

Another approach to account for right censoring and left truncation present in partnership duration data was employed by Burington et al. (2010) using the TPLE. The TPLE assigns non-zero mass only at event times just as does the standard Kaplan-Meier estimator. By assigning mass in this way, the TPLE makes the additional assumption that no probability mass need be placed in intervals that lie between the left-end of a censoring interval and the left end of a truncating interval (see Figure 1.2). Assigning zero mass to such intervals avoids the problem of having multiple maxima of the nonparametric likelihood and thereby simplifies estimation. However, Frydman (1994) demonstrated that consistent estimation may require mass to be placed in regions where the TPLE does not. Frydman modified Turnbull's nonparametric estimator of the distribution function so that it correctly accommodates interval-censored and truncated data (Turnbull, 1976). She showed that support depends not only on the censoring intervals, as (Turnbull, 1976) described, but also on the truncation intervals. Thus, the implications of assumptions that restrict the support of the NPMLE, as does the TPLE, require more consideration. Further discussion illustrating the differences between the TPLE and Turnbull's estimator with Frydman's correction (TEFC) will be presented in later sections.

The primary aims of this paper are to identify sampling conditions necessary to obtain a consistent estimator of the distribution of partnership durations from retrospectively collected survey data and to apply these insights to an analysis of the relationship duration distribution. The rest of the paper is organized as follows. The next section describes the partnership duration dataset, and the sampling scheme that motivated the methods presented in this paper. Section 3 discusses the conditions under which the NPMLE of the relationship duration distribution for RCLT data is unique and consistent.

We present the conditions regarding the size of the sampling window and the censoring and truncation time distributions that are necessary for the TPLE to be consistent. Section 4 presents conditions for consistency of parameter estimates from a Cox proportional hazards model where distribution of duration is not completely identifiable due to restrictions on the support of the truncation distribution. Section 5 examines the validity of a key assumption–the quasi-independence of truncation time and failure time–that is also necessary for consistency of the TPLE from RCLT data. This section also discusses a spline regression model. Section 6 provides a discussion.

1.2 Relationship Data

1.2.1 Sampling Schemes for Partnership Data

Partnership surveys may collect information on a fixed number of most recent relationships or alternatively, information may be collected on all relationships that are ongoing during a fixed time window called the sampling window (Burington et al., 2010). In such partnership surveys, participants are repeatedly interviewed to provide detailed information regarding their prior partnerships including age at sexual debut as well as start and end time of their sexual partnerships. This general method of collecting cross-sectional partnership duration yields data that are length-biased, as longer relationships are more likely to be observed. One can avoid this observation bias only by prospectively following participants over their lifetime starting from the time of their sexual debut. Such a design would provide incident sampling that allows complete observation of partnerships, but is obviously infeasible.

We note that this sampling scheme for partnership studies often combines prevalent and incident sampling. In our application dataset from a pilot study in Mochudi, Botswana, data are collected cross-sectionally among those whose relationships have been ongoing within the 12 month period before the interview date. Such prevalent sampling schemes do not generally obtain information about the time of initiation of partnerships that ended before the window described above; hence the duration of such relationships are truncated. In contrast data on subjects who initiate partnerships after the left endpoint of the

window are observed as in incident sampling. Estimates of the distribution using only the latter are unbiased but are censored by the end of the sampling window.

1.2.2 Statistical Notation for Relationship Data

Let T_f and T_d represent the calendar time of relationship formation and dissolution, respectively and let τ represent the calendar time of interview. T_f and T_d are considered random while the interview date τ is taken to be fixed. The relationship duration times are defined as $X = T_d - T_f$. If the relationship is ongoing at the time of the interview (i.e. T_d is after τ), the length of X is censored at $C = \tau - T_f$ where C is known as the right censoring time. The observed duration variable is $Y = \min(X, C)$; we define $\delta = I(X \leq C)$. The truncation time is defined as follows:

$$T = \begin{cases} 0 & : \text{ if } \tau - w - T_f \leq 0 \text{ (for durations not subject to truncation)} \\ \tau - w - T_f & : \text{ if } \tau - w - T_f > 0 \text{ (for durations subject to truncation)} \end{cases}$$

where w is the length of the sampling window. If $X < T$, we do not observe the relationship. See Figure 1.1 for an illustration of the time course of partnerships and the sampling scheme.

The observed data are assumed to be realizations of independent and identically distributed (iid) (T_i, Y_i, δ_i) for $i = 1, ..., n$ that meet this sampling requirement. Note the special nature of the relationship between the truncation and censoring times, $C = T + w$ for all truncated durations. Throughout the rest of the paper, we will assume X follows the distribution function F.

1.3 Support Characterization and Consistency

This section characterizes the support for the estimator of the distribution of duration, $\hat{F}$, and investigates the conditions required for consistency in a variety of settings with discontinuities in the truncation and censoring distributions.

1.3.1 Turnbull's Estimator with Frydman's Correction (TEFC) and TPLE

Turnbull (1976) developed a nonparametric estimator of the failure-time distribution function of random variable X for arbitrarily censored and truncated data. Truncation

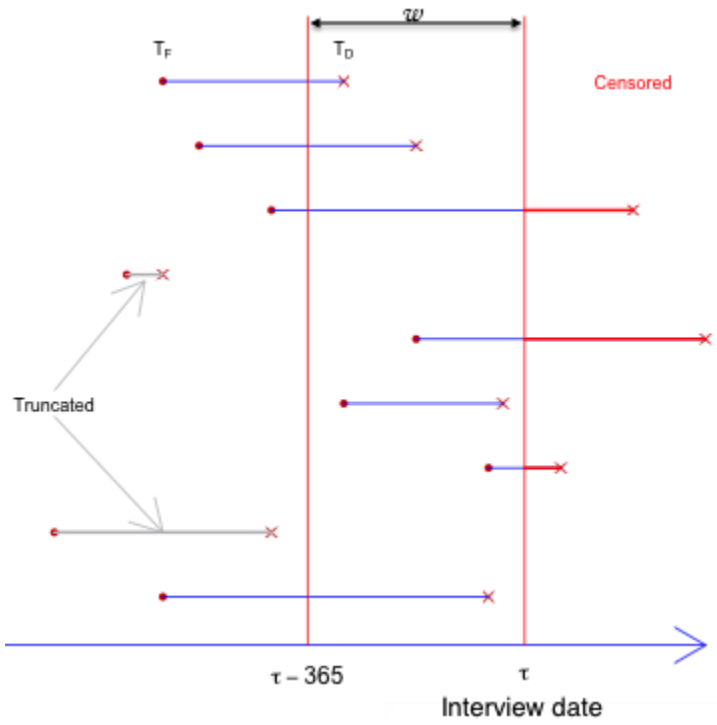


Figure 1.1: **Illustration of the time course of partnerships and the sampling scheme**. T_f, T_d and τ represent the calendar times of relationship formation, relationship dissolution, and interview, respectively. Partnerships ongoing within the sampling window, w, are observed; partnerships that end prior to the sampling window are not observed and therefore truncated. The sampling window in this figure is 1 year (365 days).

implies that independent observations are sampled from $F(x) = Pr(X \leq x | X \in B_i)$, where the set B_i is the truncation interval, defined as $B_i = [V_i, U_i]$. In the case of left truncation $B_i = [T_i, \infty)$. In addition, each X_i can be censored by $A_i = [L_i, R_i]$. In the case of exact observations $X_i = x_i$, we set $A_i = [x_i, x_i]$. For right censored data we have $A_i = [C_i, \infty)$. Note that the sets (A_i, B_i) are assumed to be independent of X_i. Turnbull's likelihood (Turnbull 1976) for the observed data is given by

$$L(F) = \prod_{i=1}^{n} \frac{P_F(A_i)}{P_F(B_i)}$$

which can be simplified as

$$L(\mathbf{s}) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{m} \alpha_{ij} s_j}{\sum_{j=1}^{m} \beta_{ij} s_j}. \tag{1.1}$$

where α_{ij} =I$([q_j, p_j] \in A_i)$, β_{ij} =I$([q_j, p_j] \in B_i)$, $q_j \in \mathcal{L} = \bigcup_{i=1}^{n} \{L_i, U_i\}$, $p_j \in \mathcal{R} = \bigcup_{i=1}^{n} \{R_i, V_i\}$, $s_j = F(p_j+) - F(q_j-)$ and $j \in \{1, \cdots, m\}$.

Frydman (1994) pointed out that the support of $\hat{F}$ is made up of the union of disjoint intervals $[q_j, p_j]$. Thus, finding the NPMLE of F reduces to maximizing the likelihood in equation (2.1) with respect to $\mathbf{s} = (s_1, \cdots, s_m)$. The constraints for maximization are $\sum_{i=1}^{n} s_j = 1$ and $0 \leq s_j \leq 1$. By construction the intervals $[q_j, p_j]$ cannot contain any other members of $\mathcal{R}$ or $\mathcal{L}$. Frydman's characterization of the support applies in the general case of interval censored, grouped as well as truncated data. For the special case of partnership data which are left truncated and right censored, the support of can be simplified as follows:

1. for exactly observed relationship duration, say x_i, we set $q_j = p_j = x_i$.

2. for right censored observation that is immediately followed by a left truncated observation, we set $q_j = c_i$, and $p_j = v_i$ provided $c_i < v_i$.

Note that in step (2), we obtain an interval where there could be non-zero mass. Henceforth, interval of this type will be referred as region where events are not observable (RENO).

By contrast, the TPLE estimator for the survival function, $S(x) = 1 - F(x)$, is given by

$$\prod_{y_{(i)} \leq x} (1 - \frac{\sum_{j=1}^{n} I(y_j = y_{(i)})}{D_i})$$

where $D_i = \sum_{j=1}^{n} I(t_j \leq y_{(i)} \leq y_j)$, $y_{(1)}, \cdots, y_{(k)}$ are distinct ordered observed failure times. From the form of the TPLE, it can be seen that the support of $\hat{F}$ is restricted to exact times of failure. Unlike the TPLE, Turnbull's NPMLE with Frydman's correction (TEFC) allows for placement of mass in RENO.

1.3.2 Example of disagreement between TEFC and and TPLE

We consider a simple numerical example to illustrate the difference between the support of the TEFC and TPLE. The data for the example are displayed in Figure 1.2. Partnership durations are reported from three individuals; two of them (observations 1 and 3 in the figure) are reported as $x = 4$ months and $x = 9$ months. These relationships are taken to be sampled conditional on being greater than 3 and 8 months, respectively. The censored

relationship (observation 2) is reported to be at least 6 months and is sampled conditional on being greater than 1 month.

As shown in Figure 1.2(a), application of steps (1) and (2) from the previous section suggests that TEFC put mass at $x = 4$ and $x = 9$ as well as the shaded region, [6,8]. Note that the left and right endpoints of the shaded region are a right censored event time that is followed by a left truncation time with no intervening events; there is no guarantee of there is not mass in this region. From Figure 1.2(a), the shaded region reflects knowledge that the relationship duration is greater than 6 months for observation 1. However, observation 3 implies that relationship durations less than 8 months, the truncation time, are possible.

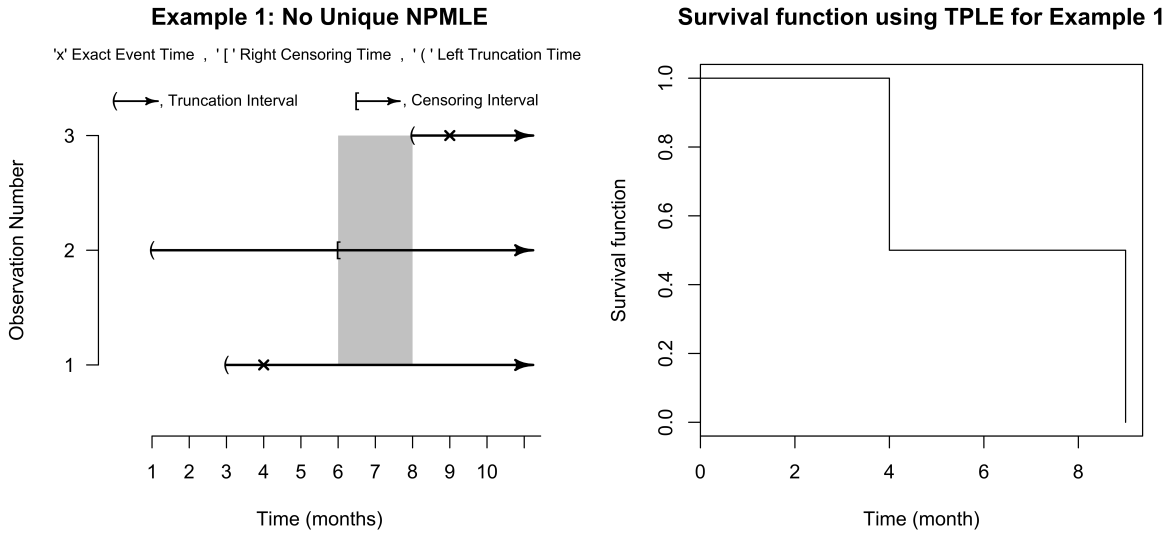
The likelihood for this example is:

$$L(\mathbf{s}) = \left(\frac{s_1}{s_1 + s_2 + s_3}\right)\left(\frac{s_2 + s_3}{s_1 + s_2 + s_3}\right)\left(\frac{s_3}{s_3}\right) \tag{1.2}$$

where $s_1, s_2,$ and s_3 are the masses at 4, $[6, 8]$, and 9, respectively. Since $s_1 + s_2 + s_3 = 1$, the likelihood simplifies to $L(\mathbf{s}) = s_1(s_2 + s_3)$, which is maximized at $s_1 = \frac{1}{2}$ and $s_2 + s_3 = \frac{1}{2}$. TEFC is not unique for this simple problem as there are infinitely many s_2 and s_3 satisfying the above constraints. The TPLE shown in Figure 1.2(b) is obtained from the above likelihood by setting mass in the shaded region to 0 (i.e. $s_2 = 0$).

1.3.3 Intrinsic versus Ignorable RENO and implications for consistency

In practice, our dataset can have two types of RENO, both of which are formed between a right censored observation and the left-truncation times that immediately follows it. An ignorable RENO has width that converges to 0 as $n \to \infty$; an intrinsic RENO does not. If the truncation distribution is continuous, the width of the RENO converges to 0 as $n \to \infty$, yielding ignorable RENO. For example the RENO shown in Figure 1 can be considered as ignorable as the width of the RENO (i.e. the shaded region) converges to 0 as $n \to \infty$. However, as illustrated by Figure 1.3.3, if the truncation distribution is discontinuous intrinsic RENOs can arise. In this setting, the support of X is continuous, the truncation support is discontinuous, and the censoring support is equal to the truncation support



(a) (b)

Figure 1.2: Example of disagreement between TEFC and TPLE for the data in Figure 1.2(a). The line segments starting with '(' correspond to the the truncation interval for each observation. The support of the TEFC includes the shaded region from $[6, 8]$ in addition to the support of the TPLE $\{4, 9\}$ which is shown in Figure 1.2(b).

shifted by a fixed constant $w < 0.5$. The width of the RENO in this example approaches $0.5 - w$ as $n \to \infty$; no values between $0.5 + w$ and 1 are observable irrespective of the sample size. Any observation x_i that falls within this window will either be censored to $0.5 + w$ if $t_i \in (0, 0.5)$ or will be excluded if $t_i \in (1, 1.5)$.

1.3.4 Consistency of the TPLE when there is intrinsic RENO

The examples above illustrate why there can be no unique NPMLE in settings with at least one RENO. In addition, $\hat{S}_{tple}(x)$ may not be a consistent estimator for the entire distribution of the duration variable when there are one or more RENOs. As illustrated by simulation in the next section, if a RENO exists between x_1 and x_2 then it is possible that

$$\sup_{x \in [0, x^*)} |\hat{S}_{tple}(x) - S(x)| \not\xrightarrow{p} 0,$$

where $x_1 \leq x_2 < x^*$ and x^* lies in the interior of the support of the duration distribution function F and the censoring distribution function G. The latter condition is equivalent

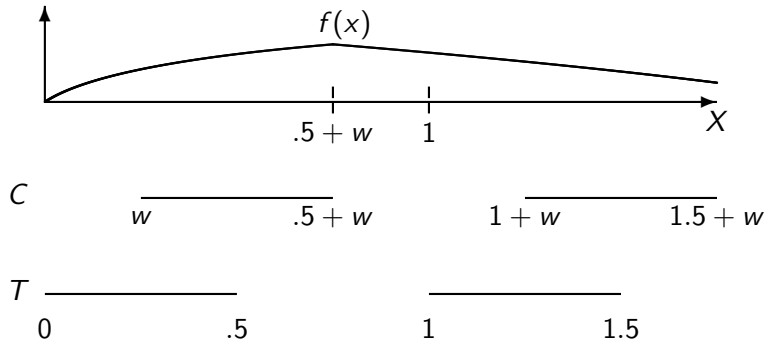


Figure 1.3: Example of Intrinsic RENO: X has continuous support but C and T do not. No duration values, x_i, between $.5+w$ and 1 are observable whenever the size of the sampling window is smaller than the size of the discontinuity in the support of the truncation times (i.e. $w \leq .5$).

to $1 - F(x^*) > 0$ and $1 - G(x^*) > 0$, and hence $F(x) > 0$ and $G(x) > 0$ for $0 \leq x \leq x^*$. Nonetheless, $\hat{S}_{tple}(x)$ has some useful asymptotic properties.

Theorem 1.3.1. *Given the presence of a RENO between x_1 and x_2 and the conditions $x_1 \leq x_2 < x^*$, $1-F(x^*) > 0$ and $1-G(x^*) > 0$, $\hat{S}_{tple}(x)$ is consistent for features of the distribution of $S(x)$ as follows:*

$$\sup_{x\in[0,x_1]} |\hat{S}_{tple}(x) - S(x)| \xrightarrow{p} 0 \text{ and } \sup_{x\in(x_2,x^*)} |\hat{S}_{tple}(x|x > x_2) - S(x|x > x_2)| \xrightarrow{p} 0$$

The proof is provided in the appendix and relies on the work of Tsai et al. (1987) and Lai and Ying (1991) who showed that the conditional distribution $S(x|X > a) = P(X > x|X > a)$ can be estimated consistently for $a \leq X \leq b$, where a is the lower boundary on the support of the left-truncation variable T and b is the upper boundary on the right-censoring variable C.

1.3.5 Simulation Study

We use simulation to illustrate the conditions under which the TPLE is consistent in the presence of RENO. In order to create a sufficiently large RENO, we will generate our dataset using the following procedure:

1. For durations subject to truncation (i.e. $T_i > 0$), truncation time is drawn from a mixture of two uniform distributions,

$$T_i = \begin{cases} V_i \text{ where } V_i \in \text{U}(0, 0.5) \text{ with probability } p \\ V_i \text{ where } V_i \in \text{U}(1, 1.5) \text{ with probability } 1-p \end{cases}$$

2. The duration variable will be sampled from the exponential distribution with CDF $1 - e^{-\lambda x}$ where $\lambda \in \{.25, .5, .75, 1\}$.

3. If $X_i \geq T_i$ keep data. Otherwise, discard the data since it is not possible to observe it.

4. Repeat steps 1 and 2 until $\frac{N}{2}$ truncated subjects are obtained.

5. For durations not subject to truncation, truncation time is $T_i = 0$. About 50% or $\frac{N}{2}$ of the durations will not be subject to truncation as is the case in our dataset obtained from Mochudi, Botswana.

6. Generate the censoring times as

$$C_i = \begin{cases} T_i + w & : \text{ if } T_i > 0 \text{ (for durations subject to truncation)} \\ V_i \in \text{U}(0, w) & : \text{ if } T_i = 0 \text{ (for durations not subject to truncation)} \end{cases}$$

   $C_i =$ for $i \in \{1, \cdots, N\}$ where w is the size of the sampling window.

7. Lastly, define $Y_i =$min(C_i, X_i) and $\delta_i =$I$[X_i \leq C_i]$.

The final dataset contains $(T_i, C_i, Y_i, \delta_i)$ for $i \in \{1, \cdots, N\}$. Note the width of the sampling window w is varied in order to control the size of the RENO and investigate its impact on consistency.

We consider the following set of parameters: $\lambda = 0.25, N = 10^4$ and $w \in \{.1, .2, \cdots, 1\}$. Figure 1.4 illustrates the difference between TPLE $S_{tple}(x)$ and $S(x)$ and the impact of the size of the RENO. We can estimate $S(x)$ unconditionally since the lower boundary of the support of the truncation times, $\tau^* = 0$.

In addition, the simulations suggest that even if there are discontinuities in the truncation and censoring support, the TPLE will converge to the true distribution provided that there is not an intrinsic RENO. Thus, it is important to distinguish between discontinuities in the support of T and C that do and do not result in intrinsic RENOs. This distinction can

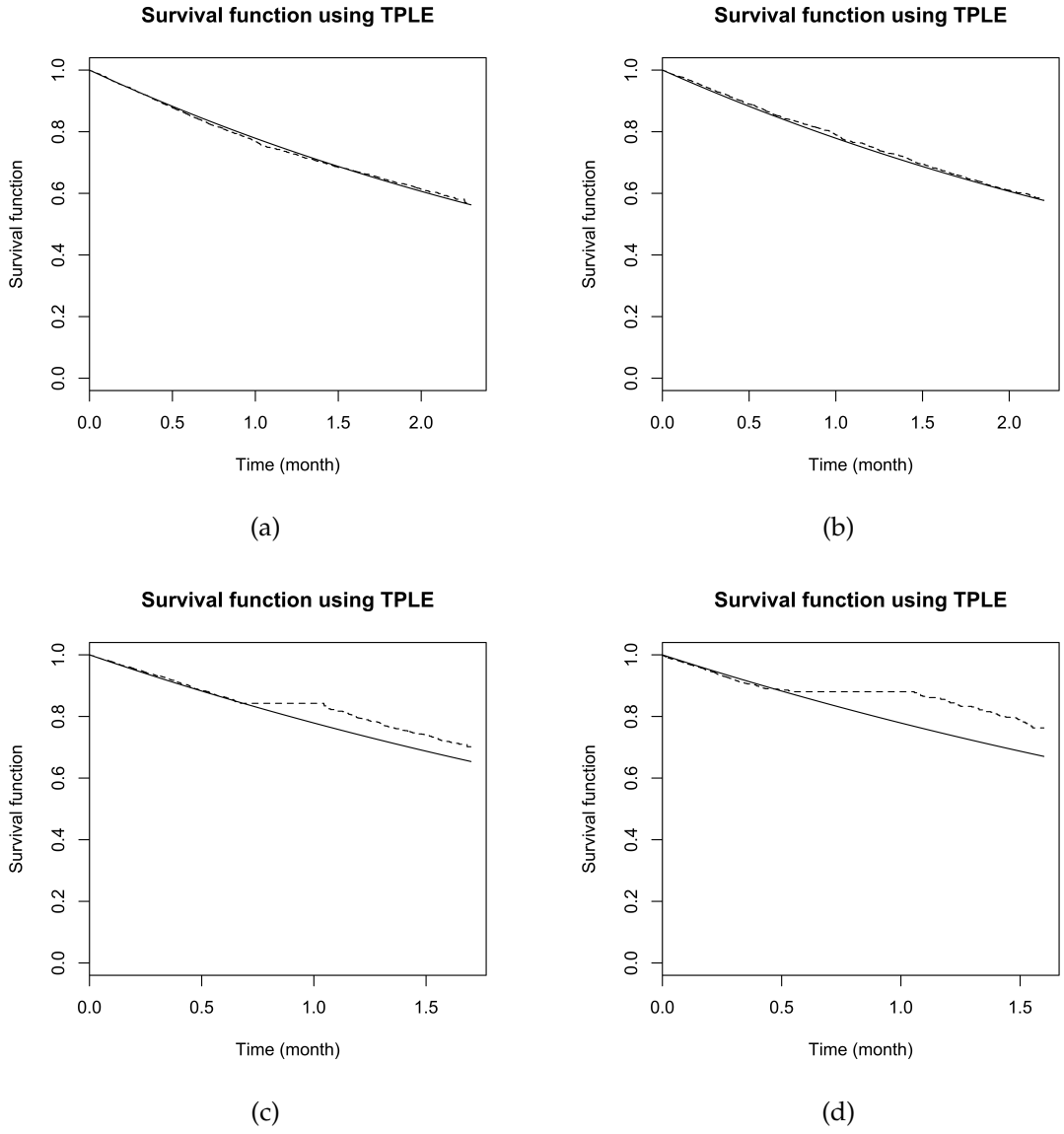


Figure 1.4: Performance of TPLE (dashed line) in the Presence of RENO: $\lambda = 0.25, N = 10^4$ and $w = .8, .7, .2, .1$ for figures (a), (b), (c), and (d), respectively. The true distribution of the relationship duration time is given by solid line. An intrinsic RENO is present whenever the size of the sampling window is smaller than the size of the discontinuity in the support of the truncation times (i.e. $w \leq .5$) as in (a) and (b).

be illustrated by comparing figures 2 and figures 1.5(a). In both, the support of C and T partially overlap, but intrinsic RENOs arise only in the setting displayed in Figure 1.3.3, because the setting of Figure 1.5(a) permits observation of duration of relationships that fall within the regions of discontinuity, namely [0.5,0.7] and [1.5,1.7]. Additional examples that illustrate the difference between ignorable and intrinsic RENOs are illustrated in

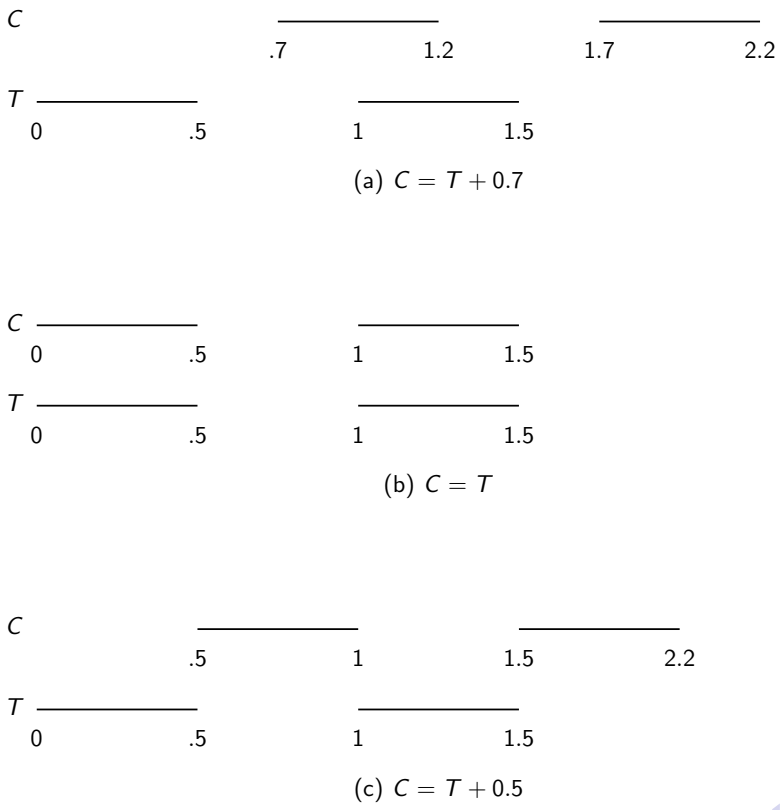


Figure 1.5: Discontinuities that do not lead to RENOs are shown in figures (a) and (c). The discontinuity shown in (b) leads to intrinsic RENO as any $x_i \in (.5, 1)$ is not observable.

figures 1.5(b) and 1.5(c).

1.4 Consistency of $\hat{\beta}$ from a Cox proportional hazard model

We examine conditions for the consistency of the estimated regression parameter, $\hat{\beta}$, from a Cox proportional hazards model. The Cox model assumes that the hazard function for X conditional on the covariate vector Z is given by

$$\lambda(x|Z) = \lambda_0(x)e^{\beta^T \mathbf{Z}} \tag{1.3}$$

where β^T is the vector of unknown coefficients and $\lambda_0(x)$ is the baseline hazard function that is independent of covariates. In the case of left-truncated X, the hazard function is

$\lambda(x|Z, X > T)$ can be simplified as

$$\lambda(x|Z, X > T) = \frac{f(x|Z, X > T)}{S(x|Z, X > T)} = \frac{\frac{f(x|Z)}{P(X>T|Z)}}{\frac{S(x|Z)}{P(X>T|Z)}} = \lambda(x|Z)$$

Additional conditioning on $X > T$ is not required as all event times are within the observable region ($X > T$). Below, we explain why it is possible to consistently estimate $\hat{\beta}$ even if $S(x)$ is not identifiable due to the presence of a RENO.

Assumptions necessary to establish the consistency of $\hat{\beta}$ are as follows:

1. (T_i, Y_i, δ_i) are i.i.d for $i \in \{1, \cdots, n\}$
2. $P(Y_i \geq T_i) > 0$
3. $X_i \perp\!\!\!\perp T_i$.
4. $X_i \perp\!\!\!\perp C_i$

Given these assumptions, the likelihood of the observed duration Y conditional on the truncation time, T is:

$$L \propto \prod_{i=1}^{n} \frac{f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}}{S(t_i)}$$

where f is the density function for the duration, S is the corresponding survival function, $\delta_i = I(x_i \leq c_i)$, $y_i = \min(x_i, c_i)$ and x_i, c_i, t_i are observed duration, censoring and truncation times, respectively. Wang et al. Wang et al. (1993) showed that the above likelihood can be factorized into the partial likelihood for β, $L_P(\beta)$ and a residual (ancillary) likelihood $L_R(\beta, \lambda_0)$ that contributes no information for the estimation of β. The partial likelihood $L_P(\beta)$ is given by

$$L_P(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta^T \mathbf{Z}_i}}{\sum_{j \in R(y_i)} e^{\beta^T \mathbf{Z}_j}} \right]^{\delta_i}$$

where $R(y_i)$ is the risk set given by $R(y_i) = \sum_{j=1}^{n} I(t_j \leq y_i \leq y_j)$. Estimators and large sample properties of β are derived based on $L_P(\beta)$. We let $a^{\otimes 0} = 1, a^{\otimes 1} = a, a^{\otimes 2} = aa^T$ and $S^{(r)}(\beta, x) = \frac{1}{n} \sum_{j=1}^{n} Q_i(x) \mathbf{Z}_j^{\otimes r} e^{\beta^T \mathbf{Z}_j}$ where $r = 0, 1$ and 2 and $Q_i(x) = I(t_i \leq x \leq y_i)$.

Also, let $v(\beta, x)$ be the limit of $V(\beta, x) = \frac{S^{(2)}(\beta,x)}{S^{(0)}(\beta,x)} - \left[\frac{S^{(2)}(\beta,x)}{S^{(0)}(\beta,x)}\right]^{\otimes 2}$ and $s^{(r)}(\beta, x)$ be the limit of $S^{(r)}(\beta, x)$. Now let

$$\sum = \int_0^\infty v(\beta, x) s^{(0)}(\beta, x) \lambda_0(x)\, dx$$

and assume $\sum$ is positive definite. Note this likelihood is similar to the usual right-censored data partial likelihood except for the difference in the definition of the risk set. Examining the form of the partial likelihood reveals that the presence of RENO will not affect the estimation of β since no observations are made within it. Therefore, following argument from Anderson et al. Andersen and Gill (1982), $\hat{\beta}$ can be shown to be consistent. The regularity conditions and additional assumptions as stated in Andersen and Gill (1982) are unaffected by the presence of RENO.

1.5 Application

1.5.1 Summary of the Mochudi Relationship Data

The relationship dataset we consider arose from an AIDS prevention pilot study conducted in the village of Mochudi, Botswana. The dataset contains information on 2268 subjects who reported at least one relationship. For the 376 subjects who had two or more sexual relationships in our dataset, we consider only the most recent, which is defined as the partnership with the most recent sexual contact. We restrict the analysis of partnerships to relationship durations that are 10 years or less; longer relationship are censored at 10 years as the reliability of recall beyond 10 years is uncertain (Nelson et al., 2010). Furthermore, for investigation of spread of STIs, it may be more useful to focus on relationships shorter than 10 years. There were 390 relationships that were greater than 10 years and only 8 (2%) of them had observed starting times. Therefore most of the information available for estimation of duration of relationships if rom relationships that are shorter than 10 years. In total, we have 2050 (90%) ongoing (censored) relationships at the time of the survey interview. In addition, 992 (44%) partnerships are untruncated since they began within in the sampling window of 12 months. Additional descriptive statistics pertaining to our population of study are provided in Tables 1 and 2.

Table 1.1: Quartiles of key covariates by HIV status

	HIV Negative N=1742	HIV Positive N=502	Combined
Age (Q1,Q2,Q3)*	(23,28,40)	(28,35,42)	(24,30,41)
Duration (X)	(.88,2,7)	(.99,3,7.99)	(.9,2,7)
Truncation (T)	(0,1,6)	(0,2,7)	(0,1,6)
Date of last sex**	(5,14,42)	(4,14,35)	(5,14,35)

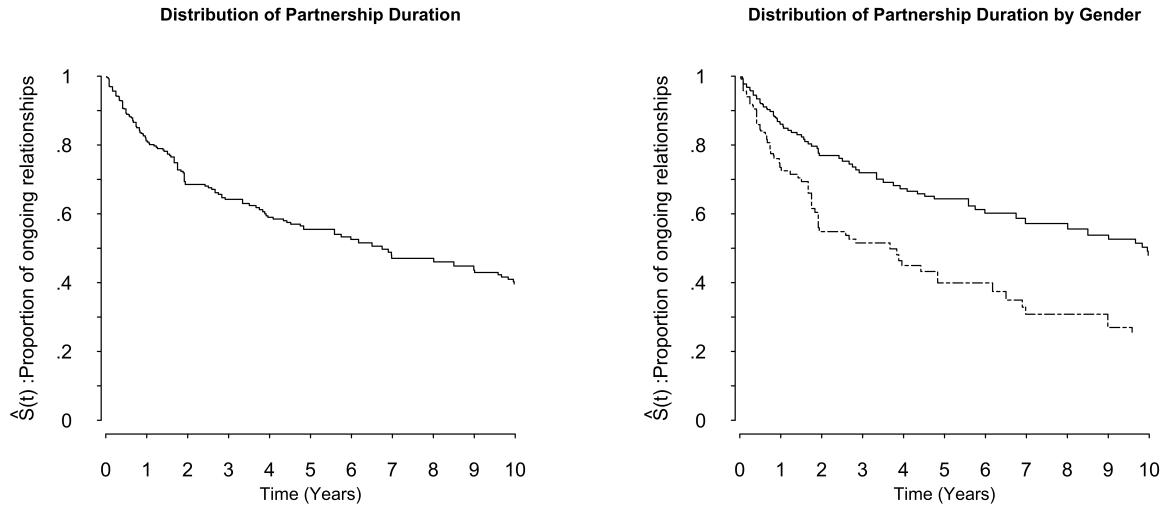
* Q1, Q2, Q3 refer to 25^{th}, 50^{th} and 75^{th} percentile of the variable measured in years.
** Refers to time (Days) from the last sexual contact to the interview date

Table 1.2: Median of key covariates by gender

	Male N=790	Female N=1478
Age (Median)	(24,31,43)	(24,30,40)
Duration (X)	(.66,1.4,6)	(.99,3,8)
Truncation (T)	(0,.48,5)	(0,2,7)
Date of last sex	(4,14,60)	(5,14,31)

1.5.2 Evaluating the quasi-independence assumption via Kendall's Tau

Validity of the TPLE depends on the quasi-independence assumption, i.e. the independence of the truncation and duration variables within the observable region. This assumption allows for the factorization of the joint density of the failure time and truncation time within in the observable region (Martin and Betensky, 2005). If this assumption of quasi-independence is violated, the construction of the likelihood may not be valid and the estimates derived from it may be biased. Keiding and Moeschberger (1992) showed that the nature of the bias in the product limit estimate depends on the correlation between truncation and event time. Unlike the independence assumption of censoring and failure times, it is possible to test for the independence of truncation and failure times nonparametrically (Martin and Betensky, 2005; Tsai, 1990). This assumption can be tested within the observable region (i.e. the region where $X_i \geq T_i$) since we have both pairs (X_i, T_i). The scatter plot of (X_i, T_i)(Figure 1.7) below provides a graphical check of the independence assumption of the truncation and duration times within the observable region. We examined Kendall's Tau statistic for the dependence of the duration and truncation variables and did not find evidence of association (Kendall's Tau Statistic=-0.005 ,



(a) (b)

Figure 1.6: (a) TPLE estimate of overall duration of duration of relationship (b) TPLE estimates of duration of relationship by gender; females (solid line) and males (dashed line).

p-value=0.0003). Conditions under which duration and truncation variables can be dependent are briefly discussed in the appendix.

1.5.3 Illustration of RENOs in the Mochudi relationship dataset

Figure 1.6(a) displays the TPLE estimate of the distribution of most recent partnerships. Applying the steps identified in Section 3.1, we observe several ignorable RENOs of various sizes with the largest RENO being around 55 days wide. As discussed above, ignorable RENOs do not affect consistency of estimation of the distribution. Figure 1.6(a) shows that 50% of the most recent relationships are at least 6.75 years long. Figure 1.6(b) shows a significant difference in the distribution of partnership durations by gender (p-value$<$.0001); the median duration of relationships for women and men were 9.95 and 3.67 years, respectively.

As discussed in prior sections, if the size of the sampling window w is less than the gap in the support of the truncation times we have a region in which consistent estimation of the distribution function for X is not possible. For the Mochudi relationship dataset $w = 365$

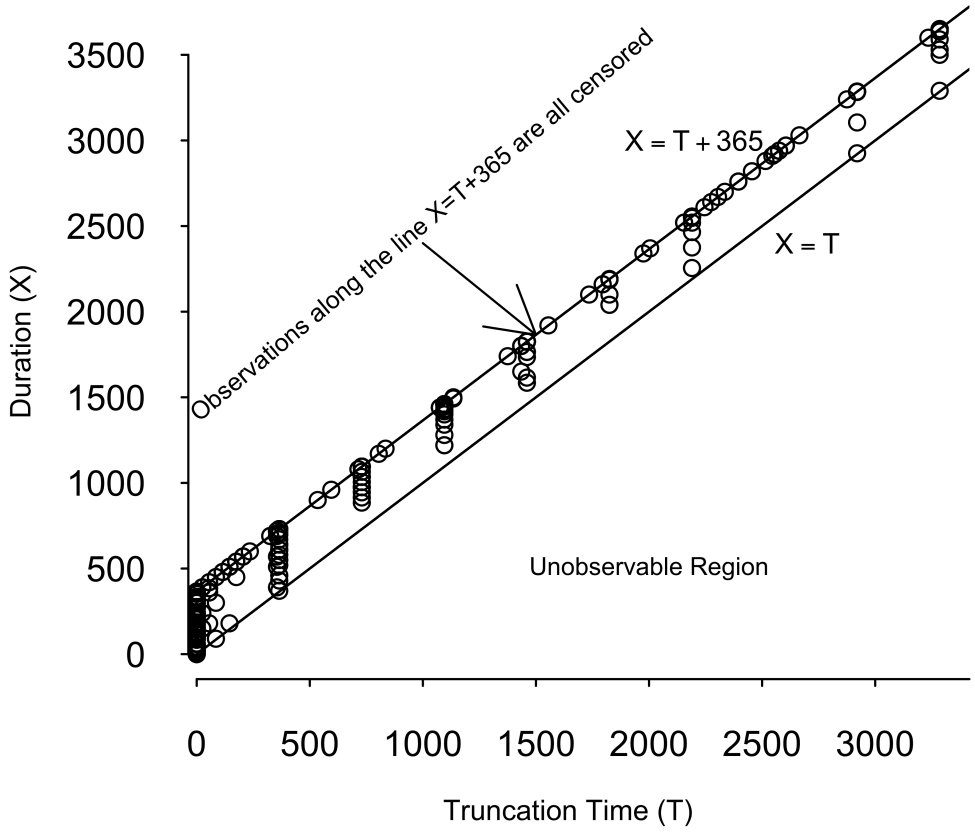


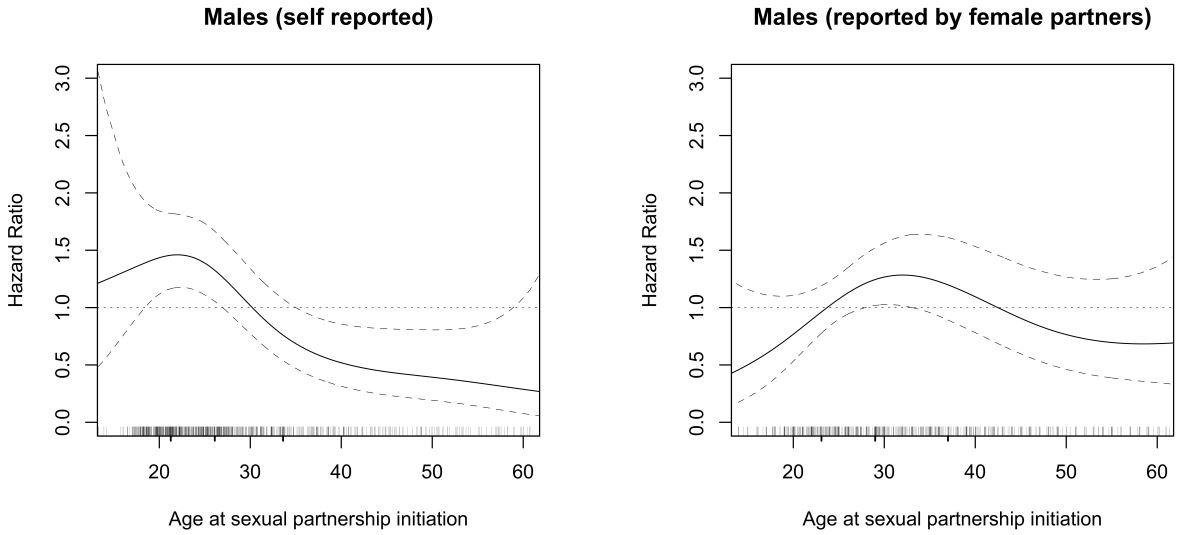
Figure 1.7: The observable region for the duration dataset

days, implying that intrinsic RENOs occur only if there is a gap that exceeds 365 days in the support of the truncation times or equivalently, the calendar time of relationship initiations. Examples of populations where such conditions might apply include those that experience mass circumcision campaigns that prevent young men from initiating relationships, or those where there is seasonal migration of young men due to work (i.e. farming, mining) (Lurie et al., 2003, 1997) or those that are characterized by cultural or religious norms that prohibit relationship formation for a period of time. Conditions that would lead to initiation time gap of 365 days or more are unlikely for the community from which this data is sampled, though it may be possible within age subgroups. Hence it is very unlikely that the consistency of estimation is compromised by the presence of RENOs.

1.5.4 Spline Modeling of Age Effect on Duration

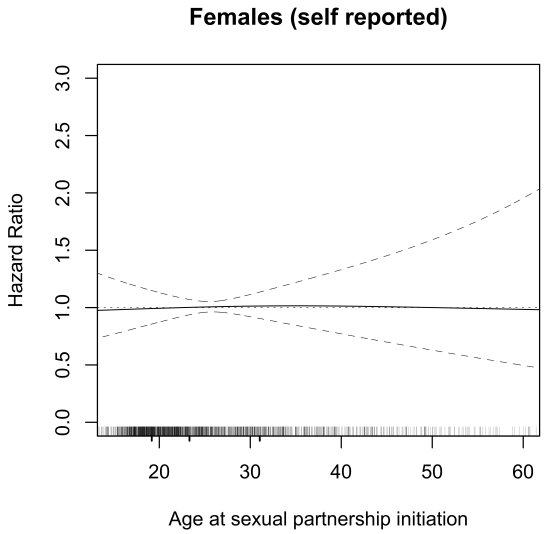
We use penalized smoothing spline to model the effect of age on the duration of relationship controlling for HIV and concurrency status. The analyses make use of the approach of Therneau and Grambsch (2000) to characterize linear and nonlinear effects of age on relationship duration. The results, displayed in Figures 1.8(a) and 1.8(c), show that the hazard of relationship termination decreases with age at relationship initiation among males. For this group, there is a significant linear, but not nonlinear, effect of age at relationship initiation (p=0.0006). However, for females (Figure 1.8(c)), the hazard of relationship termination does not vary much across age at relationship initiation. We see no significant linear (p=0.99) or nonlinear (p=0.24) effect of age at relationship initiation. We note that that the partnerships reported in our samples do not come from a closed population; hence there is no constraint that these curves be consistent.

The effect of the partner-reported age of males (i.e. reported by females) on the duration of relationship is very different from the effect of self-reported age of males (compare figures 1.8(a) and 1.8(b)). There was a significant linear association with age at partnership initiation for self-reported, but not partner-reported, age. A similar discrepancy is observed when comparing self-reported vs partner-reported age in figures 1.8(c) and 1.8(d). Lastly, it is worth noting that whereas figures 1.8(a) and 1.8(d) look similar as expected,

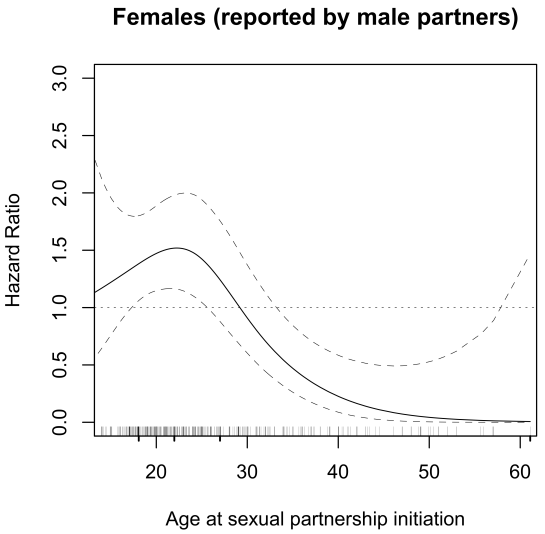


(a)

(b)



(c)



(d)

Figure 1.8: Spline regression modeling the effect of age at start of sexual partnership on duration. Dashed lines represent 95 % CI for Hazard Ratio while dotted lines represent where Hazard Ratio is equal to 1.

figures 1.8(b) and 1.8(c) do not, reflecting the fact that distribution of self-reported and partner-reported ages are similar for women but not for men.

1.6 Discussion

This paper identifies the sampling condition necessary to obtain a consistent estimator of distribution of partnership durations from retrospectively collected survey data; the partnership sampling window must be large enough to avoid potential gaps in the relationship formation times that may lead to regions where events are not observable. As shown analytically and via simulation, the presence of RENOs will lead to inconsistent estimation of the duration distribution function. Provided that truncation and duration times are independent within the observable region and that the size of the partnership sampling window is bigger than the gaps in the truncation times consistent estimation of the distribution of partnership duration is possible using either the standard TPLE or Lai and Ying's version of the TPLE. If the duration and truncation variables are dependent, consistent estimation is not possible, but these variables may be conditionally independent given a covariate like age. If the conditional independence assumption is satisfied, it is possible to obtain valid estimates of duration distribution conditional on age using TPLE. This paper also addressed how RENOs may influence estimation of covariate effects from a Cox proportional hazard model. The regression coefficients explaining the effect of a covariate on the hazard of relationship termination can be consistent even if consistent estimation of the distribution function is not possible provided that conditions listed in section 1.4 and the references therein are satisfied.

The steps for identifying RENOs have been outlined and illustrated by example. In addition, covariate effects on the distribution partnership duration are incorporated by using spline models. The results from our models show marked differences in the effect of age at initiation between men and women. Such gender difference have been noted in other populations in prior studies (Helleringer et al., 2011; Nnko et al., 2004). As noted in Nnko et al. (2004), one explanation for this phenomenon is that women may tend to underreport short non-marital relationships, but not more stable (e.g. marital) relationships. Our application dataset supports the tendency of women to report relationships with older men; the median of the reported age distribution for men (i.e. reported by women) is about 3 years greater than the median of this distribution reported by men. And these

relationships with older men are perhaps more likely to be stable and durable. Moreover, for the group of men the women in our study are reporting to have relationship with, the effect of age at partnership initiation is not associated with hazard of relationship termination. This is strikingly different from the association observed when analyzing the self-reported age at the initiation of relationships for the men in our study (compare figures 1.8(a) and (b)). In addition, for the group of women reported by the men in our study (note this group of women are not observed directly in our study), the effect of age at partnership initiation is associated with hazard of relationship termination. Again, this differs from the association observed when analyzing the self-reported data for the women in our study (compare figures 1.8(c) and 1.8(d)). Such differences may reflect the tendency of men to have shorter relationships outside of the community as well as the tendency of women to report only stable relationships.

Estimation and Modeling of Partnership Transition Probabilities

Yared Gurmu and Victor De Gruttola

2.1 Introduction

The rate of sexual partnership formation and dissolution is important in the investigation of dynamics of sexually transmitted infections (STI). These rates are an integral part of mathematical models that examine the epidemiology and control of STI (May et al., 1988; Anderson and Garnett, 2000). For example, the basic reproductive number of an STI, critical and useful metric associated with the size of an epidemic, is a function of sexual partner acquisition rate (Anderson and Garnett, 2000). Therefore, it is critical that these rates are properly estimated and modeled. The difficulty in properly estimating these rates lies in the fact that sexual history data is usually retrospectively collected and is often fraught with missing data; the total number of partnerships, the formation and dissolution time of relationships (consequently the duration of relationships) may be censored, truncated or subject to recall bias (Burington et al., 2010). Cross-sectional sexual history data is self reported and constitutes information on relationships that are currently ongoing, or have ended within a fixed time period before the sexual history interview date. Such sexual history data are inherently length-biased as longer relationships are more likely to be reported. In addition, such data can be right censored if the partnership is ongoing at the end of the partnership survey.

In order to address some of these challenges, this paper develops a Markov framework for the estimation and modeling of transition probabilities into and out of relationship states that characterize essential features of the incomplete sexual partnership process. Nelson et al. (2010) propose an imputation based framework to estimate partnership formation rates when data are obtained from incomplete sexual partnership history. Their estimation approach assumes that partnerships for whom start and end times were not obtained were uniformly distributed between the time of sexual debut and the earliest described partnership, and sensitivity analysis is employed to examine impact of this assumption on estimation of partnership formation rates. To avoid making explicit assumptions about the start and end time of unobserved relationships, this paper proposes a stochastic expectation maximization algorithm (stEM) coupled with rejection-sampling scheme in order to estimate transition rates from states of celibacy, monogamy, and con-

currency. This approach accommodates data from a retrospective sampling scheme and can utilize other available information from the sexual history data, such as total number of relationships within a specified period of time or total lifetime number of partnerships. Such information may be available even when the sojourn time of each of the partnerships in the different states is not.

This paper is organized as follows. The next section describes the motivating example, the Markov process formulation of the problem and various observation schemes under which sexual history data can be obtained. Section 3 discusses the stochastic EM (stEM) algorithm and its implementation, and section 4 presents simulation results for the performance of the stEM algorithm in our context. Section 5 discusses modeling of transition probabilities to permit dependence of the probabilities on time and other covariates. Section 6 presents applications to data obtained from a sexual behavior survey, and section 7 provides a discussion.

2.2 Markov Model Formulation for the Relationship Data

The motivating example arose from a survey that collected sexual behavior information from a cohort of HIV patients from an HIV Treatment and Care study in KwaZulu-Natal (KZN), South Africa. Information included age of sexual debut, the total lifetime number of partnerships, the duration and time of dissolution of partnerships. These data were collected biannually for a period of up to 3 years (i.e. 6 rounds of interviews were conducted in the process of collecting these data). The durations of all relationship that are still ongoing on the date of the last interview are right censored. Our interest lies in applying a Markov model to estimate rates of transition from states of celibacy, monogamy, and concurrency using incomplete sexual history data. One interesting feature of the sexual history survey instrument used in this study is that it obtains information on relationships that ended outside the sampling window if no relationships were reported to be ongoing during that window (i.e. 6 months prior to the baseline interview date). By contrast, most surveys of sexual history data gather information on duration and end time of relationships only for those relationships that are ongoing within the sampling window.

The sampling scheme used to collect data in the study we consider leads to additional duration information that can be utilized for efficient estimation of transition parameters of the Markov process.

The model shown in Figure 2.1 represents a three-state Markov process of relationship formation and dissolution. At the time of sexual debut a person enters state 1, monogamy. Then, depending on whether the person terminates the relationship or starts a new one the person will move to a state 0 (celibacy) or 2 (concurrency), respectively; information about the relationship formation and dissolution process is available until the time of the last survey. Concurrency is defined as having overlapping durations of partnerships in which sexual relationship with a new partnership is initiated before the termination of a preexisting sexual partnership. Note that all forward transitions in Figure 2.1 represent relationship formation, i.e., a transition from celibacy to monogamy or from monogamy to concurrency. Similarly, all backward transitions represent dissolution of a relationship, i.e. movement from concurrency to monogamy or from monogamy to celibacy.

This simple Markov model was chosen because it retains essential features of the relationship formation and dissolution process during the life course of an individual. Specifically, this approach allows for the estimation and modeling of the rates at which people move in and out of state of concurrency and other states of interest by using all available information outside the window. This is particularly of interest as concurrency may be an important risk factor for STI transmission as suggested by both mathematical models and empirical studies (Doherty et al., 2006; Morris and Kretzschmar, 1997; Watts and May, 1992; Koumans et al., 2001; Rosenberg et al., 1999).

2.2.1 Markov Model Notation

Let $X_i(t)$ represent the relationship status of person i at time t and let $\{X_i(t), t = 0, 1, \cdots, \mathcal{T}_i\}$ be the person's complete sexual history process until the date of the last survey,$\mathcal{T}_i$. A discrete time representation of the Markov process is used since the duration data collected is of discrete nature. The relationship states of the process $X(t)$ are $\{0, 1, 2\}$ corresponding to states of celibacy, monogamy and concurrency, respectively. Transition

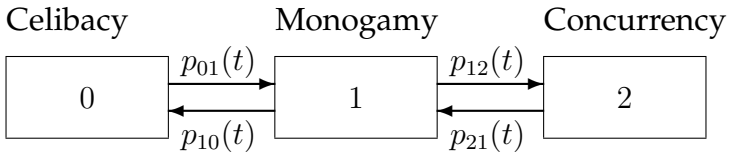


Figure 2.1: Markov Model Representation of Sexual History Data

from state to state is governed by the transition probability matrix

$$\boldsymbol{P(t)} = \begin{pmatrix} p_{00}(t) & p_{01}(t) & 0 \\ p_{10}(t) & p_{11}(t) & p_{12}(t) \\ 0 & p_{21}(t) & p_{22}(t) \end{pmatrix}$$

where $p_{ij}(t) = P(X(t+1) = j | X(t) = i) \; \forall \, t$.

2.2.2 Information available from sexual history data

Example of duration data and Markov model representation

Figure 2.2 illustrates the information available from surveys that collect retrospective partnership histories. We represent the situation in which an individual reported having two relationships within the sampling window (see Figure 2.2(a) and 2.2(b)), and one of the relationships is terminated before the end of the survey period, $\mathcal{T}$, while the other one is ongoing at the end of the survey period (i.e. right censored duration). We assume the availability of full information on $X(t)$ from $\mathcal{T} - w$ until the time of the last survey $\mathcal{T}$, where w is the total period over which the survey is conducted. Prior to $\mathcal{T} - w$ only partial information is available about the state of the relationship process. For example, in both figures 2.2(a) and 2.2(b) it is known that the first relationship started at T^*. Therefore, T^* is a relationship formation time, which implies $X(T^*) = 1$ and $X(T^* - 1) = 0$ or $X(T^*) = 2$ and $X(T^* - 1) = 1$. As T^* must be a transition time, either the individual has just entered either the state of monogamy or the state of concurrency. If an individual reports having two or more relationships at $\mathcal{T} - w$, then the last time of transition into a state of concur-

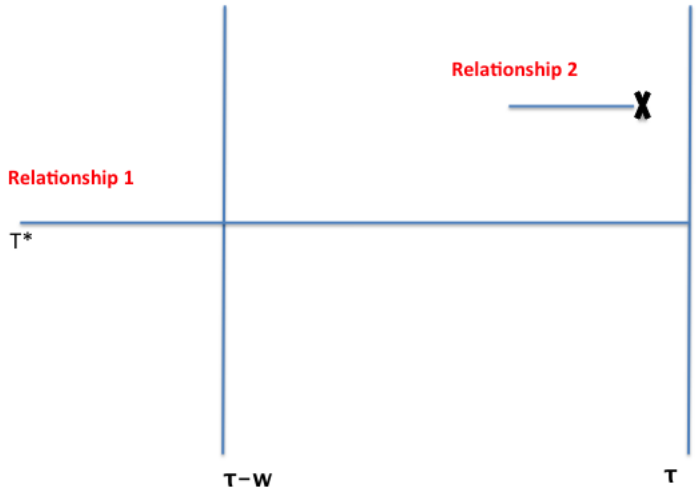
rency will also be known. In addition, as shown in Figure 2.3(b), the sample path of the process will be fully observed on the interval $[T_2^*, \mathcal{T}]$. Similarly, if an individual reports having no relationship at $\mathcal{T} - w$, then the individual is asked to report the duration and date of termination of his or her partnership. Figures 2.4(a) and 2.4(b) provide illustration of this scenario.

Lastly, Figure 2.2 also provides insight regarding the connection between the relationship data collected in partnership surveys and the Markov process representation of such data. Although relationship duration data represented in Figure 2.2(a) and 2.2(b) differ, they are both represented by the sample path as shown in 2.2(c). This figure illustrates that the relationship formation and dissolution process does not uniquely map to specific start and stop times of each relationship. Further discussion and a modeling framework that addresses this issue are presented in Chapter 3.

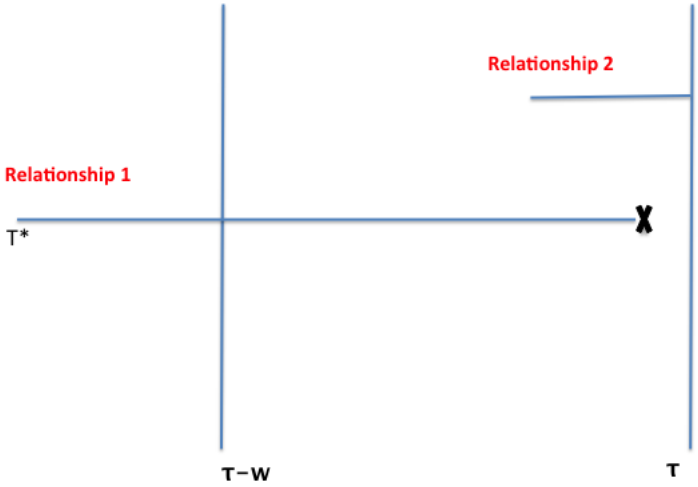
By design, partnership surveys collect sexual history data for relationships that are ongoing during the sampling window. Therefore, most such surveys omit data on sexual partnerships that end prior to the window. In the Markov process representation of the partnership data, this corresponds to not observing $X(s)$ for all $s < t$ if $X(t) = 0$ prior to the sampling window, $\mathcal{T} - w$. However, it is important to note that such missing data only occurs prior to the sampling window; within in the sampling window, $X(t)$ is fully observed. In the context of time-homogeneous Markov processes, such missingness does not lead to biased estimates of the transition probabilities as unbiased estimation can be performed by combining data across the sampling windows (Bee, 2005; Guttorp and Minin, 1995). However, the estimates obtained by using data from the sampling window only do not make as efficient use of available information as those estimates that make use of retrospective data.

2.3 Methods for estimating transition probabilities

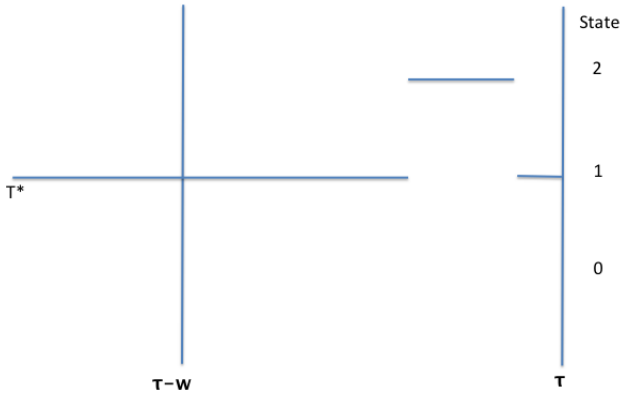
This section discusses estimation of transition rates assuming a Markov process, when the process is partially observed. We begin with a discussion of how maximum likelihood estimation (MLE) of the transition probabilities proceeds in the case of the complete data.



(a)

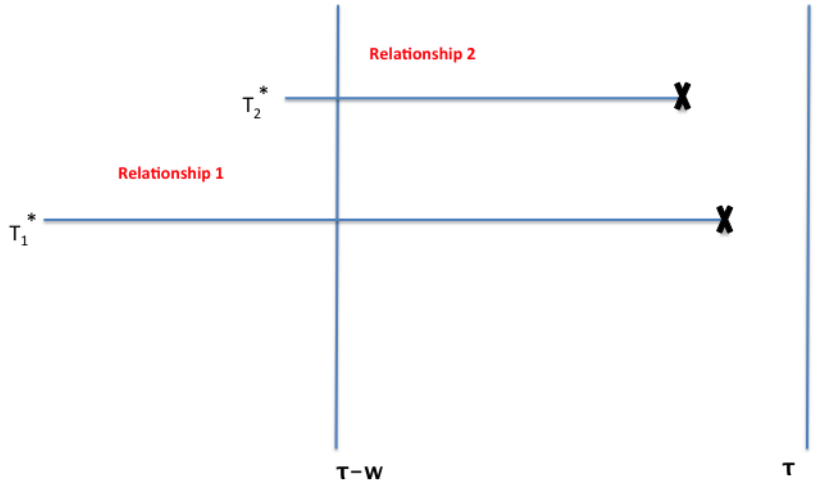


(b)

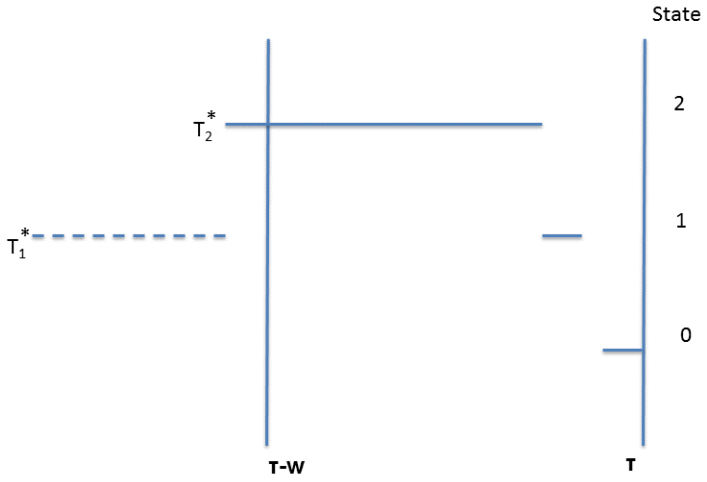


(c)

Figure 2.2: Two different duration data (a) and (b) have identical sample path ($1 \to 2 \to 1$) as shown in (c)



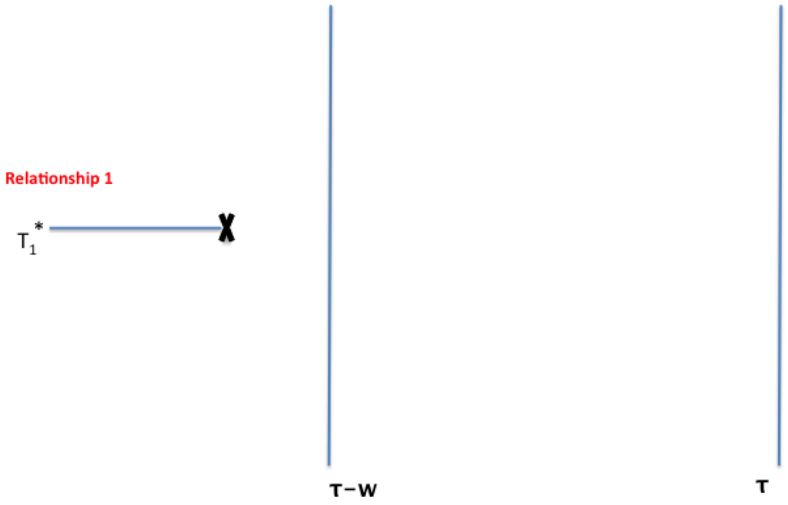
(a)



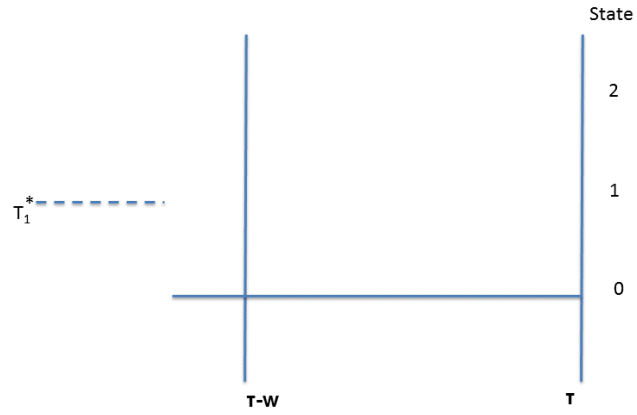
(b)

Figure 2.3: Duration representation (a) and corresponding Markov representation (b). Dashed line in (b) reflects the uncertainty in the state of the process between T_1^* and T_2^*

The imputation of missing data and calculation of MLEs will make use of a stochastic version of the expectation-maximization algorithm (EM) of Dempster et al. (Dempster et al., 1977) and Celeux et al. (Celeux and Diebolt, 1985; Gilks et al., 1996). The proposed version of the stochastic EM incorporates a rejection-sampling feature to draw samples conditional on the observed data.



(a)



(b)

Figure 2.4: Duration representation (a) and corresponding Markov representation (b). Dashed line in (b) reflects the uncertainty in the state of the process

2.3.1 Maximum Likelihood Estimation for Complete Data

Let $\{X_i(t), t = 0, 1, \cdots, \mathcal{T}_i\}$ represent the complete sample path for individual i. Individual i's contribution to the complete data likelihood can be formulated as follows:

$$
\begin{aligned}
L_i(\mathbf{p}) &= P(X_i(0) = x_0, X_i(1) = x_1, \cdots, X_i(\mathcal{T}_i) = x_{\mathcal{T}_i}) \\
&= P(X_i(0) = x_0) * P(X_i(1) = x_1 | X_i(0) = x_0) \\
&\cdots * P(X_i(\mathcal{T}_i) = x_{\mathcal{T}_i} | X_i(\mathcal{T}_i - 1) = x_{\mathcal{T}_i - 1}, \cdots, X_i(0) = x_0) \\
&= P(X_i(0) = x_0) * P(X_i(1) = x_1 | X_i(0) = x_0) \\
&\cdots * P(X_i(\mathcal{T}_i) = x_{\mathcal{T}_i} | X_i(\mathcal{T}_i - 1) = x_{\mathcal{T}_i - 1}) \\
&\propto \prod_{t=0}^{\mathcal{T}_i - 1} P(X_i(t+1) = x_{t+1} | X_i(t) = x_t) = \prod_{j=0}^{S} \prod_{i=0}^{S} p_{ij}^{N_{ij}(\mathcal{T}_k)}
\end{aligned}
$$

where $N_{ij}(\mathcal{T}_i) = \sum_{t=0}^{\mathcal{T}_i - 1} I(X_t = i, X_{t+1} = j) = \#\{i \text{ to } j \text{ transitions until } \mathcal{T}_i\}$ and the process transitions through states $\{0, 1, \cdots, S\}$. Note that the above likelihood is a product of multinomial likelihoods given the current state with parameters $N_{ij}(\mathcal{T}_i)$ and p_{ij}. Therefore, conditioning on the initial state probability, $P(X_i(0) = x_0)$, the MLEs can be calculated as

$$
\hat{p}_{ij} = \frac{N_{ij}(\mathcal{T}_i)}{\sum_{j=0}^{S} N_{ij}(\mathcal{T}_i)}
$$

And for K independent observations of the process, the individual likelihoods can be combined to get the MLEs as follows:

$$
\hat{p}_{ij} = \frac{\sum_{k=1}^{K} N_{ij}(\mathcal{T}_k)}{\sum_{k=1}^{K} \sum_{j=0}^{S} N_{ij}(\mathcal{T}_k)}
$$

Asymptotic properties of the MLEs are shown in Guttorp (Guttorp and Minin, 1995) and Billingsley (Billingsley, 1961).

If the entire process had been observed for each subject in the survey, the transition rates could be estimated explicitly using the simple formulas shown above. However, if the process was observed only at time points $\{t_0, t_1, \cdots, t_N\} \subseteq \{0, 1, \cdots, \mathcal{T}_i\}$ the observed data likelihood is:

$$
\prod_{n=0}^{N} P(X_i(t_{n+1}) = x_{t_{n+1}} | X_i(t_n) = x_{t_n})
$$

is not a simple function of the one-step transition parameters, p_{ij}. Rather, $P(X_i(t_{n+1}) = x_{t_{n+1}} | X_i(t_n) = x_{t_n}) = \sum_{x_{t_n}, \cdots, x_{(t_{n+1}-1)}} p_{x_{t_n} x_{(t_n+1)}} \cdots p_{x_{t_{n+1}} x_{(t_{n+1}-1)}}$ which involves $t_{n+1} - t_n - 1$ integrals. Therefore, maximization of the observed data likelihood may be cumbersome as $t_{n+1} - t_n - 1$ gets larger, because simple closed form solutions do not exit. In particular,

procedures such as PROC NLP (in SAS) for constrained maximization of the nonlinear equations arising from the observed data likelihood may occasionally lead to negative estimates of the transition probabilities as reported in Yeh et al. (Yeh et al., 2010). Because the complete data likelihood has simple closed form MLEs, the problem of maximizing the observed data likelihood lends itself well to a stochastic version of the EM algorithm that uses rejection sampling to calculate MLEs under different sampling schemes.

2.3.2 Stochastic EM algorithm

The E-step of the EM algorithm may often be intractable or cumbersome due to high dimensional integration. In addition, the EM estimate may be sensitive to the initial value of the unknown parameter. Given these difficulties, (Celeux and Diebolt, 1985; Gilks et al., 1996) provide a stochastic version of the EM algorithm as an alternative to standard EM algorithm. For a general discussion of the deterministic EM algorithm in the context of observational data arising from discrete time Markov chain, we refer the reader to Sherlaw-Johnson et al. (1995).

The stochastic EM (stEM) algorithm replaces the E-step with a stochastic step (S-step) whereby a sample is drawn from the conditional distribution of the missing data given the current parameter estimate and the observed data. Then, combining the observed data with the sample drawn at the S-step, a pseudo-complete dataset is obtained which can be maximized to obtain a new parameter estimate for the next iteration. The Stochastic step and the Maximization step are iterated until convergence. Thus, the stEM algorithm is implemented by alternating the following steps:

1. Find an initial value of $\mathbf{p}$, $\mathbf{p^{(0)}}$

2. *S (Stochastic) -Step*: at the k^{th} step, draw samples from $f(\mathbf{Z}|\mathbf{Y}, p^{(k)})$–the distribution of the missing data given the observed data and the current estimate of the parameter $\mathbf{p^{(k)}}$

3. *Maximization Step:* determine the new estimate $\mathbf{p^{(k+1)}}$ as the maximizer of the complete data likelihood in step 2.

4. Iterate steps 2 and 3 until convergence

As discussed in Diebolt and Ip (1996); Diebolt and Celeux (1993) the sequence of $p^{(k)}$ obtained from iterating the above 2 steps is a Markov chain that converges to a stationary distribution that is centered around the MLE of our parameter. The stochastic EM (stEM) estimate, denoted by $\tilde{p}$, is obtained from the mean of the sequence of $p^{(k)}$ after an appropriate burn-in period. In the applications discussed in this paper, the stEM estimate is calculated by averaging estimates from the last 1000 iterations. In addition, in order to allow the sequence of $p^{(k)}$ to settle and approach the stationary distribution, a burn-in period of 100 iterations is used. As discussed in Diebolt and Ip (1996), and the references therein, the stEM and EM estimators are asymptotically equivalent.

Outlined below is a version of the Stochastic EM algorithm in which the S-step implements a rejection sampling approach so a draw can be made from $f(\mathbf{Z}|\mathbf{Y}, p^{(k)})$. Note that given $\mathbf{Y}$ and a value for $p^{(k)}$ the missing data distribution is completely determined. Therefore, the S-step eliminates the need to evaluate complicated nonlinear likelihood expressions that appear in the observed data likelihood and the E-step of the deterministic EM algorithm.

Stochastic EM steps

Let T^* be the earliest observed non-zero time of realization of the process. In the KZN partnership data, T^* can be a partnership initiation time of the earliest reported relationship. For simplicity of notation, we drop the subscript i.

1. S-step:

   Given $\mathbf{Y}, T^*$ and $\mathcal{T}$, generate a new process, $Z(t)$, with initial transition probability $\mathbf{p} = \mathbf{p_0}$, which can be set to the MLE using data only on the process observed within the sampling window from $t = \mathcal{T} - w$ to $t = \mathcal{T}$. Rejection sampling is employed to assure that $Z(t)$ satisfies the following matching criteria:

   (a) $Z(0) = Y(0)$. (The initial state of the partnership process is known.)

   (b) $Z(T^*) \in \{1, 2\}$ since T^* is formation time.

   (c) $Z(t)$ is identical to $Y(t)$ within the sampling window $\{\mathcal{T} - w, \mathcal{T} - w + 1, \cdots, \mathcal{T}\}$.

(d) if $Y(\mathcal{T}-w)=2$ then $Z(t)=Y(t)$ for $t\in\{T_2^*,T_2^*+1,\cdots,\mathcal{T}\}$ where T_2^* is the last transition time to state 2 before $\mathcal{T}-w$.

(e) Match the process at the earliest time of sexual contact T_{init}. In other words, $Z(t)=X(t)$ for $t\leq T_{init}$.

(f) If the process entered state 0 prior to $\mathcal{T}-w$ and is in state 1 or 2 at $\mathcal{T}-w$, $Z(t)\in\{1,2\},\forall t\in\{T^*,T^*+1,\cdots\mathcal{T}-w\}$ where T^* is defined as above.

(g) If the process is in state 0 at $\mathcal{T}-w$ ($X(\mathcal{T}-w)=0$), then the $1\rightarrow 0$ transition time and the start time of this relationship will be known: individuals report the duration and termination date of their last partnership if they report no relationship on the date of the baseline interview ($\mathcal{T}-w$). $Z(t)$ and $Y(t)$ will be identical from the time state 0 is entered most recently.

(h) The total number of formation times ($0\rightarrow 1$ and $1\rightarrow 2$ transitions) for $Z(t)$ has to match the total life-time number of partnerships reported in the partnerships dataset.

2. M-Step:

   Maximize the likelihood of the pseudo-complete data comprising the imputed process ($\mathbf{Z}$) and the observed process ($\mathbf{Y}$) as follows:

$$\mathbf{p}^{(1)}=\max_{\mathbf{p}}(\ \log(\mathcal{L}(\mathbf{p};(\mathbf{Z},\mathbf{Y})|\mathbf{Y},\mathbf{p}^{(\mathbf{k})})))$$

Further details on the above matching criteria can be obtained from section 2.2.2. The matching criteria of this procedure are general enough that they can accommodate various types of partnership surveys. For example, if a partnership survey does not collect the time of earliest sexual contact or it is missing from the data collected, part (e) of the S-step can be omitted assuming first sexual initiation occurs after the age of 15 (time origin of the Markov chain). Alternatively, if researchers collect sexual history data only if the individual has ongoing relationships within in the sampling window, part (e) and (f) of S-step can be omitted in order to account for the lack of information outside the sampling window.

Investigating information on the total number of partnerships

If the lifetime number of partnerships is provided in the partnership survey, that information can be utilized in maximum likelihood estimation because of its inclusion in the sufficient statistic for the Markov process. The sufficient statistics for the parameter p_{ij} of our Markov model are the number of i to j transitions and total number of transitions out of state i. Specifically, knowing the total lifetime number of partnerships is equivalent to knowing the total number of $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions assuming all new concurrent relationships are formed from the state of monogamy and not within the state of concurrency. In other words, if a person is in a state of concurrency the only way to form another new relationship is first to break up the concurrent relationship, enter the state of monogamy, and then enter the state of concurrency. Without making this assumption, there does not seem to be a direct connection between the sufficient statistic for our Markov model and information on the life time number of relationships. If this assumption is not tenable, other approaches, which include a larger number of states, must be developed as described in the discussion section.

The Markov assumption simplifies sampling from the missing data distribution given the observed data, $f(\mathbf{Z}(\mathbf{t})|\mathbf{Y}, p^{(k)})$. Furthermore, rejection sampling by utilizing information on the total number of forward transitions restricts list of possible sample paths even when sojourn times are not known.

2.3.3 Variance of the stEM estimate

Variance estimation relies on an approach of Louis (1982) who formalized the missing information principle of Orchard et al. (1972) in the context of the EM algorithm. This principle can be summarized as:

Observed information=Complete information-missing information

Applying this principle, Louis provides the following identity:

$$\frac{\partial^2 l_{obs}(p;y)}{\partial p\, \partial p^T} = \mathrm{E}_p\left(\frac{\partial^2 l_c(p;x)}{\partial p\, \partial p^T}\Big|y\right) - \mathrm{cov}_p\left(\frac{\partial l_c(p;x)}{\partial p}\Big|y\right)$$

where l_{obs} and l_c are the log likelihood of the observed and complete data, respectively, and p is a vector of the parameters of interest. Given the above identity, parametric bootstrap approach (Efron and Tibshirani, 1994) can now be used to calculate the theoretical covariance and mean functions of the information and score of the complete data likelihood. The following steps are implemented in the calculation of the variance of stEM estimator $\tilde{p}$:

1. Generate bootstrap samples

   (a) For each observed process $\mathbf{Y_i}$, simulate a realization of the process from $f(\mathbf{Z}|\mathbf{Y}, \tilde{p})$ to get a pseudo-complete process $\mathbf{X_i} = \{X_i(t), t = 0, 1, \cdots, \mathcal{T}_i\}$.

   (b) Repeat 1a for all reported partnerships $i = 1, \cdots, K$ to obtain a pseudo-complete dataset.

   (c) Generate M pseudo-complete datasets by repeating the above steps.

2. Calculate the sample mean, the estimate of the first term on the right hand side of the identity, as:

$$\frac{1}{M}\sum_{m=1}^{M}\left(\frac{\partial^2 l_c^m(p;x)}{\partial p\,\partial p^T}\right)$$

   where $l_c^m(p;x)$ is the complete data log likelihood for the m^{th} simulated process and is given by

$$l_c^m(p;x) \propto \sum_{k=1}^{K}\sum_{j=0}^{2}\sum_{i=0}^{2} p_{ij}^{N_{ij}(\mathcal{T}_k)}$$

The elements of the information matrix necessary for the above computation are:

$$\begin{aligned}
\frac{\partial^2 l_c(p;x)}{\partial p_{01}^2} &= \sum_{k=1}^{K} - \left(\frac{N_{01}(\mathcal{T}_k)}{p_{01}^2} + \frac{N_{00}(\mathcal{T}_k)}{(1-p_{01})^2} \right) \\
\frac{\partial^2 l_c(p;x)}{\partial p_{10}^2} &= \sum_{k=1}^{K} - \left(\frac{N_{10}(\mathcal{T}_k)}{p_{10}^2} + \frac{N_{11}(\mathcal{T}_k)}{(1-p_{10}-p_{12})^2} \right) \\
\frac{\partial^2 l_c(p;x)}{\partial p_{10}\, \partial p_{12}} &= \sum_{k=1}^{K} - \frac{N_{11}(\mathcal{T}_k)}{(1-p_{10}-p_{12})^2} \\
\frac{\partial^2 l_c(p;x)}{\partial p_{12}^2} &= \sum_{k=1}^{K} - \left(\frac{N_{12}(\mathcal{T}_k)}{p_{12}^2} + \frac{N_{11}(\mathcal{T}_k)}{(1-p_{10}-p_{12})^2} \right) \\
\frac{\partial^2 l_c(p;x)}{\partial p_{21}^2} &= \sum_{k=1}^{K} - \left(\frac{N_{21}(\mathcal{T}_k)}{p_{21}^2} + \frac{N_{22}(\mathcal{T}_k)}{(1-p_{21})^2} \right)
\end{aligned}$$

3. Calculate the sample covariance, the estimate of the second term on the right hand side of the identity, as:

$$\frac{1}{M} \sum_{m=1}^{M} \left[\frac{\partial l_c^m(p;x)}{\partial p} - \bar{l} \right] \left[\frac{\partial l_c^m(p;x)}{\partial p} - \bar{l} \right]^T$$

where

$$\bar{l} = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial l_c^m(p;x)}{\partial p}.$$

The score equations necessary for calculating the sample covariance are:

$$\begin{aligned}
\frac{\partial l_c(p;x)}{\partial p_{01}} &= \sum_{k=1}^{K} \left(\frac{N_{01}(\mathcal{T}_k)}{p_{01}} - \frac{N_{00}(\mathcal{T}_k)}{1-p_{01}} \right) \\
\frac{\partial l_c(p;x)}{\partial p_{10}} &= \sum_{k=1}^{K} \left(\frac{N_{10}(\mathcal{T}_k)}{p_{10}} - \frac{N_{11}(\mathcal{T}_k)}{1-p_{10}-p_{12}} \right) \\
\frac{\partial l_c(p;x)}{\partial p_{12}} &= \sum_{k=1}^{K} \left(\frac{N_{12}(\mathcal{T}_k)}{p_{12}} - \frac{N_{11}(\mathcal{T}_k)}{1-p_{10}-p_{12}} \right) \\
\frac{\partial l_c(p;x)}{\partial p_{21}} &= \sum_{k=1}^{K} \left(\frac{N_{21}(\mathcal{T}_k)}{p_{21}} - \frac{N_{22}(\mathcal{T}_k)}{1-p_{21}} \right)
\end{aligned}$$

Note that all the steps in the parametric bootstrap are carried out setting $p = \tilde{p}$, the stEM estimate. Finally, the variance covariance of the stEM estimates can be calculated as usual

by inverting the estimate of the observed information obtained above.

2.4 Time Dependent Markov Chain and Modeling of Transition Rates

The partnership formation and dissolution process may not be homogeneous over time. In order to incorporate time variation, we model dependence on time according to the following multinomial logistic model:

$$log\frac{p_{ij}(t,t+1)}{p_{ii}(t,t+1)} = \alpha_{ij} + \beta_{ij} * t$$

where $i, j = 1, \cdots, S, i \neq j$ and $\sum_j p_{ij}(t,t+1) = 1$ for $t \in \{0, 1, \cdots, \mathcal{T}-1\}$. Note that the above model implies

$$p_{ij}(t) = p_{ii}(t) * exp\{\alpha_{ij} + \beta_{ij} * t\}$$

where

$$p_{ii}(t) = \frac{1}{1 + \sum_{j,j\neq i} exp\{\alpha_{ij} + \beta_{ij} * t\}}$$

In the setting of partnership survey data, the time origin 0 can correspond to the age at which the sexual history process begins. In sexual behavior surveys, this age is usually around 15 years old (Nelson et al. (2010), Mochudi Pilot Study). Therefore, we can incorporate the effect of age in our modeling by noting that $t = age - 15$ and each additional unit of time changes the logit of the transition probability by β_{ij}.

The stEM procedure can be used to estimate the time-dependent transitions of the parameter. The complete data likelihood can be written as a product of multinomial likelihoods at each time point since the transition probability matrix $P(t)$ is a stochastic matrix at each time point. An individual's contribution to the complete data likelihood is given as follows:

$$l_c \propto \sum_{t=0}^{\mathcal{T}-1} \sum_{j=0}^{S} \sum_{i=0}^{S} p_{ij}(t,t+1)^{N_{ij}(t,t+1)}$$

where $p_{ij}(t,t+1)$ is as given above. In the context of the time dependent transition rates, Maximum Likelihood estimation does not yield simple closed form solution due to the logistic link. Estimation of p_{ij} and β_{ij} involves maximization of a likelihood that is non-

linear in the parameters. Although the stEM procedure and the standard error computation proceeds in the same way as in the time-homogeneous case, the M-step of the procedure will be considerably slower due to increased number of parameters and the complicated form of the likelihood. Standard multinomial software in R can be useful for the maximization of the likelihood as long as β_{ij} varies with i (i.e. different covariate effects are assumed for the different transition types). Otherwise, if the different multinomial likelihoods (formed conditional on the starting states i) share common parameters, direct maximization of the likelihood must be performed and standard multinomial software may not be used. In order to improve the efficiency of the optimization of the likelihood, analytical gradient and Hessian functions may be provided to optimization routines in R. The above model formulation is flexible and allows for hypothesis testing. For example, we can test whether there is differential effect of time on the transition probabilities for different types of transition. This corresponds to testing the hypothesis $\beta_{ij} = \beta, \forall i, j$ using a standard likelihood ratio test given the stEM estimates under the null and alternative hypothesis. Another hypothesis may be that there is a quadratic effect of age on the transition probabilities as sexual partnership formation rates increase until early adulthood and decrease as most people age. Transition rates of the Markov chain may also depend on additional covariates such as gender and HIV status. Further discussion of modeling the effect of covariates on transition probability is discussed in the application section.

2.5 Simulation Study

A simulation study was carried out to evaluate the performance of stEM under sampling conditions that were similar to those in our applications dataset. In the sexual behavior survey from South Africa, participants were asked bi-annually to provide detailed information regarding their prior partnerships including age at sexual debut, and start and end time of their sexual partnerships. The sampling window–the total period over which the interview is conducted–is denoted by w. The partnership history process is generated according to the Markov model described in section 2.2. 100 subjects ($N = 100$) start the sexual partnership process in state 0 at the time origin ($time = 0$). The initial state

distribution probabilities are (1,0,0) for states 0,1,2 respectively; all begin the process in a state of celibacy (state 0). The transition probabilities corresponding to each state are as follows: $p_{01} = .2, p_{10} = .5, p_{12} = .4$ and $p_{21} = .6$. Given these initial state and transition probabilities, a subject's complete sexual history is generated following a Markov model. Each subject is observed for a period of $\mathcal{T}_i$ units of time where $\mathcal{T}_i \sim Unif(1, 30)$ and therefore the corresponding complete sample path is $\{X_i(t), t = 0, 1, \cdots, \mathcal{T}_i\}$. $\mathcal{T}_i$ is the end of the sampling window and corresponds to the time of the final interview in our partnership dataset. Finally, the observed data can be obtained from the complete data by imposing the following sampling scheme that was present in our application dataset:

1. If the process $X(t)$ entered state 0 before $\mathcal{T} - w$ and exited state 0 at $T_{exit} < \mathcal{T} - w$, all history before T_{exit} is set to missing. This happens because all relationships that were not long enough to make it into the sampling window are not reported except in the case where $X(\mathcal{T} - w) = 0$.

2. If the process is in state 0 at $\mathcal{T} - w$ ($X(\mathcal{T} - w) = 0$),then the $1 \to 0$ transition time and the start time of this relationship is known. This corresponds to the fact that individuals report the duration and date of termination of their last partnership if they report no relationship at the date of the baseline interview ($\mathcal{T} - w$).

3. After eliminating the unobservable history, we randomly pick T^*, earliest reported relationship formation time, uniformly between T_{exit} and $\mathcal{T} - w$. Note that T^* has to be a $0 \to 1$ or $1 \to 2$ transition time since it is reported as a time of partnership initiation. (See section 2.2.2 for further explanation.)

4. If the process is in state 1 at $\mathcal{T} - w$ ($X(\mathcal{T} - w) = 1$), then all data between T^* and $\mathcal{T} - w$ is also set to missing as the transition of the process relationships between these two time points is unobservable. This condition arises since a person who only reports 1 long relationship starting at T^* could have had multiple other short relationships that started and ended between T^* and $\mathcal{T} - w$.

5. If the process is in state 2 at $\mathcal{T} - w$ ($X(\mathcal{T} - w) = 2$), then the time point between T^* and T_2^* is set to missing where T_2^* is the time of the last $1 \to 2$ transition before

the sampling window. In this case, sojourn time in the most recent visit to state 2 is known.

In addition, note that the process cannot enter state 0 after T^* since there is an observed reported relationship indicating $X(t) \in \{1, 2\}$ for $t \geq T^*$. This modifies the transition probabilities that are used for imputation after T^*. As an example, the probability of transitioning from state 1 to 2 (conditional on not allowing for transitions to state 0) can be calculated as follows for $t \geq T^*$:

$$
\begin{aligned}
&P(X(t) = 2 | X(t-1) = 1, X(t) \neq 0) \\
&= P(X(t) = 2 | X(t-1) = 1, X(t) \in \{1, 2\}) \\
&= \frac{P(X(t) = 2, X(t-1) = 1, X(t) \in \{1, 2\})}{P(X(t-1) = 1, X(t) \in \{1, 2\})} \\
&= \frac{P(X(t) = 2, X(t-1) = 1)}{P(X(t-1) = 1, X(t) \in \{1, 2\})} \\
&= \frac{P(X(t) = 2, X(t-1) = 1)}{P(X(t-1) = 1, X(t) = 1) + P(X(t-1) = 1, X(t) = 2)} \\
&= \frac{p_{12}(t-1, t)}{p_{12}(t-1, t) + p_{11}(t-1, t)} \\
&= \frac{p_{12}(t-1, t)}{1 - p_{10}(t-1, t)}
\end{aligned}
$$

where the second from last equality is obtained by dividing with $P(X(t-1) = 1)$.

2.5.1 Stationary transition probabilities

Once the observed data are generated following the above steps, the stEM algorithm is implemented to estimate transition probabilities. The performance of the stEM algorithm is compared to the WW estimator for sampling windows of $w = 3$ and $w = 10$ (see Tables 2.1 and 2.2). In each run of the stEM algorithm, there were 1100 iterations of the S and M-step each with the first 100 iterations discarded for burn-in period. Empirical bias and empirical standard deviation of the stEM and WW estimators were computed across simulations. In addition the average of the standard deviation of each stEM estimate obtained from the Louis' formula is provided for comparison with the asymptotic variance of WW. Lastly, the empirical mean square error (MSE) is presented. As expected, the stEM

and WW estimators have similar point estimates but the former have lower variance. The estimates of variance based on Louis' formula are similar to the empirical estimate of the variance. Overall, both the stEM and WW estimates have lower MSE with increasing w because more information is available for estimation as the sampling window gets longer and more partnership transitions are captured.

Table 2.1 Comparison of within window (WW) and stochastic EM (stEM)
Time homogeneous model with w=3 and N=100

Parameter	p_{01}		p_{10}		p_{12}		p_{21}	
Truth	0.2		0.5		0.4		0.6	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	-0.0005	0.0033	-0.0107	0.0125	-0.0032	-0.0082	0.0117	-0.0090
Louis SE[1]	0.0137	0.0297	0.0385	0.0619	0.0367	0.0603	0.0496	0.0753
Empirical SE[2]	0.0150	0.0289	0.0402	0.0476	0.0371	0.0574	0.0537	0.0696
MSE	0.0002	0.0008	0.0017	0.0024	0.0014	0.0033	0.0029	0.0048

Louis SE[1] is the average of the Louis standard error estimates across all runs. It is computed for the stEM estimator only. The corresponding entry for the WW estimator is the asymptotic SE.

Empirical SE[2] is the sample standard deviation of the stEM estimates across all the runs.

Table 2.2 Comparison of within window (WW) and stochastic EM (stEM)
Time homogeneous model with w=10 and N=100

Parameter	p_{01}		p_{10}		p_{12}		p_{21}	
Truth	0.2		0.5		0.4		0.6	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	0.0004	0.0019	-0.0003	-0.0034	-0.0004	0.0040	0.0063	0.0041
Louis SE[1]	0.0127	0.0173	0.0304	0.0357	0.0293	0.0351	0.0381	0.0442
Empirical SE[2]	0.0113	0.0125	0.0312	0.0375	0.0281	0.0363	0.0412	0.0491
MSE	0.0001	0.0002	0.0009	0.0014	0.0008	0.0013	0.0017	0.0024

Louis SE[1] is the average of the Louis standard error estimates across all runs. It is computed for the stEM estimator only. The corresponding entry for the WW estimator is the asymptotic SE.

Empirical SE[2] is the sample standard deviation of the stEM estimates across all the runs.

2.5.2 Time dependent transition probabilities

For time dependent transition rates, the Markov chain is generated following the multinomial logistic model provided in section 2.4. The intercepts α_{ij} in the model are logit transformations of the baseline probabilities of transitions and are the same as in the

time-homogeneous case ($p_{01} = .2, p_{10} = .5, p_{12} = .4$ and $p_{21} = .6$) . The number of subjects (realizations of the process) is increased to 1000 in order to accommodate for the increase in the number of estimated parameters in this case. The time-effect parameters are $\beta_{01} = -0.1, \beta_{10} = -0.05, \beta_{12} = 0.05,$ and, $\beta_{21} = 0.2$. When the stEM algorithm is executed there were 1100 iterations of the S and M-step each with the first 100 iterations discarded for burn-in period. Tables 2.3 and 2.4 presents a comparison of the stEM and WW estimators for sampling windows $w = 3$ and $w = 10$, respectively. As in the time-homogeneous case, the stEM estimates are more efficient and have lower MSE compared to the WW estimates. Both the stEM and WW estimators have lower MSE with the longer sampling window which contains more transition data for estimation purpose.

Table 2.3 Comparison of within window (WW) and stochastic EM (stEM)
Time inhomogeneous model with W=3, N=1000

Time effect

Parameter	**β_{01}**		**β_{10}**		**β_{12}**		**β_{21}**	
Truth	-0.01		-0.05		0.05		0.02	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	-0.00170	0.00273	-0.00398	-0.00453	0.00100	-0.00029	0.00913	0.01414
Louis SE[1]	0.00587	0.01168	0.01553	0.02411	0.01374	0.02141	0.03839	0.04898
Empirical SE[2]	0.00746	0.01311	0.01393	0.02585	0.01396	0.02212	0.03477	0.05543
MSE	0.00006	0.00017	0.00021	0.00067	0.00019	0.00048	0.00126	0.00320

Intercept

Parameter	**α_{01}**		**α_{10}**		**α_{12}**		**α_{21}**	
Truth	-1.386294		1.609438		1.386294		.405465	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	-0.00567	-0.02216	-0.05809	0.09660	-0.03872	0.10424	0.00337	-0.02038
Louis SE[1]	0.04650	0.12979	0.17936	0.34246	0.17102	0.32714	0.26320	0.43277
Empirical SE[2]	0.05124	0.11605	0.15484	0.40326	0.18690	0.38366	0.24630	0.45275
MSE	0.00259	0.01361	0.02675	0.16788	0.03556	0.15438	0.06042	0.20454

Louis SE[1] is the average of the Louis standard error estimates across all runs. It is computed for the stEM estimator only. The corresponding entry for the WW estimator is the asymptotic SE.
Empirical SE[2] is the sample standard deviation of the stEM estimates across all the runs.

Table 2.4 Comparison of within window (WW) and stochastic EM (stEM)
Time inhomogeneous model with W=10, N=1000

Time effect

Parameter	β_{01}		β_{10}		β_{12}		β_{21}	
Truth	-0.01		-0.05		0.05		0.02	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	-0.00001	0.00097	0.00123	0.00094	-0.00041	-0.00366	-0.00112	-0.00277
Louis SE[1]	0.00565	0.00709	0.01268	0.01483	0.01143	0.01341	0.02248	0.02522
Empirical SE[2]	0.00454	0.00585	0.01281	0.01722	0.01176	0.01542	0.01836	0.01914
MSE	0.00002	0.00003	0.00016	0.00029	0.00014	0.00025	0.00033	0.00037

Intercept

Parameter	α_{01}		α_{10}		α_{12}		α_{21}	
Truth	-1.386294		1.609438		1.386294		.405465	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	-0.00480	0.00295	-0.04853	0.02166	-0.02851	0.02959	0.01314	-0.01694
Louis SE[1]	0.04572	0.07181	0.14596	0.19260	0.13841	0.18475	0.18836	0.23447
Empirical SE[2]	0.04226	0.06482	0.14558	0.22720	0.14490	0.23288	0.18660	0.22906
MSE	0.00176	0.00411	0.02302	0.05080	0.02128	0.05375	0.03485	0.05253

Louis SE[1] is the average of the Louis standard error estimates across all runs. It is computed for the stEM estimator only. The corresponding entry for the WW estimator is the asymptotic SE.
Empirical SE[2] is the sample standard deviation of the stEM estimates across all the runs.

2.6 Analysis of relationship transitions in KZN dataset

The application below considers the setting where the total number of relationships within the past year is known but the start and end time of a subset of the relationships is not. Data arose from sexual behavior survey at an HIV/AIDS clinic cohort in KZN, South Africa conducted bi-annually for 3 years. Data collected included duration of up to 3 relationships that were ongoing in the last 6 months, date of last sex, and total number of partnerships on-going within the last year from baseline. Figure 2.5 below illustrates an example of such data where an individual reported having a total of 2 relationships but only 1 of the relationships made it into the sampling window.

For our application, the stEM algorithm of section 3.3 can be adapted to estimate transition probabilities. Although the $Stochastic$ and $Maximization$ steps remain the same in this application, the Markov chain is initiated at time $b-12$; therefore, we need to calculate the distribution of states at $b-12$ conditional on the observed data and the distribution of states at $b-6$. This can be accomplished by employing the following Markov chain

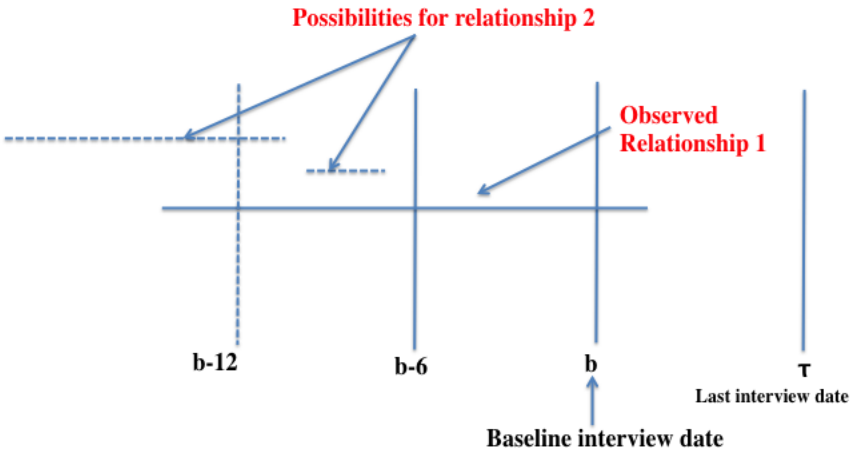


Figure 2.5: Illustration of partnership data where 2 relationships are reported ongoing within the last year but only 1 of them is exactly known.

identity:

$$(\boldsymbol{\pi_{0(t)}, \pi_{1(t)}, \pi_{2(t)}}) \prod_{i=t}^{t+n} \boldsymbol{P(i)} = (\boldsymbol{\pi_{0(t+n)}, \pi_{1(t+n)}, \pi_{2(t+n)}})$$

where $\boldsymbol{\pi_{i(t+n)}} = P(X(t+n) = i)$ is the probability of being in state i at time $t+n$ and $\boldsymbol{P(t)}$ is the transition probability matrix at time t. For the time-homogeneous case, the above identity simplifies to

$$(\boldsymbol{\pi_{0(t)}, \pi_{1(t)}, \pi_{2(t)}})\boldsymbol{P^n} = (\boldsymbol{\pi_{0(t+n)}, \pi_{1(t+n)}, \pi_{2(t+n)}}) \tag{2.1}$$

In order to obtain the state distribution at $b-12$, the above identity can be rearranged as:

$$(\pi_{0(b-12)}, \pi_{1(b-12)}, \pi_{2(b-12)}) = (\pi_{0(b-6)}, \pi_{1(b-6)}, \pi_{2(b-6)}) * \left(\prod_{t=b-12}^{(b-12+6)} \boldsymbol{P}(t) \right)^{-1} \tag{2.2}$$

Given an initial estimate of the state distribution at $b-12$, $(\pi_{0(b-12)}, \pi_{1(b-12)}, \pi_{2(b-12)})$, and an initial transition probability matrix, $\mathbf{P} = \mathbf{P^{(0)}}$, we can implement the stEM algorithm. We illustrate the steps of the implementation using Figure 2.5 as an

example. First, we determine the distribution of states at $b-12$; possible states at $b-12$ are 1 and 2 as there is already one ongoing relationship reported at that time. Therefore, we can calculate the conditional probability

$$P(X(b-12)=1|X(b-12)\in\{1,2\})=\frac{\pi_{1(b-12)}}{\pi_{1(b-12)}+\pi_{2(b-12)}}$$

where $\pi_{1(b-12)}$ and $\pi_{2(b-12)}$ are the respective probabilities of being in state 1 and state 2 at time $b-12$. Similarly,

$$P(X(b-12)=2|X(b-12)\in\{1,2\})=1-P(X(b-12)=1|X(b-12)\in\{1,2\})$$

as there are only two possible states the process can be in at time $b-12$. Second, we impute all missing data (between $b-12$ and $b-6$) by drawing samples from the missing data distribution. Note 2.1 can be utilized to obtain the distribution of states at the next time point given the current state and transition probability matrix. In the imputation process, the following two possibilities have to be considered: either relationship 2 ended between $b-12$ and $b-6$ (if $X(b-12)=2$) or relationship 2 began and ended between $b-12$ and $b-6$ (if $X(b-12)=1$). In this example, the imputation accepts only those draws that satisfy 1 of these 2 possibilities. At the end of this step, a pseudo-complete dataset is obtained. Third, we maximize the likelihood of the pseudo-complete data set to get a new estimate $\mathbf{P^{(1)}}$. These three steps are iterated until convergence. In summary, we can follow the steps below to implement stEM algorithm for estimation of transition probabilities:

1. Obtain initial estimates of $(\pi_{0(b-12)},\pi_{1(b-12)},\pi_{2(b-12)})$ based on the identity provided in 2.2. Note that the initial transition probability $P^{(0)}$ can be estimated from data available within the window and the initial state distribution at $b-6$ can be computed empirically from the observed data since relationships ongoing within 6 months of baseline are captured within the sampling window.

2. *Stochastic Step*: Given the current estimate of $\boldsymbol{P^{(k)}}$, $(\pi^{(k)}_{0(b-12)},\pi^{(k)}_{1(b-12)},\pi^{(k)}_{2(b-12)})$ and the observed data including the total number of relationships within the past year, draw samples from the distribution of the missing

data in order to obtain a pseudo complete data set. (See section 3.3 for further details on the stochastic step)

3. *Maximization Step*: Update the new estimates $\boldsymbol{P}^{(\boldsymbol{k+1})}$ as the maximum of the likelihood for the pseudo complete data from step 2.

4. Iterate steps 1-3 until convergence

Tables 2.5 - 2.8 summarize the results of fitting the following models to our dataset:

1. *Model 1: Time-homogeneous Markov model*: results are summarized in Tables 2.5 and 2.6 for males and females respectively.

2. *Model 2: Time dependent transition-probability model within each gender group*

$$log\frac{p_{ij}(t,t+1)}{p_{ii}(t,t+1)} = \alpha_{ij} + \beta_{ij} * t$$

   Results are summarized in Tables 2.7 and 2.8 for males and females respectively.

The variance estimates in Tables 2.5 through 2.8 show that the stEM estimates have lower variance compared to the WW estimates since the stEM algorithm uses more information for estimation. The reduction in the estimation of variance seemed to depend on the type of model that was considered. In the case of the time-homogeneous transition probabilities (Tables 2.5 and 2.6), there was as much as 15% reduction in variance whereas in the case of the case of covariate dependent transition probabilities (Table 2.6) there was as much as 25% reduction in standard error. We note that inference regarding the transition probabilities is not significantly affected by assuming stationary transition probabilities (Table 2.5). However, as shown in Table 2.7 the time effect associated with $2 \rightarrow 1$ transition probability achieves statistical significance under the stEM method but not the WW method.

Tables 2.7 and 2.8 summarize the impact of time on the partnership transition probabilities for each gender. For men, the odds of transitioning out of any relationship state decreased with each additional month. For women, the odds of transitioning from celibacy to monogamy and from monogamy to celibacy decreased with each additional month.

Table 2.5 Stochastic EM (stEM) versus within window(WW) estimation of transition rates when the total number of relationships within one year of baseline is known

Time homogenous model for men

	P_{01}		P_{10}	
	stEM	WW	stEM	WW
Estimate	0.01540	0.01540	0.01782	0.01788
SE	0.00123	0.00132	0.00110	0.00120
CI	(0.01299 , 0.01782)	(0.01281 , 0.01799)	(0.01566 , 0.01998)	(0.01553 , 0.02023)

	P_{12}		P_{21}	
	stEM	WW	stEM	WW
Estimate	0.00177	0.00174	0.03892	0.03180
SE	0.00035	0.00036	0.00407	0.00428
CI	(0.00108 , 0.00247)	(0.00103 , 0.00245)	(0.03094 , 0.04689)	(0.02342 , 0.04019)

Table 2.6 Stochastic EM (stEM) versus within window(WW) estimation of transition rates when the total number of relationships within one year of baseline is known

Time homogeneous model for women

	p_{01}		p_{10}	
	stEM	WW	stEM	WW
Estimate	0.01567	0.01557	0.01843	0.01822
SE	0.00125	0.00133	0.00113	0.00122
CI	(0.01321 , 0.01813)	(0.01296 , 0.01817)	(0.01621 , 0.02065)	(0.01584 , 0.0206)

	p_{12}		p_{21}	
	stEM	WW	stEM	WW
Estimate	0.00177	0.00177	0.02714	0.02676
SE	0.00035	0.00039	0.00332	0.00385
CI	(0.00108 , 0.00246)	(0.001 , 0.00253)	(0.02063 , 0.03364)	(0.01921 , 0.0343)

However, the odds of transitioning from monogamy to concurrency and from concurrency to monogamy increased with each additional month. We note that for both men and women, the impact of time on the probabilities of transitioning out of the state of monogamy (either to celibacy or concurrency) was not significant.

Table 2.7 Stochastic EM (stEM) versus within window(WW) estimation of transition rates when the total number of relationships within one year of baseline is known

Time inhomogeneous model for men

Intercept

	α_{01}		α_{10}	
	stEM	WW	stEM	WW
Estimate	-3.1702	-2.8478	-3.8017	-3.6554
SE	.2523	.3368	.2102	.2730
CI	(-3.66478 , -2.67563)	(-3.50798 , -2.18758)	(-4.21375 , -3.38965)	(-4.19043 , -3.12029)

	α_{12}		α_{21}	
	stEM	WW	stEM	WW
Estimate	-5.0244	-4.8318	-2.3661	-3.7278
SE	.5620	.8268	.2370	.5358
CI	(-6.12594 , -3.92276)	(-6.45232 , -3.21121)	(-2.83064 , -1.90156)	(-4.7779 , -2.6776)

Time-effect

	β_{01}		β_{10}	
	stEM	WW	stEM	WW
Estimate	-.0380	-.0497	-.0062	-.0109
SE	.0115	.0145	.0086	.0105
CI	(-0.06052 , -0.01543)	(-0.07816 , -0.0212)	(-0.02304 , 0.01065)	(-0.03145 , 0.00965)

	β_{12}		β_{21}	
	stEM	WW	stEM	WW
Estimate	-.0659	-.0743	-.0392	.0101
SE	.0344	.0449	.0150	.0203
CI	(-0.13345 , 0.00156)	(-0.16229 , 0.01365)	(-0.06853 , -0.00987)	(-0.02957 , 0.04986)

The above analysis was performed separately for each gender because of differences in partnership reporting patterns; the data in Table 2.9 shows that women report far fewer partnership breakups compared to men. Additionally, when examining the goodness of fit of the time-inhomogeneous model assumption for each gender (see Figure 2.6), the model fit for women seems poor compared to men. The goodness of fit was checked using a statistic that naturally arises from the stEM algorithm. In short, the idea behind this procedure is to compare the observed and expected number of trials until a match

Table 2.8 Stochastic EM (stEM) versus within window(WW) estimation of transition rates when the total number of relationships within one year of baseline is known

Time inhomogeneous model for women

Intercept

	α_{01}		α_{10}	
	stEM	WW	stEM	WW
Estimate	-3.7202	-3.3557	-3.8455	-3.7386
SE	.1739	.2202	.1335	.1767
CI	(-4.06111 , -3.37932)	(-3.78734 , -2.924)	(-4.10704 , -3.58392)	(-4.08487 , -3.39231)

	α_{12}		α_{21}	
	stEM	WW	stEM	WW
Estimate	-6.5453	-6.1995	-3.7966	-3.6280
SE	.4574	.5413	.2865	.3649
CI	(-7.44184 , -5.64869)	(-7.2605 , -5.1385)	(-4.35809 , -3.23503)	(-4.34316 , -2.91284)

Time-effect

	β_{01}		β_{10}	
	stEM	WW	stEM	WW
Estimate	-.0213	-.0338	-.0085	-.0117
SE	.0072	.0089	.0055	.0068
CI	(-0.03543 , -0.00714)	(-0.05124 , -0.01639)	(-0.01932 , 0.00237)	(-0.02507 , 0.00158)

	β_{12}		β_{21}	
	stEM	WW	stEM	WW
Estimate	.0111	.0003	.0191	.0136
SE	.0167	.0196	.0110	.0134
CI	(-0.02174 , 0.04389)	(-0.03815 , 0.03875)	(-0.00257 , 0.04069)	(-0.01254 , 0.03984)

(i.e. a simulated sample path matches all observed data). This procedure is similar in spirit to the Quantile-Quantile plot and further details are given in the model validation section of Chapter 3.

Table 2.9 Number of relationship dissolutions during the 6 month before the sampling

	Gender	
Number of dissolutions	**F**	**M**
0	448	155
1	4	8
2	0	9
3	0	1

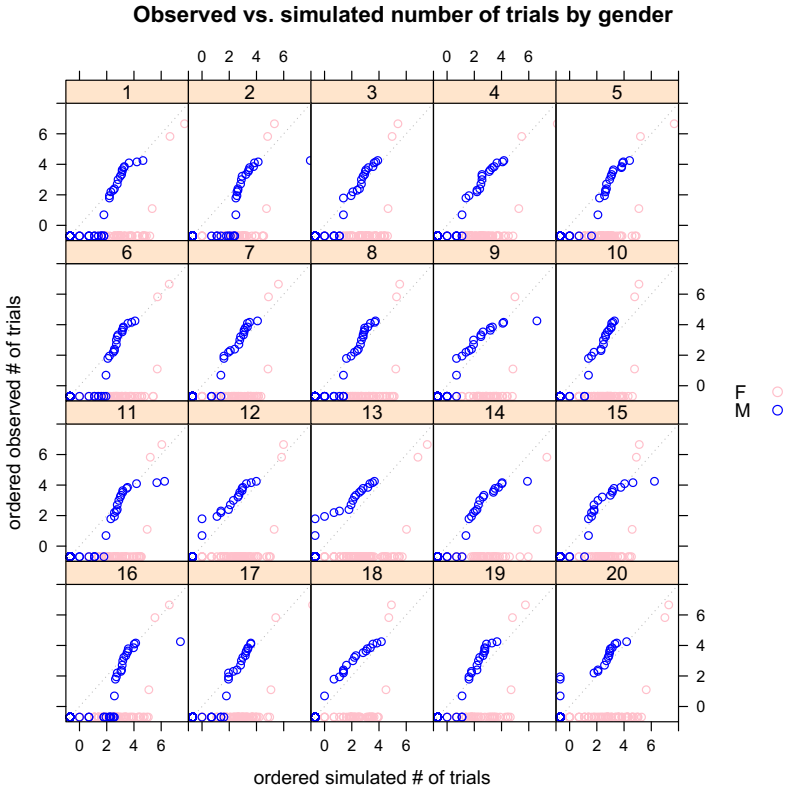


Figure 2.6: Observed vs expected number of trials from fitting time-inhomogeneous Markov model to the application data separately for each gender. Each of the 20 panels represent a plot of the observed versus expected number trials (obtained from a single run of step 2 of the validation procedure)

2.7 Discussion

This paper describes a Markov framework to model and estimate transition probabilities from duration data collected under a complex sampling scheme. Estimation of partnership transition parameters is achieved by using a stochastic expectation maximization algorithm coupled with a rejection sampling scheme. The algorithm provided here is sufficiently flexible enough to be to accommodate a variety of sampling schemes that arise in collection of retrospective data. In our setting, the stEM algorithm permits utilization of information outside the sampling window; as shown in the simulation study as well as in the application to KZN data, the stEM estimate had considerably lower SE and consequently far lower MSE compared to the WW estimator.

The application discussed in this paper provides a practical illustrative example of the value of utilizing information outside the sampling window when estimating and modeling transition probabilities. Specifically, the result of Table 2.7 indicates that one could arrive at different conclusions from hypothesis testing of parameters since the time effect associated with $2 \rightarrow 1$ transition probability was statistically significant at the 0.05 level under the stEM method but not the WW method. Overall the results of our analysis suggest that with increasing age both genders had lower odds of transitioning into and out of the state of monogamy, and men had increased of odds of ending concurrent relationships.

The analysis of our application data was restricted to the time frame beginning one year prior to baseline and ending with the last date of interview. Therefore, this analysis did not utilize information on total life-time number of partnerships as well as partnership history data earlier than 1 year from the baseline interview date. The reason that we did not make use of this information was the lack of fit of our models to these data; the number of reported partners was not compatible with the data on partnership formation and dissolution. A more complex model would be required to utilize sexual history data across the lifetime of each individual; such a model must take into account the likely error in reporting lifetime number of partners. When fitting models using data across the lifetime, it is important to check the validity of the assumption that concurrent individuals

cannot form new relationships without first ending 1 of the two ongoing relationships (See section 3.3.1 for details). For our application dataset, only about 1% of the subjects reported 3 or more overlapping partnerships ongoing within the sampling window and therefore this assumption is plausible. If this assumption is not tenable, one of the following two approaches is suggested in order to utilize information on the total lifetime number of partnerships. First, increase the number of states in the Markov model by further compartmentalizing the concurrency state to allow transitions into states with more than 2 partnerships at a time. This approach equates lifetime number of partnerships to $n_{01} + n_{12} + n_{23} + n_{34} + \cdots$ where n_{ij} is the total number of transitions from being in a state with i ongoing partnerships to being in a state with j ongoing partnerships. Second, keep the simplified 3-state Markov model, but adjust the reported lifetime number of partnerships by subtracting expected number of 3 or more simultaneous relationships. Besides checking the aforementioned assumptions, two caveats should be kept in mind when analyzing longer periods (or entire lifetime) of retrospectively collected sexual history data . First, such data is subject to recall bias, which is likely to be worse for longer periods of recall (James et al., 1991; Ellish et al., 1996; Nelson et al., 2010). In addition, counts of lifetime sexual partnerships may suffer from measurement errors arising from such phenomenon as heaping (Fenton et al., 2001; Weinhardt et al., 1998). Sensitivity analysis may be useful to examine the impact of such bias on the estimation of transition probabilities.

There are various ways in which our work presented can be extended. First, the methods considered in this paper were applied to a discretely measured partnership history data; analogous methods for the case of continuous time partnership history data will be required. Second, other possible approaches for modeling time dependent transition rates can be considered, such as those based on piecewise-constant Markov models, as suggested in (Kay, 1986). Such an approach divides the time axis $[0, \mathcal{T}]$ into several intervals across which transition rates are allowed to vary. Likelihood ratio-tests may be used in order to compare if the piecewise-constant model is a better fit for the data compared to the time-homogeneous model.

Finally, the Markov assumption in this paper can be relaxed by, for example, employing

a semi-markov model, which would permit transition probabilities to depend on the sojourn times in the current state. This has the advantage that the duration of the current relationship can be inversely correlated with the probability of dissolution (Felmlee et al., 1990). The suitability of the Markov assumption made in this paper and the goodness of fit of our model is further explored in Chapter 3 where a more general combined Markov and logistic regression model is employed to characterize the sexual history process.

Markov and Logistic Regression Framework for Modeling Relationship Patterns

Yared Gurmu and Victor De Gruttola

3.1 Introduction

Modeling and estimation of patterns of sexual partnership formation and dissolution is important in investigation of the epidemic dynamics of sexually transmitted infections (STI). Various features of the partnership process such as overlap of sexual partnerships, gap between partnerships, duration of partnerships as well as patterns of relationship formations and dissolutions need to be properly estimated for accurate modeling of STIs (Morris et al., 2007, 2010; Foxman et al., 2006). The difficulty in properly modeling and estimating these features lies in the fact that sexual history data is usually retrospectively collected and is often fraught with missing data; the total number of partnerships, the formation and dissolution time of relationships (consequently the duration of relationships) may be censored, truncated or subject to recall bias (Burington et al., 2010).

Sexual partnership data are commonly obtained through retrospective surveys that collect information on relationships that are ongoing during a fixed time window. Estimation of partnership transition probabilities from partnership data collected under such sampling scheme is discussed in Chapter 2. The Markov modeling framework from Chapter 2 is able to capture essential features of the relationship history through three discrete states of celibacy, monogamy and concurrency. The stochastic expectation maximization algorithm discussed there is useful for estimating transition probabilities into and out of the relationship states mentioned above. Although the Markov framework is useful for characterizing transition through discrete states of the sexual history process, this framework did not incorporate information on the type of concurrency pattern that may have occurred. In this paper, we extend the framework presented in Chapter 2 by incorporating logistic regression model to classify the type of concurrency that occurred prior to the time when the Markov process exited concurrency state. As different concurrency patterns may impact network structures, incorporation of such information allows investigation of the impact of concurrency patterns on STI transmission dynamics (Goodreau et al., 2012; Gorbach et al., 2002; Kretzschmar et al., 2010). Proper estimation of concurrency types as well as the transition probabilities from incomplete partnership data aids in understanding STI spread and in developing targeted intervention to prevent

transmission.

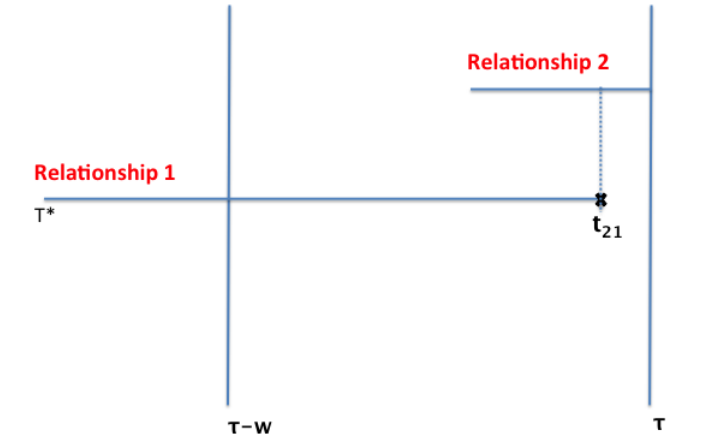
As in Chapter 2, the focus of this chapter will be on performing efficient estimation of model parameters by utilizing additional sexual history data that is available outside of the sampling window. Such data may include total number of relationships within a specified period of time, start and end time of a subset of relationships partnerships. Methods discussed in this paper are a generalization of the framework discussed in Chapter 2 and further details on the sampling scheme as well as partnership data can be found there. This paper is organized as follows. The next section describes the Markov and logistic modeling framework. Section 3 discusses the stochastic EM algorithm and its implementation, and section 4 presents simulation results for the performance of stEM. Section 5 discusses model validation strategies based on statistic that arise from the stEM procedure. Section 6 presents applications to data obtained from a sexual behavior survey, and section 7 provides a discussion.

3.2 Methods

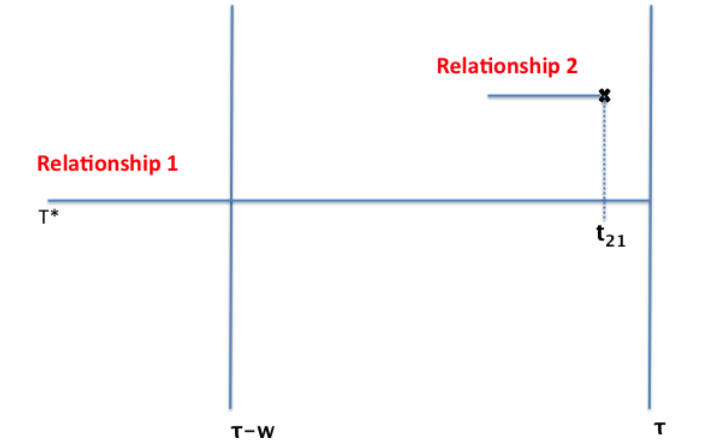
3.2.1 Markov model and logistic regression framework for partnership duration data

The Markov model discussed in this chapter is described in detail in Chapter 2 section 2.2. We extend this Markov model to include the type of concurrency pattern that occurred in a person's sexual relationship history. Such an extension is necessary because the Markov process for relationship transition does not identify which relationship ended when the process makes $2 \to 1$ transition. Figure 3.1 illustrates the connection between the duration data collected in partnership surveys and the Markov process representation. Although the concurrency pattern of the partnership durations represented in Figures 3.1 (a) and (b) differ, they are both represented by the same sample path ($1 \to 2 \to 1$) as shown in (c). The patterns in 3.1 (a) and (b) are referred to as transitional and embedded concurrency, respectively. As shown in 3.1 (a) transitional concurrency occurs when a second partnership begins before an earlier (older) partnership is terminated and the second partnership continues after the first one is terminated. Embedded concurrency (Figure 3.1 (a)) occurs

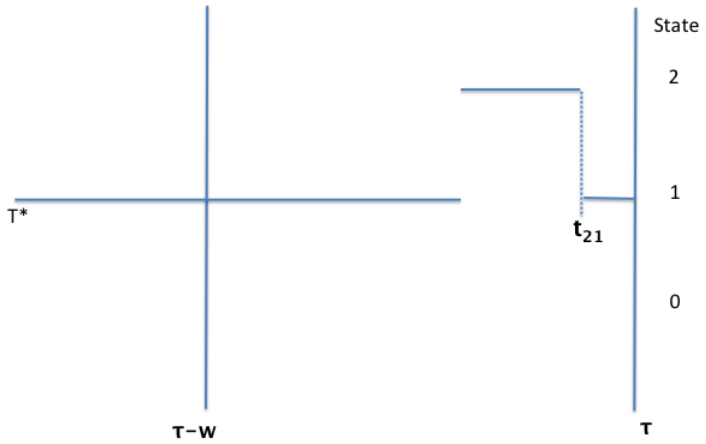
when the older partnership persists and the new side-partnership ends first.



(a)



(b)



(c)

Figure 3.1: Two different duration data (a) and (b) have identical sample path ($1 \to 2 \to 1$) as shown in (c)

As the Markov model is not sufficient to fully characterize sexual relationship history, we define the following indicators of concurrency that will permit unique mapping of the Markov data to the corresponding duration data. Let Y_{ik} be defined as follows:

$$Y_{ik} = \begin{cases} 1 & : k^{th} \text{ concurrency for individual } i \text{ is transitional} \\ 0 & : k^{th} \text{ concurrency for individual } i \text{ is embedded} \end{cases}$$

Going back to Figure 3.1 (c), an observed sample path of $1 \rightarrow 2 \rightarrow 1$ and Y=1, corresponds to transitional concurrency while an observed sample path of $1 \rightarrow 2 \rightarrow 1$ and Y=0, corresponds to embedded concurrency. Therefore, in order to fully characterize the duration data, we need both the Markov process data and the concurrency indicator. The implications of the Markov assumption on the partnership durations of an individual is discussed in the appendix. As described there, depending on the partnership formation pattern of an individual, the Markov assumption may imply independent partnership durations or positively correlated partnership durations.

The combined Markov and logistic framework described here provides quantitative information on the essential features of the relationship process (i.e. duration, gap, overlap of sexual partnerships, as well as the timing and patterns of relationship formations and dissolutions) which are important in the modeling of the spread of STIs (Morris et al., 2007, 2010; Foxman et al., 2006). For example, a gap in the formation process can be described as the sojourn time in state 0 after a $1 \rightarrow 0$ transition, and duration can be formulated as the sum of the sojourn times in states 1 and 2 as described in the appendix section.

3.2.2 Estimation of transition probability and concurrency pattern parameters

Let $\mathbf{X_i}$=$\{X_i(t), t = 0, 1, \cdots, \mathcal{T}_i\}$ represent the complete sample path for individual i. Let $\mathbf{Y_i} = \{Y_{ik}, k = 1, 2, \cdots, n_{21i}\}$ represent the concurrency pattern indicator for person i that experiences n_{21i} transitions from state 2 to 1.Given both $\mathbf{X_i}$ and $\mathbf{Y_i}$, we can uniquely map from the Markov process to the relationship history. For the purpose of this paper, we assume Y_{ik} to have Bernoulli distribution with parameter $\pi(\mathbf{x_{ik}})$ that depends on vector

of covariates at the k^{th} concurrency (i.e. k^{th} visit to state 2). The conditional probability of observing transitional concurrency will be given by the following logistic regression:

$$logit\left(P(Y_{ik}|x_{i1k}, x_{i2k})\right) = \gamma_0 + \gamma_1 * x_{i1k} + \gamma_2 * x_{i2k} \tag{3.1}$$

where X_{i1k} is the sojourn time in state 1 before the k^{th} visit to state 2 and and X_{i2k} is the sojourn time in state 2 at the k^{th} visit. In the context of time-homogeneous Markov, it can be shown that X_{i1k} and X_{i2k} have Geometric distribution with parameters p_{11} and p_{22}, respectively. In the case of time-inhomogeneous Markov, the distribution of X_{i1k} depends on the transition probabilities at the time steps at which state 1 is entered. For example, if state 1 was just entered at time t then the sojourn time in in state 1 during this particular visit is distributed as follows:

$$P(X_{i1k} = x_{i1k}) = p_{11}(t, t+1) * p_{11}(t, t+2) * \cdots * p_{11}(t, t+x_{i1k}-2) * (1 - p_{11}(t, t+x_{i1k}-1))$$

Similar reasoning can be used to determine the distribution of X_{i2k} in the case of the time-inhomogeneous process.

3.2.3 Maximum Likelihood Estimation

Given the complete Markov data $\mathbf{X_i}$ and the indicators for transitional concurrency $\mathbf{Y_i}$, individual $i's$ contribution to the likelihood for the duration data is

$$\begin{aligned}\mathcal{L}(\mathbf{p}, \gamma; \mathbf{X}, \mathbf{Y}) &= \mathcal{L}_1(\mathbf{p}; \mathbf{X})\mathcal{L}_2(\gamma; \mathbf{Y}, \mathbf{X}) \\ &\propto \left(\prod_{j=0}^{S}\prod_{i=0}^{S} p_{ij}^{N_{ij}(\mathcal{T}_k)}\right)\prod_{k=1}^{n_{21i}}\left(\pi(\mathbf{x_{ik}})^{y_{ik}} * [1 - \pi(\mathbf{x_{ik}})]^{1-y_{ik}}\right)\end{aligned}$$

where $\mathbf{p}$ is the transition parameter matrix, $\gamma = (\gamma_{\mathbf{0}}, \gamma_{\mathbf{1}}, \gamma_{\mathbf{2}})$ is vector of parameters of the logistic model and $N_{ij}(\mathcal{T}_i) = \sum_{t=0}^{\mathcal{T}_i - 1} I(X_t = i, X_{t+1} = j) = \#\{2 \text{ to } 1 \text{ transitions until } \mathcal{T}_i\}$ and the process transitions through states $\{0, 1, \cdots, S\}$ (See Chapter 2 for additional details). Note that the complete data likelihood is a product of a Markov chain likelihood($\mathcal{L}_1$) and a logistic regression likelihood ($\mathcal{L}_2$). Because the two likelihood do not share parameters, we can independently estimate the model parameters when the sexual histories are fully observed. For the Markov portion of the likelihood ($\mathcal{L}_1$) we have a sim-

ple closed form estimate (see Chapter 2) in the complete data scenario. For the logistic regression piece($\mathcal{L}_2$), parameters can be estimated using Newton-Raphson or iteratively re-weighted least squares approaches.

In practice, the partnership process is only partially observed outside of the sampling window.Therefore, the observed data likelihood $\mathcal{L}(\mathbf{p}, \gamma|\mathbf{X}_{obs}, \mathbf{Y}_{obs})$ cannot be expressed in-terms of simple functions of the one-step transition parameters, p_{ij} and is no longer separable. This suggests the use of the Expectation-Maximization (EM) algorithm because conditioning on the parameter estimates and observed data the complete data likelihood is separable.

3.2.4 Stochastic EM for estimation of transition probability and concurrency pattern parameters

Maximum likelihood estimation makes use of a stochastic version of the expectation-maximization algorithm (EM) of Dempster et al. (Dempster et al., 1977) and Celeux et al. (Celeux and Diebolt, 1985; Gilks et al., 1996). Stochastic versions of the EM have been applied in various settings including those where the E-step of the EM algorithm is intractable. The stochastic EM algorithm replaces missing values with samples drawn from the conditional distribution of the missing data given the current parameter estimate and the observed data. This replacement produces a pseudo-complete dataset that can then be maximized to obtain a new parameter estimate for the next iteration of the procedure. In our context, given the pseudo-complete Markov data set, $\mathbf{X}$, the logistic regression covariate data X_{i1k} and X_{i2k} as well as n_{21i} will be treated as known. Then, pseudo-complete concurrency indicator data, $\mathbf{Y}$, can be obtained by imputing $Y_{ik} \sim Bernoulli(\pi(x_{ik}))$ for missing Y_{ik} values. Given the pseudo-complete $(\mathbf{X}, \mathbf{Y})$ direct maximization is straightforward and yields a new parameter estimate for the next iteration of the algorithm.

Details of the stochastic EM algorithm for estimating transition parameters, $\mathbf{p}$, and concurrency pattern parameters $\gamma = (\gamma_{\mathbf{0}}, \gamma_{\mathbf{1}}, \gamma_{\mathbf{2}})$ is given as follows:

1. Based on within window data, calculate initial estimates of γ^0 and $\mathbf{p^0}$.

2. Stochastic-step I: Obtain pseudo-complete Markov data set, $\mathbf{X}$

At the j^{th} iteration, draw samples from the distribution of the missing data given the observed data and the current estimate of the transition parameter $\mathbf{p}^{(j)}$. Rejection sampling is utilized to match on relationship start times, the state of the process at $\mathcal{T} - w$ and total number of relationships ongoing within the past year from baseline. This results in a pseudo-complete dataset for the Markov process.

3. Maximization-step I:

   Determine the new parameter estimate $\mathbf{p}^{(j+1)}$ as the maximizer of the pseudo-complete Markov data.

4. Stochastic-step II: Obtain pseudo-complete logistic data set, $\mathbf{Y}$

   Given the pseudo-complete Markov chain dataset, we can now impute the concurrency indicator conditional on the covariates corresponding to the sojourn times in states 1 and 2. The concurrency indicator at the k^{th} $2 \rightarrow 1$ transition will be sampled from $Y_{ik} \sim Bernoulli(\pi(x_{ik}))$ where

$$\pi(x_{ik}) = P(Y_{ik} = 1|x_{ik}) = \frac{e^{\gamma_0^{j-1} + \gamma_1^{j-1} * x_{i1k} + \gamma_2^{j-1} * x_{i2k}}}{1 + e^{\gamma_0^{j-1} + \gamma_1^{j-1} * x_{i1k} + \gamma_2^{j-1} * x_{i2k}}}$$

5. Maximization-step II:

   The above step (4) will result in a pseudo-complete concurrency indicator data. Based on this data we can obtain new estimates of $(\gamma_0^j, \gamma_1^j, \gamma_2^j)$ by fitting the logistic regression model in equation 3.1.

6. Iterate steps 2 through 5 until convergence

3.2.5 Variance of the stEM estimate

Variance estimation makes use of the method of Louis discussed in (Louis, 1982) :

$$\frac{\partial^2 l_{obs}(\theta; x_{obs}, y_{obs})}{\partial\theta\, \partial\theta^T} = \mathrm{E}_\theta \left(\frac{\partial^2 l_c(\theta; x, y)}{\partial\theta\, \partial\theta^T} \Big| x_{obs}, y_{obs} \right) - \mathrm{cov}_\theta \left(\frac{\partial l_c(\theta; x, y)}{\partial\theta} \Big| x_{obs}, y_{obs} \right)$$

where l_{obs} and l_c are the log likelihood of the observed and complete data, respectively, and θ =(p, γ) is a vector of the parameters of interest. As shown above, this approach accounts for the loss of information due to missing data by subtracting from the information

from the complete data.

The bootstrap approach (Efron and Tibshirani, 1994) can be used to calculate the theoretical covariance and mean functions of the information and score of the complete data likelihood. The following steps are implemented in the calculation of the variance of stEM estimators $\tilde{p}$ and $\tilde{\gamma}$:

1. Generate bootstrap samples

   (a) For each individual's observed partnership history process $\mathbf{X_{i,obs}}, \mathbf{Y_{i,obs}}$, simulate a realization of the Markov process conditional on the observed data using the stEM estimate $\tilde{p}$). This results in a pseudo-complete Markov data $\mathbf{X_i} = \{X_i(t), t = 0, 1, \cdots, \mathcal{T}_i\}$. Given the pseudo-complete Markov chain dataset, we can now obtain pseudo-complete concurrency indicator data $\mathbf{Y_i}$ following stochastic step II discussed in prior section.

   (b) Repeat 1a for all individuals $i = 1, \cdots, K$ in our partnership dataset to obtain pseudo-complete Markov chain and logistic dataset for all individuals in our study.

   (c) Generate M pseudo-complete datasets by repeating the above steps.

   Given the bootstrap samples, the next two steps will be utilized to calculate the right hand side of the Louis' identity.

2. Calculate the sample mean, the estimate of the first term on the right hand side of the Louis' identity, as:

$$\frac{1}{M}\sum_{m=1}^{M}\left(\frac{\partial^2 l_c^m(\theta; x, y)}{\partial\theta\,\partial\theta^T}\right)$$

   where $l_c^m(\theta; x, y)$ is the complete data log likelihood for the m^{th} simulated process. As discussed in the previous section, the complete data log-likelihood l_c^m is a sum of Markov and logistic log likelihoods that do not share parameters; hence, the Hessian matrix $\frac{\partial^2 l_c^m(\theta;x,y)}{\partial\theta\,\partial\theta^T}$ will be block diagonal as follows:

$$H = \left[\begin{array}{c|c} A & 0 \\ \hline 0 & D \end{array}\right]$$

with A corresponding to the Markov parameters and D corresponding logistic parameters. The elements of the Hessian matrix correspond to A can be computed by taking the second derivative for the Markov likelihood (see Chapter 2 for details). The elements of the Hessian matrix H corresponding to D are easily calculated by taking second derivative of a logistic regression log-likelihood.

3. Calculate the sample covariance, the estimate of the second term on the right hand side of the identity, as:

$$\frac{1}{M}\sum_{m=1}^{M}\left[\frac{\partial l_c^m(\theta;x,y)}{\partial\theta}-\bar{l}\right]\left[\frac{\partial l_c^m(\theta;x,y)}{\partial\theta}-\bar{l}\right]^T$$

where

$$\bar{l}=\frac{1}{M}\sum_{m=1}^{M}\frac{\partial l_c^m(\theta;x,y)}{\partial\theta}.$$

Again, the score equations necessary for calculating the sample covariance can be separated into the score equation for the Markov likelihood and the score equation for the logistic regression likelihood.

Note that all the steps in the parametric bootstrap are carried out setting $\theta=\tilde{\theta}$, the stEM estimate. Finally, the variance covariance of the stEM estimates can be calculated as usual by inverting the estimate of the observed information obtained above.

3.3 Simulation

A simulation study was carried out to evaluate the performance of stEM under sampling conditions that were similar to those in our applications dataset. The sampling window– the total period over which the survey is administered is denoted by w and is allowed to be $w=15$ or $w=30$. The number of subjects (realization of the process) is $N=500$ and $N=2000$ for the homogenous and time-inhomogeneous models, respectively. The transition parameters for the time-homogeneous model are $p_{01}=.2, p_{10}=.5, p_{12}=.4$ and $p_{21}=.6$. For time dependent transition rates, the Markov chain is generated according to fol-

lowing multinomial logistic model

$$log\frac{p_{ij}(t,t+1)}{p_{ii}(t,t+1)} = \alpha_{ij} + \beta_{ij} * t$$

where α_{ij} are logit transformations of the baseline probabilities of transitions and are the same as in the time-homogeneous case. The time-effect parameters are $\beta_{01} = .07, \beta_{10} = -0.05, \beta_{12} = -0.03,$ and, $\beta_{21} = -0.01$. The initial state distribution probabilities are (1,0,0) for states 0,1,2 respectively. Given the initial state and transition probabilities, a subject's complete sexual history ($\{X_i(t), t = 0, 1, \cdots, \mathcal{T}\}$) is generated up until the end of the sampling window $\mathcal{T} \sim U(1, 100)$. Given the complete Markov process data, the concurrency classifier data is generated following the logistic regression model in 3.1 with coefficients close to what we estimated in our application dataset ($\gamma_0 = 4, \gamma_1 = -.2, \gamma_2 = -.2$). Finally, the observed data can be obtained from the complete data by imposing the sampling scheme that was present in our application dataset:

1. If the process $X(t)$ entered state 0 before $\mathcal{T} - w$ and exited state 0 at $T_{exit} < \mathcal{T} - w$, all history before T_{exit} is set to missing: only relationships that lasted long enough to make it into the sampling window are reported except in the case where $X(\mathcal{T} - w) = 0$.

2. If the process is in state 0 at $\mathcal{T} - w$ ($X(\mathcal{T} - w) = 0$), then the $1 \to 0$ transition time and the start time of this relationship will be known. This corresponds to the fact that individuals report the duration and date of termination of their last partnership if they report no relationship at the date of the baseline interview ($\mathcal{T} - w$).

3. After eliminating the unobservable history, we randomly pick T^*, earliest reported relationship formation time, uniformly between T_{exit} and $\mathcal{T} - w$. Note that T^* has to be a $0 \to 1$ or $1 \to 2$ transition time since it is reported as a time of partnership initiation.

4. If the process is in state 1 at $\mathcal{T} - w$ ($X(\mathcal{T} - w) = 1$), then all data between T^* and $\mathcal{T} - w$ are also set to missing as the transition of the process relationships betweens these two time points is unobservable due to the sampling scheme. This condition

arises since a person who only reports 1 long relationship starting at T^* could have had multiple other short relationships that started and ended between T^* and $\mathcal{T}-w$.

5. If the process is in state 2 at $\mathcal{T}-w$ ($X(\mathcal{T}-w)=2$), then the time point between T^* and T_2^* is set to missing where T_2^* is the time of the last $1 \rightarrow 2$ transition before the sampling window. In this case, sojourn time in the most recent visit to state 2 will be known.

As the focus of the application dataset begins with the period that starts one year prior to baseline interview ($b-12$) and ends with the last date of interview $\mathcal{T}$ (see Figure 2.5), the matching procedure for the Stochastic-step matches on the total number of partnerships since $b-12$, and the observed state distribution at $b-6$. Additionally, if $T^* \geq b-12$, the imputation procedure has to ensure T^* is $0 \rightarrow 1$ or $1 \rightarrow 2$ transition time. If $T^* \leq b-12$, then the state distribution at $b-12$ should be calculated conditional on $X(b-12) \in 1,2$ (see Chapter 2 for additional details regarding imputations after T^*).

Once the observed data are generated following the above steps, the stEM algorithm is implemented to estimate transition probabilities and concurrency model parameters. The stEM estimate is compared to the within window (WW) estimate which relies solely on data collected during the sampling window. In each run of the stEM algorithm, there were 1100 iterations of the S and M-step each with the first 100 iterations discarded for burn-in period. Empirical bias, standard deviation (SD), the mean Louis' standard deviation and mean square error (MSE) are presented in the tables below. Our results in Tables 3.1-3.4 show that the stEM estimates have lower variance compared to the WW estimates. Both estimates perform well with longer sampling period or increased sample size.

Table 3.1 Comparison of within window (WW) and stochastic EM (stEM)
Time homogeneous model with W=15, N=500

Transition probabilities

Parameter	p_{01}		p_{10}		p_{12}		p_{21}	
Truth	0.1		0.2		0.5		0.6	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	.000395	-.000666	.003458	.002317	.000816	-.002319	.001242	-.006275
Louis SE[1]	.003206	.005329	.006711	.010207	.008875	.012730	.010040	.013729
Empirical SE[2]	.004376	.005397	.007395	.008864	.011149	.012941	.013081	.014493
MSE	.000019	.000030	.000067	.000084	.000125	.000173	.000173	.000249

Concurrency parameters

Parameter	γ_0		γ_1		γ_2	
Truth	4		-0.2		-0.2	
	stEM	WW	stEM	WW	stEM	WW
Bias	.077217	.116000	.020412	.031070	.020845	.029459
Louis SE[1]	.314658	.417468	.162078	.212920	.099684	.140625
Empirical SE[2]	.317529	.465223	.143229	.224810	.104564	.154532
MSE	.106787	.229888	.020931	.051505	.011368	.024748

Louis SE[1] is the average of the Louis standard error estimates across all runs. It is computed for the stEM estimator only. The corresponding entry for the WW estimator is the asymptotic SE.
Empirical SE[2] is the sample standard deviation of the stEM estimates across all the runs.

Table 3.2 Comparison of within window (WW) and stochastic EM (stEM)
Time homogeneous model with W=30, N=500

Transition Parameters

Parameter	p_{01}		p_{10}		p_{12}		p_{21}	
Truth	0.1		0.2		0.5		0.6	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	-.000458	.000674	.000053	-.004962	-.002763	-.001389	-.001781	.002532
Louis SE[1]	.002513	.003763	.005160	.007195	.006612	.009001	.007313	.009680
Empirical SE[2]	.003363	.003705	.006673	.007191	.009083	.009832	.010570	.010913
MSE	.000012	.000014	.000045	.000076	.000090	.000099	.000115	.000126

Concurrency parameters

Parameter	γ_0		γ_1		γ_2	
Truth	4		-0.2		-0.2	
	stEM	WW	stEM	WW	stEM	WW
Bias	-.055086	-.063342	.006748	.007137	.014000	.023220
Louis SE[1]	.227218	.310975	.100630	.140236	.074937	.099362
Empirical SE[2]	.261240	.276958	.088824	.113526	.073453	.094182
MSE	.071281	.080718	.007935	.012939	.005591	.009409

Louis SE[1] is the average of the Louis standard error estimates across all runs. It is computed for the stEM estimator only. The corresponding entry for the WW estimator is the asymptotic SE.
Empirical SE[2] is the sample standard deviation of the stEM estimates across all the runs.

Table 3.3 Comparison of within window (WW) and stochastic EM (stEM)
Time inhomogeneous model with W=15, N=2000

Time effect

Parameter	β_{01}		β_{10}		β_{12}		β_{21}	
Truth	0.07		-0.05		-0.03		-0.01	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	-.000243	-.000202	.000805	.000863	-.000109	-.000036	.000095	.000077
Louis SE[1]	.001591	.003638	.003288	.004629	.001382	.001892	.001586	.002123
Empirical SE[2]	.002691	.003846	.003103	.003961	.001398	.001641	.001749	.001784
MSE	.000007	.000014	.000010	.000016	.0000020	.0000026	.00000307	.00000311

Intercept

Parameter	α_{01}		α_{10}		α_{12}		α_{21}	
Truth	-2.197225		-.405465		.510826		.405465	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	-.042185	-.007944	-.017397	-.042971	.001433	-.023824	.022984	.018349
Louis SE[1]	.025104	.088745	.057439	.130317	.037581	.081005	.042791	.089890
Empirical SE[2]	.054425	.106838	.072839	.122876	.038918	.093846	.050579	.082538
MSE	.004668	.011192	.005608	.016442	.001517	.009081	.003086	.006922

Concurrency Indicators

Parameter	γ_0		γ_1		γ_2	
Truth	4		-0.2		-0.2	
	stEM	WW	stEM	WW	stEM	WW
Bias	-.020959	-.026557	.001779	.002650	.005569	.007673
Louis SE[1]	.158239	.238381	.048451	.061729	.038320	.077731
Empirical SE[2]	.171499	.256818	.052984	.081927	.044506	.070070
MSE	.029851	.066661	.002810	.006719	.002012	.004969

Louis SE[1] is the average of the Louis standard error estimates across all runs. It is computed for the stEM estimator only. The corresponding entry for the WW estimator is the asymptotic SE.
Empirical SE[2] is the sample standard deviation of the stEM estimates across all the runs.

Table 3.4 Comparison of within window (WW) and stochastic EM (stEM)
Time inhomogeneous model with W=30, N=2000

Time effect

Parameter	β_{01}		β_{10}		β_{12}		β_{21}	
Truth	0.07		-0.05		-0.03		-0.01	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	.000530	-.001088	.000852	-.000348	.000035	-.000141	-.000741	.000353
Louis SE[1]	.001217	.002640	.001536	.002360	.001276	.001558	.001087	.001547
Empirical SE[2]	.002148	.003335	.002336	.002795	.001352	.001899	.001327	.001566
MSE	.000005	.000012	.000006	.000008	.0000018	.0000024	.0000023	.0000025

Intercept

Parameter	α_{01}		α_{10}		α_{12}		α_{21}	
Truth	-2.197225		-0.405465		0.510826		0.405465	
	stEM	WW	stEM	WW	stEM	WW	stEM	WW
Bias	-.024639	.026318	-.005085	-.019780	.001935	.008595	.002269	-.006297
Louis SE[1]	.021805	.064177	.040226	.091968	.026664	.057409	.029875	.063440
Empirical SE[2]	.043303	.077264	.062335	.097462	.044650	.052251	.039470	.066971
MSE	.002420	.006463	.003912	.009574	.001997	.002713	.001563	.004375

Concurrency Indicators

Parameter	γ_0		γ_1		γ_2	
Truth	4		-0.2		-0.2	
	stEM	WW	stEM	WW	stEM	WW
Bias	.005877	.012928	.001684	.002002	-.002478	-.003078
Louis SE[1]	.116167	.171108	.026066	.041517	.028604	.055979
Empirical SE[2]	.119998	.171255	.034237	.050129	.030146	.044995
MSE	.014434	.029496	.001175	.002517	.000915	.002034

Louis SE[1] is the average of the Louis standard error estimates across all runs. It is computed for the stEM estimator only. The corresponding entry for the WW estimator is the asymptotic SE.

Empirical SE[2] is the sample standard deviation of the stEM estimates across all the runs.

3.4 Model validation

As our estimation procedure involves rejection sampling, we recommend using the number of trials until a simulated sample path is accepted as a test statistic to check model fit. In the section that follows we provide details of the validation approach. The basic idea behind this procedure is to compare the observed and expected number of trials until a match; below we describe a method for calculating the expected number. If the hypothesized model matches the data generating mechanism, we expect a slope close to 1 when the observed number of trials is plotted against the expected number of

trials. The procedure below outlines a simulation based approach for performing model validation.

1. **Data Generation I — Compute observed number of trials**

   (a) Simulate complete data under the *true* model

   (b) Apply sampling scheme to obtain observed data

   (c) Apply stEM to the observed data to obtain an estimate of the transition matrix $\hat{\mathbf{p}}_1$
   (When stEM is applied, imputation is performed under the *hypothesized* model)

   (d) Store the number of trials until a match from the last iteration of the stochastic EM. Let this quantity be Y^{match}. (This quantity will be plotted on Y-axis).

2. **Data Generation II — Compute expected number of trials**

   (a) Generate complete data under *hypothesized* model using $\hat{\mathbf{p}}_1$ and observed τ_i from data generation I

   (b) Apply sampling scheme to obtain observed data

   (c) Apply stEM to observed data to obtain pstEM2 $\hat{\mathbf{p}}_2$

   (d) Store the number of trials until acceptance.

   (e) Repeat above steps M times and average the number of trails until acceptance across the M simulations to obtain the expected number of trials until a match. Let this quantity be X^{match}.(This quantity will be plotted on X-axis).

3. Plot the fitted line from regression of Y^{match} against X^{match}.

Alternatively, it is possible to plot the number of trials from each of the simulations in step 2e against the Y^{Match}, to obtain M scatter plots in order to assess goodness of fit. In addition, examples discussed in this section applied log transformation to both the observed and expected number of trials as these statistics are often skewed.

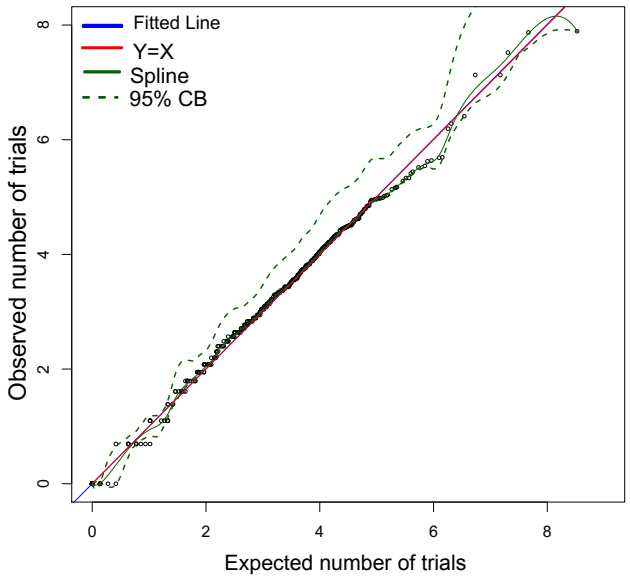
A penalized spline model is fit to the ordered observed and expected statistics. The 95% confidence band for the spline curve (dashed lines) are calculated by bootstrapping the observed and expected statistics 1000 times. When implementing the validation steps in practice, we apply the validation procedure starting with step 1(c) as we do not know the true data generating mechanism.

Figures 3.2 and 3.3 provide two different scenarios under which the model fitting assumptions are violated. In Figure 3.2, the true data generating mechanism is a time-inhomogeneous Markov process following the parameters given in the simulation section above. In Figure 3.3, the true data generating mechanism is a second order Markov process where the transition probability depends on both the current and prior state of the process. In both examples when stEM estimation is performed under the wrong model, the distribution of the observed vs expected statistics deviate sharply from the line $y = x$. However, when estimation is performed under the correct model, the scatter plot follows the line $y = x$ closely. Further examples of the implementation of the validation steps can be found in the data analysis section.

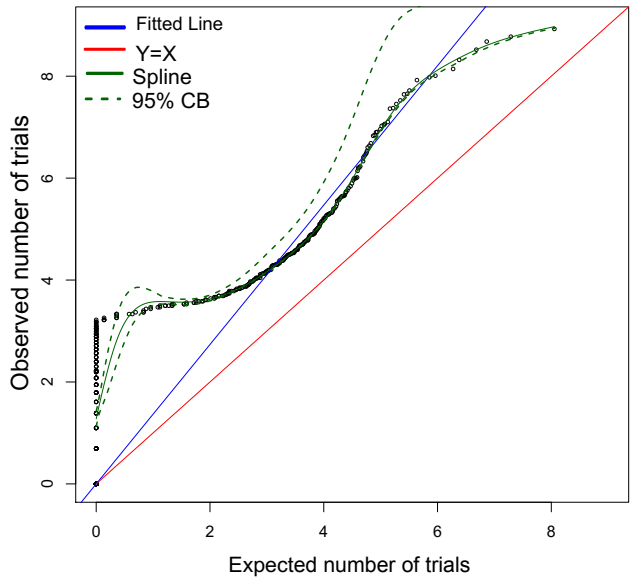
3.5 Analysis of relationship patterns in KZN dataset

We apply the proposed methods to analyze sexual history data collected from an HIV treatment and care study in KZN, South Africa. The data was collected through a survey that was conducted bi-annually for 3 years and included information on duration of up to 3 relationships that were ongoing in the last 6 months, date of last sex, and total number of partnerships on-going within the last year from baseline. The application considers the setting where the total number of relationships during the year prior to the first interview date is known but the start and (possibly censored) end time of a subset of the relationships is not. Further details about the application dataset and relevant assumptions can be found in Chapter 2.

The observed duration data was modeled using a 3-state time-inhomogeneous Markov model to characterize relationship transitions and a logistic regression model to describe the patterns of concurrency. Table 3.5 presents a comparison of the stEM and WW param-

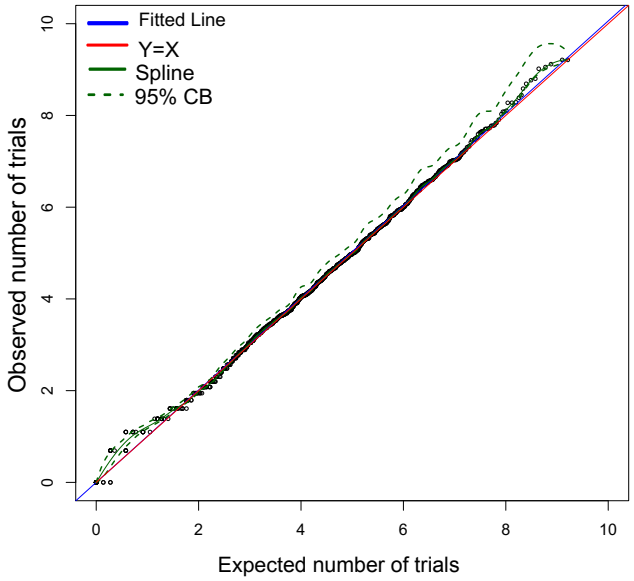


(a)

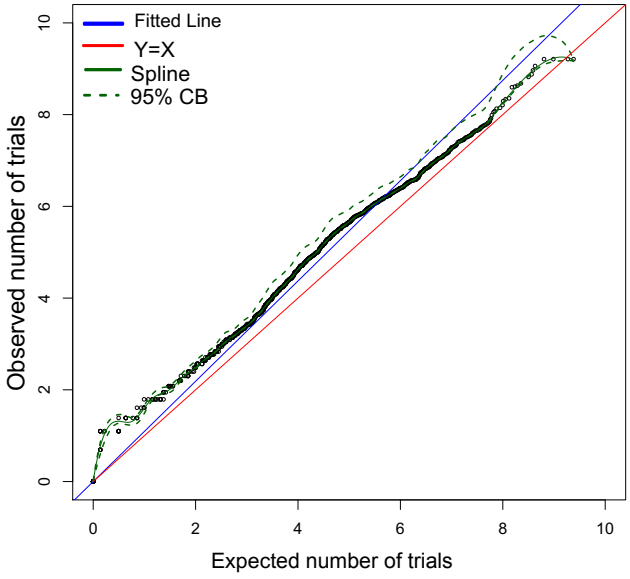


(b)

Figure 3.2: Observed vs expected number of trials when there is no model violation (Figure 3.2(a), Time-inhomogeneous model) and when there is model violation (Figure 3.2(b), Time Homogeneous model). The expected number trials until a match are calculated by averaging the number of trials until a match across $M = 30$ simulations.



(a)



(b)

Figure 3.3: Observed vs expected number of trials when there is no model violation (Figure 3.3(a), Time-inhomogeneous model) and when there is model violation (Figure 3.3(b), Time Homogeneous model). The expected number trials until a match are calculated by averaging the number of trials until a match across $M = 30$ simulations.

eter estimates for partnership transitions and concurrency patterns. Because of concerns regarding apparent under-reporting among women as revealed in our validation plot, the analysis presented in Table 3.5 restricted the data for women to what is available within the sampling window. For both the relationship transition and concurrency model parameters, the stEM estimates were noted to have as much as 30% lower standard error compared to the WW estimates. When examining the impact of age on relationship transitions, we found that with increasing age subjects had lower odds of transitioning out of their current relationship state; these results are consistent even whether the analysis is performed adjusting for gender effects or not. Of note, the odds of transitioning from concurrency to monogamy decreased by roughly 2% (OR=.979, 95% CI: [.961, .997]) for each additional month. The results from the concurrency pattern analysis show that the odds of a transitional concurrency (ending the older relationship) decreased with increasing length of time spent in a state of concurrency as well as monogamy just before entering the concurrent state. The odds of ending the partnership that started first decreased by about 20% (OR=.8, 95% CI: [.638, 1.015]) for each additional month a person stays in both the old and new partnerships (i.e. concurrent state). The odds of ending the older partnership decreased by about 15%(OR=.85, 95 % CI:[.718,1.001]) for each additional month a person remained in the first relationship prior to the start of the second concurrent relationship.

Table 3.5 Stochastic EM (stEM) versus within window (WW) estimates of transition rates and concurrency patterns when the total number of relationships within one year of baseline is known*

Time Inhomogeneous Markov Model

Intercept

	α_{01}		α_{10}	
	stEM	WW	stEM	WW
Estimate	-3.7218	-3.3252	-3.9572	-3.8695
SE	0.1124	0.1962	0.0864	0.1572
CI	(-3.94212 , -3.5015)	(-3.70972 , -2.94063)	(-4.12661 , -3.78776)	(-4.17757 , -3.56148)
	α_{12}		α_{21}	
	stEM	WW	stEM	WW
Estimate	-5.4924	-5.2721	-3.1191	-3.0754
SE	0.2147	0.4114	0.1522	0.2972
CI	(-5.9131 , -5.07167)	(-6.07854 , -4.46569)	(-3.41748 , -2.82071)	(-3.65789 , -2.49283)

Time Effect

	β_{01}		β_{10}	
	stEM	WW	stEM	WW
Estimate	-0.0241	-0.0373	-0.0017	-0.0040
SE	0.0063	0.0077	0.0046	0.0056
CI	(-0.03635 , -0.01182)	(-0.05243 , -0.02223)	(-0.01066 , 0.00732)	(-0.01493 , 0.00697)
	β_{12}		β_{21}	
	stEM	WW	stEM	WW
Estimate	-0.0321	-0.0359	-0.0190	-0.0141
SE	0.0097	0.0172	0.0065	0.0118
CI	(-0.05106 , -0.01304)	(-0.06963 , -0.00224)	(-0.03181 , -0.00614)	(-0.0373 , 0.00902)

Gender Effect (Male=1, Female=0)

	$male_{01}$		$male_{10}$	
	stEM	WW	stEM	WW
Estimate	0.1557	0.1095	-0.0447	-0.0838
SE	0.1317	0.1990	0.0982	0.1494
CI	(-0.10243 , 0.4138)	(-0.28057 , 0.49967)	(-0.23715 , 0.14774)	(-0.37663 , 0.20908)
	$male_{12}$		$male_{21}$	
	stEM	WW	stEM	WW
Estimate	0.2843	-0.0382	0.6016	0.1879
SE	0.2548	0.4189	0.1603	0.2764
CI	(-0.21503 , 0.78362)	(-0.85922 , 0.78276)	(0.2874 , 0.91572)	(-0.35372 , 0.72959)

Concurrency Indicators

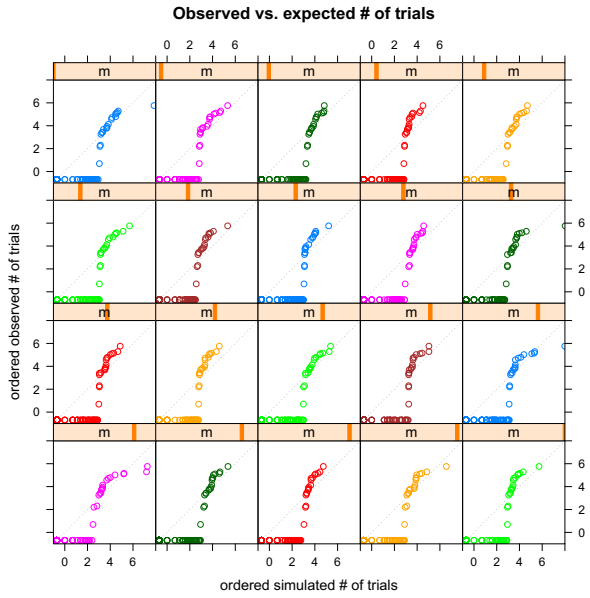
	γ_0		γ_1		γ_2	
	stEM	WW	stEM	WW	stEM	WW
Estimate	4.77	4.00	-0.1659	-0.1699	-0.2181	-0.2206
SE	1.98	2.02	0.0873	0.1176	0.1201	0.1334
CI	(0.899, 8.64)	(0.044 , 7.95)	(-0.337, 0.005)	(-0.400 , 0.061)	(-0.453 , 0.0173)	(-0.482, 0.041)

*Data for women is restricted to what was observed within the sampling window

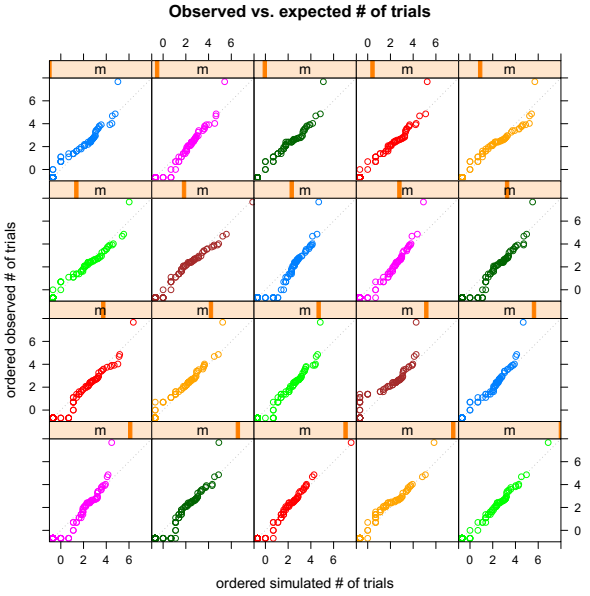
Model fit was assessed by comparing the observed number of trials until a match with the expected number of trials until a match following the procedure outlined in section 3.4. First, we considered the goodness of fit a model that combined data across both genders. For this model, when stEM estimation is performed assuming the application data arose from the time-inhomogeneous model (3.4, (a)), the model validation scatter plots show a violation of this assumption; in order to assess the violation of the assumption, this plot can be compared to the adjacent plot (b) that is generated from a time-inhomogeneous Markov model with the same parameters as in our application dataset. Second, we considered the goodness of fit of separate models for each gender. This allows us to assess the contribution of each gender to the overall model violation. As shown in (2.6) a gender difference in the model fit is noted; the model fit for women seems poor compared to men. As the validation plots show the observed number of trials until a match is far fewer than expected for women who report no relationship dissolution (i.e. have complete data) in the 6 months prior to the sampling window. This underreporting by women can also be seen by examining Table 2.9 where women reported only 4 dissolutions compared to 29 by men within the same period. Given this reporting pattern, the analysis (presented in Table 3.5) restricted the data reported by women to what was observed during the sampling window. Lastly, the goodness of fit of the logistic model was evaluated using the Hosmer-Lemeshow test (Hosmer and Lemeshow, 1980) for the data observed within the window. This test did not reject the null hypothesis that the fitted model is adequate ($\chi^2_8 = 12.9$, p-value$= .115$).

3.6 Discussion

This paper describes a Markov and logistic regression framework to characterize the partnership transition and concurrency pattern parameters from incomplete and retrospectively sampled duration data. Estimation of these parameters is achieved by using a stochastic expectation maximization algorithm coupled with a rejection sampling scheme. The algorithm provided here is sufficiently flexible enough to accommodate a variety of sampling schemes that arise in collection of retrospective data. In our setting, the stEM



(a)



(b)

Figure 3.4: Observed vs expected number of trials combining data across both genders for (a) the application dataset and (b) for data generated from time-inhomogeneous Markov Chain using the parameters estimated from the application dataset. Each of the 20 panels represent a plot of the observed versus expected number trials (obtained from a single run of step 2 of the validation procedure)

algorithm permits utilization of information outside the sampling window; as shown in the simulation study as well as in the application to KZN data, the stEM estimate had

lower SE and MSE compared to the WW estimator.

When the modeling framework was applied to duration data from a cohort of HIV patients, it was noted that with increasing age subjects had lower odds of forming new relationships or dissolving their current one; all of the transition parameter estimates were negative indicating subjects had decreased odds of transitioning out of their current relationship state. In addition, the concurrency pattern analysis suggested that subjects with concurrent partnerships had lower odds of dissolving the older partnership (transitional concurrency) for every additional month spent in a state of concurrency or in the preceding state of monogamy with the older partnership; in other words, the newer (side) partnership was at an increased odds of dissolving for every additional month spent in states of concurrency as well as monogamy. Prior research has suggested that concurrent partnerships with longer overlap have greater impact on the spread of STIs than those with shorter overlaps (Morris et al., 2010). In our context, longer periods of overlap are at increased odds of being classified as embedded concurrency adjusting for the period of time spent in monogamous state. Further risk factors may need to be examined to determine correlates of embedded concurrency and how it is associated with STI transmission at the individual or community level.

The validation method introduced in this paper reveals partnership reporting differences between men and women and underscores the importance of performing the validation procedure at the covariate-level in addition to hypothesis testing effect of the covariate on each transition type. When model validation was performed for each gender (Figure 2.6), our results showed a difference in the model fit by gender and that women may perhaps report far fewer relationship dissolutions than men. This finding is consistent with the literature as similar reporting differences have been found in prior studies (Helleringer et al., 2011; Nnko et al., 2004). Figure 2.6 also shows a low level of under-reporting of relationship breakups for men. One way to adjust this small amount of under-reporting that is observed among men could involve multiple imputation to impute what appears to be missing data (unreported breakups). This approach might assign to each person in each of the M imputed data sets (for example, $M = 20$ in Figure 2.6, the number of breakups that are expected in the simulated trials for those observations that have no breakups.

This approach allows us to use the validation method to suggest the possible number of missing breakups. If we were to iterate this procedure, it may also be turned into yet another level of stochastic EM and can be further investigated in the future. In addition to the validation procedure described in this paper, alternative validation procedures can also be developed based on the results of the implications of the Markov assumption on the covariance between duration of relationships which is outlined in the appendix. These results are valid for first order time-homogeneous processes and can be extended to the case of time-inhomogeneous processes.

Proper modeling and estimation of temporal features of the relationship process (i.e. duration, gap, overlap of sexual partnerships, as well as the timing and patterns of relationship formations and dissolutions) is important in the modeling of the spread of STIs (Morris et al., 2007, 2010; Foxman et al., 2006). The framework presented in this paper provides quantitative information on these features of the relationship process and enhance our understanding of STI transmission when properly incorporated in mathematical and network models of disease spread. The method presented in this paper can be extended to allow for correlation in concurrency patterns across the lifetime of an individual. If study participants experience recurring episodes of concurrency over a period of time, a generalized estimating equation (GEE) approach (Zeger and Liang, 1986) can be implemented to account for correlation.

Acknowledgement

The authors would like to thank Nuala McGrath and Africa Center Demographic Information System for making available the dataset used to illustrate methods developed in chapters 2 and 3.

References

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics* 1100–1120.

ANDERSON, R. M. and GARNETT, G. P. (2000). Mathematical models of the transmission and control of sexually transmitted diseases. *Sexually transmitted diseases* **27** 636–643.

ASGHARIAN, M., M'LAN, C. E. and WOLFSON, D. B. (2002). Length-biased sampling with right censoring: An unconditional approach. *Journal of the American Statistical Association* **97** 201–209.

ASGHARIAN, M., WOLFSON, D. B. ET AL. (2005). Asymptotic behavior of the unconditional npmle of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics* **33** 2109–2131.

BEE, M. (2005). Estimating rating transition probabilites with missing data. *Statistical Methods and Applications* **14** 127–141.

BILLINGSLEY, P. (1961). Statistical methods in markov chains. *The Annals of Mathematical Statistics* 12–40.

BURINGTON, B., HUGHES, J. P., WHITTINGTON, W. L., STONER, B., GARNETT, G., ARAL, S. O. and HOLMES, K. K. (2010). Estimating duration in partnership studies: issues, methods and examples. *Sexually transmitted infections* **86** 84–89.

CELEUX, G. and DIEBOLT, J. (1985). The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly* **2** 73–82.

CHEN, M. I., GHANI, A. C. and EDMUNDS, J. (2008). Mind the gap: the role of time between sex with two consecutive partners on the transmission dynamics of gonorrhea. *Sexually transmitted diseases* **35** 435–444.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38.

DIEBOLT, J. and CELEUX, G. (1993). Asymptotic properties of a stochastic em algorithm for estimating mixing proportions. *Stochastic Models* **9** 599–613.

DIEBOLT, J. and IP, E. H. (1996). Stochastic em: method and application. In *Markov chain Monte Carlo in practice*. Springer, 259–273.

DOHERTY, I. A., SHIBOSKI, S., ELLEN, J. M., ADIMORA, A. A. and PADIAN, N. S. (2006). Sexual bridging socially and over time: a simulation model exploring the relative effects of mixing and concurrency on viral sexually transmitted infection transmission. *Sexually transmitted diseases* **33** 368–373.

EFRON, B. and TIBSHIRANI, R. J. (1994). *An introduction to the bootstrap*. CRC press.

ELLISH, N. J., WEISMAN, C. S., CELENTANO, D. and ZENILMAN, J. M. (1996). Reliability of partner reports of sexual history in a heterosexual population at a sexually transmitted diseases clinic. *Sexually transmitted diseases* **23** 446–452.

FELMLEE, D., SPRECHER, S. and BASSIN, E. (1990). The dissolution of intimate relationships: A hazard model. *Social Psychology Quarterly* 13–30.

FENTON, K. A., JOHNSON, A. M., MCMANUS, S. and ERENS, B. (2001). Measuring sexual behaviour: methodological challenges in survey research. *Sexually transmitted infections* **77** 84–92.

FOXMAN, B., NEWMAN, M., PERCHA, B., HOLMES, K. K. and ARAL, S. O. (2006). Measures of sexual partnerships: lengths, gaps, overlaps, and sexually transmitted infection. *Sexually transmitted diseases* **33** 209–214.

FRYDMAN, H. (1994). A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *Journal of the Royal Statistical Society. Series B (Methodological)* 71–74.

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice* **1** 19.

GOODREAU, S. M., CASSELS, S., KASPRZYK, D., MONTAÑO, D. E., GREEK, A. and MORRIS, M. (2012). Concurrent partnerships, acute infection and hiv epidemic dynamics among young adults in zimbabwe. *AIDS and Behavior* **16** 312–322.

GORBACH, P. M., STONER, B. P., ARAL, S. O., WHITTINGTON, W. L. and HOLMES, K. K. (2002). It takes a village: understanding concurrent sexual partnerships in seattle, washington. *Sexually transmitted diseases* **29** 453–462.

GUTTORP, P. and MININ, V. N. (1995). *Stochastic modeling of scientific data*. CRC Press.

HELLERINGER, S., KOHLER, H.-P., KALILANI-PHIRI, L., MKANDAWIRE, J. and ARMBRUSTER, B. (2011). The reliability of sexual partnership histories: implications for the measurement of partnership concurrency during surveys. *AIDS (London, England)* **25** 503.

HOSMER, D. W. and LEMESHOW, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods* **9** 1043–1069.

JAMES, N. J., BIGNELL, C. J. and GILLIES, P. A. (1991). The reliability of self-reported sexual behaviour. *Aids* **5** 333–336.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53** 457–481.

KAY, R. (1986). A markov model for analysing cancer markers and disease states in survival studies. *Biometrics* 855–865.

KEIDING, N. and MOESCHBERGER, M. (1992). *Independent delayed entry*. Springer.

KOUMANS, E. H., FARLEY, T. A., GIBSON, J. J., LANGLEY, C., ROSS, M. W., MCFARLANE, M., BRAXTON, J. and ST LOUIS, M. E. (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** 497–503.

KRETZSCHMAR, M., WHITE, R. G. and CARAËL, M. (2010). Concurrency is more complex than it seems. *AIDS (London, England)* **24** 313.

LAI, T. L. and YING, Z. (1991). Estimating a distribution function with truncated and censored data. *The Annals of Statistics* 417–442.

LOUIS, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 226–233.

LURIE, M., HARRISON, A., WILKINSON, D. and KARIM, S. A. (1997). Circular migration and sexual networking in rural kwazulu/natal: implications for the spread of hiv and other sexually transmitted diseases. *Health Transition Review* 17–27.

LURIE, M. N., WILLIAMS, B. G., ZUMA, K., MKAYA-MWAMBURI, D., GARNETT, G. P., STURM, A. W., SWEAT, M. D., GITTELSOHN, J. and KARIM, S. S. A. (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.

MARTIN, E. C. and BETENSKY, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional kendall's tau. *Journal of the American Statistical Association* **100**.

MATSON, P. A., CHUNG, S.-E. and ELLEN, J. M. (2012). When they break up and get back together: length of adolescent romantic relationships and partner concurrency. *Sexually transmitted diseases* **39** 281.

MAY, R. M., ANDERSON, R. M. and IRWIN, M. (1988). The transmission dynamics of human immunodeficiency virus (hiv)[and discussion]. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **321** 565–607.

MORRIS, M., EPSTEIN, H. and WAWER, M. (2010). Timing is everything: international variations in historical sexual partnership concurrency and hiv prevalence. *PloS one* **5** e14092.

MORRIS, M., GOODREAU, S. and MOODY, J. (2007). Sexual networks, concurrency, and std/hiv. *Sexually Transmitted Diseases. New York: McGraw-Hill* 109–126.

MORRIS, M. and KRETZSCHMAR, M. (1997). Concurrent partnerships and the spread of hiv. *Aids* **11** 641–648.

NELSON, S. J., HUGHES, J. P., FOXMAN, B., ARAL, S. O., HOLMES, K. K., WHITE, P. J. and GOLDEN, M. R. (2010). Age-and gender-specific estimates of partnership formation and dissolution rates in the seattle sex survey. *Annals of epidemiology* **20** 308–317.

NNKO, S., BOERMA, J. T., URASSA, M., MWALUKO, G. and ZABA, B. (2004). Secretive females or swaggering males?: An assessment of the quality of sexual partnership reporting in rural tanzania. *Social science & medicine* **59** 299–310.

ORCHARD, T., WOODBURY, M. A. ET AL. (1972). A missing information principle: theory and applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*. The Regents of the University of California.

QIN, J. and SHEN, Y. (2010). Statistical methods for analyzing right-censored length-biased data under cox model. *Biometrics* **66** 382–392.

ROSENBERG, M. D., GURVEY, J. E., ADLER, N., DUNLOP, M. B. and ELLEN, J. M. (1999). Concurrent sex partners and risk for sexually transmitted diseases among adolescents. *Sexually transmitted diseases* **26** 208–212.

SEN, P. K. (1987). What do the arithmetic, geometric and harmonic means tell us in length-biased sampling? *Statistics & probability letters* **5** 95–98.

SHERLAW-JOHNSON, C., GALLIVAN, S. and BURRIDGE, J. (1995). Estimating a markov transition matrix from observational data. *Journal of the Operational Research Society* 405–410.

THERNEAU, T. M. and GRAMBSCH, P. M. (2000). *Modeling survival data: extending the Cox model*. Springer Science & Business Media.

TSAI, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika* **77** 169–177.

TSAI, W.-Y., JEWELL, N. P. and WANG, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **74** 883–886.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)* 290–295.

VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics* 616–620.

VARDI, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* **76** 751–761.

WANG, M.-C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association* **84** 742–748.

WANG, M.-C., BROOKMEYER, R. and JEWELL, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics* 1–11.

WANG, R., GOYAL, R., LEI, Q., ESSEX, M. and DE GRUTTOLA, V. (2014). Sample size considerations in the design of cluster randomized trials of combination hiv prevention. *Clinical Trials* 1740774514523351.

WATTS, C. H. and MAY, R. M. (1992). The influence of concurrent partnerships on the dynamics of hiv/aids. *Mathematical biosciences* **108** 89–104.

WEINHARDT, L. S., FORSYTH, A. D., CAREY, M. P., JAWORSKI, B. C. and DURANT, L. E. (1998). Reliability and validity of self-report measures of hiv-related sexual behavior: progress since 1990 and recommendations for research and practice. *Archives of sexual behavior* **27** 155–180.

WOODROOFE, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics* 163–177.

YEH, H.-W., CHAN, W., SYMANSKI, E. and DAVIS, B. R. (2010). Estimating transition probabilities for ignorable intermittent missing data in a discrete-time markov chain. *Communications in Statistics Simulation and Computation®* **39** 433–448.

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

Appendix A

Consistency of the TPLE estimator when there is RENO

We need to demonstrate

$$\sup_{x\in[\tau^*,x_1]} |\hat{S}_{tple}(x) - S(x|x > \tau^*)| \xrightarrow{p} 0 \tag{A.1}$$

$$\sup_{x\in(x_1,x_2)} |\hat{S}_{tple}(x) - S(x_1)| \xrightarrow{p} 0 \tag{A.2}$$

$$\sup_{x\in(x_2,x^*)} |\hat{S}_{tple}(x|X > x_2) - S(x|X > x_2)| \xrightarrow{p} 0 \tag{A.3}$$

where x^* lies in the interior of the support of the duration distribution function F and the censoring distribution function G, and $\tau^* = \inf\{t : H(t) > 0\}$ where $H(t)$ is the distribution function for the truncation times. Now, recall that TPLE estimate of $S(x)$ is

$$\hat{S}_{tple}(x) = \prod_{y_i\leq x}(1 - \frac{d_i}{R_i})$$

where $d_i = \sum_{j=1}^{n} I(y_j = y_{(i)}), R_i = \sum_{j=1}^{n} I(t_j \leq y_{(i)} \leq y_j)$, and $y_{(1)}, \cdots, y_{(k)}$ are distinct ordered observed failure times and $t_{(1)}, \cdots, t_{(k)}$ are the corresponding truncation times.

Proof of Theorem 1.3.1:

Since there is no intrinsic RENO between $[\tau^*, x_1]$, claim (A.1) above follows directly from the results of Tsai et al. (1987) or Lai and Ying (1991). In order to show claim (A.2), note that $\hat{S}_{tple}(x), x_1 \leq x \leq x_2$ can be factorized as follows:

$$\begin{aligned}\hat{S}_{tple}(x) &= \prod_{y_i \le x_1}(1-\tfrac{d_i}{R_i}) * \prod_{x_1 < y_i \le x}(1-\tfrac{d_i}{R_i}) \\ &= \hat{S}_{tple}(x_1) * \tfrac{\hat{S}_{tple}(x)}{\hat{S}_{tple}(x_1)}\end{aligned}$$

Since there are no observation made in the interval (x_1, x), the TPLE puts a mass of zero and hence $\hat{S}_{tple}(x_1) = \hat{S}_{tple}(x)$ which proves claim (A.2).

Finally, claim (A.3) can be proven by factorizing $\hat{S}_{tple}(x), \forall x > x_2$ as follows :

$$\begin{aligned}\hat{S}_{tple}(x) &= \prod_{y_i \le x_1}(1-\tfrac{d_i}{R_i}) * \prod_{x_1 \le y_i \le x_2}(1-\tfrac{d_i}{R_i}) * \prod_{x_2 < y_i \le x}(1-\tfrac{d_i}{R_i}) \\ &= \hat{S}_{tple}(x_1) * \tfrac{\hat{S}_{tple}(x_2)}{\hat{S}_{tple}(x_1)} * \tfrac{\hat{S}_{tple}(x)}{\hat{S}_{tple}(x_2)}\end{aligned}$$

Since there are no observation made in the interval (x_1, x_2), the TPLE puts a mass of zero and hence $\hat{S}_{tple}(x_1) = \hat{S}_{tple}(x_2)$. Thus,

$$\hat{S}_{tple}(x) = \hat{S}_{tple}(x_1) * \frac{\hat{S}_{tple}(x)}{\hat{S}_{tple}(x_2)}.$$

Now consider the TPLE estimator, $\hat{S}_*(x)$, which is constructed based only on observations $y_i > x_2$. Note that this estimator can equivalently be represented as

$$\hat{S}_*(x) = \prod_{x_2 < y_i \le x}(1-\frac{d_i}{R_i}) = \frac{\hat{S}_{tple}(x)}{\hat{S}_{tple}(x_2)}$$

where the second equality follows from the factorization of $\hat{S}_{tple}(x)$ as shown above. Since $\hat{S}_*(x)$ is a TPLE the results of Tsai et al. (1987) can be applied to show $\hat{S}_*(x)$ uniformly converges to $S(x|X > 2)$ i.e.

$$\sup_{x \in [x_2, x^*]} |\hat{S}_*(x) - \frac{S(x)}{S(x_2)}| \xrightarrow{p} 0 \ldots\ldots\ldots\ldots\ldots\ldots (1a)$$

Now observe that the only difference between $\hat{S}_*(x)$ and $\hat{S}_{tple}(x)$ is the presence of the extra term $\hat{S}_{tple}(x_1)$ in the equation for $\hat{S}_{tple}(x)$. Thus, we get

$$\hat{S}_{tple}(x) = \hat{S}_{tple}(x_1) * \hat{S}_*(x), \forall x > x_2.$$

Since $\hat{S}_{tple}(x_1) \xrightarrow{p} S(x_1)$ an application of Slutsky's theorem to (1a) leads to:

$$\sup_{x \in [\tau^*, x]} |\hat{S}_{tple}(x) - S(x)\frac{S(x_1)}{S(x_2)}| \xrightarrow{p} 0.$$

Lastly, observing $\hat{S}_{tple}(x_2) = \hat{S}_{tple}(x_1)$, and once again applying Slutsky's theorem to the above expression, we obtain the desired result:

$$\sup_{x \in (x_2, x^*)} |\hat{S}_{tple}(x|X > x_2) - S(x|X > x_2)| \xrightarrow{p} 0.$$

Appendix B

A Look at Within-Window (WW) Estimator from a Truncation Perspective

For a time-homogeneous Markov process, the within window (WW) estimator described in this paper can be seen as arising from a process where the last transition time before the sampling window is observed and the sojourn time in the last state entered just before the sampling window is known to be left truncated. In order to see this, let this last transition time, last state and sojourn time be denoted by T^l ,l and S . By definition of T^l, the sojourn time in the last state before the sampling window has to be long enough to make it into the sampling window. Let the next time of transition (i.e. after the beginning of the sampling window) be T^{l+1} and let the state entered at time T^{l+1} be $l+1$. Now the likelihood of our observed data is:

$$\begin{aligned} &\propto P(S = (T^{l+1} - T^l) | S \geq \mathcal{T} - w - T^l) \\ &* \prod_{t=T^{l+1}}^{\mathcal{T}_i - 1} P(X_i(t+1) = x_{t+1} | X_i(t) = x_t) \end{aligned} \tag{B.1}$$

We note that the sojourn time in S is geometrically distributed with parameter $1 - p_{ll}$, where p_{ll} is the probability of staying in the last state l or equivalently probability of making transition from state l to state l. Since geometric distribution is memoryless the first part of the above likelihood can be simplified as

$$P(S = (T^{l+1} - T^l)|S \geq \mathcal{T} - w - T^l) = \frac{\left(p_{ll}^{(T^{l+1}-T^l)-1} * (1 - p_{ll})\right)}{p_{ll}^{(\mathcal{T}-w-T^l)}}$$
$$= \prod_{t=\mathcal{T}-w}^{T^{l+1}-1} P(X(t+1) = x_{t+1}|X(t) = x_t)$$

where $x_t = l$ for $t \leq \mathcal{T} - w \leq T^{l+1} - 1$. Now substituting for the first part of equation B.1 with the simplified version, we get the following:

$$\propto \prod_{t=\mathcal{T}-w}^{\mathcal{T}-1} P(X(t+1) = x_{t+1}|X(t) = x_t) \tag{B.2}$$

which is the likelihood contribution for the data from within the window only. Therefore, we can view the WW estimator which is based on observations made within in the window only, as arising from a process where the sojourn time in the state just before the sampling window is left truncated.

Appendix C

Implication of the Markov Chain framework on the association between the duration of partnerships

The following section examines how the independence assumption for the sojourn times of the Markov model impacts the correlation between the various relationship durations. Let S_j^i denote the sojourn time in state i at the j^{th} visit. Let D_j represent the duration of the j^{th} relationship ordered by the formation time. We review common partnership formation patterns that arise in partnership data and investigate the implications of the time-homogeneous Markov assumption.

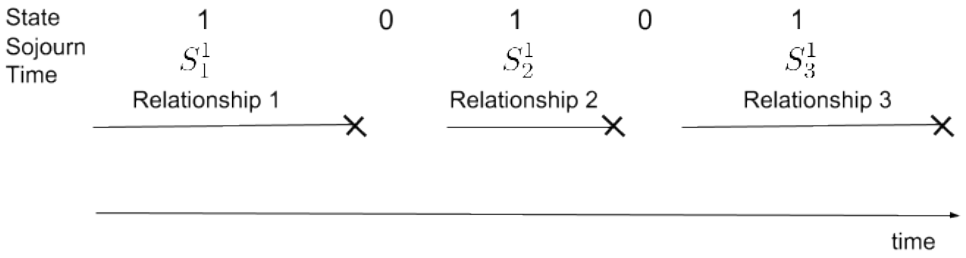
Case I: Serial Monogamy

In the case of serial monogamy the sojourn times in state 1 are identical to duration of the partnership as shown in Figure 4.1(a). Since the sojourn times in state i, S^i are independent and have a geometric distribution with parameter $1-p_{ii} = 1-\sum_{i\neq j} p_{ij}$, the duration of relationships will also be independent and geometrically distributed with parameter $1-p_{00} = p_{01}$.

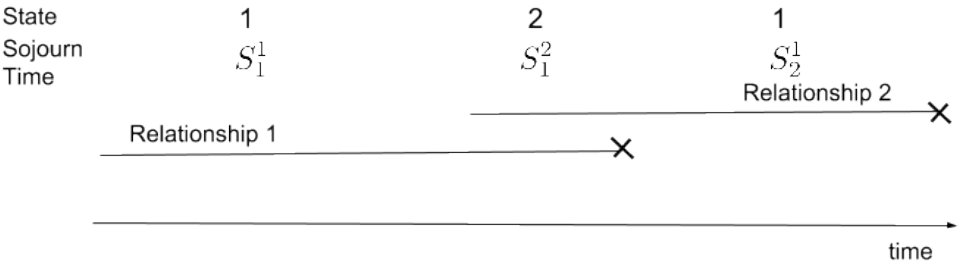
Case II: Transitional Concurrency

When we have an individual that has relationships that overlap in a transitional-concurrency manner (see Figure 4.1(b)), the independence assumption of the sojourn times implies that the relationship durations will be positively correlated as explained below. Let S_1^1 and S_2^1 represent so sojourn time in state 1 at the first visit and the second visit, respectively. Let S_1^2 represent the sojourn time in state 2 at the first visit. We now

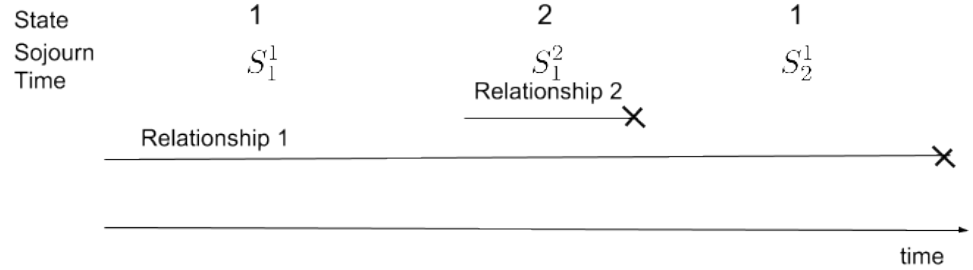
(a) Serial monogamy



(b) Transitional concurrency



(c) Embedded concurrency



(d) Embedded-transitional concurrency

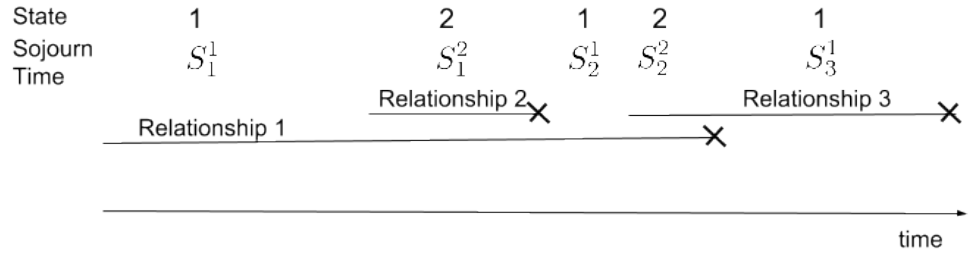


Figure 4.1: Duration and Markov chain data by partnership types

want to to know how the independence assumption of the sojourn times affects the association between duration of relationship 1 ($S_1^1 + S_1^2$) and relationship 2 ($S_1^2 + S_2^1$). In order to see if the duration of relationship 1 is independent of duration of relationship 2, we can calculate the covariance of the two durations, $D_1 = S_1^1 + S_1^2$ and $D_2 = S_1^2 + S_2^1$:

$$\begin{aligned} &COV(D_1, D_2) = COV(S_1^1 + S_1^2, S_1^2 + S_2^1) \\ &= COV(S_1^1, S_1^2) + COV(S_1^1, S_2^1) + COV(S_1^2, S_1^2) + COV(S_1^2, S_2^1) \\ &= VAR(S_1^2) = \frac{1 - p_{21}}{p_{21}^2} > 0 \end{aligned}$$

The third equality holds since the sojourn times are independent and the last equality holds since S_1^2 is Geometric (p_{21}). Therefore, the duration of relationship 1 and relationship 2 are positively correlated.

Case III: Embedded Concurrency

In the context of embedded concurrency (see Figure 4.1(c)), the duration of the first and second relationships are $D_1 = S_1^1 + S_1^2 + S_2^1$ and $D_2 = S_1^2$, respectively. We can now determine the covariance between the two relationships following similar reasoning as above.

$$\begin{aligned} &COV(D_1, D_2) = COV(S_1^1 + S_1^2 + S_2^1, S_1^2) \\ &= COV(S_1^1, S_1^2) + COV(S_1^2, S_1^2) + COV(S_2^1, S_1^2) \\ &= VAR(S_1^2) = \frac{1 - p_{21}}{p_{21}^2} > 0 \end{aligned}$$

Once again, we note that the covariance between the two durations of relationships is positive. In addition, it appears that whether the concurrency is embedded or transitional the covariance between two relationships stays the same.

Case IV: Embedded Concurrency followed by Transitional Concurrency

We now examine the association between duration of relationships in the context of an embedded concurrency followed by (see Figure 4.1(d)). The durations of the three relationships are given by: $D_1 = S_1^1 + S_1^2 + S_2^1 + S_2^2$, $D_2 = S_1^2$ and $D_3 = S_2^2 + S_3^1$, respectively. Using similar reasoning as in the prior sections, we conclude the covariance between the

different durations is:

$$COV(D_1, D_2) = VAR(S_1^2) = \frac{1 - p_{21}}{p_{21}^2} > 0$$
$$COV(D_1, D_3) = VAR(S_2^2) = \frac{1 - p_{21}}{p_{21}^2} > 0$$
$$COV(D_2, D_3) = 0$$

We note that relationship 1 is positively correlated with both relationships 2 and 3. Although both relationships 2 and 3 are concurrent with relationship 1, the covariance between the two relationships is 0 as the two relationships do not overlap relative to one another.

In general, it appears that durations that overlap will have a positive covariance and those that do not overlap (relative to one another) will have zero covariance. In addition since the duration of a relationship can be determined as the sum of independent (not necessarily identical) geometrically distributed sojourn times, the marginal distribution of durations can be determined using convolution. In summary, the Markov model assumption has important implications on the distribution of the duration of relationships as discussed in the above cases. Therefore, simulations of duration data in a Markov chain framework have to account for the dependence between the Markov chain model and duration data model.