Technical Report to

"A BIVARIATE INTERVAL CENSORSHIP MODEL FOR PARTNERSHIP FORMATION"

By Linda Yuet-Yee Wong and Qiqing Yu*

Department of Economics and Department of Mathematical Sciences,

Binghamton University,

Binghamton, NY 13902, USA

Current version: 1/12/2006

AMS 1991 subject classification: Primary 62 J 05; Secondary 62 G05.

Key Words: bivariate mixed interval-censored data, missing data, generalized MLE, exact observations, asymptotic properties.

Abstract: We consider a statistical problem of estimating a bivariate age distribution of newly formed partnership. The study is motivated by a type of data that consist of uncensored, right-censored, left-censored, interval-censored and missing observations in the coordinates of a bivariate random vector. A model is proposed for formulating such type of data. A feasible algorithm to estimate the generalized MLE (GMLE) of the bivariate distribution function is also proposed. We establish asymptotic properties for the GMLE and apply the method to the data set.

* corresponding author: Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902, USA, Fax:+6077772450.

Email address: qyu@math.binghamton.edu (Qiqing Yu)

1. Introduction

We consider estimation of a joint distribution function of a bivariate random vector with interval-censored data. The problem is applied to a marriage dataset, and so the variables of interests are the ages at first marriage of a couple, denoted by $\mathbf{X}$. We observe data on individuals' marital status over a period of 20 years. The data consist of uncensored, right-censored, left-censored, interval-censored and missing observations. We use the estimate of the joint distribution function to estimate the spousal age correlation and the marginal distributions of males' and females' ages at first marriage.

A large literature examines age at first marriage, (see, for example, Goldin and Katz [1], Loughran [2], Bloom and Bennett [3], and Keeley [4]). While marriage data naturally come in bivariate coordinates, typical empirical studies consider univariate analysis and assume a normal distribution for age. Furthermore, existing studies typically analyze age at first marriage at a point in time, including individuals from diverse backgrounds and ignoring the dynamics in individuals' changing marital status. Some recent studies apply univariate duration analysis to assess age at first marriage such as Aassve *et al* [5] and Berrington and Diamond [6]. However, these studies do not tackle all the censoring aspects of the panel data and ignore the joint feature of partnership. Because spousal age correlation significantly deviates from zero, a bivariate analysis seems more appropriate.

Transition models are widely used to analyze panel data. Some applications, such as estimation of transition probabilities and first passage times in match/partnership formation, require knowledge of the exact timing of the transition and the joint observations of a match (Wong [7,8]). Very often in matching data sets, the transition information of some agents is either missing or is only known to lie within a certain interval (interval censored). Existing methods handle these data by dropping unknown or missing event timing or missing partners' information, and often assume a completely parametric form of the predictor

for the first passage times.

We propose a bivariate interval censorship model which allows observations in each variate of the random vector $\mathbf{X}$ be uncensored, right censored, left censored, interval censored and missing. The model is applicable to non-parametric, parametric and semi-parametric approaches associated with some meaningful covariates. In this paper, we consider the non-parametric approach and assume that the joint distribution function F_0 of $\mathbf{X}$ is totally unknown. We propose to estimate F_0 by the generalized (non-parametric) maximum likelihood estimator (GMLE). We propose a feasible algorithm to obtain the GMLE. We then establish the consistency and asymptotic normality of the GMLE under a set of reasonable assumptions.

Our approach builds on the statistical literature that examines univariate and multivariate interval-censored data. For univariate interval-censored data, Groeneboom and Wellner [9] propose a univariate case 2 interval censorship model in which there are exactly two follow-ups for each individual. Under this model, observations either fall before the first follow-up time, between the two follow-up times or after the second follow-up time. In reality, there was rarely a follow-up study in which each individual was followed exactly twice. Schick and Yu [10] propose a univariate mixed case interval censorship model, which allows the number of follow-ups to be random. While these two models do not allow uncensored observations, Yu, Wong and Li [11] extend the mixed case model to a univariate mixed interval-censorship model. The model allows both interval-censored observations and uncensored observations.

For multivariate interval-censored data, Wong and Yu [12] propose a multivariate case 2 interval censorship model. Betensky and Finkelstein [13] also consider- non-parametric estimation with bivariate interval-censored data. Van der Vaart and Wellner [14] and Yu *et al* [15] establish consistency of the GMLE under a multivariate mixed case model. Again,

these models do not allow uncensored observations.

The major difference between the aforementioned studies and our paper lies on the unique structure of our dataset. The dataset contains observations that are uncensored, right censored, left censored, and interval censored. It also contains missing observations. That is, the information on a particular variate is missing completely, either because a respondent did not give the information on the spouse or the information was obviously ridiculous. Missing data without covariates do not contribute any information in the univariate case. Thus they can be deleted from the data set and the univariate interval censorship models do not consider such missing data. However, missing data in a variate of the bivariate random vector do contain information. Thus, the main innovative feature of our paper is the formation of a multivariate interval censorship model that allows all types of exact, censored and missing data and that fits the marriage data.

We make two additional contributions. First, we propose a feasible and efficient algorithm for finding the collection of all MIs. Second, we propose a method to simplify the self-consistent algorithm. Typical methods for obtaining the GMLE based on univariate interval-censored data is the self-consistent algorithm (see Turnbull [16]). The algorithm is based on finding all maximal intersections of the observed intervals (for the definition of the maximal intersection (MI), we refer to Section 3). The self-consistent algorithm can easily be extended to the multivariate interval censoring. However, there are two computational problems which need to be addressed. The first problem is that it is difficult to find the collection of all MIs in a multivariate interval censoring setting. To the best of our knowledge, Betensky and Finkelstein [13] is the first study that discusses an algorithm for finding MIs with bivariate interval-censored data. However, their algorithm is less efficient than ours and is for a different type of data (see Remark 4 of our paper). The second problem is that for a large dataset such as our marriage dataset (with 11,774 observations), the original

self-consistent algorithm can be time consuming and may not be feasible. In this paper, we propose to simplify the self-consistent algorithm.

Our results show that the marginal distribution of age at the first marriage has a single peak and is skewed to the right, in contrast to the common assumption of normal distribution in regression analyses. More intriguingly, our results indicate a serious problem when ignoring all kinds of censored and missing data, and the bivariate nature of a matching data. We compare our results with those that consider only univariate case and drop all censored and missing data. For example, using the same data source as that in our paper, Gould ([17], p.5, p.23, and Figure 3a on p.42) shows that the age at first marriage distribution has an increasing slope, which is in stark contrast with our results.

In Section 2, we describe the data and propose a statistical partnership model. Section 3 presents a method to estimate the GMLE of the joint distribution function. In Section 4, we present the joint distribution of the data and illustrate the idea via a subset of the data. In Section 5. we prove the consistency and the asymptotic normality of the GMLE under certain regularity assumptions. Section 6 is a concluding remark.

2. Model Description

The data for this analysis are from the National Longitudinal Survey of Youth 1979-98 (NLSY). The 1979-98 cross-sectional and supplemental samples consist of 11,774 respondents, who were between the ages of 14 and 22 in 1979. The samples were core nationally representative random samples. Interviews were conducted yearly from 1979 through 1994; since then data were recorded bi-annually.

Because our focus is on age at first marriage, we use only the first marriage spell, even though longer marriage histories are available. In particular, the variables of interests are the first marriage ages of these youths and their (future) spouses. This is a bivariate random vector.

There are two ways to determine when marriage occurs for respondents. First, we use the reported ages at first marriage. This results in exact data. Second, if there is no report on the ages at first marriage, we impute the ages using data on the respondent's marital status. Starting from 1978, respondents were asked about their marital status; if there was a change from singlehood to marriage, we have interval-censored data of the first marriage. There were 9009 out of 11,774 respondents who were ever married, of which 8891 cases contain the known age at first marriage of the respondent but not the spouse after applying the methods described above, and the remaining 118 cases contain no information concerning age at first marriage.

The ages of spouses in responses may not refer to their ages at first marriage. So, from the age variable we may only know that the spouses' first marriage happened at or prior to the current marriage year. Another problem relating to spouses' ages is that it is poorly recorded. Among the 9009 ever marriage cases, only 1535 cases contain spouses' age data. Among those non-missing data, 11.8 percent were below 15 years of age, and 6.91 percent were below age 8. To make it reasonable, we assume that ages that fall below 12 at the first marriage are missing.

Thus, we have exact observations on respondents whose first marriage age is known. In addition, three types of censored observations are present.

a. Some respondents were left-censored, with a starting marital status in 1979 being not single. Only their ages in 1979 were recorded.

b. Before 1994, some respondents were single, became non-responsive for a couple of years in their annual responses and then became married or divorced. Some but not all did not give their ages at first marriage in their responses. Thus the first marriage ages were interval censored for this portion of respondents, but were exact for the rest.

c. Between 1994 and 1998, respondents changed their marital status in the biannual re-

sponses (even though they did not become non-responses). Some but not all did not give their ages at first marriage in their responses. Thus their first marriage ages were also interval censored.

d. By the end of the follow-up, some individuals had not married yet; thus we have right-censored observations on these individuals.

e. If the first marriage age of a respondent was right censored, the information on the spouse was missing. Otherwise, the age of the spouse of a respondent may or may not have been reported (even though they did not become non-respondents). If it was reported, it was not reported whether it was the age at the spouse's first marriage. Thus we only have left-censored observations.

In order to formulate a model for this data set, we make use of the following notations. Imagine that we have a couple and a bivariate random vector $\mathbf{X} = (X_1, X_2)^t$, where $\mathbf{X}^t$ is the transpose of the vector $\mathbf{X}$ and X_1 and X_2 are the first marriage ages of the male and female, respectively. (Either the female is the spouse of the male respondent at his first marriage or the male is the spouse of the female respondent at her first marriage.) For $i = 1, 2$, our model is a mixture of various case k models and a right censorship model. In particular, Let K_1 and K_2 be random nonnegative integers. For $i = 1, 2$, if $K_i = 0$, X_i is subject to right censoring, namely, there is a random variable $Y_{i,0,0}$ such that we observe $(\min(X_i, Y_{i,0,0}), \mathbf{1}(X_i \leq Y_{i,0,0}))$, where $\mathbf{1}(\cdot)$ is an indicator function. If $K_i > 0$ then X_i is subject to a case K_i model, with the age at the j-th follow-up $Y_{i,K_i,j}$, $j = 1, \ldots, K_i$. Thus, if $K_1 > 0$ and $K_2 > 0$, K_1 and K_2 are the numbers of follow-ups for the male and the female whose exact first marriage ages were not reported, respectively. $Y_{1,K_1,j}$ and $Y_{2,K_2,j}$ are the ages of the male and female at the j-th follow-up time, respectively. For convenience, let $Y_{i,K_i,0} = -\infty$ (though we could also let $Y_{i,K_i,0} = 0$) and $Y_{i,K_i,K_i+1} = \infty$ for $i = 1, 2$ and $K_i \geq 1$. To deal with the situation that the information was missing, if $K_i = 1$, there are

two possibilities for $Y_{i,K_i,1}$:

(1) $Y_{i,K_i,1} \in (0,\infty)$. It corresponds to the univariate case 1 interval censoring.

(2) $Y_{i,K_i,1} = 0$. It corresponds to missing information on X_i and we have $X_i \in (0,\infty]$.

Denote $\mathbf{K} = (K_1, K_2)$, $\mathbf{Y} = \{(Y_{i,K_i,j} : \ i = 1,2; j = 1, ..., K_i\} \cup \{Y_{1,0,0}, Y_{2,0,0}\}$. We assume that the observable random vector is (L_1, R_1, L_2, R_2), where for $i = 1$ or 2,

$$(L_i, R_i) = \begin{cases} (X_i, X_i) & \text{if } K_i = 0 \text{ and } X_i \le Y_{i,0,0} \text{ (exact)}, \\ (Y_{i,0,0}, \infty) & \text{if } K_i = 0 \text{ and } X_i > Y_{i,0,0} \text{ (right-censoring)} \\ (0,\infty) & \text{if } K_i = 1 \text{ and } Y_{i,1,1} = 0 \text{ (missing)}, \\ (Y_{i,K_i,j-1}, Y_{i,K_i,j}) & \text{if } K_i \ge 1,\ Y_{i,K_i,1} > 0,\ Y_{i,K_i,j-1} < X_i \le Y_{i,K_i,j}, \\ & \quad j = 1, ..., K_i + 1 \text{ (interval censoring)}. \end{cases} \tag{2.1}$$

Notice that the first condition in (2.1) covers no censoring, the second condition covers right censoring, the third condition covers missing cases, and finally interval censoring and left (right) censoring are covered in the fourth condition. We also make use of the following assumptions.

A1. $\mathbf{X}$ and ($\mathbf{K}$, $\mathbf{Y}$) are independent.

A2. $P\{K_i = 0\} > 0$, $i = 1, 2$.

A1 is a typical identifiability condition. Assumption (2.1) together with Assumption A1 formulate a general bivariate interval censorship model that allows possible exact observations and missing observations. It includes the models studied by Wong and Yu [12] and van der Vaart and Wellner [14], which do not allow exact observations and missing data, by letting $P\{K_i = 0\} = 0$ and $P\{K_i = 1 \text{ and } Y_{i,1,1} = 0\} = 0$, $i = 1, 2$.

Assumption (2.1) together with Assumptions A1 and A2 formulate a general bivariate interval censorship model that allows exact observations. A2 emphasizes that there are exact observations. Assumption A2 distinguishes the current model from the models studied by Wong and Yu [12] and van der Vaart and Wellner [14].

We want to estimate the joint distribution function F_0 of (X_1, X_2), where

$$F_0(\mathbf{x}) = P\{X_1 \le x_1, X_2 \le x_2\} \text{ and } \mathbf{x} = (x_1, x_2).$$

For convenience, we make use of the observable rectangle $\mathcal{I}$, that is,

$$\mathcal{I} = \begin{cases} [L_1, R_1] \times (L_2, R_2] & \text{if } L_1 = R_1 \text{ and } L_2 < R_2, \\ (L_1, R_1] \times (L_2, R_2] & \text{if } L_1 < R_1 \text{ and } L_2 < R_2, \\ (L_1, R_1] \times [L_2, R_2] & \text{if } L_1 < R_1 \text{ and } L_2 = R_2, \end{cases}$$

3. Method of estimation

Let $(L_{i1}, R_{i1}, L_{i2}, R_{i2}, \mathcal{I}_i)$ be i.i.d. copies of $(L_1, R_1, L_2, R_2, \mathcal{I})$. Kiefer and Wolfowitz [18] first introduced the concept of the generalized likelihood function. Under the multivariate interval censoring, the generalized likelihood function becomes $\Lambda_n = \prod_{i=1}^n \mu_F(\mathcal{I}_i)$, where μ_F is the measure induced by an unknown distribution function F, *i.e.*, $\mu_F(\mathcal{I}_i) = \int_{(x,y)\in\mathcal{I}_i} dF(x, y)$. Let τ_1 be the maximum finite value of L_1 and R_1 and τ_2 be the maximum finite value of L_2 and R_2. Let τ_{01} and τ_{02} be the smallest possible age for a male and female to marry, respectively.

Define a *maximal intersection*, A, with respect to the $\mathcal{I}_i$'s to be a nonempty finite intersection of the $\mathcal{I}_i$'s such that for each i, $A \cap \mathcal{I}_i = \emptyset$ or A. Let $\{A_1, ..., A_m\}$ be the collection of all possible distinct MIs. For our marriage study data set, typically, $[x, x] \times [y, y]$ (the intersection of observations $[x, x] \times (0, \infty]$ and $(0, \infty] \times [y, y]$) is an MI. Moreover, $(\tau_1, \infty] \times [y, y]$ is another MI (the intersection of observations $(\tau_1, \infty] \times (0, \infty]$ and $(0, \infty] \times [y, y]$) .

Using an argument similar to Wong and Yu [12], it can be shown that the GMLE of $F_0(\mathbf{x})$ which maximizes the generalized likelihood function, Λ_n, must assign all the probability masses to the sets $A_1, ..., A_m$. Thus it suffices to maximize the generalized likelihood function of the following form:

$$\Lambda_n = \prod_{i=1}^n \mu_F(\mathcal{I}_i) = \prod_{i=1}^n [\sum_{j=1}^m \mathbf{1}(A_j \subset \mathcal{I}_i) s_j], \tag{3.1}$$

where $\mathbf{s}$ $(= (s_1, ..., s_{m-1})^t) \in D_s$, $s_m = 1 - s_1 - \cdots - s_{m-1}$, and $D_s = \{\mathbf{s};\ s_i \geq 0, s_1 + \cdots + s_{m-1} \leq 1\}$. Denote the GMLE of $\mathbf{s}$ by $\hat{\mathbf{s}}$ and that of F_0 by $\hat{F}_n$.

The s_js can be obtained by the self-consistent algorithm described by Turnbull [16] for univariate interval-censored data as follows: Let $s_j^{(0)} = 1/m$ for $j = 1, ..., m$. Denote $\delta_{ij} = \mathbf{1}(A_j \subset \mathcal{I}_i)$. At the h-step, update s_j by

$$s_j^{(h)} = \sum_{i=1}^{n} \frac{1}{n} \frac{\delta_{ij} s_j^{(h-1)}}{\sum_{k=1}^{m} \delta_{ik} s_k^{(h-1)}}, \ j = 1, ..., m, \ h \geq 1. \tag{3.2}$$

Repeat until the s_j's converge. The justification of the convergence of this method for multivariate interval-censored data is similar to that given in Turnbull [16] for univariate data. The algorithm is easy to implement. A more efficient algorithm may be obtained by mimicking the algorithms for univariate interval-censored data discussed in Wellner and Zhan [19].

Given a GMLE $\hat{\mathbf{s}}$, the GMLE of $F_0(\mathbf{x})$ is not uniquely defined on an MI unless the MI is a singleton. A GMLE of $F_0(\mathbf{x})$ can be obtained as follows:

$$\hat{F}_n(\mathbf{x}) = \sum_{A_j \subset [0,x_1] \times [0,x_2]} \hat{s}_j. \tag{3.3}$$

The GMLE of $\mathbf{s}$ may not be unique under multivariate interval censoring, however, the GMLE of $\mu_F(\mathcal{I}_i)$ is uniquely determined for each i (see Yu, Wong and He [20]).

Hereafter, we address two issues concerning the empirical implementation of computing the GMLE. We first propose an algorithm for searching for the MIs, and then a method to simplify the self-consistent algorithm.

The following algorithm is a feasible algorithm for implementing any bivariate interval-censored data. The algorithm can be generalized to handle multivariate interval-censored data.

1. *(***Search all the MIs of the observable intervals corresponding to** (L_{i1}, R_{i1})**s***). Partition the observations (L_{i1}, R_{i1})s into two groups: the group of all exact observations (i.e. $L_{i1} = R_{i1}$) and the group of all interval-censored observations (i.e.*

$L_{i1} < R_{i1}$). *Let* $e_1 < e_2 < \cdots < e_{m_0}$ *be all the distinct exact observations. Let* $l_1 < l_2 < \cdots < l_{m_1}$ *be all the distinct values of* $L_{i1}s$ *such that* $L_{i1} < R_{i1}$ *and* $r_1 < r_2 < \cdots < r_{m_2}$ *be all the distinct values of* $R_{i1}s$ *such that* $L_{i1} < R_{i1}$. *Let* (l_{i_1}, r_{i_1}), ..., (l_{i_k}, r_{i_k}) *be all possible intervals such that*

(a) l_{i_1}, ..., $l_{i_k} \in \{l_1, ..., l_{m_1}\}$,

(b) r_{i_1}, ..., $r_{i_k} \in \{r_1, ..., r_{m_2}\}$,

(c) $(l_{i_j}, r_{i_j}) \cap \{l_1, ..., l_{m_1}, r_1, ..., r_{m_2}\}$ *is empty for each* j.

Delete the pairs (l_{i_j}, r_{i_j}) *for which the interval* $(l_{i_j}, r_{i_j}]$ *contains some* e_i *given above. Denote the remaining pairs by* (p_1, q_1), ..., (p_{m_3}, q_{m_3}) *and denote* $p_{m_3+1} = q_{m_3+1} = e_1$, ..., *and* $p_{m_4} = q_{m_4} = e_{m_o}$. *Notice that* $a_1 = (p_1, q_1]$, ..., $a_{m_3} = (p_{m_3}, q_{m_3}]$, $a_{m_3+1} = [p_{m_3+1}, q_{m_3+1}]$, ..., $a_{m_4} = [p_{m_4}, q_{m_4}]$ *are all the MIs of the observable intervals corresponding to* $(L_{i1}, R_{i1})s$. *By reordering, without loss of generality (WLOG), we can assume that* $a_1 \leq a_2 \leq \cdots \leq a_{m_4}$, *that is, the endpoints of these MIs* $a_i s$ *satisfy* $p_1 \leq q_1 \leq p_2 \leq \cdots p_{m_4} \leq q_{m_4}$.

2. *(***Search all the MIs of the observable intervals corresponding to** (L_{i2}, R_{i2})**s**). *The method is the same as in Step 1. Let* u_i *and* v_i, $i = 1$, ..., m_5, *be the endpoints of the resulting MIs* $b_i s$, *and assume that* $u_1 \leq v_1 \leq u_2 \leq v_2 \leq \cdots \leq u_{m_5} \leq v_{m_5}$.

3. **A substitution of MIs**. *Let* A_1^*, .., A_m^* *be all the distinct product sets of the form* $a_i \times b_j$. *These* A_h^* *may not be MIs of* $\mathcal{I}_i s$, *but they can play the role of the MIs in finding a GMLE of* F. *It can be shown that each* A_j^* *is either a subset of an MI or does not intersect with each MI. One can stop here and take these* $A_j^* s$ *as substitutions of the MIs, or go to the next step to find all the real MIs.*

4. **Search for real MIs**.

4.1. For each A_j, *one can find the smallest intersection* B_j *of* $\mathcal{I}_i s$ *which contains* A_j *in the following steps.*

4.1.1. If $A_j^* \subset \mathcal{I}_1$, let $B_j = \mathcal{I}_1$.

4.1.2. For $i = 2$, ..., n, if $A_j^* \subset \mathcal{I}_i$ then update B_j by $B_j \cap \mathcal{I}_i$. That is, let $B_j \cap \mathcal{I}_i$ be the new B_j. By definition, each side of the intersection of $B_j \cap \mathcal{I}_i$ is the closest one, among the two corresponding sides of B_j and $\mathcal{I}_i$, to the rectangle A_j^*. B_j obtained at the end of this step is a potential MI that contains A_j.

4.2. For each B_j resulting from Step 4.1.2, check whether it is an MI. It can be shown that if B_j is an MI, then it either does not intersect with all the other B_k, $k \neq j$, or is a subset of some B_k. Thus if there exists a B_k such that $B_j \cap B_k \neq \emptyset$ and $B_j \cap B_k \neq B_j$, then B_j is not an MI.

4.3. Let A_1, ..., A_m be all the distinct MIs resulting from Step 4.2.

Remark 3. Two methods are proposed in the above algorithm for finding the MIs. One method is to find all real MIs and the second is to find a substitution of the collections of all MIs (ending at Step 3). The advantage of the second method is the symmetry of the A_j^*s, which are arranged in a rectangular array of rectangles. The disadvantage of the method is that it may increase the computational burden.

Application of formula (3.2) in implementing the self-consistent algorithm can be time-consuming when the sample contains over 10,000 observations. To overcome this difficulty, we propose a method to simplify the self-consistent algorithm.

We recognize that some distinctive observations bear the same data information in our sample. In other words, the data set contains many replications. By reordering the observable rectangles, without loss of generality, we can assume that $\mathcal{I}_1$, ..., $\mathcal{I}_M$ are all the distinct I_is and the number of replications are N_1, ..., N_M, respectively. Then (3.2) can be replaced by

$$s_j^{(h)} = \sum_{i=1}^{M} \frac{N_i}{n} \frac{\delta_{ij} s_j^{(h-1)}}{\sum_{k=1}^{m} \delta_{ik} s_k^{(h-1)}}, \ j = 1, ..., m, \ h \geq 1. \tag{3.4}$$

For the marriage study, $M = 724$, which is much smaller than $n = 11,774$.

Remark 4. The self-consistent algorithm (3.2) and the definition of the MIs are similar to those proposed in Wong and Yu [12], Betensky and Finkelstein [13] and van der Vaart and Wellner [14]. Wong and Yu [12], and van der Vaart and Wellner [14] do not discuss the algorithm for finding all MIs, Betensky and Finkelstein [13] propose an algorithm. There are two differences between thei algorithm and ours.

(1) Data forms are different. In their set-up, observed intervals in each coordinate are of the form $[a, b]$, where as in our set-up they are of the form either $(a, b]$ or $[c, c]$.

(2) They develop their algorithm directly from the definition of the MIs, without considering efficiency. In fact, one reason that Wong and Yu [12] do not discuss the algorithm in their paper is that an algorithm can be formed directly from the definition of the MIs. The algorithm proposed in this paper is faster.

4. Data analysis on the marriage study

4.1. Data analysis on a subsample

The sample contains 11,774 observations. For illustration purpose, we present a sample of 100 observations shown in Table 1. Let us explain the entries in the first few rows of the table.

1. Among the 7 cases, a female respondent was 22 years old in 1979. She was not married by the end of the study.
2. A female respondent was 19 years old in 1979. She reported that she was married and her spouse was 19 years old in 1979.
3. Among 8 cases, a female respondent was 21 years old in 1979. She reported that she was divorced, but did not report her former spouse's age.
4. A male respondent was 17 years old in 1979. He reported that he was married but did not report his spouse's age.

Put Tables 1 and 2 here.

The GMLE of the joint survival function $S(x,y) = P\{X_1 > x, X_2 > y\}$ is given in Table 2. The entries in the first column are the ages of the males. The entries in the first row are the ages of the females. In Table 2, the first entry in each cell corresponding to ages (i,j) is the GMLE of the survival function $S(i,j)$ and the second entry is its standard deviation.

It is worth mentioning that for the current sub-sample, since the sample size is small, the information matrix $\hat{J}$ $(= -\frac{\partial^2 \mathcal{L}_n(\hat{F}_n)}{\partial \mathbf{S} \partial \mathbf{S}^t})$ is singular. Thus, we cannot use $\hat{J}^{-1}$ to estimate the covariance matrix of $\hat{F}_n$. The standard deviation of the GMLE in Table 2 were obtained by the procedure developed in Yu, Wong and He [20]. We skip the details.

Since the largest observations for males and females are 40 and 39, respectively, it is not appropriate to compute the correlation between X_1 and X_2. However, we can compute the conditional correlation between X_1 and X_2, given $X_1 \leq 40$ and $X_2 \leq 39$. The conditional correlation is 0.59.

We want to emphasize that the current sample is not a random sample of the original dataset, in a sense that we collected certain typical cases for illustration only. Thus, it is not surprising that there is a mismatch of the survival function in Table 2 and the density function of the whole data set displayed in §4.2.

4.2. Data analysis on the full data set

We computed the GMLE of F_0 and the correlation coefficient between X_1 and X_2, which is 0.82. The program was written in C (language) and the computation was performed on a Pentium III personal computer. Even though the size of the marriage data is large $(n = 11774)$, it took less than 5 minutes to obtain these estimates. The graph of the GMLE of the joint density function of (X_1, X_2) is given in Figure 1. The shape of the distribution does not resemble a typical joint normal distribution.

Note from Figure 1 that there is a sharp jump in the joint density function f at $(42, 44)$,

and similarly at the marginal density functions (see Figure 2). This is mainly due to the property of the GMLE and the fact that (1) the largest observations in the data set for male and female are 42 and 44, respectively, and they are both exact observations, and (2), there was a portion of the population that were not married by the end of the study. It is well known that the GMLE is not stable at the endpoints. For the current analysis, we should ignore the value of the distribution at the endpoints.

Put Figures 1 and 2 here

Figure 2 presents a plot of the marginal density for males and females. The vertical lines correspond to the median age at first marriage. The median age difference is 2. After the peak age of the first marriage for males, the two densities become more similar.

Note that the marginal density is equivalent to the density of the waiting time with a shift. The marginal density is skewed to the right and does not follow a normal distribution, as assumed widely in marriage studies.

These results differ remarkably from Gould [17], who uses the same data set as ours to study the marriage and career decisions of young men. Gould reduces the sample to male only cases and ignores interval censoring issues, the sample size is $n = 2155$ (Gould ([17], p.5 and p.23)), out of total of 11,774. His plot of the distribution of age at first marriage (Gould ([17], Figure 3a, p.42)) shows an upward-sloping curve, becoming flat only after age 30! The plot is peculiar not only because it is non-intuitive to have more people married at older ages, but also it does not at all resemble what we have found in the data.

Further, the result of the waiting time to the first marriage distribution (and the hazard rate, not shown) having a single peak indicates that an exponential model cannot match the data, as used widely in the theoretical marriage-search literature. So, caution must be taken when analyzing transition data, whether it is in marriage, the entry of firms, merger, jobs or firms turnover and so on.

5. Asymptotic properties of the GMLE

We shall establish the asymptotic properties of the GMLE making use of the following assumptions.

A3. (L_1, R_1, L_2, R_2) takes on finitely many values and F_0 is discrete.

A4. $P\{X_1 > \tau_1, X_2 > \tau_2\} > 0$, $P\{X_1 < \tau_{01} \text{ or } X_2 < \tau_{02}\} = 0$, $P\{(X_1, X_2) = (x, y)\} > 0$, $P\{X_1 = x, X_2 > \tau_2\} > 0$ and $P\{X_1 > \tau_1, X_2 = y\} > 0$, for each $x \in [\tau_{01}, \tau_1]$ and $y \in [\tau_{02}, \tau_2]$.

In this study, all data in NLSY were rounded off to integers (number of years). There were 11,774 respondents and the follow-up lasted for 20 years. Thus we can assume that A3 holds. A4 says that there are still men (or women) having their first marriage beyond the age of τ_1 (or τ_2). Denote $\mathcal{A}$ the collection of the subsets of the forms $(\tau_1, \infty] \times (\tau_2, \infty]$, $[x, x] \times (\tau_2, \infty]$, $(\tau_1, \infty] \times [y, y]$, $[x, x] \times [y, y]$, where $x \in [\tau_{01}, \tau_1]$ and $y \in [\tau_{02}, \tau_2]$.

Theorem 1. *Under assumptions A1 and A3, $\mu_{\hat{F}_n}(\mathcal{I}_i) \to \mu_{F_0}(\mathcal{I}_i)$ a.s. for each i.*

Theorem 2. *Under assumptions A1, A2 and A3 the GMLE $\hat{F}_n(x, y)$ is strongly consistent at each $(x, y) \in \{x \le \tau_1, y \le \tau_2\}$.*

Theorem 3. *Under assumptions A1, A2, A3 and A4, $\sqrt{n}\begin{pmatrix} \hat{s}_1 - s_1^o \\ \vdots \\ \hat{s}_{m-1} - s_{m-1}^o \end{pmatrix}$ is asymptotically normal with mean 0 and dispersion matrix J^{-1}, where $\mathbf{s}_j^o = \mu_{F_0}(A_j)$. A strongly consistent estimator of J is given by $\hat{J} = -\frac{\partial^2 \mathcal{L}_n(\hat{F}_n)}{\partial \mathbf{S} \partial \mathbf{S}^t}$. Furthermore, $\sqrt{n}[\hat{F}_n(\mathbf{x}) - F_0(\mathbf{x})]$ is asymptotically normally distributed for all $\mathbf{x} \in \mathcal{A}$. A consistent estimate of the asymptotic variance of $\hat{F}_n(\mathbf{x})$ is $\frac{1}{n}\mathbf{c}^t \hat{J}^{-1}\mathbf{c}$, where $\mathbf{c}$ is a $(m-1) \times 1$ vector with the i-th entry $c_i = \mathbf{1}(A_i \subset [0, x_1] \times [0, x_2])$ unless $F_0(\mathbf{x}) = 1$.*

Under assumptions A1, A2, A3 and A4, the problem becomes an estimation problem of multinomial distribution and it follows from standard argument that the GMLE $\hat{F}_n$ is asymptotically normally distributed and the convergence rate is in $\sqrt{n}$. The proofs are very similar to the proofs in Wong and Yu [12], and thus they are put in a technical report

(see Wong and Yu [21]). Under the assumptions in Theorem 3, the GMLE $\hat{F}_n$ is also asymptotically efficient. The proof of this assertion is straightforward and is omitted.

6. Conclusion

We have proposed a feasible algorithm to estimate the GMLE of a bivariate random vector with uncensored, right-censored, left censored, interval censored and missing data. We consider a non-parametric approach and the GMLE is shown to be consistent and asymptotically normal The algorithm proposed in this paper, that aims at searching for all the MIs and simplifying the self-consistent algorithm, substantially reduces the computational burden.

Throughout the paper we have focused on developing the GMLE for the bivariate age distribution and the algorithm for computation. Because the model is new, we feel that we should first understand the properties of the estimator and its implementation. The discrete assumption we have imposed in the paper arises logically from the marriage data set. However, the method proposed is also valid for continuous random variables. The proofs of the theorems for the continuous random variables are more difficult to construct. We intend to work on it in a future project.

Results reflect that ignoring interval-censored data and the bivariate aspect of the data may produce an erroneous picture about the age distribution. Because empirical studies in workers' job transition and the formation of new firms (univariate cases), or marriage market transition and merger activities (bivariate cases) have gained much attention recently, we hope to offer a new technology for analyzing data so that more appropriate structure can be put to advance our understanding in interesting phenomena. Selectivity issues relating to match formation are outside the scope in this paper and is left for future work. The model can also be extended to include multi-dimensional traits, covariates, or parametric and semi-parametric analyses. We leave such topics for future work.

Acknowledgements The authors thank the Editor and two referees for helpful comments, and Ms. Cuixian Chen for pointing out several typos.

References

[1] C. Goldin, L.F. Katz, The Power of the Pill: Oral Contraceptives and Women's Career and Marriage Decisions. *J. of Political Economy*, 110 (2002) 730-70.

[2] D. Loughran, The Effect of Male Wage Inequality on Female Age at First Marriage. *Review of Economics and Statistics*, 84 (2002) 237-50.

[3] D. Bloom, N. Bennett, Marriage Patterns in the United States. *Harvard Institute for Economic Research Discussion Paper, 1147* (1985).

[4] M.C. Keeley, An Analysis of the Age Pattern of First Marriage. *International Economic Review*, 2 (1979) 527-44.

[5] A. Aassve, S. Burgess, A. Chesher, C. Propper, Transitions from Home to Marriage of Young Americans. *J. of Applied Econometrics*, 17 (2002) 1-23.

[6] A. Berrington, I. Diamond, Marriage or Cohabitation: A Competing Risks Analysis of First-Partnership Formation among the 1958 British Birth Cohort. *J. of Roy. Statist. Soc. Ser.* A. 163 (2000) 127-51.

[7] L.Y. Wong, On Estimation of a Two-Sided Matching Model. *Contributions to Economic Analysis: Structural Models of Wage and Employment Dynamics* edited by Bunzel, H., B.J. Christensen, G. Neumann, and J-M Robin. Published by Elsevier ScienceNorth-Holland. Forthcoming (2006)

[8] L.Y. Wong, Structural Estimation of Marriage Models. *J. of Labor Economics* 21 (2003) 699-728.

[9] P. Groeneboom, J.A. Wellner, Information bounds and non-parametric maximum likelihood estimation. *Birkhäuser Verlag, Basel.* (1992).

[10] A. Schick, Q.Q. Yu, Consistency of the GMLE with mixed case interval-censored data.

Scan. J. of Statist. 27 (2000)45-55.

[11] Q.Q. Yu, G.Y.C. Wong, L.X. Li, Asymptotic properties of self-consistent estimators with mixed interval-censored data. *Ann. Inst. Statist. Math.*, 53 (2001) 469-486.

[12] G.Y.C. Wong, Q.Q. Yu, Generalized MLE Of a joint distribution function with multivariate interval-censored data. *J. of Multi. Anal.*, 69 (1999) 155-166.

[13] R.A. Betensky, D.M. Finkelstein, A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, 18 (1999) 3089-3100.

[14] A. van der Vaart, J.A. Wellner, Preservation theorems for Glivenko-Cantelli and Uniform Glivenko-Cantelli Class. *High Dimensional probability II.* E. Gine, D.M. Mason and J.A. Wellner, editors. Birkhäuser Boston. 115-133 (2000).

[15] S.H. Yu, Q.Q. Yu, G.Y.C. Wong, Consistency of the generalized MLE with multivariate mixed case interval-censored data. *Journal of Multivariate Analysis*, (accepted) (2005).

[16] B.W. Turnbull, The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B. 38 (1976) 290-295.

[17] E. Gould, Marriage and Career: The Dynamic Decisions of Young Men. Hebrew University (unpublished paper) (2003).

[18] J. Kiefer, J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* 27 (1956) 887-906.

[19] J.A. Wellner, Y. Zhan, A hybrid algorithm for computation of the NPMLE from censored data. *JASA*, 92 (1997) 945-959.

[20] Q.Q. Yu, G.Y.C. Wong, Q.M. He, Estimation of a joint distribution function with multivariate interval-censored data when the non-parametric MLE is not unique. *Biometrical Journal*, 42 (2000) 747-763.

[21] L.Y.Y. Wong, Q.Q. Yu, Technical report to "A bivariate interval censorship model for partnership formation". http://math.binghamton.edu/qyu (2006).

Table 1. A sample of the Data Set				
L_1	R_1	L_2	R_2	N_i
0	100*	41	100	7
0	19	0	19	1
0	100	0	21	8
0	17	0	100	1
0	17	16	16	1
0	19	0	16	1
0	20	13	13	1
0	100	0	17	1
0	100	39	39	2
17	17	0	31	1
18	20	0	100	1
40	40	0	100	1
41	100	0	100	15
31	31	0	14	1
0	100	21	23	1
0	25	24	24	4
0	20	18	18	18
0	20	0	100	5
0	34	24	24	1
0	37	29	29	1
0	100	22	24	1
30	30	0	27	1
24	24	0	16	1
0	33	24	24	1
18	18	0	20	2
0	20	30	30	1
0	16	14	14	1
0	100	22	30	1
0	27	28	28	2
0	27	18	18	3
29	29	0	15	1
21	21	0	30	1
0	100	18	19	9
0	30	19	19	3

* 100 is in place of ∞ to illustrate right censoring and $(0, 100)$ represents a missing observation.

ages	13	14	16	18	19	23	24	28	29	30	39
16	0.938	0.921	0.894	0.578	0.347	0.319	0.229	0.203	0.189	0.176	0.137
	0.028	0.032	0.036	0.063	0.067	0.065	0.059	0.057	0.056	0.055	0.049
17	0.745	0.728	0.724	0.504	0.289	0.262	0.205	0.188	0.181	0.176	0.137
	0.118	0.119	0.119	0.111	0.109	0.107	0.088	0.071	0.059	0.055	0.049
18	0.550	0.533	0.528	0.470	0.289	0.262	0.205	0.188	0.181	0.176	0.137
	0.163	0.164	0.164	0.131	0.109	0.107	0.088	0.071	0.059	0.055	0.049
19	0.449	0.431	0.427	0.427	0.256	0.229	0.191	0.181	0.176	0.176	0.137
	0.086	0.086	0.086	0.086	0.078	0.075	0.065	0.058	0.055	0.055	0.049
21	0.396	0.379	0.374	0.374	0.203	0.176	0.176	0.176	0.176	0.176	0.137
	0.074	0.074	0.074	0.074	0.059	0.055	0.055	0.055	0.055	0.055	0.049
24	0.387	0.374	0.374	0.374	0.203	0.176	0.176	0.176	0.176	0.176	0.137
	0.074	0.074	0.074	0.074	0.059	0.055	0.055	0.055	0.055	0.055	0.049
29	0.381	0.374	0.374	0.374	0.203	0.176	0.176	0.176	0.176	0.176	0.137
	0.074	0.074	0.074	0.074	0.059	0.055	0.055	0.055	0.055	0.055	0.049
30	0.333	0.327	0.327	0.327	0.203	0.176	0.176	0.176	0.176	0.176	0.137
	0.062	0.062	0.062	0.062	0.059	0.055	0.055	0.055	0.055	0.055	0.049
31	0.327	0.327	0.327	0.327	0.203	0.176	0.176	0.176	0.176	0.176	0.137
	0.062	0.062	0.062	0.062	0.059	0.055	0.055	0.055	0.055	0.055	0.049
40	0.307	0.307	0.307	0.307	0.192	0.166	0.166	0.166	0.166	0.166	0.128
	0.061	0.061	0.061	0.061	0.044	0.042	0.042	0.042	0.042	0.042	0.042

Table 2. The GMLE of the survival function

Proofs

We shall establish consistency and asymptotic normality of the GMLE studied in Wong and Yu's paper "A BIVARIATE INTERVAL CENSORSHIP MODEL FOR PARTNERSHIP FORMATION" in this part.

A.1. Consistency of the GMLE

By assumption A3, τ_1 and τ_2 are both finite. Under the finite discrete assumption, there are finitely many possible values of $\mathcal{I}$, say M of them. WLOG, we can assume that the first M $\mathcal{I}_i$'s are all the distinct possible values.

Theorem 1. *Under assumptions A1 and A3,* $\mu_{\hat{F}_n}(\mathcal{I}_i) \to \mu_{F_0}(\mathcal{I}_i)$ *a.s. for each* i.

Proof of Theorem 1. Let $N_i = \sum_{j=1}^{n} \mathbf{1}(\mathcal{I}_j = \mathcal{I}_i)$. We shall first show that

$$P\{\mathcal{I} = \mathcal{I}_i\} = \mu_{F_0}(\mathcal{I}_i) g(\mathcal{I}_i), \text{ where } g \text{ does not depends on } F_0.$$

Here, abusing notation, we treat $\mathcal{I}_i$ as a fixed value, instead of a random rectangle. Let f_0 be the density function of (X_1, X_2). In particular, by (2.1) there are 3 types of observations $\mathcal{I}_i$. The first type is of form $\mathcal{I}_i = [t,t] \times (a,b]$, where t is a positive integer, a and b are non-negative integers or ∞,

$$\begin{aligned}
&P\{\mathcal{I} = \mathcal{I}_i\} \\
=&\sum_{u>a}^{b} f_0(t,u) P\{K_1 = 0, Y_{1,K_1,0} \geq t, K_2 \geq 1, Y_{2,K_2,j-1} < u \leq Y_{2,K_2,j}, j \in \{1, ..., K_2+1\}\} \\
=&\mu_{F_0}(\mathcal{I}_i) P\{K_1 = 0, Y_{1,K_1,0} \geq t, K_2 \geq 1, Y_{2,K_2,j-1} = a, Y_{2,K_2,j} = b, j \in \{1, ..., K_2+1\}\}
\end{aligned}$$

Thus, $g(\mathcal{I}_i) = P\{K_1 = 0, Y_{1,K_1,0} \geq t, K_2 \geq 1, Y_{2,K_2,j-1} = a, Y_{2,K_2,j} = b, j \in \{1, ..., K_2+1\}\}$.

The second type is of the form $\mathcal{I}_i = (a_1, b_1] \times (a_2, b_2]$. It can be derived in a similar manner as for the first type that

$$g(\mathcal{I}_i) = P\{K_i \geq 1, Y_{i,K_i,j_i-1} = a_i, Y_{i,K_i,j_i} = b_i, j_i \in \{1, ..., K_i+1\}, i = 1 \text{ and } 2\}.$$

The third type is of the form $\mathcal{I}_i = (a, b] \times [t, t]$ with

$$g(\mathcal{I}_i) = P\{K_2 = 0, Y_{2,K_2,0} \geq t, K_1 \geq 1, Y_{1,K_1,j-1} = a, Y_{1,K_1,j} = b, j \in \{1, ..., K_1 + 1\}\}.$$

The normalized log likelihood function is

$$\mathcal{L}_n(F) = \frac{1}{n} \sum_{i=1}^{M} N_i \ln \mu_F(\mathcal{I}_i).$$

Let $Ł(F) := E(\mathcal{L}_n(F))$. Then

$$\begin{aligned} Ł(F) &= \sum_{i=1}^{M} \mu_{F_0}(\mathcal{I}_i) g(\mathcal{I}_i) \ln \mu_F(\mathcal{I}_i) \\ &= \sum_{i=1}^{M} \mu_{F_0}(\mathcal{I}_i) g(\mathcal{I}_i) \ln\{\mu_F(\mathcal{I}_i) g(\mathcal{I}_i)\} - \sum_{i=1}^{M} \mu_{F_0}(\mathcal{I}_i) g(\mathcal{I}_i) \ln g(\mathcal{I}_i). \end{aligned} \tag{6.1}$$

It is important to notice that the second summand in (6.1) does not depend on F. Since

$$\sum_{i=1}^{M} \mu_{F_0}(\mathcal{I}_i) g(\mathcal{I}_i) \ln\{\mu_{F_0}(\mathcal{I}_i) g(\mathcal{I}_i)\} \text{ is finite,}$$

by the Shannon-Kolmogorov inequality and (6.1), we have

$$Ł(F) < Ł(F_0) \text{ unless } \mu_F(\mathcal{I}_i) = \mu_{F_0}(\mathcal{I}_i) \text{ for } i = 1, ..., m. \tag{6.2}$$

By the strong law of large numbers (SLLN),

$$\begin{aligned} \underline{\lim}_{n\to\infty} \mathcal{L}_n(\hat{F}_n) &\geq \underline{\lim}_{n\to\infty} \mathcal{L}_n(F_0) \text{ (as } \hat{F}_n \text{ maximizes } \mathcal{L}_n(\cdot)) \\ &= \underline{\lim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i \text{ (where } W_i = \ln \mu_{F_o}(\mathcal{I}_i)) \\ &= E(W_1) \text{ (by SLLN, as } W_i\text{s are i.i.d.)} \\ &= Ł(F_0) \text{ almost surely.} \end{aligned} \tag{6.3}$$

Let Ω' denote the event on which $\underline{\lim}_{n\to\infty} \mathcal{L}_n(\hat{F}_n) \geq Ł(F_0)$. Fix an $\omega \in \Omega'$, let F^* be a limit point of $\hat{F}_n(\cdot,\omega)$ in the sense that $\mu_{\hat{F}_{k_n}}(\mathcal{I}_i) \to \mu_{\hat{F}^*}(\mathcal{I}_i)$ for each i and for some subsequence $\{k_n\}$ of positive integers tending to infinity. Since $\ln\mu_F(\mathcal{I}_i) < 0$, it follows that

$$\begin{aligned} Ł(F_0) \leq \underline{\lim}_{n\to\infty} \mathcal{L}_n(F) &= \underline{\lim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{M} N_i \ln\mu_F(\mathcal{I}_i) && \text{(by (6.3))} \\ &\leq \sum_{i=1}^{M} \mu_{F_0}(\mathcal{I}_i) g(\mathcal{I}_i) \ln\mu_{F^*}(\mathcal{I}_i) && \text{(by Fatou's lemma)} \\ &= Ł(F^*). \end{aligned}$$

That is, $Ł(F^*) \geq Ł(F_0)$. By the foregoing inequality and (6.2), we conclude that $Ł(F^*) = Ł(F_0)$ and consequently $\mu_{F^*}(\mathcal{I}_i) = \mu_{F_0}(\mathcal{I}_i)$ for all i. Since ω is arbitrary in Ω', F^* is an arbitrary limiting point of $\hat{F}_n$ and Ω' has probability one, the theorem is proved. □

Theorem 2. *Under assumptions A1, A2 and A3 the GMLE $\hat{F}_n(x,y)$ is strongly consistent at each $(x,y) \in \{x \leq \tau_1, y \leq \tau_2\}$.*

Proof of Theorem 2. It can be shown that under assumptions A1, A2 and A3 $P\{X_1 \text{ is uncensored} | X_1 = t \leq \tau_1\} > 0$ and $P\{X_2 \text{ is uncensored} | X_2 = t \leq \tau_2\} > 0$.

By Theorem 1 $\mu_{\hat{F}_n}(\mathcal{I}_i) \to \mu_{F_0}(\mathcal{I}_i)$ a.s. for each $\mathcal{I}_i$ of form $\mathcal{I}_i = [t,t] \times (0,c]$. Given $(x,y) \in \{x \leq \tau_1, y \leq \tau_2\}$, $\hat{F}_n(x,y) = \sum_{t\leq x} \mu_{\hat{F}_n}([t,t] \times (0,y])$. There are only finitely many summands in the foregoing summation by assumption A3 Thus $\hat{F}_n(x,y) \to F_0(x,y)$ a.s. □

A.2. Asymptotic normality of the GMLE.

For the marriage study, it is reasonable to assume that

A4. $P\{X_1 > \tau_1, X_2 > \tau_2\} > 0$, $P\{X_1 < \tau_{01} \text{ or } X_2 < \tau_{02}\} = 0$, $P\{(X_1,X_2) = (x,y)\} > 0$, $P\{X_1 = x, X_2 > \tau_2\} > 0$ and $P\{X_1 > \tau_1, X_2 = y\} > 0$, for each $x \in [\tau_{01}, \tau_1]$ and $y \in [\tau_{02}, \tau_2]$.

A4 says that there are still men (or women) having their first marriage beyond the age of τ_1 (or τ_2). Under assumptions A1, A2, A3and A4, the problem becomes an estimation

problem of multinomial distribution and it follows from standard argument that the GMLE $\hat{F}_n$ is asymptotically normally distributed and the convergence rate is in $\sqrt{n}$.

Denote $\mathcal{A}$ the collection of the subsets of the forms $(\tau_1, \infty] \times (\tau_2, \infty]$, $[x, x] \times (\tau_2, \infty]$, $(\tau_1, \infty] \times [y, y]$, $[x, x] \times [y, y]$, where $x \in [\tau_{01}, \tau_1]$ and $y \in [\tau_{02}, \tau_2]$. It can be shown that if n is large enough, each subset in $\mathcal{A}$ is an MI. In fact, $[x, x] \times [y, y]$ is the intersection of observable rectangles $[x, x] \times (0, \infty]$ and $(0, \infty] \times [y, y]$; $[x, x] \times (\tau_2, \infty]$ is the intersection of $[x, x] \times (0, \infty]$ and $(0, \infty] \times (\tau_2, \infty]$; $(\tau_1, \infty] \times [y, y]$ is the intersection of $(0, \infty] \times [y, y]$ and $(\tau_1, \infty] \times (0, \infty]$; $(\tau_1, \infty] \times (\tau_2, \infty]$ is the intersection of $(0, \infty] \times (\tau_2, \infty]$ and $(\tau_1, \infty] \times (0, \infty]$.

Let $s_j^o = \mu_{F_0}(A_j)$ for $A_j \in \mathcal{A}$. Then by assumptions A3 and A4, $s_j^o > 0$ for all j. Verify that

$$Ł(F) = \sum_{h=1}^{M} g(\mathcal{I}_h) \sum_{j=1}^{m} s_j^o \mathbf{1}(A_j \subset \mathcal{I}_h) \ln \sum_j s_j \mathbf{1}(A_j \subset \mathcal{I}_h). \tag{7.1}$$

Let

$$p_h = g(\mathcal{I}_h) \sum_{j=1}^{m} s_j^o \mathbf{1}(A_j \subset \mathcal{I}_h).$$

We can rewrite (7.1) as

$$Ł(F) = \sum_{h=1}^{M} p_h \ln \sum_{j=1}^{m} s_j \mathbf{1}(A_j \subset \mathcal{I}_h) = \sum_{h=1}^{M} p_h \ln \sum_{j=1}^{m} s_j \delta_{hj}.$$

By assumptions A3 and A4, $p_h > 0$, $h = 1, ..., M$. Set $J = -E(\frac{\partial^2 \mathcal{L}_n(F_0)}{\partial \mathbf{S} \partial \mathbf{S}^t})$, where $\frac{\partial \mathcal{L}_n}{\partial \mathbf{S}}$ is an $(m-1) \times 1$ vector and $\frac{\partial^2 \mathcal{L}_n}{\partial \mathbf{S} \partial \mathbf{S}^t}$ is an $(m-1) \times (m-1)$ matrix. Verify that

$$J = nE(\frac{\partial \mathcal{L}_n(F_0)}{\partial \mathbf{s}} \frac{\partial \mathcal{L}_n(F_0)}{\partial \mathbf{s}^t}) = -\frac{\partial^2 Ł}{\partial \mathbf{s} \partial \mathbf{s}^t} = (\sum_{h=1}^{M} p_h \frac{(\delta_{hi} - \delta_{hm})(\delta_{hj} - \delta_{hm})}{(\sum_{k=1}^{m} \delta_{hk} s_k^o)^2})_{(m-1)\times(m-1)}$$

and $J = UU^t$, where $U = \begin{pmatrix} \frac{(\delta_{11} - \delta_{1m})\sqrt{p_1}}{\sum_{k=1}^{m} \delta_{1k} s_k^o} & \cdots & \frac{(\delta_{M1} - \delta_{Mm})\sqrt{p_M}}{\sum_{k=1}^{m} \delta_{Mk} s_k^o} \\ \vdots & \cdots & \vdots \\ \frac{(\delta_{1(m-1)} - \delta_{1m})\sqrt{p_1}}{\sum_{k=1}^{m} \delta_{1k} s_k^o} & \cdots & \frac{(\delta_{M(m-1)} - \delta_{Mm})\sqrt{p_M}}{\sum_{k=1}^{m} \delta_{Mk} s_k^o} \end{pmatrix}$.

We now show that J is nonsingular. Let $\mathbf{x}_j$ be the upper-right vertex of A_j, $j = 1, ..., m-1$. By reordering the $\mathcal{I}_j$'s, WLOG, we can assume that the upper-right vertex of $\mathcal{I}_i$ is equal to $\mathbf{x}_i$, $i = 1, ..., m-1$. Thus $\mathcal{I}_i \cap A_j = \emptyset$ for $j > i$, $i = 1, ..., m-1$. Then the matrix U has the upper triangle matrix form

$$U = \begin{pmatrix} \frac{\sqrt{p_1}}{s_1^o} & \cdot & \cdots & \cdot & \cdots & \frac{(\delta_{M1}-\delta_{Mm})\sqrt{p_M}}{\sum_{k=1}^m \delta_{Mk} s_k^o} \\ 0 & \frac{\sqrt{p_2}}{s_2^o+\delta_{21}s_1^o} & \cdots & \cdot & \cdots & \frac{(\delta_{M2}-\delta_{Mm})\sqrt{p_M}}{\sum_{k=1}^m \delta_{Mk} s_k^o} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\sqrt{p_{m-1}}}{s_{m-1}^o+\sum_{k=1}^{m-2} \delta_{(m-1)k} s_k^o} & \cdots & \frac{(\delta_{M(m-1)}-\delta_{Mm})\sqrt{p_M}}{\sum_{k=1}^m \delta_{Mk} s_k^o} \end{pmatrix}.$$

Recall $s_i^o > 0$ and $p_i > 0$ for $i = 1, ..., m-1$. It follows that the matrix U is of full rank and $J = UU^t$ is nonsingular.

It is easy to verify that

$$\frac{\partial^2 \mathcal{L}_n(\hat{F}_n)}{\partial \mathbf{s} \partial \mathbf{s}^t} \to E(\frac{\partial^2 \mathcal{L}_n(F_0)}{\partial \mathbf{s} \partial \mathbf{s}^t}) = -J.$$

It thus follows that

$$\frac{\partial \mathcal{L}_n(\hat{F}_n)}{\partial \mathbf{s}} = \frac{\partial \mathcal{L}_n(F_0)}{\partial \mathbf{s}} - J\Delta_n + o_p(\|\Delta_n\|),$$

where Δ_n is the $(m-1)$-dimensional column vector with entries $\hat{s}_i - s_i^o = \mu_{\hat{F}_n}(A_i) - \mu_{F_0}(A_i)$, $i = 1, \ldots, m-1$. Let $\Omega_n = \{\inf_{i \le m} \hat{s}_i = 0\}$. Verify that

$$0 = \frac{\partial \mathcal{L}_n(\hat{F}_n)}{\partial \mathbf{s}} \text{ except on the event } \Omega_n,$$

and by Theorem 1 and assumptions A1, A2, A3 and A4, $\mathrm{P}(\Omega_n) \to 0$ as $n \to \infty$. It follows from the central limit theorem that $\sqrt{n}\frac{\partial \mathcal{L}_n(F_0)}{\partial \mathbf{s}}$ is asymptotically normal with mean 0 and dispersion matrix J. This shows that $\Delta_n = J^{-1}\frac{\partial \mathcal{L}_n(F_0)}{\partial \mathbf{s}} + o_p(n^{-1/2})$. Thus we have the following result.

Theorem 3. *Under assumptions A1, A2, A3 and A4,* $\sqrt{n}\begin{pmatrix} \hat{s}_1 - s_1^o \\ \vdots \\ \hat{s}_{m-1} - s_{m-1}^o \end{pmatrix}$ *is asymptotically normal with mean 0 and dispersion matrix* J^{-1}. *A strongly consistent estimator of*

J is given by $\hat{J} = -\frac{\partial^2 \mathcal{L}_n(\hat{F}_n)}{\partial \mathbf{S} \partial \mathbf{S}^t}$. Furthermore, $\sqrt{n}[\hat{F}_n(\mathbf{x}) - F_0(\mathbf{x})]$ is asymptotically normally distributed for all $\mathbf{x} \in \mathcal{A}$. A consistent estimate of the asymptotic variance of $\hat{F}_n(\mathbf{x})$ is $\frac{1}{n}\mathbf{c}^t \hat{J}^{-1}\mathbf{c}$, where $\mathbf{c}$ is a $(m-1) \times 1$ vector with the i-th entry $c_i = \mathbf{1}(A_i \subset [0, x_1] \times [0, x_2])$ unless $F_0(\mathbf{x}) = 1$.

Under the assumptions in Theorem 3, the GMLE $\hat{F}_n$ is also asymptotically efficient. The proof of this assertion is straightforward and is omitted.