Statistica Sinica **11**(2001), 1047-1068

NONPARAMETRIC ESTIMATION OF MARGINAL DISTRIBUTIONS UNDER BIVARIATE TRUNCATION WITH APPLICATION TO TESTING FOR AGE-OF-ONSET ANTICIPATION

Jian Huang, Veronica J. Vieland and Kai Wang

University of Iowa

Abstract: Bivariate truncated data arise from the study of age-of-onset anticipation for diseases with variable age of onset in which children tend to develop clinical disease at younger ages than their affected parents. To test for age-of-onset anticipation using affected parent-child pair data, it is of interest to estimate the marginal distributions of the age-of-onset for both parents and children. However, the observed ages of onset in both parents and children are right-truncated by their current ages. In this report, we proposed a nonparametric estimator of the marginal distributions of a bivariate distribution based on right-truncated data. This estimator is shown to be consistent under appropriate conditions. A noniterative algorithm is given to compute the proposed estimator. Finite sample behavior of the estimator is investigated via simulation. An example is given to illustrate the use of the proposed estimator in testing of age-of-onset anticipation in bipolar affective disorder.

Key words and phrases: Age of onset, anticipation, bivariate data, consistency, marginal distribution, truncation.

1. Introduction

Bivariate truncated data arise from the study of genetic diseases with variable age of onset, when there is a tendency for children to develop clinical disease at younger ages than their affected parents (age-of-onset anticipation or AOA). It is now confirmed in several neurologic diseases, including myotonic dystrophy (Höweler, Busch, Geraedts, Niermeijer and Staall (1989)), Huntington's disease (The Hungtinton's Disease Collaborative Research Group (1993)), and Machado-Joseph disease (DeStefano, Cupples, Maciel, Gaspar et al. (1996)), that age of onset is negatively correlated with the length of unstable expanding DNA-triplet repeats at the disease locus. The possibility of mapping a gene for a complex disorder by screening the genome for large repeat polymorphisms (Schalling (1993)) adds incentive for clinical investigators to assess the possibility of AOA in the disorders they study.

Reports of AOA now exist in bipolar disorder (McInnis, McMahon, Stine and Ross (1993)), facioscapulohumeral muscular dystrophy (Zatz et al. (1995)), schizophrenia (Bassett and Honer (1994)), rheumatoid arthritis (Deighton, Heslop, McDonagh, Walker and Thomson (1994)), hereditary dentatorubral-pallidoluysian atrophy (Sano et al. (1994)). Most recently, Paterson, Kennedy and Petronis (1996) found strong statistical evidence of AOA for breast cancer, colon cancer, Alzheimer's disease, and maturity-onset diabetes mellitus. On the face of it, the high "hit" rate for detecting AOA among these diverse diseases is grounds for suspicion, and a careful scrutiny of the statistical methods employed supports this skepticism.

The particular statistical methods used in testing for AOA across these various reports have varied. A simple and frequently used procedure is simply to carry out a paired t-test of the hypothesis that the mean age of onset in parents is higher than the mean age of onset in children, using affected parent-child pairs (see e.g., Horwitz, Goode and Jarvik (1996), Paterson et al. (1996), Zatz et al. (1995), Myers et al. (1985)). Heiman, Hodge, Wickramaratne and Hsu (1996) showed that this test is inappropriate, for the following simple reason. In this set up, both the parent and the child in each pair are affected, and therefore, whatever the current age of each may be, the age of onset must be prior to the current age. Therefore the age of onset distribution in parents and children, respectively, is right truncated relative to the population distribution: only individuals with age of onset prior to current age are eligible for inclusion. Since children are younger than parents, the truncation effect is more pronounced in the children than in the parents. Failure to take into account the truncation effect can lead to a striking deflation of p-values. Heiman et al. (1996) showed, using simulations, that the propensity of the ordinary paired t-test to reject the hypothesis of no AOA when in fact it is true (i.e., when there is in fact no AOA) can be extremely high, depending on the underlying age of onset distribution. For some of the models they considered, (false) rejection rates could be as high as 100%.

The truncation effect can be described using a bivariate right-truncation model. Huang and Vieland (1997) proposed a semiparametric model and a conditional likelihood approach to compare the age-of-onset distributions under bivariate truncation. In that model, the distribution of the ages at interview need not be specified since the conditional likelihood (conditioning on the ages at interview) does not involve this distribution. The joint distribution of ages of onset in parents and children is assumed to be bivariate normal. In this paper, we propose a nonparametric estimator of the joint distribution based on truncated bivariate data, without specifying the form of the age-of-onset or the current age distribution.

The univariate truncation problem has been studied by many authors, see for example Woodroofe (1985), and Keiding and Gill (1990). In the univariate setting, the nonparametric maximum likelihood estimator has an explicit form using a product integral similar to the Kaplan-Meier estimator for right censored data (Kaplan and Meier (1965)). Estimation of a bivariate distribution when observations are subject to right censoring has received much attention recently. See, for example, Campbell (1981), Dabrowska (1988), Prentice and Cai (1993), Lin and Ying (1993), Gill, Van der Laan and Wellner (1995), Van der Laan (1996), and the references cited by these authors. There has also been some recent work on estimating a bivariate distribution when observations are subject to truncation. For example, Gürler (1996) considered bivariate estimation when a single component of the bivariate data is subject to truncation. The method of Gürler (1996) cannot be extended to the case where both components are truncated. Van der Laan (1996) studied the self-consistency estimator of a bivariate survival function when both components are subject to truncation. However, he did not consider the finite sample behavior of the self-consistency estimator. Our simulations show that simple use of the marginal distributions of the bivariate self-consistency estimator results in severely upward biased estimators of the underlying marginal distributions. This large bias makes the marginal distribution estimators directly based on the self-consistency estimator unsuitable for comparing the distributions in our example of testing age-of-onset anticipation.

In this paper, we propose a nonparametric estimator of the marginal distribution that does not have the large upward bias. In Sections 2 and 3 below, we first describe the bivariate truncation model and the proposed nonparametric estimators of the two marginal distributions. In Section 4, a noniterative algorithm is given for computing the estimator. We also present some simulation results. To illustrate an application of the proposed estimator, we propose a test for the equality of two marginal medians based on the estimated medians obtained from the the bivariate estimator. This test provides a nonparametric approach to testing for age-of-onset anticipation. Section 5 contains sufficient conditions for the consistency of the proposed estimator. The proofs are put together in the appendices. Discussions and further problems are included in Section 6.

2. The Bivariate Right Truncation Model

Let T_1 and T_2 be the parent's and child's age of onset, respectively. Let C_1 and C_2 be the parent's and child's age at interview (current age), respectively. Let the joint distribution functions of (T_1, T_2) and (C_1, C_2) be F and G, respectively. We assume that the pairs (T_1, T_2) and (C_1, C_2) are independent of each other in the population.

For an affected parent-child pair to be included in the sample, their disorders have to be manifested before the time they are examined. Therefore $T_1 \leq C_1$

and $T_2 \leq C_2$ and the joint distribution of the observed age of onset and the age at interview is given by

$$P(T_1 \leq t_1, T_2 \leq t_2, C_1 \leq c_1, C_2 \leq c_2 | T_1 \leq C_1, T_2 \leq C_2). \tag{2.1}$$

3. A Nonparametric Estimator

In this section, we propose a nonparametric estimator of F based on truncated bivariate data. First we construct an initial estimator. This estimator is shown to be consistent under appropriate conditions, see Section 5. However, the finite sample behavior of its corresponding marginal distribution estimators is not satisfactory. We update this initial estimator based on an equation derived from the truncation model. The resulting estimator has much better finite sample performance according to our simulations. It is also shown to be consistent.

3.1. The self-consistency estimator of the bivariate distribution

Let $t_1 \geq 0$ and $t_2 \geq 0$ be fixed. Let $K^*(t_1,t_2) = P(T_1 \leq t_1 < C_1, T_2 \leq t_2 < C_2 | T_1 \leq C_1, T_2 \leq C_2)$. This function plays an important role in the construction of our estimator, the motivation for using it is given in Appendix A. Let $\alpha = P(T_1 \leq C_1, T_2 \leq C_2)$. By the independence assumption between (T_1,T_2) and (C_1,C_2) in the population, we have

$$\begin{aligned} K^*(t_1,t_2) &= \alpha^{-1} P(T_1 \leq t_1 < C_1, T_2 \leq t_2 < C_2) \\ &= \alpha^{-1} F(t_1,t_2) \int_{(t_1,\infty)} \int_{(t_2,\infty)} dG(c_1,c_2). \end{aligned} \tag{3.2}$$

Let G^* be the distribution function of the observable ages at interview, given by

$$\begin{aligned} G^*(t_1,t_2) &= P(C_1 \leq t_1, C_2 \leq t_2 | T_1 \leq C_1, T_2 \leq C_2) \\ &= \alpha^{-1} \int_{[0,t_1]} \int_{[0,t_2]} F(c_1,c_2) dG(c_1,c_2). \end{aligned} \tag{3.3}$$

Combining equations (3.2) and (3.3) gives

$$K^*(t_1,t_2) = F(t_1,t_2) \int_{(t_1,\infty)} \int_{(t_2,\infty)} \frac{dG^*(c_1,c_2)}{F(c_1,c_2)}. \tag{3.4}$$

It is shown in the appendix that, if the support of F is contained in the support of G (this is more precisely defined in Section 5) plus mild extra conditions, a distribution function F_* satisfies (3.4) if and only if $F_* \equiv F$. This implies that F can be identified through equation (3.4). Note that even in the univariate case, a similar condition for the identifiability of F is needed, see Woodroofe (1985).

Let $(T_{1i}, T_{2i}, C_{1i}, C_{2i}), 1 \leq i \leq n$, be the observed data subject to the condition that $T_{1i} \leq C_{1i}$ and $T_{2i} \leq C_{2i}$. The bivariate function K^* can be estimated directly by its corresponding empirical function

$$K_n^*(t_1, t_2) = n^{-1} \sum_{i=1}^{n} 1_{[T_{1i} \leq t_1 < C_{1i}, T_{2i} \leq t_2 < C_{2i}]}. \tag{3.5}$$

In addition, the empirical version of G^* is simply the empirical distribution function of the observed ages at interview: $G_n^*(t_1, t_2) = n^{-1} \sum_{i=1}^{n} 1_{[C_{1i} \leq t_1, C_{2i} \leq t_2]}$.

It is natural to define a nonparametric estimator $F_n^{(0)}$ which simply solves the empirical version of (3.4):

$$K_n^*(t_1, t_2) = F_n^{(0)}(t_1, t_2) \int_{(t_1,\infty)} \int_{(t_2,\infty)} \frac{dG_n^*(c_1, c_2)}{F_n^{(0)}(c_1, c_2)}. \tag{3.6}$$

This is exactly the equation that defines the self-consistency estimator of F and can also be derived as the score equation based on the nonparametric likelihood of the data, see Van der Laan (1996).

However, (3.6) does not have a unique solution, since if $F_n^{(0)}$ satisfies (3.6), so does $kF_n^{(0)}$ for a constant k. At the population level, the identifiability of F through (3.4) is partly because F equals 1 at a point in the support of G (with the assumption that the support of F is contained in the support of G), so that k must be 1. See the proofs in Appendix B. Therefore we specify $F_n^{(0)}$ to be 1 at the point $(C_{1,n+1}, C_{2,n+1}) \equiv (\max_{1\leq i\leq n}\{C_{1i}\} + (1/n), \max_{1\leq i\leq n}\{C_{2i}\} + (1/n))$:

$$F_n^{(0)}(C_{1,n+1}, C_{2,n+1}) = 1. \tag{3.7}$$

We also include this point in a slightly different version of the empirical distribution function corresponding to G^*, given by

$$G_n^{\#}(t_1, t_2) = (n+1)^{-1} \sum_{i=1}^{n+1} 1_{[C_{1i} \leq t_1,\, C_{2i} \leq t_2]}.$$

For any $t_1 < C_{1,n+1}$ and $t_2 < C_{2,n+1}$, we define an initial estimator $F_n^{(0)}(t_1, t_2)$ to be the solution to

$$\begin{aligned} K_n^*(t_1, t_2) &= F_n^{(0)}(t_1, t_2) \int_{(t_1,\infty)} \int_{(t_2,\infty)} \frac{dG_n^{\#}(c_1, c_2)}{F_n^{(0)}(c_1, c_2)} \\ &= \frac{n}{n+1} F_n^{(0)}(t_1, t_2) \int_{(t_1,\infty)} \int_{(t_2,\infty)} \frac{dG_n^*(c_1, c_2)}{F_n^{(0)}(c_1, c_2)} + \frac{1}{n+1} F_n^{(0)}(t_1, t_2), \end{aligned} \tag{3.8}$$

where the second equality in (3.8) follows from (3.7).

The choice of $(C_{1,n+1}, C_{2,n+1})$ is based on the following considerations: (a) for the identifiability of F, it is necessary to assume that the support of F is contained in the support of G (so this choice will not cause asymptotic bias in $F_n^{(0)}$); (b) the choice of $(C_{1,n+1}, C_{2,n+1})$ is not sensitive in the sense that any other choice of $(C_{1,n+1}, C_{2,n+1})$ such that $C_{1,n+1} > \max_{1\le i\le n}\{C_{1i}\}$ and $C_{2,n+1} > \max_{1\le i\le n}\{C_{2i}\}$ will result in the same estimation of $F_n^{(0)}$ at points (t_1, t_2) such that $t_1 \le \max_{1\le i\le n}\{C_{1i}\}$ and $t_2 \le \max_{1\le i\le n}\{C_{2i}\}$, with at least one strict inequality. This can be seen from the second equality of (3.8).

In the case of no truncation, (C_1, C_2) puts total mass at (∞, ∞), then K_n^* simplifies to the usual empirical distribution function of the $(T_{1i}, T_{2i}), i = 1, \ldots, n$, and the integral on the right side of (3.8) reduces to 1. Therefore, $F_n^{(0)}$ simplifies to the empirical distribution function of the complete data.

As shown in Section 5, $F_n^{(0)}$ enjoys the desired asymptotic properties, i.e., consistency and asymptotic normality under appropriate conditions. Our simulations suggest that it performs reasonably well in the interior of the data set for moderate sample sizes ($n = 50, 100$). Here the interior refers to the interior of the smallest polygon containing all the observed pairs (T_{1i}, T_{2i}) and $(C_{1i}, C_{2i}), i = 1, \ldots, n$, plotted in the (two-dimensional) first quadrant. However, simulation shows that the performance of $F_n^{(0)}$ on the boundary points, or at the points close to the boundary, is rather poor. In particular, $F_n^{(0)}(t_1, \infty)$ and $F_n^{(0)}(\infty, t_2)$ are severely biased upwards, and estimators of the marginal distributions of F based on $F_n^{(0)}$ will be severely biased. Therefore, we derive another expression for F given in (3.12) below. Based on this expression, we use $F_n^{(0)}$ as an initial estimator and obtain an improved estimator, particularly on the boundary points.

3.2. The proposed estimator

Let

$$\begin{aligned} K^+(t_1, t_2) &= P(T_1 \le t_1 \le C_1, T_2 \le t_2 \le C_2 | T_1 \le C_1, T_2 \le C_2) \\ &= \alpha^{-1} P(C_1 \ge t_1, C_2 \ge t_2) F(t_1, t_2). \end{aligned} \tag{3.9}$$

Note that K^+ is close to K^*. In fact, if the distribution function of (T_1, T_2, C_1, C_2) is continuous, then K^+ is the same as K^*. Let $S^*(t_1, t_2) = P(T_1 > t_1, T_2 > t_2 | T_1 \le C_1, T_2 \le C_2) = \alpha^{-1} \int_{(t_1,\infty)} \int_{(t_2,\infty)} P(C_1 \ge c_1, C_2 \ge c_2) dF(c_1, c_2)$. Then

$$dS^*(t_1, t_2) = \alpha^{-1} P(C_1 \ge t_1, C_2 \ge t_2) dF(t_1, t_2). \tag{3.10}$$

Dividing (3.10) by (3.9) gives

$$\frac{dS^*(t_1, t_2)}{K^+(t_1, t_2)} = \frac{dF(t_1, t_2)}{F(t_1, t_2)}. \tag{3.11}$$

It follows that

$$F(t_1,t_2)=\int_{[0,t_1]}\int_{[0,t_2]}\frac{F(s_1,s_2)}{K^+(s_1,s_2)}dS^*(s_1,s_2). \tag{3.12}$$

Let the empirical versions of K^+ and S^* be

$$K_n^+(t_1,t_2)=\frac{1}{n}\sum_{i=1}^n 1_{[T_{i1}\le t_1\le C_{1i},\,T_{i2}\le t_2\le C_{2i}]},\ \text{and}\ S_n^*(t_1,t_2)=\frac{1}{n}\sum_{i=1}^n 1_{[T_{1i}>t_1,\,T_{2i}>t_2]}.$$

Our proposed estimator of F is

$$F_n(t_1,t_2)=\int_{[0,t_1]}\int_{[0,t_2]}\frac{F_n^{(0)}(s_1,s_2)}{K_n^+(s_1,s_2)}dS_n^*(s_1,s_2). \tag{3.13}$$

If F_1 and F_2 are the marginal distribution functions of F, their estimators are

$$F_{1n}(t_1)=\int_{[0,t_1]}\int_{[0,\infty)}\frac{F_n^{(0)}(s_1,s_2)}{K_n^+(s_1,s_2)}dS_n^*(s_1,s_2) \tag{3.14}$$

and

$$F_{2n}(t_2)=\int_{[0,\infty)}\int_{[0,t_2]}\frac{F_n^{(0)}(s_1,s_2)}{K_n^+(s_1,s_2)}dS_n^*(s_1,s_2). \tag{3.15}$$

From (3.14) and (3.15), the influence of the points on or close to the boundary of the smallest polygon that contains all the data points on the marginals F_{1n} and F_{2n} is considerably smaller than that on the marginals based on $F_n^{(0)}$.

4. Computation and Examples

We describe an approach for computing the initial estimator $F_n^{(0)}$, the main task in computing the final estimator F_n and its marginals.

Since $K_n^*(t_1,t_2)$ is determined at the $2n$ points consisting of $(T_{1i},T_{2i}), i=1,\ldots,n$ and $(C_{1i},C_{2i}), i=1,\ldots,n$, the initial nonparametric estimator $F_n^{(0)}$ is determined by its values at these $2n$ points.

In view of (3.8), it is natural to use the following steps to solve (3.8) for $F_n^{(0)}$ inductively, since if the values of $F_n^{(0)}$ are known on points $\{(C_{1i},C_{2i}):C_{1i}>t_1,C_{2i}>t_2\}$, the value of $F_n^{(0)}$ at (t_1,t_2) can be solved via (3.8).

1. Define a new observation by

$$(T_{1,n+1},T_{2,n+1},C_{1,n+1},C_{2,n+1})=(0,0,\max_{1\le i\le n}\{C_{1i}\}+(1/n),\max_{1\le i\le n}\{C_{2i}\})+(1/n)).$$

   Add it to the data set as the $(n+1)$th observation. Assign values of $F_n^{(0)}$ on this new observation: $F_n^{(0)}(T_{1,n+1},T_{2,n+1})=0,\ F_n^{(0)}(C_{1,n+1},C_{2,n+1})=1.$

2. Calculate $K_n^*(t_1,t_2)$ for $(t_1,t_2)=(C_{1i},C_{2i})$ and $(T_{1i},T_{2i}), i=1,\ldots,n$, using (3.5).
3. Search for a point (t_1,t_2) in $\{(C_{1i},C_{2i}),(T_{1i},T_{2i}),i=1,\ldots,n\}$ such that all the values of $F_n^{(0)}$ are known in $\{(C_{1i},C_{2i}):C_{1i}>t_1,C_{2i}>t_2,1=1,\ldots,n+1\}$. Such point (t_1,t_2) always exists. This can be done as follows. Choose a point (t_1,t_2) in $\{(C_{1i},C_{2i}),(T_{1i},T_{2i}),i=1,\ldots,n\}$. If (t_1,t_2) does not satisfy the condition, there must be a point $(s_1,s_2)\in\{(C_{1i},C_{2i}):C_{1i}>t_1,C_{2i}>t_2,i=1,\ldots,n+1\}$ at which the value of $F_n^{(0)}$ is not known. Then (s_1,s_2) can be checked to see if it satisfies the condition. Since $F_n^{(0)}(C_{1,n+1},C_{2,n+1})=1$ is known, and $(C_{1,n+1},C_{2,n+1})\in\{(C_{1i},C_{2i}):C_{1i}>t_1,C_{2i}>t_2,i=1,\ldots,n+1\}$ for any point (t_1,t_2) in $\{(C_{1i},C_{2i}),(T_{1i},T_{2i}),i=1,\ldots,n\}$, the above process produces the required (t_1,t_2).

   The value of $F_n^{(0)}$ at (t_1,t_2) is then solved as

$$F_n^{(0)}(t_1,t_2)=\frac{(n+1)K_n^*(t_1,t_2)}{1+\sum_{\{(C_{1i},C_{2i}):C_{1i}>t_1,C_{2i}>t_2\}} n(F_n^{(0)}(C_{1i},C_{2i}))^{-1}}.$$

4. Repeat previous steps until the values of $F_n^{(0)}$ are known for all $2n$ points (C_{1i},C_{2i}) and $(T_{1i},T_{2i}), i=1,\ldots,n$.

Finally, F_n, F_{1n} and F_{2n} can be computed using (3.13), (3.14) and (3.15). These can be written as:

$$F_n(t_1,t_2)=\frac{1}{n}\sum_{T_{1i}\le t_1,\,T_{2i}\le t_2}\frac{F_n^{(0)}(T_{1i},T_{2i})}{K_n^+(T_{1i},T_{2i})}, \tag{4.16}$$

$$F_{1n}(t_1)=\frac{1}{n}\sum_{T_{1i}\le t_1,\,T_{2i}<\infty}\frac{F_n^{(0)}(T_{1i},T_{2i})}{K_n^+(T_{1i},T_{2i})}, \tag{4.17}$$

and

$$F_{2n}(t_2)=\frac{1}{n}\sum_{T_{1i}<\infty,\,T_{2i}\le t_2}\frac{F_n^{(0)}(T_{1i},T_{2i})}{K_n^+(T_{1i},T_{2i})}. \tag{4.18}$$

4.2. Simulation examples

We examine the performance of the proposed estimators in three bivariate survival distributions used by Prentice and Cai (1992). The three distributions for (T_1,T_2) are: (i) T_1 and T_2 are independent unit exponentials; (ii) (T_1,T_2) follows Gumbel's distribution: $F(t_1,t_2)=1-\{e^{-t_1}+e^{-t_2}-e^{-(t_1+t_2)}[1+(1-e^{-t_1})(1-e^{-t_2})]\}$; (iii) (T_1,T_2) follows Clayton-Oakes distribution: $F(t_1,t_2)=1-[e^{-t_1}+e^{-t_2}-(e^{4t_1}+e^{4t_2}-1]^{-1/4}$. In all three situations, C_1 and C_2 have independent exponential distributions with mean 2. According to Prentice and Cai (1992), the Gumbel model demonstrates a fairly weak positive dependence while the Clayton model demonstrates a fairly strong positive dependence.

In each situation, we carry out 1000 replications with sample sizes $n = 50$ and 100. Because our emphasis is on estimation of the marginal distribution functions, we only include the results of the estimated marginal distribution functions F_{1n} and F_{2n} at the 10TH, 30TH, 50TH, 70TH, and 90TH percentiles of the marginal distribution of T_1 (T_2). The results are given in Tables 1-6 as $q_{.1}, q_{.3}, q_{.5}, q_{.7}$ and $q_{.9}$.

Table 1. Estimated marginal distribution function: T_1, T_2 are independent exponential with mean 1; C_1, C_2 are independent exponential with mean 2; $n = 50$.

T_1 (T_2)		$F_{1n}(T_1)$	$F_{2n}(T_2)$	$F_{1n}^{(v)}(T_1)$	$F_{2n}^{(v)}(T_2)$
$q_{.1}$	mean	0.1118	0.1116	0.1182	0.1191
	std	0.0480	0.0535	0.0950	0.1083
$q_{.3}$	mean	0.3289	0.3280	0.3659	0.3680
	std	0.0811	0.0873	0.1954	0.1992
$q_{.5}$	mean	0.5387	0.5409	0.6310	0.6241
	std	0.1054	0.1096	0.2222	0.2245
$q_{.7}$	mean	0.7468	0.7486	0.8423	0.8473
	std	0.1183	0.1170	0.1766	0.1770
$q_{.9}$	mean	0.9304	0.9319	0.9781	0.9794
	std	0.0892	0.0871	0.0700	0.0691

Table 2. Estimated marginal distribution function: T_1, T_2 are independent exponential with mean 1; C_1, C_2 are independent exponential with mean 2; $n = 100$.

T_1 (T_2)		$F_{1n}(T_1)$	$F_{2n}(T_2)$	$F_{1n}^{(v)}(T_1)$	$F_{2n}^{(v)}(T_2)$
$q_{.1}$	mean	0.1063	0.1070	0.1056	0.1096
	std	0.0339	0.0338	0.0807	0.0931
$q_{.3}$	mean	0.3190	0.3177	0.3643	0.3873
	std	0.0623	0.0591	0.1785	0.2032
$q_{.5}$	mean	0.5271	0.5273	0.6315	0.6494
	std	0.0812	0.0764	0.2115	0.2137
$q_{.7}$	mean	0.7368	0.7375	0.8529	0.8688
	std	0.0927	0.0879	0.1619	0.1538
$q_{.9}$	mean	0.9286	0.9312	0.9824	0.9863
	std	0.0788	0.0748	0.0632	0.0519

Note. F_{1n} and F_{2n} are the proposed estimators; $F_{1n}^{(v)}$ and $F_{2n}^{(v)}$ are estimators based on the bivariate estimator considered by Van der Laan (1996).

Table 3. Estimated marginal distribution function: T_1, T_2 are from the Gumbel model; C_1, C_2 are independent exponential with mean 2; $n = 50$.

T_1 (T_2)		$F_{1n}(T_1)$	$F_{2n}(T_2)$	$F_{1n}^{(v)}(T_1)$	$F_{2n}^{(v)}(T_2)$
$q_{.1}$	mean	0.1096	0.1108	0.1182	0.1123
	std	0.0429	0.0417	0.1072	0.0890
$q_{.3}$	mean	0.3294	0.3292	0.3775	0.3666
	std	0.0833	0.0800	0.2011	0.1862
$q_{.5}$	mean	0.5414	0.5404	0.6396	0.6366
	std	0.1017	0.1051	0.2193	0.2183
$q_{.7}$	mean	0.7517	0.7474	0.8545	0.8622
	std	0.1145	0.1125	0.1672	0.1653
$q_{.9}$	mean	0.9329	0.9301	0.9806	0.9827
	std	0.0849	0.0868	0.0665	0.0619

Table 4. Estimated marginal distribution function: T_1, T_2 are from the Gumbel model; C_1, C_2 are independent exponential with mean 2; $n = 100$.

T_1 (T_2)		$F_{1n}(T_1)$	$F_{2n}(T_2)$	$F_{1n}^{(v)}(T_1)$	$F_{2n}^{(v)}(T_2)$
$q_{.1}$	mean	0.1096	0.1108	0.1182	0.1123
	std	0.0429	0.0417	0.1072	0.0890
$q_{.3}$	mean	0.3294	0.3292	0.3775	0.3666
	std	0.0833	0.0800	0.2011	0.1862
$q_{.5}$	mean	0.5414	0.5404	0.6396	0.6366
	std	0.1017	0.1051	0.2193	0.2183
$q_{.7}$	mean	0.7517	0.7474	0.8545	0.8622
	std	0.1145	0.1125	0.1672	0.1653
$q_{.9}$	mean	0.9329	0.9301	0.9806	0.9827
	std	0.0849	0.0868	0.0665	0.0619

Note. F_{1n} and F_{2n} are the proposed estimators; $F_{1n}^{(v)}$ and $F_{2n}^{(v)}$ are estimators based on the bivariate estimator considered by Van der Laan (1996).

Results indicate that the performance of our proposed estimators of the marginal distribution functions are satisfactory. However, we point out that the marginal probabilities are slightly over-estimated, though the biases are relatively small in comparison with the standard error of the estimators. Upward biases in the estimated marginal probabilities result from the large upward biases of the initial estimates on the boundary points, because these over-estimates are parts of the summations that estimate the marginal probabilities (see (4.17) and (4.18)).

Table 5. Estimated marginal distribution function: T_1, T_2 are from the Clayton model; C_1, C_2 are independent exponential with mean 2; $n = 50$.

T_1 (T_2)		$F_{1n}(T_1)$	$F_{2n}(T_2)$	$F_{1n}^{(v)}(T_1)$	$F_{2n}^{(v)}(T_2)$
$q_{.1}$	mean	0.1114	0.1106	0.1092	0.1071
	std	0.0417	0.0404	0.0787	0.0753
$q_{.3}$	mean	0.3301	0.3288	0.3823	0.3819
	std	0.0785	0.0776	0.1849	0.1946
$q_{.5}$	mean	0.5405	0.5369	0.6537	0.6536
	std	0.1025	0.1010	0.2054	0.2116
$q_{.7}$	mean	0.7479	0.7453	0.8706	0.8665
	std	0.1194	0.1167	0.1514	0.1546
$q_{.9}$	mean	0.9336	0.9341	0.9844	0.9869
	std	0.0920	0.0926	0.0580	0.0558

Table 6. Estimated marginal distribution function: T_1, T_2 are from the Clayton model; C_1, C_2 are independent exponential with mean 2; $n = 100$.

T_1 (T_2)		$F_{1n}(T_1)$	$F_{2n}(T_2)$	$F_{1n}^{(v)}(T_1)$	$F_{2n}^{(v)}(T_2)$
$q_{.1}$	mean	0.1088	0.1082	0.1078	0.1066
	std	0.0294	0.0292	0.0724	0.0786
$q_{.3}$	mean	0.3227	0.3210	0.3968	0.3843
	std	0.0587	0.0589	0.1949	0.1783
$q_{.5}$	mean	0.5316	0.5303	0.6669	0.6666
	std	0.0800	0.0793	0.2018	0.1993
$q_{.7}$	mean	0.7359	0.7382	0.8771	0.8784
	std	0.0911	0.0935	0.1440	0.1426
$q_{.9}$	mean	0.9304	0.9298	0.9868	0.9885
	std	0.0804	0.0814	0.0478	0.0450

Note. F_{1n} and F_{2n} are the proposed estimators; $F_{1n}^{(v)}$ and $F_{2n}^{(v)}$ are estimators based on the bivariate estimator considered by Van der Laan (1996).

A referee suggested comparing the performance of the proposed estimator with an existing one. Therefore, for each simulation model, we also calculated the marginal distribution based on the bivariate estimator considered by Van der Laan (1996). The later estimators are denoted by $F_{1n}^{(v)}$ and $F_{2n}^{(v)}$ in the tables. Comparing the proposed marginal distribution estimator and the estimators based on the bivariate estimator of Van der Laan (1996), we see that the proposed estimator has smaller bias and standard error at $q_{.1}, q_{.3}, q_{.5}$ and $q_{.7}$. At $q_{.9}$, the proposed estimator has smaller bias, but larger standard error, due to concentration at .98. For the models and sample sizes considered, the proposed marginal estimator has better performance than the estimator based on the bivariate estimator of Van der Laan (1996). Comparing the estimated marginal probabilities with sample sizes $n = 50$ and $n = 100$, we see that, for $n = 100$, the finite sample biases are uniformly smaller. This suggests that bias decreases as sample size increases, as should be the case for any reasonable estimators.

4.3. Testing age-of-onset anticipation based on marginal medians

In testing for age-of-onset anticipation for diseases with variable age of onset, a frequently used method is the paired t-test for a difference in the means using the observed ages of onset of affected parent-child pairs. As mentioned earlier, this test does not take into account the truncation effect. The nonparametric estimators proposed in Section 3 can be used for testing age-of-onset anticipation. Our proposal is to compare the estimated medians of the marginals F_{1n} and F_{2n}. Besides the usual advantage of using medians in possibly skewed distributions, there are two reasons for using medians rather than means in the present situation. First, calculation of means requires integration with respect to F_{1n} or F_{2n} over the whole support when, for finite sample size, the observed data may be far from covering the whole support. Second, the performance of F_{1n} and F_{2n} at the tails is not as good as that in the middle, again due to truncation.

Let m_1 and m_2 be the medians of $F_1(t_1)$ and $F_2(t_2)$ respectively. If $m_1 > m_2$, then it indicates that anticipation exists. To test the null hypothesis $H_0 : m_1 = m_2$, let $\hat{m}_1$ and $\hat{m}_2$ be the medians of $F_{1n}(t_1)$ and $F_{2n}(t_2)$ respectively. By the functional delta method (Gill (1989)), assuming that the marginal densities are not zero at m_1 or m_2, we expect that $n^{1/2}(\hat{m}_1 - \hat{m}_2)$ is approximately distributed as $N(m_1 - m_2, \sigma^2)$, where σ^2 is implicitly determined by the joint distribution $F(t_1, t_2)$. Without an explicit expression for the asymptotic variance function of F_n, we use the nonparametric bootstrap (Efron (1979)) to estimate the standard error of $\hat{m}_1 - \hat{m}_2$. Notice that we are not conducting a bootstrap test, in which bootstrap samples need to be simulated under the null hypothesis, see Efron and Tibshirani (1993). In the present bivariate truncation problem, it appears difficult to simulate bootstrap samples under the restriction of the null hypothesis.

Bootstrap estimation of the standard error of $\hat{m}_1 - \hat{m}_2$ is done as follows. Let $X_i = (T_{1i}, T_{2i}, C_{1i}, C_{2i}), i = 1, \ldots, n$.

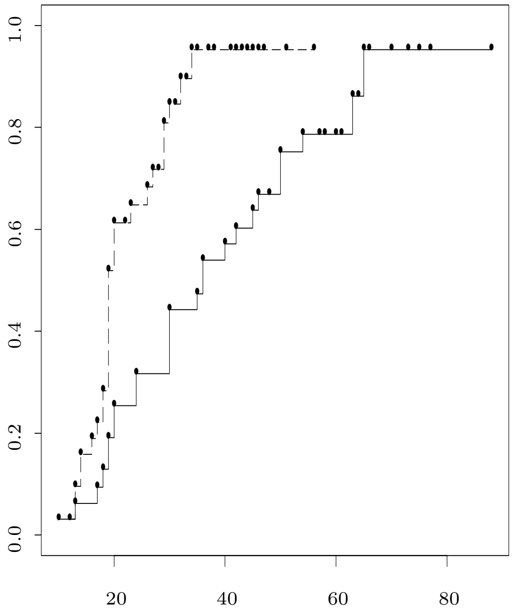
(i) Randomly generate B samples of size n, with replacement from the observed data $X_1, \ldots, X_n$, where each X_i is treated as a sampling unit.

(ii) Based on the bth bootstrap sample, compute the marginal estimators F_{1n}^{*b} and F_{2n}^{*b} using the same procedure as in computing F_{1n} and F_{2n}. Compute the medians $\hat{m}_1^{*b}$ and $\hat{m}_2^{*b}$ of F_{1n}^{*b} and F_{2n}^{*b}, respectively. Let $\Delta^{*b} = \hat{m}_1^{*b} - \hat{m}_2^{*b}$.

(iii) The bootstrap estimator of the standard error of $\hat{m}_1 - \hat{m}_2$ is given by $\widehat{\sigma}^2 = \frac{1}{B-1} \sum_{b=1}^{B} (\Delta^{*b} - \overline{\Delta}^{*b})^2$, where $\overline{\Delta}^{*b} = B^{-1} \sum_{b=1}^{B} \Delta^{*b}$.

4.4. Example: bipolar affective disorder data

We use the bipolar affective disorder data analyzed by McInnis et al. (1993) to illustrate the proposed test. 125 families were ascertained via eighteen hundred Probands screened for bipolar I or bipolar II found 125 fanilies in Baltimore and Iowa City. Ascertainment and evaluation criteria are described in detail by McInnis et al. (1993). Ages of onset and current ages of 34 parent-child pairs from 34 families among these 125 are available to us. The remaining 91 families were excluded because they either showed clinical evidence of bilineality or did not have at least one interviewed, affected individual in each of two successive generations. Among the 34 parent-child pairs, there were 9 pairs for which either ages of onset or current ages were not available, leaving 25 pairs with complete data on age of onset and current age.



Age of onset in years
Solid line: parents, dashed line: children

Figure 1. Estimated marginal distribution functions for bipolar data.

Figure 1 shows the estimated marginal distribution functions of the ages of onset of children and parents. Examination of the figure suggests that children tend to have younger age of onset than parents. In addition to examining the data in an exploratory fashion, a statistical test based on estimated medians can be carried out. Estimated median ages of onset in parents and children are 35 and 19, respectively. Using the bootstrap approach with sample size 1000, the estimated standard error of the difference in estimated medians is 8 (rounded from 8.04). Thus the standardized value of the difference in median ages of onset divided by 8 is $z = 2$, an approximate p-value of 0.023 for a one-sided test. This suggests a statistically moderate significant difference between the median ages of onset in parents and children.

One should be extremely cautious about interpretation of the estimates and the test result given above. First, the sample size is small. Second, although the truncation effect is adjusted for in the estimates, other ascertainment biases may exist that could produce false evidence in favor of anticipation. Specifically, because these data were originally obtained for the purposes of genetic linkage analysis, ascertainment was unsystematic and favored families with multiple affected individuals. Elsewhere (Huang and Vieland (1998)) we have shown that this type of sampling can inflate the Type I error rate of age-of-anticipation tests.

5. Asymptotic Properties

We now present some asymptotic results for the estimators $F_n^{(0)}$ and F_n. Proofs are given in the appendix. For any fixed $(t_1, t_2) \in R^{+2} \equiv [0, \infty) \times [0, \infty)$, let $\tau_1^F(t_1) = \inf\{s_2 : F(t_1, s_2) = F_1(t_1)\}$, $\tau_2^F(t_2) = \inf\{s_1 : F(s_1, t_2) = F_2(t_2)\}$ and $\tau_1^G(t_1) = \inf\{s_2 : G(t_1, s_2) = G_1(t_1)\}$, $\tau_2^G(t_2) = \inf\{s_1 : G(s_1, t_2) = G_2(t_2)\}$, We say that the support of F is upper-contained in the support of G if, for any (t_1, t_2), $\tau_1^F(t_1) \leq \tau_1^G(t_1)$ and $\tau_2^F(t_2) \leq \tau_2^G(t_2)$. For consistency of the estimator, we assume the following conditions:

(A1) The support of F is upper-contained in the support of G;

(A2) G is continuous, or

(A2*) G is discrete with finitely many mass points.

Assumption (A1) is crucial for the identifiability of F in the bivariate truncation model, similar to the assumption needed for identifiability in the univariate truncation model. The assumption that G is either continuous or discrete with finitely many mass points simplifies consistency proofs. (A2) or (A2*) should be satisfied in most practical situations.

Here we concentrate on the case when G is continuous. For the discrete case, it can be shown that under (A1) and (A2*), both $F_n^{(0)}$ and F_n are strongly consistent. Furthermore, let $n^{1/2}(F_n^{(0)} - F)$ or $n^{1/2}(F_n - F)$ denote the vector of the values of $n^{1/2}(F_n^{(0)} - F)$, or $n^{1/2}(F_n - F)$ evaluated at the finitely-many mass points of F. Then both $n^{1/2}(F_n^{(0)} - F)$ and $n^{1/2}(F_n - F)$ converge in distribution to multivariate normal distributions. These limit distributions have mean zero, but their covariance matrices are in general different. Asymptotic covariance matrices do not appear to have simple and explicit form. If it is desirable to treat the data as discrete in a specific situation and if the sample size is reasonably large, it is probably better to estimate F and G by directly maximizing the likelihood function. Then the variance can be estimated based on the observed information.

Theorem 5.1. *Suppose that* (A1) *and* (A2) *hold. Then for any* $(t_1, t_2) \in R^{+2}$,

$$F_n^{(0)}(t_1, t_2) \rightarrow_{a.s.} F_0(t_1, t_2), \quad \textit{and} \quad F_n(t_1, t_2) \rightarrow_{a.s.} F_0(t_1, t_2), \quad \textit{as } n \rightarrow \infty. \tag{5.19}$$

Let $[0, \tau] \equiv [0, \tau_1] \times [0, \tau_2]$, where τ_1 and τ_2 are two positive numbers. Let $D[0, \tau]$ be the space of bivariate functions which are right continuous and have left limits on $[0, \tau]$, see for example, Neuhaus (1971). We equip this space with the supremum norm $||\cdot||_\infty$, i.e., for any $g \in D[0, \tau]$, $||g||_\infty = \sup_{t \in [0,\tau]} |g(t)|$. The convergence in distribution below is according to Hoffmann-Jørgensen (1984); see e.g., Van der Vaart and Wellner (1996) for a description. For the asymptotic normality in the case of continuous F, the following extra conditions are needed.

(A3) the support of G is a bounded interval $[0, \tau]$.

(A4) F is continuous, and G has a continuous density function g with g/F being bounded on $[0, \tau]$.

Theorem 5.2. *Suppose that* (A1), (A2), (A3) *and* (A4) *hold. Let* $A(t) = \int_t^\tau dG(u)/F(u)$. *Then* $\sqrt{n}A\,(F_n^{(0)} - F_0) \Rightarrow_D Z$, *as* $n \rightarrow \infty$ *on* $[\tau_0, \tau]$, *where* $\tau_0 > 0$ *and where* Z *is a Gaussian process on* $[\tau_0, \tau]$.

Theorem 5.1 justifies our proposed estimator F_n in the sense that it is consistent. Theorem 5.2 states that the $F_n^{(0)}$ is asymptotically normal. However, its finite sample behavior is not satisfactory. We are not able to rigorously prove the weak convergence of F_n, although Theorem 5.2 and heuristics indicate that F_n should be asymptotically normal. The main technical difficulties are: (a) we are not able to prove the weak convergence of $F_n^{(0)}$ on $[0, \tau]$, only on $[\tau_0, \tau]$ for some $\tau_0 > 0$; and (b) the denominator $K_n^+(s_1, s_2)$ inside the integral in (3.13) converges to 0 as $(s_1, s_2) \rightarrow \tau$. However, simulations suggest that F_n outperforms $F_n^{(0)}$ with finite sample size. In particular, F_n outperforms $F_n^{(0)}$ on the boundary points, which appears to be true in general (i.e., not just in our simulation models). The better performance of F_n on the boundary points is particularly important to estimation of marginal distributions. Therefore, we prefer F_n although we have not been able to demonstrate its asymptotic distribution.

6. Discussion

A main difficulty in the present estimation problem is that there do not appear to exist straightforward estimators of marginal distributions. At least we have not been able to construct such estimators without first estimating the joint distribution. This is in contrast to estimation of a distribution function based on bivariate right-censored or bivariate data with a single component subject to truncation, in which marginal distributions can be easily estimated, which in turn facilitates estimation of the joint distribution. In this regard, bivariate truncation problem appears to be more difficult.

On the other hand, if the marginals of F can be estimated easily, perhaps via information outside of the data then, based on (3.11), we can derive an estimating equation for F. Let

$$H(t_1, t_2) = \int_{[0,t_1]} \int_{[0,t_2]} \frac{dS^*(s_1, s_2)}{K^+(s_1, s_2)}.$$

Then $dF(t_1, t_2) = F(t_1, t_2) dH(t_1, t_2)$ by equation (3.12). Let $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$. We have

$$S(t_1, t_2) = \int_{(t_1,\infty)} \int_{(t_2,\infty)} F(s_1, s_2) dH(s_1, s_2).$$

Since $S(t_1, t_2) = 1 - F_1(t_1) - F_2(t_2) + F(t_1, t_2)$, it follows that

$$F(t_1, t_2) = F_1(t_1) + F_2(t_2) - 1 + \int_{(t_1,\infty)} \int_{(t_2,\infty)} F(s_1, s_2) dH(s_1, s_2).$$

For known F_1, F_2 and H, this is an inhomogeneous Volterra equation with an unique solution F, see for example Kontorovich and Akilov (1982, p.396). A related equation arises in the bivariate right-censorship model as described by Gill, Van der Laan and Wellner (1995), in which marginal distributions can be estimated by using the Kaplan-Meier estimator. However, this is not the case in the present bivariate truncation model. Therefore, the approaches of Dabrowska (1988) and Prentice and Cai (1992) for estimating the joint distribution based the bivariate right-censored data, and the Volterra integral approach described in Gill, Van der Laan and Wellner (1995), do not seem to apply.

For bivariate right-censored data there exist several competing estimators of the distribution function, see for example, Campbell (1981), Dabraskow (1988), Prentice and Cai (1992), Gill, Van der Laan and Wellner (1995), and Van der Laan (1996). For bivariate data when a single component is subject to truncation, Güler (1996) described three different estimators. This may also be the case for bivariate data when both components are subject to truncation. It would be of interest to search for estimators for both the bivariate distribution and its marginals different from the ones considered in this paper.

Acknowledgements

We are grateful to Dr. Melvin G. McInnis and his colleagues at the Johns Hopkins University for providing us with the age of onset information from their bipolar affective disorder data. This work is supported in part by the NIMH grant K01-01541 [JH] and R01-52841, K02-01432 [VJV].

Appendices

Appendix A: Motivation for using K^* in constructing $F_n^{(0)}$

The use of K^* in constructing $F_n^{(0)}$ may not be obvious at first look. It arises in our attempt to express the distributions of the observed ages of onset and the observed ages at interview in terms of F. The distribution function of the observable ages of onset is given by

$$F^*(t_1, t_2) = P(T_1 \le t_1, T_2 \le t_2 | T_1 \le C_1, T_2 \le C_2). \tag{A.20}$$

Since $dG^*(t_1, t_2) = \alpha^{-1} F(t_1, t_2) dG(t_1, t_2)$, (A.20) can be rewritten as

$$\begin{aligned} F^*(t_1, t_2) &= \alpha^{-1} \int P(T_1 \le t_1 \wedge c_1, T_2 \le t_2 \wedge c_2) dG(c_1, c_2) \qquad \text{(A.21)} \\ &= \int \frac{F(t_1 \wedge c_1, t_2 \wedge c_2)}{F(c_1, c_2)} dG^*(c_1, c_2) \\ &= G^*(t_1, t_2) + \xi_1^*(t_1, t_2) + \xi_2^*(t_1, t_2) + F(t_1, t_2) \int_{t_1}^{\infty} \int_{t_2}^{\infty} \frac{dG^*(c_1, c_2)}{F(c_1, c_2)}, \end{aligned}$$

where

$$\xi_1^*(t_1, t_2) = \int_{t_1}^{\infty} \int_0^{t_2} \frac{F(t_1, c_2)}{F(c_1, c_2)} dG^*(c_1, c_2) = P(C_1 > t_1, T_1 \le t_1, C_2 \le t_2 | T_1 \le C_1, T_2 \le C_2),$$

$$\xi_2^*(t_1, t_2) = \int_0^{t_1} \int_{t_2}^{\infty} \frac{F(c_1, t_2)}{F(c_1, c_2)} dG^*(c_1, c_2) = P(C_2 > t_2, T_2 \le t_2, C_1 \le t_1 | T_1 \le C_1, T_2 \le C_2).$$

It can be verified that

$$K^*(t_1, t_2) \equiv F^*(t_1, t_2) - G^*(t_1, t_2) - \xi_1^*(t_1, t_2) - \xi_2^*(t_1, t_2). \tag{A.22}$$

From (A.21), we obtain $K^*(t_1, t_2) = F(t_1, t_2) \int_{(t_1, \infty)} \int_{(t_2, \infty)} \frac{dG^*(c_1, c_2)}{F(c_1, c_2)}$, which is exactly (3.4).

Appendix B: Proofs

Proof of Theorem 5.1

Without loss of generality, we assume that the supports of both F and G are $[0, \infty) \times [0, \infty)$. To simplify notations, let $t = (t_1, t_2), c = (c_1, c_2)$ and write $\int_t^{\infty} = \int_{t_1}^{\infty} \int_{t_1}^{\infty}$.

By the multivariate version of Helly's Selection Theorem, there exists a non-decreasing, right-continuous function F_* such that any subsequence of $F_n^{(0)}$ has a further subsequence converging to F_* at every continuity point t of F_*. (Billingsley (1986, p.392)). If we can show that $F_* = F$ for every t, then the whole sequence converges to F pointwise.

If $F(t) = 0$, then by the definition of K_n^*, $K_n^*(t) = 0$ with probability one. Therefore $F_n^{(0)}(t) = 0$ with probability one. So we need only prove (5.19) for t satisfying $F(t) > 0$. By the definition of K^*, (3.4), $K^*(t) > 0$. This implies $F_*(t) > 0$, since otherwise the left side of (3.8) converges to $K^*(t) > 0$ by consistency of empirical distribution functions, whereas the right side converges to 0.

Step (i). We first prove that

$$F_n^{(0)}(t) \int_t^\infty \frac{dG_n^{\#}(c)}{F_n^{(0)}(c)} \rightarrow_{a.s.} F_*(t) \int_t^\infty \frac{dG^*(c)}{F_*(c)}.$$

By the second equality of (3.8), it suffices to show that

$$F_n^{(0)}(t) \int_t^\infty \frac{dG_n^{*}(c)}{F_n^{(0)}(c)} \rightarrow_{a.s.} F_*(t) \int_t^\infty \frac{dG^*(c)}{F_*(c)}.$$

Write

$$\begin{aligned}
&F_n^{(0)}(t) \int_t^\infty \frac{dG_n^{*}(c)}{F_n^{(0)}(c)} - F_*(t) \int_t^\infty \frac{dG^*(c)}{F_*(c)} \\
&= \int_t^\infty \frac{F_n^{(0)}(t)}{F_n^{(0)}(c)} d(G_n^* - G^*)(c) + \int_t^\infty \left[\frac{F_n^{(0)}(t)}{F_n^{(0)}(c)} - \frac{F_*(t)}{F_*(c)} \right] dG^*(c) \\
&\equiv A_{1n} + A_{2n}.
\end{aligned}$$

Now A_{1n} converges to zero by the uniform convergence of empirical measures and A_{2n} converges to zero by the Dominated Convergence Theorem.
Step (ii). We now show that $F_* \equiv F$. Since $K_n^*(t) \rightarrow_{a.s.} K^*(t)$ for every t, and $K^*(t) = F(t) \int_t^\infty \frac{dG^*(c)}{F(c)}$, it follows that

$$F(t) \int_t^\infty \frac{dG^*(c)}{F(c)} = F_*(t) \int_t^\infty \frac{dG^*(c)}{F_*(c)}. \tag{A.23}$$

Since $dG^*(t) = \alpha^{-1} F(t) dG(t)$, we have $F(t) \int_t^\infty dG(c) = F_*(t) \int_t^\infty \frac{F(c)}{F_*(c)} dG(c)$. Let $H(t) = F(t)/F_*(t)$. We have

$$H(t) \int_t^\infty dG(c) = \int_t^\infty H(c) dG(c). \tag{A.24}$$

By the assumption that G is continuous, H is continuous. We show that $H(t) = k_0$ for all $t > 0$, where k_0 is a constant independent of t. If this is not true, let t_M and t_m be points where $H(t)$ achieves global maximum and minimum, respectively. It suffices to show that $H(t_m) = H(t_M)$. Suppose not and consider two cases.

(a) Both t_m and t_M are finite, then either

$$H(t_m)\int_{t_m}^{\infty} dG(c) < \int_{t_m}^{\infty} H(c)dG(c) \quad \text{or} \quad H(t_M)\int_{t_M}^{\infty} dG(c) > \int_{t_M}^{\infty} H(c)dG(c),$$

which contradicts (A.24).

(b) At least one of t_m and t_M is not finite. We only consider the case when $t_m = (t_1^*, \infty)$. So $\lim_{t_2^*\to\infty} H(t_1^*, t_2^*) \le H(t_1, t_2)$ for all (t_1, t_2). We can take t_1^* to satisfy $\lim_{t_2^*\to\infty} H(t_1^*, t_2^*) < H(t_1, t_2)$ for all $t_1 > t_1^*$. Then there exists a large $M_0 > 0$ such that $H(t_1^*, M_0) \le H(t_1, t_2)$ for all $t_1 > t_1^*, t_2 > M_0$, where strict inequality holds in a subset B_0 of $\{(t_1, t_2) : t_1 > t_1^*, t_2 > M_0\}$ with $\int_{B_0} dG > 0$. Therefore, $H(t_1^*, M_0)\int_{t_1^*}^{\infty}\int_{M_0}^{\infty} dG(c_1, c_2) < \int_{t_1^*}^{\infty}\int_{M_0}^{\infty} H(c_1, c_2)dG$ (c_1, c_2). This again is a contradiction.

It follows that $H(t) = k_0$ for all $t > 0$. Since $\lim_{t\to\infty} H(t) = \lim_{t\to\infty}$ $[F(t)/F_*(t)] = 1$, we have $k_0 = 1$. This implies that $F_*(t) = F(t)$ for all $t > 0$. Continuity of F_* and F implies that $F_*(t) = F(t)$ for all $t \ge 0$, and $F_n^{(0)}$ is consistent.

Fianlly, consistency of F_n follows directly from the strong consistency of S_n^* and K_n^+, and (3.13).

Proof of Theorem 5.2. The result can be proved by verifying the conditions of the theorem on the asymptotic distribution of infinite-dimensional M-estimators of Van der Vaart (1995). Here we give a direct proof, since it takes less space.

By (3.4), (3.8) and some straightforward calculation, we have

$$\frac{n+1}{n}[K_n^*(t)-K^*(t)] = [F_n^{(0)}(t)-F(t)]\int_t^{\tau}\frac{dG_n^*(c)}{F_n^{(0)}(c)} - F(t)\int_t^{\tau}\frac{F_n^{(0)}(c)-F(c)}{F(c)F_n^{(0)}(c)}dG^*(c) + F(t)\int_t^{\tau}[F_n^{(0)}(c)]^{-1}d(G_n^* - G^*)(c) + \frac{1}{n}F_n(t). \tag{A.25}$$

We ignore the last term on the right side of (A.25) since it is of order lower than $n^{-1/2}$. Let

$$A_n(t) = \int_t^{\tau}\frac{dG_n^*(c)}{F_n^{(0)}(c)}, \quad A(t) = \int_t^{\tau}\frac{dG^*(c)}{F(c)},$$

$$D_n(t) = \frac{n+1}{n}[K_n^*(t) - K^*(t)] - F(t)\int_t^{\tau} F^{-1}(c)d(G_n^* - G^*)(c),$$

and $\phi_n(t) = [F_n^{(0)}(t) - F(t)]A(t)$. It follows from (A.25) and $dG^* = \alpha^{-1}FdG$ that

$$D_n(t) = \phi_n(t) - F(t)\int_t^{\tau}\frac{\phi_n(c)}{\alpha F(c)A(c)}dG(c) + r_{1n}(t) + r_{2n}(t), \tag{A.26}$$

where

$$r_{1n}(t) = [F_n^{(0)}(t) - F(t)][A_n(t) - A(t)] + \alpha^{-1}F(t)\int_t^\tau [F(c) - F_n^{(0)}(c)]([F_n^{(0)}(c)]^{-1} \\ -[F(c)]^{-1})dG(c),$$

$$r_{2n}(t) = F(t)\int_t^\tau [(F_n^{(0)}(c))^{-1} - (F(c))^{-1}]d(G_n^* - G^*)(c).$$

By the strong consistency of $F_n^{(0)}$, $\sup_{\tau_0 \le t \le \tau} |r_{1n}(t)| = o_p(1)||F_n^{(0)} - F||_{\tau_0}^\tau$. Since the class of bivariate distribution functions is universal Donsker, $F(\tau) > 0$ and $F_n^{(0)}$ is consistent, and an empirical process indexed by a Donsker class is asymptotically equicontinuous, it follows that $\sup_{\tau_0 \le t \le \tau} |r_{2n}(t)| = o_p(n^{-1/2})$. Let the operator Γ with domain $D^2[\tau_0, \tau]$ be defined by

$$\Gamma h(t) = h(t) - \alpha^{-1}F(t)\int_t^\tau \frac{g_0(c)}{F(c)A(c)}h(c)dc.$$

This operator closely resembles the Volterra operator, its continuous invertibility can be proved the same way as in Kantorovich and Akilov (1982, p.396). Notice that assumption (iii) is needed here. By (A.26), we have $(n^{1/2}\phi_n(t) = n^{1/2}\Gamma^{-1}D_n(t) + o_p(1)$. By the weak convergence of empirical processes, $n^{1/2}D_n$ converges in distribution to a Gaussian process in $D[\tau_0, \tau]$. Since Γ^{-1} is a linear operator, $n^{1/2}\phi_n$ converges to a Gaussian process in $D[\tau_0, \tau]$. This completes the proof.

References

Asherson, P. and Owen, M. (1994). Anticipation in mental illness. *Amer. J. Hum. Genet.* **54**, 386-387.

Bassett, A. and Honer, W. (1994). Evidence for anticipation in schizophrenia. *Amer. J. Hum. Genet.* **54**, 864-870.

Bell, J. (1947). Dystrophia myotonica and allied disease: light on an old problem. In *Treasury of Human Inheritance* **4** (Edited by LS Penrose), 343-410. Cambridge University Press, Cambridge.

Campbell, G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika* **68**, 417-442.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statitst.* **16**, 1475-1489.

Deighton, C., Heslop, P., McDonagh, J., Walker, D. and Thomson, G. (1994). Does genetic anticipation occur in familial rheumatoid arthritis. *Ann. Rheum. Dis.* **53**, 833-835.

DeStefano, A. L., Cupples, A., Maciel, P., Gaspar, C. et al. (1996). A familial factor independent of CAG repeat length influences age at onset of Machado-Joseph disease. *Amer. J. Hum. Genet.* **59**, 119-127.

Gill, R. D. (1989). Non- and Semi-parametric maximum likelihood estimators and the von-Mises method (part I). *Scand. J. Statist.* **16**, 97-128.

Gürler, Ü. (1996). Bivariate estimation with right-truncated data. *J. Amer. Statist. Assoc.* **91**, 1152-1165.

Harper, P.S., Harley, H. G., Reardon, W. and Shaw, D. J. (1992). Anticipation in myotonic dystrophy: new light on an old problem. *Amer. J. Hum. Genet.* **51**, 10-16.

Heiman, G. A., Hodge, S. E., Wickramaratne, P. and Hsu, H. (1996). Age-at-interview bias in anticipation studies: computer simulations and an example with panic disorders. *Psych. Genet.* **2**, 61-66.

Heimbuch, R. C., Matthysse, S. and Kidd, K. (1980). Estimating age-of-onset distribution for disorders with variable onset. *Amer. J. Hum. Genet.* **32**, 564-574.

Horwitz, M., Goode, E. L. and Jarvik, G. P. (1996). Anticipation in familial leukemia. *Amer. J. Hum. Genet.* **59**, 990-998.

Höweler, C. F., Busch, H. F. M., Geraedts, J. P. M., Niermeijer, M. F. and Staall, A. (1989). Anticipation in myotonic dystrophy: fact or fiction? *Brain* **112**, 779-797

Huang, J. and Vieland, V. J. (1997). A new statistical test for age of onset anticipation: with application to bipolar disorder. *Genetic Epidemiology* **14**, 1091-1096.

Vieland, V. J. and Huang, J. (1998). Statistical evaluation of age-of-onset anticipation: a new test and evaluation of its behavior in realistic applications. *Amer. J. Hum. Genet.* **62**, 1212-1227.

Lin, D. Y. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika* **80**, 573-581.

The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **112**, 971-983.

Mandel, J. L. (1994). Trinucleotide disease on the rise. *Nat. Genet.* **7**, 453-455.

McInnis, M. G., McMahon, F. J., Stine, O. C. and Ross, C. A. (1993). Anticipation in bipolar affective disorder. *Amer. J. Hum. Genet.* **53**, 385-390.

Mott, F. W. (1910). Hereditary aspects of nervous and mental disease. *British Medical Journal* **2**, 1013-1020.

Myers, R. H., Cupples, L. A., Schoenfeld, M., D'Agostino, R. B., Terrin, N. C., Goldmakher, N. and Wolf, P. A. (1985). Maternal factors in onet of Huntington disease. *Amer. J. Hum. Genet.* **37**, 511-523.

Neuhaus, G. (1971). On weak convergence of stochastic processes with multidimensional time parameters. *Ann. Math. Statist.* **42**, 1285-1295.

Paterson, A. D., Kennedy, J. L. and Petronis, A. (1996). Evidence for genetic anticipation in non-Mendelian diseases. *Amer. J. Hum. Genet.* **59**, 264-268.

Prentice, R. L. and Cai, J. (1992). Covariance and survival function estimation using censored multivariate failure time data. *Biometrika* **79**, 495-512.

Sano, A., Yamauchi, N., Kakimoto, Y., Komure, O., Kawai, J., Hazama, F., Kuzume, K., Sano, N. and Kondo, I. (1994). Anticipation in hereditary dentatorubral-pallidoluysian atrophy. *Human Genetics* **93**, 699-702.

Suthers, G. K., Huson, S. M. and Davies, K. E. (1992). Instability versus predictability: the molecular diagnosis of myotonic dystrophy. *J. Med. Genet.* **14**, 761-765.

Van der Laan, M. J. (1995). Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Statist.* **24**, 596-627.

Van der Laan, M. J. (1996). Nonparametric estimation of the bivariate survival function with truncated data *J. Multivariate Anal.* **58**, 107-131.

Van der Vaart, A. W. (1995). Efficiency of infinite dimensional M-estimators. *Statistica Neerlandica* **49**, 9-30.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer Verlag, New York.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13**, 163-177.

Zatz, M, Maria, S. K., Passo-Bueno, M. A., Vainzof, M., Camplotto, S., Cerqueira, A., Wijmenga, C., Padberg, G. and Frants, R. (1995). High proportion of new mutations and possible anticipation in Brazilian facioscapulohumeral muscular dystrophy families. *Amer. J. Hum. Genet.* **56**, 99-105

Department of Statistics and Actuarial Science, 241 Schaeffer Hall, University of Iowa, Iowa City, IA 52242, U.S.A.

E-mail: jian-huang@uiowa.edu

Departments of Biostatistics and Psychiatry, University of Iowa, Iowa City, IA 52242, U.S.A.

E-mail: veronica-vieland@uiowa.edu

Department of Biostatistics, University of Iowa, Iowa City, IA 52242, U.S.A.

E-mail: kai-wang@uiowa.edu

(Received April 2000; accepted March 2001)