Bivariate Estimation With Truncated Data and Applications to AIDS

Ülkü Gürler*
Bilkent University
Industrial Engineering Department
06533 Bilkent, Ankara, TURKEY
e-mail:ulku@bilkent.edu.tr

June 1994

AMS 1990 subject classifications:62G05, 62G20

Key words and phrases: Bivariate distribution, right truncation, reverse-hazard, TR-AIDS.

———————————————————————

⋆: This work is completed while the author was visiting the Institut de Statistique, Universite Catholique de Louvain, Belgium, in the academic year 1993-1994.

Abstract

Bivariate estimation with survival data has received considerable attention recently. Most of such work however focuses on random censoring models. Random truncation is another common feature of survival data, which we will consider in this study. Random truncation may arise if the time origin of the events under study precede the observation period. Trunsfusion related AIDS (TR-AIDS) data on the incubation period of the disease is a commonly given example. More precisely, in random right truncation model one observes the i.i.d. samples of (Y, T) only if $(Y \leq T)$, where Y is the variable of interest, and T is an independent variable which prevents the complete observation of Y. If the variable of interest is T, then it is called the random left truncation. Suppose (Y, X) is a bivariate vector of random variables, with joint distribution function given by $F(y, x)$. In a recent study, Gürler(1994) presented nonparametric estimators for $F(y, x)$, when Y is subject to left truncation. In the present study, bivariate estimation methods are discussed for right trunaction model. The bivariate reverse-hazard estimators are introduced and methods for estimating bivariate d.f. are discussed, two of which are based on the reverse-hazard vector. The behavior of these estimators are inspected and compared w.r.t. several aspects. The simplest of these estimators, which is also a proper bivariate d.f. turns out to dominate the others in the simulations too. An i.i.d. representation for this estimator is presented from which a.s. consistency and weak convergence follows. The main components of the asymptotic covariance structure are also provided. Finally, some applications of the suggested estimator are presented via the the TR-AIDS data on the incubation time.

1 Introduction

Randomly truncated data frequently arise in medical studies. Other application areas include economics, insurance and astronomy(see eg. Lynden-Bell(1971), Kaminsky(1987)). In a broad sense, random truncation corresponds to biased sampling, where only a partial or incomplete data is available about the variable of interest. A typical realization can occur as follows: suppose individuals or items may experience two consecutive events in time, an initiating one at t and aterminating one at s. Usually, the statistical interest is the duration between the two. Random truncation may occur, if the observation period starts after the initiating event. In particular, consider the following AIDS example: suppose t is the time of infection with HIV and s is the time of the onset of AIDS. If the observation period is terminated at time T_e, then only those individuals for whom $Y = t - s \leq T = T_e - s$ can be observed, and right truncation occurs. Here Y corresponds to the incubation time of the disease. In AIDS cohort studies, a group of patients who are infected with HIV but not yet developed AIDS are followed for a certain time. If the recruitment starts at time T_o and the follow-up is terminated at T_e, then, only the individuals for whom $Y = t - s \geq T = T_o - s$ are observed, and random left truncation occurs. In the course of the follow-up, these individuals are also subject to right censoring, which natuarally leads to left truncation and right censoring (LTRC). This incomplete structure of the data obviously creates bias in estimation, and this bias is particularly important at the early stages of the disease, when sufficient historical data has not yet accumulated. For the analysis of AIDS related data, see e.g. Lagakos, Barraj and DeGruttola(1988) and Medley et al.(1988) among many others.

Consider first the univariate truncation model, where one observes the i.i.d. pairs (Y_i, T_i), $i = 1, .., n$, only if $(Y_i \leq T_i)$, where the main interest is in Y . Let F and G be the distribution functions (d.f.) of Y and T respectively. Woodroofe (1985) supplies the following identifiability condition: F and G can be estimated completely, only if $(F, G) \in R_o$, where $R_o = \{(F, G) : \ a_G \leq a_F;\ b_G \leq b_F\}$ with a_W and b_W denoting the lower and upper end points of the support of any d.f. W respectively. Then $(F, G) \in R_o$ implies $\alpha \equiv P(T \leq Y) > 0$ Here α is the proportion of the bivariate population (Y, T), which can be observed under the truncation scheme. Assuming the identifiability of F and G, their nonparametric estimators F_n, G_n are given by

$$F_n(y) \quad = \quad \prod_{i:Y_i > y} \left[1 - \frac{s(Y_i)}{nC_n(Y_i)}\right]$$

$$G_n(t) = 1 - \prod_{i:T_i \le t} \left[1 - \frac{r(T_i)}{nC_n(T_i)}\right] \tag{1}$$

where, for $u > 0$, $r(u) = \#\{i : T_i = u\}$, $s(u) = \#\{i : Y_i = u\}$, and

$$nC_n(u) = \#\{i : Y_i \le u \le T_i\}. \tag{2}$$

Lynden-Bell(1971) suggested $G_n(t)$ as the nonparametric maximum likelihood estimator of $G(t)$ in a problem that arose in astronomy. Woodroofe (1985) and Wang, Jewell and Tsai (1986) obtained consistency and weak convergence results. Chao and Lo (1988) presented an i.i.d. representation of $G_n(t)$, for which the order of the remainder term is improved by Stute(1993) and is extended to the LTRC model by Gijbels and Wang(1993), again with an improved order. Gijbels and Wang applied their results to kernel density estimation. Kernel estimators of the hazard function for LTRC data are studied by Uzunoğulları and Wang (1992). Gürler, Stute and Wang(1993), Gu and Lai (1990), Lai and Ying (1991), and Gross and Caroll(1992) extended the results for truncated/censored data in various directions. See also Keiding and Gill(1990) for a Markov -Process approach to the model. Recently, Gürler(1994) studied nonparametric estimators of the bivariate d.f. when a component is subject to left truncation. In her study, three nonparametric estimators for the bivariate d.f. are presented and their a.s consistency and weak convergence are established via strong i.i.d. representations. It is mentioned that only one of these estimators is a proper bivariate d.f., and the other two could assign negative mass, as is the case with several estimators in random censoring (see eg. Dabrowska(1988), Campbell(1981)). Despite this drawback, the other estimators are reported to have a slightly smaller bias, based on simulation studies.

In the present study, the results of Gürler(1994) is extended in the following directions: (i) Estimators of the bivariate 'reverse-hazard' vector and the bivariate d.f. are presented for right truncation model and a detailed comparison is provided. (ii) Alternative approaches for estimation are introduced and their relative performance are discussed. (iii) Finally, an analysis of TR-AIDS(Transfusion related AIDS) data is presented as an application. As such, the present work can be considered as a complimentary study of Gürler(1994), where the practical aspects of the estimators are emphasized more.

The rest of the paper is organised as follows: In Section 2, the bivariate model is introduced and nonparametric estimators are provided for the bivariate 'reverse-hazard' vector and the d.f.. In Section

3, results of the simulation studies are presented. Finally in Section 4, an application of these methods is illustrated with TR-AIDS data.

2 Bivariate Model and the Estimators

2.1 Preliminaries

Bivariate survival function is important in understanding the joint behavior of correlated lifetimes, as well as in assessing the strength of such association. Recently, there has been several studies elaborating the subject, such as Prentice and Cai(1993), Lin and Ying(1993), van der Laan(1993) and Tsai and Crowley(1994). All of these studies however consider variations of the random censoring model. Despite the important applications in survival analysis, bivariate estimation with truncated data has not received much attention in the literature so far. Following the work of Gürler(1994), we introduce below the bivariate right truncation model and discuss methods of estimation. Suppose we are interested in estimating the bivariate d.f. $F(y,x)$ of the random pair (Y,X). However, due to truncation effects, we can only observe the triplets (Y_i, X_i, T_i), $i = 1, \ldots, n$ for which $(Y_i \leq T_i)$. Here, T is a random variable, which is assumed to be independent of (Y,X), with d.f. G. The marginal d.f.'s of Y and X are denoted by F_Y and F_X respectively. For the identifiability of F, it is necessary that $a_{F_Y} \leq a_G$ and $b_{F_Y} \leq b_G$, as in the univariate model. We assume this condition holds in the rest of the paper, however in Section 4, further comments are made on this issue. The observed samples then have the transformed distribution H, given by:

$$
\begin{aligned}
H_{Y,X,T}(y,x,t) &= P\,(\,Y \leq y, X \leq x, T \leq t \mid Y \leq T) \\
&= \alpha^{-1} \int_0^t F(y \wedge u, x)\, dG(u)
\end{aligned}
$$

where α is as defined before, and $y \wedge u = \min(y,u)$. The problem is to reconstruct the latent $F(y,x)$ from the observable H and its marginals. Some of the bivariate and univariate marginals are as follows:

$$
F^*_{Y,X}(y,x) \equiv H_{Y,X,T}(y,x,\infty) = \alpha^{-1} \int_0^\infty F(y \wedge u, x)\, dG(u)\,.
$$

$$
H_{X,T}(x,t) = \alpha^{-1} \int_0^t F(u,x)\, dG(u)
$$

$$
H_{Y,T}(y,t) = \alpha^{-1} \int_0^t F_Y(y \wedge u)\, dG(u).
$$

$$F_Y^*(y) = \alpha^{-1} \int_0^\infty F_Y(y \wedge u)\, dG(u)$$
$$G_T^*(t) = H_{Y,T}(\infty, t) = \alpha^{-1} \int_0^t F_Y(u)\, dG(u).$$

Assuming the existance of the densities(denoted in lowecase letters), we get

$$f_{Y,X}^*(y,x) = \alpha^{-1}[1 - G(y)]\, f_{Y,X}(y,x)$$
$$f_Y^*(y) = \alpha^{-1}[1 - G(y)]\, f_Y(y) \tag{3}$$

Another quantity of interest, which provides insight to the truncation model is the following:

$$C(z) = \alpha^{-1}\bar{G}(z-)F_Y(z) = F_Y^*(z) - G_T^*(z) \tag{4}$$

This quantity plays an important role in the identification and the estimation of the model. The empirical counterpart, $C_n(z)$, is proportional to the 'risk-set' at time z given in (2). Unlike the usual risk-sets employed in survival analysis however, (2) is not a monotone function. In particular, it tails-off to zero at both ends and this introduces additional difficulty in the analysis.

2.2 Bivariate 'reverse-hazard' :

In univariate models, it is well known that the hazard function and the d.f. determine each other in a unique way and this correspondance is exploited to get an estimate of the d.f. via that of the hazard in survival analysis. In the bivariate case however there has been several definitions of the hazard function or the failure rate (see eg. Marshall(1975), Pons(1986)). Dabrowska(1988) presented a very nice representation of the bivariate d.f. in terms of the three component bivariate hazard vector she has introduced. Below we introduce the bivariate reverse-hazard vector which is analogous to hers and which turns out to be the natural quantity to consider in the right truncation model.

Definition: Suppose the $F(y,x)$ is continuous in both components and the bivariate density $f(y,x)$ exists. Then we define the bivariate 'reverse- hazard' vector $\tilde{\Lambda}(u,v)$ as :

$$\begin{aligned}\tilde{\Lambda}(u,v) &\equiv \{\frac{F(du,dv)}{F(u,v)}, \frac{F(du,v)}{F(u,v)}, \frac{F(u,dv)}{F(u,v)}\} \\ &\equiv \{\Lambda_{12}(du,dv), \Lambda_1(du,v), \Lambda_2(u,dv)\} \qquad (5)\end{aligned}$$

We adopted the term 'reverse-hazard' as coined by Lagakos et al.(1988) in the univariate case, since in usual hazard an immediate 'future' failure is considered, whereas the above vector relates to instantaneous 'past' failure. Gross and Carol(1992) use the term 'retro-hazard' for the same quantity in the univariate set-up. The general correspondance between the reverse-hazard vector and a bivariate survival function can be established as in Dabrowska(1988), and will not be presented here. However, when F is continuous and the density exists, we have the following relation: Let $R(y,x) = -\log F(y,x)$, then

$$F(y,x) \;=\; F_X(x)F_Y(y)exp\{-\Lambda(y,x)\} \tag{6}$$

where

$$\begin{aligned}\Lambda(y,x) &= \int_y^{b_{F_Y}}\int_x^{b_{F_X}} R(du,dv)\\ &= \int_y^{b_{F_Y}}\int_x^{b_{F_X}} \{\Lambda_{12}(du,dv) - \Lambda_1(du,v)\Lambda_2(u,dv)\}\end{aligned}$$

From the relations given in the previous section, it is not hard to verify the following

$$\Lambda_{12}(du,dv) = \frac{F^*_{Y,X}(du,dv)}{C_2(u,v)}$$

$$\Lambda_1(du,v) = \frac{F^*_{Y,X}(du,v)}{C_2(u,v)}$$

$$\Lambda_2(u,dv) = \frac{C_2(u,dv)}{C_2(u,v)}.$$

where

$$\begin{aligned}C_2(y,x) &= F^*_{Y,X}(y,x) - H_{T,X}(y-,x)\\ &= \alpha^{-1}[1-G(y)]F(y,x)\end{aligned} \tag{7}$$

Let $F^*_{n,Y,X}(y,x)$ and $H_{n,T,X}(y-,x)$ be the empirical bivariate d.f.'s of the observed pairs (Y,X) and (T,X) respectively and define the bivariate risk set at time (u,v) as

$$\begin{aligned}nC_{2,n}(u,v) &= \#\{i : Y_i \le u \le T_i\ ,\ X_i \le v\}\\ &= n[F^*_{n,Y,X}(u,v) - H_{n,T,X}(u-,v)].\end{aligned}$$

Then we have the following estimator for the reverse-hazard vector,

$$\begin{aligned}\tilde{\Lambda}_n(u,v) &= \{\frac{F^*_{n,Y,X}(\nabla u,\nabla v)}{C_{2,n}(u,v)}, \frac{F^*_{n,Y,X}(\nabla u,v)}{C_{2,n}(u,v)}, \frac{C_{2,n}(u,\nabla v)}{C_{2,n}(u,v)}\}\\ &\equiv \{\Lambda_{12,n}(\nabla u,\nabla v), \Lambda_{1,n}(\nabla u,v), \Lambda_{2,n}(u,\nabla v)\}\end{aligned} \tag{8}$$

Where, for any bivariate function W, $W(\nabla u,v) = W(u,v) - W(u-,v)$ and $W(u,\nabla v) = W(u,v) - W(u,-v)$. Remark here that the reverse-hazard vector describes the identifiable components of the truncation model and it is also of independent interest, since it describes the univariate and bivariate failure behavior in the immediate past. See eg., Pons(1986) and Gross and Carol(1992) for hazard based inferences and further applications.

2.3 Estimators of $F(y,x)$

We now present the three estimators which are right truncation counterparts of the estimators studied in Gürler(1994). The first one presented below is motivated by the relations given in (3) and (4)

$$F_n^1(y,x) = \frac{1}{n}\sum_i \frac{F_{Y,n}(Y_i)}{C_n(Y_i)} I(Y_i \le y, X_i \le x).$$

The simple structure of the above estimator is very appealing in comparison to the alternative estimators described below. Stute(1993b) and Lin and Ying(1993) obtain estimators of similar structure for censored observations with uncensored covariate, and for bivariate observations both censored by the same variable respectively. For the latter case, Lin and Ying report a very good performance for practical sample sizes. This result agrees with our findings about $F_n^1(y,x)$ in the simulations. We will discuss more about this estimator after we present the alternative ones. The relation given in (6) suggests two other estimators as given below. For i=2,3

$$F_n^i(y,x) = F_{X,n}(x)F_{Y,n}(y)\Theta_n^i(y,x),$$

where, $F_{X,n}(x)$ and $F_{Y,n}(y)$ are the marginals of $F_n^1(y,x)$, and

$$\Theta_n^2(y,x) = exp\{-\Lambda_n(y,x)\},$$

with

$$\begin{aligned}
\Lambda_n(y,x) &= \sum_{u>y}\sum_{v>x}[\Lambda_{1,n}(\nabla u, v)\Lambda_{2,n}(u,\nabla v) - \Lambda_{12,n}(\nabla u,\nabla v)] \\
&= \sum_{u>y}\sum_{v>x}\frac{1}{C_{2,n}^2(u,v)}[C_{2,n}(u,v-)F_n^*(\nabla u, v) - C_{2,n}(u,v)F_n^*(\nabla u, v-)] \\
&\equiv \sum_{u>y;v>x} h_2(u,v).
\end{aligned}$$

and,

$$\begin{aligned}
\Theta_n^3(y,x) &= \prod_{u>y;v>x}\frac{1-\Lambda_{1,n}(\nabla u,v) - \Lambda_{2,n}(u,\nabla v) + \Lambda_{12,n}(\nabla u,\nabla v)}{[1-\Lambda_{1,n}(\nabla u, v)][1-\Lambda_{2,n}(u,\nabla v)]} \\
&= \prod_{u>y;v>x}\frac{C_{2,n}(u,v)}{C_{2,n}(u,v-)}\frac{[C_{2,n}(u,v-) - F_n^*(\nabla u, v-)]}{[C_{2,n}(u,v) - F_n^*(\nabla u, v)]} \\
&\equiv \prod_{u>y;v>x} h_3(u,v)
\end{aligned}$$

It can easily be verified that $F_n^1(y,x)$ is a bivariate distribution function and its marginal for Y reduces to $F_n(y)$ given in (1). The other two estimators have the undesirable property that they may assign negative mass. Despite this disadvantage, in the simulations presented in Gürler(1994), the analogue of $F_n^2(y,x)$ is reported to perform slightly better. Dabrowska(1988) estimator for censoring corresponds to $F_n^3(y,x)$, and Pruitt(1991) demonstrates exactly the type of data points for which this estimator assigns negative mass. For censoring, he showed that the number of points which are assigned negative mass increases as n^2 and the total amount of negative mass does not disappear as $n \to \infty$. In the present case, the mass assigned by the last two estimators can be written as:

$$F_n^i(\nabla y, \nabla x) = \Theta_n^i(y,x)[D_n^{1i}(y,x)D_n^{2i}(y,x) - D_n^{3i}(y,x)] \qquad i = 2,3 \tag{9}$$

where

$$D^{12}(y,x) = F_{X,n}(x) - F_{X,n}(x-)exp\{\sum_{u>y} h_3(u,x)\}$$

$$D^{22}(y,x) = F_{Y,n}(y) - F_{Y,n}(y-)exp\{\sum_{v>x} h_3(y,v)\}$$

$$D^{32}(y,x) = F_{X,n}(x-)F_{Y,n}(y-)exp\{\sum_{v>x} h_3(y,v) + \sum_{u>y} h_2(u,x)\}(1 - exp(h_2(y,x))$$

and

$$D^{13}(y,x) = F_{X,n}(x) - F_{X,n}(x-)\prod_{u>y} h_2(u,x)$$

$$D^{23}(y,x) = F_{Y,n}(y) - F_{Y,n}(y-)\prod_{v>x} h_2(y,v)$$

$$D^{33}(y,x) = F_{X,n}(x-)F_{Y,n}(y-)\prod_{v>x} h_2(y,v).\prod_{u>y} h_2(u,x)(1 - h_3(y,x))$$

An inspection of (9) implies that both of the estimators assign nonnegative mass to the actually observed data pairs. Moreover, a detailed analysis shows that $F_n^3(y,x)$ tends to assign more negative mass than $F_n^2(y,x)$. Demonstration of such results is avoided here since it would be quite space demanding. Instead however, we present in Section 3, the simulation results which compare the proportion of the points to which negative mass are assigned. These results support the above claims. They also agree with the findings of Pruitt(1991), about the negative mass assigned by the Dabrowska(1988) estimator in the censoring case.

2.4 Large sample properties:

The large sample results for the above estimators can be obtained by employing similar methods with left truncation. We therefore provide only a sketch of the proof for the following theorem in the appendix. Moreover, since $F_n^1(y,x)$ is a bivariate d.f. and according to the simulation results, the alternative estimators are not doing any better, it seems natural to recommend to use this estimator. The theorem below therefore relates only to $F_n^1(y,x)$, but similar results could be obtained for the others. For $i = 1, \dots, n$, define

$$L_{i,n}(z) = \frac{I(Y_i > z)}{C(Y_i)} - \int_z \frac{I(Y_i \le z \le T_i)}{C^2(z)} F^*(dz)$$

and

$$\bar{L}_n(z) = \frac{1}{n}\sum_i L_{i,n}(z).$$

Theorem: Suppose $F(y,x)$ admits the density $f(y,x)$ and let $T_a = \{(y,x) : y > a > a_F;\ 0 < x < \infty\}$. Suppose also that $\int_{a_{F_Y}}^{b_{F_Y}} F(du)/[1 - G(u)]^2 < \infty$. Then, the following representation holds

$$\begin{aligned} F_n^1(y,x) - F(y,x) &= [F_n^*(y,x) - F^*(y,x)]\frac{F_Y(y)}{C(y)} \\ &- \int_0^y [F_n^*(u,x) - F^*(u,x)]d(\frac{F_Y(u)}{C(u)}) \\ &+ \int_0^y \frac{F_Y(u)}{C^2(u)}[C(u) - C_n(u) + \bar{L}_n(u)C(u)]F^*(du,x) + R_n(y,x) \\ &\equiv \bar{\xi}_n(y,x) + R_n(y,x) \end{aligned}$$

where

$$\sup_{(y,x)\in T_a} |R_n(y,x)| = O(\frac{log^3 n}{n}).$$

It can easily be seen that $\bar{\xi}_n(y,x)$ above is a sum of mean zero i.i.d. variables. The strong consistency, LIL and weak convergence to a two-time parameter Gaussian process then follows from the above theorem by standard arguments. In particular, let $\theta_n(y,x) = \sqrt{n}[F_n^1(y,x) - F(y,x)]$. Then we can write

$$\begin{aligned} \theta_n(y,x) &= \tilde{F}_n(y,x)\frac{\alpha}{\bar{G}(y)} - \int^y \tilde{F}_n(u,x)d(\frac{1}{\bar{G}(u)}) \\ &- \int^y \frac{\tilde{C}_n(u)}{C(u)}F(du,x) + \int^y \tilde{L}_n(u)C(u)F(du,x) \end{aligned}$$

where

$$\tilde{F}_n(y,x) = \sqrt{n}[F_n^*(y,x) - F^*(y,x)],$$

$$\tilde{C}_n(y) = \sqrt{n}[C_n(y) - C(y)], \quad \tilde{L}_n(y) = \sqrt{n}L_n(y)$$

It is seen that $F_n^1(y,x)$ is asymptotically unbiased and the asymptotic variance of it can be obtained via the covariance structures given below, which we don't pursue here for space considerations. A proof for the following lemma is illustrated in the appendix.

Lemma : Suppose $\int F(dz)/\bar{G}(z) < \infty$. Then, for $a_{F_Y} < u, v < b_{F_Y}$, we have

$$Cov(\tilde{F}_n(u,x), \tilde{F}_n(v,x)) = F^*(u \wedge v, x)[1 - F^*(u \vee v, x)]$$

$$Cov(\tilde{F}_n(u,x), \tilde{C}_n(v)) = C_2(u \wedge v, x)\frac{\bar{G}(v)}{\bar{G}(u \wedge v)} - C(v)F^*(u,v)$$

$$Cov(\tilde{C}_n(u), \tilde{C}_n(v)) = C(u \wedge v)[\frac{\bar{G}(u \vee v)}{\bar{G}(u \wedge v)} - C(u \vee v)]$$

$$Cov(\tilde{L}_n(u), \tilde{L}_n(v)) = \int_{u \vee v} \frac{1}{C(z)F_Y(z)} F_Y(dz)$$

$$Cov(\tilde{F}_n(u,x), \tilde{L}_n(v)) = F(u,x) - \frac{F(u \wedge v, x)}{F_Y(v)}$$

and

$$Cov(\tilde{C}_n(u), \tilde{L}_n(v)) = \begin{cases} F_Y(u)[1 - \frac{1}{F_Y(v)}] & \text{if } v \geq u \\ \int_v^u \frac{1}{F_Y(z)}\{(1 - \frac{\bar{G}(u)}{\bar{G}(z)})F(dz) - F(dz,u)\} - \bar{F}(u) & \text{if } v < u \end{cases}$$

2.5 Alternative methods:

Before closing this section we would like to discuss briefly some alternative methods of estimation. We will present below two such estimators, since their analogs for the censored data are studied extensively in the literature. See eg. Campbell(1981), Campbell and Foldes(1982), and Lo and Wang(1989), and more recently Lin and Ying(1993), and Tsai and Crowley(1994). The first one simply follows from the relations given by (3) and (7):

$$F(y,x) = F_Y(y)\, C_2(y,x)/C(y),$$

which could be estimated by

$$F_n(y,x) = F_{Y,n}(y)\, \frac{C_{2,n}(y,x)}{C_n(y)}. \tag{10}$$

This estimator is similar in spirit to the one suggested in Campbell(1981) which is based on a decomposition of the bivariate d.f. to a marginal and a conditional component. The other one is based on a path dependent line integral of the univariate hazard function as described below. For any differentiable function $h(.) \in R^k$, let $h^{(i)}(.)$ statnd for the partial derivative w.r.t. i'th component,

$i = 1, \ldots, k$. Then, letting $R(y,x) = -\log F(y,x)$, we have

$$\begin{aligned} R(y,x) &= -\int_y^\infty R^{(1)}(u,\infty)du + \int_x^\infty R^{(2)}(y,v)dv \\ &= \int_y^\infty \frac{dF_Y(u)}{F_Y(u)} + \int_x^\infty \frac{F^{(2)}(y,v)dv}{F(y,v)} \\ &= \int_y^\infty \frac{dF_Y^*(u)}{C(u)} + \int_x^\infty \frac{C_2^{(2)}(y,v)dv}{C(y,v)} \end{aligned}$$

This suggests another possible estimator as below:

$$F_n(y,x) = exp\{-R_n(y,x)\} \tag{11}$$

with

$$R_n(y,x) = \sum_{i:Y_i>y} [nC_n(Y_i)]^{-1} + \sum_{j:X_j>x} [nC_{2,n}(y,X_j)]^{-1}.I(Y_j \le y \le T_j).$$

Remarks:

1. The estimators (10) and (11) are non-decreasing in x, but not so in the y component. They could also assign negative mass as the other estimators. Further, in the preliminary simulations, these estimators were dominated significantly by F_n^1. The following comparison could give an idea: consider the ratio of the estimated MSE for F_n^1 to that of F_n^i, i=2,3. For the cases considered in the simulations of the next section, the minimum of such ratios were 0.996 indicating the similar performance of the estimators. However for the alternative estimators presented above, the minimum for the better one was as low as 0.88 and the best about 0.96 (Details of these simulations could be obtained from the author). We therefore did not further consider these estimators.

2. The extension of the recent Prentice and Cai(1992) estimator to truncation is also possible via the relations provided in this paper. As remarked in van der Laan(1994) however, this estimator performs quite similar to the Dabrowska(1988) estimator. We therefore do not expect significant gain from such an alernative. On the other hand, the efficient estimation methods suggested by van der Laan(1994) do not seem to be immediately applicable to the truncated data. Such an extension needs extra work and is worthwhile to consider. A generalization of the above 'path-dependent' estimators is considered by Tsai and Crowley(1994.

3 Simulation Study

For the simulations presented in this section, the estimators are compared with respect to two performance measures: avarage squared error(squared bias and variance), and avarage sup norm. These quantities are evaluated at the cartesian product of 40 grid points at each axis. To avoid being misled by the extra noise at the tails, this evaluation is done over a region, where 95% of the total probability lies. The following models for (Y, X) and T are considered. **Model 1**: Y, X independent, each $\sim$ exp(1), and $T \sim \exp(\mu)$. **Model 2**: (Y, X) : Bivariate Exponential, ie., $Y = \min(U_1, U_{12})$ and $X = \min(U_2, U_{12})$ where U_1, U_2, U_{12} are independent exp(1), and $T \sim \exp(\mu)$. The parameter μ of the T variable is adjusted to obtain different values for α. The results are summarized in Tables 1 and 2. The reported α's in Table 1 are the average proportion of observable pairs in 1000 replications. In Table 1, F_n^o corresponds to the bivariate empirical d.f., applied to an untruncated sample for reference purposes. It is seen that F_n^3 has slightly better bias when there is heavy truncation, whereas F_n^1 has smaller variance. In terms of the MSE, F_n^1 is always dominating, but the difference is practically negligible. Observe for instance that the ratio of the MSE of F_n^1 to that of F_n^2 varies between .998 and .9995, and to that of F_n^3 between .996 and .999. When sup morm is considered, there are cases where F_n^1 is beated by F_n^2. From Table 2, for instance it is seen that out of six cases corresponding to light truncation($\alpha \simeq 0.75$), F_n^2 behaves at least as good in four cases. The ratio of the sup norms here change between .9939 and 1.005 for F_n^2 and from .973 to .998 for F_n^3. These figures also indicate that, as long as MSE or sup norm are considered, the estimators are practically equally good. This remark is also supported by the results of the samples generated from the above models. Such a realization is presented in Figure 1. As evidenced from the graphs, it is very hard to distinguish between the estimators, and this was the case in all samples which are not illustrated here.

As to the comparison of F_n^2 and F_n^3, Table 3 presents the the amount of negative mass and the proportion of points getting them, for the two estimators. It is seen that in almost all cases, F_n^2 is significantly better. For this estimator the amount of negative mass and the proportion of the points receiving such mass tend to zero as the truncation proportion decreases. However, this is not true for F_n^3. This result agrees with the results of Pruitt(1991)for the censored case and discourages the use of F_n^3.

From all the above discussion, one can conclude that F_n^1 is the most preferable estimator on the following grounds: **(i)** it is a proper bivariate d.f. **(ii)** In most of the simulated cases its performance is dominating, and **(iii)** It has a simpler computational form. We therefore used this estimator for the real data analysis of the next section. Finally consider the effect of the truncation proportion to the estimators. Again from Table 1, it is seen that, even after correcting for the truncation effect, the efficiency loss is quite large for $\alpha \simeq 0.25$ or 0.5. For $\alpha \simeq 0.75$, however the negative effect of truncation for all sample sizes seem to decrease significantly. To have an idea about the influence of the sample size and α, results of a typical sample from Model 1 is presented in Figure 2, for which the true d.f. is as given in Figure 1. Here it can be seen that the performance is quite promising even for $n = 50$, which could be considered rather small for R^2, and $\alpha \simeq 0.45$, indicating that less than a half of the population is observable.

4 Application to AIDS Data

We now present some applications of the proposed methods to TR-AIDS(Transfusion related AIDS) data. We used the data as given in Wang(1989), which is not reproduced here for space considerations. Earlier versions of this data has also appeared and analysed in Lui et al.(1986), Medley et al.(1987), Lagakos et al.(1988), Kalbfleisch and Lawless(1989), Wang(1989) and Gross and Carol(1992). Hence the purpose of this section is to provide some complimentary findings in light of the results of the previous sections, rather than providing a complete detailed analysis. For this data, Y is the incubation time(in months) as measured from the transfusion to the diagnosis of AIDS, X is the age(in years) of the individuals at the time of the study, and T is the time(in months), from the transfusion to the end of the study (July 1986). In the studies mentioned above, the analysis is done separately for three age groups: G_1: "children", aged 1-4; G_2: "adults", aged 5-59; and, G_3: "elderly", aged 60 or more. There are 34,120 and 141 patients respectively in these groups. This classification was done due to the consideration that immuno competence was low for very young and elderly patients, as mentioned by Kalbfleisch and Lawless(1989). This issue could be handled with the bivariate methods of this paper without splitting the data as such. Nevertheless we will present the results both for combined and separate groups. Note here that for this data, it is not plausible to assume that $a_{F_Y} \leq a_G$, $b_{F_Y} \leq b_G$, and one can only estimate

$$F_c(y,x) \equiv F(y,x)/F_Y(T_o)$$

where T_o can be taken as $\max(T_1, \ldots, T_n)$. It is not possible to estimate $F_Y(T_o) \equiv c$ from the available data set. Therefore the results presented below are based on $F_c(y, x)$ and $c = 1$ is employed for the illustration of the methods, as done in Kalbfleisch and Lawless(1989), where a discussion for the assessment of the value of c is also provided. For the analysis of this section, $F_n^1(y, x)$ is used as suggested by the results of the previous section. In Figure 3, estimated $F(y, x)$ is presented for the combined and the seperate groups. It is still insightful to treat the groups seperately, as it is seen that the shape of the bivariate d.f. is quite different particularly for the 'child' group. However also notice that the sample size for this group is very small relative to the other two. The concentration of the data around the ages 50 to 60 in G_2 is indicated by the steep rise in the d.f. From the joint distribution of the incubation time and the age at infection for(combined groups), the proportion of those infected in ages 1-4, 5-59 and 60+ are found to be 0.025, 0.607 and 0.368 respectively. This joint c.d.f. is then used to calculate the conditional d.f. of the incubation time(Y) given that $X \in G_i, i = 1, 2, 3$. The mean incubation time(in months) calculated from these conditional distributions were 21.9, 70.77 and 56.64 respectively for children, adults and elderly respectively. A quantity of more interest is the regression of Y on X. Gross and Carol(1992) and Kalbfleisch and Lawless(1991) applied proportional hazard models to this data set, considering the age group as the covariate. They have detected significant difference in the survival function of these groups. So far however, to the best of our knowledge, the effect of the age at infection on the incubation period has not been investigated. Consider the model $Y = m(x) + \epsilon$ where $m(x) = E[Y|X = x]$ is the regression function and ϵ is the mean zero error term. We provide below a rather informal approach to estimate $m(x)$ by nonparametric kernel regression methods based on the estimates of the bivariate d.f. we have provided. In particular, we use the following :

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i F_n^1(\nabla Y_i, \nabla X_i) K_b(x - X_i)}{\sum_{i=1}^{n} F_{X,n}^1(\nabla X_i) K_b(x - X_i)},$$

where, K is the Gaussian kernel and b=bw(in the figures) is the smoothing parameter, with

$$K_b(x - X_i) = \frac{1}{b} K(\frac{x - X_i}{b})$$

This estimator is analogous to Nadaraya-Watson regression estimators for the i.i.d. observations and the the large sample properties of it is studied in Gürler and Gijbels(1994) for LTRC data. Figure 4 illustrates the results for combined and seperate groups, with several choices of the smoothing parameter. The results for child and the elderly group strongly suggest an inreasing trend in the former and an almost constant in the latter. The results for the adult group, however is not much indicative of such trends. In fact, the early ages in the adult group behave more like the continuation of the 'child' group, and the later ones like the beginning of the 'elderly'. This observation becomes more notable when the results for the combined group are inspected, where the whole picture becomes more clear. The graph with $b = 5$ for instance suggests an increasing trend until the age about 40, after which a decline is observed until 60's and then the curve is more or less stabilized. There is a sign of a local decline around the age 20, but it is not clear if it is just an artifact or a real thing, since the data is quite sparse in that region. In particular, for the ages 5, 6, and 11 only a single patient was observed and the next observed age was 17. This structure of the data raises the need for local bandwidths, the choice of which is another delicate issue in particular for the truncated data. Nevertheless, this analysis suggests that the split of data in the way it is done previously is not quite natural, and it is more insightful either to treat the entire group or to redefine the groups so as to better capture the structure of the regression curve.

APPENDIX

Proof of Theorem:(sketch)

Let $\Lambda_Y(dy) \equiv \Lambda_1(dy, \infty)$ be the failure rate of Y, and $\Lambda_{Y,n}(dy)$ denote the empirical counterpart of it. Then we can write

$$\begin{aligned}
\Lambda_{Y,n}(y) - \Lambda_Y(y) &= \int_y^{b_{F_Y}} [\Lambda_{Y,n}(dz) - \Lambda_Y(dz)] \\
&= \int_y^{b_{F_Y}} \left(\frac{F^*_{Y,n}(dz)}{C_n(z)} - \frac{F^*_Y(dz)}{C(z)}\right) \\
&= \int_y^{b_{F_Y}} \frac{1}{C(z)}[F^*_{Y,n}(dz) - F^*_Y(dz)] - \int_y^{b_{F_Y}} \frac{[C_n(z) - C(z)}{C_1^2(z)]} F^*_Y(dz) + R_n(y) \\
&\equiv \bar{L}_n(y) + R_n(y)
\end{aligned}$$

where

$$R_n(y) = \int_y^{b_{F_Y}} \left\{ \frac{[C_n(z) - C(z)]}{C^2(z)} [F^*_{Y,n}(dz) - F^*_Y(dz)] + \frac{[C_n(z) - C(z)]^2}{C^2(z) C_n(z)} F^*_{Y,n}(dz) \right\}$$

From the above representation and two term taylor expansion, we get

$$F_{Y,n}(y) - F_Y(y) = F_Y(y)\bar{L}_n(y) + R_n(y),$$

where the order for $R_n(y)$ is obtained as in Stute(1993a). The representaion and the result for the bivariate d.f. then follows as in Gürler(1994).

Proof of Lemma

We will illustrate one of them, since the others follow similarly.

$$Cov(\tilde{C}_n(u), \tilde{L}_n(v)) = E\{I(Y \le u \le T)[\frac{I(Y > v)}{C(Y)} - \int_v^\infty \frac{1}{C^2(z)} I(Y \le z \le T)F^*(dz)]\}$$

$$= I(v < u)E\{\frac{I(v < Y \le u \le T)}{C(Y)}\} - \int_v^\infty \frac{1}{C^2(z)} E[I(Y \le u \wedge z, T \ge u \vee z)]F^*(dz)]\}$$

$= I + II$

Note that $F^*(dz) = \alpha^{-1}\bar{G}(z)F(dz)$ and $C(z) = \alpha^{-1}\bar{G}(z)F(z)$. Then for $u < v$

$$II = F(u)\int_v d(\frac{1}{F(z)}) = F(u)[1 - \frac{1}{F(v)}]$$

and for $u > v$

$$II = -\int_v^u \frac{C(u)}{F(u)C(z)} F(dz) - \bar{F}(u)$$

The result now follows, after observing

$$I = \int_v^u \int_u^\infty \frac{1}{C(z)} F^*(dz, dt) = \int_v^u \frac{\alpha^{-1}\bar{G}(z)}{C(z)}[F_Y(dz) - F(dz, u)].$$

References

[1] Campbell G.(1981), "Nonparametric Bivariate estimation with randomly censored data," *Biometrika*, 68, 417-42

[2] Campbell, G. and Foldes, A. (1982), "Large sample properties of nonparametric bivariate estimators with censored data," in *Proceedings, International Colloquia on Nonparametric Statistical Inference, Budapest, 1980,* 23-28. North-Holland, Amsterdam.

[3] Chao, M. T., Lo, S. H. (1988),"Some representations of the nonparametric maximum likelihood estimators with truncated data," *Anns. Statist.*, 16, 661-668.

[4] Dabrowska, D.M. (1988), "Kaplan-Meier estimate on the plane," *Anns. Statist.*, 16, 1475-1489.

[5] Gijbels I., and Wang, J.L.(1993),"Strong representations of the Survival function estimator for truncated and censored data with applications,"*J. Multivariate Anal.*, 47, 210-229.

[6] Gross, S.T., and Carol, C.H.(1992),"Regression models for truncated survival data,"*Scand. J Statist*, 19, 193-213.

[7] Gürler, Ü.(1994), "Bivariate distribution function when a component is randomly truncated" truncated data with applications," *Discussion Paper* 9407, Institut de Statistique, Universite Catholique de Louvain, Belgium.

[8] Gürler, Ü. and Gijbels I. (1994), " Kernel regression with left truncated and right censored data" (in preperation)

[9] Gürler, Ü., Stute, W., Wang, J-L. (1993),"Weak and strong quantile representations for randomly truncated data with applications," *Statist. and Prob. Letters*, 17, 139-148.

[10] Kalbfleisch, J.D., and Lawless, J.F. (1989), "Inference based on Retrospective ascertainment: An analysis of the data on transfusion-related AIDS," *J. Amer. Statist. Assoc.*, 84, 360-372.

[11] Kalbfleisch, J.D., Lawless, J.F. (1991), "Regression models for right truncated data with application to AIDS incubation times and reporting lags," *Statistica Sinica*, 1, 19-32.

[12] Kaminsky, K.S.(1987), "Prediction of IBNR claim counts by modelling the distribution of report lags," *Insurance Math. Econom.*, 6, 151-159.

[13] Keiding, N. and Gill, R.D.(1990),"Random truncation models and markov processes," *Anns. Statist.,* 18, 582-602.

[14] Lagakos, S.W., Barraj. L.M., DeGruttola, V.(1988), "Nonparametric analysis of truncated survival data with application to AIDS," *Biometrika*, 75, 515-523.

[15] Lai, T.L. and Ying, Z. (1991), "Estimating a distribution function with truncated and censored data," *Anns. Statist.*, 19, 417-442.

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

[17] Lo, S.H. and Wang, J.L. (1989), "I.I.D. representations for the bivariate product limit estimators and the bootstrap versions," *J. Multivariate Anal.*, 28, 211-226.

[18] Lynden-Bell, D.(1971), "A method of allowing for known observational selection in small samples applied to 3CR quasars," *Mon. Nat. Roy. Astr. Soc.*, 155, 95-118.

[19] Medley, G.F., Billard, L., Cox, D.R. and Anderson, R.M. (1988), "The distribution of the incubation period for the acquired immuno-deficiency syndrome(AIDS)," *Proc. Roy. Soc. London Ser. B* , 223, 267-277.

[20] Pons, O.(1986),"A test of independence between two censored survival times," *Scand J Statist.,* 13, 173-185.

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

[22] Pruitt, R.C.(1991)," On negative mass assigned by the bivariate Kaplan-Meier estimator," *Anns. Statist.,* 19, 443-453.

[23] Stute, W.(1993a),"Almost sure representations of the product-limit estimator for truncated data," *Anns. Statist.,* 21, 146-156.

[24] Stute, W.(1993b),"Consistent estimation under random censorship whencovariables are present," *J. Multivariate Anal.,* 45, 89-103.

[25] Tsai, W-Y., and Crowley, J .(1994),"A note on nonparametric estimators of the bivariate survival function under univariate censoring", *Tech. Report*

[26] Uzunoğulları, Ü., Wang, J-L. (1992a), "A comparison of the hazard rate estimators for left truncated and right censored data," *Biometrika*, 79, 297-310.

[27] Wang, M.C. (1989), "A semiparametric model for randomly truncated data," *J. Amer. Statist. Assoc.*, 84, 742-748.

[28] Wang, M. C., Jewell, N. P., Tsai W. Y. (1986), "Asymptotic properties of the product limit estimate under random truncation," *Anns. Statist.*, 14, 1597-1605.

[29] Woodroofe, M.(1985), "Estimating a distribution function with truncated data," *Anns. Statist.*, 13, 163-177.

Model	n	α		F_n^o	F_n^1	F_n^2	F_n^3
		0.20	sq.b	0	2713	2712	2708
			var.	332	4089	4092	4102
		0.51	sq.b	0	4	4	4
	50		var.	338	2453	2454	2456
		0.77	sq.b	0	0	0	0
			var.	343	520	520	521
1		0.26	sq.b	0	770	768	768
			var.	150	1333	1336	1337
		0.53	sq.b	0	5	5	5
	75		var.	150	1480	1482	1481
		0.73	sq.b	0	2	2	2
			var.	151	357	357	357
		0.20	sq.b	0	2421	2420	2417
			var.	221	3708	3712	3717
		0.50	sq.b	0	4	4	4
	50		var.	224	1682	1683	1684
		0.77	sq.b	0	0	0	0
			var.	225	370	370	370
		0.25	sq.b	0	669	668	667
2			var	105	1265	1269	1273
		0.53	sq.b	0	1	1	1
	75		var.	100	883	885	884
		0.73	sq.b	0	0	0	0
			var.	103	296	296	296

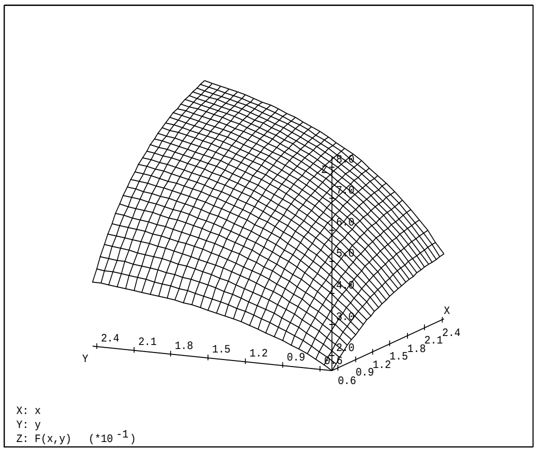
Table 1: Avarage squared bias and variance. 1000 replications, x10^{-5}.

Model	n	α	F_n^1	F_n^2	F_n^3
1	30	0.21	0.49145	0.49148	0.49223
		0.51	0.24434	0.24459	0.24614
		0.78	0.16196	0.16111	0.16302
	60	0.20	0.43521	0.43525	0.43669
		0.50	0.18247	0.18294	0.18416
		0.77	0.11253	0.11253	0.11333
	100	0.20	0.39834	0.39832	0.39953
		0.50	0.14625	0.14657	0.14732
		0.77	0.08702	0.08713	0.08749
2	30	0.26	0.39630	0.39845	0.40277
		0.54	0.22839	0.22882	0.23255
		0.74	0.15556	0.15485	0.15654
	60	0.25	0.34936	0.35087	0.35669
		0.53	0.16902	0.16966	0.17132
		0.73	0.11015	0.11014	0.11063
	100	0.25	0.31530	0.31732	0.32384
		0.53	0.13514	0.13549	0.13702
		0.73	0.08661	0.08677	0.08700

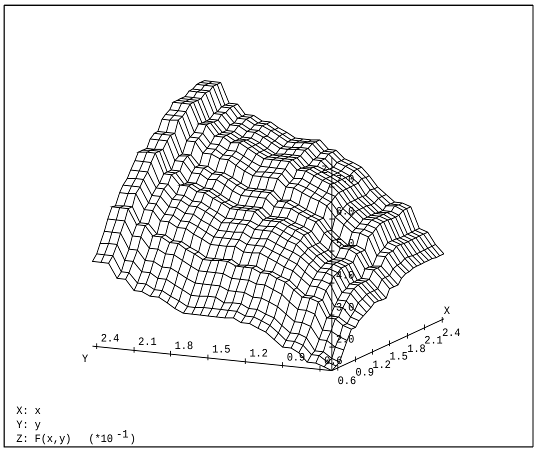
Table 2: Avarage sup $|F_n^i(y,x) - F(y,x)|$ over the grid points, 1000 replications, x10^{-5}.

Model	n	α	prop 2	amt 2	prop 3	amt 3
1	30	0.3660	0.2336	0.1972	0.2607	0.4298
		0.5956	0.2151	0.1584	0.3010	0.3500
		0.7160	0.1857	0.1090	0.3171	0.2546
		0.9103	0.1011	0.0286	0.3186	0.0922
	60	0.3618	0.2921	0.2849	0.2914	0.4738
		0.5918	0.2671	0.2202	0.3186	0.3657
		0.7165	0.2396	0.1575	0.3271	0.2643
		0.9093	0.1534	0.0437	0.3337	0.0896
2	30	0.4537	0.1850	0.1097	0.2751	0.3018
		0.6730	0.1571	0.0772	0.3068	0.2104
		0.8400	0.1046	0.0335	0.3132	0.1114
		0.9762	0.0187	0.0027	0.2929	0.0189
	60	0.4493	0.2533	0.1649	0.2997	0.3213
		0.6723	0.2149	0.1093	0.3237	0.2082
		0.8333	0.1582	0.0534	0.3204	0.1097
		0.9763	0.0431	0.0049	0.3217	0.0186

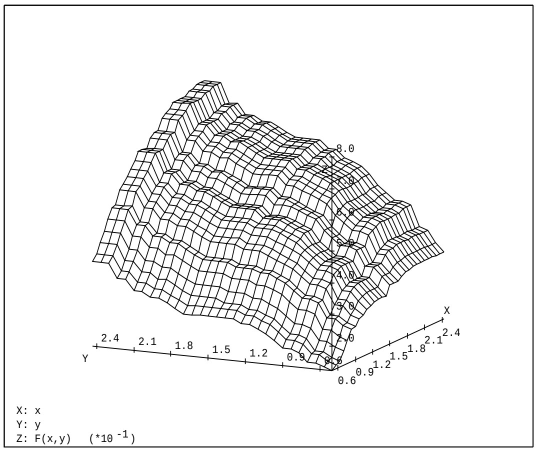
Table 3: prop i: proportion of negative mass assigned by F_n^i, amt i: total amount of negative mass assigned by F_n^i, i=2,3



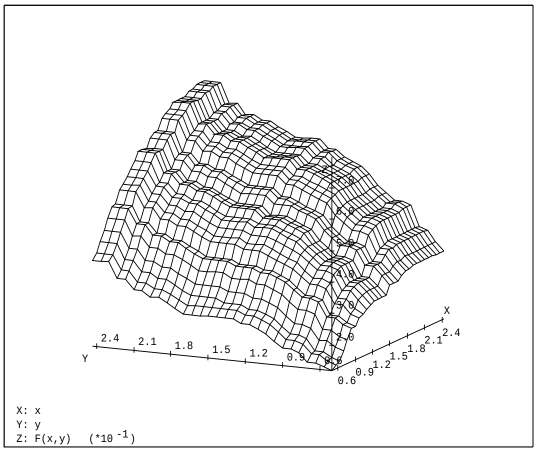
$F(y,x)$



$F_n^1(y,x)$



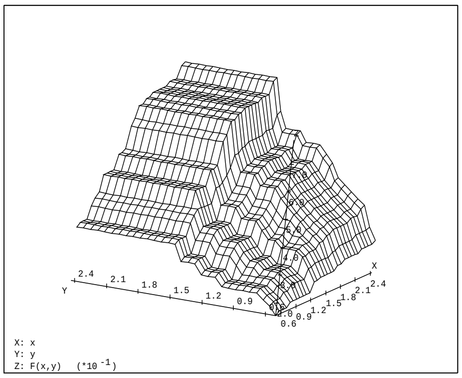
$F_n^2(y,x)$



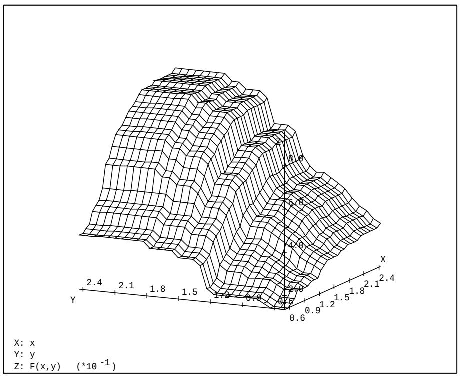
$F_n^3(y,x)$

Figure 1: Bivariate d.f. estimators. $\alpha = 0.75$, n=100; (Y,X): Independent exponential.

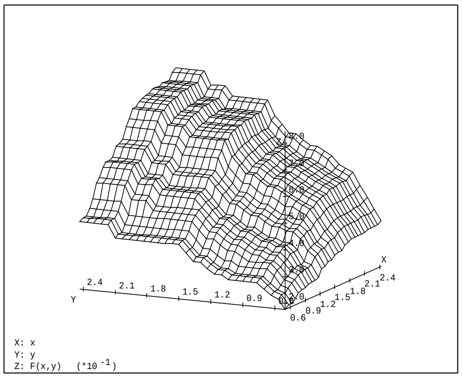
$\mu = 1.3, \alpha \simeq 0.45$ | $\mu = 1.0, \alpha \simeq 0.75$



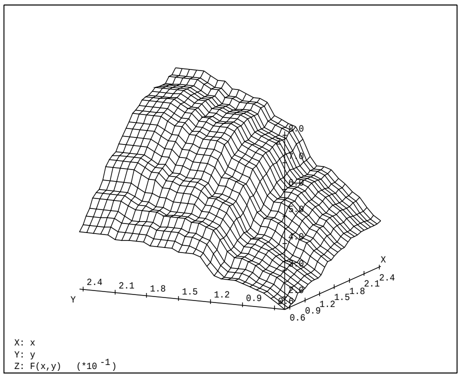
$n = 50$



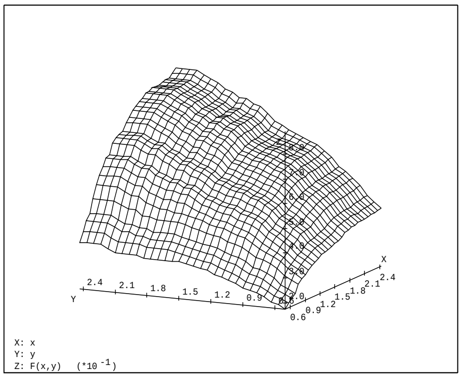
$n = 50$



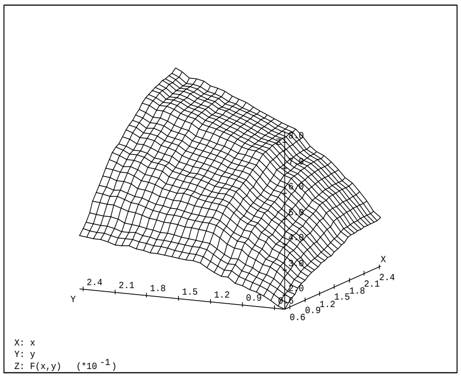
$n = 100$



$n = 100$

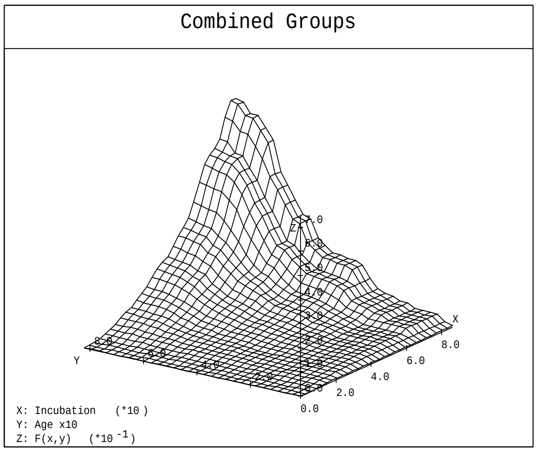


$n = 400$

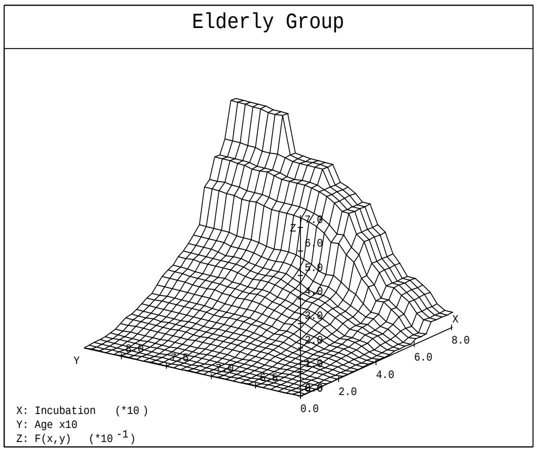


$n = 400$

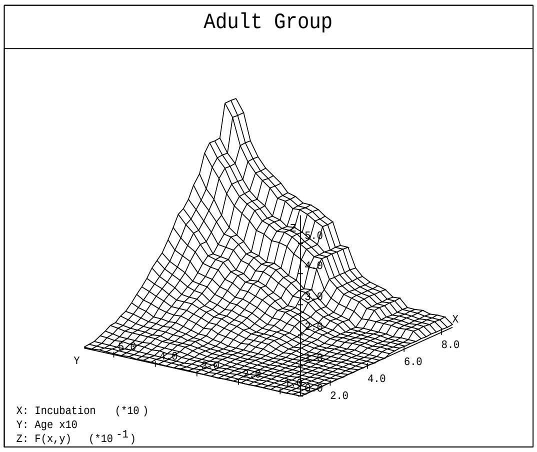
Figure 2: $F_n^1(y, x)$; (Y,X): Independent exponential.



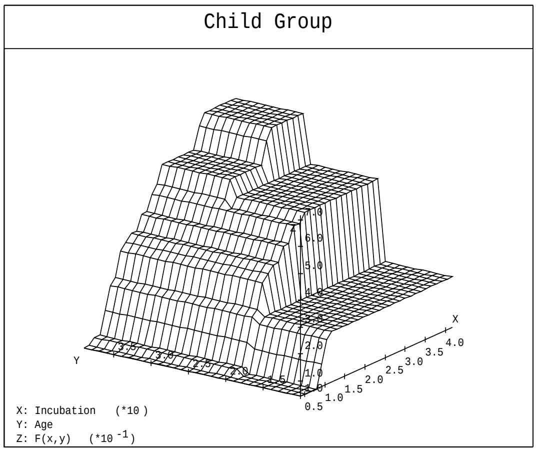
$n = 295$



$n = 141$



$n = 120$



$n = 34$

Figure 3: $F_n^1(y,x)$ for AIDS data.

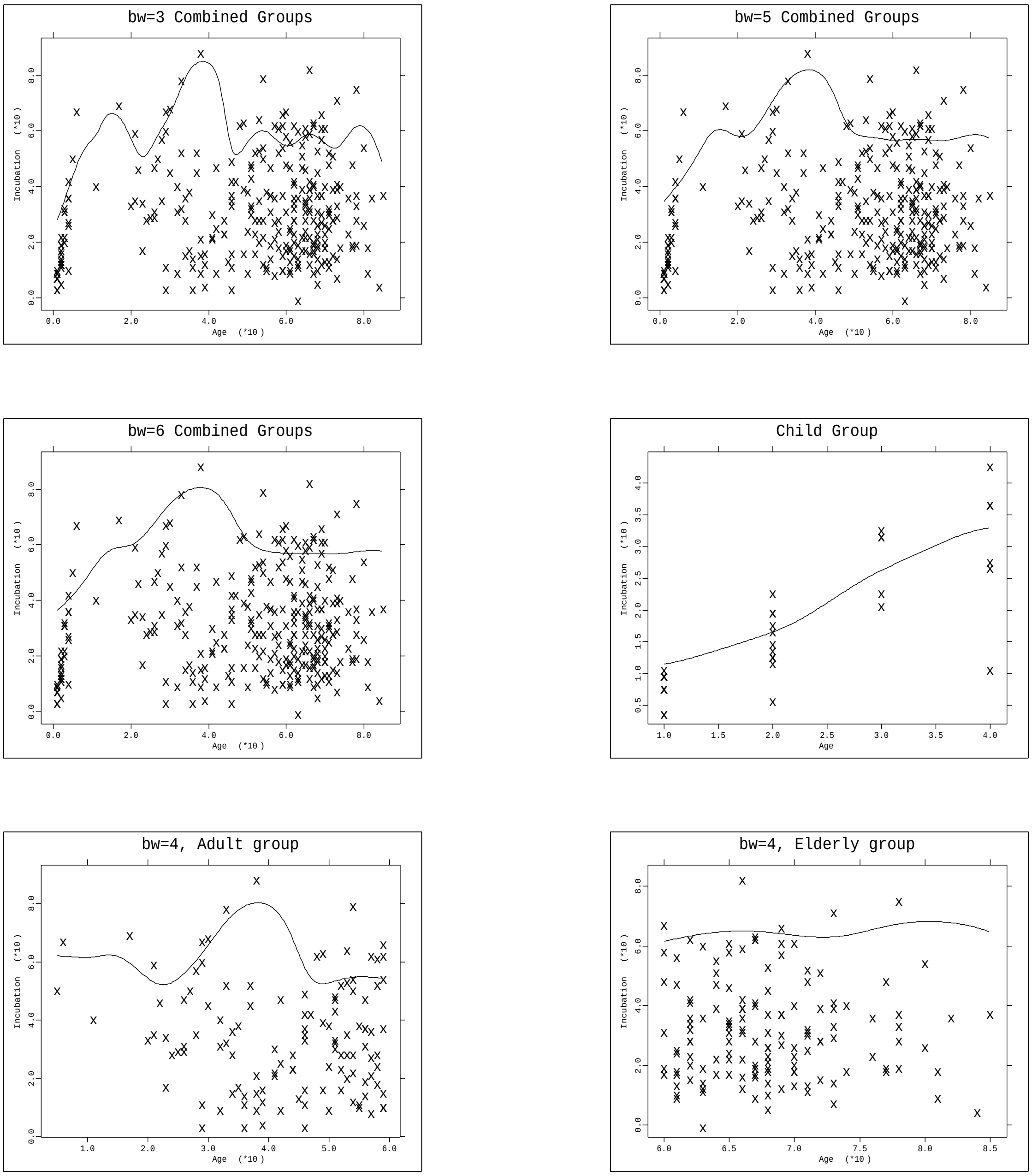


Figure 4: Kernel estimates of $E(Y|X = x)$ for combined and seperate groups.