A Nonparametric Least-Squares Estimation Method for Tumor Growth Function with Interval-Censored Observation

BY GANG CHENG, YING ZHANG

Department of Biostatistics, University of Iowa, Iowa City, IA 52242, USA

gang-cheng@uiowa.edu   ying-j-zhang@uiowa.edu

AND LIQIANG LU

School of Mathematical Science, Fudan University, Shanghai 200433, China

Shanghai Key Laboratory for Contemporary Applied Mathematics, Shanghai 200433, China

malqlu@fudan.edu.cn

SUMMARY

Study of tumor growth is an important area in cancer research. In this manuscript, we propose a nonparametric least-squares method to estimate tumor growth function for the case that tumor onset time is subject to interval censoring. Such scenarios constantly arise in animal tumorigenicity experiments and tumor screening programs, in which tumor onset time is only known within an interval made by adjacent screening times. The proposed estimator is shown asymptotically consistent under a specific metric using modern empirical process theory. Simulation studies are carried out to justify validity of the proposed method. Finally the method is applied to estimate breast tumor growth for the cohort of breast cancer patients in the state of Iowa of the United States using the data extracted from the Surveillance, Epidemiology and End Results (SEER) program.

Some key words: asymptotic consistency; interval censoring; nonparametric least-squares estimate; tumor growth/progression.

1. INTRODUCTION

Study of tumor growth is an important area in cancer research that has drawn a great attention among cancer researchers since 1970's when some worldwide tumor screening programs were developed aiming to design cancer intervention programs for reducing cancer incidence and cancer mortality. For example, women were invited to attend a breast cancer screening program in Nijmegen, Netherlands since 1975 in which serial screening mammograms were obtained to provide information on breast tumor growth (Peer et al., 1993). The knowledge of tumor growth will be critical in helping plan and evaluate such tumor screening programs effectively.

Study of tumor growth is also motivated by an animal tumorigenicity experiment (Albert & Shih, 2003) in which the objective was to compare tumorigenesis in mice injected with cell lines carrying different gene mutations in order to understand the effect of genetic mutations on tumorigenesis. In this experiment, five clones of mice were developed for each of two different gene mutations with five immunodeficient mice within each clone. These mice were scheduled to have follow-up screenings at 3 to 4-day intervals after injection. At each follow-up time, the mice were checked for the presence of a tumor and the volume of an existing tumor was measured.

Tumor growth was assessed based on these observed volumes to compare the tumor growth processes from different gene mutations.

Tumor growth should be regarded as a stochastic process due to variability in tumor progression among cancer patients. The key characteristic of tumor growth is the tumor growth function which is the expected tumor size as a function of time since the tumor onset. Various parametric models for tumor growth function have been adopted in studying tumor growth in literatures. Peer et al. (1993) calculated the growth rate defined as the tumor volume doubling time under the exponential growth model, using the tumor volumes measured from the serial mammography. Similar works were done by von Fournier et al. (1980), Norton et al. (1976) and Norton (1988) under the Gompertz growth function while Spratt et al. (1993b) used the logistic growth function to model the tumor growth. Spratt et al. (1993a) and Hart et al. (1998) summarised all those growth functions and compared the results among them. The aforementioned approach was referred to as "mathematical model approach" in Heitjan (1991), as the tumor growth function is modeled using some known mathematical functions. Non-linear regression models with random effects and autocorrelation were also utilised in Heitjan (1991) to ascertain predictors for tumor growth.

It is noted that tumor onset time is often unknown in study of tumor growth. However, it is assumed known in the methods described in the preceding paragraph mainly for mathematical convenience. Albert & Shih (2003) developed a method that jointly models tumor onset time and the growth function. Their method can be viewed as a latent variable approach: first, a discrete distribution function is used to estimate the distribution of tumor onset time (latent variable) under the framework of mixed-effects model; second, similar to Heitjan's approach, both linear and non-linear mixed-effects regression were used to model the tumor growth with tumor onset time and covariates as predictors; finally, integrating the latent variable out from both the mixed-effects tumor onset and the tumor growth models, a likelihood function can be easily formed to carry out the maximum likelihood analysis.

Though Albert-Shih's method nicely dealt with the issue of unknown tumor onset time, it is still an approach from the paradigm of parametric estimation method as the distribution of tumor growth needs to be assumed which is generally hard to justify in practice. In this manuscript, we propose to estimate the tumor growth function from the paradigm of nonparametric estimation method that not only deals with the unknown tumor onset issue but also does not need to assume the distribution of tumor growth. We will show both theoretically and numerically that the proposed method yields a consistent estimate of the true growth function.

The rest of manuscript is organised as follows: Section 2 presents the nonparametric least-squares method to estimate the tumor growth function; Section 3 describes the asymptotic consistency of the proposed estimator; Section 4 carries out simulation studies to justify validity of the proposed method; Section 5 applies the method to study breast tumor growth using the data extracted from the SEER program; Section 6 summarises the outcomes of this study and outlines the potential extension to other problems. Technical details are included in the Appendix.

2. A NONPARAMETRIC LEAST-SQUARES ESTIMATION METHOD

Let $(L, R]$ denote the random interval that contains tumor onset time T, i.e. $T \in (L, R]$. We consider the situation that there is only one measurement of tumor size and let Y denote the observed tumor size at a screening time $O \geq R$. In most of applications we encountered $O = R$. Here we allow the situation of $O > R$ for the sake of generality. Therefore, the observed data from a tumor growth process constitute

$$D = (L, R, O, Y). \tag{1}$$

In this paper, the stochastic process of tumor growth is assumed to be independent of tumor onset time, which is a common assumption made in the literatures. Suppose that $F(t)$ is the cumulative distribution function of tumor onset time. The following objective functional

$$LS(G(\cdot)) = E_{(Y,L,R,O)|L<T\leq R\leq O} \left(Y - H(O)\right)^2 ,$$

is constructed as the utility functional for deriving a nonparametric least-squares estimate of the tumor growth function, $G(t) = EY(t)$. Here $H(\cdot)$ is the expected size observed at time O and is a functional of G. It is easily shown that the functional $H(\cdot)$ that minimises $LS(G(\cdot))$ is given by

$$\begin{aligned} H(O) &= E_{Y|(L,R,O,L<T\leq R\leq O)} Y \\ &= E_{T|(L,R,O,L<T\leq R\leq O)} (E_{Y|(T,L,R,O,L<T\leq R\leq O)} Y) \\ &= E_{T|(L,R,O,L<T\leq R\leq O)} G(O-T) \\ &= \int_L^R G(O-t) \frac{1}{F(R)-F(L)} dF(t). \end{aligned}$$

Let a class of monotone nondecreasing functions $\mathcal{G}$ defined as follows,

$$\mathcal{G} = \{G : [0,\infty) \to [0,\infty), G(\cdot) \text{ is nondecreasing on } [0,\infty)\}. \tag{2}$$

then the true growth function $G_0(\cdot)$ satisfies

$$\begin{aligned} G_0(\cdot) &= arg \min_{G\in\mathcal{G}} LS\left(G(\cdot)\right) \\ &= arg \min_{G\in\mathcal{G}} E_{(Y,L,R,O)|L<T\leq R\leq O} \left(Y - \int_L^R G(O-t) \frac{1}{F(R)-F(L)} dF(t)\right)^2 , \end{aligned} \tag{3}$$

Let $D_i = (L_i, R_i, O_i, Y_i), i = 1, \ldots, n$ be the independent and identically distributed copies of D. If the cumulative distribution function of tumor onset time F is known, a nonparametric estimate of G_0, $\tilde{G}$ may be obtained by maximising the empirical version of the least-squares objective functional given by

$$\widetilde{LS}(G(\cdot)) = \sum_{i=1}^{n} \left(Y_i - \int_{L_i}^{R_i} G(O_i - t) \frac{1}{F(R_i) - F(L_i)} dF(t)\right)^2 . \tag{4}$$

In real applications, however, the distribution function of the onset time is usually unknown as exemplified in our motivating problems. We suggest to use a consistent estimate $\hat{F}$ of F to replace F in the above least-squares objective functional (4). The whole estimation procedure for G_0 is described as follows:

1. Obtain the nonparametric maximum likelihood estimate (NPMLE) $\hat{F}$ of the cumulative distribution function of tumor onset time by maximising the following likelihood function

$$L(F(t)) = \prod_{i=1}^{n} (F(R_i) - F(L_i)).$$

   using the iterative convex minorant algorithm given by Jongbloed (1998).

2. Construct the empirical version of the least-squares objective functional (3) with $\hat{F}$ replacing F for (4), that is

$$\widehat{LS}(G(\cdot)) = \sum_{i=1}^{n} \left(Y_i - \int_{L_i}^{R_i} G(O_i - t) \frac{1}{\hat{F}(R_i) - \hat{F}(L_i)} d\hat{F}(t) \right)^2 . \tag{5}$$

3. The nonparametric least-squares estimate $\hat{G}(\cdot)$ of the tumor growth function will be obtained by minimising (5), that is

$$\hat{G}(\cdot) = arg \min_{G \in \mathcal{G}} \widehat{LS} \left(G(\cdot) \right) .$$

In general, finding the minimiser of (4) for $G(\cdot)$ is a daunting job because $G(\cdot)$ appears in (4) in a convolution form. Nevertheless, the use of NPMLE of F facilitates a numerically manageable approach for estimating $G_0(\cdot)$ as the integral in (5) can be written as a finite sum with the NPMLE $\hat{F}$. As a step function, suppose the NPMLE $\hat{F}$ has jumps only at points in $S = \{s_1, s_2, \dots, s_k\}$ with $0 \le s_1 < s_2 < \cdots < s_k < \infty$, the ordered distinct observation times of the set $\{L_i, R_i, i = 1, 2, \dots, n\}$. Hence it results in (5) being rewritten as

$$\widehat{LS}(G(\cdot)) = \sum_{i=1}^{n} \left(Y_i - \sum_{s_j \in (L_i, R_i]} G(O_i - s_j) \frac{\hat{f}(s_j)}{\hat{F}(R_i) - \hat{F}(L_i)} \right)^2 , \tag{6}$$

where $\hat{f}(t)$ is the jump size of the NPMLE $\hat{F}(t)$ at time t. The nonparametric least-squares estimate (NPLSE) $\hat{G}(\cdot)$ of $G_0(\cdot)$ can be uniquely defined as a step function with jumps only at points in $V = \{v_1, v_2, \dots, v_l\}$ with $0 < v_1 < v_2 < \cdots < v_l < \infty$, the ordered distinct points of the set

$$\{(O_i - s_j) | L_i < s_j \le R_i, \text{ for } i = 1, 2, \dots, n; \ j = 1, 2, \dots, k\}$$

with nondecreasing constraints:

$$0 \le \hat{G}(v_1) \le \hat{G}(v_2) \le \cdots \le \hat{G}(v_l). \tag{7}$$

Let $\underline{G} = (G_1, G_2, \dots, G_l) \equiv (G(v_1), G(v_2), \dots, G(v_l))$. Then the objective functional (5) can be simplified to a quadratic form of $\underline{G}$

$$\widehat{LS}(\underline{G}) = A_1 - \underline{G}^T A_2 + \underline{G}^T A_3 \underline{G},$$

where $A_1 = \sum_{i=1}^{n} Y_i^2$, A_2 is the $l-$dimensional vector with the uth element given by

$$A_2[u] = \sum_{i=1}^{n} 2Y_i \frac{\hat{f}(O_i - v_u)}{\hat{F}(R_i) - \hat{F}(L_i)} \Delta_{i,u}, \text{ for } u = 1, 2, \dots, l,$$

and A_3 is the l by l symmetric matrix with the (u, u') entry given by

$$A_3[u, u'] = \sum_{i=1}^{n} \frac{\hat{f}(O_i - v_u)\hat{f}(O_i - v_{u'})}{(\hat{F}(R_i) - \hat{F}(L_i))^2} \Delta_{i,u} \Delta_{i,u'}, \text{ for } u, u' = 1, 2, \dots, l,$$

for which $\Delta_{i,u} = 1[L_i < O_i - v_u \le R_i \text{ and } O_i - v_u \in S]$ for $i = 1, 2, \dots, n, u = 1, 2, \dots, l$. Then this nonparametric estimation problem becomes a quadratic programming problem subject to the linear inequality constraints given by (7).

To calculate the NPLSE $\hat{G}$, either the iterative convex minorant algorithm (ICM) developed by Jongbloed (1998), or the projected Newton-Raphson algorithm developed by Cheng et al. (2011) can be applied. From our experiment, it appears that the projected Newton-Raphson algorithm converges much faster than the ICM algorithm and hence is adopted in our calculation.

3. Consistency of the Nonparametric Least-Squares Estimator

In this section, we state the consistency of the proposed NPLSE in a specific metric under some mild regularity conditions. As the NPLSE is a special case of M-estimation, the modern empirical process theory of M-estimation will be utilised throughout the technical arguments.

Let

$$m_{G,F}(D) = -\left(Y - \int_L^R G(O-t)\frac{1}{F(R)-F(L)}dF(t)\right)^2,$$

be a stochastic process indexed by functions G and F, where $G \in \mathcal{G}$ defined in (2) and $F \in \mathcal{F}$, a class of cumulative distribution functions. Let P and $\mathbb{P}_n$ be the underlying true probability and empirical measures, respectively, for the observed data $(D_1, D_2, \cdots, D_n)$. We define a deterministic bivariate functional $M(G,F)$ in the index set $\mathcal{G} \times \mathcal{F}$ as

$$M(G,F) = Pm_{G,F}(D)$$

and a random functional $\mathbb{M}_n(G)$ in the index set $\mathcal{G}$ as

$$\begin{aligned}\mathbb{M}_n(G) &= \mathbb{P}_n m_{G,\hat{F}}(D)\\ &= \frac{1}{n}\sum_{i=1}^{n} - \left(Y_i - \int_{L_i}^{R_i} G(O_i-t)\frac{1}{\hat{F}(R_i)-\hat{F}(L_i)}d\hat{F}(t)\right)^2,\end{aligned}$$

where $\hat{F}$ is the NPMLE of the cumulative distribution function F_0 of tumor onset time with interval-censored observations.

The NPLSE $\hat{G}$ can be therefore regarded as the M-estimator defined as

$$\hat{G}_n = arg \max_{G\in\mathcal{G}} \mathbb{M}_n(G).$$

The following regularity conditions are sufficient to warrant the consistency of the NPLSE $\hat{G}_n$:

Condition 1. The observation interval $(L, R]$ is sampled from $(L_0, R_0]$ with $0 \leq L_0 < R_0 < \infty$ and is separable in the sense that there exists a constants $\mu_0 > 0$ such that

$$P(R - L > \mu_0) = 1.$$

Condition 2. The underlying density function $f_0(t)$ of tumor onset time has a positive lower bound on $(L_0, R_0]$, i.e.

$$f_0(t) \geq \kappa_0 > 0, \text{ for } t \in (L_0, R_0].$$

Condition 3. The class for the tumor growth function $\mathcal{G}$ as defined in (2) is uniformly bounded, that is

$$\sup_{t\in[0,\tau)} G(t) \leq \nu_0, \text{ for } G \in \mathcal{G},$$

where $\nu_0 \in (0, \infty)$ and $\tau = O_0 - L_0$ is the study duration with O_0 being the latest monitoring time for tumor in the study.

Remark. Conditions 1 and 2 are reasonable in view of real applications and they directly imply that there exists a constant ξ_0 such that $F_0(R) - F_0(L) \geq \xi_0$ for the true cumulative distribution function $F_0(t)$ of tumor onset time. It further implies that there exists a constant $\rho_0 > 0$ such that $P\left(\hat{F}(R) - \hat{F}(L) \geq \rho_0\right) \to 1$ as sample size goes to infinity due to the fact that the NPMLE $\hat{F}$ converges uniformly to F_0 with probability 1 (Groeneboom & Wellner, 1992). Condition 3 is intuitively a reasonable assumption for the tumor growth function in cancer research.

A generalised L_2-norm $d(\cdot,\cdot)$ is defined for the class $\mathcal{G}$ as follow:

$$d(G_1, G_2) = \left(E_{(L,R,O)}\left(\int_L^R (G_1(O-t) - G_2(O-t))\frac{1}{F_0(R) - F_0(L)}dF_0(t)\right)^2\right)^{\frac{1}{2}}$$

$$\text{for any } \; G_1, G_2 \in \mathcal{G}.$$

THEOREM 1. *If Conditions 1-3 satisfy, the NPLSE defined as*

$$\hat{G}_n = \arg\max_{G\in\mathcal{G}} \mathbb{M}_n(G).$$

converges in probability to the true tumor growth function G_0 *in metric* $d(\cdot,\cdot)$*, that is,* $d(\hat{G}_n, G_0) \to_p 0$

The proof of Theorem 1 is given in the Appendix.

4. SIMULATION STUDY

Monte-Carlo Simulations with 1,000 repetitions for sample size $n = 100, 300,$ and 500 are employed, respectively, to study the asymptotic behavior of the proposed NPLSE. For each case, data $\{(L_i, R_i, O_i, Y_i), i = 1, \ldots, n\}$ are generated in the following manner: tumor onset time T_i is sampled from an Exponential distribution with mean 5; a series of inter-arrival screening times are generated independently from an Exponential distribution with mean 2 and the interval $(L_i, R_i]$ is selected from the adjacent screening times such that $L_i < T_i \leq R_i$. If T_i occurs before the first screening time, L_i is chosen to be 0 and R_i the first screening time. For this simulation study, the tumor is assumed observed at time $O_i = R_i$ and the tumor size Y_i at O_i is sampled from an Exponential distribution with mean $2(R_i - T_i) + 0.02$. The NPMLE $\hat{F}(t)$ and NPLSE $\hat{G}(t)$ are computed for each of the 1,000 repetitions. All computation tasks for the simulation study were performed with Intel Core 2 CPU 6600 @2.40GHZ and the computing software was developed for R 2.9.1.

In Figure 1, the three plots on the left column show the single NPMLEs of the cumulative distribution function of tumor onset time in three repetitions randomly selected from the Monte-Carlo simulation study and the three plots on the right column present the mean, 2.5 and 97.5 percentiles of the 1,000 NPMLEs of the cumulative distribution function of tumor onset time from the simulation study. The results clearly show that the NPMLE $\hat{F}(t)$ is asymptotically consistent and its variation decreases as sample size increases.

Figures 2 presents the simulation results for NPSLE $\hat{G}(t)$ when the Exponential distribution with mean of 2 is used for the independent inter-arrival screening times. Again, the three plots on the left column describe the single NPLSEs of the tumor growth function in three repetitions from the Monte-Carlo simulation study and the three plots on the right column give the mean, 2.5 and 97.5 percentiles of the 1,000 NPLSEs of the tumor growth function from the simulation study. The results demonstrate that the newly proposed NPLSE $\hat{G}(t)$ of the tumor growth function is asymptotically consistent and its variance decreases as sample size increases. However,

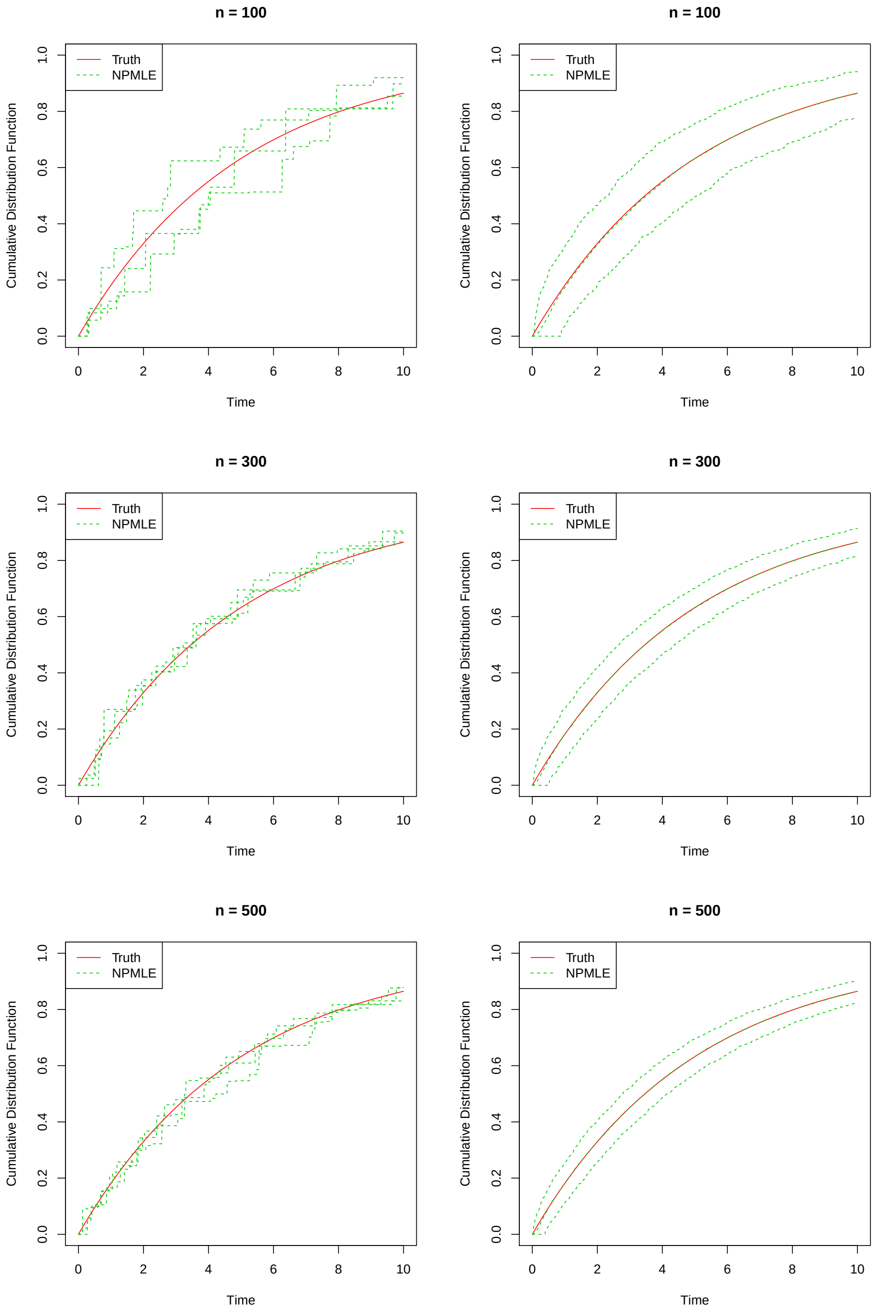

Fig. 1: The Monte-Carlo simulation study for the nonparametric maximum likelihood estimate of the cumulative distribution function of tumor onset time. Left panel: the NPMLEs in 3 random repetitions; right panel: the mean, 2.5 and 97.5 percentiles of the NPMLEs with 1,000 repetitions.

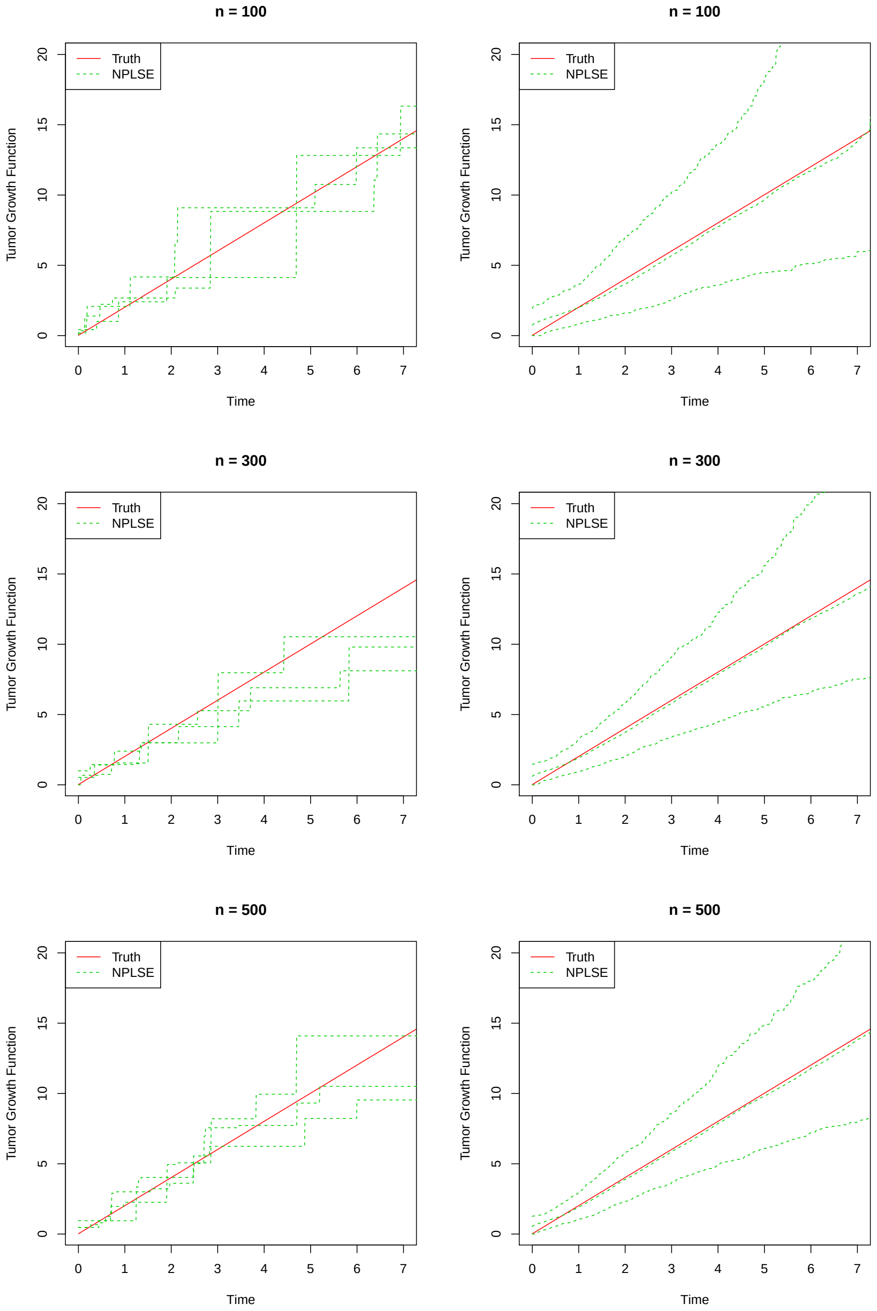

Fig. 2: The Monte-Carlo simulation study for the nonparametric least-squares estimate of the tumor growth function. Left panel: the NPLSEs in 3 random repetitions; right panel: the mean, 2.5 and 97.5 percentiles of the NPLSEs with 1,000 repetitions.

compared to the NPMLE $\hat{F}(t)$, it appears the NPLSE $\hat{G}(t)$ is more variable which may imply that $\hat{G}(t)$ converges to the true tumor growth function in a rate slower than $n^{1/3}$, the rate of convergence of $\hat{F}(t)$ (Groeneboom & Wellner, 1992).

5. Data Analysis

The proposed estimation method is applied to the breast tumor data obtained from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI). This program collects and publishes cancer incidence and survival data from population-based cancer registries covering over one quarter of the US population. The collected information by the SEER program includes patient demographics, primary tumor site, tumor stage and size at diagnosis, follow-up for survival status, etc. In this study, the following information for each case can be generated directly or indirectly from the original data: birth year, tumor diagnosis time accurate to month, tumor type and tumor size. Breast tumor is classified into four types (SEER, 2011) denoted as 0,1,2 and 4 to represent *in situ*, *localized*, *regional* and *distant* tumors, respectively. Tumor size is measured at diagnosis and is the largest diameter of the primary breast tumor in millimeters. Survival status of breast cancer subjects is also known and is accurate to month.

We study the cohort of white females in the state of Iowa who were alive in 1999. According to the registry data obtained from the SEER program, there were a total of 1,386,855 female subjects with age between 0 and 84 in 1999 and most of them were followed to December, 2008 with information available regarding their breast tumor incidence and survival status between 1973 and 2008. In this study, we focus on the cohort of Iowa white women who were at risk of breast cancer by the end of 1999. Thus 17,771 Iowa women who were diagnosed of having breast tumor prior to December 31, 1999 will be excluded from the study that results in a total of 1,369,084 women for the study cohort and among them 20,576 women had a breast tumor diagnosed between 2000 and 2008 for the first time.

5·1. *NPMLE of the Cumulative Distribution Function for Tumor Onset Time*

The focused study cohort of Iowa white women who were at risk of breast cancer on January 1, 2000 constitutes a sample of the special case in interval censored data for tumor onset time, current status data or interval-censoring case 1 data (Groeneboom & Wellner, 1992). Either $L_i = 0$ (left-censored case) for the ith subject who had a breast tumor diagnosed between January 1, 2000 and December 31, 2008 for the first time or $R_i = \infty$ (right-censored case) for the ith subject who had no breast tumor detected before December 31, 2008. We estimate the cumulative distribution function of breast tumor onset age in months. So the observation times R_i for left-censored case and L_i for the right-censored case are the number of months since birth to the time at diagnosis and the number of months from birth to December, 2008, respectively. Because only the birth year is available in the data set, the birth month is randomly sampled uniformly between 1 and 12 in order to obtain L_i and R_i in month. 1,000 NPMLEs of the cumulative distribution function of tumor onset age in month can be calculated with 1,000 repetitions of random samples for the birth month. The mean of the 1,000 NPMLEs is used as $\hat{F}$ for the subsequent analysis of tumor growth function. From $\hat{F}$ as shown in Figure 3, it appears there is almost no incidence for breast cancer prior to age of 20 years old (240 months after birth) and the incidence rate peaks between ages from 33 and 42 years old (400 to 500 months after birth) and ages between 50 and 67 years old (600 to 800 months after birth).

5·2. *NPLSE of the Tumor Growth Function*

The 20,576 Iowa white women from the aforementioned study cohort who had their first breast tumor diagnosed between January 1, 2000 and December 31, 2008 are available for the analysis of tumor growth. The subjects with the largest diameter of primary breast tumor greater than 50 mm are regarded as outliers and excluded from our analysis. The tumor growth functions for the four types of breast tumor, *in situ*, *localized*, *regional* and *distant*, are estimated, respectively, using the nonparametric least-squares estimation method developed in Section 2 with $\hat{F}$ given in Section 5·1.

Let $D_k = \{(L_{k,i}, R_{k,i}, O_{k,i}, Y_{k,i}), i = 1, \ldots, n_k\}$ denote the data sets obtained for the subjects with the kth tumor type at diagnosis for $k = 0, 1, 2, 4$. $Y_{k,i}$ is the tumor size measured at $O_{k,i} = R_{k,i}$. Because of the nature of registry data, information for tumor is only available when it was diagnosed. Hence $L_{k,i}$ could be naturally set at zero. However, we found that such working interval with the left end at zero for tumor growth will not generate a satisfactory outcome and the nonparametric least-squares estimation procedure basically results in a horizontal line at the mean tumor size for the tumor growth function. The poor estimation is largely due to the fact that the working interval for determining the age of tumor onset is too wide so that the nonparametric estimate of the conditional density for the age at tumor onset used in (6) tends to be flat which equally distributes the observed tumor size at any point in the interval. To avoid this phenomenon, we need to construct a tighter working interval that brackets the age of tumor onset.

With the intuition that a smaller tumor is expected to have a shorter growing time, we propose to construct a tighter working interval for each subject based on the individual tumor size and the growth rate for each tumor type, which may be estimated from the ordinary least-squares estimation method (OLSE) in an ad-hoc approach. Initially, we suppose that each observed tumor has the onset time within 60 months to the diagnosis as it is very unlikely that a breast tumor was undiagnosed for more than 5 years since the onset nowadays. So the estimated growth rate $\hat{\beta}_k$ for each type of breast tumor ($k = 0, 1, 2, 4$) can be obtained as the estimated slope from the OLSE model with the working interval fixed at 60 months. The estimated standard error in the OLSE model for the kth type of breast tumor is given by $\hat{\sigma}_k^2 = \sum_{i=1}^{n_k} \hat{e}_{k,i}^2/(n_k - 1)$, where

$$\hat{e}_{k,i} = Y_{k,i} - \int_{R_{k,i}-60}^{R_{k,i}} \hat{\beta}_k (R_{k,i} - t) \frac{d\hat{F}(t)}{\hat{F}(R_{k,i}) - \hat{F}(R_{k,i} - 60)}.$$

The results of OLSE for tumor growth rate of all four types are summarised in Table 1.

Then the working left end point of the observation interval for each subject is chosen as

$$L_{k,i} = R_{k,i} - \min((Y_{k,i} - e_{k,i})/\hat{\beta}_k, 60),$$

where $e_{k,i}$ is sampled from the normal distribution with mean 0 and variance $\hat{\sigma}_k^2$, and it is set as 0 if $Y_{k,i} - e_{k,i} < 0$.

With the working interval constructed as above, the NPLSEs of the tumor growth function for different types of breast tumor are computed and plotted in Figure 4. It appears that a *in situ* or *localized* tumor has a slower growth rate and a *regional* or *distant* tumor grows much rapidly.

6. DISCUSSION

In this manuscript, we developed a nonparametric estimation method aiming to study tumor growth when the exact information about tumor onset time is not available. The proposed method is shown consistent using modern empirical process theory. Extensive simulation studies are carried out to provide numerical evidence for validity of the method. This method is robust and can be generally used in estimation of a monotone function that characterises a underlying stochastic process for the case of incomplete observation that the starting time of the process is subject to interval censoring. Particularly, this method can be naturally applied to HIV/AIDS study for estimating the AIDS incubation time with both HIV infection and AIDS onset times being interval censored that has been widely researched in literatures by, for example, Padian et al. (1987), DeGruttola & Lagakos (1989), Gomez & Lagakos (1994), Gomez & Calle (1999), Tu (1995), Sun (1995) and Sun (1997).

Though not theoretically justified, the outcome from the extensive simulation study suggests that the rate of convergence of the proposed estimation procedure should not be high and is possibly well below $n^{1/3}$. The lower rate of convergence implies that a large sample is required for the method to yield a satisfactory estimate of the underlying tumor growth function.

The proposed method is developed for data structure $D = (L, R, O, Y)$, where $(L, R]$ provides the information for tumor onset time and Y is the only measurement for tumor size available at time O. This data structure fits many real scientific problems. Nonetheless, a more general data structure consisting of a random vector

$$D = \{(L, R, K, O_{K,1}, \ldots, O_{K,K}, Y_{K,1} \ldots, Y_{K,K}),$$

with K being the random number of measurements made for tumor progression and $Y_{K,j}$ the size measured at time $O_{K,j}$, is also seen in practice of cancer research, for example,Albert & Shih (2003). To accommodate this data structure, we may extend the proposed method by treating multiple measurements for the tumor progression on a same subject as independent observations of tumor size on different subjects. This will results in an ad-hoc approach for estimating the tumor growth function with low estimation efficiency. To improve the estimation efficiency, one has to account for the correlation structure among subsequent measurements of tumor size. The statistical model to be considered for possible correlations is not trivial and remains an open research problem for further investigation.

There is no formal nonparametric inference procedure for comparing tumor growth curves under the circumstance considered in this manuscript. Although a nonparametric permutation test based on the difference of the area under the estimated tumor growth curves can be constructed, the inference of this test procedure can be very time-consuming. The study of asymptotic distribution of a class of functionals of the proposed NPLSE may be similarly conducted as those in Zhang (2006) and Balakrishnan & Zhao (2009). The result can be potentially useful in developing a nonparametric test statistic for group comparison of tumor growth curves and hence is highly desired. It is, however, a technically challenging task as the rate of convergence of NPLSE is unknown and it remains an open question for future research.

Appendix: The proof of Theorem 1

We prove Theorem 1 by verifying the conditions for the general consistency theorem of M-estimation given by van der Vaart (1998) (Theorem 5·7). For this particular setting, we shall verify the following three conditions:

(a) $\sup_{G \in \mathcal{G}} |\mathbb{M}_n(G) - M(G, F_0)| \xrightarrow{P} 0$,

(b) $\sup_{G: d(G, G_0) > \epsilon} M(G, F_0) < M(G_0, F_0)$,

(c) The sequence of estimates $\hat{G}_n$ satisfying

$$\mathbb{M}_n(\hat{G}_n) \geq \mathbb{M}_n(G_0) - o_p(1).$$

using modern empirical process theory. Throughout the rest of paper, K is denoted as a universal constant that may vary from place to place.

To justify (a), we rewrite

$$\sup_{G\in\mathcal{G}}|\mathbb{M}_n(G)-M(G,F_0)|\leq\sup_{G\in\mathcal{G}}|\mathbb{M}_n(G)-M(G,\hat{F})|+\sup_{G\in\mathcal{G}}|M(G,\hat{F})-M(G,F_0)|. \qquad \text{(A1)}$$

and we show that both $\sup_{G\in\mathcal{G}}|\mathbb{M}_n(G)-M(G,\hat{F})|$ and $\sup_{G\in\mathcal{G}}|M(G,\hat{F})-M(G,F_0)|$ converge to 0 in probability. The first term in (A1) can be rewritten as

$$\begin{aligned}\sup_{G\in\mathcal{G}}|\mathbb{M}_n(G)-M(G,\hat{F})| &= \sup_{G\in\mathcal{G}}|\mathbb{P}_n m_{G,\hat{F}}(D)-Pm_{G,\hat{F}}(D)|\\ &= \sup_{m\in\mathcal{M}}|(\mathbb{P}_n-P)m_{G,\hat{F}}(D)|,\end{aligned}$$

where $\mathcal{M}=\{m_{G,\hat{F}}(D):\ G\in\mathcal{G}\}$. Hence to show $\sup_{G\in\mathcal{G}}|\mathbb{M}_n(G)-M(G,\hat{F})|$ converges to 0 in probability, it suffices to show that $\mathcal{M}$ is *P-Glivenko-Cantelli* when n is sufficiently large.

By Theorem 2·7·5 in van der Vaart & Wellner (1996), we have that for any $\epsilon>0$, $\log N_{[\,]}(\epsilon,\mathcal{G},L_1(P))\leq K\left(\frac{1}{\epsilon}\right)$. This implies that there exists a set of $\epsilon-$brackets

$$\{[l_i,u_i]: i=1,2,\ldots,exp(K\epsilon^{-1}), P(u_i-l_i)<\epsilon\}$$

and for any $G\in\mathcal{G}$, there exists some $i\in[1,exp(K\epsilon^{-1})]$ such that $l_i(t)\leq G(t)\leq u_i(t)$ for $t\in[0,\tau)$. We construct the following set of brackets for $\mathcal{M}$, $\{[m_i^l,m_i^u]: i=1,2,\ldots,exp(K\epsilon^{-1})\}$, where

$$m_i^l=-\left(Y^2-2Y\int_L^R l_i(O-t)\frac{1}{\hat{F}(R)-\hat{F}(L)}d\hat{F}(t)+\left(\int_L^R u_i(O-t)\frac{1}{\hat{F}(R)-\hat{F}(L)}d\hat{F}(t)\right)^2\right),$$

and

$$m_i^u=-\left(Y^2-2Y\int_L^R u_i(O-t)\frac{1}{\hat{F}(R)-\hat{F}(L)}d\hat{F}(t)+\left(\int_L^R l_i(O-t)\frac{1}{\hat{F}(R)-\hat{F}(L)}d\hat{F}(t)\right)^2\right).$$

It follows that

$$\begin{aligned}P(m_i^u-m_i^l) &= E_{(Y,L,R,O)}(m_i^u-m_i^l)=E_{(L,R,O)}E_{Y|(L,R,O)}(m_i^u-m_i^l)\\ &= E_{(L,R,O)}E_{Y|(L,R,O)}\left\{\int_L^R(u_i(O-t)-l_i(O-t))\frac{1}{\hat{F}(R)-\hat{F}(L)}d\hat{F}(t)\right.\\ &\quad\left.\left(2Y+\int_L^R(u_i(O-t)+l_i(O-t))\frac{1}{\hat{F}(R)-\hat{F}(L)}d\hat{F}(t)\right)\right\}\\ &= E_{(L,R,O)}\left\{\int_L^R(u_i(O-t)-l_i(O-t))\frac{1}{\hat{F}(R)-\hat{F}(L)}d\hat{F}(t)\right.\\ &\quad\left.\left(2E_{Y|(L,R,O)}Y+\int_L^R(u_i(O-t)+l_i(O-t))\frac{1}{\hat{F}(R)-\hat{F}(L)}d\hat{F}(t)\right)\right\}.\end{aligned}$$

We note that

$$E_{Y|(L,R,O)}Y=\int_L^R G_0(O-t)\frac{1}{F_0(R)-F_0(L)}dF_0(t),$$

where F_0 is the true cumulative distribution function of tumor onset time and G_0 the true tumor growth function. With Condition 3 given in Section 3, it can be made that for any $G\in\mathcal{G}$, $G(O-t)\leq\nu_0$ for $t\in(L,R]$. Because $[l_i,u_i]$ is a $\epsilon-$bracket for some $G\in\mathcal{G}$, for a sufficiently small ϵ, it is reasonable to let $l_i(O-t)\leq\nu_0, u_i(O-t)\leq 2\nu_0$ for $t\in(L,R]$. Hence it follows that $E_{Y|(L,R,O)}Y\leq$

$\nu_0 \int_L^R \frac{1}{F_0(R)-F_0(L)} dF_0(t) = \nu_0$ and

$$\int_L^R (u_i(O-t) + l_i(O-t)) \frac{1}{\hat{F}(R) - \hat{F}(L)} d\hat{F}(t) \leq 3\nu_0 \int_L^R \frac{1}{\hat{F}(R) - \hat{F}(L)} d\hat{F}(t) = 3\nu_0.$$

Therefore,

$$\begin{aligned} P(m_i^u - m_i^l) &\leq KE_{(L,R,O)} \left\{ \int_L^R (u_i(O-t) - l_i(O-t)) \frac{1}{\hat{F}(R) - \hat{F}(L)} d\hat{F}(t) \right\} \\ &\leq KE_{(L,R,O)} \left\{ \int_L^R (u_i(O-t) - l_i(O-t)) dF_0(t) \right\} = KP(u_i - l_i) \leq K\epsilon \end{aligned}$$

due to the arguments made in the **Remark** in Section 3 and the uniformly consistency of $\hat{F}$ shown by Groeneboom & Wellner (1992). This implies that $\log N_{[\,]}(\epsilon, \mathcal{M}, L_1(P)) \leq K\left(\frac{1}{\epsilon}\right)$ and hence $\mathcal{M}$ is indeed *P-Glivenko-Cantelli* by Theorem 19·4 of van der Vaart (1998).

For the second term in (A1), it can be shown that

$$\begin{aligned} |M(G, \hat{F}) - M(G, F_0)| &= \left| E_{(Y,L,R,O)}[(a-b)(2Y - a - b)] \right| \\ &\leq E_{(Y,L,R,O)}[|a-b||2Y - a - b|] = E_{(L,R,O)} E_{Y|(L,R,O)}[|a-b||2Y - a - b|] \\ &\leq E_{(L,R,O)} \left[|a-b| (E_{Y|(L,R,O)} Y + a + b) \right] \end{aligned}$$

where $a = \int_L^R G(O-t) \frac{1}{\hat{F}(R)-\hat{F}(L)} d\hat{F}(t) \leq \nu_0,$ $b = \int_L^R G(O-t) \frac{1}{F_0(R)-F_0(L)} dF_0(t) \leq \nu_0,$ and $c = E_{Y|(L,R,O)} Y = \int_L^R G_0(O-t) \frac{1}{F_0(R)-F_0(L)} dF_0(t) \leq \nu_0$. It follows that

$$|M(G, \hat{F}) - M(G, F_0)| \leq KE_{(L,R,O)}|a - b|.$$

With Conditions 1-3, we can show that

$$\begin{aligned}
E_{(L,R,O)}|a-b| &= E_{(L,R,O)}\left|\int_L^R G(O-t)\left(\frac{1}{\hat{F}(R)-\hat{F}(L)}d\hat{F}(t)-\frac{1}{F_0(R)-F_0(L)}dF_0(t)\right)\right| \\
&= E_{(L,R,O)}\left|\int_L^R G(O-t)\frac{1}{\hat{F}(R)-\hat{F}(L)}d\left(\hat{F}(t)-F_0(t)\right)\right. \\
&\quad \left.+\int_L^R G(O-t)\left(\frac{1}{\hat{F}(R)-\hat{F}(L)}-\frac{1}{F_0(R)-F_0(L)}\right)dF_0(t)\right| \\
&\leq E_{(L,R,O)}\left|\int_L^R G(O-t)\frac{1}{\hat{F}(R)-\hat{F}(L)}d\left(\hat{F}(t)-F_0(t)\right)\right| \\
&\quad +E_{(L,R,O)}\left|\int_L^R G(O-t)\frac{(F_0(R)-F_0(L))-(\hat{F}(R)-\hat{F}(L))}{(\hat{F}(R)-\hat{F}(L))(F_0(R)-F_0(L))}dF_0(t)\right| \\
&\leq KE_{(L,R)}\left|\int_L^R d\left(\hat{F}(t)-F_0(t)\right)\right| \\
&\quad +KE_{(L,R)}\left|\left((F_0(R)-F_0(L))-(\hat{F}(R)-\hat{F}(L))\right)\int_L^R dF_0(t)\right| \\
&\leq KE_{(L,R)}\left|(\hat{F}(R)-F_0(R))-(\hat{F}(L)-F_0(L))\right| \\
&\quad +KE_{(L,R)}\left|\left((\hat{F}(L)-F_0(L))-(\hat{F}(R)-F_0(R))\right)\right| \\
&\leq K\left(E_R\left|\hat{F}(R)-F_0(R)\right|+E_L\left|\hat{F}(L)-F_0(L)\right|\right).
\end{aligned}$$

Since

$$P\left\{\lim_{n\to\infty}\sup_{t\in\mathcal{R}}|\hat{F}(t)-F_0(t)|=0\right\}=1$$

as given in Gnoeneboom & Wellner (1992), using *Lebesgue's dominated convergence theorem* results in

$$E_R\left|\hat{F}(R)-F_0(R)\right| \to_p 0 \text{ and } E_L\left|\hat{F}(L)-F_0(L)\right| \to_p 0.$$

Therefore $E_{(L,R,O)}|a-b| \to_p 0$ and it follows that $|M(G,\hat{F})-M(G,F_0)| \to_p 0$. So (a) holds.

It is straightforward to show that

$$\begin{aligned}
M(G_0,F_0)-M(G,F_0) &= E_{(Y,L,R,O)}\left((Y-b)^2-(Y-c)^2\right) \\
&= E_{(L,R,O)}\left[(c-b)E_{Y|(L,R,O)}(2Y-b-c)\right] = E_{(L,R,O)}(b-c)^2 \\
&= E_{(L,R,O)}\left(\int_L^R (G(O-t)-G_0(O-t))\frac{1}{F_0(R)-F_0(L)}dF_0(t)\right)^2 \\
&= d^2(G,G_0).
\end{aligned}$$

This immediately justifies (b). Because the estimate $\hat{G}_n$ is obtained by maximising the objective function $\mathbb{M}_n(G)$ over the parameter space $\mathcal{G}$, so (c) automatically holds. Hence the proof is complete.

REFERENCES

ALBERT, P. S. & SHIH, J. H. (2003). Modeling tumor growth with random onset. *Biometrics* **59**, 897–906.

BALAKRISHNAN, N. & ZHAO, X. (2009). New multi-sample nonparametric tests for panel count data. *Annals of Statistics* **37**, 1112–1149.

Cheng, G., Zhang, Y. & Lu, L. (2011). Efficient algorithms for computing the non and semi-parametric maximum likelihood estimates with panel count data. *Journal of Nonparametric Statistics* **23**, 567–579.

DeGruttola, V. & Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to aids. *Biometrics* **45**, 1–11.

Gomez, G. & Calle, M. L. (1999). Non-parametric estimation with doubly censored data. *Journal of Applied Statistics* **26**, 45–58.

Gomez, G. & Lagakos, S. W. (1994). Estimation of the infection time and latency distribution of aids with doubly censored data. *Biometrics* **50**, 204–212.

Groeneboom, P. & Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. DMV Seminar Band 19, Birkhäuser Basel.

Hart, D., Shochat, E. & Agur, Z. (1998). The growth law of primary breast cancer as inferred from mammography screening trials data. *British Journal of Cancer* **78**, 382–387.

Heitjan, D. F. (1991). Generalized norton-simon models of tumor growth. *Statistics in Medicine* **10**, 1075–1088.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics* **7**, 310–321.

Norton, L. (1988). A gompertzian model of human breast cancer growth. *Cancer Research* **48**, 7067–7071.

Norton, L., Simon, R., Brereton, H. D. & Bogden, A. E. (1976). Predicting the course of gompertzian growth. *Nature* **264**, 542–545.

Padian, N., Marquis, L., Francis, D. P., Anderson, R. E., Rutherford, G. W., O'Malley, P. M. & Winkelstein, W. (1987). Male-to-female transmission of human immunodeficiency virus. *Journal of the American Medical Association* **258**, 788–790.

Peer, P. G. M., Dijck, J. A. A. M., Hendriks, J. H. C. L., Holland, R. & Verbeek, A. L. M. (1993). Age-dependent growth rate of primary breast cancer. *Cancer* **71**, 3547–3551.

SEER (2011). Seer research data record discription. Tech. rep., Surveillance Epidemiology and End Results, Maryland.

Spratt, J. A., von Fournier, D., Spratt, J. S. & Weber, E. E. (1993a). Decelerating growth and human breast cancer. *Cancer* **71**, 2013–2019.

Spratt, J. A., von Fournier, D., Spratt, J. S. & Weber, E. E. (1993b). Mammographic assessment of human breast cancer growth and duration. *Cancer* **71**, 2020–2026.

Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to aids studies. *Biometrics* **51**, 1096–1104.

Sun, J. (1997). Self-consistency estimation of distributions based on truncated and doubly censored survival data with applications to aids cohort studies. *Lifetime Data Analysis* **3**, 305–313.

Tu, X. M. (1995). Nonparametric estimation of survival distributions with censored initiating time, and censored and truncated terminating time: application to transfusion data for acquired immune deficiency syndrome. *Applied Statistics* **44**, 3–16.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak convergence and empirical processes: with applications to statistics*. New York: Springer-Verlag.

von Fournier, D., Weber, E., Hoeffken, W., Bauer, M., Kubli, F. & Barth, V. (1980). Growth rate of 147 mammary carcinomas. *Cancer* **45**, 2198–2207.

Zhang, Y. (2006). Nonparametric k-sample tests with panel count data. *Biometrika* **93**, 777–790.

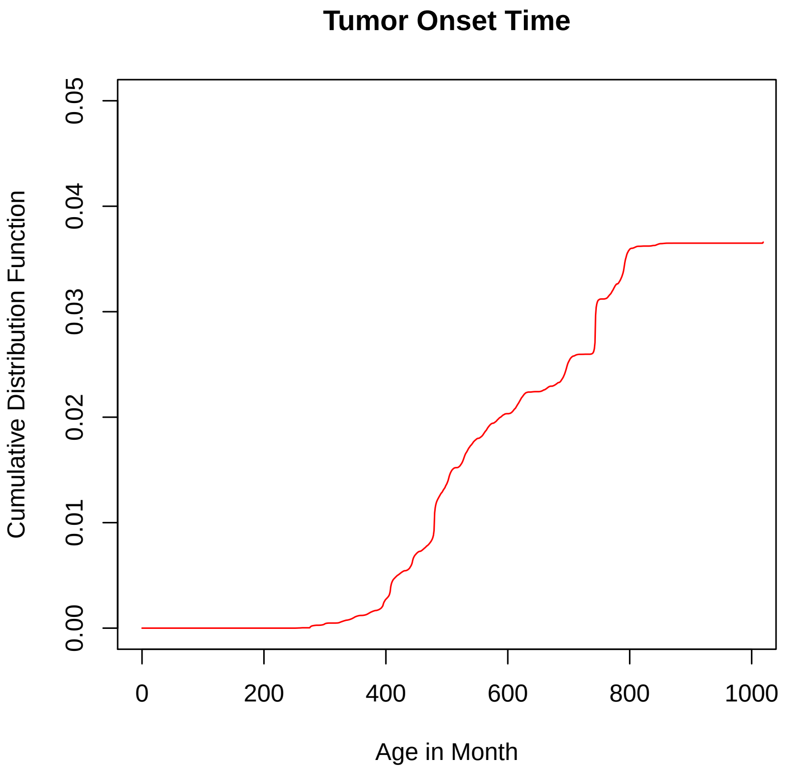

Fig. 3: The average of NPMLEs of the cumulative distribution function of breast tumor onset age in month.

Table 1: The Ordinary Least-Squares Estimates of Tumor Growth Rate

Type	$\hat{\beta}_k$	Std ($\hat{\sigma}_k$)	Expected Size with 60 month
In situ	0.2891	9.39	17
Localized	0.3847	9.69	23
Regional	0.6264	12.10	37
Distant	0.7127	13.18	42

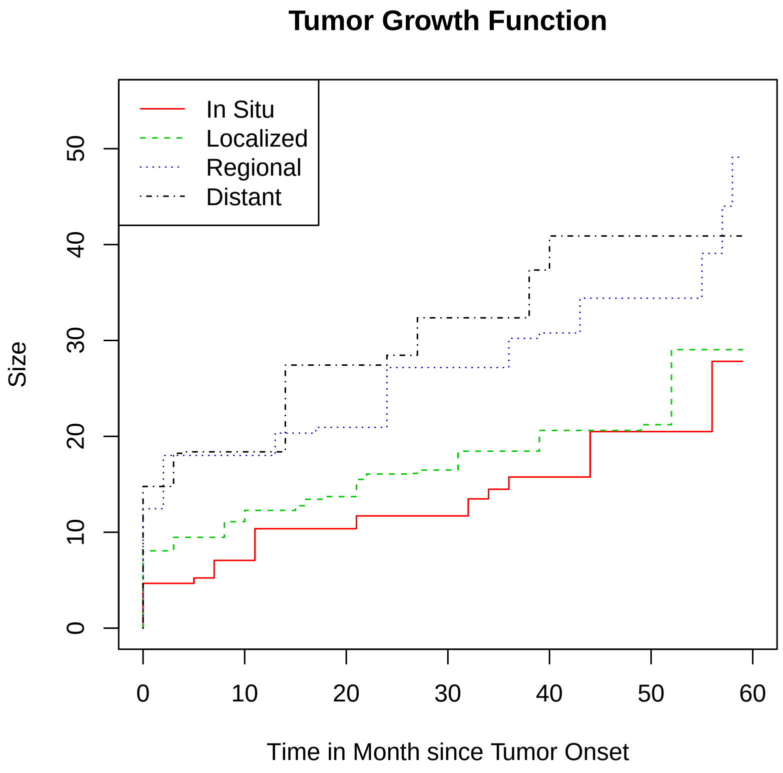

Fig. 4: The NPLSEs of breast tumor growth functions