TO: DR. GEORGE WONG

from: Yu, Qiqing

pages 4-10, total of 8 pages

Fax: 212-794-4958
Tel: 718-626-2860
Address: Dr. George Wong, 428 E 72nd Street, Strang Cancer Prevention Center, New York, NY 10021.

Dear George,

The following is a draft of the report. You can take a look and let me know how to revise it before I return to the U.S. If it is possible, we can have it done by telephone and I will ask my wife or my student Wang Jing to send you by the express mail the original version.

If you want to fax to me, use the following Fax number: 86-20-84039222 (which is China-Guangzhou-local)

(To Dr. Yu Qiqing via Dr. Wang Yanmin)

Yu qiqing,

Abstract:

The overall aim of our research proposal is to consider non-parametric estimation of the joint survival function $S(x_1, ..., x_d)$ =Pr$(X_1 > x_1, ..., X_d > x_d)$ of $d \geq 2$ intercorrelated time-to-event variables X_1, ..., X_d, each of which is subject to interval censoring. The generalized maximum likelihood estimator (GMLE) is the standard optimal procedure in survival analysis. However, it has not been studied for the multivariate interval-censored (IC) data in the literature. No closed-form expression for the GMLE is available, and asymptotic distribution theory for it is difficult. Our research efforts in the first year have been focused on (1) formulating a multivariate interval censorship model, called the case 2 multivariate interval censorship model, and establishing consistency and asymptotic normality under discrete assumptions; (2) resolving the identifiability problem in the case that the GMLE is not unique.

A. TABLE OF CONTENTS

B. INTRODUCTION

Interval-censored (IC) data are encountered in three areas of breast cancer research. The most common application is in clinical relapse follow-up studies in which the study endpoint is disease-free survival. When a patient relapses, it is usually known that the relapse takes place between two follow-up visits, and the exact time to relapse is unknown. In statistics, we say relapse time is interval censored. Interval censoring is also encountered in breast cancer registry studies in which information on family history of cancer is updated periodically. The Strang Breast Surveillance Program for women at increased risk for breast cancer, for instance, has enlisted over 800 women with complete pedigree information which is verified and updated continuously. Family history data such as age at diagnosis of a specific cancer, or a benign but risk-conferring condition, are obtained from each registrant at each update. Time to a cancer event, and definitely time to first detection of a benign condition, are at best known to fall in the time interval between the last update and age at diagnosis. A third but increasingly important area of application of interval censoring is in breast cancer chemoprevention experiments or prevention trials, which involve the observation of one or more surrogate endpoint biomarkers (SEB) over time. The scientific question of interest here is the estimation of time for the SEB to reach a target value, and time from cessation of intake of a chemopreventive agent to the loss of its protective effect. Unfortunately, the exact values of both these time variables are known only to lie in between two successive assay inspection times.

Let X denote a time-to-event variable with distribution $F(x) = Pr(X \leq x)$, or equivalently, survival function $S(x) = 1 - F(x)$. In interval censoring, X is not observed and is known only to lie in an observable interval (L, R). In our previous DOD funded grant, we have made fundamental contributions to both the theory of the generalized maximum likelihood (GML) estimation of S, and the computation in connection with the inference of GML estimator (GMLE) $\hat{S}$ of S.

Multivariate interval censoring involves $d \geq 2$ correlated X variables, each of which is subject to interval censoring. The main statistical concern here is the GML estimation of the joint survival function $S(x_1, ..., x_d) = Pr(X_1 > x_1, ..., X_d > x_d)$, and the correlations among the variables. Our interest in multivariate IC data is driven by needs arising from two related areas of breast cancer research at Strang. First, our investigators in the Strang Cancer Genetics Program want to study various patterns of familial aggregation of breast, ovarian and other forms of cancer using family history data from the Strang Breast Surveillance Program. Studies of familial early onset of breast cancer, breast-ovarian and breast-prostate associations will lead to multivariate IC data of high dimensions; therefore, a proper statistical procedure together with a feasible software to deal with such data are very much needed. Second, we are conducting a one-year chemoprevention trial of indole-3-carbinol (I3C) for breast cancer prevention. In this prevention trial we are monitoring the levels of two SEB's, a urinary estrogen metabolite ratio and a blood counterpart, both of which are subject to interval censoring. An earlier dose-ranging study of I3C conducted by Wong *et al* [1] has been published.

Statistical analysis of multivariate IC data has never been attempted. In the multivariate situation, modeling of the intercorrelated time-to-event variables and their dependency structure will require a great deal of innovative thinking; moreover, GML computation in realistic sample sizes can be prohibitively difficult.

The overall aim of this research proposal is to develop statistical inference for multivariate interval-censored data that are encountered in breast cancer chemoprevention trials employing multiple surrogate endpoint biomarkers, and in breast cancer registry follow-up studies of familial aggregation of breast and other forms of cancer. Asymptotic generalized maximum likelihood theory will be investigated and computer software package for maximum likelihood inference and Kaplan-Meier type survival plots will be implemented.

Key research accomplishments in the first year

- Completed task 1. In particular,
  We developed computer software packages for computing the GMLE.
  1.a. We developed software for identifying maximal intersections with multivariate IC data, the sofware is working but needs to be improved as it is still slow.
  1.b. We developed software for solving self-consistent equations to obtain SCE for general multivariate IC data.
- Completed task 2. In particular, we formulate a multivariate interval censorship model, called the case 2 multivariate interval censorship model. The GMLE of the distribution function is studied and its consistency and asymptotic normality are established under the case 2 multivariate interval censorship model and discrete assumptions on the censoring random vectors. This result is published in a peer-reviewed journal (see [2]).
- The GMLE of F with univariate IC data is uniquely defined at each follow-up time. However, this is no longer true in general with multivariate IC data. How to estimate the survival function and the covariance matrix of the estimator in such a case is a new practical issue in analyzing IC data. We propose a procedure in such a situation. Our method always results in a GMLE with a nonsingular sample information matrix. We also give a theoretical justification for such a procedure. This result is accepted by a peer-reviewed journal (see [7]).
- We developed software for implementing the above procedure.
- Revising papers initiated from the previous DOD grant (see [3], [4], [5] , [6]).

C. BODY

C.1. Multivariate case 2 model.

Consider nonparametric estimation of the joint survival function $S(x_1, ..., x_d) = \Pr(X_1 > x_1, ..., X_d > x_d)$ of $d \geq 2$ intercorrelated time-to-event variables X_1, ..., X_d, each of which is subject to interval censoring. For ease of presentation and without any loss of generality, we shall restrict our discussion to the bivariate case $\underline{X} = (X_1, X_2)$. There are several possible models for multivariate IC data. To start with, we had formulated the following simple model.

Let (U_i, V_i) denote the independent random censoring interval for X_i, and (L_i, R_i) denote the observable interval-censored (IC) data for X_i defined as

$$(L_i, R_i) = \begin{cases} [0, U_i] & \text{if } X_i \leq U_i, \\ (U_i, V_i] & \text{if } U_i < X_i \leq V_i, \\ (V_i, +\infty) & \text{if } X_i > V_i, \end{cases}$$

for $i = 1$, 2. This model is called a multivariate case 2 interval censorship model.

Let B_i denote any one of $[0, U_i]$, $(U_i, V_i]$ and $(V_i, +\infty)$. Therefore, a bivariate IC data point is a rectangular region in $\mathcal{R}^2$ taking one of the nine forms in $\mathcal{B} = \{B_k \times B_l : \ k, l = 1, 2, 3\}$. Given a sample of size n, the observations $(L_{i1}, R_{i1}, L_{i2}, R_{i2})$ can be represented by rectangle subsets $I_i \in \mathcal{B}$, for $i = 1$, ..., n. Define a maximal intersection (MI) A of the observable rectangles I_1, ..., I_n, to be a nonempty finite intersection of the I_i's such that $A \cap I_i = \emptyset$ or A, for each i. Let A_1, ..., A_m, denote the distinct maximal intersections with respect to I_1, ..., I_n.

The generalized likelihood function of S is given by $\Lambda_n = \mu_S(I_1) \times \cdots \times \mu_S(I_n)$, where $\mu_S(\cdot)$ is the probability measure induced by S. Wong and Yu [2] show that the GMLE $\hat{S}$, which maximizes L, must assign all the probability masses s_1, ..., s_m to A_1, ..., A_m. In general, $\hat{S}$ has to be obtained iteratively. Since $\hat{S}$ is also a self-consistent estimate (SCE), we can implement the SCE algorithm by solving for $\hat{s}_1$, ..., $\hat{s}_m$ in

$$s_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{ij} s_j}{\sum_{k=1}^{m} \delta_{ik} s_k},$$

$j = 1$, ..., m, where $\delta_{ij} = \mathbf{1}[A_j \subset I_i]$, $\mathbf{1}[\cdot]$ denoting the indicator function, and obtain an SCE of $S(\underline{x})$

$$\tilde{S}(\underline{x}) = \sum_{A_j \subset (x_1, +\infty) \times \cdots \times (x_d, +\infty)} \hat{s}_j.$$

With starting values $s_j^{(0)} = 1/m$ for all j, $\tilde{S}(\underline{x})$ is the GMLE at convergence.

We have established uniform consistency of $\hat{S}$ by proving

$$\Pr\{\lim_{n\to\infty} \sup_{\underline{x} \text{ is observable}} |\hat{S}(\underline{x}) - S(\underline{x})| = 0\} = 1$$

under the condition:

C1. S is arbitrary, the censoring vectors (U_1, V_1) and (U_2, V_2) take on countably many values.

C2. S is arbitrary, the censoring distribution G of (U, V) is continuous, and some regularity assumptions are imposed on either S or the censoring vector.

Asymptotic normality results are fundamentally important for confidence statements and hypothesis testing in data analysis. We want to prove that $\sqrt{n}(\hat{S}(\underline{x}) - S(\underline{x}))/\hat{\sigma} \xrightarrow{\mathcal{D}} N(0, 1)$, where $\hat{\sigma}^2$ is the inverse of the observed Fisher information number, or equivalently, $\hat{S}$ is both asymptotically normal and asymptotically efficient. We have established asymptotic normality for $\hat{S}$ in the following situation:

D1. S is arbitrary, (U_1, V_1) and (U_2, V_2) take on finitely many values, say $\underline{a_1}$, ..., $\underline{a_N}$, and $S(\underline{a_k}) > S(\underline{a_l})$, if $a_{k1} \leq a_{l1}$ and $a_{k2} \leq a_{l2}$ with at least one strict inequality, $\underline{a_k} = (a_{k1}, a_{k2})$ and $\underline{a_l} = (a_{l1}, a_{l2})$.

The above consistency and asymptotic normality results that we accomplished in our first year of research are published in a peer-reviewed statistical journal ([2]), except that the proof of the consistency under C2 is in a manuscript under preparing.

C.2. A new identifiability problem.

In our first year research, we encounted a new issue in the statistical analysis with multivariate interval censored data that does not occur in the univariate interval-censored data case. That is, the solution of the GMLE is not unique. Thus, we worked on solving this new problem.

When $d = 1$, it is well known that the GMLE $\hat{\mathbf{S}}$ of $\mathbf{S}$ $(=(s_1, ..., s_m)')$ is unique in the sense that the measure of an MI is uniquely determine by the GMLE. Under the assumption that all the random variables are discrete and take on finitely many values, it reduces to a parametric problem of estimating a multinomial distribution. Thus it is easy to show that $\hat{\mathbf{S}}$ is consistent. Furthermore, if $s_i > 0$ for all i, $\hat{\mathbf{s}}$ $(= (\hat{s}_1, ..., \hat{s}_{m-1}))$ is asymptotically normally distributed. Letting $s_m = 1-s_1-\cdots s_{m-1}$, an estimator of the covariance matrix of $(\hat{s}_1, ..., \hat{\mathbf{s}}_{m-1})$ is the inverse of the sample information matrix $-\left(\frac{\partial^2 \log \Lambda_n}{\partial s_i \partial s_j}\right)_{(m-1)\times(m-1)}\Big|_{\mathbf{S}=\hat{\mathbf{S}}}$.
Let

$$J_{\hat{\mathbf{s}}} = -\frac{\partial^2 \log \Lambda_n}{\partial \mathbf{s} \partial \mathbf{s}^t}\Big|_{\mathbf{s}=\hat{\mathbf{s}}} = (\sum_{h=1}^{n} \frac{(\delta_{hi} - \delta_{hM})(\delta_{hj} - \delta_{hM})}{(\mu_{\hat{F}}(\mathcal{I}_h))^2})_{(M-1)\times(M-1)}.$$

$J_{\hat{\mathbf{s}}}$ is also nonsingular and $J_{\hat{\mathbf{s}}}^{-1}$ is another consistent estimator of the covariance matrix of the GMLE $\hat{\mathbf{s}}$. Note that $\hat{F}$ is a linear function of $\hat{\mathbf{s}}$. Thus, we can estimate the covariance of $(\hat{F}(\mathbf{x}), \hat{F}(\mathbf{y}))$.

When $d \geq 2$, the above arguments are no longer true. The GMLE of $\mathbf{S}$ may not be unique and $J_{\hat{\mathbf{s}}}$ may not be positive definite. See the following bivariate examples.

Example 1 Suppose that a sample of size 4 consists of observations $(L_{i1}, R_{i1}, L_{i2}, R_{i2})$, $i = 1, ..., 4$, which equal $(1, 6, 1, 3)$, $(1, 6, 4, 6)$, $(1, 3, 1, 6)$ and $(4, 6, 1, 6)$, respectively. Then the MI's are $A_1 = (1, 3]\times(1, 3]$, $A_2 = (1, 3]\times(4, 6]$, $A_3 = (4, 6]\times(1, 3]$ and $A_4 = (4, 6]\times(4, 6]$. $\hat{\mathbf{S}}_q = q(1/2, 0, 0, 1/2) + (1-q)(0, 1/2, 1/2, 0)$, $q \in (0, 1)$, are all GMLEs of $\mathbf{S}$.

Yu, Wong and He (see [7]) showed that $J_{\hat{\mathbf{s}}}$ in Example 1 is singular.

There are two types of non-uniqueness of $\hat{F}$ for case 2 IC data. (1)the measure $\mu_{\hat{F}}$ (induced by $\hat{F}$) assigned to an MI A maynot be unique. (2) $\hat{F}$ is **always** not unique at the points in the interior of an MI. Example 1 presents an instance for both types of non-uniqueness of the GMLE. There are at least two GMLE's, say $\hat{F}$ and $\tilde{F}$, such that $\mu_{\hat{F}}(A_1) = 1/2$ but $\mu_{\tilde{F}}(A_1) = 0$ for an MI A_1, where $A_1 = (1, 3] \times (1, 3]$ is an MI. Thus $\hat{F}(3, 3) = 1/2$ and $\tilde{F}(3, 3) = 0$, where the point (3,3) is a vertex of the MI. Note that fixed $\mu_{\hat{F}}(A_i)$ for all MI's A_i, $\hat{F}$ (or $\tilde{F}$) is not uniquely determined on each A_i if $\mu_{\hat{F}}(A_i) > 0$ (or $\mu_{\tilde{F}}(A_i) > 0$). Namely, we can define $\hat{F}$ to be one with a density function $\hat{f} = \frac{1}{8}$ on $A_1 = (1, 3] \times (1, 3]$, or define to be discrete with a jump 1/2 at the point $(3, 3)$ on A_1.

Only the first type of non-uniqueness causes a problem in estimating the variance of $\hat{F}(\mathbf{x})$. If $J_{\hat{\mathbf{S}}}$ is singular, it indicates that we encounter the problem. Thus how to deal with such a situation is an important new issue in multivariate interval censoring as it does not occur in univariate interval censoring. The procedure to solve such problem is as follows.

Estimation of F_0 and the covariance matrix of a GMLE

Let $\hat{F}$ be a GMLE of F. Then it is proved (see Proposition 1 in [7]) that each solution of $\mathbf{S}$ $(= (s_1, ..., s_m)')$ to the equations

$$\sum_{j=1}^{m} \delta_{ij} s_j = \mu_{\hat{F}}(I_i),\ i = 1, ..., n,\ \sum_{j=1}^{m} s_j = 1 \text{ and } s_j \geq 0,$$

is also a GMLE of $\mathbf{S}_0$. In an obvious way, rewrite the $n+1$ equations above as a matrix form

$$\mathcal{B}\mathbf{S} = \mathbf{p}, \text{ where } \mathbf{S} \geq 0 \text{ and } \mathbf{p} = (\mu_{\hat{F}}(I_1), ..., \mu_{\hat{F}}(I_n), 1)^t.$$

Derive a GMLE $\hat{\mathbf{S}}$ using the self-consistent algorithm. If the rank r of $\mathcal{B}$ is $r = m$, then $J_{\hat{\mathbf{s}}}$ is nonsingular (see Lemma 1 in [7]) and we use $J_{\hat{\mathbf{s}}}^{-1}$ as the estimate of the covariance matrix of $(\hat{s}_1, ..., \hat{s}_{M-1})$.

If $r < m$, then there are multiple solutions $\mathbf{S}$ of $\mathcal{B}\mathbf{S} = \mathbf{p}$, where $\mathbf{S} \geq 0$ and $\mathbf{p} = (\mu_{\hat{F}}(I_1), ..., \mu_{\hat{F}}(I_n), 1)^t$, in which $\hat{F}$ is the cdf. induced by the given $\hat{\mathbf{S}}$, and we shall look for another GMLE to replace $\hat{F}$. It can be shown (see Lemma 3 in [7]) that

> there exist r linearly independent column vectors of $\mathcal{B}$, say, columns i_1, ..., i_r, such that $\mathbf{c}_o = (\mathcal{V}^t\mathcal{V})^{-1}\mathcal{V}^t\mathbf{p} \geq 0$ (write $\mathbf{c}_o^t = (c_{01}, ..., c_{0r})$), where $\mathcal{V}$ is the $(n+1) \times r$ matrix consisting of columns $i_1, ..., i_r$ of $\mathcal{B}$.

Let $\tilde{\mathbf{S}} = (\tilde{s}_1, ..., \tilde{s}_m)^t$, where $\tilde{s}_{i_j} = c_{0j}$, $j = 1$, ..., r, and $\tilde{s}_i = 0$ if $i \notin \{i_1, ..., i_r\}$. Verify that $\tilde{\mathbf{S}}$ is a GMLE of $\mathbf{S}_0$. We choose such a GMLE to replace $\hat{\mathbf{S}}$. Since there are only finitely many possible combinations of r linearly independent column vectors of $\mathcal{B}$, we have implemented a computer algorithm to obtain $\tilde{\mathbf{S}}$.

By reordering the index, without loss of generality (WLOG), we can assume

$$\tilde{\mathbf{S}}^t = (c_{01}, ..., c_{0r}, 0, ..., 0) = (\tilde{\mathbf{s}}^t, \tilde{s}_M, 0, ..., 0),$$

where M is the number of nonzero entries of $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{s}}$ the $(M-1) \times 1$ vector whose entries are all positive. Replacing $\hat{\mathbf{s}}$ by $\tilde{\mathbf{s}}$, $J_{\tilde{\mathbf{s}}}$ is nonsingular (see Lemma 1 in [7]). We propose to use the GMLE of F_0 corresponding to $\tilde{\mathbf{S}}$, denoted by

$$\tilde{F}(\mathbf{x}) = \sum_{i=1}^{M} u_i \tilde{s}_i \text{ (and } \tilde{F}(\mathbf{y}) = \sum_{i=1}^{M} v_i \tilde{s}_i),$$

where $u_j = \mathbf{1}(A_j \subset [0, x_1] \times \cdots \times [0, x_d])$ (and $v_j = \mathbf{1}(A_j \subset [0, y_1] \times \cdots \times [0, y_d])$). Then the covariance of the new GMLE $\tilde{F}$ can be estimated by

$$Cov(\tilde{F}(\mathbf{x}), \tilde{F}(\mathbf{y})) = (u_1 - u_M, ..., u_{M-1} - u_M) J_{\tilde{\mathbf{s}}}^{-1} (v_1 - v_M, ..., v_{M-1} - v_M)^t.$$

C3. REPORTABLE OUTCOMES

- **6 published articles in journals cited in the science citation index**: [2], [3], [4], [5], [6], [7].
- Computer softwares related to Task 1 installed in math.binghamton.edu.ftp/pub/qyu.

D. CONCLUSIONS

In the first year of our DOD grant, we have successfully accomplished our research objectives (Tasks 1 and 2) on the asymptotic inference of the GMLE of the survival function for multivariate interval-censored data. Under the case 2 multivariate interval censorship models, we have established consistency, asymptotic normality and asymptotic efficiency of the GMLE under the discrete condition. the result is published in a journal (see [2]).

We have successfully proposed a procedure to deal with the identiability problem that the GMLE may not be unique, and a new paper which describes our procedure has been accepted by a statistical journal [7]. We further developed a computer software to implement the new statiscal procedure.

The results which we have established provide a set of fundamentally important statistical tools for the analysis of most types of IC data that are encountered in clinical follow-up studies.

E. REFERENCES

[1] Wong, G. Y. C., Bradlow, H. L., Sepkovic, D., Mehl, S., Mailman, J. and Osborne, M. P. (1997). A dose-ranging study of indole-3-carbinol for breast cancer prevention. *Journal of Cellular Biochemistry Supplements* 28/29, 111-116.

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

[3] Yu, Q. Q., Li, L. X. and Wong, G. Y. C. (2000). On consistency of the self-consistent estimator of survival functions with interval censored data. *Scan. J. of Statist.*, 27, 35-44.

[4] Schick, A. and Yu, Q. Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scan. J. of Statist.*, 27, 45-55.

[5] Yu, Q. Q. and Li, L.X. (2000). Asymptotic properties of the GMLE of self-consistent estimators with doubly-censored data. *Acta Math. Sinica.* (Accepted).

[6] Yu, Q. Q., Wong, G. Y. C. and Li, L. X. (2000). Asymptotic properties of self-consistent estimators with mixed interval-censored data. *Ann. Inst. Statist. Math.*, (tentatively accepted).

[7] Yu, Q.Q, Wong, G.Y.C. and He, Q.M. (2000). Estimation of a joint distribution function with multivariate interval-censored data when the nonparametric MLE is not unique. *Biometrical Journal.* (Accepted). 69, 155-166.