東海大學統計研究所

碩士論文

指導教授:沈葆聖博士

在雙截資料下 Cox 模型的擬最大概似估計

Pseudo maximum likelihood estimation for Cox Model with Doubly Truncated Data

研究生:卓穎蓁

中華民國一百零五年六月

東海大學碩士班研究生

論文口試委員審定書

統計學系碩士班卓穎蓁君所提之論文

在雙截資料下 Cox Model 的擬最大概似估計

經本委員會審議，認為符合碩士資格標準。

論文口試委員召集人 戴政 （簽章）

委員 張文彤

江存琨

中華民國 105 年 06 月 24 日

謝誌

轉眼間兩年的研究所生涯即將邁入尾聲，心中滿滿不捨，感謝讓我有所蛻變的一切。首先感謝我的指導教授-沈葆聖博士，在論文撰寫期間給予的教導與督促，並且從中學習到對於研究的執著與態度。在研究所求學的這兩年，旁聽了沈老師大學部的課，老師教會我的不只是專業的知識，從老師的身上我也看到了他對工作教學上的熱忱與負責，我想我要學的還很多，在往後的人生裡只要我遇到困難時，我仍會想起老師那份堅強的意志力，絕不輕言放棄，並且用更積極的態度去面對解決事情。真的很榮幸能夠遇見老師，並在老師的薰陶下成長，不管是在課業上、求職上，老師都很願意給予意見與幫助，在此由衷的感謝您。

再來要感謝的是戴政博士和張玉媚博士百忙之中抽空來擔任我的論文口試委員，過程中提出的問題給了我許多啟發，使我受益良多，也給了一些寶貴的建議，讓我可以使論文能夠更加完善。

此外論文的完成承蒙許多人的支持以及鼓勵，讓我在曾經想要放棄之時卻又見到一絲曙光。感謝研究室的同學們，在我研究所生涯中添增了豐富的色彩，總是能帶給我無窮盡的鼓勵，給予我心靈上的滋潤並且洗滌疲憊的身心。

最後，感謝默默支持、關心、鼓勵與包容我的家人，讓我在求學過程當中無後顧之憂，忙碌之餘總是為我加油打氣。尤其是我的父親，辛苦的工作為了就是讓我有好的資源、在好的環境下沒有任何煩惱的學習。感謝有個偉大的爸爸，永遠當我的後盾，讓我可以一步一步的往前邁進，內心的所有感謝無法用言語一一表達，只能再一次地謝謝，有您真好!

Abstract

The partial likelihood (PL) function has been mainly used for proportional hazards models with censored data. The PL approach can also be used for analyzing left-truncated or left-truncated and right-censored data. However, when data is subject to double truncation, the PL approach no longer works due to the complexities of risk sets. In this article, we propose pseudo maximum likelihood approach for estimating regression coefficients and baseline hazard function for the Cox model with doubly truncated data. We propose expectation-maximization algorithms for obtaining the pseudo maximum likelihood estimators (PMLE). The consistency property of the PMLE is established. Simulations are performed to evaluate the finite-sample performance of the PMLE. The proposed method is illustrated using an AIDS data set.

Key Words: EM algorithm; Pseduo-likelihood; Double truncation; Inverse-Probability-Weighted.

Contents

1. Introduction

Doubly truncated failure-time arises if an individual is potentially observed only if its failure-time lies within a certain interval, unique to that individual. Efron and Petrosian (1999) motivated double-truncation issue using data on quasars, which are only detected when their luminosity falls between two observational limits (Lynden-Bell (1971)). Doubly truncated data also play an important role in the statistical analysis of survival times. Bilker and Wang (1996) indicated that induction times in AIDS are doubly truncated. Consider the following example:

Example: CDC AIDS Blood Transfusion Data

The AIDS Blood Transfusion Data are collected by the Centers for Disease Control (CDC), which is from a registry database, a common resource of medical data. The data consist of the time in month and only cases having either one transfusion or multiple transfusions in the same calendar month were used. The cases either diagnosed or reported after July 1, 1986 (τ_2) were not included to avoid inconsistent data and bias resulting from reporting delay. Thus, the observed data is subject to right truncation. Moreover, cases having the AIDS prior to January 1, 1982 (τ_1) were not included because HIV was unknown prior to 1982, any cases of transfusion-related AIDS before this time would not have been properly classified and thus would have been missed. Hence, in addition to right truncation, the observed data are also truncated from the left. Let T_B denote the calendar time (in years) of the initiating events (HIV infection), and T_D be the calendar time (in years) of AIDS onset. Let $T = 12(T_D - T_B)$ (in month) be the induction or incubation time from HIV infection to AIDS. Let $U = 12(\tau_1 - T_B)$ and $V = U + d_0 = 12(\tau_2 - T_B)$ (in month), where $d_0 = 12(\tau_2 - \tau_1) = 54$. Define a population as the individuals who were infected with HIV before τ_1 and develop AIDS prior to τ_2. Thus, the CDC AIDS Blood Transfusion Data can be viewed as being doubly truncated since the incubation times T's are observable only when $\tau_1 \leq T_D \leq \tau_2$ (i.e. $U \leq T \leq V$). Assume for the un-truncated individual, a $p \times 1$ vector of covariates $Z = [z_1, \ldots, z_p]^T$ is available. Figure 1 highlights all the different times for doubly truncated data described in Example.

For doubly truncated data, the nonparametric maximum likelihood estimator (NPMLE) of the distribution function of T was first studied by Efron and Petrosian (1999)). Shen (2010a) added the NPMLE for the truncation distributions

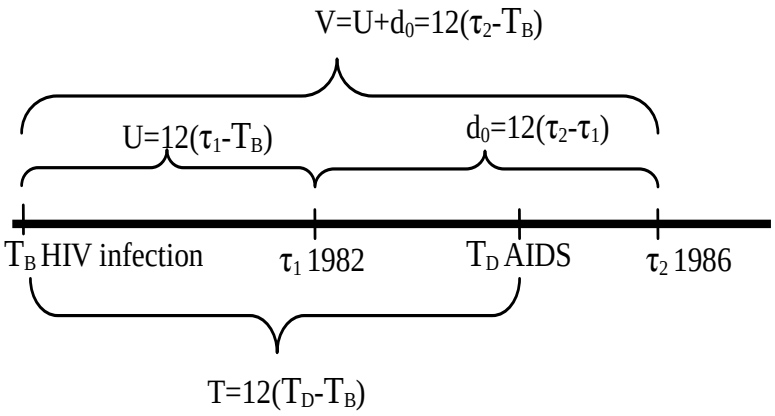

Figure 1. Schematic depiction of doubly truncated data

and established the consistency and weak convergence of the NPMLE. Although a closed-form variance estimation was not established then, this issue was solved for the event-time distribution by Emura et al.(2015a). Moreira and de Uña-Álvarez (2010a) demonstrated that the simple bootstrap method can be used to approximate the sampling distribution of the NPMLE. Zhang (2015) derived a variance estimation for both NPMLEs under the assumption that the lower and upper truncation variables are independent. Moreira, de Uña-Álvarez and Crujeiras (2010) presented the R package DTDA for analyzing truncated data, which contains three different algorithms for the approximation of the NPMLE under double truncation as well as bootstrap confidence bands. Moreira and Keilegom (2013) considered several bandwidth selection procedures for kernel density estimation of a random variable that is sampled under random double truncation.

Parametric procedures for doubly truncated have been also studied in the literature. Efron and Petrosian (1999) proposed the maximum likelihood estimator (MLE) under a parametric family, called the special exponential family (SEF). Hu and Emura (2015) developed the randomized Newton-Raphson algorithms to obtain the MLE. Emura et al. (2015b) pointed out that the classical asymptotic theory for the i.i.d. data is not suitable for studying the MLE under double-truncation. Alternatively, they formalized the asymptotic results under the independent but not identically distributed data that take into account for the between-sample het-

erogeneity of truncation variables.

Compared to the nonparametric or parametric analyses, research is much scarcer on analyses of doubly truncated data in the presence of covariates. When Z is discrete, under linear transformation models, Shen (2013) extended Zhang et al's (2005) approach to doubly-truncated and interval-censored data. However, the proposed approach requires the estimation of survival functions for each level of covariates. Shen (2016) analyzed doubly truncated data using semiparametric transformation models. He demonstrated that the extended estimating equations of Cheng et al. (1995) can be used to analyze doubly truncated data. However, his approach only considered the estimation of regression coefficients. In this article, we consider the estimation of regression coefficients and baseline hazard function for Cox model with doubly truncated data.

The partial likelihood (PL) function has been mainly used for proportional hazards models with censored data (Cox 1972, 1975). The PL approach can also be used for analyzing right-truncated (RT) or left-truncated and right censored (LTRC) data (Kalbfleisch and Lawless (1991); Wang, Brookmeyer, and Jewell (1993)). For LT/LTRC data, the PL function has an expression similar to that of the PL for right-censored data except for the definition of the risk sets. However, when data is subject to double truncation, the PL approach no longer works due to the complexities of risk sets. In Section 2, we propose pseudo maximum likelihood approach for estimating regression coefficients and baseline hazard function for the Cox model with doubly truncated data. Expectation-maximization algorithms are proposed for obtaining the pseudo maximum likelihood estimators (PMLE). We consider two cases: the case when the distribution of the truncation time is unspecified and the case when it is known up to a finite-dimensional parameter vector. The consistency property of the PMLE is established. In Section 3, simulations are performed to evaluate the finite-sample performance of the PMLE. In Section 4, the proposed estimator is illustrated using an AIDS data set.

2. The Proposed Estimator

2.1 When G is unspecified

Let $F(t|Z) = P(T \leq t|Z)$ denote the cumulative distribution function of T given Z. Let $G(t) = P(U \leq t)$ denote the cumulative distribution function of U. Suppose that the support of T does not depend on Z. Let a_F and b_F denote the left and right endpoint of support of T, and similarly, define (a_G, b_G) as the left and right endpoint of support of U. Throughout this article, for identifiability of $F(t|Z)$, we assume that $a_G \leq a_F \leq a_G + d_0, b_G \leq b_F \leq b_G + d_0$. We also assume that U is independent of (T, Z).

In Example, since $a_G = 0$, $a_G \leq a_F$ always holds. When $a_F = 0$, $a_F \leq a_G + d_0$ always holds. Furthermore, when d_0 is large enough, we have $b_F \leq b_G + d_0$, this will hold if follow-up is long enough such that $b_G + d_0$ is larger than the right endpoint of T. In Example, U denotes the time interval between the date of HIV infection and study start date and it is reasonable to assume that U is independent of (T, Z) if the pattern of infection through blood transfusion is independent of covaraites Z and incubation time T.

The proportional hazards model (Cox (1972)) is commonly used in the analysis of survival time and related data. The Cox model is given by

$$\lambda(t|Z) = \lambda(t) \exp(\beta^T Z),$$

where β is the regression parameter, $\lambda(t|Z)$ is the conditional hazard function of T given Z with $\lambda(t)$ as an arbitrary baseline hazard function.

Let $(T_1, U_1, V_1 = U_1 + d_0, Z_1), \ldots, (T_n, U_n, V_n = U_n + d_0, Z_n)$ denote the doubly truncated sample of (T, U, V, Z). As pointed out by Ren and Zhou (2011), under Cox model (2.1), given Z, the variable T satisfies Lehmann family properties (Kalbfleisch and Prentice (2002), page 97) $S(t|Z) = [S(t)]^{\exp(\beta^T Z)}$, where $S(t) = \exp(-\Lambda(t))$ is the baseline survival function and $\Lambda(t)$ is the cumulative baseline hazard function. This implies that $f(t|Z) = \exp(\beta^T Z)\lambda(t)S(t|Z)$, where $f(t|Z)$ is the density functions of $S(t|Z)$. Notice that the distribution of the truncated vector (T_i, U_i, V_i, Z_i) differs from that of the untruncated vector (T, U, V, Z), e.g. $P(T_i \leq t|Z_i) = \alpha(\beta, \Lambda, G; Z_i)^{-1} P(T \leq t, U \leq T \leq V = U + d_0|Z = Z_i)$, where $\alpha(\beta, \Lambda, G; Z_i) = P(U \leq T \leq U + d_0|Z = Z_i)$ is the untruncation probability given

$Z = Z_i$. Under the Cox model, the un-truncated probability $\alpha(Z_i)$ can be written as

$$\begin{aligned}\alpha(\beta, \Lambda, G; Z_i) &= P(U \le T \le U + d_0 | Z = Z_i) \\ &= P(T > U | Z = Z_i) - P(T > U + d_0 | Z = Z_i) \\ &= S(U | Z = Z_i) - S(U + d_0 | Z = Z_i) \\ &= \int_0^\infty [S(x|Z_i) - S(x + d_0|Z_i)] dG(x) \\ &= \int_0^\infty \left[\{\exp(-\Lambda(x)\}^{\exp(\beta^T Z)} - \{\exp(-\Lambda(x + d_0))\}^{\exp(\beta^T Z)} \right] dG(x).\end{aligned}$$

First, we consider the estimation of $G(x)$. Let $F(t) = P(T \le t)$ and $f(t)$ denote the unconditional cumulative distribution function and probability density functions of T, i.e. $F(t) = \int P(T \le t | Z = z) P_Z(z) dz$, where $P_Z(z)$ is the probability density function of Z. The inverse-probability-weighted (IPW) estimators of $G(t)$ and $F(t)$ can be obtained (see Shen (2003), Shen (2010a)) by solving the following pairs of simultaneous equations:

$$\hat{F}_n(t) = \Big[\sum_{i=1}^n \frac{1}{\hat{G}_n(T_i) - \hat{G}_n((T_i - d_0)-)} \Big]^{-1} \sum_{i=1}^n \frac{I_{[T_i \le t]}}{\hat{G}_n(T_i) - \hat{G}_n((T_i - d_0)-)},$$

and

$$\hat{G}_n(t) = \Big[\sum_{i=1}^n \frac{1}{\hat{F}_n(U_i + d_0) - \hat{F}_n(U_i-)} \Big]^{-1} \sum_{i=1}^n \frac{I_{[U_i \le t]}}{\hat{F}_n(U_i + d_0) - \hat{F}_n(U_i-)},$$

where $\hat{F}_n(t)$ and $\hat{G}_n(t)$ are the IPW estimators of $F(t)$ and $G(t)$, respectively. Notice that this approach requires that the truncated sample is a simple random sample from the truncated population, i.e. the observations with $T \ge V$. If the sampling scheme depends on covariates, the IPW estimator $\hat{F}_n(t)$ is an inconsistent estimator of $F(t)$.

Next, we consider the estimation of β and Λ. Given $\hat{G}_n$, the likelihood for F is proportional to

$$L_n(\beta, \Lambda, \hat{G}_n) = \prod_{i=1}^n \frac{f(T_i|Z_i)}{\alpha(\beta, \Lambda, \hat{G}_n; Z_i)}$$

and the log-likelihood function of $L_n(\beta, \Lambda, \hat{G}_n)$ can be expressed as

$$l_n(\beta, \Lambda, \hat{G}_n) = n^{-1} \sum_{i=1}^n \Big\{ \int_0^\infty [\beta^T Z_i + \log \lambda(t) - \Lambda(t|Z_i)] dN_i(t) - \log \alpha(\beta, \Lambda, \hat{G}_n; Z_i) \Big\},$$

where $N_i(t) = I_{[T_i \le t]}$.

Because the likelihood l_n includes the plug-in value of $\hat{G}_n$, the likelihood is called the pseudo-likelihood. The maximization of $l_n(\beta, \Lambda, \hat{G}_n)$ leads to the same difficulty as in estimating density function (no maximizer). A rout way route out of this difficulty is to extend the parameter space so that the estimator of Λ is allowed to be discrete. Thus, we relax Λ to be right-continuous and allow $\Lambda(t)$ to have jumps at the T_i's. For length-biased and right-censored data, Qin et al. (2011) proposed expectation-maximization (EM) algorithms to obtain the maximum likelihood estimation of the nonparametric and Cox models. Motivated by the approach of Qin et al. (2011), we propose an EM algorithm for obtaining the PMLE of (β, Λ) based on $l_n(\beta, \Lambda, \hat{G}_n)$. Let $t_1 < t_2 < \cdots < t_n$ be the ordered failure times for $\{T_1, \ldots, T_n\}$. We redefine $\Lambda(\cdot)$ as a step function with jumps only at the event times t_i. For $i = 1, \ldots, n$, given $O_i = (T_i, Z_i)$, let $O_i^* = \{(T_{i1}^*, U_{i1}^*), \ldots, (T_{im_i}^*, U_{im_i}^*)\}$ denote the truncated latent data corresponding to covariate Z_i. Given G, the random integer m_i then follows a geometric distribution with parameter $\alpha(\beta, \Lambda, G; Z_i)$ and

$$E[m_i|O_i] = \frac{1 - \alpha(\beta, \Lambda, G; Z_i)}{\alpha(\beta, \Lambda, G; Z_i)}.$$

We develop the EM algorithm based on the discrete version of $\Lambda(x) = \sum_{t_j \le x} \lambda_j$, where λ_j is the positive jump at time t_j for $j = 1, \ldots, n$. For notational convenience, let $f_i(t) = dF(t|Z_i)$.

E Step:

Let $T_1, \ldots, T_n$ denote the doubly-truncated sample with $U_i < T_i < U_i + d_0$. Hence, the log-likelihood based on the complete data is then

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \left\{ I_{[T_i = t_j]} + \sum_{l=1}^{m_i} I_{[T_{il}^* = t_j]} \right\} \log f_i(t_j).$$

Conditional on the observed doubly truncated data O_i,

$$E\left[I_{[T_i = t_j]} \middle| O_i \right] = I_{[T_i = t_j]}.$$

Furthermore,

$$E_{m_i}\left\{ E\left[\sum_{l=1}^{m_i} I_{[T_{il}^* = t_j]} \middle| O_i \right] \right\} = E_{m_i}\left\{ \sum_{l=1}^{m_i} P(T = t_j | T < U \text{ or } T > U + d_0, Z = Z_i) \right\}$$

$$= E_{m_i}\Big\{ m_i P(T = t_j | Z = Z_i) P(T < U \text{ or } T > U + d_0 | T = t_j, Z = Z_i)$$

$$/ P(T < U \text{ or } T > U + d_0 | Z = Z_i) \Big\}$$

$$= E_{m_i}\Big\{ m_i f_i(t_j)[1 - G(t_j) + G(t_j - d_0)] / [1 - \alpha(\beta, \Lambda, G; Z_i)] \Big\}$$

$$= \frac{f_i(t_j)[1 - G(t_j) + G(t_j - d_0)]}{\alpha(\beta, \Lambda, G; Z_i)},$$

where

$$f_i(t_j) = \exp(\beta^T Z_i)\lambda_j \exp\Big\{ - \sum_{l=1}^{j} \lambda_l \exp(\beta^T Z_i) \Big\}.$$

Hence,

$$w_{ij} = E\left[I_{[T_i = t_j]} + \sum_{l=1}^{m_i} I_{[T_{ij}^* = t_j]} \middle| O_i \right]$$

$$= I_{[T_i = t_j]} + \frac{f_i(t_j)[1 - G(t_j) + G(t_j - d_0)]}{\alpha(\beta, \Lambda, G; Z_i)}.$$

Thus, given $\hat{G}_n$, w_{ij} can be estimated by

$$\hat{w}_{ij} = I_{[T_i = t_j]} + \frac{f_i(t_j)[1 - \hat{G}_n(t_j) + \hat{G}_n(t_j - d_0)]}{\alpha(\beta, \Lambda, \hat{G}_n; Z_i)}. \tag{2.1}$$

The expected complete-data log-likelihood function conditional on the observed data O_i $(i = 1, \ldots, n)$ is as follows:

$$l_E(\beta, \lambda, \hat{G}_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{w}_{ij} \log f_i(t_j)$$

$$= \sum_{j=1}^{n} \hat{w}_{+j} \log \lambda_j + \sum_{i=1}^{n} \hat{w}_{i+} \beta^T Z_i - \sum_{l=1}^{n} \sum_{j=l}^{n} \sum_{i=1}^{n} \hat{w}_{ij} \exp(\beta^T Z_i)\lambda_l,$$

where $\hat{w}_{+j} = \sum_{i=1}^{n} \hat{w}_{ij}$ and $\hat{w}_{i+} = \sum_{j=1}^{k} \hat{w}_{ij}$.

M Step:

In the M-step, we maximize $l_E(\beta, \lambda, \hat{G}_n)$ with respect to λ_j $(j = 1, \ldots, n)$,

$$\frac{\partial l_E(\beta, \lambda, \hat{G}_n)}{\partial \lambda_j} = \frac{\hat{w}_{+j}}{\lambda_j} - \sum_{l=j}^{n} \sum_{i=1}^{n} \hat{w}_{il} \exp(\beta^T Z_i) = 0,$$

which leads to a closed form of λ_j as a function of β, given by

$$\lambda_j(\beta) = \frac{\hat{w}_{+j}}{\sum_{l=j}^{n}\sum_{i=1}^{n}\hat{w}_{il}\exp(\beta^T Z_i)}. \tag{2.2}$$

Next, we maximize $l_E(\beta, \lambda, \hat{G}_n)$ with respect to β

$$\frac{\partial l_E(\beta,\lambda,\hat{G}_n)}{\partial\beta} = \sum_{i=1}^{n}\hat{w}_{i+}Z_i - \sum_{l=1}^{n}\sum_{j=l}^{n}\sum_{i=1}^{n}\hat{w}_{ij}Z_i\exp(\beta^T Z_i)\lambda_l. \tag{2.3}$$

By inserting $\lambda_j(\beta)$ of (2.2) into (2.3), β can be solved by the following equation:

$$\sum_{i=1}^{n}\hat{w}_{i+}Z_i - \sum_{l=1}^{n}\hat{w}_{+l}\left\{\frac{\sum_{i=1}^{n}\sum_{j=l}^{n}\hat{w}_{ij}Z_i\exp(\beta^T Z_i)}{\sum_{i=1}^{n}\sum_{j=l}^{n}\hat{w}_{ij}\exp(\beta^T Z_i)}\right\} = 0.$$

Hence, given $\hat{G}_n$, one can update the expectation of the likelihood via $\hat{w}_{ij}$ in (2.1) and repeat M-step until the the estimators β and λ_j $(j = 1, \ldots, n)$ converge. We denote the PMLE by $\hat{\zeta}_n = (\hat{\beta}_n, \hat{\Lambda}_n)$ and let $\zeta_0 = (\beta_0, \Lambda_0)$ be the true value.

We require the following conditions to derive the the asymptotic properties of $\hat{\beta}_n$ and $\hat{\Lambda}_n$:

(A1) Let $[0, \tau_c] \in [0, \infty]$ such that $K(x) = G(x) - G((x-d_0)-) > \delta > 0$ for $x \in [0, \tau_c]$. Moreover, assume that (a) $\int_0^{\tau_c} G(dx)/W(x) < \infty$, where $W(x) = F_u(x+d_0) - F_u(x)$ and (b) $F_u(dx)/G(dx)$ is uniformly bounded on $[0, \tau_c]$

(A2) The true value of the hazard function $\Lambda_0(\cdot)$ is continuously differentiable, $\Lambda_0(0) = 0$ and $\Lambda_0(\tau_c) < \infty$.

(A3) The parameter β is in a compact set $\mathcal{B}$ that contains β_0.

(A4) Both $E[||Z||^2]$ and $E[|\exp(\beta^T Z)|]$ are bounded, where $||Z|| = (|z_1|^2 + \cdots + |z_p|^2)^{1/2}$.

(A5) The information matrix $-\partial^2 E[l_n(\beta, \hat{\lambda}(\cdot, \beta), G]/\partial^2\beta$ evaluated at true value β_0 is positive definite for every n.

(A6) If $P(b^T Z = c) = 1$ for some constant c, then $b = 0$.

Assumption (A1) is required for the consistency of $\hat{G}_n$. Assumptions (A2) and (A3) are required for stochastic approximation. Assumptions (A4) and (A5) are conditions for establishing asymptotic properties of the estimated coefficient under Cox

model (Andersen et al. (1993)). This implies that given G and λ the information matrix for β is positive definite. Assumption (A6) implies that there is no covariate colinearity, which ensures the model identifiability.

Note that given $\hat{G}_n$ the log-likelihood function $l_E(\beta, \lambda, \hat{G}_n)$ is strictly concave in λ. Hence, given $\hat{G}_n$, for each β in a compact set $\mathcal{B}$, we can find a unique maximizer of $\hat{\lambda}(\cdot, \beta, \hat{G}_n)$ of the likelihood function $l_E(\beta, \lambda, \hat{G}_n)$. The existence of the unique PMLE for (β, λ) follows by assumptions (A2) through (A5).

Theorem 1. Let $\hat{\zeta}_n = (\hat{\beta}_n, \hat{\Lambda}_n)$ and $\zeta_0 = (\beta_0, \Lambda_0)$. Under assumptions (A1)-(A6), the PMLE $\hat{\zeta}_n$ is consistent: $\hat{\beta}_n$ converges to β_0, and $\hat{\Lambda}_n(t)$ converges uniformly in t for $t \in [0, \tau_c]$. Furthermore, $\sqrt{n}[\zeta_n - \zeta_0]$ converges weakly to a mean zero Gaussian process.

Proof: The proof is technical and not shown here.

2.2. When $G(x) = P(U \leq x)$ is parameterized as $G(x; \theta)$

In some cases, the distribution of left truncation times, denoted by $G(x) = P(U \leq x)$, can be parameterized as $G(x; \theta)$, where $\theta \in \Theta$, Θ is a known compact set in R^q and θ is a q-dimensional vector. For prevalent data with fixed recruitment time, the truncation distribution G can be interpreted as the disease distribution. For stable diseases, the disease-onset cases are approximately uniformly distributed over the calendar time, i.e. length-biased data . For a new disease, however, one might prefer to parameterize G so that the parameterization reflects the growth of the disease over time. When $G(x)$ is parameterized as $G(x; \theta)$, Moreira and de Uña-Álvarez (2010b) and Shen (2010b) proposed a semiparametric estimator of F. Both papers demonstrated that it may be more efficient than the NPMLE of F.

Under Cox model and $G(x) = G(x; \theta)$, we have

$$\alpha(Z_i) = \alpha(\beta, \Lambda, \theta; Z_i) = \int_0^\infty H(x; \theta) f(x|Z_i) dx = \int_0^\infty [S(x|Z_i) - S(x + d_0|Z_i)] g(x; \theta) dx,$$

where $g(x; \theta)$ and the probability density function of U and $H(x; \theta) = G(x; \theta) - G(x - d_0; \theta)$. The full likelihood function of (F, G) is given by

$$L(\beta, \Lambda, \theta) = \prod_{i=1}^{n} \Big\{ dF(T_i|Z_i) g(U_i; \theta) / \alpha(\beta, \Lambda, \theta; Z_i) \Big\}.$$

The full likelihood can be written as

$$L(\beta, \Lambda, \theta) = L_m(\beta, \Lambda, \theta) \times L_c(\theta),$$

where

$$L_m(\beta, \Lambda, \theta) = \prod_{i=1}^{n} \frac{H(T_i; \theta) dF(T_i|Z_i)}{\alpha(\beta, \Lambda, \theta; Z_i)} \quad \text{and} \quad L_c(\theta) = \prod_{i=1}^{n} \frac{g(U_i; \theta)}{H(T_i; \theta)}.$$

Let $\hat{\theta}_n$ denote the MLE by maximizing $L_c(\theta)$. Given $\hat{\theta}_n$, the pseudo-likelihood for F is proportional to

$$L_n(\beta, \Lambda, \hat{\theta}_n) = \prod_{i=1}^{n} \frac{f(T_i|Z_i)}{\alpha(\beta, \Lambda, \hat{\theta}_n; Z_i)}.$$

The log-likelihood function of $L_n(\beta, \Lambda, \hat{\theta}_n)$ can be expressed as

$$l_n(\beta, \Lambda, \hat{\theta}_n) = n^{-1} \sum_{i=1}^{n} \left\{ \int_0^\infty [\beta^T Z_i + \log \lambda(t) + \log S(t|Z_i)] dN_i(t) - \log \alpha(\beta, \Lambda, \hat{\theta}_n; Z_i) \right\}.$$

We can obtain the semiparametric PMLE of (β, Λ) using the EM algorithms proposed in Section 2.1 with $\hat{G}_n(t)$ replaced by $G(t; \hat{\theta}_n)$. We denote the semiparametric MLE by $(\tilde{\beta}_n, \tilde{\Lambda}_n)$. When the parametric information is correct, it is expected the semiparametric PMLE outperforms the PMLE, but may behave badly when the assumed parametric model is far off. Moreira et al. (2014) proposed several Kolmogorov-Smirnov and Cramér-von Mises type test statistics, by which we can check if G can be parameterized as $G(t; \theta)$.

To derive the asymptotic properties of $\tilde{\beta}_n$ and $\tilde{\Lambda}_n$, we need the following conditions (B1), (B2) and conditions (A3)-(A6) of Theorem 1:

(B1) $G(x; \theta)$ is continuous in $x \in [0, \tau_c]$ for each $\theta \in \Theta$.

(B2) $\hat{\theta}_n \to \theta$ implies that $G(x; \hat{\theta}_n) \to G(x; \theta)$ for each $x \in [0, \tau_c]$.

Theorem 2. Let $\tilde{\zeta}_n = (\tilde{\beta}_n, \tilde{\Lambda}_n)$. and $\zeta_0 = (\beta_0, \Lambda_0)$. Under assumptions (B1),(B2) and (A3)-(A6), the PMLE $\tilde{\zeta}_n$ is consistent: $\tilde{\beta}_n$ converges to β_0, and $\tilde{\Lambda}_n(t)$ converges uniformly in t for $t \in [0, \tau_c]$. Furthermore, $\sqrt{n}[\tilde{\zeta}_n - \zeta_0]$ converges weakly to a mean zero Gaussian process.

Proof: By Anderson (1970), under the usual regularity conditions, $\hat{\theta}_n$ converges to θ with probability one. Similar to Lemma 3.1 of Wang (1989), under conditions

(B1) and (B2), we have with probability one, $\sup_{x\in[0,\tau]} |G(x;\hat{\theta}_n) - G(x;\theta)| \to 0$ as $n \to \infty$. The rest of proof is similar to that of Theorem and is omitted.

3. Simulation Studies

In this section, we conduct simulation studies to evaluate the performance of the proposed estimators. We generate T following the proportional hazards model with $\Lambda(t) = e^t - 1$ and $\beta = (\beta_1 = -2, \beta_2 = -3)^T$. The resulting T has the survivorship function

$$P(T > t|z_1, z_2) = e^{-(e^t-1)e^{-2z_1-3z_2}},$$

where z_1 is a bernoulli random variable with probability 0.5 and z_2 is an ordinal variable with $P(z_1^* = i) = 0.25$ for $i = 1, 2, 3, 4$. The U's are i.i.d. exponentially distributed with distribution function $G(x;\theta) = 1 - e^{-\theta x}$. The values of θ is chosen as $\theta = 0.25$. The V is generated from $V = U + d_0$, with $d_0 = 6, 9, 12$ such that the proportions of truncation(q_T) are equal to 0.58, 0.39 and 0.25, respectively. We keep the sample if $U \le T \le V$ and regenerate a sample if $T < U$ or $T > V$. The estimators $\hat{\beta}_n = (\hat{\beta}_{1n}, \hat{\beta}_{2n})^T$ and $\hat{\Lambda}_n$ are obtained using the IPW estimator $\hat{G}_n$ and EM algorithm in Section 2.1. Similarly, the estimators $\tilde{\beta}_n = (\tilde{\beta}_{1n}, \tilde{\beta}_{2n})^T$ and $\tilde{\Lambda}_n$ are obtained using the conditional maximum likelihood estimator $G(x;\hat{\theta}_n)$ and the EM algorithm in Section 2.1 with $\hat{G}_n(t)$ replaced by $G(t;\hat{\theta}_n)$. The convergence criterion is set as $||\hat{\beta}_n^{(r+1)} - \hat{\beta}_n^{(r)}|| < 0.0001$ (or $||\hat{\beta}_n^{(r+1)} - \hat{\beta}_n^{(r)}|| < 0.0001$).

For the estimation of β, Tables 1 and 2 show the mean average biases (bias) over all simulation runs, empirical standard deviations (std), respectively. Tables 1 and 2 also show the proportion of truncation $1 - P(U \le T \le V)$ (denoted by q_T). The sample sizes are chosen as 100, 200 and 400. The replication is 1000 times. We also consider the estimation of $S(t_0|1,1) = e^{-(e^{t_0}-1)e^{-5}}$, where the values of t_0 are chosen as such that $S(t_0|1,1) = 0.2(t_0 = 5.48)$, $S(t_0|1,1) = 0.5(t_0 = 4.64)$ and $S(t_0|1,1) = 0.8(t_0 = 3.53)$. Tables 3 and 4 show the mean average biases and empirical standard deviations (std) for $\hat{S}_n(t_0|1,1) = \exp\{-\hat{\Lambda}_n(t_0)e^{\hat{\beta}_{1n}+\hat{\beta}_{2n}}\}$ and $\tilde{S}_n(t_0|1,1) = \exp\{-\tilde{\Lambda}_n(t_0)e^{\tilde{\beta}_{1n}+\tilde{\beta}_{2n}}\}$, respectively.

Table 1. simulation results for bias and standard deviation of $\hat{\beta}_n$

			$\hat{\beta}_{1n}$		$\hat{\beta}_{2n}$	
d_0	n	q_T	bias	std	bias	std
6	100	0.58	0.229	0.400	0.384	0.520
6	200	0.58	0.153	0.301	0.235	0.393
6	400	0.58	0.105	0.166	0.186	0.192
9	100	0.39	0.117	0.382	0.197	0.454
9	200	0.39	0.096	0.263	0.138	0.298
9	400	0.39	0.072	0.137	0.085	0.143
12	100	0.25	0.070	0.323	0.069	0.447
12	200	0.25	0.038	0.256	0.027	0.286
12	400	0.25	0.015	0.132	0.010	0.150

Table 2. simulation results for bias and standard deviation of $\tilde{\beta}_n$

			$\tilde{\beta}_{2n}$		$\tilde{\beta}_{2n}$	
d_0	n	q_T	bias	std	bias	std
6	100	0.58	0.197	0.327	0.285	0.354
6	200	0.58	0.104	0.214	0.175	0.227
6	400	0.58	0.078	0.153	0.154	0.159
9	100	0.39	0.111	0.313	0.131	0.321
9	200	0.39	0.057	0.209	0.093	0.192
9	400	0.39	0.051	0.128	0.047	0.137
12	100	0.25	0.056	0.298	0.017	0.297
12	200	0.25	0.029	0.200	0.011	0.185
12	400	0.25	0.009	0.114	0.003	0.124

Table 3. simulation results for bias and standard deviation of $\hat{S}_n(t|1,1)$

			$\hat{S}_n(5.48\|1,1)$		$\hat{S}_n(4.64\|1,1)$		$\hat{S}_n(3.53\|1,1)$	
d_0	n	q_T	bias	std	bias	std	bias	std
6	100	0.58	0.021	0.081	-0.008	0.099	-0.015	0.089
6	200	0.58	0.007	0.054	-0.013	0.060	-0.017	0.065
6	400	0.58	0.007	0.038	-0.005	0.042	-0.013	0.037
9	100	0.39	0.029	0.079	0.009	0.100	-0.008	0.075
9	200	0.39	0.013	0.054	-0.002	0.073	-0.010	0.060
9	400	0.39	0.003	0.044	-0.004	0.051	-0.004	0.029
12	100	0.25	0.015	0.099	-0.004	0.121	-0.009	0.097
12	200	0.25	0.006	0.057	0.009	0.093	-0.006	0.072
12	400	0.25	0.003	0.048	-0.004	0.057	0.003	0.038

Table 4. simulation results for bias and standard deviation of $\tilde{S}_n(t|1,1)$

			$\tilde{S}_n(5.48\|1,1)$		$\tilde{S}_n(4.64\|1,1)$		$\hat{S}_n(3.53\|1,1)$	
d_0	n	q_T	bias	std	bias	std	bias	std
6	100	0.58	0.016	0.064	-0.006	0.076	0.014	0.069
6	200	0.58	0.006	0.041	-0.012	0.047	0.012	0.051
6	400	0.58	0.005	0.029	-0.004	0.033	0.009	0.028
9	100	0.39	0.020	0.062	0.008	0.079	0.011	0.061
9	200	0.39	0.011	0.043	-0.001	0.057	0.010	0.049
9	400	0.39	-0.005	0.036	-0.002	0.041	-0.005	0.023
12	100	0.25	0.012	0.081	-0.003	0.103	0.007	0.081
12	200	0.25	0.008	0.046	0.008	0.076	0.004	0.060
12	400	0.25	-0.004	0.039	-0.002	0.047	0.002	0.032

Based on the results of Tables 1 through 4, we have the following conclusions:

(i) For the estimation of β, the standard deviations of both $\hat{\beta}_{in}$ and $\tilde{\beta}_{in}$ increase as the proportion of truncation q_T increases. Similarly, the biases of both estimators tend to increase as the proportion of truncation q_T increases. Specifically, when $n = 100$ and truncation is severe (i.e. $q_T = 0.64$) the biases of both $\hat{\beta}_{in}$ and $\tilde{\beta}_{in}$ can be large. Their biases are small when truncation is light (i.e. $q_T = 0.15$) or sample size is large (i.e. $n = 400$).

(ii) The biases of $\tilde{\beta}_{ni}$ and $\tilde{S}_n(t_0)$ are smaller than that of $\hat{\beta}_{ni}$ and $\hat{S}_n(t_0)$, respectively, for most of the cases considered. The standard deviations of $\tilde{\beta}_{ni}$ and $\tilde{S}_n(t_0)$ are smaller than that of $\hat{\beta}_{ni}$ and $\hat{S}_n(t_0)$, respectively. The improvement in efficiency of $\tilde{\beta}_{ni}$ and $\tilde{S}_n(t_0)$ tend to increases as truncation proportion q_T increases.

4. Applications

To illustrate the proposed estimator, we analyze the CDC AIDS Blood Transfusion Data described in Example. Only cases having either one transfusion or multiple transfusions in the same calendar month are used. The data set include 295 cases diagnosed prior to July 1, 1986 (see Table 1 of Kalbfleisch and Lawless (1989)). The value of U_i (in month) is time from HIV infection to January 1, 1982; while V_i is defined as time from HIV infection to the end of study (July 1, 1986). Thus, the difference between V_i and U_i is always 54 (i.e. d_0) months. Our goal is to study the relationship between AIDS incubation time and age at infection. We treated age as categorical variable with two levels of age: 0-4 and > 4 years.

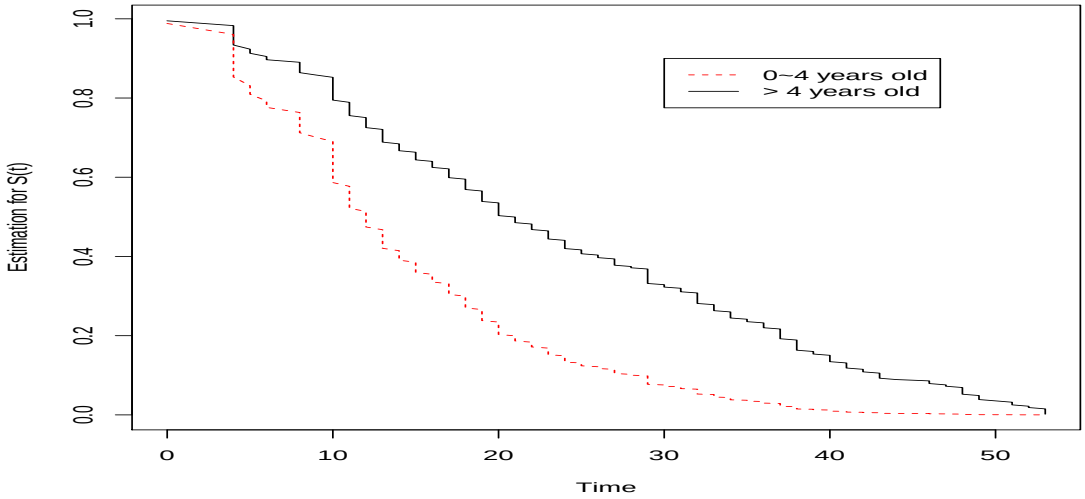

Figure 2. Plot of the Estimation for two survival function: ages 0-4 and > 4.

We fit Cox model with one covariate $z_1 = 1$ for individuals of age group 0-4 and $z_1 = 0$ otherwise. The estimated coefficient for z_1 is $\hat{\beta}_{1n} = 0.842$. Based on 500 bootstrap estimate, the corresponding bootstrap standard deviation estimator is equal to 0.216, which results in p-values < 0.001. The 95% interval estimators for β_1 is equal to $(0.419, 1.265)$. Figure 2 indicates that the survival function for age > 4 is above that for age 0-4.

5. Discussions

In this article, we have proposed estimators of regression coefficients and baseline hazards function for Cox model with doubly truncated data. We consider the case when the distribution of the truncation time is unspecified and the case when it is known up to a finite-dimensional parameter vector. Our simulation studies indicate that the proposed estimator performs adequately. The proposed method can be extended to the case when $V = U + D$, where D is a random variable. In this case, we can estimate the joint distribution function $K(x, y) = P(U \leq u, V \leq v)$ using the IPW estimator (Shen (2010)), say $\hat{K}_n(x, y)$. The truncation probability can be written as

$$\alpha(\beta, \Lambda, K; Z_i) = \int_0^\infty \int_0^y [S(x|Z_i) - S(y|Z_i)]K(dx, dy)$$

$$= \int_0^\infty \int_0^y \left[\{\exp(-\Lambda(x)\}^{\exp(\beta^T Z)} - \{\exp(-\Lambda(y))\}^{\exp(\beta^T Z)}\right] K(dx, dy).$$

The E-step can be modified as

$$E_{m_i}\left\{E\left[\sum_{l=1}^{m_i} I_{[T_{il}^*=t_j]}\middle| O_i\right]\right\}$$

$$= E_{m_i}\left\{m_i f_i(t_j)[1-K(t_j,\infty)+K(\infty,t_j)]/[1-\alpha(\beta,\Lambda,K;Z_i)]\right\} = \frac{f_i(t_j)[1-K(t_j,\infty)+K(\infty,t_j)]}{\alpha(\beta,\Lambda,K;Z_i)}.$$

Thus, given $\hat{K}_n$, w_{ij} can be estimated by

$$\hat{w}_{ij} = I_{[T_i=t_j]} + \frac{f_i(t_j)[1-\hat{K}_n(t_j,\infty)+\hat{K}_n(\infty,t_j)]}{\alpha(\beta,\Lambda,\hat{K}_n;Z_i)}.$$

The rest of algorithm is the same as the case when D is a constant. Similarly, the semiparametric approach can also be extend to the case when the distribution of $K(x,y)$ can be parameterized as $K(x,y;\theta)$.

In some cases, the Cox model may not fit adequately and other alternative models may provide more precise summarization of data. The semiparametric transformation models (Cheng et al. (1995), Chen et al. (2002), Zeng and Lin (2006)) have been proposed to allow various nonproportional hazards structures, such as proportional odds (Bennett (1983), Pettitt (1984)). Further research is required to extend the proposed method to semiparametric transformation models.

References

Anderson, E. B., (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of Royal Statistical Society Series B,* **32**, 283-301.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), Statistical Methods Based on Counting Processes, New York: Springer.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, **2**, 273-277.

Bilker, W. B. and Wang, M.-C. (1996). A semiparametric extension oft the Mann-Whitney test for randomly truncated data. *Biometrics*, **52**, 10-20.

Chen, K., Jin, Z. and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, **89**, 659-668.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, **82**, 4, 835-845.

Cox, D. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society* **B**, **34**, 187-220.

Cox D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Efron, B. and Petrosian, V., (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association*, **94**, 824-834.

Kalbfleish, J. D. and Lawless, J. F. (1989). Inferences based of retrospective ascertainment: An analysis of the data on transfusion related AIDS. *Journal of the American Statistical Association*, **84**, 360-372.

Emura, T., Konno, Y. and Michimae, H. (2015a). Statistical inference based on the nonparametric maximum likelihood estimator under double-truncation *Lifetime Data Analysis*, 21, 397-418.

Emura, T., Hu, Y.-H. and Konno, Y. (2015b). Asymptotic inference for maximum likelihood estimators under the special exponential family with double-truncation. *Statical Papers*, DOI 10.1007/s00362-015-0730-y.

Hu Y.-H. and Emura, T. (2015). Maximum likelihood estimation for a special exponential family under random double-truncation. *Computational Statistics*, **30**, 1199-1229.

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

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monograph National Royal Astronomical Society* **155**, 95-118.

Moreira, C. and de Uña-Álvarez, J. (2010a). Bootstrapping the NPMLE for doubly truncated data. . *Journal of Nonparametric Statistics*, **22**, No. 5, 567-583.

Moreira, C. and de Uña-Álvarez, J. (2010b). A semiparametric estimator of survival

for doubly truncated data. *Statistics in Medicine*, **29**, Issue 30, 3147-3159.

Moreira, C., de Uña-Álvarez, J. and Rosa M. Crujeiras, R. M. (2010b). DTDA: An R package to analyze randomly truncated data. *Journal of Statistical Software*, **37**, Issue 7, 1-20.

Moreira, C. and Van Keilegom, I. (2013). Bandwidth selection for kernel density estimation with doubly truncated data. *Computational Statistics and Data Analysis*, **61**, 107-123.

Moreira, C. and de Uña-Álvarez, J., and Van Keilegom, I. (2014). Goodness-of-fit tests for a semiparametric model under random double truncation. *Computational Statistics*, **29**, 1365-1379.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*, **22**, 712-731.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics*, **23**, 182-198.

Pettitt, A. N. (1984). Proportional odds model for survival data and estimates using ranks. *Journal of the Royal Statistical Society, Series C* **33**, 169-175.

Qin, J., Ning, J., Liu, H. and Shen, Y. (2011). Maximum likelihood estimations and EM algorithms with length-biased data. *Journal of the American Statistical Association*, **106**, 1434-1449.

Ren, J. and Zhou, M. (2011). Full likelihood inferences in the Cox model: an empirical approach. *Annals of the Institute Statistical Mathematics*, **63**, 1005-1018.

Shen, Y., Ning, J. and Qin, J. (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time model. *Journal of the American Statical Association*, **104**, 1192-1202.

Shen, P.-S. (2003). The product-limit estimate as an inverse-probability-weighted average. *Communications in Statistics: Theory and Methods*, **32**, 1119-1133.

Shen, P.-S. (2010a). Nonparametric analysis of doubly truncated data. *Annals of the Institute Statistical Mathematics*, **62**, No 5, 835-853.

Shen, P.-S. (2010b). Semiparametric analysis of doubly truncated data. *Communications in Statistics-Theory and Methods*, **39**, 3178-3190.

Shen, P.-S. (2013). Regression analysis of interval censored and doubly truncated data with linear transformation models. *Computational Statistics*, **28**, 581-596.

Shen, P.-S. (2016). Analysis of transformation models with doubly truncated data. *Statistical Methodology*, **30**, 15-30.

Wang, M.-C., (1989). A Semiparametric model for randomly truncated data. *Journal of the American Statistical Association*, 84, 742-748.

Wang M.-C., Brookmeyer, R. and Jewell N. P. (1993). Statistical models for prevalent cohort data, *Biometrics*, **49**, 1-11.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Annals of Statistics*, **13**, 163-177.

Zeng, D. and Lin, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* **93**, 627-640.

Zhang, Z., Sun, L., Zhao, X. and Sun, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics*, **33**, 61-70.

Zhang, X. (2015). Nonparametric inference for an inverse-probability-weighted estimator with doubly truncated data. *Communications in Statistics: Simulation and Computation*, **44**, 489-504.