Statistics and Application 统计学与应用, 2022, 11(5), 1168-1179
Published Online October 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.115120

Hans汉斯

右删失数据条件密度函数的非参数估计

薛　婧

山西财经大学统计学院，山西　太原

收稿日期：2022年9月19日；录用日期：2022年10月9日；发布日期：2022年10月25日

摘　要

本文考虑响应变量受到随机右删失的回归模型，在右删失数据与未受删失影响数据条件独立的情况下，构造了一种右删失数据条件密度函数的非参数估计量，进而得到该估计量的一致强相合性及其收敛速度，最后通过模拟探究了估计量的估计效果。

关键词

右删失数据，非参数估计，条件密度，一致强相合性

Nonparametric Conditional Density Estimation for Right-Censored Data

Jing Xue

School of Statistics, Shanxi University of Finance and Economics, Taiyuan Shanxi

Received: Sep. 19th, 2022; accepted: Oct. 9th, 2022; published: Oct. 25th, 2022

Abstract

In this paper, we consider a regression model in which the response is subject to random right censored. Under the assumption that survival time and censored time are conditional independent, we construct a nonparametric conditional density estimator. The uniform strong consistency of the conditional density estimator and its convergence rate are derived. Finally, we study the accuracy of the estimator.

文章引用: 薛婧. 右删失数据条件密度函数的非参数估计[J]. 统计学与应用, 2022, 11(5): 1168-1179.
DOI: 10.12677/sa.2022.115120

Keywords

Right-Censored Data, Nonparametric Estimation, Conditional Density, Strong Consistency

1. 引言

研究响应变量和解释变量之间的关系是统计分析中重要的一部分，在研究中通常建立回归函数来解释这种关系。在生存分析或可靠性研究中经常会遇到右删失数据，因此有大量的文献研究右删失数据的非参数模型。本文中记解释变量为 X，响应变量为 Y，T 为未受到删失的变量，C 为删失变量，观测样本记 Y_i，$Y_i=\min\{T_i,C_i\}$，此处 $Y_i=\min\{T_i,C_i\}$ 以及 $\delta_i=\mathbb{I}_{\{T_i\le C_i\}}$，$\mathbb{I}_A$ 为集合 A 的示性函数。在一般情况下，删失变量 C 与解释变量和响应变量 (X,T) 并不是独立的，但在给定 X 的情况下，C 与 T 条件独立，即 $T\mid X$ 与 $C\mid X$ 独立。

在非参数统计推断领域中，条件密度函数的估计不仅在给定解释变量时对响应变量的预测发挥着重要作用，并且它是探索解释与响应变量之间所有关系的基本工具。Tjøstheim (1994) [1]和 Polonik and Yao (2000) [2]对条件密度函数进行直接估计。Hyndman et al. (1996) [3]中提出了条件密度的核估计方法。Hyndman 和 Yao (2002) [4]提出了对于条件密度估计中选择平滑系数简单且有效的方法。Gooijer 和 Zerom (2003) [5]的 Nadaraya-Watson 估计量。但这些文章中所研究的都是完整的观测值，删失变量的非参数估计可以回溯到 Kaplan 和 Meier (1958) [6]提出 T 的生存函数，之后多位学者证明了此估计量的弱收敛性、一致收敛速度等性质。在 Beran (1981) [7]了在右删失数据下条件密度函数的非参数回归估计估计量。Gonzalez (1994) [8]此估计量的一些渐近性质及应用。Khardani 和 Semmar (2014) [9]在 C 与 (X,T) 独立情况下非参数条件密度函数核估计，但其对 $(C_i)_i$ 和 $(X_i,T_i)_i$ 的独立性假设太强。因此，本文将其独立性假设放松为 $(C_i)_i$ 和给定 $(X_i)_i$ 的 $(T_i)_i$ 之间的独立性假设。

本文研究的主要目的是当响应变量 Y 为右删失数据时，构造条件密度函数的非参数估计量，证明估计量的一致强相合性并得出收敛速度。在第二节中，给出本文构造的条件密度函数估计量；第三节中给出本文用到的假设及主要结论，并给出证明；最后对构造的估计量进行模拟，计算其估计误差来验证估计效果。

2. 估计量的构造

考虑n对定义在 $\mathbb{R}^d\times\mathbb{R}$ 上的独立随机变量 $(X_i,T_i)(i=1,2,\cdots,n)$，$(C_i)_{i=1,\cdots,n}$ 为独立同分布的删失随机变量，且服从未知连续分布函数G。当假设 (X_i,T_i) 与 C_i 独立时，删失变量的生存函数估计量在Kaplan和Meier (1958) [6]中被定义为

$$\bar{G}_n(t)=\begin{cases}\prod_{i=1}^{n}\left(1-\dfrac{1-\delta_{(i)}}{n-i+1}\right)^{\mathbb{I}_{\{Y_{(i)}\le t\}}}, & t<Y_{(n)}\\ 0, & t\ge Y_{(n)}\end{cases}$$

本文考虑在响应变量 Y_i 是右删失时，研究给定 $X=x$ 时T的非参数条件密度估计。删失变量的条件生

存函数在Gonzalez (1994) [8]中构造的一致收敛到 $\bar{G}(t|x)$ 的估计量被定义为

$$\bar{G}_n(t|x)=\begin{cases}\prod_{i=1}^{n}\left[\dfrac{1-\sum_{j=1}^{n}\mathbb{I}_{\{Y_j\le Y_i\}}w_j^{NW}(x)}{1-\sum_{j=1}^{n}\mathbb{I}_{\{Y_j<Y_i\}}w_j^{NW}(x)}\right]^{\beta_i(t)}, & t<Y_{(n)}\\ 0, & t\ge Y_{(n)}\end{cases} \tag{1}$$

此处 $\beta_i(t)=\mathbb{I}_{\{Y_i\le t,\delta_i=1\}}$，$w_i^{NW}(x)=\dfrac{h_n^{-1}K\left(\dfrac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n}h_n^{-1}K\left(\dfrac{x-X_i}{h_n}\right)}$。

对于 (X,Y,δ) 的观测值 $(X_i,Y_i,\delta_i),\cdots,(X_n,Y_n,\delta_i)$，本文定义条件密度函数 $\phi(t|x)$ 的非参数估计为

$$\forall x\in\mathbb{R}^{\mathrm{d}},\forall t\in\mathbb{R}\quad \hat{\phi}_n(t|x)=\frac{\sum_{i=1}^{n}h_n^{-(d+1)}\delta_i\bar{G}_n^{-1}(Y_i|X_i)L\left(\dfrac{t-Y_i}{h_n}\right)K\left(\dfrac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n}h_n^{-d}K\left(\dfrac{x-X_i}{h_n}\right)}=:\frac{\hat{g}_n(x,t)}{\ell_n(x)} \tag{2}$$

此处 $\tilde{g}_n(x,t):=\dfrac{1}{n}\sum_{i=1}^{n}h_n^{-(d+1)}\delta_i\bar{G}^{-1}(Y_i|X_i)L\left(\dfrac{t-Y_i}{h_n}\right)K\left(\dfrac{x-X_i}{h_n}\right)$，$\ell_n(x):=\dfrac{1}{n}\sum_{i=1}^{n}h_n^{-d}K\left(\dfrac{x-X_i}{h_n}\right)$。

3. 假设与主要结论

3.1. 假设

本文定义 $F(\cdot)$ 和 $G(\cdot)$ 分别为 T 和 C 的分布函数且，τ_F 和 τ_G 为生存函数 $\bar{F}$ 和 $\bar{G}$ 的上端点，假设 $\tau_F<\infty,\bar{G}(\tau_F)>0$ 并且 C 与 T 条件独立，即 $T|X$ 与 $C|X$ 独立。我们假设存在一个紧集 $\mathcal{C}\subset\mathcal{C}_0=\{x\in\mathbb{R}^{\mathrm{d}},\ell(x)>0\}$，此处的 $\ell(X)$ 是解释变量 X 的边缘密度函数，Ω 也是一个紧集使得 $\Omega\subset(-\infty,\tau],\tau<\tau_F\wedge\tau_G$。对于平滑系数 h，我们设定 $h_n\to0(n\to\infty)$。在推导中为了方便表示，将 M 定义为任意的正常数。

假设 A. 核函数 K 和 L 是 Lipschitz 连续函数并且满足紧支撑，且对所有的 $l=1,\cdots,d$，$u=(u_1,\cdots,u_d)^{\mathrm{T}}$，有

$$\int_{\mathbb{R}^{\mathrm{d}}}u_lK(u)\mathrm{d}u=0,\int_{\mathbb{R}}vL(v)\mathrm{d}v=0. \tag{3}$$

假设 B.

1) X 的边缘密度函数 $\ell(\cdot)$ 二阶可微且满足 Lipschitz 条件，对所有的 $x\in\mathcal{C}$ 且 $\Gamma>0$，有 $\ell(x)>0$。

2) (X,T) 的联合密度函数为 $g(\cdot,\cdot)$ 有界函数且二阶可导。

假设 C.

1) 数列 h_n 满足 $\lim_{n\to\infty}h_n+\log n/nh_n^d=0$。

2) 对一些 $\beta>0$，有 $\lim_{n\to\infty}n^{\beta}h_n=\infty$。

3.2. 一致强相合性及收敛速度

定理 3.1. 在假设 **A**，**B** 和 **C** 下，得到

$$\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\hat{\phi}_n(t\mid x)-\phi(t\mid x)\right|=O\left\{\max\left(\sqrt{\frac{\log n}{nh_n^{d+1}}},h_n^2\right)\right\}\quad \text{a.s.}\quad (n\to\infty)$$

证明：令

$$\tilde{g}_n(x,t):=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n^{(d+1)}}\delta_i\bar{G}^{-1}(Y_i\mid X_i)L\left(\frac{t-Y_i}{h_n}\right)K\left(\frac{x-X_i}{h_n}\right)$$

定理 3.1 的证明在以下分解的基础上进行

$$\begin{aligned}&\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\hat{\phi}_n(t\mid x)-\phi(t\mid x)\right|\\&\le\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\frac{\hat{g}_n(x,t)}{\ell_n(x)}-\frac{\tilde{g}_n(x,t)}{\ell_n(x)}\right|+\left|\frac{\tilde{g}_n(x,t)}{\ell_n(x)}-\frac{g_n(x,t)}{\ell_n(x)}\right|+\left|\frac{g(x,t)}{\ell_n(x)}-\frac{g(x,t)}{\ell(x)}\right|\\&\le\frac{1}{\inf_{x\in\mathcal{C}}\ell_n(x)}\left\{\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\hat{g}_n(x,t)-\tilde{g}_n(x,t)\right|+\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\tilde{g}_n(x,t)-g(x,t)\right|\right.\\&\left.+\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\phi(t\mid x)\right|\sup_{x\in\mathcal{C}}\left|\ell(x)-\ell_n(x)\right|\right\}\end{aligned}\tag{4}$$

引理 3.2. 在假设 **A**，**B** (2)和 **C** 下，可以得到

$$\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\tilde{g}_n(x,t)-g(x,t)\right|=O\left\{\max\left(\sqrt{\frac{\log n}{nh_n^{d+1}}},h_n^2\right)\right\}\quad \text{a.s.}\quad (n\to\infty).$$

引理 3.3. 在假设 **A**，**B** (1)和 **C** 下，可以得到

$$\sup_{x\in C}\left|\ell(x)-\ell_n(x)\right|=O\left\{\max\left(\sqrt{\frac{\log n}{nh_n^{d}}},h_n^2\right)\right\}\quad \text{a.s.}\quad (n\to\infty).$$

引理 3.4. 在假设 **A**，**B** (2)和 **C** 下，可以得到

$$\sup_{x\in C}\sup_{t\in\Omega}\left|\hat{g}_n(x,t)-\tilde{g}_n(x,t)\right|=O\left(\sqrt{\frac{\log n}{nh_n}}+h_n^2\right)\quad \text{a.s.}\quad (n\to\infty).$$

4. 模拟

在这一部分中通过计算均方误差来观察估计量 $\hat{\phi}(t\mid x)$ 的估计效果。本文利用 R 语言生成 (X_i,T_i,Y_i,δ_i) 的 n 组随机数，其中解释变量 $X_i\sim N(1,0.5^2)$，$\log T_i=X_i+e_i$，其中 $e_i\sim N(0,1)$，$\log C_i=X_i+\epsilon_i$，其中 $\epsilon_i\sim N(\mu,0.11^2)$，则我们可以获得 n 组观测值 (X_i,Y_i,δ_i)，此处 $Y_i=\min\{T_i,C_i\}$ 以及 $\delta_i=\mathbb{I}_{\{T_i\le C_i\}}$。

Spierdijk (2008) [10]中利用期望删失数据比例来计算 μ，使得可以借助 μ 来控制删失数据率。在本节模拟中设定删失数据率分别为 10%，30%，50%，计算后 $\mu=1.551,0.63,0$。这与 Kim (2010) [11]中的设定相似。

在计算中 $K(x)$ 和 $L(t)$ 均使用 Gaussian 核函数，估计量的均方误差计算公式为

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\phi}_n(Y_i\mid X_i)-\phi(Y_i\mid X_i)\right)^2.$$

删失比例分别为 10%、30%和 50%的情况下估计量的最小均方误差如表 1 所示。在计算过程中，我们生成取值在 $(0.05,2.00)$ 上，间隔为 0.01 的平滑系数，计算出使得估计量均方误差最小的平滑系数。从表 1 可以看出，随着删失比例的下降或样本容量的增加，估计量的最小均方误差减小，这与我们的预期相符，因为删失比例的下降或样本容量的增加意味着数据信息的增加。

Table 1. Minimum MSEs of estimator in the different censored rate
表 1. 不同删失比例下估计量的最小均方误差

删失比例	n	MSE
10%	100	0.009448
	150	0.007624
	200	0.004927
30%	100	0.016942
	150	0.011483
	200	0.010387
50%	100	0.024821
	150	0.016092
	200	0.013945

5. 中间结果的证明

5.1. 引理 3.2 的证明

证明：此处将其分解为两个部分

$$\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\tilde{g}_n(x,t)-g(x,t)\right|\le\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\tilde{g}_n(x,t)-\mathbb{E}\tilde{g}_n(x,t)\right|+\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\mathbb{E}\tilde{g}_n(x,t)-g(x,t)\right| \tag{5}$$

首先来证明

$$\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\mathbb{E}\tilde{g}_n(x,t)-g(x,t)\right|=O\left(h_n^2\right).$$

由于 $Y_i=\min\{T_i,C_i\}$，对于所有的可测函数 ϕ 以及所有的 $i=1,\cdots,n$，有

$$\mathbb{I}_{\{T_1\le C_1\}}\varphi(Y_1)=\mathbb{I}_{\{T_1\le C_1\}}\varphi(T_1).$$

因此可以得到

$$\mathbb{E}\tilde{g}_n(x,t)=n^{-1}\sum_{i=1}^{n}\frac{1}{h_n^{d+1}}\mathbb{E}\left\{\delta_i\bar{G}^{-1}(T_i\mid X_i)L\left(\frac{t-T_i}{h_n}\right)K\left(\frac{x-X_i}{h_n}\right)\right\}$$

$$=n^{-1}\sum_{i=1}^{n}\frac{1}{h_n^{d+1}}\mathbb{E}\left\{\mathbb{E}\left[\mathbb{I}_{\{T_i\le C_i\}}\mid X_i,T_i\right]\bar{G}^{-1}(T_i\mid X_i)L\left(\frac{t-T_i}{h_n}\right)K\left(\frac{x-X_i}{h_n}\right)\right\}.$$

由 $\bar{G}(T_i\mid X_i)=1-P(C_i\le T_i)$ 以及 $\mathbb{E}\left[\mathbb{I}_{\{T_i\le C_i\}}\mid X_i,T_i\right]=P(T_i\le C_i)$，有

$$\mathbb{E}\tilde{g}_n(x,t)=n^{-1}\sum_{i=1}^{n}\frac{1}{h_n^{d+1}}\mathbb{E}\left\{L\left(\frac{t-T_i}{h_n}\right)K\left(\frac{x-X_i}{h_n}\right)\right\}$$

$$=n^{-1}\sum_{i=1}^{n}\int_{\mathbb{R}^d}\int_{\mathbb{R}}L\left(\frac{t-T_i}{h_n}\right)K\left(\frac{x-X_i}{h_n}\right)g(X_i,T_i)\mathrm{d}X_i\mathrm{d}T_i.$$

通过换元 $u=(x-X_i)/h_n$，$v=(t-T_i)/h_n$，并将 g 在 (x,t) 处泰勒展开

$$\begin{aligned}\mathbb{E}\tilde{g}_n(x,t) &= n^{-1}\sum_{i=1}^{n}\int_{\mathbb{R}^d}\int_{\mathbb{R}} K(u)L(v)g(x-h_nu,t-h_nv)\mathrm{d}u\mathrm{d}v \\ &= n^{-1}\sum_{i=1}^{n}\int_{\mathbb{R}^d}\int_{\mathbb{R}} K(u)L(v)\Big[g(x,t)-uh_ng_1'-h_nvg_2'+\frac{1}{2}h_n^2u^2g_{11}'' \\ &\quad +\frac{1}{2}h_n^2uvg_{12}''+\frac{1}{2}h_n^2uvg_{21}''+\frac{1}{2}h_n^2v^2g_{22}''\Big]\mathrm{d}u\mathrm{d}v.\end{aligned}$$

由假设可知

$$\int K(u)\mathrm{d}u=1,\int L(v)\mathrm{d}v=1,$$

$$\int uK(u)\mathrm{d}u=0,\int vL(v)\mathrm{d}v=0.$$

于是有

$$\begin{aligned}\left|\mathbb{E}\tilde{g}_n(x,t)-g(x,t)\right| &\le n^{-1}\sum_{i=1}^{n}\left|\int_{\mathbb{R}^d}\int_{\mathbb{R}} K(u)L(v)\left[g(x-h_nu,t-h_nv)-g(x,t)\right]\mathrm{d}u\mathrm{d}v\right| \\ &\le Mn^{-1}\sum_{i=1}^{n}h_n^2 \\ &\le Mh_n^2.\end{aligned}$$

因此，

$$\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\mathbb{E}\tilde{g}_n(x,t)-g(x,t)\right|=O\left(h_n^2\right).$$

现在，证明另一部分

$$\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\tilde{g}_n(x,t)-\mathbb{E}\tilde{g}_n(x,t)\right|=O\left(\sqrt{\frac{\log n}{nh_n^{(d+1)}}}\right)$$

对任意的 $x\in\mathcal{C}$ 和 $t\in\Omega$，根据 $\tilde{k}(x)=\arg\min_k\left\|x_k-x\right\|$ 和 $\tilde{j}(t)=\arg\min_j\left|t-t_j\right|$ 选取 k 和 j，由于集合 $\mathcal{C}$ 和 Ω 的紧集性质，对于 $(x_k)_{1\le k\le\lambda_n}$ 和 $(t_j)_{1\le j\le\kappa_n}$，有

$$\mathcal{C}\subset\bigcup_{k=1}^{\lambda_n}\mathcal{B}(x_k,a_n)\text{ 和 }\Omega\subset\bigcup_{j=1}^{\kappa_n}B(t_j,b_n)$$

其中 $\lambda_n\sim a_n^{-d}$ 和 $\kappa_n\sim b_n^{-1}$ 并且 $a_n=b_n=h_n^{(d+3)/2}n^{-1/2}$。

在这个部分中，将其依照如下分解来分别证明

$$\begin{aligned}&\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\tilde{g}_n(x,t)-\mathbb{E}\tilde{g}_n(x,t)\right| \\ &\le\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\tilde{g}_n(x,t)-\tilde{g}_n(x,t_{\tilde{j}})\right|+\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\mathbb{E}\tilde{g}_n(x,t_{\tilde{j}})-\mathbb{E}\tilde{g}_n(x,t)\right| \\ &\quad+\max_j\sup_{x\in\mathcal{C}}\left|\tilde{g}_n(x,t_j)-\tilde{g}_n(x_{\tilde{k}},t_j)\right|+\max_j\sup_{x\in\mathcal{C}}\left|\mathbb{E}\tilde{g}_n(x_{\tilde{k}},t_j)-\mathbb{E}\tilde{g}_n(x,t_j)\right| \\ &\quad+\max_k\max_j\left|\tilde{g}_n(x_k,t_j)-\mathbb{E}\tilde{g}_n(x_k,t_j)\right| \\ &=:\mathcal{T}_{1,n}+\mathcal{T}_{2,n}+\mathcal{T}_{3,n}+\mathcal{T}_{4,n}+\mathcal{T}_{5,n}.\end{aligned}\tag{6}$$

首先对于$\left(\mathcal{T}_{1,n}\right)$：运用核函数 L 的 Lipschitzian 条件可以得到

$$\begin{aligned}
&\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\tilde{g}_n\left(x,t\right)-\tilde{g}_n\left(x,t_{\tilde{j}}\right)\right| \\
&\le \sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\frac{1}{n}\sum_{i=1}^{n}h_n^{-(d+1)}\delta_i\bar{G}^{-1}\left(Y_i\mid X_i\right)K_i\left(x\right)\left|L_i\left(t\right)-L_i\left(t_{\tilde{j}}\right)\right| \\
&\le \sup_{x\in\mathcal{C}}\sup_{t\in\Omega}C\left|t-t_{\tilde{j}}\right|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}h_n^{-(d+1)}\delta_i\bar{G}^{-1}\left(Y_i\mid X_i\right)K_i\left(x\right).
\end{aligned}$$

由假设可知 $K\left(u\right)$ 是紧支撑的，并且 $\bar{G}\left(t\mid x\right)$ 是删失变量的生存函数，则 $K\left(u\right)$ 与 $\bar{G}\left(t\mid x\right)$ 均为有界，因此

$$\begin{aligned}
\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\tilde{g}_n\left(x,t\right)-\tilde{g}_n\left(x,t_{\tilde{j}}\right)\right| &\le Mb_n\frac{1}{n}\sum_{i=1}^{n}h_n^{-(d+2)} \\
&\le M\frac{b_n}{h_n^{(d+2)}} \\
&= M\frac{h_n^{(d+3)/2}n^{-1/2}}{h_n^{(d+2)}} \\
&= M\sqrt{\frac{1}{nh_n^{d+1}}}.
\end{aligned}$$

由假设有

$$\left(\mathcal{T}_{1,n}\right)=O\left(\sqrt{\frac{1}{nh_n^{d+1}}}\right).$$

下面对于$\left(\mathcal{T}_{2,n}\right)$

$$\begin{aligned}
&\sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\left|\mathbb{E}\tilde{g}_n\left(x,t_{\tilde{j}}\right)-\mathbb{E}\tilde{g}_n\left(x,t\right)\right| \\
&\le \sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\frac{1}{n}\sum_{i=1}^{n}h_n^{-(d+1)}\left|\mathbb{E}\left\{\delta_i\bar{G}^{-1}\left(Y_i\mid X_i\right)K_i\left(x\right)L_i\left(t_{\tilde{j}}\right)\right\}-\mathbb{E}\left\{\delta_i\bar{G}^{-1}\left(Y_i\mid X_i\right)K_i\left(x\right)L_i\left(t\right)\right\}\right| \\
&\le \sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\frac{1}{n}\sum_{i=1}^{n}h_n^{-(d+1)}\left|\int_{\mathbb{R}^{\mathrm{d}}}\int_{\mathbb{R}}K_i\left(x\right)g\left(X_i,T_i\right)\left\{L_i\left(t_{\tilde{j}}\right)-L_i\left(t\right)\right\}\mathrm{d}X_i\mathrm{d}T_i\right| \\
&\le \sup_{x\in\mathcal{C}}\sup_{t\in\Omega}\frac{1}{n}\sum_{i=1}^{n}h_n^{-(d+1)}\left|\int_{\mathbb{R}^{\mathrm{d}}}\int_{\mathbb{R}}K_i\left(x\right)g\left(X_i,T_i\right)\right|t_{\tilde{j}}-t\left|\mathrm{d}X_i\mathrm{d}T_i\right| \\
&\le \sup_{x\in\mathcal{C}}\sup_{t\in\Omega}C\left|t_{\tilde{j}}-t\right|\frac{1}{n}\sum_{i=1}^{n}h_n^{-(d+2)}\mathbb{E}\left\{K_i\left(x\right)\right\} \\
&\le M\frac{b_n}{h_n^{d+1}}.
\end{aligned}$$

之后同理可得

$$\left(\mathcal{T}_{2,n}\right)=O\left(\sqrt{\frac{1}{nh_n^{d+1}}}\right).$$

下面对于$\left(\mathcal{T}_{3,n}\right)$

$$\begin{aligned}
&\max_j \sup_{x\in\mathcal{C}} \left|\tilde{g}_n\left(x,t_j\right)-\tilde{g}_n\left(x_{\tilde{k}},t_j\right)\right| \\
&\le \max_j \sup_{x\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} h_n^{-(d+1)}\delta_i \bar{G}^{-1}\left(Y_i \mid X_i\right) L_i\left(t_j\right)\left|K_i(x)-K_i\left(x_{\tilde{k}}\right)\right| \\
&\le \max_j \sup_{x\in\mathcal{C}} C\left|x-x_{\tilde{k}}\right| \frac{1}{n}\sum_{i=1}^{n} h_n^{-(d+1)}\delta_i \bar{G}^{-1}\left(Y_i \mid X_i\right) L_i\left(t_j\right) \\
&\le Ma_n \frac{1}{n}\sum_{i=1}^{n} h_n^{-(d+2)} \\
&\le M\frac{a_n}{\left(h_n^-\right)^{d+2}}.
\end{aligned}$$

之后同理可得

$$\left(\mathcal{T}_{3,n}\right)=O\left(\sqrt{\frac{1}{nh_n^{d+1}}}\right).$$

之后对于$\left(\mathcal{T}_{4,n}\right)$

$$\begin{aligned}
&\max_j \sup_{x\in\mathcal{C}} \left|\mathbb{E}\tilde{g}_n\left(x_{\tilde{k}},t_j\right)-\mathbb{E}\tilde{g}_n\left(x,t_j\right)\right| \\
&\le \max_j \sup_{x\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} h_n^{-(d+1)} \left|\mathbb{E}\left\{\delta_i \bar{G}^{-1}\left(Y_i \mid X_i\right) K_i\left(x_{\tilde{k}}\right) L_i\left(t_j\right)\right\}-\mathbb{E}\left\{\delta_i \bar{G}^{-1}\left(Y_i \mid X_i\right) K_i(x) L_i\left(t_j\right)\right\}\right| \\
&\le \max_j \sup_{x\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} h_n^{-(d+1)} \mathbb{E}\left\{L_i\left(t_j\right)\left|K_i\left(x_{\tilde{k}}\right)-K_i(x)\right|\right\} \\
&\le \max_j \sup_{x\in\mathcal{C}} C\left|x_{\tilde{k}}-x\right| \frac{1}{n}\sum_{i=1}^{n} h_n^{-(d+1)} \mathbb{E}\left\{L_i\left(t_j\right)\right\} \\
&\le Ma_n \frac{1}{n}\sum_{i=1}^{n} h_n^{-(d+2)}.
\end{aligned}$$

同理可得

$$\left(\mathcal{T}_{4,n}\right)=O\left(\sqrt{\frac{1}{nh_n^2}}\right).$$

最后对$\left(\mathcal{T}_{5,n}\right)$进行证明，对所有的$1\le i\le n, 1\le k\le \lambda_n, 1\le j\le \kappa_n$，令

$$U_i=U_i\left(x_k,t_j\right):=\left\{h_n^{-(d+1)}\delta_i \bar{G}^{-1}\left(Y_i \mid X_i\right) K_i(x) L_i(t)-\mathbb{E}\left[h_n^{-(d+1)}\delta_i \bar{G}^{-1}\left(Y_i \mid X_i\right) K_i(x) L_i(t)\right]\right\}.$$

由于核函数K和L以及$g(\cdot,\cdot)$均为有界函数，则有

$$\left|U_i\right|\le \frac{C}{h_n^{(d+1)}}=M.$$

以及

$$\begin{aligned}
\operatorname{Var}\left(U_i\right)&=\mathbb{E}\left(U_i^2\right)-\mathbb{E}^2\left(U_i\right) \\
&\le \frac{C}{h_n^{d+1}}\left\{\mathbb{E}\left[K_i^2(x) L_i^2(t)\right]-\mathbb{E}^2\left[K_i(x) L_i(t)\right]\right\} \\
&\le Ch_n^{d+1}\int_{\mathbb{R}^d}\int_{\mathbb{R}} K^2(u) L^2(v) g\left(x_k-uh_n,t_j-vh_n\right)\mathrm{d}u\mathrm{d}v \\
&\le \frac{C}{h_n^{(d+1)}}:=\sigma^2.
\end{aligned}$$

在此简要介绍 Berstein 不等式 Hoeffding (1994) [12] (式 2.13)：$X_i,\cdots,X_n$ 为 X 的独立样本，$\mathbb{E}X=0$，$\mathbb{E}X^2=\sigma^2$，$|X|<M$，M 为常数，且 $t\geq 0$，则

$$\mathbb{P}\left(n^{-1}\sum_{i=1}^{n}X_i\geq t\right)\leq \exp\left(-\left(\frac{nt}{M}\right)h\left(\frac{Mt}{\sigma^2}\right)\right) \tag{7}$$

其中对于所有的 $u>0$，有 $h(u)=3u/(6+2u)$。

因此，使用 Berstein 不等式，对于所有的 $\epsilon>0$ 有

$$\mathbb{P}\left\{\left|n^{-1}\sum_{i=1}^{n}U_i\right|>\epsilon\right\}\leq 2\exp\left\{-\left(\frac{n\epsilon}{M}\right)h\left(\frac{M\epsilon}{\sigma^2}\right)\right\}.$$

现在，令 $\epsilon=\epsilon_0\left(\log n/nh_n^{-(d+1)}\right)^{1/2}$，则对于任意的 (k,j) 可以得到

$$\begin{aligned}\mathbb{P}\left\{\left|n^{-1}\sum_{i=1}^{n}U_i\right|>\epsilon\right\}&\leq 2\exp\left\{\frac{2n\epsilon^2}{6\sigma^2+2M\epsilon}\right\}\\&\leq 2\exp\left\{-\frac{3\epsilon_0^2\log n}{3c+c_0}\right\}\\&\leq 2n^{-C\epsilon_0^2}.\end{aligned}$$

接下来利用 $\mathbb{P}(\cup_i A_i)\leq\sum_i\mathbb{P}(A_i)$，可以得到

$$\begin{aligned}&\mathbb{P}\left\{\max_{k=1,\cdots,\lambda_n}\max_{j=1,\cdots,\kappa_n}\left|\frac{1}{n}\sum_{i=1}^{n}U_i(x_k,t_j)\right|>\epsilon\right\}\\&\leq\sum_{k=1}^{\lambda_n}\sum_{k=1}^{\kappa_n}\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i(x_k,t_j)\right|>\epsilon\right\}\\&\leq C\lambda_n\kappa_n n^{-C\epsilon_0^2}.\end{aligned}$$

因此，选取合适的 ϵ_0 使得 $\lambda_n\kappa_n n^{-C\epsilon_0^2}$ 收敛，并且运用 Borel-Cantelli 引理可以得到

$$(\mathcal{T}_{5,n})=O\left(\sqrt{\frac{\log n}{nh_n^{(d+1)}}}\right)$$

引理 3.2 证得。

5.2. 引理 3.3 的证明

证明：首先，将其分解为两部分

$$\begin{aligned}\sup_{x\in\mathcal{C}}|\ell_n(x)-\ell(x)|&\leq\sup_{x\in\mathcal{C}}|\ell_n(x)-\mathbb{E}[\ell_n(x)]|+\sup_{x\in\mathcal{C}}|\mathbb{E}[\ell_n(x)]-\ell(x)|\\&=:\mathcal{L}_{1n}+\mathcal{L}_{2n}.\end{aligned} \tag{8}$$

第一部分 $\mathcal{L}_{1n}$ 与引理 3.2 类似，即

$$\begin{aligned}\mathcal{L}_{1n}&=\sup_{x\in\mathcal{C}}|\ell_n(x)-\mathbb{E}[\ell_n(x)]|\\&=\sup_{x\in\mathcal{C}}|\ell_n(x)-\ell_n(x_k)|+\sup_{x\in\mathcal{C}}|\mathbb{E}[\ell_n(x_k)]-\mathbb{E}[\ell_n(x)]|+\max_k|\ell_n(x_k)-\mathbb{E}[\ell_n(x_k)]|\\&=\mathcal{S}_{1,n}+\mathcal{S}_{2,n}+\mathcal{S}_{3,n}.\end{aligned}$$

对于 $(S_{1,n})$ 同样借助核函数 K 的 Lipschitzian 条件，可得

$$\begin{aligned}\mathcal{S}_{1,n} &= \sup_{x\in\mathcal{C}}\left|\ell_n(x)-\ell_n(x_k)\right| \\ &\le \sup_{x\in\mathcal{C}}\frac{1}{nh_n^d}\sum_{i=1}^{n}\left|K_i(x)-K_i(x_k)\right| \\ &\le \sup_{x\in\mathcal{C}} C\left|x-x_k\right|\frac{1}{h_n^{d+1}} \\ &\le M\frac{a_n}{h_n^{d+1}}.\end{aligned}$$

由引理 3.2 证明中对 a_n 的设定，可以得到

$$\left(S_{1,n}\right)=O\left(\sqrt{\frac{1}{nh^d}}\right).$$

对于 $\left(S_{2,n}\right)$ 同样可以得到

$$\begin{aligned}\mathcal{S}_{2,n} &= \sup_{x\in\mathcal{C}}\left|\mathbb{E}\left[\ell_n(x_k)\right]-\mathbb{E}\left[\ell_n(x)\right]\right| \\ &\le \sup_{x\in\mathcal{C}}\frac{1}{nh_n^d}\sum_{i=1}^{n}\left|\mathbb{E}\left[K_i(x_k)\right]-\mathbb{E}\left[K_i(x)\right]\right| \\ &\le \sup_{x\in\mathcal{C}}\frac{1}{nh_n^d}\sum_{i=1}^{n}\mathbb{E}\left[\left|K_i(x_k)-K_i(x)\right|\right] \\ &\le \sup_{x\in\mathcal{C}} C\left|x_k-x\right|\frac{1}{h_n^d} \\ &\le M\frac{a_n}{h_n^{d+1}}.\end{aligned}$$

因此，

$$\left(S_{2,n}\right)=O\left(\sqrt{\frac{1}{nh^d}}\right).$$

$\left(S_{3,n}\right)$ 部分同引理 3.2 的最后一部分证明，可以得到

$$\left(S_{3,n}\right)=O\left(\sqrt{\frac{\log n}{nh_n^{(d+1)}}}\right)$$

对于本引理的第二部分，同样借助泰勒展开，得到

$$\begin{aligned}\mathcal{L}_{2n} &= \sup_{x\in\mathcal{C}}\left|\mathbb{E}\left[\ell_n(x)\right]-\ell(x)\right| \\ &= \sup_{x\in\mathcal{C}}\left|\frac{1}{n}\sum_{n=1}^{n}h_n^{-d}\mathbb{E}\left[K\left(\frac{x-X_i}{h_n}\right)\right]-\ell(x)\right| \\ &\le \sup_{x\in\mathcal{C}}\frac{1}{n}\sum_{n=1}^{n}\left|\int_{\mathbb{R}^d}K(u)\ell(x-h_nu)\mathrm{d}u-\ell(x)\right| \\ &\le \sup_{x\in\mathcal{C}}\frac{1}{n}\sum_{n=1}^{n}h_n^2 \\ &\le Mh_n^2.\end{aligned}$$

则可以得到

$$\mathcal{L}_{2n} = O\left(h_n^2\right).$$

引理 3.3 证得。

5.3. 引理 3.4 的证明

证明：我们有

$$\begin{aligned}\sup_{x\in C}\sup_{t\in\Omega}\left|\hat{g}_n(x,t)-\tilde{g}_n(x,t)\right| &\le \sup_{x\in C}\sup_{t\in\Omega}\sum_{i=1}^{n}\left|\frac{1}{nh_n^{(d+1)}}\delta_i K_i(x)L_i(t)\left(\frac{1}{\bar{G}_n(Y_i\mid X_i)}-\frac{1}{\bar{G}(Y_i\mid X_i)}\right)\right| \\ &\le \frac{\sup_{x\in C}\sup_{t\le\tau_F}\left|\bar{G}_n(t\mid x)-\bar{G}(t\mid x)\right|}{\bar{G}_n(\tau\mid x)}\tilde{g}_n(x,t).\end{aligned} \tag{9}$$

又由于 $\bar{G}_n(\tau\mid x)>0$ 且 $\tilde{g}_n(x,t)$ 有界，可以得到

$$\sup_{x\in C}\sup_{t\in\Omega}\left|\hat{g}_n(x,t)-\tilde{g}_n(x,t)\right| \le M\sup_{x\in C}\sup_{t\le\tau_F}\left|\bar{G}_n(t\mid x)-\bar{G}(t\mid x)\right|.$$

根据 Gonzalez (1994) [8]中结论有

$$\sup_{x\in C}\sup_{t\in\Omega}\left|\hat{g}_n(x,t)-\tilde{g}_n(x,t)\right| = O\left(\sqrt{\frac{\log n}{nh_n}}+h_n^2\right)\quad \text{a.s.}\quad \text{as}\quad n\to\infty. \tag{10}$$

引理 3.4 得证。

6. 结论

在众多文献中对数据的独立性进行了较强的假设，即假设 C 与 (X,T) 独立，但在实际观测到的数据中其只能满足条件独立的假设。因此本文考虑在右删失数据与未受到删失影响的数据条件独立的情况下，构建了具有一致强相合性的条件密度函数的非参数估计量，并计算其收敛速度。本文在构造非参数估计量时用到了经典的 Nadaraya-Watson 核估计，对于其他核估计，例如能够减少边界效应的局部线性回归核估计，所构成的非参数估计量将在未来的工作中被讨论。

参考文献

[1] Tjøstheim, D. (1994) Non-Linear Time Series: A Selective Review. *Scandinavian Journal of Statistics*, **21**, 97-130.

[2] Polonik, W. and Yao, Q. (2000) Conditional Minimum Volume Predictive Regions for Stochastic Processes. *Journal of the American Statistical Association*, **95**, 509-519. https://doi.org/10.1080/01621459.2000.10474228

[3] Hyndman, R.J., Bashtannyk, D.M. and Grunwald, G.K. (1996) Estimating and Visualizing Conditional Densities. *Journal of Computational and Graphical Statistics*, **5**, 315-336. https://doi.org/10.1080/10618600.1996.10474715

[4] Hyndman, R.J. and Yao, Q. (2002) Nonparametric Estimation and Symmetry Tests for Conditional Density Functions. *Journal of Nonparametric Statistics*, **14**, 259-278. https://doi.org/10.1080/10485250212374

[5] De Gooijer, J.G. and Zerom, D. (2003) On Conditional Density Estimation. *Statistica Neerlandica*, **57**, 159-176. https://doi.org/10.1111/1467-9574.00226

[6] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457-481. https://doi.org/10.1080/01621459.1958.10501452

[7] Beran, R. (1981) Nonparametric Regression with Randomly Censored Survival Data. Technical Report, Technical University of California, Berkeley.

[8] Gonzalez-Manteiga W. and Cadarso-Suarez, C. (1994) Asymptotic Properties of a Generalized Kaplan-Meier Estimator with Some Applications. *Journal of Nonparametric Statistics*, **4**, 65-78. https://doi.org/10.1080/10485259408832601

[9] Khardani, S. and Semmar, S. (2014) Nonparametric Conditional Density Estimation for Censored Data Based on a

Recursive Kernel. *Electronic Journal of Statistics*, **8**, 2541-2556. https://doi.org/10.1214/14-EJS960

[10] Spierdijk, L. (2008) Nonparametric Conditional Hazard Rate Estimation: A Local Linear Approach. *Computational Statistics & Data Analysis*, **52**, 2419-2434. https://doi.org/10.1016/j.csda.2007.08.007

[11] Kim, C., Oh, M., Yang, S.J., *et al*. (2010) A Local Linear Estimation of Conditional Hazard Function in Censored Data. *Journal of the Korean Statistical Society*, **39**, 347-355. https://doi.org/10.1016/j.jkss.2010.03.002

[12] Hoeffding, W. (1994) Probability Inequalities for Sums of Bounded Random Variables. In: Fisher, N.I. and Sen, P.K., Eds., The Collected Works of Wassily Hoeffding, Springer, New York, NY, 409-426. https://doi.org/10.1007/978-1-4612-0865-5_26