東海管理評論
民國九十一年七月，第四卷，第一期，17-26

The Nonparametric MLE as an Inverse-Probability-of Truncation Weighted Average

*Pao-Sheng Shen**

Abstract

For randomly censored data, Satten and Datta (2001) showed that the Kaplan-Meier estimator can be expressed as an inverse-probability-of censoring weighted estimator. In this article, it is shown that the truncation product-limit estimate, first introduced by Lynden-Bell(1971), can also be expressed as an inverse-probability-of truncation weighted average, where the weights are related to the distribution function of truncation variables.

Keywords：Product-Limit Estimator, Random Truncation.

1. Introduction

The Kaplan-Meier estimator (product-limit estimator, PLE) for the survival function of randomly censored time-to-event data (Kaplan and Meier (1958)) is often introduced as the maximizer of a nonparametric maximum likelihood (see Kalbfleisch and Prentice (1978); Wang (1987)). In a series of papers, Robins and coworkers proposed a class of estimators using a data-reweighting scheme (Robins and Rotnitzky(1992); Robins(1993); Robins and Finkelstein(2000)). An outcome of their approach applied to survival analysis is an inverse-of probability-of censoring representation of the Kaplan-Meier estimator. Satten and Datta (2001)

* Professor of Department of Statistics, Tunghai University

give two demonstrations of this representation. In this article, it will be shown that the truncation product-limit estimate, first introduced by Lynden-Bell (1971), can also be expressed as an inverse-probability-of-truncation weighted average, where the weights are related to the distribution function of truncation variables.

2. The Truncation Product-Limit Estimator

Let U^* and V^* be the target and truncation variables with distribution functions F and G respectively. Assume that U^* and V^* are independent. Under random truncation, both U^* and V^* are observable only when $U^* \geq V^*$. Let (U_1, V_1), ..., (U_n, V_n) denote the truncated sample. Hence, $H(u,v) = P(U_i \leq u,\ V_i \leq v) = P(U^* \leq u,\ V^* \leq v \mid U^* \geq V^*)$. Let $I_{[A]}$ be the indicator function of the event A. Let $F_n^*(u) = n^{-1}\sum_{i=1}^{n} I_{[U_i \leq u]}$, $G_n^*(v) = n^{-1}\sum_{i=1}^{n} I_{[V_i \leq v]}$, and $R_n(u) = G_n^*(u) - F_n^*(u-) = n^{-1}\sum_{i=1}^{n} I_{[V_i \leq u \leq U_i]}$. The truncation product-limit estimates of F and G, first introduced by Lynden-Bell (1971), can be viewed as a nonparametric method for dealing with delayed entry of uncensored life table data, as well as truncated astronomy data (see Wang, Jewell, and Tsai (1986); Woodroofe (1985); He and Yang (1998)). The nonparametric maximum likelihood estimates (MLE) of $F(x)$ and $G(x)$ are given by

$$\hat{F}_n(x) = 1 - \prod_{u \leq x}\left[1 - \frac{F_n^*\{u\}}{R_n(u)/n}\right]$$

and

$$\hat{G}_n(x) = \prod_{v > x}\left[1 - \frac{G_n^*\{v\}}{R_n(v)/n}\right]$$

where $F_n^*\{u\}=F_n^*(u)-F_n^*(u-)$ and $G_n^*\{v\}=F_n^*(v)-F_n^*(v-)$.

Under the semiparametric model, V^* is assumed to have distribution function $G(y;\theta)$, where G is specified, $\theta\in\Theta$ and θ can be a vector. For the semiparametric model, the MLE of $F(x)$, derived by Wang (1989), is

$$\left(\sum_i \frac{1}{G(U_i;\theta)}\right)^{-1}\sum_i \frac{I_{[U_i\le x]}}{G(U_i;\theta)} \qquad \text{(2.1)}$$

Note that this weighted average (2.1) is actually the MLE described by Vardi (1985), with G a weight function. Similarly, when U^* is assumed to have distribution function $F(x;\lambda)$, where F is specified, $\lambda\in\Lambda$ and λ can be a vector. For the semiparametric model, the MLE of $G(x)$ is

$$\left(\sum_i \frac{1}{1-F(V_i;\lambda)}\right)^{-1}\sum_i \frac{I_{[V_i\le x]}}{1-F(V_i;\lambda)} \qquad \text{(2.2)}$$

In the following Sections, we give two demonstrations of the equivalence of the inverse-probability-of-truncation weighted estimator and the Lynden-Bell's (1971) estimator. First, substitution of $\hat{G}_n(U_i)$ for $G(U_i;\theta)$ in (2.1) leads to

$$\hat{F}_w(x)=\left(\sum_i \frac{1}{\hat{G}_n(U_i)}\right)^{-1}\sum_i \frac{I_{[U_i\le x]}}{\hat{G}_n(U_i)}.$$

Next, substitution of $\hat{F}_n(x)$ for $F(V_i;\lambda)$ in (2.2) leads to

$$\hat{G}_w(x)=\left(\sum_i \frac{1}{1-\hat{F}_n(V_i)}\right)^{-1}\sum_i \frac{I_{[V_i\le x]}}{1-\hat{F}_n(V_i)}.$$

In Section **3**, we show that $\hat{F}_w$ is equivalent to $\hat{F}_n$.

3. Equivalence of $\hat{F}_w$ and $\hat{F}_n$

Theorem 3.1 $\hat{F}_w = \hat{F}_n$

Proof：

Note that both $\hat{F}_w$ and $\hat{F}_n$ are step right-continuous functions. Thus, $\hat{F}_w$ and $\hat{F}_n$ are the same if the magnitudes of the jumps in the two functions are equal. The jumps occur at the distinct order statistics $U_{(1)} < U_{(2)} < \ldots < U_{(r)}$ of the sample $U_1, U_2, \ldots, U_n$. The jump in $\hat{F}_w$ at time $U_{(i)}$ is given by

$$\hat{F}_w\left(U_{(i)}\right) - \hat{F}_w\left(U_{(i-1)}\right) = \frac{d_i / \hat{G}_n\left(U_{(i)}\right)}{\sum_{j=1}^{r} d_j / \hat{G}_n\left(U_{(j)}\right)},$$

where $d_i = F_n^*\left(U_{(i)}\right) - F_n^*\left(U_{(i-1)}\right)$ for $1 \le i \le r$.

Now, by Corollary 2.4 of He and Yang（1998）, we have

$$\frac{d_i / \hat{G}_n\left(U_{(i)}\right)}{\sum_{j=1}^{r} d_j / \hat{G}_n\left(U_{(j)}\right)} = \frac{d_i \left[1 - \hat{F}_n\left(U_{(i-1)}\right)\right] / R_n\left(U_{(i)}\right)}{\sum_{j=1}^{r} d_j \left[1 - \hat{F}_n\left(U_{(j-1)}\right)\right] / R_n\left(U_{(j)}\right)}$$

Since

$$\sum_{j=1}^{r} \frac{d_j \left[1 - \hat{F}_n\left(U_{(j-1)}\right)\right]}{R_n\left(U_{(j)}\right)} = \sum_{j=1}^{r} \prod_{k=1}^{j-1} \left(\frac{R_n\left(U_{(k)}\right) - d_k}{R_n\left(U_{(k)}\right)} \right) \left(\frac{d_j}{R_n\left(U_{(j)}\right)} \right)$$

$$=\sum_{j=1}^{r}\prod_{k=1}^{j-1}\left(\frac{R_n(U_{(k)})-d_k}{R_n(U_{(k)})}\right)\left[1-\frac{R_n(U_{(j)})-d_j}{R_n(U_{(j)})}\right]$$

$$=\sum_{j=1}^{r}\left[\hat{F}_n(U_{(j)})-\hat{F}_n(U_{(j-1)})\right]=1$$

we have

$$\hat{F}_w(U_{(i)})-\hat{F}_w(U_{(i-1)})=\frac{d_i\left[1-\hat{F}_n(U_{(i-1)})\right]}{R_n(U_{(i)})}=\hat{F}_n(U_{(i)})-\hat{F}_n(U_{(i-1)}).$$

Thus, $\hat{F}_w$ and $\hat{F}_n$ are the same.

4. Equivalence of $\hat{G}_w$ and $\hat{G}_n$

Theorem 4.1 $\hat{G}_w=\hat{G}_n$

Proof：

Note that both $\hat{G}_w$ and $\hat{G}_n$ are step right-continuous functions. Thus, $\hat{G}_w$ and $\hat{G}_n$ are the same if the magnitudes of the jumps in the two functions are equal. The jumps occur at the distinct order statistics $V_{(1)}<V_{(2)}<\ldots<V_{(s)}$ of the sample $V_1,V_2,\ldots,V_n$. The jump in $\hat{G}_w$ at time $V_{(j)}$ is given by

$$\hat{G}_w(V_{(j)})-\hat{G}_w(V_{(j-1)})=\frac{f_j/\left[1-\hat{F}_n(V_{(j)})\right]}{\sum_{k=1}^{r}f_k/\left[1-\hat{F}_n(V_{(k)})\right]},$$

where $f_j=G_n^*(V_{(j)})-G_n^*(V_{(j-1)})$ for $1\le i\le r$.

Now, by Corollary 2.4 of He and Yang (1998), we have

$$\frac{f_j/\left[1-\hat{F}_n(V_{(j)})\right]}{\sum_{k=1}^{s} f_k/\left[1-\hat{F}_n(V_{(k)})\right]}=\frac{f_j\hat{G}_n(V_{(j)})/R_n(V_{(j)})}{\sum_{k=1}^{s} f_k\hat{G}_n(V_{(k)})/R_n(V_{(k)})}$$

Since

$$\sum_{k=1}^{s}\frac{f_k\left[\hat{G}_n(V_{(k)})\right]}{R_n(V_{(k)})}=\sum_{k=1}^{s}\prod_{i=1}^{k-1}\left(\frac{R_n(V_{(i)})-f_k}{R_n(V_{(i)})}\right)\left(\frac{f_k}{R_n(V_{(k)})}\right)$$

$$=\sum_{k=1}^{s}\prod_{i=1}^{k-1}\left(\frac{R_n(V_{(i)})-f_i}{R_n(V_{(i)})}\right)\left[1-\frac{R_n(V_{(i)})-f_i}{R_n(V_{(i)})}\right]$$

$$=\sum_{k=1}^{s}\left[\hat{G}_n(V_{(k)})-\hat{G}_n(V_{(k-1)})\right]=1$$

we have

$$\hat{G}_w(V_{(j)})-\hat{G}_w(V_{(j-1)})=\frac{f_j\hat{G}_n(V_{(j)})}{R_n(V_{(j)})}=\hat{G}_n(V_{(j)})-\hat{G}_n(V_{(j-1)}).$$

Thus, $\hat{G}_w$ and $\hat{G}_n$ are the same.

5. Discussion

Following recent work by Satten and Datta (2001), this article extends the weightrd-average from of the PLE to the data subject to left-truncation. We have given two demonstrations of the equivalence of the inverse-probability-of-truncation weighted average and product-limit representations of the Lynden-

Bell's estimator. In survival analysis, the weighted-average approach can lead to useful generalizations, primarily to more general censoring or truncated models where censoring or truncation need not be identically distributed.

References

He, S. and G. L. Yang (1998), "Estimation of the Truncation Probability in the Random Truncation Model." *Ann. Statist.*, 26: 1011-1027.

Kalbfleisch, J. and R. Prentice (1980), "The Statistical Analysis of Failure Time Data." New York: Wiley.

Lynden-Bell, D. (1971), "A Method of Allowing for Known Observational Selection in Small Samples Applied to 3CR Quasars." *Mon. Not. R. Astr. Soc.*, 155: 95-118.

Robins, J. M. and A. Rotnitzky (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers." in *AIDS Epidemiology-Methodological Issues*, eds. N. Jewell, K. Dietz, and V. Farewell, Boston: Brikhauser, 297-331.

_____ (1993), "Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers." *in Proceedings of the American Statistical Association-Biopharmaceutical Section*, Alexanndria, VA: ASA., 24-33.

______ and D. Finkelstein (2000), "Correcting for Non-Compliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank tests." *Biometrices*, 56: 779-788.

Statten, G. A. and S. Datta (2001), "The Kaplan-Meier Estimator as an

Inverse-Probability-of Censoring Weighted Average." *Amer. Statist. Ass.*, 55: 207-210.

Vardi, Y. (1982), "Empirical Distribution in Selection Bias Models." *Ann. Statist.*, 10: 616-620.

Wang, M.-C. (1987), "Product-Limit Estimates: a Generalized Maximum Likelihood Study." *Communi. in Statist.*, Part A-Theory and Methods, 6: 3117-3132.

______ (1989), "A Semiparametric Model for Randomly Truncated Data." *J. Amer. Statist. Ass.*, 84: 742-748.

______, Jewell, N. P. and W.-Y. Tsai (1986), "Asymptotic Properties of the Product-Limit Estimate under Random Truncation." *Ann. Statist.*, 14: 1597-1605.

______ (1991), "Nonparametric Estimation from Cross-Sectional Survival Data." *J. Amer. Statist. Ass.*, 86: 130-143.

Woodroofe, M. (1985), "Estimating a Distribution Function with Truncated Data." *Ann. Statist.*, 13: 163-167.

非參數最大估計值
表示成以截取機率為權數之平均值

沈葆聖*

摘要

對於隨機設限資料，Satten 和 Datta（2001）證明 Kaplan-Meier 估計值可以表示成以設限時間機率分配為權數的加權平均值。本文中，我們證明截取資料下，Lynden-Bell（1971）所提出的 product-limit 估計值亦可表示成以截取時間機率分配為權數的加權平均值。

關鍵詞：Product-Limit 估計值、隨機截取。

* 東海大學統計系教授