Reconstructing the Kaplan–Meier Estimator as an M-estimator

Jiaqi Gu[1], Yiwei Fan[1], and Guosheng Yin[1]

[1] *Department of Statistics and Actuarial Science, The University of Hong Kong*

Abstract

The Kaplan–Meier (KM) estimator, which provides a nonparametric estimate of a survival function for time-to-event data, has broad applications in clinical studies, engineering, economics and many other fields. The theoretical properties of the KM estimator including its consistency and asymptotic distribution have been well established. From a new perspective, we reconstruct the KM estimator as an M-estimator by maximizing a quadratic M-function based on concordance, which can be computed using the expectation–maximization (EM) algorithm. It is shown that the convergent point of the EM algorithm coincides with the traditional KM estimator, which offers a new interpretation of the KM estimator as an M-estimator. As a result, the limiting distribution of the KM estimator can be established using M-estimation theory. Application on two real datasets demonstrates that the proposed M-estimator is equivalent to the KM estimator, and the confidence intervals and confidence bands can be derived as well.

Keyword: Censored data; Confidence interval; Loss function; Nonparametric estimator; Survival curve

1 Introduction

In clinical studies, time-to-event data often arise, which record the time of an individual from entry into a study till the occurrence of an event of interest, such as the onset of illness, disease progression, or death (Altman and Bland, 1998). In the past several decades, various methods have been developed for survival analysis, including the Kaplan–Meier (KM) estimator (Kaplan and Meier, 1958), the log-rank test (Mantel, 1966) and the Cox proportional hazards model (Cox, 1972; Breslow and Crowley, 1974). As one of the most highly cited works with nearly 60,000 citations (Noorden et al., 2014), the KM estimator is the standard approach to estimating the survival function for time-to-event data. Under the independence assumption of survival time and censoring time, the KM estimator is a step function with jumps only at the observed event time points, which provides a nonparametric maximum likelihood estimator (NPMLE) for the survival function (Johansen, 1978). By comparing the KM estimators of treatment and control groups, treatment effects can be evaluated using the log-rank test or hazard ratio. In addition, the KM estimator also has broad applications in other fields, such as engineering (Huh et al., 2011), economics (Leclere, 2005) and sociology (Kaminski and Geisler, 2012).

As an NPMLE, the asymptotic theories of the KM estimator have been extensively studied in the literature. The consistency of the KM estimator has been shown by Peterson (1977), and the large-sample variance of the KM estimator at different time points was derived by Greenwood (1927). By estimating the cumulative hazard function with the Nelson–Aalen estimator, the Breslow estimator (Breslow and Crowley, 1974) has been shown to be asymptotically equivalent to the KM estimator. The KM estimator has been shown to converge in law to a Gaussian process whose covariance function can be estimated using Greenwood's formula (Breslow and Crowley, 1974; Hall and Wellner, 1980). In the Bayesian paradigm, the KM estimator has been shown to be the limit of a Bayes estimator of a squared-loss function with a noninformative Dirichlet process prior (Susarla and Ryzin, 1976).

Inspired by the M-estimator of semiparametric regression models for censored data (Zhou, 1992; Jin, 2007), we propose an M-estimator for the survival function. When the L_2 functional norm is used in the M-function, we show that the M-estimator can be obtained recursively via the expectation–maximization (EM) algorithm and the resultant estimator matches with the traditional KM estimator. Therefore, the KM estimator can be reconstructed as a special case of M-estimators, and its large-sample variance and limiting distribution are derived by M-estimation theory, together with the confidence intervals and confidence bands. An application on two real datasets is presented which further demonstrates the equivalence between the existing and M-estimation derivations.

2 Background

2.1 Kaplan–Meier Estimator of Survival Function

Let $T_1, \ldots, T_n$ be independent and identically distributed (i.i.d.) survival times with a cumulative distribution function F_0 and survival function $S_0 = 1 - F_0$. Assume that censoring times $C_1, \ldots, C_n$ are i.i.d. from a distribution G_0. The observed time of subject i is $X_i = \min(T_i, C_i)$ with a censoring indicator $\Delta_i = I(T_i < C_i)$. As usual, independence is assumed between event time T_i and censoring time C_i for $i = 1, \ldots, n$.

Let $\#\{i : \text{condition}\}$ denote the number of observations that meet the condition. When there is no censoring (i.e., $\Delta_i = 1$ for all i), the survival function S_0 is estimated by $\hat{S}(x) = \#\{i : X_i > x\}/n$, which is one minus the empirical distribution function. Kaplan and Meier (1958) extended this estimator to the case with censored observations. Let $X_{(1)} < \cdots < X_{(K)}$ be the K distinct observed event times, and then the KM estimator is $\hat{S}(x) = \prod_{k:X_{(k)}\le x} \left[1 - \#\{i : X_i = X_{(k)}; \Delta_i = 1\}/\left(\#\{i : X_i \ge X_{(k)}\}\right)\right]$, for $0 \le x \le X_{(K)}$. As a step function with jumps only at $X_{(1)} < \cdots < X_{(K)}$, the KM estimator can be interpreted as a product limit estimator of conditional probabilities, which was shown to be weakly consistent and asymptotically normal (Kaplan and Meier, 1958; Breslow and Crowley, 1974).

2.2 M-estimator

Consider i.i.d. vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ from a normal distribution $N(\boldsymbol{\theta}, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix. The sample mean $\bar{\mathbf{y}} = \sum \mathbf{y}_i/n$ is a natural estimator of $\boldsymbol{\theta}$. However, it is well-known that $\bar{\mathbf{y}}$ may perform poorly in estimating $\boldsymbol{\theta}$ if the data are originated from an unknown contaminated distribution. Inspired by the fact that $\bar{\mathbf{y}}$ minimizes $\sum \|\mathbf{y}_i - \boldsymbol{\theta}\|^2$, Huber (1964) proposed the M-estimator $\hat{\boldsymbol{\theta}}$ by minimizing an empirical criterion function, $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} M_n(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^n m_{\boldsymbol{\theta}}(\mathbf{y}_i)/n$, where $\boldsymbol{\Theta}$ is the parameter space and $m_{\boldsymbol{\theta}}(\cdot)$ is the M-function. With different M-functions measuring different types of concordance $\boldsymbol{\theta}$ and observed data, various M-estimators such as maximum likelihood, least squares, and least absolute deviation estimators, can be obtained. The asymptotic properties, including consistency and asymptotic normality, have been studied (Huber, 1967; van der Vaart, 1998), whether $\boldsymbol{\theta}$ is a parameter vector or a functional of the distribution of $\mathbf{y}_1, \ldots, \mathbf{y}_n$. M-estimators have also been extensively investigated in regression models (Koenker and Portnoy, 1990; de Menezes et al., 2021), survival analysis (Zhou, 1992; Jin, 2007) and other areas.

3 M-estimator of Survival Function

3.1 M-estimator of Survival Function without Censoring

In the case with no censoring (i.e., $\Delta_i = 1$ for all i), we consider an M-function $m_S : \mathcal{S} \to \mathbb{R}$ where $\mathcal{S} = \{S(x) : [0, \infty) \to [0, 1]; S(x) \text{ is nonincreasing}\}$. The M-estimator $\hat{S}(x)$ is obtained by maximizing a criterion function, $\hat{S}(x) = \arg\max_{S(x)\in\mathcal{S}} M_n(S) = \arg\max_{S(x)\in\mathcal{S}} n^{-1} \sum_{i=1}^n m_S(X_i)$. As the survival function $S(x)$ represents the probability $\mathbb{P}(X > x)$ for all $x > 0$, we adopt the L_2 functional norm (or a quadratic norm) between the indicator function $I(X > x)$ and the survival function $S(x)$,

$$m_S(X) = \int_0^\infty \big\{ - I(X > x)^2 + 2S(x)I(X > x) - S(x)^2 \big\} d\mu(x), \tag{1}$$

where $\mu(x)$ is a cumulative probability function. Correspondingly, the empirical criterion function is

$$\begin{aligned} M_n(S) &= n^{-1} \sum_{i=1}^n \int_0^\infty \big\{ - I(X_i > x)^2 + 2S(x)I(X_i > x) - S(x)^2 \big\} d\mu(x) \\ &= \int_0^\infty \left[-\#\{i : X_i > x\}/n + 2S(x)\#\{i : X_i > x\}/n - S(x)^2 \right] d\mu(x), \end{aligned} \tag{2}$$

which measures the concordance between the data and the survival function $S(x)$. Hence, the KM estimator

$$\hat{S}(x) = \prod\nolimits_{k:X_{(k)} \le x} \frac{\#\{i : X_i > X_{(k)}; \Delta_i = 1\}}{\#\{i : X_i \ge X_{(k)}\}} = \frac{\#\{i : X_i > x\}}{n},$$

is the maximizer of $M_n(S)$.

3.2 M-estimator of Survival Function with Censoring

When the data are subject to censoring, the M-function (2) is not applicable and thus the empirical criterion function (2) cannot be used. For a censored observation $\{X, \Delta = 0\}$, the only information available is that the event of interest has not occurred prior to X, and the integration of the M-function (1) is undefined in the interval (X, ∞). We use a truncated L_2 functional norm as the M-function for observed survival data $\{X, \Delta\}$,

$$\widetilde{m}_S(X, \Delta) = \begin{cases} m_S(X), & \text{if } \Delta = 1, \\ \displaystyle\int_0^X \big\{ - I(X > x)^2 + 2S(x)I(X > x) - S(x)^2 \big\} d\mu(x), & \text{if } \Delta = 0. \end{cases}$$

To obtain the maximizer,

$$\hat{S}(x) = \arg \max_{S(x) \in \mathcal{S}} \widetilde{M}_n(S) = \arg \max_{S(x) \in \mathcal{S}} n^{-1} \sum_{i=1}^n \widetilde{m}_S(X_i, \Delta_i), \tag{3}$$

we apply the EM algorithm as follows.

- **E-step**: Given the estimator $\hat{S}^{(g)}(x)$ at the gth iteration, compute the expectation of the empirical criterion function $M_n(S)$ by calculating the conditional probability that the event of interest occurs before x for all censored observations,

$$\begin{aligned} E\{M_n(S)|\hat{S}^{(g)}\} = &\frac{1}{n}\sum_{\Delta_i=1}\int_0^\infty \{-I(X_i > x)^2 + 2S(x)I(X_i > x) - S(x)^2\}d\mu(x) \\ &+ \frac{1}{n}\sum_{\Delta_i=0}\int_0^\infty \left\{-\frac{\hat{S}^{(g)}(\max(x, X_i))}{\hat{S}^{(g)}(X_i)} + 2S(x)\frac{\hat{S}^{(g)}(\max(x, X_i))}{\hat{S}^{(g)}(X_i)} - S(x)^2\right\}d\mu(x). \end{aligned} \tag{4}$$

- **M-step**: Compute the estimator $\hat{S}^{(g+1)}(x)$ by maximizing the expectation in (4),

$$\hat{S}^{(g+1)}(x) = \arg\max_{S(x)\in\mathcal{S}} E\{M_n(S)|\hat{S}^{(g)}\}. \tag{5}$$

The validity of this EM algorithm is guaranteed by the next theorem.

Theorem 1. *For all $\hat{S}^{(g)}(x) \in \mathcal{S}$, the quantity $E\{M_n(S)|\hat{S}^{(g)}\} - \widetilde{M}_n(S)$ is maximized when $S = \hat{S}^{(g)}$.*

Based on Theorem 1, we conclude that

$$\begin{aligned} \widetilde{M}_n(\hat{S}^{(g+1)}) &= E\{M_n(\hat{S}^{(g+1)})|\hat{S}^{(g)}\} - \Big[E\{M_n(\hat{S}^{(g+1)})|\hat{S}^{(g)}\} - \widetilde{M}_n(\hat{S}^{(g+1)})\Big] \\ &\geq E\{M_n(\hat{S}^{(g)})|\hat{S}^{(g)}\} - \Big[E\{M_n(\hat{S}^{(g)})|\hat{S}^{(g)}\} - \widetilde{M}_n(\hat{S}^{(g)})\Big] = \widetilde{M}_n(\hat{S}^{(g)}), \end{aligned}$$

and thus the M-estimator would be obtained at the convergent point of the EM algorithm.

Theorem 2. *If $\hat{S}^{(0)}$ is a nonincreasing right-continuous function with $\hat{S}^{(0)}(0) = 1$ and $\hat{S}^{(0)}(x) = \hat{S}^{(0)}(X_{(K)})$ for all $x \geq X_{(K)}$, the sequence of functions $\{\hat{S}^{(g)}, g = 0, 1, \ldots\}$ with the recursive relationship in (5) satisfies*

$$\hat{S}(x) := \lim_{g\to\infty} \hat{S}^{(g)}(x) = \prod_{k:X_{(k)}\leq x}\left(1 - \frac{\#\{i : X_i = X_{(k)}; \Delta_i = 1\}}{\#\{i : X_i \geq X_{(k)}\}}\right). \tag{6}$$

In the proofs of both theorems in the Appendix, the EM algorithm is shown to converge to the KM estimator by induction, implying that the KM estimator is an M-estimator in (3).

4 Asymptotic Properties of KM Estimator

By casting the KM estimator as an M-estimator, we can establish its asymptotic properties in the framework of M-estimation theory as introduced in Section 2.2, including the consistency, asymptotic normality, pointwise confidence intervals, and confidence bands.

4.1 Consistency and Asymptotic Normality

Define

$$\kappa_x(X,\Delta) = \begin{cases} I(X > x), & \text{if } \Delta = 1, \\ S_0(\max(x,X))\{S_0(X)\}^{-1}, & \text{if } \Delta = 0. \end{cases}$$

Consequently, it holds that $E\{\widetilde{m}_S(X,\Delta)|S_0\} = \int_0^\infty \big\{ -\kappa_x(X,\Delta) + 2S(x)\kappa_x(X,\Delta) - S(x)^2\big\} d\mu(x)$ and $\partial E\{\widetilde{m}_S(X,\Delta)|S_0\}/\partial S(x) = 2\kappa_x(X,\Delta) - 2S(x)$, for all $x > 0$.

Given that $X = \min(T,C)$ and $\Delta = I(T < C)$ where $T \sim F_0 = 1 - S_0$ and $C \sim G_0$, it is straightforward to show that $E\left\{\kappa_x(X,\Delta)\right\} = S_0(x)$ and $E\left[\partial E\{\widetilde{m}_S(X,\Delta)|S_0\}/\partial S(x)\right] = 2S_0(x) - 2S(x)$, which equals 0 when $S(x) = S_0(x)$. Thus, $S_0(x)$ is the maximizer of $M(S) = E[E\{\widetilde{m}_S(X,\Delta)|S_0\}]$ in the space of survival functions $\mathcal{S}$. In addition, $E\{\widetilde{M}_n(S)|S_0\} = \sum_{i=1}^n E\{\widetilde{m}_S(X_i,\Delta_i)|S_0\}/n$ converges almost surely to $M(S)$ by the law of large numbers. As a result, $\hat{S}(x) = \arg\max \widetilde{M}_n(S)$ converges to $S_0(x)$ in probability for all $x > 0$, leading to the consistency of $\hat{S}(x)$.

If there exists X_F where $F_0(X_F) < 1$, we have that for $0 < x_1 \leq x_2 < X_F$,

$$E\big\{\kappa_{x_1}(X,\Delta)\kappa_{x_2}(X,\Delta)\big\} = S_0(x_2)\{1 - G_0(x_1)\} + S_0(x_1)S_0(x_2)\int_0^{x_1}\{S_0(u)\}^{-1}dG_0(u)$$

by the conditional expectation formula. As a result, when $S = S_0$, we have

$$\begin{aligned} E\left[\frac{\partial E\{\widetilde{m}_S(X,\Delta)|S_0\}}{\partial S(x_1)}\frac{\partial E\{\widetilde{m}_S(X,\Delta)|S_0\}}{\partial S(x_2)}\right] =& 4S_0(x_1)S_0(x_2)\int_0^{x_1}\frac{1}{S_0^2(1-G_0)}d(1-S_0), \\ E\left[\frac{\partial^2 E\{\widetilde{m}_S(X,\Delta)|S_0\}}{\partial S(x_1)\partial S(x_2)}\right] =& -2I(x_1 = x_2). \end{aligned}$$

Thus, we have the joint asymptotic distribution of $(\hat{S}(x_1), \hat{S}(x_2))$ as

$$\sqrt{n}\begin{pmatrix}\hat{S}(x_1) - S_0(x_1) \\ \hat{S}(x_2) - S_0(x_2)\end{pmatrix} \xrightarrow{\mathcal{D}} N\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} S_0(x_1)^2\int_0^{x_1}\frac{d(1-S_0)}{S_0^2(1-G_0)} & S_0(x_1)S_0(x_2)\int_0^{x_1}\frac{d(1-S_0)}{S_0^2(1-G_0)} \\ S_0(x_1)S_0(x_2)\int_0^{x_1}\frac{d(1-S_0)}{S_0^2(1-G_0)} & S_0(x_2)^2\int_0^{x_2}\frac{d(1-S_0)}{S_0^2(1-G_0)}\end{pmatrix}\right\}. \tag{7}$$

Under the log-transformation, the pointwise asymptotic distribution of $\log \hat{S}(x)$ is

$$\sqrt{n}(\log \hat{S}(x) - \log S_0(x)) \xrightarrow{\mathcal{D}} N\left(0, \int_0^x \{S_0^2(1-G_0)\}^{-1}d(1-S_0)\right),$$

where the variance can be estimated by $\sum_{k:X_{(k)}\leq x}\Delta_{(k)}\big\{\big(\sum_{l=k}^K n_{(l)}\big)\big(\sum_{l=k}^K n_{(l)} - \Delta_{(k)}\big)\big\}^{-1}$, with $n_{(k)} = \#\{i : X_i = X_{(k)}\}$ and $\Delta_{(k)} = \#\{i : X_i = X_{(k)}; \Delta_i = 1\}$.

4.2 Comparison with Greenwood's Formula

Greenwood's formula (Greenwood, 1927) is the standard approach to computing the variance of $\hat{S}(x)$. Taking the values $S(X_{(1)}), \dots, S(X_{(K)})$ into consideration, the likelihood function based on the observed data is $L(S) = \prod_{k=1}^{K} \{S(X_{(k-1)}) - S(X_{(k)})\}^{\Delta_{(k)}} S(X_{(k)})^{n_{(k)} - \Delta_{(k)}}$, where $X_{(0)} = 0$, $X_{(K+1)} = \infty$. Decomposing the survival process as K Bernoulli trials of surviving from $X_{(k-1)}$ to $X_{(k)}$ with success probability $\pi_k = S(X_{(k)})/S(X_{(k-1)})$ $(k = 1, \dots, K)$, we have $S(X_{(k)}) = \prod_{l=1}^{k} \pi_l$ and the likelihood function $L(S) = \prod_{k=1}^{K} (1 - \pi_k)^{\Delta_{(k)}} \pi_k^{n_{(k)} - \Delta_{(k)} + \sum_{l=k+1}^{K} n_{(l)}}$. As a result, the maximum likelihood estimator of π_k is $\hat{\pi}_k = 1 - \Delta_{(k)} / (\sum_{l=k}^{K} n_{(l)})$ and the KM estimator is $\hat{S}(x) = \prod_{k: X_{(k)} \leq x} \hat{\pi}_k$. Using the binomial formula, the variance of $\hat{\pi}_k$ is $\operatorname{var}(\hat{\pi}_k) = \pi_k (1 - \pi_k) / (\sum_{l=k}^{K} n_{(l)})$ and the covariance between $\hat{\pi}_k$ and $\hat{\pi}_l$ is $\operatorname{cov}(\hat{\pi}_k, \hat{\pi}_l) = 0$ $(k \neq l)$. Applying the delta method, we have $\operatorname{var}(\log \hat{\pi}_k) = (1 - \pi_k) / (\pi_k \sum_{l=k}^{K} n_{(l)})$ and

$$\operatorname{var}(\log \hat{S}(x)) = \sum_{k: X_{(k)} \leq x} \operatorname{var}(\log \hat{\pi}_k) = \sum_{k: X_{(k)} \leq x} (1 - \pi_k) \Big/ \left(\pi_k \sum_{l=k}^{K} n_{(l)} \right),$$

which is estimated by $\sum_{k: X_{(k)} \leq x} (1 - \hat{\pi}_k) / (\hat{\pi}_k \sum_{l=k}^{K} n_{(l)}) = \sum_{k: X_{(k)} \leq x} \Delta_{(k)} / \{ (\sum_{l=k}^{K} n_{(l)}) (\sum_{l=k}^{K} n_{(l)} - \Delta_{(k)}) \}$. The variance of $\hat{S}(x)$ can be calculated by applying the delta method one more time.

The estimated variance of $\hat{S}(x)$ deduced by M-estimation theory is the same as Greenwood's formula. However, there exist differences between our derivation and Greenwood's formula: (1) Under Greenwood's formula, the variance of $\hat{\pi}_k$ is obtained conditional on the quantity $\sum_{l=k}^{K} n_{(l)}$ and thus the variance of $\log \hat{S}(x)$ is obtained conditional on the quantities $\sum_{l=k}^{K} n_{(l)}$ for $k \in \{k : X_{(k)} \leq x\}$. The reason why Greenwood's formula is able to calculate the unconditional variance of $\log \hat{S}(x)$ requires the application of martingale and empirical process theories and thus is non-trivial. In contrast, our derivation of the asymptotic distribution of $\hat{S}(x)$ is a direct application of the M-estimation theory and thus is more straightforward. (2) The derivation of the asymptotic distribution of $\hat{S}(x)$ is based on the delta method for Greenwood's formula, while our derivation only utilizes the law of large numbers and basic M-estimation theory.

4.3 Confidence Band

Because the M-estimator (3) and the asymptotic distribution (7) obtained by the EM algorithm and the M-estimation theory coincide with the KM estimator and Greenwood's formula, the asymptotic distribution of the process $\hat{Z}(x) = \sqrt{n}\{\hat{S}(x) - S_0(x)\}$ $(0 < x < X_F)$ with respect to the M-estimator $\hat{S}(x) = \arg\max_{S(x) \in \mathcal{S}} \widetilde{M}_n(S)$ converges weakly to a zero-mean Gaussian process $Z(x)$ with covariance function

$$\operatorname{cov}\{Z(x_1), Z(x_2)\} = \int_0^{x_1} \frac{S_0(x_1)S_0(x_2)}{S_0(u)^2\{1-G_0(u)\}} d\{1-S_0(u)\}, \quad 0 < x_1 \le x_2 < X_F,$$

which is the same as the results in Breslow and Crowley (1974) and Hall and Wellner (1980). Let $A(x) = \int_0^x S_0(u)^{-2}\{1-G_0(u)\}^{-1} d\{1-S_0(u)\}$. Under the log-transformation, the process $\hat{Z}^*(x) = \sqrt{n}\{\log \hat{S}(x) - \log S_0(x)\}$ converges weakly to a zero-mean Gaussian process $Z^*(x)$ with covariance function

$$\operatorname{cov}\{Z^*(x_1), Z^*(x_2)\} = \int_0^{x_1} \frac{d\{1-S_0(u)\}}{S_0(u)^2\{1-G_0(u)\}} = A(x_1), \quad 0 < x_1 \le x_2 < X_F.$$

That is, $\hat{Z}^*(x)$ converges weakly to $W(A(x))$ where $W(\cdot)$ is the Wiener process. Because $\lim_{x\to\infty} A(x) = \infty$ and the confidence band of the process $\{W(x) : 0 < x < \infty\}$ is difficult to obtain, it is more convenient to utilize the standard Brownian bridge to derive the confidence band. To connect $Z^*(x)$ with the Brownian bridge, we define $H(x) = A(x)/\{1 + A(x)\}$ with $H(0) = 0$ and $\lim_{x\to\infty} H(x) = 1$, and $H(x) = 1$ for $x \ge X_F$. As a result, we have

$$\operatorname{cov}\{Z^*(x_1), Z^*(x_2)\} = \frac{H(x_1)}{1-H(x_1)} = \frac{H(x_1)\{1-H(x_2)\}}{\{1-H(x_1)\}\{1-H(x_2)\}},$$

which is the covariance function of the zero-mean Gaussian process $B^0(H(x))/\{1-H(x)\}$ $(0 < x < X_F)$ and $B^0(\cdot)$ is the standard Brownian bridge process on $[0, 1]$. With the constant

$$c_\alpha(a, b) = \inf\left\{c : \mathbb{P}\left(\sup_{[a,b]} |B^0(\cdot)| \le c\right) \ge 1-\alpha\right\}, \quad 0 < \alpha < 1,\ 0 < a < b < 1,$$

the asymptotic $100(1-\alpha)\%$ confidence band of the survival function in the interval $[x_1, x_2] \subset [0, X_F]$ is

$$\left[\hat{S}(x)\exp\left(-\frac{c_\alpha(\hat{H}(x_1), \hat{H}(x_2))}{\sqrt{n}(1-\hat{H}(x))}\right), \hat{S}(x)\exp\left(\frac{c_\alpha(\hat{H}(x_1), \hat{H}(x_2))}{\sqrt{n}(1-\hat{H}(x))}\right)\right], \quad x_1 < x < x_2,$$

where $\hat{H}(x) = 1 - \left[1 + \sum_{k:X_{(k)}\le x} n\Delta_{(k)}\{(\sum_{l=k}^K n_{(l)})(\sum_{l=k+1}^K n_{(l)})\}^{-1}\right]^{-1}$.

5 Real Data Application

To illustrate the equivalence of the KM estimator and our M-estimator under the L_2 functional norm, we present two real data examples. The first dataset is from the diabetic retinopathy study (DRS), where 197 high-risk patients were recruited to investigate the effect of laser treatment in preventing visual loss caused by diabetic retinopathy. on diabetic retinopathy (Huster et al., 1989). For each patient, one eye was randomly chosen to receive the laser treatment and the other received no treatment, leading to $n = 394$ observations in total. The second dataset is from a lung cancer study involving the North Central Cancer Treatment Group of 228 patients with advance lung cancer (Loprinzi et al., 1994). The event of interest was death, and the study goal was to investigate whether patients' self-assessment could provide additional prognostic information to complement physicians' assessment.

We apply the M-estimation using the EM algorithm to the time-to-event data with laser treatment and control groups combined in the DRS and the lung cancer study. Figure 1 exhibits the convergence path $\{\hat{S}^{(g)}(x)\}$ and the KM estimator of the survival function. As the number of iterations g grows, the difference between $\hat{S}^{(g)}(x)$ and the KM estimator diminishes. The EM algorithm converged to the KM estimator in 50 iterations for the DRS example and 36 iterations for the lung cancer example. As shown in Figure 2, the survival function estimators, the 95% confidence intervals and 95% confidence bands computed by the M-estimation and the existing NPMLE approach for both datasets are exactly the same. The proposed M-estimator sheds new light on the KM estimator as well as providing new theoretical development in the M-estimation framework.

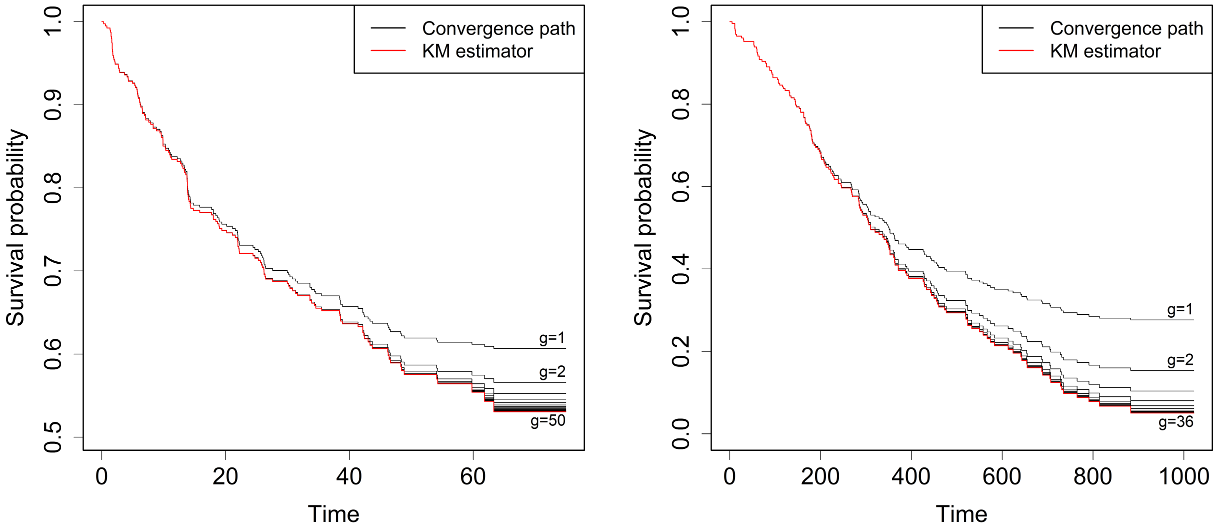

Figure 1: The convergence path (estimated survival curves) $\{\hat{S}^{(g)}(x), g = 1, 2, \ldots\}$ of the EM algorithm and the KM estimator of the survival function for the diabetic retinopathy study with EM convergence at 50 iterations (left) and the lung cancer dataset with EM convergence at 36 iterations (right).

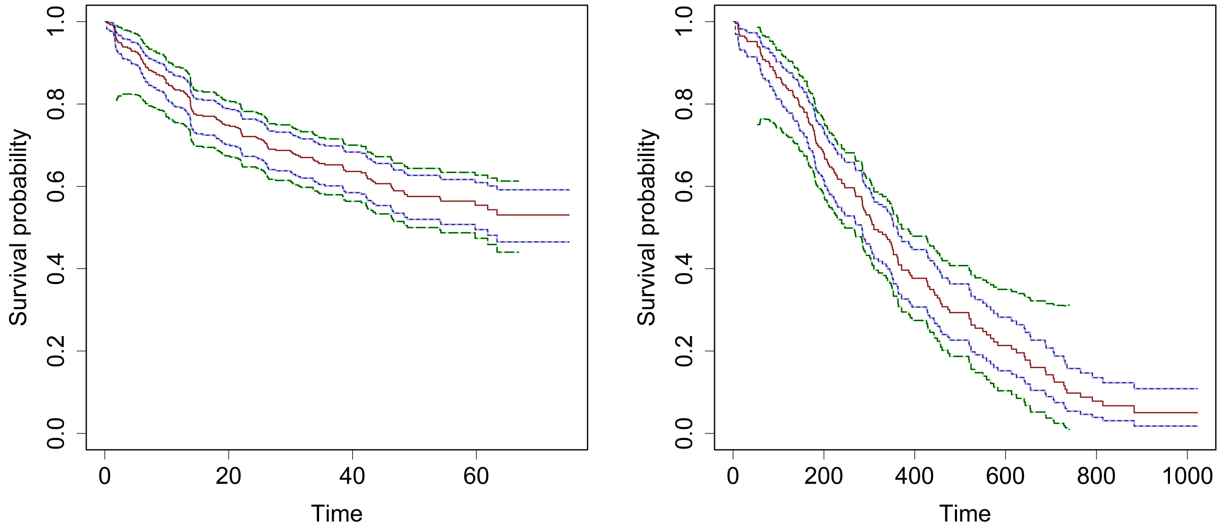

Figure 2: The KM curves (red), 95% confidence intervals (blue) and 95% confidence bands (green) for the diabetic retinopathy (left) and lung cancer datasets (right). Dashed lines for confidence intervals and confidence bands are based on M-estimation, and solid lines are based on Greenwood's formula.

6 Discussion

By casting the survival function estimation in the M-estimation framework, we reconstruct the celebrated KM estimator as an M-estimator, which can be obtained recursively via the EM algorithm. When the L_2 functional norm is used as the M-function, we prove that the EM algorithm converges to the KM estimator, implying that the KM estimator can be viewed as an M-estimator. The theoretical properties of the KM estimator, including the large-sample variance and the limiting distribution, are re-established using M-estimation theory. Simulation studies, which are omitted due to complete overlapping of the estimated curves, demonstrate the exact equivalence relationship between the KM estimator and our M-estimator, while the former has an explicit product limit estimator form and the latter is obtained using the EM algorithm. Real data applications further corroborate that the reconstructed M-estimator under the quadratic M-function is equivalent to the KM estimator and both the confidence intervals and confidence bands coincide with those obtained using Greenwood's formula.

As the KM estimator is an M-estimator under the L_2 functional norm, it is possible to develop other M-estimators of the survival function under different M-functions. One possible M-estimator is the maximum Lq-likelihood estimator (Ferrari and Yang, 2010) with the Lq-likelihood as the M-function. Because the KM estimator is a special case of the maximum Lq-likelihood estimator for $q = 1$, it is possible to use the Lq-likelihood to reduce bias in estimating the survival function for small samples. For example, under the Lq-likelihood in Ferrari and Yang (2010), if $q > 1$ ($q < 1$), more (less) weights would be assigned to later events in the M-function, leading to a smaller mean squared error in estimating $S_0(x)$ for larger (smaller) x. Moreover, the M-estimation framework sheds new light on other problems in survival analysis, such as comparison of multiple survival curves and the Cox regression analysis with covariates.

Acknowledgement

We thank the Editor, Associate Editor, and two referees for their careful reviews of our manuscript and many insightful comments that have significantly improved the paper.

Appendix

Proof of Theorem 1

By the definition of $E\{M_n(S)|\hat{S}^{(g)}\}$, we can deduce that

$$n\big[E\{M_n(S)|\hat{S}^{(g)}\} - \widetilde{M}_n(S)\big] = \sum_{\Delta_i=0}\int_{X_i}^{\infty}\Big\{-\frac{\hat{S}^{(g)}(x)}{\hat{S}^{(g)}(X_i)} + 2S(x)\frac{\hat{S}^{(g)}(x)}{\hat{S}^{(g)}(X_i)} - S(x)^2\Big\}d\mu(x)$$

$$= \sum_{\Delta_i=0}\int_{X_i}^{\infty} -\Big\{S(x) - \frac{\hat{S}^{(g)}(x)}{\hat{S}^{(g)}(X_i)}\Big\}^2 d\mu(x) + \sum_{\Delta_i=0}\int_{X_i}^{\infty}\frac{\hat{S}^{(g)}(x)^2}{\hat{S}^{(g)}(X_i)^2}d\mu(x) - \sum_{\Delta_i=0}\int_{X_i}^{\infty}\frac{\hat{S}^{(g)}(x)}{\hat{S}^{(g)}(X_i)}d\mu(x).$$

Note that $n\big[E\{M_n(S)|\hat{S}^{(g)}\} - \widetilde{M}_n(S)\big]$ consists of three terms, with the latter two terms irrelevant to $S(x)$, and the first term is maximized if and only if $S(x) = \hat{S}^{(g)}(x)$ by the Cauchy–Schwarz inequality.

Proof of Theorem 2

Recall that $n_{(k)} = \#\{i : X_i = X_{(k)}\}$ and $\Delta_{(k)} = \#\{i : X_i = X_{(k)}; \Delta_i = 1\}$. We have for $g = 1, 2, \ldots,$

$$E\{M_n(S)|\hat{S}^{(g-1)}\} = \frac{1}{n}\sum_{k=1}^{K}\int_0^{\infty}\Delta_{(k)}\big\{-I(X_{(k)} > x)^2 + 2S(x)I(X_{(k)} > x) - S(x)^2\big\}d\mu(x)$$

$$+\frac{1}{n}\sum_{k=1}^{K}\int_0^{\infty}(n_{(k)} - \Delta_{(k)})\left\{-\frac{\hat{S}^{(g-1)}(\max(x, X_{(k)}))}{\hat{S}^{(g-1)}(X_{(k)})} + 2S(x)\frac{\hat{S}^{(g-1)}(\max(x, X_{(k)}))}{\hat{S}^{(g-1)}(X_{(k)})} - S(x)^2\right\}d\mu(x).$$

Hence,

$$\hat{S}^{(g)}(x) = \arg\max_{S(x)\in\mathcal{S}} E\{M_n(S)|\hat{S}^{(g-1)}\} = \frac{1}{n}\sum_{k=1}^{K}\left\{\Delta_{(k)}I(X_{(k)} > x) + (n_{(k)} - \Delta_{(k)})\frac{\hat{S}^{(g-1)}(\max(x, X_{(k)}))}{\hat{S}^{(g-1)}(X_{(k)})}\right\}.$$

If $\hat{S}^{(g-1)}$ is a nonincreasing right-continuous function with $\hat{S}^{(g-1)}(0) = 1$, it is clear that $I(X_{(k)} > x)$ and $\hat{S}^{(g-1)}(\max(x, X_{(k)}))\{\hat{S}^{(g-1)}(X_{(k)})\}^{-1}$ are both nonincreasing right-continuous and thus $\hat{S}^{(g)}$ is a nonincreasing right-continuous function with $\hat{S}^{(g)}(0) = 1$. Given $\hat{S}^{(0)}$ is a nonincreasing right-continuous function with $\hat{S}^{(0)}(0) = 1$, by induction, $\hat{S}(x)$ is a nonincreasing right-continuous function with $\hat{S}(0) = 1$.

- For all $x \in [0, X_{(1)})$, it is obvious that $\hat{S}^{(g)}(x) = 1$ for $g = 0, 1, \ldots$ and equation (6) holds.
- We then prove that equation (6) holds at $x = X_{(1)}, \ldots, X_{(K)}$.
  - It is clear that for $g = 0, 1, \ldots,$

    $$\hat{S}^{(g)}(X_{(1)}) = \frac{1}{n}\left\{\sum_{k=1}^{K}(n_{(k)} - \Delta_{(k)}) + \sum_{k=2}^{K}\Delta_{(k)}\right\} = 1 - \frac{\Delta_{(1)}}{\sum_{k=1}^{K} n_{(k)}} = 1 - \frac{\#\{i : X_i = X_{(1)}; \Delta_i = 1\}}{\#\{i : X_i \geq X_{(1)}\}}$$

    and (6) holds at $x = X_{(1)}$.
  - Suppose that (6) holds at $x = X_{(l-1)}$ $(l = 2, \ldots, K)$. If $\hat{S}(X_{(l-1)}) = 0$, then $\hat{S}(X_{(l)}) = 0$ and (6) holds at $x = X_{(l)}$. If $\hat{S}(X_{(l-1)}) \neq 0$, by the convergence condition of the EM algorithm,

    $$\hat{S}(x) = \arg\max_{S(x)\in\mathcal{S}} E\{M_n(S)|\hat{S}\} = \frac{1}{n}\sum_{k=1}^{K}\left\{\Delta_{(k)}I(X_{(k)} > x) + (n_{(k)} - \Delta_{(k)})\frac{\hat{S}(\max(x, X_{(k)}))}{\hat{S}(X_{(k)})}\right\}, \quad (8)$$

    we have

    $$\frac{\hat{S}(X_{(l)})}{\hat{S}(X_{(l-1)})} = \frac{n_{(l)} - \Delta_{(l)} + \sum_{k=l+1}^{K} n_{(k)} + \hat{S}(X_{(l)})\sum_{k=1}^{l-1}(n_{(k)} - \Delta_{(k)})/\hat{S}(X_{(k)})}{\sum_{k=l}^{K} n_{(k)} + \hat{S}(X_{(l-1)})\sum_{k=1}^{l-1}(n_{(k)} - \Delta_{(k)})/\hat{S}(X_{(k)})}.$$

    It is clear that

    $$\frac{\hat{S}(X_{(l)})}{\hat{S}(X_{(l-1)})} = \frac{n_{(l)} - \Delta_{(l)} + \sum_{k=l+1}^{K} n_{(k)}}{\sum_{k=l}^{K} n_{(k)}} = 1 - \frac{\Delta_{(l)}}{\sum_{k=l}^{K} n_{(k)}} = 1 - \frac{\#\{i : X_i = X_{(l)}; \Delta_i = 1\}}{\#\{i : X_i \geq X_{(l)}\}}$$

    and equation (6) holds at $x = X_{(l)}$.
- Finally, we prove that equation (6) holds for $x \in (X_{(l-1)}, X_{(l)})$ $(l = 2, \ldots, K)$ and for $x \in (X_{(K)}, \infty)$.
  - For $x \in (X_{(l-1)}, X_{(l)})$ $(l = 2, \ldots, K)$, if $\hat{S}(X_{(l-1)}) = 0$, then $\hat{S}(x) = 0$ and (6) holds; otherwise, by the convergence condition (8) of the EM algorithm, we have

    $$\frac{\hat{S}(x)}{\hat{S}(X_{(l-1)})} = \frac{n_{(l)} - \Delta_{(l)} + \sum_{k=l+1}^{K} n_{(k)} + \hat{S}(x)\sum_{k=1}^{l-1}(n_{(k)} - \Delta_{(k)})/\hat{S}(X_{(k)})}{\sum_{k=l}^{K} n_{(k)} + \hat{S}(X_{(l-1)})\sum_{k=1}^{l-1}(n_{(k)} - \Delta_{(k)})/\hat{S}(X_{(k)})},$$

    implying that $\hat{S}(x)/\hat{S}(X_{(l-1)}) = 1$. Thus, equation (6) holds.

– For $x \in (X_{(K)}, \infty)$, if $\hat{S}(X_{(K)}) = 0$, then $\hat{S}(x) = 0$ and (6) holds; otherwise, for $g = 1, 2, \ldots$,

$$\frac{\hat{S}^{(g)}(x)}{\hat{S}^{(g)}(X_{(K)})} = \frac{\sum_{k=1}^{K}(n_{(k)} - \Delta_{(k)})\hat{S}^{(g-1)}(x)}{\sum_{k=1}^{K}(n_{(k)} - \Delta_{(k)})\hat{S}^{(g-1)}(X_{(K)})} = \frac{\hat{S}^{(g-1)}(x)}{\hat{S}^{(g-1)}(X_{(K)})} = \cdots = \frac{\hat{S}^{(0)}(x)}{\hat{S}^{(0)}(X_{(K)})} = 1.$$

Therefore, $\hat{S}(x)/\hat{S}(X_{(K)}) = \lim_{g\to\infty} \hat{S}^{(g)}(x)/\hat{S}^{(g)}(X_{(K)}) = 1$, and equation (6) holds.

References

Altman, D. G. and Bland, J. M. (1998). Time to event (survival) data. *British Medical Journal*, 317(7156):468–469.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

de Menezes, D., Prata, D., Secchi, A., and Pinto, J. (2021). A review on robust M-estimators for regression analysis. *Computers & Chemical Engineering*, 147:107254.

Ferrari, D. and Yang, Y. (2010). Maximum Lq-likelihood estimation. *The Annals of Statistics*, 38(2):753–783.

Greenwood, M. (1927). A report on the natural duration of cancer. *The Journal of the American Medical Association*, 88(7):507.

Hall, W. J. and Wellner, J. A. (1980). Confidence bands for a survival curve from censored data. *Biometrika*, 67(1):133–143.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233.

Huh, W. T., Levi, R., Rusmevichientong, P., and Orlin, J. B. (2011). Adaptive data-driven inventory control with censored demand based on Kaplan-Meier estimator. *Operations Research*, 59(4):929–941.

Huster, W. J., Brookmeyer, R., and Self, S. G. (1989). Modelling paired survival data with covariates. *Biometrics*, 45(1):145.

Jin, Z. (2007). M-estimation in regression models for censored data. *Journal of Statistical Planning and Inference*, 137(12):3894–3903.

Johansen, S. (1978). The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics*, 5(4):195–199.

Kaminski, D. and Geisler, C. (2012). Survival analysis of faculty retention in science and engineering by gender. *Science*, 335(6070):864–866.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Koenker, R. and Portnoy, S. (1990). M-estimation of multivariate regressions. *Journal of the American Statistical Association*, 85(412):1060–1068.

Leclere, M. J. (2005). PREFACE modeling time to event: Applications of survival analysis in accounting, economics and finance. *Review of Accounting and Finance*, 4(4):5–12.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., and Klatt, N. E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*, 12(3):601–607.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170.

Noorden, R. V., Maher, B., and Nuzzo, R. (2014). The top 100 papers. *Nature*, 514(7524):550–553.

Peterson, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association*, 72(360a):854–858.

Susarla, V. and Ryzin, J. V. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71(356):897–902.

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

Zhou, M. (1992). M-estimation in censored linear models. *Biometrika*, 79(4):837–841.