Semiparametric Observed Information for Kaplan-Meier Integrals and Nelson-Aalen Integrals

Mai Zhou
Department of Statistics, University of Kentucky,
Lexington, KY 40536 USA
Email: maizhou@gmail.com

Running Title: Semiparametric Observed Information

Abstract

We calculate the semiparametric *observed* Fisher information for a Kaplan-Meier integral based on n iid right censored observations. We show that the inverse of the observed information is equal to the Aalen-Johansen variance estimator of the Kaplan-Meier. Observed as well as expected semiparametric information for a Nelson-Aalen integral are also derived. The Cramér-Rao lower bound for the Nelson-Aalen integral (using expected Fisher information) is only achieved asymptotically. The least favorable families of distributions for these semiparametric estimation/testing problems are explicitly identified.

MSC 2010 Subject Classification: Primary 62N02; secondary 62G05, 62B10.

Key Words and Phrases: Observed Fisher information, Nelson-Aalen estimator, Cramér-Rao lower bound, Least favorable family of distributions.

1 Introduction

The Kaplan-Meier (1958) paper [9] for estimating the survival curve is the most cited statistical paper in history. Even among all the scientific papers, it is in the top 20.

The Kaplan-Meier estimator are commonly taught in (bio)statistics courses and related computational software are readily available. Among the important properties of the estimator, it is known that the Kaplan-Meier estimator is consistent and asymptotically efficient. However, the efficient result for the Kaplan-Meier needs some math backgrounds usually not available in master level courses. For example, the càdlàg functions and $D[0, \tau]$ space. Also, this is an asymptotic result so that the limit of stochastic processes and Brownian motion or Gaussian Process are essential. See the convolution theorems in Bickel, Klaassen, Ritov and Wellner (1993), Wellner (1982).

A more interesting situation is to estimate a finite dimensional parameter of interest within a nonparametric model. Usually, those models are called *semiparametric models.* For many examples of semiparametric models, see Chapter 3 and 4 of Bickel, Klaassen, Ritov and Wellner (1993); Chapter 25 of van der Vaart (1998); Chapter 4 of Kosorok (2008) and Chapter 4 and 5 of Tsiatis (2006). The information lower bound for estimating a finite dimensional parameter, while having an infinite dimensional nuisance parameter, is a major topic discussed in the above books.

While the above books all discussed the *expected* information for large n, we shall compute the *observed* information for fixed n here with the Kaplan-Meier integrals, and observed as well as expected information for the Nelson-Aalen integrals.

Observed Fisher information calculations for classic parametric models are common [6]. For semiparametric models, Murphy and van der Vaart (1999) [23], working on the profile likelihood, show that discrete derivatives can be used to calculate a consistent observed information.

We follow the scheme of Stein (1956) [14] in which the semiparametric information is defined by the infimum of all informations for the parametric sub-models contained in the nonparametric model. Our approach presented in this paper works with finite dimensional estimators in the form of the Kaplan-Meier integrals and we never need to work with limit for $n \to \infty$. We show that the observed Fisher information for the Kaplan-Meier integrals can be calculated for fixed sample size n. Moreover, we find that the Aalen-Johansen variance estimator of the Kaplan-Meier equals to the inverse of the observed Fisher information. This is a 'sample version' of the Cramér-Rao lower bound been reached. The actual Cramér-Rao lower bound (using expected information) for finite n is unknown to this author. However, asymptotically, the lower bound is reached.

Furthermore, the least favorable family of distributions for the semiparametric estimation problem are explicitly identified.

Parallel results on observed information for the Nelson-Aalen integrals are also obtained. In addition, with the help of counting process martingale tools, we also calculated the semiparametric expected Fisher information for the the Nelson-Aalen integrals. It turns out that the Nelson-Aalen integrals only reach the Cramér-Rao lower bound in the asymptotic sense.

In parametric likelihood analysis, the observed Fisher information may give rise to a better normal approximation for the distribution of the Maximum Likelihood Estimator (than using the expected information), see Efron and Hinkley (1978). Similar phenomena also occur in other estimation problems, see Lindsay and Li (1997), Walker (1987), Tierney and Kadane (1986), and Savalei (2010) among others. Therefore calculating observed information is useful even if the expected information is available. A key difference is that the expected information is non-random, while the observed information is data dependent and thus random.

We also identified in this paper the least favorable parametric sub-model for estimating the Kaplan-Meier integrals (also for the Nelson-Aalen integrals). Those parametric family of distributions are very useful in a number of places: see for example DiCiccio and Romano (1990) for calculating nonparametric re-sampling confidence limits. The least favorable family is also intimately connected to the empirical likelihood analysis (Owen, 2001, Ch. 9); and useful in the semiparametric Bayesian analysis. Roughly speaking, the least favorable family reduces the semiparametric estimation/testing problem to a parametric problem. We shall not pursue these topics here.

We end this section with some notation and definition. For n iid right censored observations: (T_i, δ_i) $i = 1, 2, \cdots, n$, where $T_i = \min(X_i, C_i)$, $\delta_i = I[X_i \le C_i]$, we assume the lifetimes X_i are iid with CDF $F(t)$ and the censoring variable C_i is independent of X_i. We assume $F(t)$ is continuous but the distribution $G(t)$ of C_i can have jumps.

The nonparametric likelihood function for F based on the n iid right censored observations is (see for example, Kaplan and Meier 1958):

$$L(F) = \prod_{\delta_i=1} \Delta F(t_i) \prod_{\delta_i=0} [1 - F(t_i)] \ . \tag{1}$$

Kaplan and Meier (1958) also showed that among all CDFs, continuous or discrete, the one CDF that we now call the Kaplan-Meier estimator maximizes the above likelihood:

$$1 - \hat{F}_{km}(t) = \prod_{s \le t} \left(1 - \frac{\Delta N(s)}{R(s)} \right) \ , \tag{2}$$

where $N(t) = \sum_{i=1}^n I[T_i \le t,\, \delta_i = 1],\ R(t) = \sum_{i=1}^n I[T_i \ge t]$.

Therefore, the Kaplan-Meier estimator is a nonparametric maximum likelihood estimator (NPMLE). The difference between NPMLE and regular MLE's is that the parameter here is the entire CDF (infinite dimensional), and regular MLEs are for finite dimensional parameters.

2 Information for the Kaplan-Meier Integrals

The Kaplan-Meier estimator is an estimator of the unknown distribution (survival) function $F(t)$ which is infinite dimensional. We instead will look at the Kaplan-Meier integrals of a given functions $g(t)$. We first look at just one single integral case.

2.1 Information Number

One way to extract a one-dimensional feature out of the infinite dimensional CDF $F(t)$, is to take a linear functional

$$\mu = \int g(t) dF(t) \ . \tag{3}$$

For example the mean μ of the CDF is a one-dimensional parameter. Here $g(t)$ is a function we pick to extract the feature we want. If $g(t) = I[t \le 3]$ then the one-dimensional feature is $F(3)$; if $g(t) = t$ then the feature is the mean value of F, etc.

We want to compute the observed information contained in the censored data likelihood $L(F)$, defined in (1), at $F = \hat{F}_{km}$, for estimating μ.

It is well known that the observed information is related to the (negative) second derivative of the log likelihood, i.e. we need to compute the second derivative of $\log L(F)$ at

$F = \hat{F}_{km}$. According to Stein (1956) [14] this computation can be done first for a 1-parametric sub-family of distributions contained in the nonparametric model and then minimize over all such sub-families.

We shall compute the derivative of $\log L(F)$ with respect to μ as a composite function as follows:

$$\log L(F) = \log L(F_{\lambda(\mu)}) \ ,$$

with F_λ and $\lambda(\mu)$ to be defined below.

Since we only need to compute the derivative at $F = \hat{F}_{km}$, we define a parametric sub-family of distributions that passing through $\hat{F}_{km}$:

$$\Delta F_\lambda(t_i) = \Delta \hat{F}_{km}(t_i)[1 - \lambda f(t_i)] \ , \quad i = 1, \cdots, n \ ; \tag{4}$$

where the parameter $\lambda \in (-a, a)$ for some $a > 0$. Clearly, when $\lambda = 0$ the distribution $F_{\lambda=0}$ goes back to $\hat{F}_{km}$. We are going to compute the derivative at $F = \hat{F}_{km}$ i.e. $\lambda = 0$.

Similar parametric sub-family of distributions like (4) were used by Bickel, Klaassen, Ritov and Wellner (1993) Chapter 3, and van der Vaart (1998) page 364, among others. Intuitively, λ is the magnitude of change and f is the direction of change of F_λ away from the Kaplan-Meier $\hat{F}_{km}$ in the relation (4).

Since $\sum_{i=1}^n \Delta F_\lambda(t_i)$ must be one (for all λ) as is true for all CDFs, we must require $f(t_i)$ in (4) to satisfy

$$\sum_{i=1}^n f(t_i) \Delta \hat{F}_{km}(t_i) = 0. \tag{5}$$

We also assume $0 < \sum_{i=1}^n f^2(t_i) \Delta \hat{F}_{km}(t_i) < \infty$.

First, we look at the function $\lambda = \lambda(\mu)$. Since the specific one-dimension feature we are looking at is $\mu = \int g(t) dF(t)$, this leads to

$$\int g(t) dF_\lambda(t) = \int g(t)[1 - \lambda f(t)] d\hat{F}_{km}(t) = \mu \ .$$

Or, re-arrange terms and write it as a sum, we get an equation for λ

$$\lambda \sum_{i=1}^n g(t_i) f(t_i) \Delta \hat{F}_{km}(t_i) = \sum_{i=1}^n g(t_i) \Delta \hat{F}_{km}(t_i) - \mu \ .$$

$$\lambda = \frac{\sum_{i=1}^n g(t_i) \Delta \hat{F}_{km}(t_i) - \mu}{\sum_{i=1}^n g(t_i) f(t_i) \Delta \hat{F}_{km}(t_i)} \ . \tag{6}$$

Therefore,

$$\frac{\partial \lambda}{\partial \mu} = \left[- \sum_{i=1}^n g(t_i) f(t_i) \Delta \hat{F}_{km}(t_i) \right]^{-1} \ , \qquad \frac{\partial^2 \lambda}{(\partial \mu)^2} = 0 \ .$$

Next, we compute the partial derivative of $\log L(F_\lambda)$ with respect to λ. We first substitute F in the likelihood (1) by F_λ and then take the derivatives.

Let us denote

$$u(\lambda) \overset{\text{def}}{=} \log L(F_\lambda)\,.$$

It is easy to check (we do not need it here, but it is reassuring) that $u^{'}(0) = 0$. Without calculation, we can also explain why $u^{'}(0) = 0$: this log likelihood $u(\lambda)$ achieves its maximum value at $\lambda = 0$ (the Kaplan-Meier), therefore the derivative at $\lambda = 0$ must be 0.

Definition: (Advanced times) For any function $g(s)$ and CDF $F(s)$, the advanced time transformation $\bar{g}(s)$ is defined by

$$\bar{g}(s) = \frac{\int_{(s,\infty)} g(x) dF(x)}{1 - F(s)}\ . \tag{7}$$

Actually, we should probably write it as $\bar{g}_F(s)$ instead of $\bar{g}(s)$ since the definition uses F. What we use in this paper, is for $F = \hat{F}_{km}$. The references of the advanced time are Efron and Johnstone (1990) and Akritas (2000).

Long and tedious calculation/simplification show (see Appendix for details)

$$\frac{u^{''}(0)}{n} = -\sum_{i=1}^{n} [f(t_i) - \bar{f}(t_i)]^2 [1 - \hat{G}_{km}(t_i-)] \Delta \hat{F}_{km}(t_i)\ ,$$

where $\hat{G}_{km}$ is the Kaplan-Meier estimator of the censoring distribution, defined similarly by (2) except replacing $N(t)$ by $N_c(t) = \sum_i I[T_i \le t,\, \delta_i = 0]$. Notice we used the *'advanced time'* $\bar{f}(\cdot)$ above.

Putting the two derivatives together, by the chain rule, the second derivative of $\log L(F)$ with respect to μ (at $\lambda = 0$, or equivalently at $\hat{F}_{km}$) is

$$\frac{\partial^2 \log L(F)}{(\partial \mu)^2}\ |_{\lambda=0} = \frac{-n \sum_i [f(t_i) - \bar{f}(t_i)]^2 [1 - \hat{G}_{km}(t_i-)] \Delta \hat{F}_{km}(t_i)}{[\sum_i g(t_i)\, f(t_i) \Delta \hat{F}_{km}(t_i)]^2}\ . \tag{8}$$

For further simplifications we need to use the variance/covariance identity of advanced-times (see Lemma 1 in Appendix). Using Lemma 1, we can write the derivative (8) as

$$\frac{\partial^2 \log L(F)}{(\partial \mu)^2}\ |_{\lambda=0} = \frac{-n \sum [f(t_i) - \bar{f}(t_i)]^2 [1 - \hat{G}_{km}(t_i-)] \Delta \hat{F}_{km}(t_i)}{\left\{ \sum [g(t_i) - \bar{g}(t_i)]\, [f(t_i) - \bar{f}(t_i)] \Delta \hat{F}_{km}(t_i) \right\}^2}\ . \tag{9}$$

Finally by the Cauchy-Schwarz inequality (see Lemma 2 in Appendix) we find the minimum (or infimum) over all f of the second order derivative

$$\inf_f \frac{-\partial^2 \log L(F)}{(\partial \mu)^2}\ |_{\lambda=0} = \frac{n}{\sum_{i=1}^{n} \frac{[g(t_i) - \bar{g}(t_i)]^2}{1 - \hat{G}_{km}(t_i-)} \Delta \hat{F}_{km}(t_i)} = \mathcal{J}(\mu, \hat{F}_{km})\,. \tag{10}$$

This is the "observed Fisher information", denoted by $\mathcal{J}(\mu, \hat{F}_{km})$, for estimating μ, at $\lambda = 0$ or at $F = \hat{F}_{km}$, obtained using Stein's (1956) method. May be this result deserve to be stated separately as a Theorem, indeed this is just a special case for the Theorem 1 later for $r \geq 1$ parameters.

In the above calculation when we use the Cauchy-Schwarz inequality to find the infimum, it also gives an easy way to identify the f that achieves the infimum in (10). This f have the smallest information and thus also gives rise to the 'least favorable' sub-family of distributions via (4). For definition and more discussion and applications of 'least favorable' subfamily of distributions, see Stein (1956), Bickel, Klaassen, Ritov and Wellner (1993), van der Vaart (1998), DiCiccio and Romano (1990), Owen (2001), or Efron and Tibshirani (1993) section 22.7. This least favorable f satisfies

$$f(t_i) - \bar{f}(t_i) \propto \frac{g(t_i) - \bar{g}(t_i)}{1 - \hat{G}_{km}(t_i-)}, \qquad (\text{a.s. } \hat{F}_{km}).$$

Remark: Given $g(\cdot)$, we can use the above to determine $f(t_i)$ recursively starting from the last observation.

Example 1: Notice the Kaplan-Meier estimator $\hat{F}_{km}(\tau_0)$ can be obtained as the integral

$$\hat{F}_{km}(\tau_0) = \int I[x \leq \tau_0] d\hat{F}_{km}(x)\,,$$

similar to (3) with a function $g(x) = I[x \leq \tau_0]$. Here $\mu = F(\tau_0) = \int I[t \leq \tau_0] dF(t)$.

The observed information (for $\mu = \int g dF$) we obtained on the right hand side of (10), is

$$\mathcal{J} = \left(\sum_{i=1}^{n} \frac{[g(t_i) - \bar{g}(t_i)]^2}{1 - \hat{G}_{km}(t_i-)} \frac{\Delta \hat{F}_{km}(t_i)}{n} \right)^{-1}. \tag{11}$$

Plug $g(x) = I[x \leq \tau_0]$ into (11) we see that

$$g(t_i) - \bar{g}(t_i) = \begin{cases} 0, & \text{if } t_i > \tau_0; \\ 1 - \frac{\hat{F}_{km}(\tau_0) - \hat{F}_{km}(t_i)}{1 - \hat{F}_{km}(t_i)}, & \text{if } t_i \leq \tau_0 \end{cases}$$

and we have

$$[g(t_i) - \bar{g}(t_i)]^2 = \begin{cases} 0, & \text{if } t_i > \tau_0; \\ \frac{[1 - \hat{F}_{km}(\tau_0)]^2}{[1 - \hat{F}_{km}(t_i)]^2}, & \text{if } t_i \leq \tau_0\,. \end{cases}$$

The observed information for estimating $\mu = F(\tau_0)$, at $\hat{F}_{km}$ is thus

$$\mathcal{J}(F(\tau_0)) = \left([1 - \hat{F}_{km}(\tau_0)]^2 \sum_{t_i \leq \tau_0} \frac{1}{1 - \hat{G}_{km}(t_i-)} \frac{\Delta \hat{F}_{km}(t_i)}{n[1 - \hat{F}_{km}(t_i)]^2} \right)^{-1}. \tag{12}$$

After some simplification, we finally have

$$\mathcal{J}(F(\tau_0)) = \left([1 - \hat{F}_{km}(\tau_0)]^2 \sum_{t_i \le \tau_0} \frac{\Delta N(t_i)}{(R(t_i) - \Delta N(t_i))^2} \right)^{-1} . \tag{13}$$

On the other hand, the Greenwood formula for estimating the variance of the Kaplan-Meier estimator $\hat{F}_{km}(\tau_0)$ is

$$\text{Greenwood} = [1 - \hat{F}_{km}(\tau_0)]^2 \sum_{t_i \le \tau_0} \frac{\Delta N(t_i)}{R(t_i)(R(t_i) - \Delta N(t_i))} . \tag{14}$$

However, aside from the popular Greenwood, there are several alternative estimators of the variance also proposed and studied in the literature, see for example [10] and [21] and [22]. We would like to highlight one of the modifications of Greenwood: the variance estimator of Aalen-Johansen (1978) [1]:

$$\text{Aalen-Johansen} = [1 - \hat{F}_{km}(\tau_0)]^2 \sum_{t_i \le \tau_0} \frac{\Delta N(t_i)}{(R(t_i) - \Delta N(t_i))^2} . \tag{15}$$

Recently, [22] compared *ten* different estimators of $\text{Var}(\hat{F}_{km}(\tau_0))$ by simulations. They found the Greenwood tend to under estimate the true variance, particularly when τ_0 at the tail of the distribution. Apparently Aalen-Johansen > Greenwood. But the difference is small and as sample size increases, if the underlying survival distribution is continuous, the difference goes to zero.

Compare (13) and (15), we have shown the equality

$$\frac{1}{\mathcal{J}(F(\tau_0))} = \text{Aalen-Johansen} . \tag{16}$$

We wanted to point out that the equality above "**almost**" state that the Kaplan-Meier estimator reached the Cramér-Rao lower bound for variances of unbiased estimators. We term this a '**sample version**' of the lower bound been reached.

The reasons we say "almost" are

> (1) In the 'real' Cramér-Rao lower bound, we should be using the expected information and the true variance of an estimator. Here the observed information is only an estimator of the expected information and the Aalen-Johansen is only an estimated variance of the Kaplan-Meier.
>
> Nevertheless, these are good estimators of their true values, and so at least we have shown the Cramér-Rao lower bound is approximately reached for finite n and asymptotically the Cramér-Rao lower bound is reached.
>
> (2) The Kaplan-Meier estimator may not be an unbiased estimator in general. But we know the bias is extremely small and this is not too much of a concern.

Therefore, this serves as a good evidence that the variance of $\hat{F}_{km}(\tau_0)$ is as small as it can be in the Cramér-Rao sense — i.e. efficient.

Remark Akritas (2000) showed that the Kaplan-Meier integrals are asymptotically normally distributed with a variance σ^2. If the lifetimes X_i have a continuous distribution $F(t)$, we can easily see that (asymptotically reaching the lower bound)

$$\lim_{n\to\infty} \frac{n}{\mathcal{J}(\mu, \hat{F}_{km})} = \sigma^2 \, .$$

2.2 Information Matrix

We may extract a finite number, r, of features from a CDF with r integrals. The r ($r \geq 1$) parameters are $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_r)$ which are defined as

$$(\mu_1, \cdots, \mu_r) = \left(\int g_1(t) dF(t), \cdots, \int g_r(t) dF(t) \right) \, .$$

We denote $\boldsymbol{g}(t) = (g_1(t), \cdots, g_r(t))$, and $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_r)$.

The calculation of observed information matrix is similar to the one parameter case. We only give an outline here.

We define the r-parameter subfamily of distributions as

$$\Delta F_\lambda(t_i) = \Delta \hat{F}_{km}(t_i)[1 - \boldsymbol{\lambda} \cdot \boldsymbol{f}(t_i)] \, , \quad i = 1, \cdots, n \, . \tag{17}$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_r)$ and $\boldsymbol{f}(t) = (f_1(t), \cdots, f_r(t))$ and the product $\boldsymbol{\lambda} \cdot \boldsymbol{f}(t_i)$ is the inner product $\sum_{k=1}^r \lambda_k f_k(t_i)$. We require $\forall k = 1, \cdots, r$; $\sum_i f_k(t_i) \Delta \hat{F}_{km}(t_i) = 0$ due to the condition of summation to one for probability (F_λ as a probability). We also require $0 < \sum f_k^2(t_i) \Delta \hat{F}_{km}(t_i) < \infty$ to rule out $f \equiv 0$.

Let us define three $r \times r$ matrices:

$$\sum = (\sigma_{uv}) = \left(\sum_{i=1}^n [g_u(t_i) - \bar{g}_u(t_i)][g_v(t_i) - \bar{g}_v(t_i)] \frac{\Delta \hat{F}_{km}(t_i)}{1 - \hat{G}_{km}(t_i -)} \right) \, , \tag{18}$$

$$A = (a_{uv}) = \left(\sum_{i=1}^n [g_u(t_i) - \bar{g}_u(t_i)][f_v(t_i) - \bar{f}_v(t_i)] \Delta \hat{F}_{km}(t_i) \right) \, ,$$

$$B = (b_{uv}) = \left(\sum_{i=1}^n [f_u(t_i) - \bar{f}_u(t_i)][f_v(t_i) - \bar{f}_v(t_i)][1 - \hat{G}_{km}(t_i -)] \Delta \hat{F}_{km}(t_i) \right) \, .$$

We first take the partial derivatives of $\log L(F)$ with respect to $\boldsymbol{\lambda}$ and (after simplifications) get

$$\frac{\partial^2 \log L(F)}{\partial \lambda_u \partial \lambda_v}|_{\lambda=0} = -nB \ .$$

The r parameters for the subfamily of distributions defined in (17) are calculated as

$$\sum_{i=1}^{n} g_k(t_i)[1 - \boldsymbol{\lambda} \cdot \boldsymbol{f}(t_i)]\Delta \hat{F}_{km}(t_i) = \mu_k \ , \qquad k = 1, \cdots, r.$$

This can be written as (recall $\sum f_k(t_i)\Delta \hat{F}_{km}(t_i) = 0$, and Lemma 1)

$$A\boldsymbol{\lambda} = \boldsymbol{\tau} - \boldsymbol{\mu}$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_r)$ and $\tau_k = \sum_{i=1}^{n} g_k(t_i)\Delta \hat{F}_{km}(t_i))$. Taking partial derivative in the above equation with respect to $\boldsymbol{\mu}$, we see

$$\frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}} = -A^{-1} \ , \qquad \frac{\partial^2 \lambda}{(\partial \mu)^2} = 0 \ .$$

Direct calculation (see Appendix) show the negative of the second derivative matrix of $\log L(F)$ with respect to $\boldsymbol{\mu}$ is

$$n(A^{-1})^\top B A^{-1} \ .$$

By the matrix version of the Cauchy-Schwarz inequality (see Appendix), we can show

$$\inf_{(f_1, \cdots, f_r)} n(A^{-1})^\top B A^{-1} = n \ \Sigma^{-1} \ . \tag{19}$$

We summarize the above results into the following theorem.

Theorem 1 Suppose we have n iid right censored observations (T_i, δ_i) as specified in section 1. The nonparametric likelihood function for $F(t)$, based on the n iid (T_i, δ_i), is given by (1). Define $r \geq 1$ parameters $(\mu_1, \cdots, \mu_r)$ by $(\int g_1(t)dF(t), \cdots, \int g_r(t)dF(t))$ where $g_k(t)$ are given functions.

Then the observed Fisher information matrix for estimating $(\mu_1, \cdots, \mu_r)$ at $F = \hat{F}_{km}$ is

$$\mathcal{J}(\boldsymbol{\mu}, \hat{F}_{km}) = n\Sigma^{-1} \ ,$$

where Σ is defined in (18), and $\hat{F}_{km}$ is the Kaplan-Meier estimator.

Furthermore, the least favorable subfamily of distributions for estimating $\boldsymbol{\mu}$ is given by (17) with the $\boldsymbol{f}$ specified by

$$f_k(t_i) - \bar{f}_k(t_i) \propto \frac{g_k(t_i) - \bar{g}_k(t_i)}{1 - \hat{G}_{km}(t_i-)} , \quad i = 1, \cdots, n; \ k = 1, \cdots, r; \qquad (\text{a.s. } \hat{F}_{km}) \ ,$$

where $\bar{f}$ and $\bar{g}$ are advanced times with respect to $\hat{F}_{km}$. □

3 Information for the Nelson-Aalen Hazard Integrals

The calculations of the information number for the parameter $\theta =$*Nelson-Aalen integrals* are easier since the jumps of a discrete cumulative hazard do not have to sum to one, and counting process martingale tools are readily available. For the observed information, the calculation is carried out in subsection 3.1. We also compute in subsection 3.2 the expected information for the hazard integrals, defined as (infimum over submodels of) the expectation of the square of the first derivative of the log likelihood.

3.1 Observed information number

Given the same n iid right censored data as described in section 1, the infinite dimensional parameter here is the unknown cumulative hazard function, $\Lambda(t)$, of the lifetimes X_i. A one dimensional feature θ of the $\Lambda(t)$ is defined by

$$\int g(t) d\Lambda(t) = \theta \tag{20}$$

where we assume the function $g(t)$ is such that the integral is finite.

The *Poisson* log likelihood for $\Lambda(t)$, based on the n iid right censored data (T_i, δ_i) and assuming the Λ is discrete, can be written as (see for example Zhou (2016) page 23 equation (2.2) and page 24 equation (2.3)):

$$\log L_1(\Lambda) = \sum_{i=1}^{n} \left(\Delta N(t_i) \log \Delta\Lambda(t_i) - \sum_{j=1}^{n} \Delta\Lambda(t_j) I[t_j \le t_i] \right) , \tag{21}$$

where $N(t)$ and $R(t)$ were defined previously near equation (2). The NPMLE of $\Lambda(t)$ here is the Nelson-Aalen estimator,

$$\widehat{\Lambda}_{na}(t) = \sum_{t_i \le t} \frac{\Delta N(t_i)}{R(t_i)} \ .$$

For a given $h(t)$, [1] we define a parametric subfamily of cumulative hazard functions, $\Lambda_\lambda(t)$, by its jumps

$$\Delta\Lambda_\lambda(t_i) = \Delta\widehat{\Lambda}_{na}(t_i)[1 - \lambda h(t_i)] \ . \tag{22}$$

When $\lambda = 0$, we have $\Lambda_\lambda = \widehat{\Lambda}_{na}$. We want to compute the derivatives at $\lambda = 0$.

For this subfamily of cumulative hazard functions, the parameter θ defined above is then

$$\theta = \sum_{i=1}^{n} g(t_i)\Delta\widehat{\Lambda}_{na}(t_i) - \lambda \sum_{i=1}^{n} g(t_i)h(t_i)\Delta\widehat{\Lambda}_{na}(t_i) \ .$$

[1] In fact this function can be random, but must be predictable.

From the above equation, we have

$$\frac{\partial \lambda}{\partial \theta} = \left\{ - \sum_{i=1}^{n} g(t_i) h(t_i) \Delta \hat{\Lambda}_{na}(t_i) \right\}^{-1} , \qquad \frac{\partial^2 \lambda}{(\partial \theta)^2} = 0 \ .$$

To calculate the derivative of $\log L_1(\Lambda)$ with respect to λ, at $\lambda = 0$, we first write out the log likelihood using the parametric subfamily of hazard functions specified above. Then direct calculation show

$$\frac{\partial^2 \log L_1(\Lambda_\lambda)}{(\partial \lambda)^2} |_{\lambda=0} = - \sum_{i=1}^{n} h^2(t_i) \Delta N(t_i) = - \sum_{i=1}^{n} h^2(t_i) R(t_i) \Delta \widehat{\Lambda}_{na}(t_i) \ .$$

By the chain rule, the second order derivative of $\log L_1(\Lambda)$ with respect to θ, at $\Lambda = \widehat{\Lambda}_{na}$ or at $\lambda = 0$, is

$$\frac{\partial^2 \log L_1(\Lambda)}{(\partial \theta)^2} \ |_{\lambda=0} = \frac{- \sum_{i=1}^{n} h^2(t_i) R(t_i) \Delta \widehat{\Lambda}_{na}(t_i)}{\left[\sum_{i=1}^{n} g(t_i) h(t_i) \Delta \widehat{\Lambda}_{na}(t_i) \right]^2} \ .$$

Finally, using the Cauchy-Schwarz inequality, we see

$$\inf_h \ - \frac{\partial^2 \log L_1(\Lambda)}{(\partial \theta)^2} \ |_{\lambda=0} = \frac{1}{\sum_{i=1}^{n} \frac{g^2(t_i) \Delta \widehat{\Lambda}_{na}(t_i)}{R(t_i)}} \ . \tag{23}$$

We shall call this the (semiparametric) observed information (for parameter θ at $\widehat{\Lambda}_{na}$).

The h that achieve the infimum in (23) is least favorable. The least favorable subfamily of distributions for the estimation problem at hand is given by (22) with an h function satisfy

$$h(t_i) \ \propto \ \frac{g(t_i)}{R(t_i)} , \qquad \text{(a.s. } \widehat{\Lambda}_{na}\text{)}. \tag{24}$$

Theorem 2 Suppose we have n iid right censored observations (T_i, δ_i) as specified in section 1. The nonparametric (Poisson version of) log likelihood based on the right censored observations for the unknown cumulative hazard function $\Lambda(t)$ (of X_i) is given as in (21).

Define a parameter θ by $\theta = \int g(t) d\Lambda(t)$ for a given function $g(t)$. Then the observed Fisher information contained in the log likelihood (21) for estimating θ, at $\Lambda = \widehat{\Lambda}_{na}$ is

$$\mathcal{J}(\theta, \widehat{\Lambda}_{na}) = \left\{ \sum_{i=1}^{n} \frac{g^2(t_i) \Delta \widehat{\Lambda}_{na}(t_i)}{R(t_i)} \right\}^{-1} ,$$

where $\widehat{\Lambda}_{na}(t)$ is the Nelson-Aalen estimator and $R(t)$ is defined in section 1.

The least favorable subfamily of cumulative hazard functions for estimating θ is given by (22) with h satisfy (24). We want to point out that this least favorable f is a predictable function.

A multi-parameter version of this theorem is straightforward. We omit to save space. □

We recall that the NPMLE of the parameter θ based on the Nelson-Aalen estimator is $\hat{\theta} = \int g(t) d\widehat{\Lambda}_{na}(t)$. And its variance can be estimated by

$$\sum_{i=1}^{n} \frac{g^2(t_i)\Delta N(t_i)}{[R(t_i)]^2} \, , \tag{25}$$

see for example Aalen (1978), Klein (1991) [10].

We want to point out that:

$$\left\{ \mathcal{J}(\theta, \widehat{\Lambda}_{na}) \right\}^{-1} = \left\{ \text{Aalen variance estimator (25) for } \hat{\theta} = \int g(t) d\widehat{\Lambda}_{na}(t) \right\} \, .$$

This almost says the estimator $\hat{\theta}$ reached the Cramér-Rao lower bound, except these are the estimators of the true variance and estimator of the expected information. Same comments for the Kaplan-Meier observed information (section 2.1) also apply here.

3.2 Expected information

As a comparison we compute the expected semiparametric information in $\log L(\Lambda)$ for estimating $\theta = \int g(t) d\Lambda(t)$. Assume the true model (or true cumulative hazard function) $\Lambda_0(t)$ is continuous.

The expected information is related to the derivative of $\log L(\Lambda)$ at the true parameter, therefore we define a submodel passing through the true model Λ_0, indexed by a finite parameter $\eta \in \mathbb{R}$, by

$$d\Lambda_\eta(t) = d\Lambda_0(t)[1 - \eta h(t)] \ . \tag{26}$$

Please note, the derivatives here need to be computed at the true Λ_0 which is assumed continuous. While the observed information is based on the derivatives computed at $\widehat{\Lambda}_{na}$ which is discrete. Therefore the log likelihood we use in section 3.1 for observed information is (21), assuming discrete Λ. The log likelihood we use here for expected information is (27) below, which is the Poisson log likelihood without the discrete assumption (see Zhou (2016) page 23, equation (2.1)):

$$\log L_2(\Lambda) = \sum_{i=1}^{n} \Delta N(t_i) \log d\Lambda(t_i) - \sum_{i=1}^{n} \Lambda(t_i) \, . \tag{27}$$

The first derivative of $\log L_2(\Lambda_\eta)$ with respect to η at $\eta = 0$ is

$$\frac{\partial}{\partial \eta} \log L_2(\Lambda_\eta)|_{\eta=0} = -\sum_{i=1}^{n} h(t_i)\Delta N(t_i) + \sum_{i=1}^{n} \int_0^{t_i} h(u) d\Lambda_0(u) \,.$$

Exchange the order of summation and integral in the second term on the right, we have

$$\frac{\partial}{\partial \eta} \log L_2(\Lambda_\eta)|_{\eta=0} = -\int_0^\infty h(t) dN(t) + \int_0^\infty R(u)h(u)d\Lambda_0(u)$$

$$= -\left(\int_0^\infty h(t) d \left[N(t) - \int_0^t R(u)d\Lambda_0(u) \right] \right) = -M(\infty) \;. \qquad \text{say}$$

Notice the expression inside the square bracket of the last integral above is a counting process martingale $M^+(t)$, thus the whole integral is also a martingale evaluated at infinity, provided $h(t)$ is a predictable function.

On the other hand, if we multiply $g(t)$ on both side of the equation (26) and integrate, we have

$$\theta = \theta_0 - \eta \int h(t) d\Lambda_0(t) \,.$$

Consequently, we see that the derivative

$$\frac{\partial \eta}{\partial \theta} = \frac{-1}{\int g(t)h(t)d\Lambda_0(t)} \;.$$

Combine the two derivatives using chain rule, we obtain the (first) derivative of $\log L_2$ with respect to θ.

$$\frac{\partial \log L_2(\Lambda_\eta)|_{\eta=0}}{\partial \eta} \frac{\partial \eta}{\partial \theta} = \frac{M(\infty)}{\int g(t)h(t)d\Lambda_0(t)} \,.$$

The variance of this derivative gives the expected information for the particular submodel of cumulative hazards defined in (26).

The variance or second moment of it can be computed by first compute the predictive variation process of the martingale, then taking an expectation of the predictive variation: i.e. $\text{Var}M(\infty) = \mathbb{E}\langle M(\infty)\rangle$. Therefore the second moment of the numerator above is

$$\mathbb{E} \int_0^\infty h^2(t) R(t) d\Lambda_0(t) = \int_0^\infty h^2(t) n[1 - F_0(t-)][1 - G_0(t-)] d\Lambda_0(t) \;.$$

Thus the expected information number for this sub-model is

$$\frac{\int_0^\infty h^2(t) n[1 - F_0(t-)][1 - G_0(t-)] d\Lambda_0(t)}{\left[\int_0^\infty g(t)h(t)d\Lambda_0(t)\right]^2} \,.$$

The semiparametric expected information number is defined (see for example Stein (1953)) as the infimum over all sub-models of the above information. In view of the above calculations, we have

$$\mathbf{I}_e(\theta, \Lambda_0) = \inf_h \frac{\int_0^\infty h^2(t) n[1-F_0(t-)][1-G_0(t-)] d\Lambda_0(t)}{\left[\int_0^\infty g(t)h(t)d\Lambda_0(t)\right]^2}$$

$$= \left[\int_0^\infty \frac{g^2(t)d\Lambda_0(t)}{n(1-F_0(t-))(1-G_0(t-))}\right]^{-1} .$$

where the last equality is due to the Cauchy-Schwarz inequality.

The h function that achieves the infimum represents the sub-model with smallest information. Thus the least favorable subfamily is given by (26) with

$$h(t) \propto \frac{g(t)}{n(1-F_0(t-))(1-G_0(t-))} \quad , \qquad \text{(a.s. } \Lambda_0\text{)} . \tag{28}$$

Theorem 3 Suppose we have n iid right censored observations (T_i, δ_i) as specified in section 1. Assume the true cumulative hazard function $\Lambda_0(t)$ of lifetimes X_i is continuous. The nonparametric (Poisson version of) log likelihood based on the right censored observations for a cumulative hazard function $\Lambda(t)$ is given in (27).

Define a parameter θ by $\theta = \int g(t)d\Lambda(t)$ for a given function g. Then the semiparametric expected Fisher information contained in the log likelihood (27) for estimating θ, at true model Λ_0 is

$$\mathbf{I}_e(\theta, \Lambda_0) = \left\{ \int \frac{g^2(t)d\Lambda_0(t)}{n(1-F_0(t-))(1-G_0(t-))} \right\}^{-1} ,$$

where $F_0(t)$ is the true lifetime distribution, $G_0(t)$ is the true censoring distribution.

The least favorable subfamily of cumulative hazard functions for estimating θ is given by (26) with an h function satisfy (28).

A multi-parameter version of this theorem is straightforward. □

Is the Cramér-Rao Lower Bound being reached here? On the other hand, the variance of the Nelson-Aalen integral can be calculated as the mean of the martingale predictable variation process

$$\mathrm{Var}(\int g d\widehat{\Lambda}_{na}) = \mathbb{E} \int \frac{g^2(t)d\Lambda_0(t)}{R(t)} .$$

Since

$$\mathbb{E}\frac{1}{R(t)} \geq \frac{1}{\mathbb{E}R(t)} \qquad \text{and} \qquad \mathbb{E}R(t) = n(1-F_0(t-))(1-G_0(t-)) ,$$

we see that the Nelson-Aalen integral has a variance that is larger than the Cramér-Rao information lower bound[2]. But the two are fairly close. Since as $n \to \infty$, $n/R(t)$ has a non-random limit, the lower bound is attained asymptotically. This can also be confirmed by the martingale central limit theorem for $\sqrt{n}\int g(t)d(\widehat{\Lambda}_{na}(t) - \Lambda_0(t))$ and inspect the asymptotic variance.

4 Appendix

Lemma 1 (Variance/Covariance identity using advanced time) We have, for any $g(t)$ and any CDF F

$$\int [g(t) - \mathbb{E}g]^2 dF(t) = \int [g(t) - \bar{g}(t)]^2 dF(t) \ ,$$

where $\mathbb{E}g = \int g(t)dF(t)$, provided these integrals are well defined. Similarly, we have a covariance identity (where $h(t)$ is another function):

$$\int [g(t) - \mathbb{E}g][h(t) - \mathbb{E}h]dF(t) = \int [g(t) - \bar{g}(t)][h(t) - \bar{h}(t)]dF(t) \ .$$

Proof: The variance identity was proved in Efron and Johnstone (1990). The covariance identity can be similarly proved, see their equation (2.5). □

Lemma 2 Integral version of the Cauchy-schwarz inequality:

$$\int_\Omega \xi^2(x)w^2(x)dv \int_\Omega \frac{\eta^2(x)}{w^2(x)}dv \ \geq \ \left[\int_\Omega \xi(x)\eta(x)dv\right]^2$$

here dv is a measure on Ω and $w \neq 0$. The inequality becomes an equality if and only if (aside from a multiplicative constant) $\xi w = \eta/w$ (a.s. dv).

Lemma 3 (Matrix Cauchy-Schwarz Inequality) For matrices A, B and Σ defined in section 2, we have, for any $\boldsymbol{f}(t) = (f_1(t), \cdots, f_r(t))$ as defined before,

$$(A^{-1})^\top BA^{-1} \geq \Sigma^{-1}$$

where $\geq$ means the matrix inequality for positive definite matrices. The equality is achieved when, for $k = 1, \cdots, r$

$$f_k(t) - \bar{f}_k(t) = \frac{g_k(t) - \bar{g}_k(t)}{1 - \hat{G}_{km}(t-)} \ , \qquad (\text{a.s. } \hat{F}_{km}) \ .$$

[2] Since $1/x$ is strictly convex, by Jensen inequality as long as $n/R(t)$ is not a constant, the information lower bound is strictly smaller.

Proof: See Tripathi (1999) for a proof of the matrix Cauchy-Schwarz inequality. □

Next we give some details in calculating the second derivative, $u^{''}(0)$. We restrict the calculation for a single parameter. The r parameter case is similar. Substitute $F = F_\lambda$ into the log empirical likelihood, we have

$$\log L(F_\lambda) = \sum_{\delta_i=1} \log \Delta\hat{F}_{km}(t_i)[1-\lambda f(t_i)] \;+ \sum_{\delta_i=0} \log \left(\sum_{s_j > t_i} \Delta\hat{F}_{km}(s_j)[1-\lambda f(s_j)] \right) \;.$$

The second derivative of $u(\lambda) = \log L(F_\lambda)$ at $\lambda = 0$ can be calculated

$$-\frac{u^{''}(0)}{n} = \sum_{i=1}^{n} f^2(t_i)\frac{\delta_i}{n} + \;\sum_{i=1}^{n} \frac{1-\delta_i}{n} \frac{\left(\sum_{t_j>t_i} f(t_j)\Delta\hat{F}_{km}(t_j)\right)^2}{[1-\hat{F}_{km}(t_i)]^2} \;.$$

Leave the second sum on the right hand side unchanged and apply the self-consistency identity (Lemma 27 of Zhou (2016)) to the first sum on the right hand side, we have

$$= \sum_{i=1}^{n} f^2(t_i)\Delta\hat{F}_{km}(t_i) - \sum_{i=1}^{n} \frac{1-\delta_i}{n} \frac{\sum_{t_j>t_i} f^2(t_j)\Delta\hat{F}_{km}(t_j)}{1-\hat{F}_{km}(t_i)}$$

$$+ \;\sum_{i=1}^{n} \frac{1-\delta_i}{n} \frac{\left(\sum_{t_j>t_i} f(t_j)\Delta\hat{F}_{km}(t_j)\right)^2}{[1-\hat{F}_{km}(t_i)]^2} \;.$$

The two sums with $1-\delta_i$ can be combined into one sum and the terms for each i become the variance of f with respect to the (conditional) distributions: $P_i = \Delta\hat{F}_{km}(t_j)/[1-\hat{F}_{km}(t_i)], \quad t_j > t_i$. Therefore

$$= \sum_{i=1}^{n} f^2(t_i)\Delta\hat{F}_{km}(t_i) - \sum_{i=1}^{n} \frac{1-\delta_i}{n} \frac{\sum_{t_j>t_i} \left[f(t_j) - \mathbb{E}^i f\right]^2 \Delta\hat{F}_{km}(t_j)}{1-\hat{F}_{km}(t_i)} \;,$$

where $\mathbb{E}^i$ denote the expectation with respect to the conditional distribution P_i. Notice that $\mathbb{E}^i f = \bar{f}(t_j)$, where the advanced time is defined using the Kaplan-Meier.

The first sum above can also be written as a variance (with respect to the Kaplan-Meier), since $\sum f(t_i)\Delta\hat{F}_{km}(t_i) = 0$. From here, we use the advanced time identity to re-write the variances and get

$$= \sum_{i=1}^{n} \left[f(t_i) - \bar{f}(t_i)\right]^2 \Delta\hat{F}_{km}(t_i) - \sum_{i=1}^{n} \frac{1-\delta_i}{n} \frac{\sum_{t_j>t_i} \left[f(t_j) - \bar{f}(t_j)\right]^2 \Delta\hat{F}_{km}(t_j)}{1-\hat{F}_{km}(t_i)} \;.$$

We then use the self-consistency identity for the Kaplan-Meier again to reduce the above to

$$= \sum_{i=1}^{n} \left[f(t_i) - \bar{f}(t_i) \right]^2 \frac{\delta_i}{n} \ .$$

The final step is using the identity (see bottom of page 72, Zhou (2016))

$$\Delta \hat{F}_{km}(t_i) = \frac{\delta_i}{n(1 - \hat{G}_{km}(t_i-))}$$

to get

$$-\frac{u''(0)}{n} = \sum_{i=1}^{n} \left[f(t_i) - \bar{f}(t_i) \right]^2 \left[1 - \hat{G}_{km}(t_i-) \right] \Delta \hat{F}_{km}(t_i) \ .$$

□

References

[1] Aalen, O.O. and Johansen, S. (1978) An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5, 141–150.

[2] Akritas, M. (2000). The central limit theorem under censoring. *Bernoulli,* 6: 1109–1120.

[3] Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* The Johns Hopkins University Press.

[4] DiCiccio, T.J. and Romano, J.P. (1990). Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review*, 58, 59–76.

[5] Owen, A. (2001). *Empirical Likelihood* Chapman & Hall/CRC Press.

[6] Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher Information. *Biometrika.* 65 (3): 457–487.

[7] Efron, B. and Johnstone, I. (1990). Fisher's information in terms of the hazard rate. *Ann. Statist.*, 18, 38–62.

[8] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap* Chapman & Hall/CRC Press.

[9] Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of tha American Statistical Association* 53, 457–481.

[10] Klein, J.P. (1991). Small sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators. *Scandinavian Journal of Statistics*, 18, 333–340.

[11] Kosorok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer, New York.

[12] Lindsay, B. and Li, B. (1997). On second-order optimality of the observed Fisher information. *Ann. Statist.* 25 (5) 2172–2199.

[13] Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods*, 15(4):352–367. doi: 10.1037/a0020143.

[14] Stein, C. (1956). Efficient nonparametric testing and estimation, *Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1, 187–195.

[15] Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, Volume 81, 82–86.

[16] Tripathi, G. (1999). A matrix extension of the Cauchy-Schwarz inequality. *Ecomomics Letters*, 63, 1–3.

[17] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data.* Springer, New York.

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

[19] Walker, A.M. (1969). On the asymptotic behavior of posterior distributions. JRSSB, 31, 80–88.

[20] Zhou, M. (2016). *Empirical Likelihood Methods in Survival Analysis.* CRC Press, Taylor & Francis Group, Boca Raton.

[21] Borkowf, C.B. (2005). A simple hybrid variance estimator for the Kaplan-Meier survival function. *Statistics in Medicine* Volume24, Issue6 30 March 2005 Pages 827–851.

[22] Khan HN, Zaman Q, Azmi F, Shahzada G and Jakovljevic M (2022). Methods for Improving the Variance Estimator of the Kaplan-Meier Survival Function, When There Is No, Moderate and Heavy Censoring - Applied in Oncological Datasets. *Front. Public Health*, Vol. 10, 793648. https://doi.org/10.3389/fpubh.2022.793648

[23] Murphy, S. and van der Vaart, AW. (1999). Observed information in semi-parametric models. *Bernoulli*, 5(3), 381–412.

[24] Wellner, J. (1982). Asymptotic optimality of the product limit estimator. *The Annals of Statistics*, 10(2) 595–602.