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

Mai Zhou
University of Kentucky, Lexington, KY, 40536 USA

Abstract

The well-known Kaplan-Meier estimator is a nonparametric maximum likelihood estimator. We calculate the semiparametric *observed* Fisher information for a Kaplan-Meier integral based on n iid right censored observations. The inverse of the information is seen equal to the variance of the Kaplan-Meier integral for large n. Observed as well as expected semiparametric information for a Nelson-Aalen integral are also derived. The least favorable families of sub-models for both integrals are 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.

Corresponding Author: Mai Zhou Email: maizhou@gmail.com

Running Title: Semiparametric Observed Information

1 Introduction

For n iid right censored observations: $(T_i,\, \delta_i)\ \ i = 1, 2, \cdots, n$, 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.

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

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

The (Fisher) information for an infinite dimensional parameter is difficult since the new concept of derivatives of $\log L(F)$ with respect to a CDF needs to be defined and the corresponding theory needs to be developed. Strong math background is required. The standard reference on this topic is Bickel, Klaassen, Ritov & Wellner (1993). 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 & 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 right censored data. We follow the scheme of Stein (1956). It turns out that the observed Fisher information for a Kaplan-Meier integral is much easier to calculate than the expected information. We obtain the exact value of the observed information for finite n. No approximation, no limit for n. Only the classic derivatives are used.

In parametric MLE analysis, the observed Fisher information often gives rise to a better normal approximation for the distribution of the MLE (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 the least favorable parametric sub-model for estimating the Kaplan-Meier integrals (also for the Nelson-Aalen integrals). Those parametric family of distributions is very useful in a number of places: see for example DiCiccio and Romano (1990) for calculating nonparametric resampling confidence limits. The least favorable family is also intimately connected to the empirical likelihood analysis (Owen, 2001, Ch. 9). We shall study these topics in other places. Roughly speaking, the least favorable family reduces the nonparametric problem to a parametric problem.

Similar calculations are also carried out in this paper for the Nelson-Aalen integrals, where both expected and observed Fisher information are worked out for finite sample sizes n. In this case, the two informations are exactly equal if we replace the true CDF with the Kaplan-Meier estimator and the true cumulative hazard with the Nelson-Aalen estimator.

2 Information for Kaplan-Meier Integrals

2.1 Information Number

One way to reduce the (infinite) dimension, or to extract a one-dimensional feature of the infinite dimensional parameter of CDF, is to take a functional

$$\mu = \int g(t) dF(t) \ ,$$

i.e. μ 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 likelihood $L(F)$ at $F = \hat{F}_{km}$, for estimating μ.

By well known theory, 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 $F = \hat{F}_{km}$. We shall compute the derivative of $L(F)$ with respect to μ as a composite function as follows:

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

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

Since we only need to compute the derivative at the $F = \hat{F}_{km}$, we define a parametric subfamily of distributions (passing through $\hat{F}_{km}$):

$$\Delta F_\lambda(t_i) = \Delta \hat{F}_{km}(t_i)[1 - \lambda f(t_i)] \ , \quad i = 1, \cdots, n \ ; \tag{3}$$

where the parameter $\lambda \in (-a,\, a)$ for some $a > 0$. We assume $0 < \sum_{i=1}^{n} f^2(t_i)\Delta\hat{F}_{km}(t_i) < \infty$. We are to compute the derivative at $\lambda = 0$. The equation (3) says that we only look at those F that are dominated by the Kaplan-Meier. Similar parametric sub-family of distributions were used by Bickel, Klaassen, Ritov & 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 from the Kaplan-Meier $\hat{F}_{km}$.

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

$$\sum_{i=1}^{n} f(t_i)\Delta\hat{F}_{km}(t_i) = 0. \tag{4}$$

Next, 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

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

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

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

$$\lambda = \frac{\sum_{i=1}^{n} g(t_i)\Delta\hat{F}_{km}(t_i) - \mu}{\sum_{i=1}^{n} g(t_i)f(t_i)\Delta\hat{F}_{km}(t_i)} \ . \tag{5}$$

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 $u(\lambda) = \log L(F_\lambda)$ with respect to λ. We first substitute F in the likelihood (1) by F_λ then take the derivatives.

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.

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 replace $N(t)$ by $N_c(t) = \sum I[T_i \leq t,\, \delta_i = 0]$.

We call the attention of the reader to the use of *'advanced time'* $\bar{f}$ here: a transformation used by Efron & Johnstone (1990) and Akritas (2000). We also give its definition in the Appendix.

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

$$\frac{\partial^2 \log L(F)}{(\partial\mu)^2} \mid_{\lambda=0} = \frac{-n \sum [f - \bar{f}]^2 [1 - \hat{G}_{km}(t_i-)] \Delta \hat{F}_{km}(t_i)}{[\sum g(t_i)\, f(t_i) \Delta \hat{F}_{km}(t_i)]^2} \ . \tag{6}$$

For further simplifications we need

Lemma 1 *We have*

$$\sum_{i=1}^{n} g(t_i)\, f(t_i) \Delta \hat{F}_{km}(t_i) = \sum_{i=1}^{n} [g(t_i) - \bar{g}(t_i)][f(t_i) - \bar{f}(t_i)] \Delta \hat{F}_{km}(t_i) \ . \tag{7}$$

Proof: In view of (4),

$$\sum g(t_i)\, f(t_i) \Delta \hat{F}_{km}(t_i) = \sum [g(t_i) - \mathbf{E}g]\, [f(t_i) - \mathbf{E}f] \Delta \hat{F}_{km}(t_i)$$

where $\mathbf{E}g$, $\mathbf{E}f$ are mean with respect to the Kaplan-Meier. Now the covariance between g and f can be written as the right hand side of (7) above using advanced times, a fact which can be proved similar to Efron & Johnstone (1990), equation (2.5). $\square$

Using the Lemma 1, we can write the derivative (6) as

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

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

$$\inf_f \frac{-\partial^2 \log L(F)}{(\partial\mu)^2} \mid_{\lambda=0} = \frac{n}{\sum_{i=1}^{n} \frac{[g(t_i) - \bar{g}(t_i)]^2}{1 - \hat{G}_{km}(t_i-)} \Delta \hat{F}_{km}(t_i)} \ . \tag{9}$$

This is the "observed Fisher information" for estimating μ, at $\lambda = 0$ or $F = \hat{F}_{km}$, as described in Stein (1956).

According to Akritas (2000), the asymptotic distribution of the Kaplan-Meier integral, $\int g(t) d[\hat{F}_{km}(t) - F(t)]$, is normal with mean zero and a variance well approximated by the inverse of the observed information above. In fact, if you replace the Kaplan-Meier in (9)

by the true CDF and summation by integral, you get the asymptotic variance given in Akritas (2000). This asymptotic equality shows that the Kaplan-Meier integral achieves the information bound, and therefore is an (asymptotic) efficient estimator of μ.

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 (9). This f also gives rise to the 'least favorable' subfamily of distributions via (3). For definition and more discussion 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

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

2.2 Information Matrix

We may extract a finite number, r, of features from a CDF with r integrals. The r ($r > 1$) parameters are $\mu = (\mu_1, \cdots, \mu_r)$ 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

$$\Delta F_\lambda(t_i) = \Delta \hat{F}_{km}(t_i)[1 - \boldsymbol{\lambda} \cdot \boldsymbol{f}(t_i)] , \quad i = 1, \cdots, n . \tag{10}$$

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 $\sum_{k=1}^r \lambda_k f_k(t_i)$.

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

$$\sum = (\sigma_{uv}) = \left(\sum_{i=1}^n [g_u(t_i) - \bar{g}_u(t_i)][g_v(t_i) - \bar{g}_v(t_i)] \frac{\Delta \hat{F}_{km}(t_i)}{1 - \hat{G}_{km}(t_i-)} \right) , \tag{11}$$

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

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

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 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 $\sum f_k(t_i)\Delta \hat{F}_{km}(t_i) = 0$)

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

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

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

Matrix version of the Cauchy-Schwarz inequality (see Appendix) then show

$$\inf_{(f_1, \cdots, f_r)} n(A^{-1})^\top B A^{-1} = n \ \Sigma^{-1} \ . \tag{12}$$

Similar to Akritas (2000), we can show the asymptotic distribution of $\sqrt{n} \int \boldsymbol{g}(t) d[\hat{F}_{km}(t) - F_0(t)]$ is an r-variate normal with mean zero and a variance matrix V well approximated by Σ in (11) (just replace the Kaplan-Meier by the true CDF in Σ, and replace the sum by integration, you obtain V).

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 $F(t)$, the unknown CDF of X_i, is given by (1). Define 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

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

where Σ is defined in (11), and $\hat{F}_{km}$ is the Kaplan-Meier estimator.

Furthermore, the least favorable subfamily of distributions for estimating $\boldsymbol{\mu}$ is given by (10) with the $\boldsymbol{f}$ specified by

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

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

Remark The semiparametric expected information is more difficult, since the (exact) expectation/variance of the Kaplan-Meier estimator etc. are not available. We will have to resort to approximations. On the other hand, when data are non-censored, the semiparametric expected information can be computed.

3 Information for the Nelson-Aalen Hazard Integrals

The calculations for the information of the *Nelson-Aalen hazard integrals* are in fact easier since the jumps of a discrete cumulative hazard do not have to sum to one, and counting process martingale results 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. This is the definition of semiparametric information bound used by most books on semiparametric estimation. We call it the expected information here.

3.1 Observed information

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

$$\int g(t) d\Lambda(t) = \theta \tag{13}$$

where we assume the function g is such that the integral is finite.

The *Poisson* log likelihood for $\Lambda(t)$, based on the n iid right censored data is (see for example Zhou (2016)):

$$\log L_1(\Lambda) = \sum_{i=1}^{n} \left(\Delta N(t_i) \log \Delta\Lambda(t_i) - \sum_{j=1}^{n} \Delta\Lambda(t_j) I[t_j \le t_i] \right) . \tag{14}$$

The NPMLE of $\Lambda(t)$ here is the Nelson-Aalen estimator,

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

We define the parametric subfamily of cumulative hazard functions by

$$\Delta\Lambda_\lambda(t_i) = \Delta\widehat{\Lambda}_{na}(t_i)[1 - \lambda h(t_i)] \ . \tag{15}$$

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

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

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

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

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

$$\inf_h \ -\frac{\partial^2 \log L_1(\Lambda)}{(\partial\theta)^2} \ |_{\lambda=0} = \frac{1}{\sum_{i=1}^{n} \frac{g^2(t_i)\Delta\widehat{\Lambda}_{na}(t_i)}{R(t_i)}} \ . \tag{16}$$

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

It is seen that the estimator of parameter θ based on the Nelson-Aalen, $\hat{\theta} = \int g(t)d\widehat{\Lambda}_{na}(t)$, has a variance well approximated by the inverse of the observed information derived above (see for example Klein (1991)).

The least favorable subfamily of distributions for the estimation problem at hand is given by (15) with an h satisfy

$$h(t_i) \propto \frac{g(t_i)}{R(t_i)}, \qquad (\text{a.s. } \widehat{\Lambda}_{na}). \tag{17}$$

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 lifetimes X_i) is given as in (14).

Define a parameter θ by $\theta = \int g(t)d\Lambda(t)$ for a given function g. Then the observed Fisher information contained in the log likelihood (14) for estimating θ, at $\Lambda = \widehat{\Lambda}_{na}$ is

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

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 (15) with h satisfy (17).

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

3.2 Expected information

As a comparison we will 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, and define a submodel by

$$d\Lambda_\eta(t) = d\Lambda_0(t)[1 - \eta h(t)] \,. \tag{18}$$

We recall the log likelihood for right censored data is given in (14). However, here we assumed the true model is continuous so the log likelihood is

$$\log L_2(\Lambda) = \sum_{i=1}^{n} \Delta N(t_i) \log \Delta\Lambda(t_i) - \sum_{i=1}^{n} \Lambda(t_i) \,. \tag{19}$$

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

$$= -\int_0^\infty h(t)dN(t) + \int_0^\infty R(u)h(u)d\Lambda_0(u)$$

$$= -\left(\int_0^\infty h(t)d\left[N(t) - \int_0^t R(u)d\Lambda_0(u)\right]\right) \ .$$

Notice the last expression is a (counting process) martingale evaluated at infinity. 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. Therefore the second moment of the above is

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

It is easy to see that the derivative of

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

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

The expected information is defined (see for example Bickel, Klaassen, Ritov, & Wellner (1993)) as the infimum over all sub-models of the second moment of the first derivative of log likelihood with respect to θ. In view of the above calculations, we have

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

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

Therefore the expected information is the right hand side of above. And the least favorable subfamily is given by (18) with

$$h(t) \ \propto \ \frac{g(t)}{n(1-F_0(t))(1-G_0(t-))} \quad , \qquad (\text{a.s. } \Lambda_0) \ . \tag{20}$$

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 (19).

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 (19) for estimating θ, at true model Λ_0 is

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

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 (18) with h satisfy (20).

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

Compare the observed and expected information, we see that the observed information is exactly equal to the expected information if we replace the unknown $\Lambda_0(t)$ by $\widehat{\Lambda}_{na}(t)$, $F_0(t)$ by $\hat{F}_{km}(t)$ and $G_0(t)$ by $\hat{G}_{km}(t)$ (recall $[1 - \hat{F}_{km}(t)][1 - \hat{G}_{km}(t)] = R(t)/n$). A similar relation between the two least favorable subfamilies also hold.

References

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

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

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

[4] 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.

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

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

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

[8] 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.

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

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

[11] Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data *Psychological Methods*, 15(4):352-67. doi: 10.1037/a0020143.

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

[13] 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.

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

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

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

[17] Walker, A.M. (1969). On the asymptotic behaviour of posterior distributions. JRSSB, 31, 80-88.

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

4 Appendix.

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

$$\bar{g}(s) = \frac{\int_{(s,\infty)} g(x)dF(x)}{1 - F(s)} \ . \tag{21}$$

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), Akritas (2000).

Lemma 2 (Variance identity with advanced time) We have, for any g and any CDF F

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

where $\mathbf{E}g = \int g(t)dF(t)$.

Proof: See Efron and Johnstone (1990) equation (2.5).

Lemma 3 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 maximum of right hand side is achieved if and only if (aside from a multiplicative constant) $\xi w = \eta/w$ (a.s. dv). It is easy to check, when this happens we have the equality holds.

Lemma 4 (Matrix Cauchy-Schwarz Inequality) For matrices A, B and Σ defined in section 2, we have

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

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

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

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

Next we calculate the second derivative. We give the calculation for a single parameter, r parameter case is similar (and may be found in Zhou (2016)). 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: $P_i = \Delta\hat{F}_{km}(t_j)/[1-\hat{F}_{km}(t_i)], \quad t_j > t_i$:

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

where $\mathbf{E}^i$ denote the expectation with respect to the conditional distribution P_i. Notice that $\mathbf{E}^i f = \bar{f}(t_j)$, 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 self-consistency identity can be found in Zhou (2016) Lemma 27. 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

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

□