Empirical Likelihood Based on Synthetic Right Censored Data

Wei Liang[a], Hongsheng Dai[b]

[a]*School of Mathematical Sciences, Xiamen University, China*
[b]*Department of Mathematical Sciences, University of Essex, UK*

Abstract

In this paper, we develop a Mean Empirical Likelihood (MeanEL) method for right censored data. This MeanEL approach is based on traditional empirical likelihood methods but uses synthetic data to construct an EL ratio statistics, which is shown to have a χ^2 limiting distribution. Different simulation studies show that the MeanEL confidence intervals tend to have more accurate coverage probabilities than other existing Empirical Likelihood methods. Theoretical comparisons of different EL methods are also provided under a general framework.

Keywords:
Censoring; Mean Empirical Likelihood; Survival Analysis

1. Introduction

Liang et al. (2019) proposed a Mean Empirical Likelihood (MeanEL) method based on synthetic pairwise mean data. Empirical simulation results in Liang et al. (2019) showed that this MeanEL method provides better results for heavy-tail or highly-skewed distributions and for exponentially tilted likelihood. However, theoretical comparisons of MeanEL and other existing EL methods, such as Bartlett correction Empirical Likelihood (BEL) in DiCiccio et al. (1991), the adjusted empirical likelihood (AEL) in Chen et al. (2008) and extended empirical likelihood method (EEL) in Taso and Wu (2013), were not established in Liang et al. (2019). This paper will extend such MeanEL approach to right-censored data analysis. Theoretical justification on why using such synthetic data can provide better coverage probability accuracy is also discussed in this paper.

Assume that independent and identically distributed random observations $T_1, T_2, \cdots, T_n$ with an unknown distribution function $F(t)$ are subject to right censoring, so that we only observe

$$Z_i = \min(T_i, C_i), \quad \eta_i = I_{\{T_i \le C_i\}}, \quad i = 1, 2, \cdots, n, \tag{1.1}$$

where C_1, C_2, $\cdots$, C_n are censoring times with distribution G, independent of survival times T. We are interested in the estimation problem for a parameter $\theta = \theta(F)$. The true parameter value θ_0 is a unique solution of the equation

$$\mathbf{E}\, g(T, \theta) = 0 \tag{1.2}$$

for some function g. In this paper, we focus on estimating equations having true parameter value θ_0 as the unique solution, since such estimating equations will provide (asymptotically) unbiased estimates. There are many such examples of g in the literature and the solution's existence and uniqueness are discussed therein (Newey and Smith, 2004). Different function g corresponds to different parameter of interests. For example, if we choose $g(t,\,\theta) = m(t) - \theta$, then θ is the expectation of $m(T)$, i.e. $\theta = \mathbf{E}[m(T)] = \int m(t)\,\mathrm{d}F(t)$. Other examples include: [1.] $g(t,\,\theta) = (t - t_0 - \theta)I_{\{t>t_0\}}$ corresponding to θ being the mean residual life time at given time t_0; [2.] $g(t,\,\theta) = I_{\{t>t_0\}} - \exp(-\theta)$ corresponding to θ being the cumulative hazard function at given time t_0; [3.] $g(t,\,\theta) = I_{\{t\le\theta\}} - t_0$ corresponding to θ being the quantile function at given time t_0.

Based on synthetic data introduced in Liang et al. (2019), if T is observed, the pairwise mean synthetic data set can be defined as,

$$\mathcal{M} = \left\{ \frac{g(T_i,\,\theta) + g(T_j,\,\theta)}{2} : 1 \le i \le j \le n \right\}, \tag{1.3}$$

which can also be written as $\mathcal{M} = \{M_1(\theta), M_2(\theta), \cdots, M_{N_0}(\theta)\}$ with $N_0 = n(n+1)/2$. Based on the data set (1.3), the MeanEL ratio for θ is

$$\mathcal{R}(\theta) = \sup\left\{ \prod_{k=1}^{N_0} N_0 p_k \,\middle|\, \sum_{k=1}^{N_0} p_k M_k(\theta) = 0,\ \sum_{k=1}^{N_0} p_k = 1, p_k \ge 0, k = 1, 2, \cdots, N_0 \right\}. \tag{1.4}$$

Under some regularity assumptions, Liang et al. (2019) proved the mean empirical log-likelihood ratio $\mathcal{L}(\theta_0) = -2\log\mathcal{R}(\theta_0)/(n+1) \to \chi^2(1)$, in dist. Therefore, the $(1-\alpha)$ confidence interval can be constructed as $I = \{\theta : \mathcal{L}(\theta) < \chi^2_\alpha(1)\}$.

However, the above approach is not readily available under censoring, since we only observe $(Z_i,\,\eta_i)$ instead of T_i and we cannot pairwise index variable η_i directly. Therefore, we need to develop a new approach to construct, under right censoring, a synthetic data set, an estimating equation and a MeanEL ratio for θ.

This paper is organised as follows. In Section 2, we will present the MeanEL methodologies for right censored data and show that the MeanEL still has a limiting χ^2 distribution, which can be used to construct a MeanEL-based confidence interval. Simulation studies are presented in Section 3 and they demonstrate that MeanEL outperforms the existing methods, especially for heavy-tail distributions. Section 4 provides a real data analysis. A theoretical high-order accuracy justification of different methods are provided in Section 5.

2. Methodology for censored data

Let $Z_{1:n} \le Z_{2:n} \le \cdots \le Z_{n:n}$ be the ordered Z-values and $\eta_{[i:n]}$ be the concomitant of the ith order statistic, that is $\eta_{[i:n]} = \eta_j$ if $Z_{i:n} = Z_j$. Let $G_i = G(Z_{i:n})$, $g_i(\theta) = g(Z_{i:n},\,\theta)$,

$\delta_i = \eta_{[i:n]}$. We define the pairwise mean data set as

$$\mathscr{M}_C = \left\{ \left(\frac{g_i(\theta) + g_j(\theta)}{2},\, \delta_i \delta_j \right) :\, 1 \le i < j \le n \right\}.$$

In this new data set $\mathscr{M}_C$, only those observations satisfying $\delta_i = \delta_j = 1$ can be treated as uncensored. The following equation can be easily proved,

$$\mathbf{E}\left(\frac{(g_i(\theta_0) + g_j(\theta_0))\, \delta_i \delta_j}{2(1 - G_i)(1 - G_j)} \right) = 0, \quad i < j. \tag{2.1}$$

Based on this equation, the MeanEL ratio can be defined as

$$\mathcal{R}_C(\theta_0) = \sup \left\{ \prod_{k=1}^{N} N p_k \,\middle|\, \sum_{k=1}^{N} p_k W_k = 0,\, \sum_{k=1}^{N} p_k = 1,\, p_k \ge 0,\, k = 1,\, 2,\, \cdots,\, N \right\},$$

where $W_1, W_2, \cdots, W_N$ represent the items within the brackets of equation (2.1) and $N = n(n-1)/2$.

The above likelihood is not available since G is unknown. Therefore we here consider using the Kaplan-Meier estimator $\hat{G}$ to replace G, $\hat{G}(t) := 1 - \prod_{i=1}^{n} \left[1 - \frac{1 - \delta_i}{n - i + 1} \right]^{I_{\{Z_{i:n} \le t\}}}$. Let $\hat{G}_i = \hat{G}(Z_{i:n})$, and then the observed quantity for W_k ($k = 1, \cdots, N$) becomes

$$W_{nk} := \frac{(g_i(\theta_0) + g_j(\theta_0))\, \delta_i \delta_j}{2(1 - \hat{G}_i)(1 - \hat{G}_j)}, \quad k = 1, \cdots, N; i, j = 1, \cdots, n. \tag{2.2}$$

Then the MeanEL ratio can be rewritten as

$$\hat{\mathcal{R}}_C(\theta_0) = \sup \left\{ \prod_{k=1}^{N} N p_k \,\middle|\, \sum_{k=1}^{N} p_k W_{nk} = 0,\, \sum_{k=1}^{N} p_k = 1,\, p_k \ge 0,\, k = 1,\, 2,\, \cdots,\, N \right\}.$$

Its large sample results are given in the following Theorem.

Theorem 2.1. *Assume* $\boldsymbol{E}\left(\frac{\delta\, g(Z, \theta_0)}{1 - G(Z)} \right)^4 < \infty$, $\boldsymbol{E}\left(\frac{\delta}{1 - G(Z)} \right)^4 < \infty$, *the mean empirical log-likelihood ratio*

$$\mathcal{L}_C(\theta_0) = -2 \log \hat{\mathcal{R}}_C(\theta_0)/n$$

is asymptotically a scaled χ^2 *random variable, that is*

$$\frac{\sigma_2^2}{\sigma_1^2}\, \mathcal{L}_C(\theta_0) = -\frac{2\sigma_2^2}{n\sigma_1^2} \log \hat{\mathcal{R}}_C(\theta_0) \to \chi^2(1), \quad \textit{in dist.} \tag{2.3}$$

where $\sigma_1^2 = \int g^2(t, \theta_0) F(t-) G(t-)^{-1} d\Lambda(t)$ *with* $\Lambda(t)$ *as the cumulative hazard function for* T,

and $\sigma_2^2 = \frac{2}{N}\sum_{k=1}^{N} W_{nk}^2$.

Proof. From Liang et al. (2019), Theorem 2.1 follows easily from Lemma Appendix A.1 and the fact from Chapter 3 of Fleming and Harrington (1991) that

$$\begin{aligned}\sigma_1^2 &= \lim_{n\to\infty} \mathrm{var}\left[\sqrt{n}\int g(t,\,\theta_0)\mathrm{d}\hat{F}(t)\right] \\ &= \lim_{n\to\infty} \mathbf{E}\left[\sqrt{n}\int g(t,\,\theta_0)\hat{F}(t-)\left(\mathrm{d}\hat{\Lambda}(t)-\mathrm{d}\Lambda(t)\right)\right]^2 \\ &= \lim_{n\to\infty} \mathbf{E}\int g(t,\,\theta_0)^2\hat{F}(t-)^2\frac{n}{\bar{Y}(t)}\mathrm{d}\Lambda(t) = \int g(t,\,\theta_0)^2\frac{F(t-)}{G(t-)}\mathrm{d}\Lambda(t)\end{aligned}$$

where $\bar{Y}(t) = \sum_{i=1}^{n} I_{\{Z_i\geq t\}}$ and $\mathrm{d}\hat{\Lambda}(t) = \bar{Y}(t)^{-1}\mathrm{d}Q(t)$, $Q(t) = \sum_{i=1}^{n} I_{\{Z_i\leq t,\,\delta_i=1\}}$. □

A consistent estimator for σ_1^2 is $\hat{\sigma}_1^2 = \int g^2(t,\hat{\theta})\hat{F}(t-)\hat{G}(t-)^{-1}\mathrm{d}\hat{\Lambda}(t)$, with $\hat{F},\hat{G}$ as the product-limit estimators, and $\hat{\theta}$ as the solution of $\int g(t,\,\theta)\,\mathrm{d}\hat{F}(t) = 0$. In the asymptotic result (2.3), if we replace σ_2^2 by

$$\hat{\sigma}_{2A}^2 = \frac{2}{N}\sum_{i<j}\left(\frac{\delta_i\,\delta_j\,(g_i(\hat{\theta})+g_j(\hat{\theta}))}{2(1-\hat{G}_i)(1-\hat{G}_j)}\right)^2,$$

then an α-level MeanEL confidence interval for θ can be constructed as follows

$$I_A = \left\{\theta:\ \frac{\hat{\sigma}_{2A}^2}{\hat{\sigma}_1^2}\ \mathcal{L}_C(\theta) \leq \chi_\alpha^2(1)\right\}. \tag{2.4}$$

We may also replace σ_2^2 by $\hat{\sigma}_{2B}^2 = \hat{\sigma}_{22}^2\hat{\sigma}_{20}^2 + \hat{\sigma}_{21}^4$ in (2.3), where

$$\hat{\sigma}_{20}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{(1-\hat{G}_i)^2},\qquad \hat{\sigma}_{21}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\,g_i(\hat{\theta})}{(1-\hat{G}_i)^2},\qquad \hat{\sigma}_{22}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\,g_i^2(\hat{\theta})}{(1-\hat{G}_i)^2}.$$

Then another α-level MeanEL confidence interval for θ can be constructed as follows

$$I_B = \left\{\theta:\ \frac{\hat{\sigma}_{2B}^2}{\hat{\sigma}_1^2}\ \mathcal{L}_C(\theta) \leq \chi_\alpha^2(1)\right\}. \tag{2.5}$$

3. Simulation studies

For a given sample size n, we generate lifetime observations $T_1, T_2, \cdots, T_n$ from a specific distribution F and censoring time observations $C_1, C_2, \cdots, C_n$ from certain censoring distribution G. Then, based on the simulated data, we can compare the performance of IC-confidence interval (He et al., 2016), ScaledEL-confidence interval (Wang and Jing, 2001) and MeanEL-confidence intervals I_A, I_B proposed in the previous section.

Table 1: Coverage probabilities for Uniform(0, 1) Distributions

		$C \sim$ U(0, 5/2)				$C \sim$ U(0, 5/3)				$C \sim$ U(0, 5/4)			
		IC	Scaled	Mean-A	Mean-B	IC	Scaled	Mean-A	Mean-B	IC	Scaled	Mean-A	Mean-B
20	0.95	0.9350	0.9280	**0.9570**	0.9414	0.9348	0.9280	**0.9563**	0.9384	0.8638	0.8397	**0.8877**	0.8186
	0.90	0.8812	0.8785	0.9127	**0.8901**	0.8806	0.8745	**0.9097**	0.8876	0.8048	0.7785	**0.8389**	0.7549
30	0.95	0.9392	0.9361	0.9586	**0.9493**	0.9439	0.9399	0.9533	**0.9501**	0.9029	0.8873	**0.9281**	0.8754
	0.90	0.8871	0.8860	0.9114	**0.8969**	0.8899	0.8873	**0.8997**	0.9156	0.8439	0.8267	**0.8789**	0.8119
40	0.95	0.9464	0.9438	0.9604	**0.9535**	**0.9493**	0.9476	0.9523	0.9561	0.9174	0.9075	**0.9435**	0.9049
	0.90	0.8939	0.8926	0.9152	**0.9041**	**0.8956**	0.8942	0.9073	0.9083	0.8644	0.8489	**0.8949**	0.8430
50	0.95	**0.9489**	0.9469	0.9608	0.9561	**0.9476**	0.9451	0.9600	0.9555	0.9269	0.9201	**0.9497**	0.9198
	0.90	**0.8977**	0.8963	0.9143	0.9065	**0.8963**	0.8958	0.9141	0.9057	0.8715	0.8637	**0.9045**	0.8615
100	0.95	**0.9497**	0.9504	0.9571	0.9548	**0.9494**	0.9435	0.9567	0.9543	0.9430	0.9398	0.9593	**0.9457**
	0.90	0.8986	**0.8994**	0.9089	0.9053	**0.9003**	0.9005	0.9088	0.9056	0.8867	0.8859	0.9136	**0.8944**

The boldface results are the most accurate coverage probabilities among EL methods.

Table 2: Coverage probabilities for Weibull(1, 10) Distributions

		$C \sim$ Exp(4.3)				$C \sim$ Exp(2.7)				$C \sim$ Exp(1.86)			
		IC	Scaled	Mean-A	Mean-B	IC	Scaled	Mean-A	Mean-B	IC	Scaled	Mean-A	Mean-B
20	0.95	0.9291	0.9268	**0.9528**	0.9433	0.9226	0.9168	**0.9484**	0.9367	0.9175	0.9036	**0.9414**	0.9258
	0.90	0.8747	0.8764	**0.8943**	0.9082	0.8664	0.8637	**0.9025**	0.8839	0.8621	0.8553	**0.8973**	0.8761
30	0.95	0.9390	0.9374	0.9555	**0.9497**	0.9346	0.9291	**0.9537**	0.9456	0.9361	0.9269	**0.9541**	0.9443
	0.90	0.8860	0.8875	0.9087	**0.9002**	0.8794	0.8772	**0.9051**	0.8933	0.8836	0.8782	0.9096	**0.8960**
40	0.95	0.9421	0.9407	0.9558	**0.9510**	0.9453	0.9415	0.9583	**0.9543**	0.9396	0.9318	0.9561	**0.9487**
	0.90	0.8865	0.8875	0.9047	**0.8973**	0.8910	0.8894	0.9111	**0.9020**	0.8868	0.8807	0.9064	**0.8955**
50	0.95	0.9421	0.9410	0.9543	**0.9511**	0.9456	0.9424	0.9569	**0.9531**	0.9435	0.9369	0.9563	**0.9506**
	0.90	0.8895	0.8901	0.9041	**0.8989**	**0.8928**	0.8916	0.9097	0.9092	0.8894	0.8855	0.9071	**0.8984**
100	0.95	0.9466	0.9464	0.9518	**0.9504**	**0.9493**	0.9481	0.9557	0.9533	**0.9488**	0.9467	0.9551	0.9528
	0.90	0.8942	0.8949	**0.9010**	0.8986	0.8974	0.8964	0.9055	**0.9021**	**0.8993**	0.8975	0.9041	0.9077

The boldface results are the most accurate coverage probabilities among EL methods.

In our simulation, the parameter of interests, θ, is the mean of T, therefore the estimating equation is $g(T, \theta) = T - \theta$. Uniform, Weibull and LogNorm distributions are considered as the underlying lifetime distribution F. Let Unif(0, c) and Exp(c) be different censoring distributions G, where the value c determines the censoring proportion. We set c to be three different values to achieve a 20%, 30% and 40% censoring proportion respectively. Based on $20,000$ sets of simulated data, we construct IC confidence intervals, Scaled confidence intervals, Mean-A and Mean-B confidence intervals. The coverage probabilities with Uniform and Weibull distributions are summarized in Table 1 and Table 2. We also plot in Figure 1 the results of coverage probabilities with T following LogNorm distribution and nominal level set at 0.95.

From these Tables and Figure 1, the following results are noted.

1. As the sample size n increases, all coverage probabilities converge to the nominal level. When the sample size n is fixed, coverage probabilities of all methods decrease as the censoring proportion increases. In most cases, coverage probabilities of MeanELs are closer to the nominal level than the other two methods.

2. The coverage probabilities of Mean-A (using (2.4)) are a little higher than that of Mean-B (using (2.5)). More specifically, for Uniform distribution, the coverage probabilities of Mean-B are closer to the nominal level than that of Mean-A when censoring proportion

is small, but Mean-A performs better than Mean-B when censoring proportion is large. For heavy-tailed distribution LogNorm(0, 1/4), the performance of Mean-A is the best among all different methods.

3. Different distributions of T lead to different results. For Uniform distribution, IC is more accurate than others when censoring proportion is low, while MeanEL performs better when censoring proportion is high. For Weibull distribution, Mean-B is closer to nominal level when sample size is small.

In Figure 1, Mean-A performs much better than other EL methods when censoring proportion is large. This is what we expected. The tail properties of survival times and censoring times are very important to the accuracy of the estimators. Under heavy-censoring there may not have enough observations for the tails, in particular for small sample sizes. MeanEL method actually uses each data point more than once. Using replicated i.i.d. data will not make a great deal of difference when each observation is of the same weight. However, because EL-type approaches give different weights for all observations, using observations (in the tails) more than once will give a great advantage for the estimator.

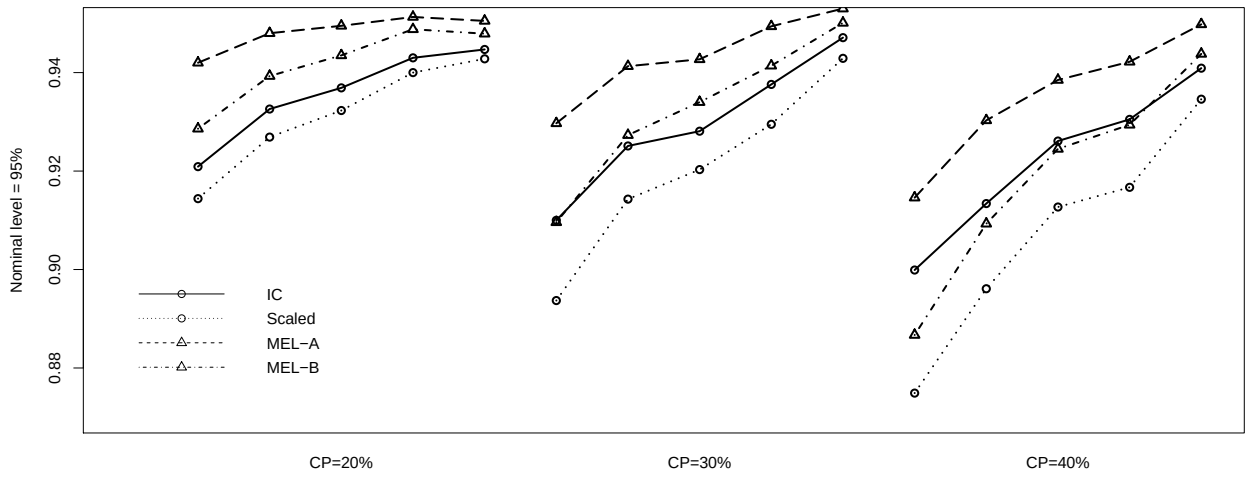

Figure 1: For LogNorm(0, 0.25), coverage probabilities under different censoring proportion. Here the five points (from left to right) in each line represent coverage probabilities when sample sizes are $n = 20$; 30; 40; 50; 100. 'Scaled' represents the method in Wang and Jing (2001), 'IC' represents the method in He et al. (2016), 'MEL-A' represents MeanEL-A, and 'MEL-B' represents MeanEL-B.

All of our simulation studies confirm that MeanEL outperforms IC and Scaled EL methods under right censoring, in terms of the expected coverage probabilities. Of course, the proposed MeanEL approach requires much heavier computational cost because of the enlarged synthetic dataset. Therefore, for light-tail or symmetric distributions, such as uniform distributions

Table 3: Average Lengths of Confidence Intervals for different Distributions under nominal level= 0.90.

	IC	Scaled	Mean-A	Mean-B	IC	Scaled	Mean-A	Mean-B	IC	Scaled	Mean-A	Mean-B
Unif(0, 1)	$C \sim$ U(0, 5/2)				$C \sim$ U(0, 5/3)				$C \sim$ U(0, 5/4)			
20	0.2181	0.2188	0.2529	0.2379	0.2265	0.2206	0.2699	0.2449	0.2335	0.2186	0.2782	0.2392
30	0.1803	0.1809	0.2087	0.2002	0.1886	0.1853	0.2358	0.2203	0.1988	0.1866	0.2651	0.2369
40	0.1568	0.1576	0.1819	0.1765	0.1644	0.1622	0.2075	0.1965	0.1750	0.1658	0.2483	0.2258
50	0.1407	0.1414	0.1614	0.1577	0.1472	0.1460	0.1885	0.1805	0.1579	0.1509	0.2368	0.2181
100	0.1000	0.1003	0.1110	0.1098	0.1041	0.1040	0.1322	0.1293	0.1115	0.1089	0.1791	0.1698
Weibull(1, 10)	$C \sim$ Exp(4.3)				$C \sim$ Exp(2.7)				$C \sim$ Exp(1.86)			
20	0.0905	0.0927	0.1017	0.0971	0.0961	0.0984	0.1096	0.1035	0.1026	0.1046	0.1180	0.1096
30	0.0749	0.0761	0.0819	0.0796	0.0794	0.0805	0.0881	0.0849	0.0854	0.0864	0.0964	0.0919
40	0.0653	0.0661	0.0706	0.0691	0.0693	0.0700	0.0761	0.0742	0.0744	0.0749	0.0826	0.0798
50	0.0584	0.0589	0.0624	0.0613	0.0618	0.0623	0.0674	0.0660	0.0665	0.0669	0.0737	0.0717
100	0.0414	0.0416	0.0439	0.0436	0.0439	0.0440	0.0475	0.0470	0.0470	0.0472	0.0516	0.0510
LogN(0, 1/4)	$C \sim$ Exp(4.6)				$C \sim$ Exp(2.9)				$C \sim$ Exp(2)			
20	0.2049	0.2068	0.2360	0.2241	0.2161	0.2160	0.2514	0.2351	0.2271	0.2231	0.2615	0.2393
30	0.1709	0.1715	0.1982	0.1914	0.1805	0.1789	0.2135	0.2039	0.1924	0.1881	0.2306	0.2166
40	0.1490	0.1492	0.1758	0.1714	0.1573	0.1558	0.1907	0.1842	0.1685	0.1647	0.2085	0.1987
50	0.1336	0.1335	0.1597	0.1564	0.1419	0.1404	0.1761	0.1712	0.1517	0.1481	0.1945	0.1870
100	0.0950	0.0948	0.1204	0.1192	0.1009	0.1000	0.1386	0.1366	0.1086	0.1067	0.1570	0.1539

(Table 1), MeanEL is not very attractive since it does not improve much on the accuracy of coverage probabilities but has longer computing time. We also present the average lengths of confidence intervals in Table 3. From this Table, we can see that, unsurprisingly the average length of confidence intervals constructed by MeanEL are a little longer than others.

4. Real Data Analysis

In this section, we compare our proposed methods with existing methods using the primary biliary cirrhosis(PBC) dataset, which is described in Fleming and Harrington (1991) and originates from a Mayo Clinic trial between 1974 to 1984. It contains the survival time of 312 patients and the status variable, which indicates if the patients' survival times are censored. We use this dataset to illustrate our proposed method described in Section 2. Figure 2 presents the 95% confidence intervals for the mean survival time base on four different methods, Mean-A, Mean-B described in Section 2 and the methods in Wang and Jing (2001) and He et al. (2016). All confidence intervals produced by Mean-A, IC and Scale methods contain the Maximum Empirical Likelihood point estimate value 3286. Mean-A provides similar confidence interval as Scale but performs much better than IC. Mean-B does not work as well as Mean-A, because the estimate $\hat{\sigma}^2_{2B}$ is not as good as $\hat{\sigma}^2_{2A}$. This can be explained by that $\hat{\sigma}^2_{2A}$ is actually asymptotically equivalent to a U-statistic, which has the minimum variance among all unbiased estimators (Lee, 1990). Also when deriving $\hat{\sigma}^2_{2B}$, we actually omitted the last term in equation (A.1), which makes $\hat{\sigma}^2_{2B}$ to have larger bias than $\hat{\sigma}^2_{2A}$.

5. Theoretical comparisons

In this section, we present a theoretical comparison of MeanEL and other EL methods. Define $A_k = n^{-1}\sum_{i=1}^n g_i^k(\theta_0) - \alpha_k$ with $\alpha_k = \mathbf{E}\, g_i^k(\theta_0)$. Following Liu and Chen (2010), we

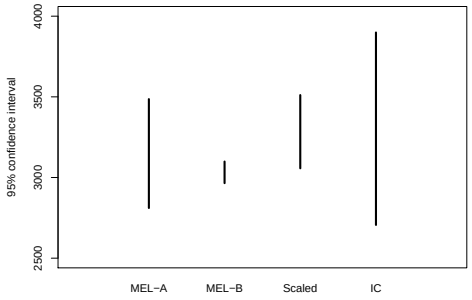

Figure 2: Confidence intervals based on different methods. 'Scaled' represents the method in Wang and Jing (2001), 'IC' represents the method in He et al. (2016), 'MEL-A' represents MeanEL-A, and 'MEL-B' represents MeanEL-B.

assume that $\alpha_2 = 1$, then the original EL can be written as

$$\mathcal{R}_O(\theta_0) = n(R_1 + R_2 + R_3)^2 + O_p(n^{-\frac{3}{2}}),$$

and the Bartlett correction uses the corrected statistics

$$\mathcal{R}_B(\theta_0) = n\left(1 - \frac{b}{n}\right)(R_1 + R_2 + R_3)^2 + O_p(n^{-\frac{3}{2}}), \tag{5.1}$$

where $b = \frac{1}{2}\alpha_4 - \frac{1}{3}\alpha_3^2$, $R_1 = A_1$, $R_2 = \frac{1}{3}\alpha_3 A_1^2 - \frac{1}{2}A_1 A_2$ and $R_3 = \frac{3}{8}A_1 A_2^2 + \frac{4}{9}\alpha_3^2 A_1^2 - \frac{5}{6}\alpha_3 A_1^2 A_2 + \frac{1}{3}A_1^2 A_3 - \frac{1}{4}\alpha_4 A_1^3$. Then the corrected statistic $\mathcal{R}_B(\theta_0)$ gives second order accuracy, i.e. $P\{\hat{R}_B(\theta_0) \le z\} = P\{\chi_1^2 \le z\} + O(n^{-2})$. Based on this idea, we now present the accuracy of MeanEL in the following theorem.

Theorem 5.1. *The MeanEL ratio $\mathcal{R}(\theta_0)$, defined in (1.4), can be written as*

$$\mathcal{R}(\theta_0) = n\left(R_1^{(1)} + R_2^{(1)} + R_3^{(1)}\right)^2 + O_p(n^{-\frac{3}{2}})$$

where $R_1^{(1)} = R_1$, $R_2^{(1)} = R_2$ and $R_3^{(1)} = R_3 - \frac{1}{8}A_1\left((\alpha_3 A_1 - A_2)^2 + 2A_1^2 + \frac{4}{n+1}\right)$. We can further write

$$\mathcal{R}(\theta_0) = n\left(1 - \frac{b_1}{n}\right)(R_1 + R_2 + R_3)^2 + O_p(n^{-\frac{3}{2}})$$

with $b_1 = \frac{n}{4}(\alpha_3 A_1 - A_2)^2 + \frac{n}{2}A_1^2 + 1$ and $\boldsymbol{E}(b_1) = \frac{1}{4}\alpha_4 - \frac{\alpha_3^2}{4} + \frac{5}{4}$.

Proof. See Appendix. □

We have $b - \mathbf{E}(b_1) = \frac{1}{4}\alpha_4 - \frac{1}{12}\alpha_3^2 - \frac{5}{4}$. The formula of b, defined in (5.1), implies for

distributions with kurtosis α_4 much larger than skewness α_3 (such as log-normal), Bartlett correction will be significantly better than original EL. The new MeanEL approach is actually equivalent to using "Bartlett correction constant" $\mathbf{E}(b_1)$, such that $\mathbf{E}(b_1) \in (0, b)$. Therefore, MeanEL will always be better than the original empirical likelihood (correction constant to be 0), for such distributions. Also the advantage of MeanEL is that it does not need to estimate the correction constant b.

We also compared the results of using pair-wise mean data and the results of using three-value mean data $((g_i(\theta) + g_j(\theta) + g_k(\theta))/3)$ in the supplementary file. Both theoretical justifications and simulation studies show that using the pair-wise mean data will provide better results than using three-value mean data.

Appendix A. Proof of Theorem 2.1

Lemma Appendix A.1. *Assume* $\boldsymbol{E}\left(\frac{\delta\, g(Z,\theta_0)}{1-G(Z)}\right)^4 < \infty$, $\boldsymbol{E}\left(\frac{\delta}{1-G(Z)}\right)^4 < \infty$, *we have*

$$
\begin{aligned}
&(i) \qquad \max_{1\le k\le N} |W_{nk}| = o_p(n^{1/2}).\\
&(ii) \qquad \sqrt{n}\left(\frac{1}{N}\sum_{k=1}^{N} W_{nk}\right) \to N(0,\, \sigma_1^2), \quad \textit{in dist.}\\
&(iii) \qquad \frac{1}{N}\sum_{k=1}^{N} W_{nk}^2 = O_p(1).\\
&(iv) \qquad \frac{1}{N}\sum_{k=1}^{N} W_{nk}^2 = \frac{1}{2}\hat{\sigma}_{2A}^2 + o_p(1), \qquad \frac{1}{N}\sum_{k=1}^{N} W_{nk}^2 = \frac{1}{2}\hat{\sigma}_{2B}^2 + o_p(1).
\end{aligned}
$$

Proof. Part (i) follows form Liang et al. (2019) and the result $\sup_{0\le z\le Z_{n:n}} \left|\frac{\hat{G}(z)-G(z)}{1-\hat{G}(z)}\right| = O_p(1)$ in Zhou (1992). We provide a sketch of proofs for other parts of the Lemma here and details can be found in the supplementary file.

For part (ii), we rewrite $N^{-1}\sum_{k=1}^{N} W_{nk}$ as

$$
\frac{1}{N}\sum_{k=1}^{N} W_{nk} \;=\; \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{1-\hat{G}_i}\right)\left(\frac{1}{n-1}\sum_{i=1}^{n}\frac{\delta_i}{1-\hat{G}_i}\right) - \frac{1}{n(n-1)}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{(1-\hat{G}_i)^2}.
$$

Noting that

$$
\begin{aligned}
&\left|\frac{1}{n(n-1)}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{(1-\hat{G}_i)^2}\right| \le \left|\frac{1}{n(n-1)}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{(1-G_i)^2}\right| \max_i \left|\frac{1-G_i}{1-\hat{G}_i}\right|^2 = O_p(n^{-1}),\\
&n^{-1}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{1-\hat{G}_i} = \int_0^\infty g(t,\,\theta_0)\,\mathrm{d}\hat{F}(t) = \int_0^\infty g(t,\,\theta_0)\,\mathrm{d}(\hat{F}(t)-F(t)),
\end{aligned}
$$

and $n^{-1}\sum_{i=1}^{n}\frac{\delta_i}{1-\hat{G}_i}=\int_0^\infty \mathrm{d}\hat{F}(t)=1$, we get

$$\sqrt{n}\left(\frac{1}{N}\sum_{k=1}^{N}W_{nk}\right)=\sqrt{n}\int_0^\infty g(t,\,\theta_0)\,\mathrm{d}(\hat{F}(t)-F(t))+o_p(1).$$

Following Corollary 1.2 in Stute (1995) we have the conclusion (ii).

For part (iii) we rewrite $N^{-1}\sum_{k=1}^{N}W_{nk}^2$ as

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}W_{nk}^2 &= \frac{n}{2(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i^2(\theta_0)}{(1-\hat{G}_i)^2}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{(1-\hat{G}_i)^2}\right)\\ &\quad+\frac{n}{2(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{(1-\hat{G}_i)^2}\right)^2-\frac{1}{n(n-1)}\sum_{i=1}^{n}\frac{\delta_i g_i^2(\theta_0)}{(1-\hat{G}_i)^4}. \qquad \text{(A.1)}\end{aligned}$$

The last term of (A.1) is bounded above by $\left|\frac{1}{n(n-1)}\sum_{i=1}^{n}\frac{\delta_i g_i^2(\theta_0)}{(1-G_i)^4}\right|\max_i\left|\frac{1-G_i}{1-\hat{G}_i}\right|^4=O_p(n^{-1})$ and can be omitted. In addition, we have

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i^2(\theta_0)}{(1-\hat{G}_i)^2}-\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i^2(\theta_0)}{(1-G_i)^2}\right|=\left|\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{G}_i-G_i)(1-\hat{G}_i+1-G_i)}{(1-\hat{G}_i)^2}\frac{\delta_i g_i^2(\theta_0)}{(1-G_i)^2}\right|\\ \le\;&\left(\max_i\left|\frac{\hat{G}_i-G_i}{1-\hat{G}_i}\right|+\max_i\left|\frac{(\hat{G}_i-G_i)(1-G_i)}{(1-\hat{G}_i)^2}\right|\right)\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i^2(\theta_0)}{(1-G_i)^2}=O_p(1),\end{aligned}$$

and

$$\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{(1-\hat{G}_i)^2}-\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{(1-G_i)^2}\right|=O_p(1),\quad \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{(1-\hat{G}_i)^2}-\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{(1-G_i)^2}\right|=O_p(1).$$

Therefore, part (iii) is proved.

For part (iv), since $\hat{\theta}$ is a consistent estimator of θ_0 and

$$\frac{1}{N}\sum_{k=1}^{N}W_{nk}^2-\frac{1}{2}\hat{\sigma}_{2A}^2=\frac{1}{4N}\sum_{i<j}\frac{\delta_i\,\delta_i}{(1-\hat{G}_i)^2(1-\hat{G}_j)^2}\left[(g_i(\theta_0)+g_j(\theta_0))^2-(g_i(\hat{\theta})+g_j(\hat{\theta}))^2\right],$$

we have $\frac{1}{N}\sum_{k=1}^{N}W_{nk}^2=\frac{1}{2}\hat{\sigma}_{2A}^2+o_p(1)$. Also from equation (A.1) we have,

$$\frac{1}{N}\sum_{k=1}^{N}W_{nk}^2-\frac{1}{2}\hat{\sigma}_{2B}^2 = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i^2(\theta_0)}{(1-\hat{G}_i)^2}-\hat{\sigma}_{22}^2\right)\hat{\sigma}_{20}^2 \qquad \text{(A.2)}$$

$$+\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{(1-\hat{G}_i)^2}-\hat{\sigma}_{21}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i g_i(\theta_0)}{(1-\hat{G}_i)^2}+\hat{\sigma}_{21}^2\right)+o_p(1).$$

Since $\hat{\theta}$ is a consistent estimator of θ_0, then both the first term and the second term of (A.2) are $o_p(1)$. Therefore we have $\frac{1}{N}\sum_{k=1}^{N}W_{nk}^2=\frac{1}{2}\hat{\sigma}_{2B}^2+o_p(1)$. This implies that $\lim_{n\to\infty}\hat{\sigma}_{2A}^2=\lim_{n\to\infty}\hat{\sigma}_{2B}^2=\lim_{n\to\infty}\frac{2}{N}\sum_{k=1}^{N}W_{nk}^2$. □

Note that, the assumptions required by this Lemma are always satisfied based on the commonly-used regularity condition for the tails of the survival and censoring distributions (Stute and Wang, 1993), which guarantees the survival function F is identifiable.

Appendix B. Proof of Theorem 5.1

Proof. Let $B_k=N_0{}^{-1}\sum_{l=1}^{N_0}M_l^k(\theta_0)-\beta_k$ with $\beta_k=\mathbf{E}\,M_l^k(\theta_0)$. Following Liu and Chen (2010),

$$\begin{aligned}
&N_0{}^{-1}\left(2\sum_{k=1}^{N_0}\log(1+\lambda M_k(\theta_0))\right)=2n^{-1}\mathcal{R}(\theta_0)\\
=\;&\frac{1}{\beta_2}B_1^2+\frac{2\beta_3}{3\beta_2^3}B_1^3-\frac{1}{\beta_2^2}B_1^2B_2+\left(\frac{2\beta_3^2}{\beta_2^4}-\frac{\beta_3^2}{\beta_2^3}-\frac{\beta_4}{2\beta_2^4}\right)B_1^4+\frac{2}{3\beta_2^3}B_1^3B_3\\
&+\left(\frac{2\beta_3}{\beta_2^2}-\frac{4\beta_3}{\beta_2^3}\right)B_1^3B_2+\left(\frac{2}{\beta_2^2}-\frac{1}{\beta_2}\right)B_1^2B_2^2+O_p\left(n^{-5/2}\right) \qquad \text{(B.1)}
\end{aligned}$$

Under $\alpha_1=0$ and the assumption $\alpha_2=1$, we have

$$\begin{aligned}
&\beta_1=0,\quad \beta_2=\frac{1}{2},\quad \beta_3=\frac{1}{4}\alpha_3,\quad \beta_4=\frac{1}{8}\alpha_4+\frac{3}{8},\\
&B_1=A_1,\quad B_2=\frac{1}{2}A_2+\frac{1}{2}A_1^2+\frac{1}{2(n+1)}+O_p\left(n^{-2/3}\right),\\
&B_3=\frac{1}{4}A_3+\frac{3}{4}A_1+\frac{3}{4}A_1A_2+\frac{3\alpha_3}{4(n+1)}+O_p\left(n^{-2/3}\right).
\end{aligned}$$

Substitute these equations into (B.1) to get

$$\begin{aligned}
n^{-1}\mathcal{R}(\theta_0)\;=\;&A_1^2+\frac{2}{3}\alpha_3A_1^3-A_1^2A_2+\left(\frac{3}{4}\alpha_3^2-\frac{1}{2}\alpha_4-\frac{1}{2}\right)A_1^4\\
&+\frac{2}{3}A_1^3A_3-\frac{3}{2}\alpha_3A_1^3A_2+\frac{3}{4}A_1^2A_2^2-\frac{A_1^2}{n+1}+O_p\left(n^{-5/2}\right).
\end{aligned}$$

Assuming $\mathcal{R}(\theta_0) = n\left(R_1^{(1)} + R_2^{(1)} + R_3^{(1)}\right)^2 + O_p(n^{-\frac{3}{2}})$, where $R_1^{(1)} = O_p(n^{-1/2})$, $R_2^{(1)} = O_p(n^{-1})$ and $R_3^{(1)} = O_p(n^{-3/2})$, then we have

$$R_1^{(1)} = A_1, \quad R_2^{(1)} = \frac{1}{3}\alpha_3 A_1^2 - \frac{1}{2}A_1 A_2,$$
$$R_3^{(1)} = \left(\frac{23}{72}\alpha_3^2 - \frac{1}{4}\alpha_4 - \frac{1}{4}\right)A_1^3 + \frac{1}{3}A_1^2 A_3 - \frac{7}{12}\alpha_3 A_1^2 A_2 + \frac{1}{4}A_1 A_2^2 - \frac{A_1}{2(n+1)}.$$

Further, we write $n\left(R_1^{(1)} + R_2^{(1)} + R_3^{(1)}\right)^2 + O_p(n^{-\frac{3}{2}}) = n\left(1 - \frac{b_1}{n}\right)(R_1 + R_2 + R_3)^2 + O_p(n^{-\frac{3}{2}})$. Expand and simplify this equation with $R_1 = R_1^{(1)}, R_2 = R_2^{(1)}$, then the theorem is proved. □

References

Chen J., Variyath A.M. and Abraham B. (2008). Adjusted empirical likelihood and its Properties. *Journal of Computational Graphical Statistics* **17**, 426–443.

DiCiccio T., Hall P. and Romano J. (1991) Empirical likelihood is Bartlett-correctable. *Annals of Statistics* **19**, 1053–1061.

Fleming T.R. and Harrington D.P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

He S.Y., Liang W., Shen J.S. and Yang G. (2016). Empirical Likelihood for Right Censored Lifetime Data. *Journal of Amreican Statistical Association* **514**, 646-655.

Lee A. J. (1990). *U-Statistics: Theory and Practice*. New York: Dekker.

Liang W., Dai H. and He S.Y. (2019). Mean empirical likelihood. *Computational Statistics and Data Analysis* **138**, 155-169.

Liu Y. and Chen J. (2010). Adjusted Empirical Likelihood with High-Order Precision. *The Annals of Statistics*, **38**, 1341-1362.

Newey W. K. and Smith, R.J. (2004). Higher Order Properties of GMM Generalized Empirical Likelihood Estimators. *Econometrica,* **72**, 219–255.

Stute W. and Wang J. L. (1993). The Strong Law under Random Censorship. *Annals of Statistics,* **21**, 1591–1607.

Stute W. (1995). The central limit theorem under random censorship. *Annals of Statistics* **23(2)**, 422–439.

Taso M. and Wu F. (2013). Empirical likelihood on the full parameter space. *Annals of Statistics* **41**, 2176-2196.

Wang Q.H. and Jing B.Y. (2001). Empirical likelihood for a class of functions of survival distribution with censored data. *Annals of the Institute of Statistical Mathematics* **53**, 517-527.

Zhou M. (1992). Asymptotic normality of the synthetic data regression estimator for censored survival data. *Annals of Statistics*. **20**, 1002-1021.