Empirical Likelihood Ratio With Arbitrarily Censored/Truncated Data by EM Algorithm

Mai ZHOU

The empirical likelihood is a general nonparametric inference procedure with many desirable properties. Recently, theoretical results for empirical likelihood with certain censored/truncated data have been developed. However, the computation of empirical likelihood ratios with censored/truncated data is often nontrivial. This article proposes a modified self-consistent/EM algorithm to compute a class of empirical likelihood ratios for arbitrarily censored/truncated data with a mean type constraint. Simulations show that the chi-square approximations of the log-empirical likelihood ratio perform well. Examples and simulations are given in the following cases: (1) right-censored data with a mean parameter; and (2) left-truncated and right-censored data with a mean type parameter.

Key Words: Constrained maximization; Self consistency; Wilks theorem.

1. INTRODUCTION

The empirical likelihood method was first proposed by Thomas and Grunkemeier (1975) to obtain better confidence intervals in connection with the Kaplan-Meier estimator. Owen (1988, 1990) and many others developed this into a general methodology. It has many desirable statistical properties; see Owen (2001). A crucial step in carrying out the empirical likelihood ratio method is to find the maximum of the log-empirical likelihood function (LELF) under some constraints. In all the articles mentioned above, this is achieved by using the Lagrange multiplier method, which reduces the maximization over $n-1$ variables to a set of k equations. Furthermore, k is small and fixed as n goes to infinity. These equations can be solved easily, and thus the empirical likelihood ratio can be obtained.

Recently, the empirical likelihood ratio method has been shown to work with certain censored/truncated data involving a weighted mean or hazard parameter. Pan and Zhou (1999) showed that, for right-censored data, the empirical likelihood ratio with a mean or hazard constraint also has a chi-square limit (Wilks theorem). Murphy and van der Vaart

Mai Zhou is Associate Professor, Department of Statistics, University of Kentucky, Lexington, KY 40506 (E-mail: mai@ms.uky.edu).

Journal of Computational and Graphical Statistics, Volume 14, Number 3, Pages 643–656
DOI: 10.1198/106186005X59270

(1997) demonstrated, among other things, that Wilks theorem also holds for doubly censored data. For truncated data, we refer readers to Li (1995) for a similar result.

These theoretical results painted a bright future for application of the empirical likelihood in the analysis of censored/truncated data. However, maximizing the censored/truncated empirical likelihood under mean constraints remains difficult, as a simple Lagrange multiplier calculation is often not available (see Examples 1 and 2 later). The articles cited above that studied the theoretical properties of the censored empirical likelihood ratio do not offer a viable computational method either.

This article proposes a modified self-consistent/EM algorithm to compute the censored/truncated data empirical likelihood ratio under mean type constraints. The proposed algorithm can handle very general types of censored/truncated data as described by Turnbull (1976). It can also handle either weighted mean or hazard constraints, but here the focus is on the weighted mean constraints.

We have implemented this algorithm in R software (Gentleman and Ihaka 1996) for right-, left-, or doubly censored data, and for left-truncated, right-censored data with a mean type constraint. It is available as a user contributed package, called `emplik` on CRAN http://cran.r-project.org. See the functions `el.cen.EM()` and `el.ltrc.EM()` inside the `emplik` package.

Section 2 begins with a description of the proposed algorithm and proceeds to show that the proposed self-consistent equation is equivalent to the log-likelihood equation. Section 3 gives some specific calculations of the E-step for commonly seen types of censored/truncated data. Section 4 gives examples and simulations of the empirical likelihood ratio computations. Section 5 contains some further discussion.

We end this section with two specific examples that introduce the notation and setup of censored data, empirical likelihood, and the mean constraint. The computation of the empirical likelihood ratio in these two examples can be handled easily by the proposed EM algorithm, but is otherwise difficult to accomplish.

Example 1. Suppose iid observations $X_1, \ldots, X_n \sim F(\cdot)$ are subject to right censoring, so that we only observe

$$T_i = \min(X_i, C_i)\ ; \quad \delta_i = I_{[X_i \le C_i]}, \quad i = 1, \ldots, n; \tag{1.1}$$

where C_i is the censoring time for X_i. We assume that C_i is independent of X_i.

The log-empirical likelihood function (LELF) for $F(\cdot)$ based on the censored observations (T_i, δ_i) is

$$L(p) = \text{LELF} = \sum_{i=1}^{n} \left[\delta_i \log p_i + (1 - \delta_i) \log \left(\sum_{T_j > T_i} p_j \right) \right], \tag{1.2}$$

where $p_i = \Delta F(T_i) = F(T_i) - F(T_i-)$.

To compute the empirical likelihood ratio for testing the hypothesis H_0: mean$(g(X)) = \mu$, where $g(\cdot)$ and μ are given, we need to find the maximum of the above LELF with respect

to p_i under the constraints

$$\sum_{i=1}^{n} p_i g(T_i) = \mu \,, \qquad \sum_{i=1}^{n} p_i = 1 \,, \qquad p_i \geq 0 \,. \tag{1.3}$$

Arguments similar to those of Li (1995) show that the maximization will force the $p_i = 0$ except where T_i is uncensored. We focus on finding those nonzero p_is. Straightforward application of the Lagrange multiplier method leads to the equations

$$\frac{\delta_i}{p_i} + \sum_{k=1}^{n} (1-\delta_k) \frac{I_{[T_k < T_i]}}{\sum_{T_j > T_k} p_j} - \lambda[g(T_i) - \mu] - \gamma = 0 \,; \quad \text{for } \delta_i = 1 \,,$$

which are not easy to solve for the p_i's.

Example 2. Let $X_1, \ldots, X_n$ be iid positive random variables denoting lifetimes. Let the CDF of X_i be F. The censoring mechanism is such that X_i is observable if and only if it lies inside the interval $[Z_i, Y_i]$. The Z_i and Y_i are positive random variables, independent of X_i, with continuous distribution functions G_L and G_R, respectively, and $Z_i \leq Y_i$ with probability one.

The lifetime X_i is said to be left-censored if $X_i < Z_i$ and right-censored if $X_i > Y_i$. The available information may be expressed by a pair of random variables: $T_i,\ \delta_i$, where

$$T_i = \max(\min(X_i, Y_i), Z_i) \quad \text{and} \quad \delta_i = \begin{cases} 1 & \text{if} \quad Z_i \leq X_i \leq Y_i \\ 0 & \text{if} \quad X_i > Y_i \\ 2 & \text{if} \quad X_i < Z_i \end{cases} \qquad i = 1, \ldots, n. \tag{1.4}$$

See Chang and Yang (1987).

The log-empirical likelihood for the lifetime distribution F is

$$L(p) = \sum_{\delta_i = 1} \log p_i + \sum_{\delta_i = 0} \log \left(\sum_{T_j > T_i} p_j \right) + \sum_{\delta_i = 2} \log \left(\sum_{T_j < T_i} p_j \right) \,, \tag{1.5}$$

where $p_i \geq 0$ and $\sum p_i = 1$.

With or without mean constraints, the Lagrange multipliers do not simplify the maximization of this empirical likelihood.

2. CONSTRAINED CDF WITH ARBITRARILY CENSORED/TRUNCATED OBSERVATIONS

Before we get into the details, we first describe our (modified) EM computational algorithm in a generic manner.

Given: censored/truncated data, a mean constraint equation for the CDF.

Step 0. (Initialization): Pick an initial CDF F which may or may not satisfy the mean constraint. $F(\cdot)$ must be discrete with support sets as described by Turnbull (1976) and modified by Alioum and Commenges (1996).

Step 1. (E-step): Find the conditional probability with respect to F that an observation is equal to X_i, given the censored/truncated information. This step is the same as the E-step of Turnbull (1976) and will produce pseudo observations X_i and weights w_i.

Step 2. (M-step): With the (X_i, w_i) from the E-step above, find a new CDF estimate F (or equivalently p_i's) by using formulas (A.2) and (A.3) in the Appendix. This new CDF will satisfy the mean constraint.

Step 3. Iterate Steps 1–2 until convergence.

The convergence criterion for the iteration can be based on the values of the log-empirical likelihood, which should increase at each iteration. When the values of the log-empirical likelihood no longer increase, we stop the iteration.

Next we show that the solution of the EM algorithm is equivalent to the constrained maximization of the log-likelihood. We use the same setup and notation as Turnbull (1976).

Suppose X is a random variable whose CDF $F(\cdot)$ is to be estimated. The observations are pairs of sets: $A_i, B_i, i = 1, \ldots, n$ whose relation to the CDF is as follows.

Suppose independent random variables, X_i, are drawn from the conditional distribution functions $P(X \le t | X \in B_i)$. Here, each B_i is called a truncation set. Furthermore, X_i is censored into the set A_i; that is, we only know that $X_i \in A_i$. We suppose the sets A_i, B_i can be written as unions of disjoint intervals. When X_i is observed exactly, then A_i is a single point. When B_i is the entire sample space, then there is no truncation.

The empirical (or nonparametric) likelihood pertaining to the CDF of X, based on the censored and truncated observations (A_i, B_i), is

$$\prod_i \frac{P_F(X \in A_i)}{P_F(X \in B_i)}.$$

See Turnbull (1976, equation (3.6)).

Turnbull (1976) and Alioum and Commenges (1996) identified the nonoverlapping intervals, $[q_j, r_j]$, where the NPMLE of F may have positive probability masses. Pick any point t_j inside $[q_j, r_j]$ to represent the interval. For definiteness let us denote the *midpoints* of those intervals by t_j and the corresponding probability masses by $s_j, j = 1, \ldots, m$.

With the probabilities s_j and obviously defined indicator functions α_{ij}, β_{ij}, we can write $P_F(X \in A_i) = \sum_{j=1}^m \alpha_{ij} s_j$ and so on, and the empirical likelihood as

$$\prod_i \frac{\sum_{j=1}^m \alpha_{ij} s_j}{\sum_{j=1}^m \beta_{ij} s_j}.$$

We seek to maximize the empirical likelihood with respect to the probabilities s_j with an additional mean constraint

$$\sum_{j=1}^m s_j g(t_j) = \mu. \tag{2.1}$$

The EM algorithm, when convergent, will give a solution to Equations (2.2):

$$\pi_j^*(s) = s_j \quad j = 1, 2, \ldots m, \tag{2.2}$$

where $s = (s_1, \ldots, s_m)$ and

$$\pi_j^*(s) = \frac{\sum_{i=1}^n \{\mu_{ij}(s) + \nu_{ij}(s)\}}{M(s) + \lambda(g(t_j) - \mu)} \ . \tag{2.3}$$

In the above equations the function $g(\cdot)$ and constant μ are assumed given (when the constraint is given); λ is the solution of the following equation:

$$0 = \sum_{j=1}^m \frac{(g(t_j) - \mu) \times \sum_{i=1}^n \{\mu_{ij}(s) + \nu_{ij}(s)\}}{M(s) + \lambda(g(t_j) - \mu)} \ . \tag{2.4}$$

The quantities $M(s)$, μ_{ij}, ν_{ij} are defined the same way as in Turnbull (1976); they are

$$M(s) = \sum_i \sum_j (\mu_{ij} + \nu_{ij}) \ ,$$

$$\mu_{ij} = \mu_{ij}(s) = E_s I_{\{X_i \in [q_j, r_j]\}} \ ,$$

$$\nu_{ij} = E_s J_{ij} \ ,$$

where E_s is the expectation with respect to the probability s; J_{ij} is the number of "ghosts" of X_i that have values in $[q_j, r_j]$.

The "ghost" of X_i can be described as follows. Because of truncation, each observation $X_i = x_i$ can be considered a remnant of a group, the size of which is unknown and all (except the one observed) with X-values in B_i^c. They can be thought of as X_i's "ghosts."

Equations (2.3) and (2.4) are the results of Theorem A.1 in the Appendix.

We claim that the constrained maximization can be achieved by iterations based on the self-consistent Equation (2.2). We now show the equivalence claim. To maximize the log-likelihood under the mean constraint (2.1) and $\sum s_j = 1$, we proceed by the Lagrange multiplier. Taking partial derivatives of the target function

$$G = \sum_{i=1}^n \left\{ \log \left(\sum_{j=1}^m \alpha_{ij} s_j \right) - \log \left(\sum_{j=1}^m \beta_{ij} s_j \right) \right\} - \gamma \left(\sum_{j=1}^m s_j - 1 \right) - \lambda \left(\sum_{j=1}^m s_j (g(t_j) - \mu) \right)$$

with respect to the s_j, we get

$$d_j^*(s) = \sum_{i=1}^n \left\{ \frac{\alpha_{ij}}{\sum_{k=1}^m \alpha_{ik} s_k} - \frac{\beta_{ij}}{\sum_{k=1}^m \beta_{ik} s_k} \right\} - \gamma - \lambda(g(t_j) - \mu) \ . \tag{2.5}$$

For s to be the (constrained) NPMLE, those partial derivatives must be zero. Multiplying each of the partial derivatives $d_j^*(s)$ by s_j and then summing over j, we get $\gamma = 0$.

The left side of the self-consistent Equation (2.2) can then be written as

$$\pi_j^*(s) = \frac{s_j}{M(s) + \lambda(g(t_j) - \mu)} \left\{ d_j^*(s) + \lambda(g(t_j) - \mu) + \sum_{i=1}^{n} \left(\sum_{k=1}^{m} \beta_{ik} s_k \right)^{-1} \right\} .$$

Calculations similar to those in Turnbull (1976) give

$$\pi_j^*(s) = \left\{ 1 + \frac{d_j^*(s)}{M(s) + \lambda(g(t_j) - \mu)} \right\} s_j .$$

Therefore, the self-consistent Equation (2.2) becomes

$$\left\{ 1 + \frac{d_j^*(s)}{M(s) + \lambda(g(t_j) - \mu)} \right\} s_j = s_j .$$

From this set of equations we can prove (using arguments similar to those of Turnbull 1976) the following theorem.

Theorem 1. *The solution of the constrained log-likelihood equation $d_j^*(s) = 0$ is equivalent to the solution of the self-consistent Equation (2.2).*

The word "equivalent" in Theorem 1 means that if s_j is a solution to $d_j^*(s) = 0$, then it is also a solution to the equation $\pi_j^*(s) = s_j$, and vice versa.

A good initial F to start the EM calculation is the NPMLE without the constraint (if available). In the case of right-censored data, this is the Kaplan-Meier (1958) estimator. If such an initial F is not available, as in the case of doubly censored data, a distribution with equal probability masses on all possible jump locations t_j can be used.

Remark: The convergence property of this modified EM algorithm is very similar to that of Turnbull (1976), provided we pick the μ value inside the range of $g(t_j)$.

When there is only censoring and no truncation, the negative LELF is clearly convex in s. With a constraint that is linear in s, it is easy to see that there is a unique maximizer, and thus there is a unique solution to the self-consistent equations.

Remark: For two or more independent samples, the maximization of the empirical likelihood with constraints of the type

$$\int g_1(t) dF_1(t) = \int g_2(t) dF_2(t)$$

can be handled similarly by the modified self-consistent/EM algorithm.

3. SOME SPECIAL CASES

This section gives some explicit formulas useful in calculating the E-step for some common types of censored/truncated data.

3.1 RIGHT-, LEFT-, OR DOUBLY CENSORED CASE

E-step: Given F, the weight, w_j, at location t_j can be computed as

$$\sum_{i=1}^{n} E_F\left[I_{[X_i=t_j]}|T_i,\delta_i\right] = w_j\ .$$

We only need to compute the weights for those t_j where either (1) t_j is a jump point for the given distribution F; or (2) t_j is an uncensored observation. In many cases (1) and (2) coincide (e.g., the Kaplan-Meier estimator). The weights for other locations are obviously zero. Also, when T_i is uncensored, the conditional expectation is trivial.

Suppose the ith observation, T_i, is right censored: $\delta_i = 0$. The E-step above can be computed as follows:

$$E_F[I_{[X_i=t_j]}|T_i,\delta_i] = \frac{\Delta F(t_j)}{1-F(T_i)} \quad \text{for} \quad t_j > T_i$$

and $E_F[\cdot] = 0$ for $t_j \le T_i$.

For an uncensored observation T_i, it is obvious that $E_F[\cdot] = 1$ when $t_j = T_i$ and $E_F[\cdot] = 0$ for any other t_j.

For a left-censored observation T_i, the E-step above can be computed as follows:

$$E_F[I_{[X_i=t_j]}|T_i,\delta_i] = \frac{\Delta F(t_j)}{F(T_i)} \quad \text{for} \quad t_j < T_i$$

and $E_F[\cdot] = 0$ for $t_j \ge T_i$.

3.2 LEFT-TRUNCATED AND RIGHT-CENSORED DATA

We describe in some detail the left-truncated and right-censored observation case, because this is a commonly seen data type in practice.

For left truncated observations, there is an explicit expression for the NPMLE of the CDF, the Lynden-Bell estimator (see Li 1995). Li (1995) studied the empirical likelihood when the parameter is the probability $F(T_0)$ for a given T_0. For left-truncated and right-censored observations, there is also an explicit NPMLE of the CDF (Tsai, Jewell, and Wang 1987). However, no explicit formula exists to compute the NPMLE under the mean constraint, so we need the EM algorithm described here.

Suppose the observations we have are $(Y_1, T_1, \delta_1), \ldots, (Y_n, T_n, \delta_n)$, where the Y's are the left truncation times and the T's are the (possibly right-censored) lifetimes. Denote by X the lifetime before censoring/truncation. The censoring indicator δ assumes the usual meaning that $\delta = 1$ if T is uncensored and $\delta = 0$ if T is right censored. Because of truncation, for all i, we have $(T_i > Y_i)$. We assume that Y is independent of X and both distributions are unknown.

The NPMLE of the CDF puts positive probabilities only at the locations of observed, uncensored T_i's. Denote those locations by t_j. The log-empirical likelihood pertaining to

the distribution of X is (see, e.g., Tsai, Jewell, and Wang 1987)

$$L(p) = \sum_{i:\delta_i=1} \left(\log p_i - \log \left(\sum_{T_j > Y_i} p_j \right) \right) + \sum_{i:\delta_i=0} \left(\log \left(\sum_{T_j > T_i} p_j \right) - \log \left(\sum_{T_j > Y_i} p_j \right) \right).$$

E-step: Given a current estimate $F(\cdot)$ that has positive probabilities only at t_j, we compute the weight

$$w_j = \sum_{i=1}^{n} E_F[I_{[X_i=t_j]}|T_i, \delta_i] + \sum_{i=1}^{n} I_{[t_j<Y_i]} \Delta F(t_j)/P_F(X > Y_i) \, .$$

The above can be written more explicitly by noticing that (1) the E_F part can be computed in exactly the same way as in the example for the censored data case, and (2) the second term is (without summation)

$$I_{[t_j<Y_i]} \Delta F(t_j)/P_F(X > Y_i) = \frac{I_{[t_j<Y_i]} p_j}{\sum_k I_{[t_k>Y_i]} p_k} \, ,$$

where $p_j = \Delta F(t_j)$.

4. SIMULATIONS AND EXAMPLES OF EMPIRICAL LIKELIHOOD RATIO COMPUTATION

4.1 COMPUTATION OF EMPIRICAL LIKELIHOOD RATIO

Once the constrained NPMLE of probabilities are computed by the EM algorithm, we can plug them into the log-likelihood, as in (1.2) or (1.5), to get the censored/truncated log-empirical likelihood with mean constraint easily. This in turn allows us to compute the empirical log-likelihood ratio statistic:

$$\begin{aligned} -2 \log R(H_0) &= -2 \log \frac{\max_{H_0} L(p)}{\max_{H_0+H_1} L(p)} && (4.1) \\ &= 2 \left[\log(\max_{H_0+H_1} L(p)) - \log(\max_{H_0} L(p)) \right] \\ &= 2 \left[\log(L(\tilde{p})) - \log(L(\hat{p})) \right] \, . && (4.2) \end{aligned}$$

Here, $\tilde{p}$ is the NPMLE of probabilities without a mean constraint; $\hat{p}$ is the NPMLE of probabilities under the mean constraint of the null hypothesis.

In some cases, there may be faster methods available to compute $\tilde{p}$. A case in point is the Kaplan-Meier estimator for right-censored data. If not, we can always use Turnbull's (1976) EM algorithm to compute $\tilde{p}$.

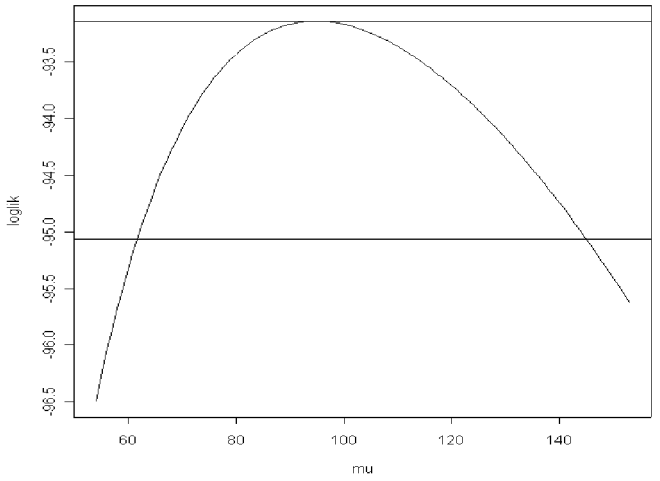

Figure 1. Log-likelihood for μ near maximum.

Examples with real data and simulation results are reported below to illustrate the usefulness of the proposed EM method and also to illustrate the small to medium sample performance of the chi square approximation. The software used are the R functions `el.cen.EM()` and `el.ltrc.EM()` from the `emplik` package.

4.2 CONFIDENCE INTERVAL, REAL DATA, RIGHT CENSORED

The first example concerns Veteran's Administration lung cancer study data (e.g., available from the R package `survival`). We took the subset of survival times from treatment 1 and the small cell group. There are two right-censored observations. The survival times are: 30, 384, 4, 54, 13, 123+, 97+, 153, 59, 117, 16, 151, 22, 56, 21, 18, 139, 20, 31, 52, 287, 18, 51, 122, 27, 54, 7, 63, 392, 10.

We use the EM algorithm to compute the log-empirical likelihood under the constraint mean$(F) = \mu$ for various values of μ. The log-empirical likelihood has a maximum when $\mu = 94.7926$, which is the mean computed from the Kaplan-Meier estimator.

The 95% confidence interval for the mean survival time is $\{\mu | -2 \log R(\mu) < \chi^2_1(.95)\}$, which is seen here to be $[61.70948, 144.912]$ (Figure 1) since the log-empirical likelihood was $3.841/2 = \chi^2_1(0.95)/2$ below the maximum value $(= -93.14169)$ both when $\mu = 61.70948$ and $\mu = 144.912$. We see that the confidence interval is not symmetric around the MLE. In general, confidence intervals obtained by inverting empirical likelihood ratio tests are not necessarily symmetric, which can improve the coverage for skewed data compared to the Wald type confidence intervals.

4.3 SIMULATION: RIGHT-CENSORED DATA

It is generally believed that, for smaller sample sizes, likelihood ratio/chi square based inference is more accurate than those obtained by Wald method. The Q-Q plot (Figure

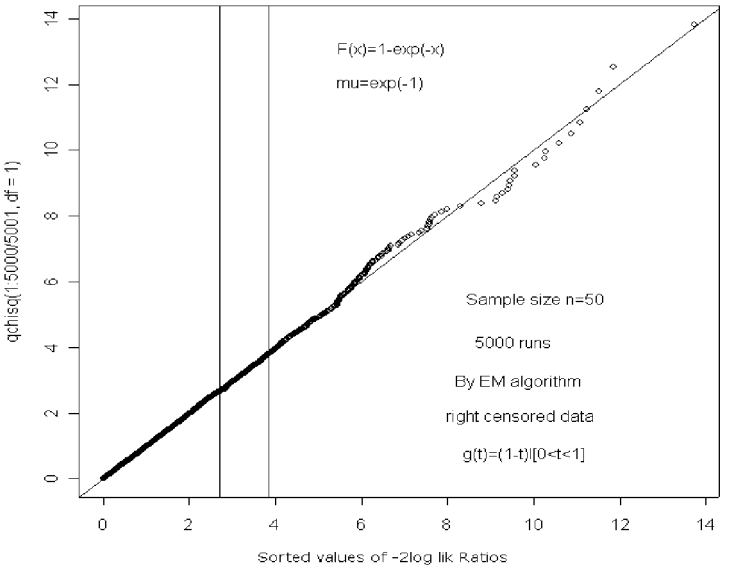

Figure 2. A Q-Q plot for right-censored likelihood ratio.

2) from the following simulation shows that the chi-square distribution is a pretty good approximation to the distribution of the -2 log-empirical likelihood ratio statistic for right-censored data involving a mean parameter.

We randomly generated 5,000 right-censored samples, each of size $n = 50$ as in Equation (1.1), where $X \sim \exp(1)$, $C \sim \exp(.2)$, and $g(t) = I_{[t\le 1]}$ or $g(t) = (1-t)I_{[t\le 1]}$. The constraint is $\int_0^1 g(t)d(1-\exp(-t)) = \mu$. Both plots look similar; we only show here the one corresponding to $g(t) = (1-t)I_{[t\le 1]}$.

We computed 5,000 empirical likelihood ratios. Each ratio is obtained as in (4.2) by using the Kaplan-Meier estimator's jumps as $(\tilde{p})$, and $(\hat{p})$ given by the modified EM method proposed in this article. The Q-Q plot, based on 5,000 empirical likelihood ratios and χ_1^2 percentiles, is shown in Figure 2. Two vertical lines were drawn at the points 3.84 and 2.71, which are the critical values of χ_1^2 with nominal levels 5% and 10%. From the Q-Q plot, we can see that the χ_1^2 approximation is pretty good because the sorted -2log-likelihood ratios line up closely to χ_1^2 percentiles. Only at the far upper tail of the distributions, the differences are visible. This implies that, for confidence intervals with confidence levels up to 98% or 99%, the approximate coverage probabilities are pretty close to the nominal; when confidence levels are over 99.7%, then the approximations are not as good.

4.4 SIMULATION AND EXAMPLE: LEFT-TRUNCATED, RIGHT-CENSORED CASE

We generate left truncation times Y as shifted exponential random variables, $\exp(4) - .1$. We generate lifetimes X as $\exp(1)$ random variables and censoring times C as $\exp(.15)$ random variables. The truncation probability $P(Y > X)$ is around 13.4%. The censoring

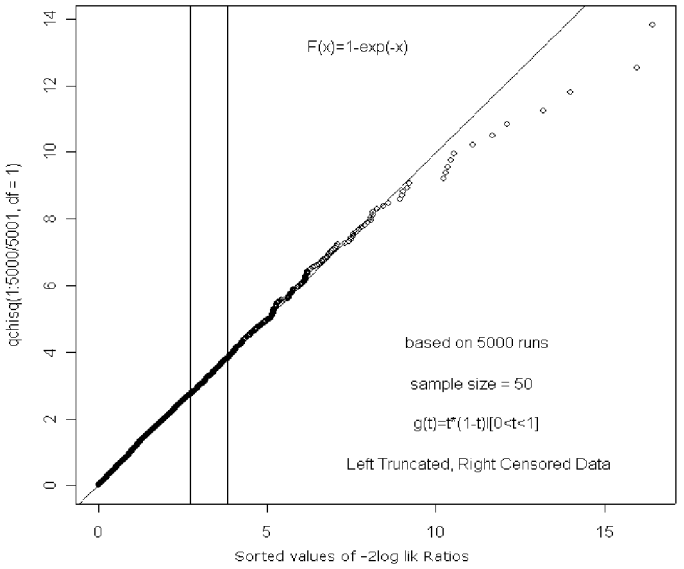

Figure 3. Q-Q plot of -2log-likelihood ratios vs. $\chi^2_{(1)}$ percentiles for sample size 50.

probability $P(X > C)$ is around 13%.

The triplets, (Y, X, C), are rejected unless $Y < X$ and $Y < C$. If the triplet is not rejected, we return Y, $T = \min(X, C)$ and $\delta = I_{[X \leq C]}$. Fifty triplets $\{Y,\ T,\ \delta\}$ are generated each time a simulation is run. The function $g(t)$ we used is $t(1-t)I_{[0<t<1]}$. The mean of this function is $\mu = (3e^{-1} - 1)$. Figure 3 shows the Q-Q plot for this simulation. Again, we see that the chi-square distribution is a very good approximation to the distribution of the empirical likelihood ratio. Only in the far upper tail the quality of the approximation deteriorates.

Finally, let us look at a small dataset taken from the book of Klein and Moeschberger (1997, tab. 1.7, p. 16). There, the survival times of female psychiatric inpatients are reported as follows: Y =(51, 58, 55, 28, 25, 48, 47, 25, 31, 30, 33, 43, 45, 35, 36), T =(52, 59, 57, 50, 57, 59, 61, 61, 62, 67, 68, 69, 69, 70, 76), and δ =(1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1). The mean computed from the Tsai-Jewell-Wang (1987) estimate is 64.4375. The plot of -2 log-likelihood ratio against various μ values is similar to Figure 1 and is omitted. The -2 log-likelihood ratio has a minimum of zero at $\mu = 64.4375$, as it should be. A 95% confidence interval for μ includes those values for which the -2log-likelihood ratio is less than 3.84. In this case, it is $[59.5702302, 69.1899978]$.

5. DISCUSSION

The computational algorithm proposed in this article covers a wide variety of censored/truncated data cases as described by Turnbull (1976). It enables us to compute the

NPMLE of the CDF under a mean type constraint. This in turn enables us to compute the -2 log-empirical likelihood ratio in such cases. Coupled with the empirical likelihood theory (Wilks theorem), the ratio can be used to draw inference about functionals of the CDF, and we see good small to medium sample performance in the examples/simulations.

Multivariate versions of the proposed EM algorithm and Theorem 1 are obviously possible. There, we have k constraints: $\sum_i s_i g_j(t_i) = \mu_j, j = 1, \ldots, k$. In fact, this is also implemented in the `emplik` package as function `el.cen.EM2`.

Another method often used to search for a maximum is the Newton type method. For maximizing censored/truncated empirical likelihoods, a Lagrange multiplier reduction is often not available, and the Newton type method has to work with $n - 1$ variables, which grows at the rate of sample size n. Moreover, the Newton method involves matrices of size n by n. Inverting such matrices makes things worse. See Chen and Zhou (2001) for more details. On the other hand, the memory requirement of the EM method is linear in n.

On a desktop PC with 3.06 GHz CPU, 512 MB RAM, we recorded the following times:

- Sample size 2,000 (25% right censored). EM: 2 seconds; Newton: 20 seconds.
- Sample size 4,000 (25% right censored). EM: 5 seconds; Newton: more than 5 minutes.
- On a notebook computer (Celeron 2.2 GHz CPU, 512 MB RAM), we tested the EM method with even larger sample sizes:
  - Sample size 10,000 (25% right censored). EM: 55 seconds.
  - Sample size 20,000 (25% right censored). EM: 4 minutes.

The theory for the asymptotic properties of the constrained NPMLE and the empirical likelihood ratio lags behind the computation. There has yet to be a theory of empirical likelihood ratio that covers all the censored/truncated data cases described by Turnbull (1976), but see Owen (2001), Murphy and van der Vaart (1997), and Banerjee and Wellner (2001) for some known special cases.

APPENDIX: CONSTRAINED MAXIMIZATION OF THE EMPIRICAL LIKELIHOOD WITH UNCENSORED, WEIGHTED OBSERVATIONS

Suppose we have independent (uncensored, not truncated) observations $X_1, \ldots, X_n$ from distribution $F(\cdot)$. Associated with the observations are nonnegative weights $w_1, \ldots, w_n$. The meaning of the weights is such that if $w_i = 2$, then there are two observations with value X_i, and so on. We shall allow the weights to be fractions for our applications later.

The empirical likelihood based on the weighted observations is $\prod (p_i)^{w_i}$, and the log-empirical likelihood is

$$\sum w_i \log p_i \, . \tag{A.1}$$

Theorem A.1. *Suppose $g(\cdot)$ is a given function that satisfies $max_i g(X_i) > \mu > min_i g(X_i)$. The maximization of the log-empirical likelihood (A.1) with respect to the p_i and subject to the constraints*

$$\sum p_i = 1\,, \qquad \sum g(X_i)p_i = \mu$$

is given by the formula

$$p_i = \frac{w_i}{\sum_j w_j + \lambda(g(X_i) - \mu)}\,, \tag{A.2}$$

where λ is the solution of the equation

$$\sum_i \frac{w_i(g(X_i) - \mu)}{\sum_j w_j + \lambda(g(X_i) - \mu)} = 0\,. \tag{A.3}$$

The solution of (A.3) is unique and the p_i's given above constitute a proper probability.

A multivariate version of this theorem also holds, in which we have k constraints $\sum_i g_j(X_i)p_i = \mu_j, j = 1, \ldots, k$. In this case, g, λ and μ in the above formula are understood to be vectors of length k.

This theorem can be easily proved by using the Lagrange multiplier and much of it is contained in the proof of Theorem 1 of Owen (1990). We omit the details.

[Received October 2002. Revised September 2004.]

REFERENCES

Alioum, A., and Commenges, D. (1996), "A Proportional Hazards Model for Arbitrarily Censored and Truncated Data," *Biometrics*, 52, 512–524.

Banerjee, M., and Wellner J. A. (2001), "Likelihood Ratio Tests for Monotone Functions," *The Annals of Statistics*, 29, 1699–1731.

Chang, M. N., and Yang, G. L. (1987), "Strong Consistency of a Non-parametric Estimator of the Survival Function with Doubly Censored Data," *The Annals of Statistics*, 15, 1536–1547.

Chen, K., and Zhou, M. (2001), "Computing Censored Empirical Likelihood by Sequential Quadratic Programing," Technical Report, Department of Statistics, University of Kentucky.

Gentleman, R., and Ihaka, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Kaplan, E., and Meier, P. (1958), "Non-parametric Estimator From Incomplete Observations," *Journal of American Statistical Association*, 53, 457–481.

Klein, J. P., and Moeschberger, M. L. (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer.

Li, G. (1995), "Nonparametric Likelihood Ratio Estimation of Probabilities for Truncated Data," *Journal of the American Statistical Association*, 90, 997–1003.

Murphy, S., and van der Vaart, A. (1997), "Semiparametric Likelihood Ratio Inference," *The Annals of Statistics*, 25, 1471–1509.

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75 237–249.

——— (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18, 90–120.

——— (2001), *Empirical Likelihood*, London: Chapman & Hall.

Pan, X. R., and Zhou, M. (1999), "Using One Parameter Sub-family of Distributions in Empirical Likelihood with Censored Data," *Journal of Statistical Planning and Inference*, 75, 379–392.

Thomas, D. R., and Grunkemeier, G. L. (1975), "Confidence Interval Estimation of Survival Probabilities for Censored Data," *Journal of the American Statistical Association*, 70, 865–871.

Tsai, W.-Y., Jewell, N. P., and Wang, M.-C. (1987), "The Product Limit Estimate of a Survival Curve Under Right Censoring and Left Truncation," *Biometrika*, 74, 883–886.

Turnbull, B. (1976), "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society*, Ser. B, 290–295.