Statistica Sinica Preprint No: SS-2015-0410.R1	
Title	Comparison of Extended Empirical Likelihood Methods: Size and Shape of Test Based Confidence Regions
Manuscript ID	SS-2015-0410.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0410
Complete List of Authors	Mi-Ok Kim and Mai Zhou
Corresponding Author	Mi-Ok Kim
E-mail	miok.kim@ucsf.edu, irismiokkim@gmail.com
Notice: Accepted version subject to English editing.	

Statistica Sinica

COMPARISON OF EXTENDED EMPIRICAL LIKELIHOOD METHODS: SIZE AND SHAPE OF TEST BASED CONFIDENCE REGIONS

Mi-Ok Kim and Mai Zhou

Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco CA 94143, U.S.A.
Department of Statistics, University of Kentucky, Lexington KY, 40536-0082, U.S.A.

Abstract: Empirical likelihood is a general non-parametric inference methodology. It uses likelihood principle in a way that is analogous to that of parametric likelihoods. In a wide range of applications the methodology was shown to provide likelihood ratio statistics that have limiting chi-square distributions and observe a non-parametric version of Wilks theorem. Amongst recent extensions of the empirical likelihood are the analysis of censored data, longitudinal data and semi-parametric regression model. However, this property of Wilks theorem only remained true in some but not in others. This motivates our discussion of relative optimality of extended empirical likelihood methods. We compare extended empirical likelihood methods and evaluate their relative optimality by comparing the confidence regions provided by inverting the likelihood ratio tests. We show that those extension methods with its likelihood ratio statistic observing the Wilks theorem provides the smallest confidence region. Specific examples are provided for the case of censored data analysis and estimating equations involving nuisance parameters.

Key words and phrases: Empirical Likelihood, Likelihood Ratio, Wilks Confidence Region, Scaled Chi Square Distribution.

1. Introduction

Empirical likelihood is a general non-parametric inference method using likelihood principle in a way that is analogous to that of parametric likelihoods. Since its introduction by Owen (1988, 1990), the methodology has been extended to the analysis of censored data (Li, Li & Zhou (2005)), longitudinal data (e.g., Wang, Qian & Carroll (2010)) and semi-parametric regression model (e.g., Shi & Lau

(2000)). In a wide range of applications as outlined in Owen (2001) empirical likelihood ratio statistics were shown to have limiting chi-square distributions, that is, observe a Wilks theorem. This property remains true with some recent extended methods but not with others, however. This motivates our discussion of relative optimality of extended methods. Specific examples of those extended empirical likelihoods are described in section 3.

In this paper we focus our attention on extensions of empirical likelihood method that have the likelihood function maximized at the same location, thus yielding equivalent maximum empirical likelihood estimators, but provide different likelihood ratio statistics. We discuss the relative optimality of the methods by comparing the confidence regions provided by inverting the likelihood ratio tests. We show that the extended empirical likelihood that yields the likelihood ratio statistic observing the Wilks theorem provides the smallest confidence region.

The remainder of this paper proceeds as follows. Section 2 formally defines a set-up where multiple empirical likelihood methods can be compared. The section also provides asymptotic results for the comparison of the corresponding confidence regions. Section 3 describes specific examples for the cases of censored data analysis and estimating equations involving nuisance parameters. Section 4 provides empirical results. Section 5 concludes. All the proofs are deferred to the Appendix.

2. Comparison of Various Empirical Likelihoods via Its Confidence Regions

Specifying extended empirical likelihood requires details. For example, various ways of constructing the likelihoods are available for censored data and accordingly different extended extended empirical likelihoods are specified. We defer presentation of specific examples including censored data case to section 3 and provide general results in this section.

Given a random sample $\mathcal{Z}_n = \{Z_i\}_{i=1}^n$ with a common cumulative distribution function F_0 and $Z_1 \in R^k$, we consider an (extended) empirical likelihood defined for some discrete cumulative distribution function F (denoted by $EL(\mathcal{Z}_n, F)$). We subsequently define the maximizer $F_n = argmax_F \log EL(\mathcal{Z}_n, F)$. By the theory of the methodology, the empirical likelihood gives rise to the em-

pirical likelihood ratio test for a finite parameter $\theta = T(F) \in R^p$ with the test statistic

$$-2\log ELR_n(\theta) = -2[\max_{F\in\mathcal{F}} \log EL(\mathcal{Z}_n, F) - \log EL(\mathcal{Z}_n, F_n)] \ ,$$

where $\mathcal{F}$ is a set of discrete cumulative distribution functions that satisfy the constraint $T(F) = \theta$. The true value of the parameter is given as $\theta_0 = T(F_0)$, and the maximum empirical likelihood estimator is given as $\hat{\theta}_n = T(F_n)$.

We invert the likelihood ratio test and obtain a confidence region $C_n = \{\,\theta|\, -2\log ELR_n(\theta) \le c\}$, where the critical value c is chosen according to the (asymptotic) distribution of $-2\log ELR_n(\theta)$ under null hypothesis for some desired confidence level. The asymptotic distribution under null hypothesis is given by the following results: under some regularity conditions

$$\begin{aligned} n^{1/2}(\hat{\theta}_n - \theta_0) &\to U \qquad (2.1)\\ -2\log ELR_n(\theta_0) &= n(\hat{\theta}_n - \theta_0)^\top V^{-1}(\hat{\theta}_n - \theta_0) + o_p(1)\,, \end{aligned}$$

where U is a $p-$variate random variable with $U \sim N_p(0, W)$, and W and V are $p \times p$ (semi-)positive definite matrices. Without loss of generality we consider two methods that provide the same maximum empirical likelihood estimators, $\hat{\theta}_{1n} = \hat{\theta}_{2n}$, but different likelihood ratio statistics $ELR_{1n}(\theta)$ and $ELR_{2n}(\theta)$. We suppose that the asymptotic results in (2.1) hold with $V = V_1$ and $V = V_2$ with $WV_1^{-1} = I$ and $WV_2^{-1} \neq I$ for $ELR_{1n}(\theta)$ and $ELR_{2n}(\theta)$ respectively. By the results of Hjort, McKeague & Keilegom (2009) and Qin & Lawless (1994), the empirical likelihood with $V = V_1$ admits a chi-square with p degrees of freedom as the limiting distribution of $-2\log ELR_{1n}(\theta)$, whereas the empirical likelihood with $V = V_2$ admits a limiting distribution characterized by weighted sum of p independent chi-squares of degree of freedom 1. This distribution is often called "the scaled chi square distribution", especially in the case of df=1 (Wang & Jing (2001), Wang & Li (2002), Ren (2008)).

We consider respectively defined confidence regions with the two methods at the same confidence level as follows:

$$C_{1n} = \{\,\theta|\, -2\log ELR_{1n}(\theta) \le c_1\}\,, \qquad C_{2n} = \{\,\theta|\, -2\log ELR_{2n}(\theta) \le c_2\}\,,$$

where the critical values c_1 and c_2 are appropriately chosen by the limiting distributions. C_{1n} and C_{2n} are centered at the same value in the sense that

$-2\log ELR_{1n}(\hat{\theta}_n) = -2\log ELR_{2n}(\hat{\theta}_n) = 0$ where $\hat{\theta}_n$ denotes the common maximum empirical likelihood estimator. The following theorem shows the asymptotic relative optimality of $C_{1n}(\theta)$.

Theorem 1. *Assume regularity conditions under which the results in (2.1) holds for $ELR_{1n}(\theta)$ and $ELR_{2n}(\theta)$ respectively. Given a confidence level $1-\alpha_n \in (0,1)$ with $\alpha_n \to 0$, there exist N such that for $n > N$ we have*

$$Vol(C_{1n})/Vol(C_{2n}) < 1\,,$$

where $Vol(C)$ for any measurable set C in R^p is defined as $Vol(C) = \int_{R^p} I_{[t\in C]}\mu(dt)$ with μ the Lebesgue measure in R^p.

Theorem 1 implies that the empirical likelihood method that provides the likelihood ratio statistic which is only incompletely self-studentizing is not optimal.

Remark 1. In several papers (e.g., Wang & Jing (2001), Wang & Li (2002)), it is suggested that one may *adjust* the empirical likelihood ratio of type $-2\log ELR_{2n}$ by multiplying a factor, that will turn the asymptotic null distribution into a regular chi square, and may even improve the performance of the resulting confidence regions. While the adjustment may improve the level of confidence of the resulting confidence region, it is clear that if the multiplying factor is (asymptotically) a constant and does not depend on θ, then the resulting confidence regions are the same as before, in the sense that the shape and orientation of the resulting confidence region is the same as before. In our examples later, we set the confidence level by simulating the null distribution, eliminating the need of the adjustment. To achieve optimality, the multiplying factor must depend on θ and be the value $\log ELR_{1n}(\theta)/\log ELR_{2n}(\theta)$ $= (\hat{\theta}-\theta_0)^\top V_1^{-1}(\hat{\theta}-\theta_0)/(\hat{\theta}-\theta_0)^\top V_2^{-1}(\hat{\theta}-\theta_0) + o_p(1)$. Obviously, computing this factor is equivalent to computing the statistic $-2\log ELR_{1n}$, if not more difficult.

3. Extended Empirical Likelihoods

In this section we present concrete examples of extended methods that provide likelihood ratio statistics that do not observe the non-parametric version of the Wilks theorem. In each of these examples the construction of the likelihood does not match the structure of data. The ill-constructed likelihoods provide the

likelihood ratio statistics that are only incompletely self-studentizing. For contrast, we also attempt to give a properly defined empirical likelihood that does observe the Wilks theorem.

3.1. Analysis of Estimating Equations Involving Nuisance Parameters

We suppose that the unknown parameters $\theta \in R^p$ is related to the common cumulative distribution function F_0 via a $p-$variate vector of estimating functions involving finite dimensional nuisance parameters $\psi \in R^q$, $m(\cdot, \theta, \psi)$, so $E\{m(Z_1, \theta_0, \psi_0)\} = 0$ for the true values θ_0 and ψ_0. We also suppose a $q-$variate vector of estimating functions $h(\cdot, \theta, \psi)$ exists that defines the true parameter values θ_0 and ψ_0 uniquely as a solution to the following estimating equations jointly: $E\{m(Z_1, \theta_0, \psi_0)\} = 0\,,\ \ E\{h(Z_1, \theta_0, \psi_0)\} = 0\,$. We consider the following empirical likelihood jointly for θ and ψ:

$$EL_n(\theta, \psi) = \max\left\{ \prod_{i=1}^n w_i \,\middle|\, \sum_{i=1}^n w_i \begin{pmatrix} m(Z_i, \theta, \psi) \\ h(Z_i, \theta, \psi) \end{pmatrix} = 0\,, \sum_{i=1}^n w_i = 1\,, 0 \le w_i \right\},$$

where w_i denotes the point mass assigned to the $i-$th observation Z_i. Without the constraints this likelihood is maximized by $w_i = n^{-1}$, $i = 1, \ldots, n$, and the maximum empirical likelihood estimator $(\hat{\theta}_n, \hat{\psi}_n)$ is given by the solution to the following equations: $n^{-1}\sum_{i=1}^n m(Z_i, \theta, \psi) = 0$ and $n^{-1}\sum_{i=1}^n h(Z_i, \theta, \psi) = 0\,$.

Two methods have been proposed for the inference of θ alone. Qin & Lawless (1994) used the profile likelihood $EL_{1n}(\theta) = \max_\psi EL_n(\theta, \psi)\,$. The others used the so-called plug-in method (Hjort, McKeague & Keilegom (2009)) with the following likelihood:

$$EL_{2n}(\theta) = \max\left\{ \prod_{i=1}^n w_i \,\middle|\, \sum_{i=1}^n w_i m(Z_i, \theta, \tilde{\psi}) = 0\,, \sum_{i=1}^n w_i = 1\,, 0 \le w_i \right\},$$

where $\tilde{\psi}$ is a $\sqrt{n}-$consistent estimator, so that $E\{m(Z_i, \theta_0, \tilde{\psi})\} = 0$. Without externally given $\tilde{\psi}$ available, a common choice is the maximum empirical likelihood estimator $\hat{\psi}_n$. With this particular choice, the plug-in empirical likelihood has $\hat{\theta}_n$ as the maximum empirical likelihood estimator and $\max_\theta EL_{2n}(\theta) = \prod_{i=1}^n n^{-1}$, same as the profile likelihood method. The likelihood ratio statistics are given

respectively:

$$ELR_{1n}(\theta) = EL_{1n}(\theta)/\prod_{i=1}^{n} n^{-1}\,, \quad ELR_{2n}(\theta) = EL_{2n}(\theta)/\prod_{i=1}^{n} n^{-1}\,. \tag{3.1}$$

The following theorem shows that results in (2.1) hold for the profile and the plug-in method.

Theorem 2. *Assume conditions under which the below results hold:*

$$n^{1/2}\begin{pmatrix}\hat{\theta}_n - \theta_0 \\ \hat{\psi}_n - \psi_0\end{pmatrix} \longrightarrow N_p(0,\Sigma)\,,$$

in distribution, where $\Sigma = \{S_{12}S_{11}^{-1}S_{21}\}^{-1}$ *with* $S_{12} = S_{21}^{\top} = E\begin{pmatrix}\partial m/\partial\theta & \partial m/\partial\psi \\ \partial h/\partial\theta & \partial h/\partial\psi\end{pmatrix}$, $S_{11} = E\begin{pmatrix}mm^{\top} & mh^{\top} \\ hm^{\top} & hh^{\top}\end{pmatrix}$, *and the expectations taken at* $\theta = \theta_0$ *and* $\psi = \psi_0$. *Also assume that for any* $\tilde{\psi}$ *with* $||\tilde{\psi}-\psi_0|| = O_p(n^{-1/2})$, $n^{-1}\sum_{i=1}^{n} m(Z_i,\theta_0,\tilde{\psi})m^{\top}(Z_i,\theta_0,\tilde{\psi}) \to E(mm^{\top})$. *Suppose* $\Sigma_{(jk)}$, $j,k = 1,2$, *denotes the blocks of* Σ *that correspond to* θ *and* ψ. *Then, the results in (2.1) hold for both* $-2\log ELR_{1n}(\theta_0)$ *and* $-2\log ELR_{2n}(\theta_0)$ *with*

$$U \sim N_p(0,\Sigma_{(11)})\,,\ W = V_1 = \Sigma_{(11)}\,,\ V_2 = \Big[E(\partial m/\partial\theta)^{\top}\{E(mm^{\top})\}^{-1}E(\partial m/\partial\theta)\Big]^{-1}\,.$$

Specifically, $-2\log ELR_{2n}(\theta_0) \longrightarrow \sum_{j=1}^{p} c_j\xi_j$ *in distribution, where* ξ_j *are independent chi-squared distributed random variables with degree of freedom 1 and* $c_1,\ldots,c_p$ *are eigenvalues of* V_2W^{-1}.

The weighted χ^2 asymptotic results stems from the ill-construction of the likelihood: the likelihood of $EL_{2n}(\theta)$ takes the same form as that of Owen's likelihood for the case of independent data, whereas $m(Z_i,\theta,\hat{\psi}_n)$, $i = 1,\ldots,n$, are not independent anymore as they all involve the same $\hat{\psi}_n$.

3.2. Censored Data Analysis

We suppose that $\{Z_i\}_{i=1}^n = \{(T_i,\delta_i)\}_{i=1}^n$ where (T_i,δ_i) denote right censored observations from lifetime variables $\{Y_i\}$ with a common cumulative distribution function F_0 subject to random censoring as follows:

$$T_i = \min(Y_i,C_i) \quad \text{and} \quad \delta_i = I_{[Y_i\le C_i]} \quad \text{for} \ \ i = 1,\ldots,n\,.$$

The censoring variable C_i are assumed independent of the Y_i with a cumulative distribution function G_0. We suppose the true value θ_0 of an unknown parameter $\theta \in R^p$ is related to F_0 via estimating equations such that

$$E\{g(Y_i)\} = \theta_0\,,$$

where $g = (g_1, \ldots, g_p)$ are p functions. We examine three empirical likelihood methods that have been proposed in literature for the inference of θ.

Censored Likelihood Method: The first method uses the empirical likelihood defined for the censored data. Kaplan & Meier (1958), Owen (2001) and others have defined the empirical likelihood of the censored data $\{Z_i\}_{i=1}^n$ for some cumulative distribution function F as

$$EL(\mathcal{Z}_n, F) \;=\; \prod_{i=1}^{n} \{\Delta F(T_i)\}^{\delta_i} \{ \sum_{j:T_j>T_i} \Delta F(T_j)\}^{1-\delta_i}$$

where $\Delta F(t) = F(t+) - F(t-)$ is the jump of F at t or a probability mass assigned to the point t. As in the case of uncensored case, the definition *assumes* a discrete $F(\cdot)$ with possible jumps only at the observed times. With $w_i = \Delta F(T_i)$, $i = 1, \ldots, n$, the likelihood above can be written in term of the jumps and the log likelihood is

$$\log EL_{1n}(F) = \sum_{i=1}^{n} \left\{ \delta_i \log w_i + (1-\delta_i) \log \sum_{j=1}^{n} w_j I_{[T_j>T_i]} \right\}\,, \qquad (3.2)$$

where $I_{[\cdot]}$ denotes an indicator function. We note that the likelihood is maximized at the well known Kaplan-Meier estimator (Kaplan & Meier (1958)) with $w_i = \Delta \hat{F}_{KM}(T_i)$, $i = 1, \ldots, n$. Using the empirical likelihood of the censored data in (3.2) Zhou (2011) proposed the empirical likelihood for θ as

$$EL_{1n}(\theta) = \max_{w_i} \left\{ \prod_{i=1}^{n} w_i^{\delta_i} \left(\sum_{j=1}^{n} w_j I_{[T_j>T_i]} \right)^{1-\delta_i} \middle| \sum_{i=1}^{n} w_i g(T_i) = \theta\,,\; \sum_{i=1}^{n} w_i = 1\,, 0 \le w_i \right\}\,.$$

The maximum empirical likelihood estimator $\hat{\theta}_n$ is given by the equation,

$$\sum_{i=1}^{n} g(T_i) \Delta \hat{F}_{KM}(T_i) = \hat{\theta}_n\,. \qquad (3.3)$$

The empirical likelihood ratio statistic is accordingly given as

$$-2\log ELR_{1n}(\theta) = -2\log\{EL_{1n}(\theta)/EL_{1n}(\hat{\theta}_n)\}.$$

Synthetic data and uncensored empirical likelihood method: Recall the parameter θ is defined as $Eg(Y_i) = \theta$. The second method was motivated by the following lemma, which can be proved easily.

Lemma 1. *For any function $g(\cdot)$ such that $Eg(Y_i)$ is well defined, we consider $X_i = \{g(T_i)\delta_i\}/\{1-G_0(T_i)\}$, where $G_0(\cdot)$ is the cumulative distribution function of censoring variable C_i. Then, we have $Eg(Y_i) = E(X_i)$.*

If G_0 were known, $\{X_i\}_{i=1}^n$ or the synthetic data would be completely observed and the following empirical likelihood could be considered:

$$EL_{2n}(\theta) = \max_{w_i}\left\{\prod_{i=1}^n w_i \,\middle|\, \sum_{i=1}^n w_i X_i = \theta\,,\ \sum_{i=1}^n w_i = 1\,, 0 \le w_i\right\}.$$

The first and second methods utilize different likelihoods. $EL_{1n}(\theta)$ is based on the likelihood of censored data $\{(T_i,\delta_i)\}_{i=1}^n$ with w_i denoting the point mass of a discrete cumulative function of Y_i. $EL_{2n}(\theta)$ is based on the likelihood of completely observed $\{X_i\}_{i=1}^n$ with w_i denoting the point mass of a cumulative distribution function of X_i, assuming the knowledge of the censoring time distribution G_0.

However, usually $G_0(t)$ or $1-G_0(t)$ is not known, and X_i are not available. An alternative method based on the synthetic data uses an estimator $\hat{G}(t)$, often the Kaplan-Meier estimator $\hat{G}_{KM}$, instead and defines the likelihood as

$$EL^*_{2n}(\theta) = \max_{w_i}\left\{\prod_{i=1}^n w_i \,\middle|\, \sum_{i=1}^n w_i X^*_i = \theta\,,\ \sum_{i=1}^n w_i = 1\,, 0 \le w_i\right\},$$

where $X^*_i = \{g(T_i)\delta_i\}/\{1-\hat{G}_{KM}(T_i)\}$. The replacement, however, creates a problem that X^*_i are not independent since all of them involve $\hat{G}_{KM}(T_i)$.

Since $\hat{G}_{KM}(t)$ is uniformly consistent for $G_0(t)$, the mean of the (no longer independent) sequence $\{X^*_i\}$ is still asymptotically same as $Eg(Y)$, but the variance of the sequence $\{X^*_i\}$ is different from the variance of X_i. To be precise, under some regularity conditions, we can show that as $n\to\infty$, $\sqrt{n}(\bar{X}-\theta_0)\longrightarrow N(0,\sigma_1^2)$

and $\sqrt{n}(\bar{X}^* - \theta_0) \longrightarrow N(0, \sigma_2^2)$ in distribution with $\sigma_1^2 > \sigma_2^2$, if $g(\cdot)$ is a scalar function ($p = 1$). This result similarly holds in a multivariate case ($p > 1$). We refer to Srinivasan & Zhou (1994) for the calculation and the expression of the asymptotic variances or variance-covariance matrix. Like the plug-in method, the second synthetic data method uses the empirical likelihood method developed for independent identical sequences, but applies it to the sequence $\{X_i^*\}$ that is not independent and identical.

Interestingly, if we use the Kaplan-Meier estimator $\hat{G}_{KM}(t)$ in the above, the empirical likelihood $EL_{2n}^*(\theta)$ is maximized when θ takes the value $\hat{\theta}_n$ defined in (3.3), as $\Delta\hat{F}_{KM}(T_i) = \delta_i / \left[n\{1 - \hat{G}_{KM}(T_i)\}\right]$. Therefore, even though the likelihoods $EL_{1n}(\theta)$ and $EL_{2n}^*(\theta)$ look different, they maximize at the same F, and thus the confidence regions obtained by inverting the tests, are centered at the same point $\hat{\theta}_n$. Accordingly the likelihood ratio statistic is given as

$$-2\log ELR_{2n}^*(\theta) = -2\log\{EL_{2n}^*(\theta)/EL_{2n}^*(\hat{\theta}_n)\}\,.$$

This method is also a plug-in method as we plug $\hat{G}_{KM}(t)$ for $G_0(t)$. It differs from the plug-in method described in Section 3.1, however, in the sense that the problem has been switched from original data to synthetic data before plug-in and the parameter that got plugged is of infinite dimensional ($G_0(t)$).

Weighted empirical likelihood: This third method was proposed by Ren (2001, 2008). Given *fixed* weights $v_i \geq 0$, a weighted empirical likelihood is defined as follow:

$$EL_{3n}(w_1, \ldots, w_n) = \prod_{i=1}^{n} (w_i)^{v_i}\,, \qquad \sum_i w_i = 1\,.$$

It is not hard to show that EL_{3n} is maximized when $w_i = v_i / \sum_j v_j$. With the particular choice of $v_i = \Delta\hat{F}_{KM}(T_i)$ the maximizer of EL_{3n} is $w_i^* = \Delta\hat{F}_{KM}(T_i)$, the jump of Kaplan-Meier. Thus, with this choice of weights, the confidence regions obtained by inverting the likelihood ratio test defined below are also centered at the same value $\hat{\theta}_n$ defined in (3.3).

The weighted empirical likelihood under the null hypothesis can be computed by solving the constrained maximization problem: given $v_i = \Delta\hat{F}_{KM}(T_i)$,

$$EL_{3n}(\theta) = \max_{w_i} \left\{ \prod_{i=1}^{n} (w_i)^{v_i} \,\middle|\, \sum_{i=1}^{n} w_i \delta_i g(T_i) = \theta\,,\ \sum_{i=1}^{n} w_i = 1\,, 0 \leq w_i \right\}\,.$$

The likelihood ratio statistic is given as $-2\log ELR_{3n}(\theta) = -2\log\{EL_{3n}(\theta)/EL_{3n}(\hat{\theta}_n)\}$. In this likelihood, observations i and j are supposed to be independent, from which the product form of the empirical likelihood rises. The weights v_i are supposed to be fixed, non-random. With $v_i = \Delta\hat{F}_{KM}(T_i)$, however, this is violated. Theorem below shows that results in (2.1) hold for all three methods described above.

Theorem 3. *Assume mild regularity conditions under which the Kaplan-Meier integral is consistent and asymptotically normal. Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \longrightarrow U\,, \tag{3.4}$$

in distribution, where $U \sim N(0, \Sigma)$. *Furthermore,*

$$-2\log ELR_{1n}(\theta_0) = n(\hat{\theta}_n - \theta_0)^\top V_{1n}^{-1}(\hat{\theta}_n - \theta_0) + o_p(1)\,, \tag{3.5}$$

$$-2\log ELR^*_{2n}(\theta_0) = n(\hat{\theta}_n - \theta_0)^\top V_{2n}^{-1}(\hat{\theta}_n - \theta_0) + o_p(1)\,, \tag{3.6}$$

$$-2\log ELR_{3n}(\theta_0) = n(\hat{\theta}_n - \theta_0)^\top V_{3n}^{-1}(\hat{\theta}_n - \theta_0) + o_p(1)\,, \tag{3.7}$$

where $\lim_{n\to\infty} V_{1n} = \Sigma$, $\lim_{n\to\infty} V_{2n} \neq \Sigma$ *and* $\lim_{n\to\infty} V_{3n} \neq \Sigma$.

It follows form Theorem 3 and Hjort, McKeague & Keilegom (2009) that $-2\log ELR_{1n}(\theta_0) \to \chi^2_p$ in distribution whereas the Wilk's theorem does not hold for $-2\log ELR^*_{2n}(\theta_0)$ and $-2\log ELR_{3n}(\theta_0)$. Theorems 1 and 3 together show that the confidence region provided by inverting the test $-2\log ELR_{1n}(\theta_0)$ by the censored empirical likelihood method is asymptotically optimal and in particular better than those confidence regions obtained from ELR^*_{2n} and ELR_{3n}.

4. Empirical Studies

We illustrate the relative optimality of confidence regions using two simulation studies. Confidence regions calibrated to yield same coverage probabilities were obtained by inverting the likelihood ratio tests described above.

In each case the extended likelihood methods which yield the likelihood ratio statistics that do not observe the Wilk's theorem are computationally challenging since the asymptotic distributions need to be estimated. For example, Theorem 2 implies that the extended likelihood method using $ELR_{2n}(\theta)$ require estimating the scale parameters $c_1, \ldots, c_p$ or the eigenvalues of V_2W^{-1}. In real applications

bootstrap may be used to calibrate converge probabilities. In the below simulations we used 10,000 Monte Carlo samples to calibrate the associated confidence regions to have 90% coverage probabilities. In contrast the confidence regions obtained by inverting the likelihood ratio statistics that observe the Wilk's theorem do not need numerical calibration and were constructed based on the 90-th percentiles of the corresponding chi-square limiting distributions. The comparisons provide empirical confirmation of Theorem 1.

4.1. Simulation Studies 1

We used the following regression model with heteroscedastic errors and applied the profile likelihood and the plug-in method described in section 3.1 for the inference of two slope parameters:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i x_{2i}\,,$$

where $x_{1i} \sim N(0,1)$, $x_{2i} \sim \chi^2_1$ and $e_i \sim N(0,1)$. Here the parameters of interest is $\theta = (\beta_1, \beta_2)$ with the true value $\theta_0 = (1, 0.5)$ and the intercept is the nuisance parameter ($\psi = \alpha$). Consider $Z_i = (Z_{1i}, Z_{2i})$ with $Z_{1i} = y_i$ and $Z_{2i} = (x_{1i}, x_{2i})$. The estimating functions $m(Z_i, \theta, \psi)$ and $h(Z_i, \theta, \psi)$ are given respectively as $m(Z_i, \theta, \psi) = (Z_{1i} - \theta^\top Z_{2i})Z_{2i}$ and $h(Z_i, \theta, \psi) = (Z_{1i} - \theta^\top Z_{2i})$. Figure 1 shows empirically estimated null distributions of the profile likelihood ratio statistic $-2\log ELR_{1n}(\theta_0)$ and the plug-in likelihood ratio statistic $-2\log ELR_{2n}(\theta_0)$ in (3.1) based on 5,000 Monte Carlo samples. The Q-Q plots show that the null distribution of the plug-in likelihood ratio statistic remains deviating from a chi-squared distribution with degree of freedom 2 even in a large sample case.

Figure 2 shows an example of confidence regions. The confidence regions provided by the plug-in empirical likelihood method are larger relative to those by the profile likelihood and shaped differently.

4.2. Simulation Studies 2

We used censored data T_i and δ_i from the following censored data setting: lifetime Y_i and censoring times C_i are independent identical with $Y_i \sim \exp(1)$ and $C_i \sim 0.1 + \exp(0.6)$. The two expectations we are going to test are: $EY_i = a_1$, $EI_{[Y_i>0.5]} = a_2$, where the parameters of interest are $\theta = (a_1, a_2)$ with the true value $\theta_0 = (1, e^{-0.5})$.

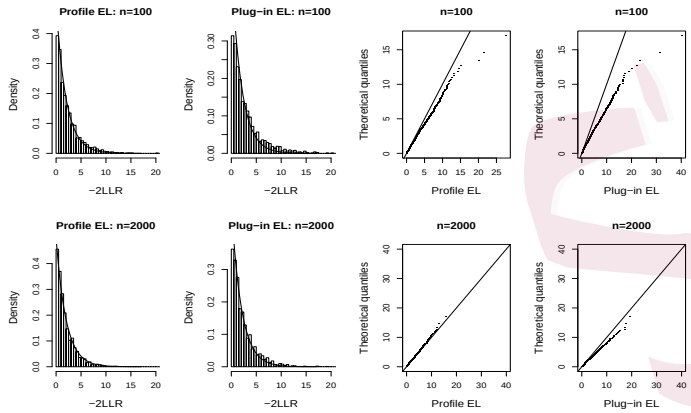

Figure 1: Empirically estimated null distributions of the profile and the plug-in likelihood ratio statistic for the case of estimating equations with nuisance parameters

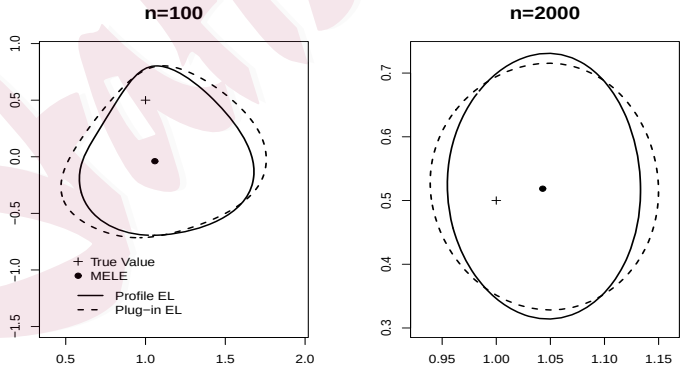

Figure 2: Confidence regions drawn for a sample of the case of estimating equations with nuisance parameters

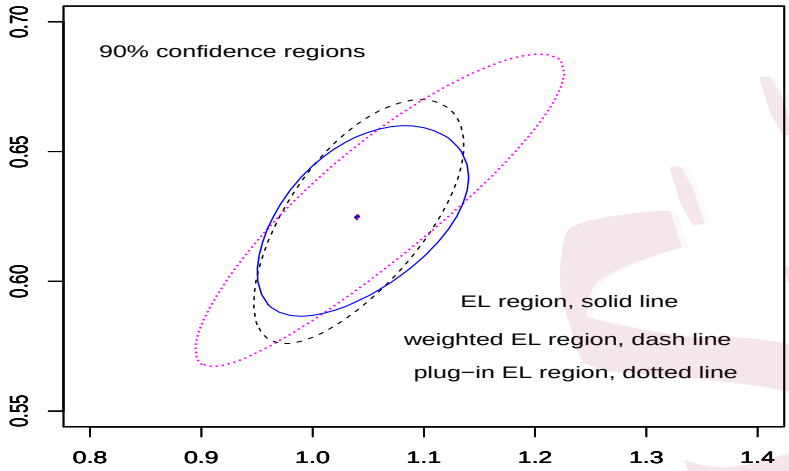

Figure 3: Confidence regions drawn for a sample data of $n = 1,000$ subject to 37% censoring

Figure 3 shows three confidence regions drawn for a sample data with the sample size $n = 1,000$ subject to 37% censoring. As implied by Theorem 3, the orientation and shape of the confidence regions by the plug-in (using $-2\log ELR^{*}_{2n}(\theta_0)$) and the weighted empirical likelihood methods (using $-2\log ELR_{3n}(\theta_0)$) are very different from the confidence region by the censored empirical likelihood method ($-2\log ELR_{1n}(\theta_0)$). Also, the confidence region by the censored empirical likelihood method is the one with the smallest area.

5. Conclusion

In this paper we consider several extensions of empirical likelihood method with some providing likelihood ratio statistics that observe a nonparametric version of Wilks theorem similarly as the Owen's empirical likelihood ratio statistic and others not. Limiting to the extensions that provide the same maximum empirical likelihood estimators, we evaluate relative optimality of the methods by comparing the confidence regions obtained by inverting the likelihood ratio tests. We show that the extension method which yields the likelihood ratio statistic observing the Wilks theorem provides the smallest confidence region.

Acknowledgements First author is supported by US NSF Grant DMS-1007535. Second author is supported by US NSF Grant DMS-1007666.

6. Appendix

Proof of Theorem 1 We consider

$$C_{1n}^* = \{\,\theta \big| n(\hat{\theta}-\theta)^\top V_1^{-1}(\hat{\theta}-\theta) \le c_1\}\,, \qquad C_{2n}^* = \{\,\theta \big| n(\hat{\theta}-\theta)^\top V_2^{-1}(\hat{\theta}-\theta) \le c_2\}\,,$$

where $n^{1/2}(\hat{\theta}-\theta) \to N_p(u, W)$ for $\theta = \theta_0 + n^{-1/2}u$ with a $p-$ variate constant vector u. It follows from (2.1) that $Vol(C_{1n}^*) = Vol(C_{1n}) + \epsilon_{1n}$ and $Vol(C_{2n}^*) = Vol(C_{2n}) + \epsilon_{2n}$ where $\epsilon_{1n}, \epsilon_{2n} \to 0$. WLOG, we chose the α_n such that $\Phi^{-1}(1-\alpha_n) = O(\sqrt{n})$, then the limiting volumes do not vanish as $n \to \infty$. By Lemma 2 below, as $n \to \infty$, $Vol(C_{1n}^*)/Vol(C_{2n}^*) \approx Vol(C_{2n}^*)/Vol(C_{1n}^*) < 1$ unless $V_2 = W$.

Remark 2. If we let $\alpha_n = \alpha$ (fixed), both confidence regions would have volumes shrink down to zero as n grows and hence the relative optimality is in the asymptotic sense of the confidence level increasing with n.

Lemma 2. *(i) Consider two types of confidence regions based on an estimator $\hat{\theta}$ that has distribution $N(\theta_0, I)$: $C_1 = \{\theta :\ (\hat{\theta}-\theta)^\top(\hat{\theta}-\theta) < c\}$, $C_2 = \{\theta :\ (\hat{\theta}-\theta)^\top B(\hat{\theta}-\theta) < c^*\}$, where $B \neq I$ is a symmetric non-negative definite matrix. When both confidence region have same coverage probability $1-\alpha \in (0,1)$, C_1 is superior to C_2 in the sense that $Vol(C_1) < Vol(C_2)$.*

(ii) If $\hat{\theta}$ is normally distributed with $N(\theta_0, \Sigma)$, then the optimal confidence region for the parameter θ_0 is $\{\theta :\ (\hat{\theta}-\theta)\Sigma^{-1}(\hat{\theta}-\theta) < c\}$.

Remark 3. Obviously if $B = \sigma^2 I$, then the two confidence regions in Lemma 2 (i) are identical. The constant σ^2 will be canceling out when adjust the cut off level c^*. So the Lemma should be understood as to use any symmetric non-negative definite matrix other than $\sigma^2 I$ type.

Remark 4. If confidence regions are allowed not to center at $\hat{\theta}$ then there may be better regions available (Stein phenomena) than those in Lemma 2. We only consider a limited class of confidence regions that are centered at the same $\hat{\theta}$ since they are derived from extended empirical likelihood ratio tests.

Proof of Lemma 2 First we note that (ii) is an easy consequence of (i), so proving (i) suffices. Note that the coverage probability of C_1 is $\int_{\|\theta_0 - x\|_2 < c} f(x)dx$,

where $f(x)$ denotes the density of $\hat{\theta}$. Define a distance by $d_B(\theta_0, \hat{\theta}) = (\theta_0 - \hat{\theta})^\top B(\theta_0 - \hat{\theta})$. The coverage probability of C_2 is $\int_{d_B(\theta_0 - x) < c^*} f(x)dx$. By substituting $y = \hat{\theta} - \theta_0$ with $y \sim N(0, I)$, the two integrals become

$$\int_{\|y\|_2 < c} f(y)dy \quad \text{and} \quad \int_{d_B(y) < c^*} f(y)dy \; .$$

Denote the regions of the two integrations above as $R_1 = \| \; y \; \|_2 < \; c$ and $R_2 = d_B(y) < c^*$. Since the normal density function $f(y)$ is a strict monotone decreasing function of $\| \; y \; \|$, the requirement of equal coverage probability

$$\int_{R_1} f(y)dy = \int_{R_2} f(y)dy \tag{6.1}$$

forces $Vol(R_2) > Vol(R_1)$. In other words, R_1 is the region with highest $f(y)$ values that integrates to a given α value, any other region that integrate to the same α value will have to include regions with some lower $f(y)$ values, thus a larger $Vol(R_2)$.

It is obvious that the inequality is strict unless R_2 coincide with R_1.

Proof of Theorem 2 Define $Q_n = \left(n^{-1}\sum_{i=1}^n m(Z_i, \theta_0, \psi_0), n^{-1}\sum_{i=1}^n h(Z_i, \theta_0, \psi_0)\right)$. From the proofs of Theorem 1 and Corollary 5 of Qin and Lawless (1994) we have

$$\begin{aligned} n^{1/2}(\hat{\theta}_n - \theta_0) &= \begin{pmatrix} \Sigma_{(11)} & \Sigma_{(12)} \end{pmatrix} S_{21} S_{11}^{-1} (n^{1/2} Q_n) + o_p(1) \, , \\ n^{1/2}(\hat{\psi}_n - \psi_0) &= \begin{pmatrix} \Sigma_{(21)} & \Sigma_{(22)} \end{pmatrix} S_{21} S_{11}^{-1} (n^{1/2} Q_n) + o_p(1) \, . \end{aligned} \tag{6.2}$$

where $S_{21(2)} = S_{12(2)}^\top = E\left\{(\partial m/\partial\psi)^\top, (\partial h/\partial\psi)^\top\right\}$. After some matrix manipulation we have

$$-2\log ELR_{1n}(\theta_0) = (n^{1/2}Q_n)^\top S_{11}^{-1} S_{12} \begin{pmatrix} \Sigma_{(11)} \\ \Sigma_{(21)} \end{pmatrix} (\Sigma_{(11)})^{-1} [\Sigma_{(11)}, \Sigma_{(12)}] S_{21} S_{11}^{-1} (n^{1/2} Q_n).$$

On the other hand, by the standard arguments of empirical likelihood methods,

$$-2\log ELR_{2n}(\theta_0) = \tilde{\lambda}^\top \left\{E(mm^\top)\right\} \tilde{\lambda} - \hat{\lambda}^\top \left\{E(mm^\top)\right\} \hat{\lambda} + o_p(1) \, , \tag{6.3}$$

where $\hat{\lambda}$ and $\tilde{\lambda}$ are defined by solutions to the equations

$$0 = n^{-1}\sum_{i=1}^n \frac{m(Z_i, \hat{\theta}_n, \hat{\psi}_n)}{1 + \hat{\lambda}^\top m(Z_i, \hat{\theta}_n, \hat{\psi}_n)} \, , \quad 0 = n^{-1}\sum_{i=1}^n \frac{m(Z_i, \theta_0, \hat{\psi}_n)}{1 + \tilde{\lambda}^\top m(Z_i, \theta_0, \hat{\psi}_n)} \, .$$

From (6.2) and (6.3) we have

$$-2\log ELR_{2n}(\theta_0) = n(\hat{\theta}_n - \theta_0)^\top \left(E\frac{\partial m}{\partial \theta}\right)^\top \left\{E(mm^\top)\right\}^{-1} \left(E\frac{\partial m}{\partial \theta}\right)(\hat{\theta}_n - \theta_0) + o_p(1)\,.$$

Definitions of S_{12}, S_{21}, S_{11} and $\Sigma_{(11)}$ complete the proof.

Proof of Theorem 3 When $p = 1$ the result (3.4) is given by Akritas (2001). Zhou (2011) show that for $p > 1$ the equation (3.4) also holds with the jkth element of Σ defined by

$$\sigma_{jk} = \int [g_j(x) - \bar{g}_j(x)][g_k(x) - \bar{g}_k(x)]\frac{dF_0(x)}{1 - G_0(x-)}\,, \tag{6.4}$$

where $\bar{g}_j(t)$ denoting the 'advanced time transformation' of g_j in Efron and Johnstone (1990).

Zhou (2011) also showed that the equation (3.5) holds with the jkth element of V_{1n} given by

$$\hat{\sigma}_{jk} = \sum_{i=1}^{n} [g_j(T_i) - \bar{g}_j(T_i)][g_k(T_i) - \bar{g}_k(T_i)]\frac{\Delta \hat{F}_{KM}(T_i)}{1 - \hat{G}_{KM}(T_i)}\,.$$

Since V_{1n} is just Σ with the unknown F_0 and G_0 replaced by their Kaplan-Meier estimators and the well known fact that the Kaplan-Meier estimators are uniformly consistent, $V_{1n} \to \Sigma$ in probability as $n \to \infty$. We refer to Zhou (2011) for details.

It follows from Owen (2001, p. 220-221) (with X_i^* in place of X_i there) that

$$-2\log ELR_{2n}^*(\theta_0) = U_{2n}^\top V_{2n}^{-1} U_{2n} + o_p(1)$$

where

$$U_{2n} = \sqrt{n}[\bar{X}^* - \theta_0] = \sqrt{n}(\hat{\theta}_n - \theta_0)\,,\ V_{2n} = n^{-1}\sum_{i=1}^{n}(X_i^* - \theta_0)(X_i^* - \theta_0)^\top\,.$$

Since $\hat{G}_{KM}(t)$ is uniformly consistent for $G_0(t)$,

$$V_{2n} = n^{-1}\sum_{i=1}^{n}(X_i - \theta_0)(X_i - \theta_0)^\top + o_p(1) = Var(X_1) + o_p(1)\,.$$

By the results in Zhou (1992) and Srinivasan and Zhou (1994), $var(X_1) \neq \Sigma$ however. This suffices to show (3.6).

As for the equation (3.7), when $p = 1$, Ren (2001) showed that the limiting distribution of this test statistic with $\theta = \theta_0$ is a scaled chi-square distribution with degree of freedom 1. To obtain a multivariate version (with $p > 1$), we modify the proof of Ren along the steps of Owen (2001, p. 220-221) to incorporates weights, and derive the following:

$$-2\log ELR_{3n}(\theta) = U_n^\top V_{3n}^{-1} U_n + o_p(1)$$

where

$$U_n = \sqrt{n}[\sum_{i=1}^{n} g(T_i)\Delta\hat{F}_{KM}(T_i) - \theta_0] = \sqrt{n}(\hat{\theta}_n - \theta_0)$$

and V_{3n} is a matrix with the jkth element given by

$$v_{jk} = \sum_{i=1}^{n}(g_j(T_i) - \theta_0)(g_k(T_i) - \theta_0)\Delta\hat{F}_{KM}(T_i) \ .$$

An application of the law of large number for the Kaplan-Meier integral gives

$$v_{jk} = \int (g_j(t) - \theta_0)(g_k(t) - \theta_0)d\hat{F}_{KM}(t) \to \int (g_j(t) - \theta_0)(g_k(t) - \theta_0)dF_0(t)$$

in probability as $n \to \infty$. Since Σ depends on the censoring distribution G_0, while $\lim V_{3n}$ does not, $\lim_{n\to\infty} V_{3n} \neq \Sigma$.

References

Akritas, M. (2000). The central limit theorem under censoring. *Bernoulli* **6**, 1109-1120.

Efron, B. & Johnstone, I. M. (1990). Fisher's information in terms of the hazard rate. *Ann. Statist.* **18**, 38-62.

Hjort, N. L., McKeague, I. W. & Keilegom, I. V. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37**, 1079-1111.

Kaplan, E. L. & Meier, P. (1958). Nonparametric Estimation From Incomplete Observations. *J. Amer. Statist. Assoc.* **53**, 457-481.

Li, G., Li, R. & Zhou, M. (2005). Empirical likelihood in survival analysis. In *Contemporary Multivariate Analysis and Design of Experiments* Edited by J. Fan and G. Li. pp. 337-350. The World Scientific Publisher.

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

Owen, A. (1990). Empirical Likelihood Ratio Confidence Regions. *Ann. Statist.* **18**, 90-120.

Owen, A. (2001). *Empirical Likelihood*. New York: CRC Press.

Qin, J. & Lawless, J. (1994). Empirical Likelihood and General Estimating Equations. *Ann. Statist.* **22**, 300-325.

Ren, J. (2001). Weighted empirical likelihood ratio confidence intervals for the mean with censored data. *Ann. Inst. Statist. Math.* **53**, 498-516.

Ren, J. (2008). Weighted empirical likelihood in some two-sample semiparametric models with various types of censored data. *Ann. Statist.* **36**, 147-166.

Shi, J. & Lau, T.-S. (2000). Empirical Likelihood for Partially Linear Models. *J. Multivariate Anal.* **72**, 132-148.

Srinivasan, C. & Zhou, M. (1994). Linear regression with censoring. *J. Multivariate Anal.* **49**, 179-201.

Wang, S., Qian, L. & Carroll, R. (2010). Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika* **97**, 79-93.

Wang, Q. H. & Jing, B. Y (2001). Empirical likelihood for a class of functionals of survival distribution with censored data *Ann. Inst. Statist. Math.* **53**, 517-527.

Wang, Q. H. & Li, G. (2002). Empirical likelihood semiparametric regression analysis under random censorship *J. Multivariate Anal.* **83**, 469-489

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

Zhou, M. (2011). A Wilks theorem for the censored empirical likelihood of several means. *Preprint* www.ms.uky.edu/∼mai/research/Note3.1.pdf

Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco CA 94143, U.S.A.

E-mail: (miok.kim@ucsf.edu)

Department of Statistics, University of Kentucky, Lexington KY, 40536-0082, U.S.A.

E-mail: (mai@ms.uky.edu)