Published in final edited form as:

Scand Stat Theory Appl. 2015 September ; 42(3): 798–812. doi:10.1111/sjos.12137.

Empirical Likelihood for Censored Linear Regression and Variable Selection

Tong Tong Wu¹, Gang Li², and Chengyong Tang³

¹Department of Biostatistics and Computational Biology, University of Rochester

²Department of Biostatistics, University of California, Los Angeles

³Department of Statistics, Fox School of Business, Temple University

Abstract

The linear regression model for right censored data, also known as the accelerated failure time model using the logarithm of survival time as the response variable, is a useful alternative to the Cox proportional hazards model. Empirical likelihood as a nonparametric approach has been demonstrated to have many desirable merits thanks to its robustness against model misspecification. However, the linear regression model with right censored data cannot directly benefit from the empirical likelihood for inferences mainly due to dependent elements in estimating equations of the conventional approach. In this paper, we propose an empirical likelihood approach with a new estimating equation for linear regression with right censored data. A nested coordinate algorithm with majorization is used for solving the optimization problems with nondifferentiable objective function. We show that the Wilks theorem holds for the new empirical likelihood. We also consider the variable selection problem with empirical likelihood when the number of predictors can be large. Since the new estimating equation is nondifferentiable, a quadratic approximation is applied to study the asymptotic properties of penalized empirical likelihood. We prove the oracle properties and evaluate the properties with simulated data. We apply our method to a SEER small intestine cancer dataset.

Keywords

Accelerated failure time model; Coordinate descent algorithm; High-dimensional data analysis; Linear regression model; Oracle property; Variable selection; Wilks' theorem

1 Introduction

The empirical likelihood (Owen 1988, 2001) is an influential computation-intensive statistical approach. The conventional parametric likelihood approach plays a fundamental role in practical data analysis for its convenient and efficient estimation and inferences. It,

Corresponding Author: Tong Tong Wu, Ph.D., 4169 Saunders Research Building, Department of Biostatistics and Computational Biology University of Rochester, tongton wu@urmc.rochester.edu.

Supplementary Materials

All the proofs of theorems are provided in the online supplementary file (<http://www.urmc.rochester.edu/biostat/people/faculty/wu-tongtong.cfm>).

however, relies on correctly assuming some parameterized distributions. Being a nonparametric likelihood approach in the sense of being distributional assumption free, the empirical likelihood shares some convenient merits of parametric likelihood (e.g. χ^2 distributed empirical likelihood ratio) and has many desirable advantages in deriving estimators and confidence sets for unknown parameters, including objective determination of the shape of the confidence region (Owen 1988), seamless incorporation of auxiliary or prior information (Qin and Lawless 1994), Bartlett correctability (Chen and Cui 2006, 2007), and many others. In the past two decades, methods using the empirical likelihood approach have been actively and extensively developed with applications in numerous scientific areas; see Owen (2001) for an overview and Chen and Van Keilegom (2009) for a review.

Among those previous works, the empirical likelihood method has been extended to linear regression with right censored time to event data; see, for example, Qin and Jing (2001), Li and Wang (2003), Zhou (2005), Lu and Liang (2006), Zhou and Li (2008), Zhao and Yang (2012), and Li and Lu (2009). For instance, Qin and Jing (2001) and Li and Wang (2003) proposed an approximate complete-data empirical likelihood to estimate the regression coefficients based on the synthetic data approach of Koul et al. (1981). Zhou and Li (2008) and Fang et al. (2013) developed empirical likelihood methods based on the estimating equation of the Buckley-James estimator (Buckley and James 1979) using a censored-data empirical likelihood and an approximate complete-data empirical likelihood to deal with censored data, respectively. However, each of these methods has its limitations. The synthetic data approach makes a strong assumption of independence between censoring time C and survival time and covariates (Y, X) . The approximate complete-data empirical likelihood does not have the standard χ^2 limiting distribution and requires an adjustment in order to use the standard chi-square approximation (Li and Wang 2003; Li and Lu 2009). The censored-data empirical likelihood (Zhou and Li 2008) has the standard χ^2 limiting distribution. However, it is more computationally intensive and more difficult to cover problems such as incorporating auxiliary information and constructing confidence regions for linear combinations of the regression coefficients. Zheng et al. (2012) introduced an EL method based on derived influence functions of parameters of interest with plugging-in consistent nuisance parameter estimators. This influence function based EL method was proved to retain the nonparametric Wilks property under appropriate conditions.

With the current trend of statistical methodological development, empirical likelihood has been investigated in scenarios with increasing complexity and data dimensionality (Hjort et al. 2009; Chen and Van Keilegom 2009; Chen et al. 2009). Tang and Leng (2010) and Leng and Tang (2012) extend the empirical likelihood approach to high-dimensional data by selecting important variables from the feature space. They apply the empirical likelihood approach in conjunction with appropriate penalty functions on the magnitude of the parameters of interest and construct a framework to build sparse models. All developments require more intensive and demanding computations in solving practical problems with empirical likelihood to continue enjoying those nonparametric merits. On the other hand, algorithms for implementing empirical likelihood are relatively less developed in the recent literature. Tang and Wu (2013) develop two versions of nested coordinate descent algorithms to improve practical implementation of empirical likelihood. As for the linear regression

with right censored data discussed above, the empirical likelihood and penalized empirical likelihood approaches remain less investigated partly because the estimating equations generally involve non-smooth functions of the parameter of interest. On the other hand, the merits of the penalized empirical likelihood approach such as parsimonious model construction and the oracle properties are desirable in analyzing right censored survival data.

The first goal of this paper is to introduce a new empirical likelihood approach for right censored linear regression that possesses its conventional nice properties. The second goal is to build parsimonious and robust right censored models with empirical likelihood when the number of covariates is large. Specifically, we propose a complete-data empirical likelihood in combination with an alternative approach to the Buckley-James estimator. Unlike the previous methods in which the estimating equation is a sum of dependent terms, our estimating equation is a sum of approximately i.i.d. terms. Consequently, our empirical likelihood method naturally inherits all nice features of the complete data empirical likelihood method. For example, its empirical likelihood ratio has the standard χ^2 limiting distribution. We consider an approach with appropriate penalization to selecting important variables when the number of variables is large. However, the new estimating equation is nondifferentiable, so we use the quadratic approximation developed by Parente and Smith (2010) to study the asymptotic properties of the penalized empirical likelihood. After the treatment, this penalized right censored empirical likelihood regression model will be proved to have the nice oracle properties (Fan and Li 2001).

Both our proposed method and the method in Zheng et al. (2012) are plug-in empirical likelihood-based methods. However, the main result – Theorem 2.1 in Zheng et al. (2012) – does not generally apply to right censored data problems because it requires a continuously differentiable estimating function. However, the Buckley-James estimating function and the estimating function considered in our paper are not differentiable. In summary, our approach is different than Zheng et al. (2012) in the following ways: (1) our method is an estimating equation based approach; (2) the estimation function is not required to be continuously differentiable in our method; (3) the Buckley-James estimation function is re-expressed as approximately the sum of i.i.d. random variables, which allows us to apply the standard complete-data empirical likelihood method for censored regression setting.

The rest of the paper is organized as follows. A new empirical likelihood method for right censored linear regression is introduced in Section 2. This complete-data empirical likelihood combined with a new estimating equation is extended to variable selection for many covariates. Its asymptotic properties are also investigated. The computational issue is briefly discussed in the end of this section. Simulation experiments are conducted to evaluate its performance in finite sample size and variable selection in Section 3. Section 4 is the SEER (Surveillance, Epidemiology, and End Results) small intestine cancer example using our method. A brief discussion is provided in Section 5. Regularity conditions can be found in the Appendix. All the proofs of theorems are provided in the online supplementary file (<http://www.urmc.rochester.edu/biostat/people/faculty/wu-tongtong.cfm>).

2 Empirical Likelihood for Linear Regression with Right Censored Data

2.1 Estimating Equation for Right Censored Linear Regression

Let Y and C denote times to event and censoring, respectively. The observed survival time T is $\min(Y, C)$ and the censoring indicator is $\delta = \mathbb{I}(Y \leq C)$, which equals 1 if $Y \leq C$ and 0 otherwise. The survival time T is in the range $[0, \zeta]$, where ζ denotes the end of a study. Suppose $X \in \mathbb{R}^p$ is a covariate vector associated with the true event time. Conditionally on X , Y and C are assumed to be independent. Let $\{(X_i, Y_i, C_i, \delta_i), i = 1, \dots, n\}$ be n independent samples of $\{(X, Y, C, \delta)\}$ and $\{(X_i, T_i, \delta_i), i = 1, \dots, n\}$ be observed samples. For the sake of simplicity, denote Y_i as the log transformed version of the survival time, e.g. $\log(Y_i)$. The linear regression model of the survival time on covariates is given by

$$Y_i = X_i^T \beta + \epsilon_i \quad (1)$$

where ϵ_i is the i th random error for the observation following some unknown distribution $F(t) = P(\epsilon_i \leq t)$ and is independent of X_i and C_i . We assume no intercept in model (1) to ensure F is identifiable.

Define Y_i^* to be the conditional expectation of the underlying survival time Y_i on (T_i, δ_i, X_i) .

Thus

$$Y_i^* = E(Y_i | T_i, \delta_i, X_i) \quad (2)$$

$$= \delta_i T_i + (1 - \delta_i) \left[X_i^T \beta + \frac{\int_{T_i - X_i^T \beta}^{\infty} t dF(t)}{1 - F(T_i - X_i^T \beta)} \right].$$

Or, it is equivalent to re-write (2) as

$$\begin{aligned} Y_i^* - X_i^T \beta &= \delta_i (T_i - X_i^T \beta) + (1 - \delta_i) \frac{\int_{T_i - X_i^T \beta}^{\infty} t dF(t)}{1 - F(T_i - X_i^T \beta)} \\ &\equiv \delta_i \epsilon_i^\beta + (1 - \delta_i) \frac{\int_{\epsilon_i^\beta}^{\infty} t dF^\beta(t)}{1 - F^\beta(\epsilon_i^\beta)}, \end{aligned} \quad (3)$$

where $\epsilon_i^\beta = T_i - X_i^T \beta$ is the i th censored residual, and $F^\beta(t)$ is the distribution function of ϵ_i^β depending on β . It is easy to see that

$$E(X_i - \mu_X)(Y_i^* - X_i^T \beta) = 0, \quad (4)$$

so we denote $(X_i - \mu_X)(Y_i^* - X_i^T \beta)$ by $W_i(\beta, F^\beta, \mu_X)$. If μ_X and F^β are known, then the sample version of (4) is $\frac{1}{n} \sum_{i=1}^n W_i(\beta, F^\beta, \mu_X)$. Therefore, it is possible to solve the following estimating equation for β

$$\Psi_n(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i(\beta, F^\beta, \mu_X) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu_X)(Y_i^* - X_i^T \beta) = 0. \quad (5)$$

Unfortunately, μ_X and F^β are unknown in practice so the estimating equation (5) is not useful.

In order to derive a useful estimating equation for β , we note that, for a fixed β , the probability distribution function F^β can be estimated by the Kaplan-Meier estimator $\hat{F}_n^\beta$ based on the observed right censored residuals $\epsilon_i^\beta = T_i - X_i^T \beta$'s. If we replace F^β and μ_X in $W_i(\beta, F^\beta, \mu_X)$ with $\hat{F}_n^\beta$ and $\bar{X} = \sum_i X_i/n$, the estimating equation (5) becomes

$$\Psi_n(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i(\beta, \hat{F}_n^\beta, \bar{X}) = 0, \quad (6)$$

which is the original estimating equation of the Buckley-James estimator (Buckley and James 1979; Ritov 1990). However, because the summands in (6) are not independent, the standard empirical likelihood results for complete data without censoring do not apply directly. Additional efforts are needed to develop the empirical likelihood method based on (6) (Zhou and Li 2008; Li and Lu 2009). Below we consider an alternative estimating equation in which the summand are approximately i.i.d., which will allow adaptation of standard complete data empirical likelihood methods.

Instead of solving the Buckley-James family of estimating equation (6), in which the summands are not independent because of the Kaplan-Meier estimate $\hat{F}_n^\beta$, we consider the Tsiatis family of estimating equation (Tsiatis 1990):

$$\Phi_n(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int w(t) \{X_i - \bar{X}(t)\} dM_i(t), \quad (7)$$

where $w(t) = t - \frac{\int_t^\infty u dF^\beta(u)}{1 - F^\beta(t)}$, $N_i(t) = I(e_i^\beta \leq t, \delta_i = 1)$ is the counting process for the i th censored residual $e_i^\beta = T_i - X_i^T \beta$, $dM_i(t) = dN_i(t) - R_i(t) d\Lambda(t)$ with $\Lambda(t) = -\log[1 - F(t)]$, $\bar{X}(t) = \sum_{i=1}^n R_i(t) X_i / \sum_{i=1}^n R_i(t)$, and $R_i(t) = I(e_i^\beta \geq t)$ is the at-risk process for the i th censored residual. It follows Proposition 4.1 of Ritov (1990) that $\Psi_n(\beta) = \Phi_n(\beta) + o_p(1)$.

Therefore, we can solve $\Phi_n(\beta) = 0$ for β .

Since $w(t)$ still contains the unknown distribution function $F^\beta(t)$, we therefore propose the following estimating equation by replacing $F^\beta(t)$ by the Kaplan-Meier estimate $\hat{F}_n^\beta(t)$ based on the censored residuals:

$$\Gamma_n(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n S_i(\beta), \quad (8)$$

Where

$$S_i(\beta) = \int \hat{w}(t) \{X_i - \bar{X}(t)\} d\hat{M}_i(t), \quad (9)$$

$\hat{w}(t) = t - \frac{\int_t^\infty u d\hat{F}_n^\beta(u)}{1 - \hat{F}_n^\beta(t)}$, and $d\hat{M}_i(t) = dN_i(t) - R_i(t) d\hat{\Lambda}(t)$ with the Nelson-Aalen estimator based

on the censored residuals $\hat{\Lambda}(t)$ for $\Lambda(t)$. We prove in Lemma 1 in the Appendix that $\Gamma_n(\beta) = \Phi_n(\beta) + o_p(1)$, therefore, instead of solving (6) or (7), we can equivalently solve (8) for β . We also prove that $\Gamma_n(\beta)$ is a sum of approximately i.i.d. random variables, so the results of complete-data empirical likelihood can apply. This new estimating equation leads to the following definition of empirical likelihood for β . By solving the estimating equation $\Gamma_n(\beta) = 0$ for β , we define an M-estimator $\hat{\beta}$ satisfying $\Psi_n(\hat{\beta}) = o_p(1)$. It is straightforward from Theorem 5.1 in Ritov (1990), $\hat{\beta}$ is $\sqrt{n}$ -consistent and asymptotically normal.

2.2 Empirical Likelihood for Right Censored Linear Regression and Asymptotic Properties

We now consider constructing empirical likelihood for β for convenient and robust statistical inferences. With the estimating function (8), we define the empirical likelihood for β as follows:

$$L(\beta) = \sup \left\{ \prod_{i=1}^n p_i \left| \sum_{i=1}^n p_i S_i(\beta) = 0, p_i \geq 0, \sum_{i=1}^n p_i = 1 \right. \right\}. \quad (10)$$

The empirical likelihood ratio is given by

$$\ell(\beta) = -2 \log \left[L(\beta) \right] - 2n \log n = 2 \sum_{i=1}^n \log \left(1 + \lambda^T S_i(\beta) \right) \quad (11)$$

where λ is the Lagrange multiplier $\lambda \in \mathbb{R}^r$ satisfying

$$\sum_{i=1}^n \frac{S_i(\beta)}{1 + \lambda^T S_i(\beta)} = 0. \quad (12)$$

Since $S_i(\beta)$ are independent and identically distributed asymptotically, we show that the Wilks theorem holds for the above empirical likelihood.

Theorem 1. *Under the regularity conditions in the Appendix, we have as $n \rightarrow \infty$,*

$$\ell(\beta_0) \rightarrow \chi_p^2$$

in distribution, where p is the dimensionality of β .

The proof of Theorem 1 is given in the Appendix. Correspondingly, we can define the profile empirical likelihood for a subset of regression coefficients as follows. Let

$\beta = (\beta_1^T, \beta_2^T)^T$, where β_1 is a sub-vector of β being the components of parameter of interest.

The profile empirical likelihood for β_1 is defined by

$$\ell^*(\beta_1) = \inf_{\beta_2} \ell(\beta). \quad (13)$$

Then we can show that the profile empirical likelihood ratio is also central χ^2 distributed.

Theorem 2. *Under the regularity conditions in the Appendix, we have as $n \rightarrow \infty$,*

$$\ell^*(\beta_{10}) \rightarrow \chi_q^2$$

in distribution, where q is the dimensionality of β_1 .

With Theorem 2, the $1 - \alpha\%$ confidence region for β_1 can be constructed by

$$\left\{ \beta_1 : \ell^*(\beta_1) \leq \chi_{q, \alpha}^2 \right\} \quad (14)$$

where $\chi_{q, \alpha}^2$ is the upper α -level quantile of the χ^2 distribution with q degrees of freedom.

In a short summary, Theorems 1 and 2 show that with approximately independent estimating equations, the empirical likelihood ratio and profile empirical likelihood ratio enjoy the Wilks property so that our empirical likelihood approach can be more conveniently applied for solving practical problems with right censored data.

2.3 Penalized Empirical Likelihood for Variable Selection

To build robust learning models and identify relevant predictors to the response variable, one can use the variable selection technique of penalizing the objective function. In this section, we show that variable selection can also be conducted in empirical likelihood for linear regression models on right censored data. The SCAD penalty (Fan and Li 2001) is used as a demonstration in this paper:

$$p'_\tau(|\beta|) = \tau \left\{ I(|\beta| \leq \tau) + \frac{(a\tau - \beta)_+}{(a-1)\tau} I(\beta > \tau) \right\},$$

where τ is the tuning constant controlling the strength of shrinkage, and $a = 3.7$ following the recommendation in Fan and Li (2001).

We define the penalized empirical likelihood estimator to be the minimizer of

$$\ell_p(\beta) = \ell(\beta) + \sum_{i=1}^p p_\tau(|\beta_{j_i}|) = \sum_{i=1}^n \log(1 + \lambda^T S_i(\beta)) + \sum_{i=1}^p p_\tau(|\beta_{j_i}|) \quad (15)$$

where λ is the Lagrange multiplier satisfying (12). As pointed out by Tang and Leng (2010), involving the penalty function on β has no impact on the evaluation of the empirical likelihood ratio for given β , which conveniently facilitates the calculation of the penalized empirical likelihood.

Since $S_i(\beta)$ involves β in an indicator function, it is not differentiable. Therefore, standard results in Qin and Lawless (1994) for empirical likelihood with estimating equations do not straightforwardly apply. In addition, the results in Leng and Tang (2012) for penalized empirical likelihood with estimation equations also do not carry forward automatically. As shown in the Appendix, the quadratic approximation of Parente and Smith (2010) for empirical likelihood is used for studying the asymptotic properties of the penalized empirical likelihood with nondifferentiable estimating functions.

To choose the optimal value for the tuning parameter τ , we combine our penalized empirical likelihood method with three information criteria: BIC of Schwarz (1978), BICC of Wang et al. (2009), and EBIC of Chen and Chen (2008). The three BIC-type criteria are defined as

$$\begin{aligned} \text{BIC}(\tau) &= -2\ell_p(\beta_\tau) + \log(n) \cdot \text{df}_\tau, \\ \text{BICC}(\tau) &= -2\ell_p(\beta_\tau) + \max\{1, \log \log(p)\} \log(n) \cdot \text{df}_\tau, \\ \text{EBIC}(\tau) &= -2\ell_p(\beta_\tau) + [\log(n) + 2\log(p)] \cdot \text{df}_\tau. \end{aligned}$$

where β_τ is the estimate of β with τ being the tuning parameter, and df_τ is the number of nonzero coefficients in β_τ . The optimal τ minimizes the BIC criteria. In general, all the three BIC-type criteria work well. One can use BIC for its well-known model selection consistency.

Without loss of generality, assume the truth $\beta_0 = (\beta_{10}^T, 0^T)^T$, where $\beta_{10} \in \mathbb{R}^s$ denotes the true nonzero components. Now suppose the penalized estimator $\hat{\beta}$ is the minimizer of (15) and is arranged in the same way, i.e., $\hat{\beta} = (\hat{\beta}_1^T, \hat{\beta}_2^T)^T$, where $\hat{\beta}_1 \in \mathbb{R}^s$ and $\hat{\beta}_2 \in \mathbb{R}^{p-s}$. We prove the selection consistency in Theorem 3.

Theorem 3. *Under the regularity conditions in the Appendix, we have $P(\hat{\beta}_2 = 0) \rightarrow 1$ as $n \rightarrow \infty$.*

Define $g_i(\beta) = \int w(t) \{X_i - \bar{X}(t)\} dN_i(t)$, $g(\beta) = E\{g_i(\beta)\}$, $G = E\{g_i(\beta)g_i^T(\beta)\}$, and $Q = E\{g_i(\beta)g_i^T(\beta)\}$. Let $\Omega = (G^T Q^{-1} G)^{-1}$, which is the asymptotic efficiency (Ritov 1990) of $\tilde{\beta}$, i.e., as $n \rightarrow \infty$,

$$\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, \Omega)$$

where $\tilde{\beta}$ is the unpenalized estimator that solves $\Gamma_n(\beta) = 0$. We can decompose the variance-covariance matrix Ω as a block matrix according to the arrangement of β_0 as

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}.$$

The efficiency of $\hat{\beta}_1$ is summarized in the following theorem.

Theorem 4. *Under the regularity conditions in the Appendix, we have as $n \rightarrow \infty$,*

$$\sqrt{n}(\hat{\beta}_1 - \beta_{10}) \rightarrow N(0, \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21})$$

in distribution.

From Theorem 4, it is clear that the penalized estimate $\hat{\beta}_1$ is more efficient than the unpenalized estimate by comparing their variances $\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$ and Ω_{11} . The improvement of the asymptotic variance is seen as a gain by using the penalized empirical likelihood. By estimating the unimportant components in the parameter as zero, the penalized empirical likelihood approach achieves better variance of the important elements, which is analogous to the situation of using empirical likelihood with prior information $\beta_2 = 0$; see Qin and Lawless (1995).

2.4 Computation

Though the empirical likelihood and penalized empirical likelihood have good properties for analyzing right censored data, dedicated effort is needed for handling the optimizations in the empirical likelihood approach. It is well known that the computation of empirical likelihood is difficult, especially when the dimension of covariates is moderate to high. A recent paper of Tang and Wu (2013) develops two versions of computational algorithms to tackle this problem. One is a nested coordinate descent algorithm with two layers to estimate λ and β separately, and the other is to couple the nested coordinate descent algorithm with the MM (majorization-minimization) algorithm (Hunter and Lange 2004; Lange 2004; Wu and Lange 2010). The latter is to find a surrogate matrix to approximate the Hessian matrix of the empirical likelihood and can further simplify the computation. Since $S_i(\beta)$ is not differentiable, the Newton-Raphson type algorithms for optimizing with respect to β is not feasible. Thus, we propose to solve the optimization with the coordinate decent MM algorithm. Combining with the MM algorithm will also improve the instability problem in the computation of EL methods.

There are two layers of optimizations. The first one is to evaluate $\ell(\beta)$ in (11) for a given β . Note that a pseudo-logarithm function is used throughout this paper to overcome the bounded support problem of the logarithm function (Owen 2001):

$$\log_{\star}(z) = \begin{cases} \log(z) & \text{if } z \geq \epsilon; \\ \log(\epsilon) - 1.5 + 2z/\epsilon - z^2/(2\epsilon^2) & \text{if } z \leq \epsilon, \end{cases}$$

where ϵ is usually chosen to be $1/n$. This pseudo-logarithm function is twice differentiable and is more stable in solving empirical likelihood than the original logarithm function. In our study, we use $\ell_{\star}(\beta) = \sum_{i=1}^n \log_{\star}[1 + \lambda^T S_i(\beta)]$ as a surrogate for $\ell(\beta)$. We apply the coordinate decent algorithm proposed in Tang and Wu (2013) for evaluating $\ell(\beta)$ for implementing our method. In practice, the pseudo-logarithm function does not change the asymptotic properties of empirical likelihood.

The second layer of optimization is respect to β or part of its components for evaluating the profile empirical likelihood and penalized empirical likelihood. We propose to apply majorization by using the quadratic bound principle:

$$\begin{aligned} \ell_{\star}(\beta) &= \ell_{\star}(\tilde{\beta}) + d\ell_{\star}(\tilde{\beta})(\beta - \tilde{\beta}) + (\beta - \tilde{\beta})^T \left\{ \int_0^1 d^2 \ell_{\star}[\tilde{\beta} + t(\beta - \tilde{\beta})](1-t) dt \right\} (\beta - \tilde{\beta}) \\ &\leq \ell_{\star}(\tilde{\beta}) + d\ell_{\star}(\tilde{\beta})(\beta - \tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})^T B(\beta - \tilde{\beta}) \\ &\equiv h(\beta | \tilde{\beta}) \end{aligned}$$

where the matrix B satisfies $B \geq d^2 \ell(\beta)$ (i.e., $B - d^2 \ell(\beta)$ is nonnegative definite). Since the pseudo-logarithm function is twice differentiable and has a bounded curvature $-\lambda \lambda^T / \epsilon^2$, it is easy to define the B matrix.

3 Simulation Examples

3.1 Asymptotic Distribution of Empirical Likelihood Ratio

In this example, we examine the distribution of the empirical likelihood ratio $\ell(\beta)$ evaluated at the true value β_0 and compare with χ^2 distributions. We generate $n = 100$ random samples from multivariate normal distributions in $\mathbb{R}^3$ with mean 0 and the pairwise covariance

$$\text{Cov}(X_{ij}, X_{ik}) = \begin{cases} 1 & \text{for } j = k \\ \rho & \text{for } j \neq k \end{cases}$$

for $i = 1, \dots, n$. Both independent ($\rho = 0$) and correlated cases ($\rho = 0.5$ and 0.9) are considered. The true survival time Y_i is generated by

$$E(Y_i | X_i) = X_i^T \beta,$$

where $\beta^T = (3, 2, 1)^T$. The censoring time is generated from a normal distribution with mean 1 and variance 5^2 . The average censoring rate is 45%. We repeat the simulation for 100 times.

Quantile-quantile (QQ) plots in Figure 1 compare the empirical likelihood ratios vs. χ_3^2 distribution. It is clear that the empirical distribution from our method have almost identical to a χ_3^2 distribution. This shows the validity of the new empirical likelihood with approximate central χ^2 distribution. As for comparison, we show the QQ plots in the second row of Figure 1 for the EL based on Buckley-James (BJ) estimating equation such that

$$\sum_{i=1}^n X_i (Y_i^* - X_i^T \beta) = 0,$$

where Y_i^* is defined in equation (2). Obviously, the QQ plots based on the Buckley-James estimating equation are quite different than a straight line, which means the empirical likelihood ratios are far away from a central χ_3^2 distribution.

3.2 χ^2 Distributions for Profiled Empirical Likelihood

In this simulation example, we examine the empirical distributions of $\ell(\beta_2)$ evaluated at the truth β_{20} by varying its size s . The empirical coverage of the confidence region (14) is also checked.

The predictors variables X are generated similarly as the previous example with different correlation coefficient ρ . The true survival time Y_i is generated by

$$E(Y_i | X_i) = X_i^T \beta,$$

where $\beta^T = (5, 4, 3, 2, 1)^T$. The censoring time is generated from a normal distribution with mean 1 and variance 5^2 . The average censoring rate is 48%.

We examine $s = 1, \dots, 3$ corresponding to the profiled empirical likelihood for s component(s) in β . Without loss of generality, β_2 is chosen to be the first s components of β . We fix β_2 at its truth β_{20} and calculate (13) using the nested coordinate decent algorithm (Tang and Wu 2013). As the simulation is repeated for 500 times, we record the value $2\ell(\beta_{20})$.

The coverage probabilities, i.e., the frequency such that $2\ell(\beta_{20}) > \chi_{1,0.95}^2$, based on the 500 replicates at different ρ 's and s 's are calculated and summarized in Table 1. The empirical distribution of $2\ell(\beta_{20})$, which asymptotically follows a χ_s^2 distribution, are compared to those of a χ_s^2 distribution by examining the QQ plots in Figure 2 at different ρ 's and s 's. It is obvious from both Table 1 and Figure 2 that the nested optimization based on coordinate descent is satisfactory, in both maintaining the nominal level of coverage and approximating the asymptotic distributions.

3.3 Penalized Empirical Likelihood

We illustrate the applications of empirical likelihood in variable selection in right censored linear regression. The SCAD penalty (Fan and Li 2001) is used in (15).

The predictor X_i , $i = 1, \dots, n$ is generated from a multivariate normal distribution with mean zero and a compound symmetry variance-covariance matrix with $\sigma = 1$ and $\rho = 0.5$. The true survival time Y_i is generated by

$$E(Y_i | X_i) = X_i^T \beta,$$

where $\beta^T = (3, 1.5, 0, 0, 2, 0, \dots, 0)^T$. The censoring time is generated from a normal distribution with mean 1 and variance 5^2 . The average censoring rate is 48%.

Table 2 reports the results for $n = 100, 200$, $p = 10, 20$, and $\rho = 0.5$. Our penalized empirical likelihood method is combined with three information criteria – BIC, BICC, and EBIC – for selecting the tuning parameter τ and compared with the R package *bujar* developed by Wang and Wang (2010). The column of N_{nonzero} reports the average total number of selected predictors and The column of N_{true} reports the average total number of selected true predictors from (X_1, X_2, X_5) . The last column reports the average model error (ME) defined by

$$\text{ME} = (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta),$$

where Σ is the variance-covariance matrix of the predictors. Standard errors are reported in the parentheses. The model error of the oracle model (i.e., the model using the true predictors) is also reported in the last column of each setting. Both the PEL and *bujar* methods select all three true predictors in all settings. When the predictors are correlated, the

R package *bujar* selects more false positives. The model errors of the penalized empirical likelihood methods are smaller than *bujar* and are closed to the Oracle estimation.

4 Analysis of SEER Small Intestine Data

In this section, we apply our method to the small intestine data set from the Surveillance, Epidemiology, and End Results (SEER 1973–2002) Program supported by the National Cancer Institute (NCI). Small intestine cancer is a relatively rare cancer compared to other gastrointestinal malignancies such as gastric cancer (stomach cancer) and colorectal cancer (Terry and Santora 2006). Surgery and radiation therapy are the most commonly used treatments for small intestine cancer. However, exposure to radiation may cause damage to cellular DNA and increase the risk for certain types of tumors. For example, Hashibe et al. (2005) showed that radiotherapy for first primary oral cancer is a significant risk factor for cancer development.

The goal of the analysis is to determine the efficacy of surgery and radiation therapy for small intestine cancer. The data set includes 2,669 patients who were diagnosed with small intestine cancer (ICD9 = 152) as a single or first primary tumor in the period from 1973–2002. Each of the 2,699 patients has complete information on surgery status (yes or no), radiation therapy (yes or no), and the survival time, which is defined as the time from the diagnosis of the first primary small intestine cancer to the diagnosis of the second primary cancer or death. Covariate information is also available on age at the first primary cancer diagnosis (age < 60 or ≥ 60), gender, race (white or other), stage (distant or other), and tumor grade (III or IV, or other). All these covariates are binary and have been previously shown to affect cancer development. Patients with survival times less than 6 months are considered as synchronous cases and are excluded from the study.

Li and Wu (2010) showed that the Cox model does not fit this dataset well as the proportionality assumption does not hold. We therefore fit a linear regression model to this dataset. Table 3 shows the estimates of regression coefficients. The variables *Surgery*, *race*, *stage*, and *grade* are selected and have nonzero coefficients by both penalized empirical likelihood (PEL) and *bujar*. The penalized empirical likelihood methods show that a white small intestine cancer patient who is at distant stage and grade I or II has longer survival time. Surgery will elongate small intestine patients' survival time. The variable *radiation* is selected by *bujar* only. We argue that the penalized empirical likelihood results are more reliable since *radiation* was found to have no significant impact on the survival time by Li and Wu (2010).

In conclusion, controlling for the other variables, surgery for first primary small intestine cancer appears to reduce the risk of second primary cancer or death significantly. Radiotherapy for first primary small intestine cancer does not seem to affect the development of second primary cancer.

5 Discussion

Our result is based on the assumption that the censored errors are i.i.d.. The method will fail, similar to the Buckley-James estimator, when the i.i.d. assumption does not hold. The

covariates dimension can grow with the sample size n . However, the converging rate could be slow for two reasons. First, the converging rate is slow in the empirical likelihood framework with diverging number of parameters, even for complete data, due to its nonparametric nature (Leng and Tang 2012). Second, the estimating equation is discontinuous and the estimated regression coefficients are an M-estimator with an unconventional form. It is also unclear whether the quadratic approximation developed in Parente and Smith (2010) continue to apply or not. The converging rate is likely to be slower than the situation considered in Leng and Tang (2012). We therefore do not discuss the case of diverging number of covariates in this paper, and hope to address the problem in a future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

Wu's research is supported in part by NSF Grant CCF-0926181. Li's research is supported in part by NIH grants CA016042, U54 RR031268-01, and P01AT003960. The authors thank the editor, associate editor, and referees for their helpful comments and suggestions

Appendix:: Regularity Conditions

Define $g_i(\beta) = \int w(t) \{X_i - \bar{X}(t)\} dN_i(t)$, $g(\beta) = E\{g_i(\beta)\}$, $G = E\{g_i(\beta)g_i(\beta)^T\}$, and $Q = E\{g_i(\beta)g_i(\beta)^T\}$. Our theoretical analysis assumes the following conditions:

- C.1** X_i has compact support with mean μ_X .
- C.2** F^β has finite Fisher information.
- C.3** $\beta \in \mathcal{B}$ where $\mathcal{B}$ is compact. The truth β_0 is in the interior of $\mathcal{B}$, and β_0 is the unique solution to $0 = g(\beta)$, and the matrix Q is nonsingular.
- C.4** $g(\beta)$ is differentiable at β_0 , the matrix G is of full rank. In addition let $v_n = \sqrt{\{\hat{g}(\beta) - g(\beta)\}}$ where $\hat{g}(\beta) = n^{-1} \sum_{i=1}^n g_i(\beta)$ then for any $\delta_n \rightarrow 0$

$$\sup_{\|\beta - \beta_0\| \leq \delta_n} \frac{\|v_n(\beta) - v_n(\beta_0)\|}{1 + \sqrt{n}\|\beta - \beta_0\|} \xrightarrow{p} 0.$$

Here conditions C.1 and C.2 are the ones from Ritov (1990) for analyzing right censored data, conditions C.3 is the technical device to ensure consistency of the root of $0 = \hat{g}(\beta)$, and C.4 is from Parente and Smith (2010) for handling nonsmooth function $g(\beta)$. Essentially, it is assumed that the expectation of the nonsmooth criterion function of the parameter is continuous and differentiable. The second part is required for establishing the asymptotic normality of $\hat{\beta}$, and a similar condition can be found in Ritov (1990).

References

- Buckley J and James L (1979), "Linear regression with censored data," *Biometrika*, 66, 429–436.
- Chen J and Chen Z (2008), "Extended Bayesian information criterion for model selection with large model space," *Biometrika*, 95, 759–771.
- Chen SX and Cui HJ (2006), "On Bartlett correction of empirical likelihood in the presence of nuisance parameters," *Biometrika*, 16, 1101–1115.
- Chen SX and Cui HJ (2007), "On the second order properties of empirical likelihood with moment restrictions," *Journal of Econometrics*, 141, 492–516.
- Chen SX, Peng L, and Qin YL (2009), "Effects of data dimension on empirical likelihood," *Biometrika*, 96, 711–722.
- Chen SX and Van Keilegom I (2009), "A review on empirical likelihood methods for regression (with discussions)," *Test*, 18, 415–447.
- Fan J and Li R (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.
- Fang KT, Li G, Lu X, and Qin H (2013), "An Empirical Likelihood Method for Semiparametric Linear Regression with Right Censored Data," *Computational and Mathematical Methods in Medicine*, 2013, Article ID 469373.
- Hashibe M, Ritz B, Le AD, Li G, Sankaranarayanan R, and Zhang Z-F (2005), "Radiotherapy for oral cancer as a risk factor for second primary cancers," *Cancer Letters*, 220, 185–195. [PubMed: 15766594]
- Hjort NL, McKeague I, and Van Keilegom I (2009), "Extending the scope of empirical likelihood," *The Annals of Statistics*, 37, 1079–1111.
- Hunter DR and Lange K (2004), "A tutorial on MM algorithms," *American Statistician*, 58, 30–37.
- Koul H, Susarla V, and Van Ryzin J (1981), "Regression analysis with randomly right-censored data," *Ann. Statist*, 9, 1276–1288.
- Lange K (2004), *Optimization*, New York: Springer-Verlag.
- Leng C and Tang CY (2012), "Penalized empirical likelihood and growing dimensional general estimating equations," *Biometrika*, 99, 703–716.
- Li G and Lu X (2009), "Comments on: A review on empirical likelihood methods for regression," *Test*, 18, 463–467. [PubMed: 21152356]
- Li G and Wang QH (2003), "Empirical likelihood regression analysis for right censored data," *Statistica Sinica*, 13, 51–68.
- Li G and Wu TT (2010), "Semiparametric additive risks regression for two-stage design survival studies," *Stat Sin*, 20, 1581–1607. [PubMed: 21931467]
- Lu W and Liang Y (2006), "Empirical likelihood inference for linear transformation models," *Journal of Multivariate Analysis*, 97, 1586–1599.
- Owen AB (1988), "Empirical likelihood ratio confidence intervals for a single function," *Biometrika*, 75, 237–249.
- Owen AB (2001), *Empirical Likelihood*, New York: Chapman and Hall/CRC.
- Parente PM and Smith RJ (2010), "GEL Methods For Nonsmooth Moment Indicators," *Econometric Theory*, 27, 74–113.
- Qin G and Jing BY (2001), "Empirical likelihood for censored linear regression," *Scandinavian Journal of Statistics*, 28, 661–673.
- Qin J and Lawless J (1994), "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300–325.
- Qin J and Lawless J (1995), "Estimating equations, empirical likelihood and constraints on parameters," *The Canadian Journal of Statistics*, 23, 145–159.
- Ritov Y (1990), "Estimation in a Linear Regression Model with Censored Data," *Ann. Statist*, 18, 303–328.
- Schwarz G (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

- SEER (1973–2002), “Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Public-Use Data,” National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch.
- Tang CY and Leng C (2010), “Penalized high dimensional empirical likelihood,” *Biometrika*, 97, 905–920.
- Tang CY and Wu TT (2013), “Nested Coordinate Descent Algorithms for Empirical Likelihood,” *Journal of Statistical Computation and Simulation*, in press.
- Terry SM and Santora T (2006), “Benign Neoplasm of the Small Intestine (<http://www.emedicine.com/med/topic2652.htm>),”
- Tsiatis AA (1990), “Estimating Regression Parameters Using Linear Rank Tests for Censored Data,” *Ann. Statist.*, 18, 354–372.
- Wang H, Li B, and Leng C (2009), “Shrinkage tuning parameter selection with a diverging number of parameters,” *Journal of the Royal Statistical Society, Series B*, 71, 671–683.
- Wang Z and Wang C (2010), “Buckley-James Boosting for Survival Analysis with High-Dimensional Biomarker Data,” *Statistical Applications in Genetics and Molecular Biology*, 9, 1–33.
- Wu TT and Lange K (2010), “The MM Alternative to EM,” *Statistical Science*, 25, 492–505.
- Zhao Y and Yang S (2012), “Empirical likelihood confidence intervals for regression parameters of the survival rate,” *Journal of Nonparametric Statistics*, 24, 59–70.
- Zheng M, Zhao Z, and Yu W (2012), “Empirical likelihood methods based on influence functions,” *Statistics and Its Interface*, 5, 355–366.
- Zhou M (2005), “Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model,” *Biometrika*, 92, 492–498.
- Zhou M and Li G (2008), “Empirical likelihood analysis of the Buckley-James estimator,” *Journal of Multivariate Analysis*, 99, 649–664. [PubMed: 19018294]

Figure 1:

QQ Plots of empirical likelihood ratios vs. χ^2_3 for $n = 100$. The first row is the EL ratios based on our proposed method and the second row is based on the Buckley-James (BJ) estimating equation. The first column is for the case when $\rho = 0$; the second column is for $\rho = 0.5$; and the third column is for $\rho = 0.9$.

Figure 2:

QQ Plot of quantiles of profiled empirical likelihood vs. χ^2_s for $n = 100$ based on 500

replicates. (a) $s = 1, \rho = 0$; (b) $s = 1, \rho = 0.5$; (c) $s = 1, \rho = 0.9$; (d) $s = 2, \rho = 0$; (e) $s = 2, \rho = 0.5$; (f) $s = 2, \rho = 0.9$; (g) $s = 3, \rho = 0$; (h) $s = 3, \rho = 0.5$; (i) $s = 3, \rho = 0.9$.

Table 1:

Coverage probabilities (CP) of profiled empirical likelihood based on 500 replicates

s	n	ρ	CP	n	ρ	CP	n	ρ	CP
1	100	0	0.966	200	0	0.966	500	0	0.956
		0.5	0.934		0.5	0.956		0.5	0.944
		0.9	0.920		0.9	0.936		0.9	0.940
2	100	0	0.936	200	0	0.966	500	0	0.966
		0.5	0.902		0.5	0.958		0.5	0.972
		0.9	0.924		0.9	0.958		0.9	0.942
3	100	0	0.920	200	0	0.938	500	0	0.966
		0.5	0.920		0.5	0.948		0.5	0.956
		0.9	0.892		0.9	0.936		0.9	0.952

Table 2:

Simulation results of our proposed penalized empirical likelihood (PEL) and Rpackage *bujar* for penalized linear regression based on 100 replicates.

(n, p)	ρ	Method	N_{nonzero}	N_{true}	ME
(100, 10)	0.5	PEL-BIC	3.59 (0.08)	3 (0)	0.212 (0.023)
		PEL-BICC	3.59 (0.08)	3 (0)	0.212 (0.023)
		PEL-EBIC	3.30 (0.06)	3 (0)	0.188 (0.022)
		bujar	4.47 (0.10)	3 (0)	0.281 (0.015)
		Oracle			0.067 (0.005)
(200, 10)	0.5	PEL-BIC	3.09 (0.03)	3 (0)	0.038 (0.004)
		PEL-BICC	3.09 (0.03)	3 (0)	0.038 (0.004)
		PEL-EBIC	3.05 (0.02)	3 (0)	0.038 (0.004)
		bujar	4.08 (0.10)	3 (0)	0.189 (0.010)
		Oracle			0.028 (0.003)
(200, 20)	0.5	PEL-BIC	3.66 (0.1)	3 (0)	0.061 (0.005)
		PEL-BICC	3.64 (0.1)	3 (0)	0.061 (0.005)
		PEL-EBIC	3.24 (0.06)	3 (0)	0.067 (0.008)
		bujar	4.97 (0.13)	3 (0)	0.203 (0.010)
		Oracle			0.025 (0.002)

Table 3:

Result of SEER small intestine data set with 7 covariates using unpenalized empirical likelihood, penalized empirical likelihood (PEL) and R package *bujar*

Variable	Unpenalized	PEL-BIC	PEL-BICC	PEL-EBIC	bujar
surgery	0.54	2.33	2.33	2.33	1.87
radiation	0.1	0	0	0	-0.07
age60	-0.28	0	0	0	0
male	0.08	0	0	0	0
white	1.36	2.99	2.99	2.99	2.63
stage-distant	1.62	2.92	2.92	2.92	2.53
grade34	-0.57	-1.94	-1.94	-1.94	-1.48

Table 4:

Selection frequencies based on 100 random subsets of 500 out of 2669 SEER small intestine samples

Variable	PEL-BIC	PEL-BICC	PEL-EBIC	bujar
surgery	80	80	53	100
radiation	4	4	2	37
age60	9	9	6	29
male	6	6	3	15
white	99	99	94	100
stage-distant	100	100	99	100
grade34	83	83	54	99