Empirical Likelihood Method for Ratio Estimation

By

Bin Dong

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics - Biostatistics

Waterloo, Ontario, Canada, 2011

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Empirical likelihood, which was pioneered by Thomas and Grunkemeier (1975) and Owen (1988), is a powerful nonparametric method of statistical inference that has been widely used in the statistical literature. In this thesis, we investigate the merits of empirical likelihood for various problems arising in ratio estimation. First, motivated by the smooth empirical likelihood (SEL) approach proposed by Zhou & Jing (2003), we develop empirical likelihood estimators for diagnostic test likelihood ratios (DLRs), and derive the asymptotic distributions for suitable likelihood ratio statistics under certain regularity conditions. To skirt the bandwidth selection problem that arises in smooth estimation, we propose an empirical likelihood estimator for the same DLRs that is based on non-smooth estimating equations (NEL). Via simulation studies, we compare the statistical properties of these empirical likelihood estimators (SEL, NEL) to certain natural competitors, and identify situations in which SEL and NEL provide superior estimation capabilities.

Next, we focus on deriving an empirical likelihood estimator of a baseline cumulative hazard ratio with respect to covariate adjustments under two nonproportional hazard model assumptions. Under typical regularity conditions, we show that suitable empirical likelihood ratio statistics each converge in distribution to a χ^2 random variable. Through simulation studies, we investigate the advantages of this empirical likelihood approach compared to use of the usual normal approximation. Two examples from previously published clinical studies illustrate the use of the empirical likelihood methods we have described.

Empirical likelihood has obvious appeal in deriving point and interval estimators for time-to-event data. However, when we use this method and its asymptotic critical value to construct simultaneous confidence bands for survival or cumulative hazard functions, it typically necessitates very large sample sizes to achieve reliable coverage accuracy. We propose using a bootstrap method to recalibrate the critical value of the sampling distribution of the sample log-likelihood ratios. Via simulation studies, we compare our EL-based bootstrap estimator for the survival function with EL-HW and EL-EP bands proposed by Hollander *et al.* (1997) and apply this method to obtain a simultaneous confidence band for the cumulative hazard ratios in the two clinical studies that we mentioned above.

While copulas have been a popular statistical tool for modeling dependent data in recent years, selecting a parametric copula is a nontrivial task that may lead to model misspecification because different copula families involve different correlation structures. This observation motivates us to use empirical likelihood to estimate a copula nonparametrically. With this EL-based estimator of a copula, we derive a goodness-of-fit test for assessing a specific parametric copula model. By means of simulations, we demonstrate the merits of our EL-based testing procedure. We demonstrate this method using the data from Wieand *et al.* (1989).

In the final chapter of the thesis, we provide a brief introduction to several areas for future research involving the empirical likelihood approach.

Acknowledgements

I would like to thank all the people who made this possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The likelihood method is one of the most powerful tools in statistical inference. For parametric models, Wilks (1938) showed that under suitable regularity conditions the likelihood ratio statistic converges in distribution to a chi-squared random variable as the sample size, n, increases. Therefore, we can use the likelihood ratio statistic to test hypotheses and construct confidence intervals in parametric model settings. However, when the underlying probability model is misspecified, the maximium likelihood estimator (MLE) obtained from parametric likelihood can be biased and inefficient. Thus, as an alternative, statistical researchers have explored using the principles of the likelihood method in nonparametric contexts.

Empirical likelihood is a nonparametric method which was first described by Thomas and Grunkemeier (1975). In that pioneering paper, they employed a nonparametric likelihood ratio idea to construct pointwise confidence intervals for the survival function. Subsequently, their idea was extended by Owen (1988), who proposed the method of empirical likelihood for estimating a univariate mean and various other statistics. Since then, empirical likelihood has been widely applied to numerous problems in statistical inference; see Owen (2001) for details.

Unlike its parametric counterpart, the empirical likelihood method does not assume the data come from a known family of distributions. Therefore, it avoids the model misspecification problem that confronts parametric analysis. Instead this empirical method of inference defines the likelihood to be the product of probability masses at observed data points, $\prod_i P(X_i)$. Therefore, by finding the nonparametric maximum likelihood estimator, which consists of the point masses that maximize the empirical likelihood function, we can define the analogue, for empirical likelihood, of the likelihood ratio statistic. As Owen (1988) demonstrated, the empirical likelihood ratio statistics for various parameters, $\theta(F)$, of an unknown distribution function F each have an asymptotic χ^2 distribution under certain regularity conditions. Consequently, we can use the empirical likelihood ratio statistic to carry out statistical inference in a way that is completely analogous to using the parametric likelihood ratio statistic in the parametric setting.

Since empirical likelihood (EL) makes use of the flexibility and effectiveness of the likelihood method, its approach to the problem of estimation has many unique properties such as range preserving, data-determined asymmetric confidence interval, Bartlett correctable, better coverage probability for small samples compared to alternative estimators based on other nonparametric methods. As Owen (2001) demonstrates, EL may easily incorporate known constraints on parameters, and adjust for biased sampling schemes. It is also easier to combine data from multiple sources, with possibly different distributions. A further advantage of EL is that it can be combined with estimating equations to obtain a more efficient estimator. Therefore, the EL method has been extensively used not only for complete data but also for censored and truncated data.

Before presenting my work in empirical likelihood for ratio estimation for complete and right-censored data, I will first provide a summary of some key results in empirical likelihood theory.

1.1 Key Results in Empirical Likelihood

Definition 1. *Let $X_1, ..., X_n \in \mathbb{R}$. The empirical cumulative distribution function (ECDF) of $X_1, ..., X_n$ is*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x} \ , \tag{1.1}$$

for $-\infty < x < \infty$.

Theorem 1. *Let $X_1, ..., X_n$ be i.i.d random variables with a common cumulative distribution function (CDF) F_0. The nonparametric likelihood of the CDF F,*

$$L(F) = \prod_{i=1}^{n} \{F(X_i) - F(X_i-)\}, \tag{1.2}$$

is maximized by the ECDF of $X_1, ..., X_n$.

Proof. See Theorem(2.1) in Owen (2001). □

Definition 2. *Let $T_1, ..., T_n$ be i.i.d lifetimes with CDF $F(t) = P(T_i < t)$. Let $C_1, ..., C_n$ be censoring times with CDF $G(t) = P(C_i < t)$. Assume, further, that the lifetimes and the censoring times are independent. Under the random censorship model, we observe only $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i < C_i), i = 1, ..., n$. The EL of F*

is

$$EL(F) = \prod_{i=1}^{n} [\Delta F(X_i)]^{\delta_i} [1 - F(X_i)]^{1-\delta_i}, \tag{1.3}$$

where $\Delta F(X_i) = F(X_i) - F(X_i^-)$.

Definition 3. *Let* $\Delta\Lambda(t) = \frac{\Delta F(t)}{1-F(t^-)}$ *be the hazard function for the CDF* $F(t)$*, with* $\Lambda(t)$ *as the corresponding cumulative hazard function. The EL of* Λ *is*

$$EL(\Lambda) = \prod_{i=1}^{n} [\Delta\Lambda(X_i)]^{\delta_i} \exp\{-\Lambda(X_i)\}. \tag{1.4}$$

Note that the expression (1.4) is not the exact likelihood function of Λ, but a Poisson extension of the exact likelihood function; see Murphy (1995) for the details. It can be shown that among all cumulative distribution functions, the Kaplan-Meier estimator maximizes the empirical likelihood in expression (1.3), and the Nelson-Aalen estimator is the nonparametric maximum likelihood estimator (NPMLE) of Λ in expression (1.4).

Definition 4. *For a distribution function* F*, let* F_n *be the NPMLE for* F_0*, the true distribution function. We define the empirical likelihood ratio to be*

$$R(F) = \frac{L(F)}{L(F_n)}, \tag{1.5}$$

for $F \in \Gamma$*, a set of all distribution functions in* $\Re$.

Definition 5. *Suppose that we are interested in a parameter* $\theta = T(F)$ *for some function* T *of distributions. The profile empirical likelihood ratio function of* θ *is*

$$R(\theta) = \sup\{R(F) | T(F) = \theta, F \in \Gamma\}. \tag{1.6}$$

For example, if we are interested in estimating μ, the population mean for a single-sample inference problem, using only probability distributions w_i with $\sum_{i=1}^n w_i = 1$, the profile empirical likelihood ratio function for μ is

$$R(\mu) = \max\left\{\prod_{i=1}^n nw_i | \sum_{i=1}^n w_i X_i = \mu, w_i \geq 0, \sum_{i=1}^n w_i = 1\right\}$$

Theorem 2. *Let $X_1, ..., X_n$ be i.i.d random variables with distribution function F_0. Let $\mu_0 = E(X_i)$, and suppose that $Var(X_i) < \infty$. Then $-2\log(R(\mu_0))$ converges in distribution to χ_1^2 as $n \to \infty$.*

Proof. See Theorem (2.2) in Owen (2001) □

Therefore, the corresponding $100(1-\alpha)\%$ empirical likelihood confidence region for μ is

$$\begin{aligned}&\{\mu| -2\log(R(\mu)) \leq q_{1-\alpha}\}\\&= \left\{\textstyle\sum_{i=1}^n w_i X_i | -2\sum_{i=1}^n \log(nw_i) \leq q_{1-\alpha}, w_i \geq 0, \sum_{i=1}^n w_i = 1\right\}.\end{aligned}$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the χ_1^2 distribution.

Theorem 3. *For i.i.d random vectors $X_1, ..., X_n$ in $\Re^d$ with mean μ_0, we can similarly define the empirical likelihood ratio function $R(\mu)$ for the multivariate mean and the corresponding confidence region. Provided $X_1, ..., X_n$ have a finite variance-covariance matrix with rank $q > 0$, $-2\log R(\mu_0)$ converges in distribution to a χ_q^2 random variable as $n \to \infty$.*

Proof. See Theorem (3.2) in Owen (2001). □

Since estimating equations are widely used for estimating population parameters and deriving the corresponding statistics, we next consider combining empirical likelihood with estimating equations.

Let $m(X, \theta) = 0$ be an estimating equation for θ_0. Define

$$R(\theta) = \max\left\{ \prod_{i=1}^{n} nw_i | \sum_{i=1}^{n} w_i m(X_i, \theta) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\}$$

to be the empirical likelihood ratio function for θ.

Theorem 4. *Suppose $Var\{m(X_i, \theta_0)\}$ is finite with rank $q > 0$. If θ_0 satisfies $E\{m(X, \theta_0)\} = 0$, then $-2 \log R(\theta_0) \xrightarrow{D} \chi^2_q$ as $n \to \infty$.*

Proof. See Theorem (3.4) in Owen (2001). □

Now consider the empirical likelihood for the cumulative hazard function under the constraint $\int g_n(t, \theta) d\Lambda(t) = 0$, where $g_n(t)$ is a stochastic function and θ is the parameter of interest.

Theorem 5. *Let $T_1, ..., T_n$ be i.i.d lifetimes with CDF $F(t) = P(T_i < t)$, and $C_1, ..., C_n$ be censoring times with CDF $G(t) = P(C_i < t)$ as described in Definition 2. Suppose that $g_n(t)$ is a sequence of predictable functions with respect to the filtration F_t, and $g_n(t) \xrightarrow{P} g(t)$ with*

$$0 < \int \frac{|g(x)|^m d\Lambda(x)}{(1 - F(x))(1 - G(x))} < \infty, m = 1, 2.$$

Let $\hat{\Lambda}_n(t)$ be the Nelson-Aalen estimator of $\Lambda(t)$. Then

$$-2\log\frac{\sup_{\Lambda} EL(\Lambda)}{EL(\hat{\Lambda}_n(t))} \xrightarrow{D} \chi_1^2 \qquad as \;\; n \to \infty.$$

Proof. See Theorem 2 in Pan and Zhou (2002). □

The reminder of this thesis is organized as follows. In chapter 2, we propose empirical likelihood estimators for diagnostic test likelihood ratios, and obtain the asymptotic distributions for the corresponding likelihood ratio statistics under certain regularity conditions. Using simulation studies, we also compare the statistical properties of these EL estimators to certain natural competitors. In chapter 3 we derive an empirical likelihood estimator of a baseline cumulative hazard ratio with respect to covariate adjustments under two nonproportional hazard model assumptions. We show that the empirical likelihood ratio statistics each converge in distribution to a χ^2 random variable under suitable regularity conditions. Via simulation, we explore the advantages of this empirical likelihood approach compared to the usual normal approximation to this problem in statistical inference. We investigate use of the bootstrap to estimate simultaneous confidence bands for the survival and cumulative hazard functions in chapter 4. By comparing our EL-based bootstrap with several natural estimator competitors in terms of coverage probabilities at the nominal level of 95% in a simulation study, we discover the merits of our method, especially when the sample sizes are small. We apply this bootstrap method to the two real datasets and obtain the simultaneous confidence band of the adjusted cumulative hazard ratios for all t in a time interval of interest. In chapter 5 we derive an empirical likelihood-based estimator for two-dimensional copulas. Using this EL-based estimator we are able to develop a goodness-of-fit test to check the suitability of a parametric model of interest. In the final chapter, we outline some avenues for future research involving the empirical likelihood method.

Chapter 2

Empirical Likelihood for Diagnostic Test Likelihood Ratios

2.1 Introduction

Diagnostic test likelihood ratios (DLRs), which are important characteristics used to interpret a diagnostic test outcome, have been reported in the clinical and epidemiologic literature for several decades. These ratios provide valuable information about the predictive properties of a diagnostic test, while having the attractive feature of being independent of the prevalence of disease in the study population.

For any diagnostic test, we assume there are two subgroups in the study population, the disease-free and diseased individuals, respectively. The diagnostic test likelihood ratios (DLRs) that correspond to positive and negative outcomes of a given test are

$$\rho_+ = \frac{\Pr(\text{positive test outcome}|\text{diseased})}{\Pr(\text{positive test outcome}|\text{disease-free})} = \frac{\text{sensitivity}}{\text{1-specificity}} \ ,$$

and

$$\rho_- = \frac{\text{Pr(negative test outcome}|\text{diseased)}}{\text{Pr(negative test outcome}|\text{disease-free)}} = \frac{\text{1-sensitivity}}{\text{specificity}} \ .$$

DLRs are ratios of conditional probabilities that we can use to calculate the posterior odds in favour of disease, given the actual test result and the prior odds. A value of ρ_+ greater than one indicates the degree to which disease is more likely given a positive test result. A value of ρ_- that belongs to the interval $(0, 1)$ indicates that the patient is less likely to have disease if a negative test result is observed.

For a binary diagnostic test we assume two subpopulations, the disease-free and diseased groups, labeled 1 and 2, respectively. Let X_i represent the number of positive diagnostic test results observed in the n_i members of group $i, (i = 1, 2)$, Let $1 - p_1$ be the test specificity and p_2 be the sensitivity; then

$$\rho_+ = p_2/p_1 \quad \text{and} \quad \rho_- = (1 - p_2)/(1 - p_1).$$

Since

$$p_1 = (1 - \rho_-)/(\rho_+ - \rho_-), \qquad p_2 = \rho_+(1 - \rho_-)/(\rho_+ - \rho_-)$$

is a 1-1 transformation for $\rho_+ > 1$ and $0 < \rho_- < 1$, the corresponding log-likelihood function is

$$\begin{aligned} l(\rho_+, \rho_-) = \quad & x_2 \log \rho_+ + (x_1 + x_2) \log(1 - \rho_-) - (n_1 + n_2) \log(\rho_+ - \rho_-) \\ & + (n_2 - x_2) \log \rho_- + (n_1 + n_2 - x_1 - x_2) \log(\rho_+ - 1) \end{aligned}$$

from which we can obtain MLEs and the information matrix. Furthermore, we can use score, Wald, or likelihood ratio statistics to obtain marginal confidence intervals or a joint confidence region for ρ_+ and ρ_-.

When there are covariates that may influence the accuracy of the diagnostic test, Leisenring and Pepe (1998) proposed a regression method that allows for direct assessment of covariate effects on DLRs for binary diagnostic tests. They used the GEE method to estimate the regression coefficients even for clustered or unbalanced data. However their method does not accommodate continuous test results which also commonly arise in practice.

For continuous-scale diagnostic tests we can use parametric, semi-parametric, and nonparametric methods to estimate the DLRs. But when the model is misspecified, these parametric and semi-parametric estimators can be biased and inefficient. Therefore, we consider using a nonparametric method to estimate DLRs. Since the DLRs, regarded as functions of the cdfs from the disease-free and diseased groups may be smooth, kernel estimation is a natural nonparametric method to consider. In a study of the receiver operating characteristic (ROC) curve, which is also a function of the two cdfs from disease-free and diseased groups, Lloyd and Yong (1999) showed that the kernel estimator for ROC has smaller mean squared errors than the empirical estimator. This result encouraged us to use a kernel method to estimate the DLRs. Also, Claeskens *et al.*(2003) described a smooth empirical likelihood method based on kernel estimating equations to obtain an estimator of ROC that retains a high degree of efficiency and coverage accuracy compared to other nonparametric estimators. This motivated us to consider adapting their method to the problem of estimating DLRs.

Smooth empirical likelihood is a nonparametric method that combines the empirical likelihood function (see Owen, 2001) with kernel estimating equations. It was first proposed by Zhou and Jing (2003) for estimating differences of quantiles, and also advocated by other authors. Chen, Peng and Zhao (2009) applied this approach to copulas. We adapted the smooth empirical likelihood (SEL) method to estimate DLRs and obtain the SEL estimators and their corresponding asymptotic distributions. From simulation studies we found that the SEL estimator is more efficient than both the kernel and empirical estimators, and the SEL interval estimate has higher coverage accuracy than its kernel and empirical competitors. However since the SEL method involves selecting a suitable bandwidth, which is often a challenging problem, the SEL method has an unavoidable drawback that may prevent its application in some situations.

To skirt this bandwidth selection problem we next consider combining the empirical likelihood function with non-smooth estimating equations. Since the empirical likelihood method involves a constrained maximization problem, instead of obtaining nuisance parameter estimates from estimating equations as regular SEL does, we derived the profile log-likelihood function by solving its dual optimization problem, which does not need smooth estimating equations. Under certain regularity conditions, we showed that the empirical log-likelihood ratio statistic converges in distribution to a χ_1^2 random variable. A second simulation study demonstrated that the non-smooth empirical likelihood (NEL) estimator outperforms the corresponding empirical estimator in term of having smaller coverage error, and hence higher coverage accuracy, especially when the sample sizes are small.

Fan, Huang and Wong (2000) show that the empirical log-likelihood ratio statistic has a chi-squared limiting distribution only if the limiting distributions of any nuisance parameter and NEL estimators exist. Using the corollary of Pakes and Pollard (1989) we showed that the nuisance parameters and NEL estimator have asymptotic normal distributions, and this justifies the NEL method. Using this method, we can adopt appropriate estimating functions, such as indicator or quantile functions, without worrying about their smoothness. This extends the application of the empirical likelihood method.

2.2 A Smooth Empirical Likelihood Estimator

2.2.1 Notation and Definitions

Suppose that $X_{11}, ..., X_{1n_1}$ and $X_{21}, ..., X_{2n_2}$ are independent random samples from the disease-free and diseased populations with distribution functions F_1 and F_2 respectively. Let $G_{h_1}(t)$ and $G_{h_2}(t)$ be kernel estimators for F_1 and F_2 with corresponding bandwidths h_1 and h_2, where $h_j = h_j(n_j) \to 0$ as $n_j \to \infty$ for $j = 1, 2$.

Without loss of generality we only consider the positive DLR, which we denote by θ. In the spirit of a binary diagnostic test, θ corresponds to the ratio of test sensitivity to the false positive rate, i.e., $1-$ specificity. In order to construct a smooth empirical estimator of θ, we define $p = (p_1, ..., p_{n_1})$ and $q = (q_1, ..., q_{n_2})$ to be two probability vectors with $\sum_{i=1}^{n_1} p_i = 1$ and $\sum_{j=1}^{n_2} q_j = 1$. Let

$$\hat{F}_{h_1,p}(\eta) = \sum_{i=1}^{n_1} p_i G_{h_1}(\eta - X_{1i}) \quad \text{and} \quad \hat{F}_{h_2,q}(\eta) = \sum_{j=1}^{n_2} q_j G_{h_2}(\eta - X_{2j}).$$

The smooth empirical likelihood of θ is

$$L(\theta) = \sup_{(p,q,\eta)} (\prod_{i=1}^{n_1} p_i)(\prod_{j=1}^{n_2} q_j), \tag{2.1}$$

subject to the following constraints:

$$\hat{F}_{h_1,p}(\eta) = \sum_{i=1}^{n_1} p_i G_{h_1}(\eta - X_{1i}) = 1 - t, \quad 0 \leq t \leq 1 \tag{2.2}$$

and

$$\hat{F}_{h_2,q}(\eta) = \sum_{j=1}^{n_2} q_j G_{h_2}(\eta - X_{2j}) = 1 - \theta t, \quad \theta > 0. \tag{2.3}$$

In constraints (2.2) and (2.3), the parameter η represents the fixed threshold that separates a positive diagnostic test outcome from its negative counterpart, and t denotes the corresponding false positive rate of the test in the disease-free population.

Using Lagrange multipliers, the log-likelihood function under constraints (2.2) and (2.3), as well as $\sum_{i=1}^{n_1} p_i = 1$ and $\sum_{j=1}^{n_2} q_j = 1$, is

$$l(\theta) = \sum_{i=1}^{n_1} \log(p_i) + \sum_{j=1}^{n_2} \log(q_j) + n_1\lambda_1\{(1-t) - \sum_{i=1}^{n_1} p_i G_{h_1}(\eta - X_{1i})\} +$$

$$n_2\lambda_2\{(1-\theta t) - \sum_{j=1}^{n_2} q_j G_{h_2}(\eta - X_{2j})\} + \lambda_3(1 - \sum_{i=1}^{n_1} p_i) + \lambda_4(1 - \sum_{j=1}^{n_2} q_j).$$

Set $\partial l(\theta)/\partial p_i = 0$ and $\partial l(\theta)/\partial q_j = 0$; we get $\lambda_3 = n_1 - n_1\lambda_1(1-t)$, $\lambda_4 = n_2 - n_2\lambda_2(1-\theta t)$, and

$$p_i = \frac{1}{n_1\{1 + \lambda_1 w_1(\eta, X_{1i})\}}, \;\; i = 1, ..., n_1,$$

$$q_j = \frac{1}{n_2\{1 + \lambda_2 w_2(\eta, X_{2j})\}}, \;\; j = 1, ..., n_2,$$

where

$$w_1(\eta, X_{1i}) = G_{h_1}(\eta - X_{1i}) - (1 - t) \text{ and } \;\; w_2(\eta, X_{2j}) = G_{h_2}(\eta - X_{2j}) - (1 - \theta t).$$

Then

$$g_1(\eta, \theta) = \sum_{i=1}^{n_1} \frac{w_1(\eta, X_{1i})}{n_1\{1 + \lambda_1 w_1(\eta, X_{1i})\}} = 0, \tag{2.4}$$

$$g_2(\eta, \theta) = \sum_{j=1}^{n_2} \frac{w_2(\eta, X_{2j})}{n_2\{1 + \lambda_2 w_2(\eta, X_{2j})\}} = 0, \tag{2.5}$$

are estimating equations for θ.

In order to obtain the smooth empirical estimators of θ, we first need to find λ_1 and λ_2 that satisfy the equations

$$\partial l(\theta)/\partial \lambda_j = \sum_{i}^{n_j} \frac{w_j(\eta, X_{ji})}{1 + \lambda_j w_j(\eta, X_{ji})} = 0, \;\; j = 1, 2, \tag{2.6}$$

and the constraints $1 + \lambda_j w_j(\eta, X_{ji}) > 1/n_j$, for $j = 1, 2$, which come from the probability requirements, $0 \le p_i \le 1$ and $0 \le q_j \le 1$, for $i = 1, ..., n_1, j = 1, ..., n_2$. Following Owen (2001), the constraint equations for λ_1 and λ_2 can be solved via the dual problem of globally minimizing

$$L_*(\lambda_j) = -\sum_{i=1}^{n_j} \log_*\{1 + \lambda_j w_j(\eta, X_{ji})\}, \quad j = 1, 2. \tag{2.7}$$

where

$$\log_*(z) = \begin{cases} \log(z), & \text{if } z \geq 1/n, \\ \log(1/n) - 1.5 + 2nz - (nz)^2/2, & \text{if } z < 1/n. \end{cases}$$

Therefore, for a given value of η we can obtain $\lambda_j = \lambda_j(\eta),\ j = 1, 2$. But η is also a nuisance parameter. We can eliminate it by $l_*(\theta) = \max_\eta \min_\lambda \{\sum_{j=1}^2 L_*(\lambda_j, \theta, \eta)\}$ to obtain the profile empirical log-likelihood of θ. Define the empirical log-likelihood ratio as

$$l_n(\theta) = -2\{\sum_{i=1}^{n_1} \log(n_1 p_i) + \sum_{j=1}^{n_2} \log(n_2 q_j)\}.$$

Then

$$l_n(\theta) = 2 \sum_{j=1}^{2} \sum_{i=1}^{n_j} \log\{1 + \lambda_j w_j(\eta, X_{ji})\},$$

is the profile empirical log-likelihood ratio statistic, and $\hat{\theta} = \arg\min_\theta l_n(\theta)$ is the smooth empirical likelihood estimator of θ.

2.2.2 Point Estimation and Confidence Intervals

We adopt the same conditions (C1-C4) that were identified by Claeskens *et al.* (2003). That is, for $j = 1, 2$ we assume that:

(C1) The density function f_j is r smooth in a neighbourhood of η, i.e., there exists an integer $r \geq 2$ such that $f_j^{(r-1)}$ exists in the neighbourhood of η. Also, f_j is continuous at η and $f_1(\eta) f_2(\eta) > 0$.

(C2) As $\min(n_1, n_2) \to \infty, n_j/(n_1 + n_2) \to \gamma_j$, where $0 < \gamma_j < 1$.

(C3) The kernel K_j is an rth-order ($r \geq 2$) kernel satisfying

$$\int s^k K_j(s)ds = \begin{cases} 1, & \text{if } k = 0, \\ 0, & \text{if } 1 \le k \le r-1, \\ c \ne 0, & \text{if } k = r. \end{cases}$$

(C4) For $j = 1, 2, n_j h_j^{4r} \to 0$, and $n_j h_j^{2r} / \log n_j \to \infty$ as $n_j \to \infty$.

Note that the smoothness requirements in condition (C1) come from the kernel estimating equations, and $f_1(\eta) f_2(\eta) > 0$ guarantees the asymptotic variance of the estimator has correct order. Condition (C2) requires that the growth rate of the two sample sizes be balancing, i.e., one sample size cannot grow too fast to dominate the other sample size. Condition (C3) gives the form of kernel functions that are usually used in nonparametric estimation of densities. Finally, condition (C4) assures the convergence rate of estimates of any nuisance parameter in estimating equations as well as the convergence rate of the profile empirical log-likelihood ratio.

Theorem 1. *Assume that conditions (C1)-(C3) hold, for fixed t, $0 \le t \le 1$. If $n_2 h_2^{2r} \to 0$, the smooth empirical likelihood estimator $\hat{\theta}$ satisfies*

$$\sqrt{n_2}\{\hat{\theta}(t) - \theta(t)\} \xrightarrow{D} N[0, \{\theta(1-\theta t)/t + n_2 f_2^2(\eta)\, Var(\tilde{\eta})/t^2\}], \tag{2.8}$$

where $Var(\tilde{\eta}) = \frac{\theta(1-\theta t)t(1-t)}{n_1 f_1^2(\eta)\theta(1-\theta t) + n_2 f_2^2(\eta)(1-t)}$.

Proof. By Qin and Lawless (1994), the log-likelihood of θ under constraints (2.2) and (2.3) acquires its maximum value at $\hat{\theta}$ in the neighborhood of radius $n^{-1/3}$ of θ. Since $\eta = F_1^{-1}(1-t) = F_2^{-1}(1-\theta t)$, in order to obtain the asymptotic result for $\hat{\theta}(t)$

we first consider the bias and variance of $\tilde{\eta}$, where $\tilde{\eta} = \arg\max_{\eta} \min_{\lambda}\{\sum_{j=1}^{2} L_*(\lambda_j, \theta, \eta)\}$. Based on Lemma 3 in Claeskens *et al.* (2003) we have

$$\begin{aligned} 0 = E[g_2(\theta)] &= E\left[\sum_{i=1}^{n_2} \frac{w_2(\tilde{\eta}, X_{2i}, \theta)}{n_2\{1 + \lambda_2(\tilde{\eta}) w_2(\tilde{\eta}, X_{2i}, \theta)\}}\right] \\ &\simeq E[w_2(\tilde{\eta}, X_{21}, \theta)] = F_2(\tilde{\eta}) - F_2(\eta) + o(h_2^r), \end{aligned}$$

so that $E(\tilde{\eta}) - \eta = o(h_2^r)$.

Similarly, using the approach followed by Claeskens *et al.* (2003) we obtain

$$\mathrm{Var}(\tilde{\eta}) = \frac{\theta(1-\theta t)t(1-t)}{n_1 f_1^2(\eta)\theta(1-\theta t) + n_2 f_2^2(\eta)(1-t)}$$

Now, consider a Taylor expansion of $g_2(\theta)$ around $\hat{\theta}$; this yields

$$\begin{aligned} g_2(\theta) &\simeq \sum_{i=1}^{n_2}\left\{\frac{w_2(\hat{\eta}, X_{2i}, \hat{\theta})}{n_2\{1 + \lambda_2(\hat{\eta}) w_2(\hat{\eta}, X_{2i}, \hat{\theta})\}} + \frac{(\theta - \hat{\theta})t}{n_2\{1 + \lambda_2(\hat{\eta}) w_2(\hat{\eta}, X_{2i}, \hat{\theta})\}^2}\right\} \\ &= 0 + \sum_{i=1}^{n_2} \frac{(\theta - \hat{\theta})t}{n_2\{1 + \lambda_2(\hat{\eta}) w_2(\hat{\eta}, X_{2i}, \hat{\theta})\}^2} \simeq (\hat{\theta} - \theta)t \end{aligned}$$

Therefore, as $n_2 h_2^{2r} \to 0$, $\hat{\theta}$ has an asymptotic normal distribution with mean θ. Using another Taylor expansion and some results from Claeskens *et al.* (2003), for $n_i \to \infty, (i = 1, 2)$ we can show that

$$\begin{aligned} E(g_2^2) &= \tfrac{1}{n_2^2} E[\textstyle\sum_{j=1}^{n_2} w_2(\tilde{\eta}, X_{2j}, \theta)^2] + f_2^2(\eta)\mathrm{Var}(\tilde{\eta}) \\ &= F_2(\eta)(1 - F_2(\eta))/n_2 + f_2^2(\eta)\mathrm{Var}(\tilde{\eta}). \end{aligned}$$

Therefore, the asymptotic variance of the smooth empirical likelihood estimator is

$AVar(\hat{\theta}) = E(g_2^2)/t^2 = \theta(1-\theta t)/n_2 t + f_2^2(\eta)\text{Var}(\tilde{\eta})/t^2$ □

Theorem 2. *Under conditions (C1)-(C4), the smooth empirical log-likelihood ratio* $l_n(\theta) \xrightarrow{D} \chi_1^2$

Proof of the theorem is similar to that provided by Claeskens *et al.* (2003) so we do not include it here. Theorem 2 is a smooth nonparametric version of Wilks' theorem for a DLR. Based on this asymptotic result, we can construct a $100(1-\alpha)\%$ confidence interval for the smooth empirical likelihood estimator as we show in the next corollary.

Let $A_{q_{1-\alpha}} = \{\theta : l_n(\theta) \leq q_{1-\alpha}\}$, where $q_{1-\alpha}$ is the $1-\alpha$ quantile of a χ_1^2 distribution.

Corollary 1. *Under conditions (C1)-(C4),*

$$P(\theta \in A_{q_{1-\alpha}}) = 1 - \alpha + o(1).$$

2.2.3 Simulation Study

In order to compare the accuracy and coverage probability of the smooth empirical likelihood estimator with its natural kernel and empirical competitors, we generated pseudorandom samples with various sample sizes from known distributions $F_j, j = 1, 2$, for the disease-free and diseased populations. We computed the smooth empirical likelihood estimator of DLR, $\hat{\theta}(t)$, for $0 < t < 1$, the false positive probability in the disease-free population.

Table 2.1: Comparison of the mean squared errors of SEL, KE, EE

The estimated mean squared errors of the smooth empirical likelihood (SEL), kernel (KE) and empirical estimators (EE) arising from different sample sizes for disease-free (n_1) and diseased (n_2) groups.

n_1	n_2	SEL	KE	EE
50	50	0.140	0.161	1.923
50	90	0.659	0.676	0.675
90	50	0.623	0.637	1.522
100	100	0.0027	0.0029	0.328

First we generated pseudorandom samples from $F_1 \sim N(6, 2)$ and $F_2 \sim N(10, 4)$ with different sample sizes. We used Gaussian kernels and the bandwidth function `bw.nrd0` in `R` in all sample settings except for the case of sample sizes $n_1 = 50, n_2 = 90$ for which the method of Sheather and Jones (1991) was used to select bandwidths. For $t_i = i/100, i = 1, 2, ..., 99$ we calculated the mean squared errors, based on the same two samples, for the SEL, kernel and empirical estimators. The study results are given in Table 2.1.

From Table 2.1 we observe that the smooth empirical likelihood estimator outperforms the kernel and empirical estimators in all cases. The relative gain in accuracy of the smooth empirical likelihood estimator compared to the kernel estimator is smaller than the corresponding relative gain with respect to the empirical estimator except for the case of $n_1 = 50, n_2 = 90$. However, the performance of the smooth empirical likelihood estimator depends on the choice of bandwidths. In the simulation study we selected Gaussian bandwidths, h_1 and h_2, that were of $O(n^{-1/5})$, which satisfies conditions (C1)-(C4).

To gain a visual impression of the smooth empirical estimator of DLR, we plot the smooth empirical likelihood estimator vs t for normal samples with sample sizes (n_1, n_2) equal to $(50, 50)$, $(50, 90)$ in Figures 2.1 and 2.2 respectively. In Figure 2.3

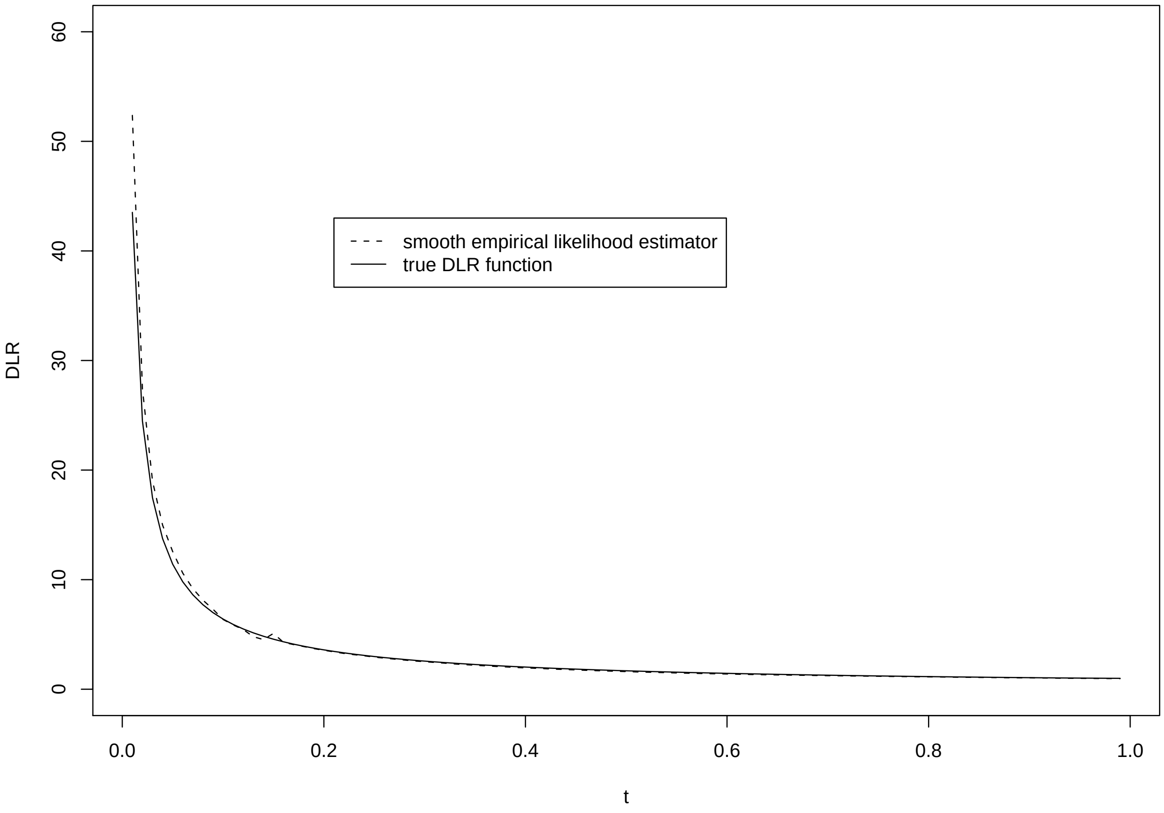

Figure 2.1: DLR for normal data with sample sizes $n_1 = n_2 = 50$

we show the smooth empirical likelihood estimator for data with $F_1 \sim \text{Exp}(1/6)$, $F_2 \sim \text{Exp}(1/10)$ and $n_1 = n_2 = 50$.

Figures 2.1 and 2.2 show that the smooth empirical estimator fits normally distributed data with equal sample sizes very well, and similar datasets with unequal sample sizes well provided $t > 0.2$. For data from exponential distributions the fit is less satisfactory when $t < 0.1$ but noticeably better when $t \geq 0.4$. As we can see in Figure 2.4, the smooth empirical likelihood estimator is a smooth function of t while the empirical estimator is distinctly jagged.

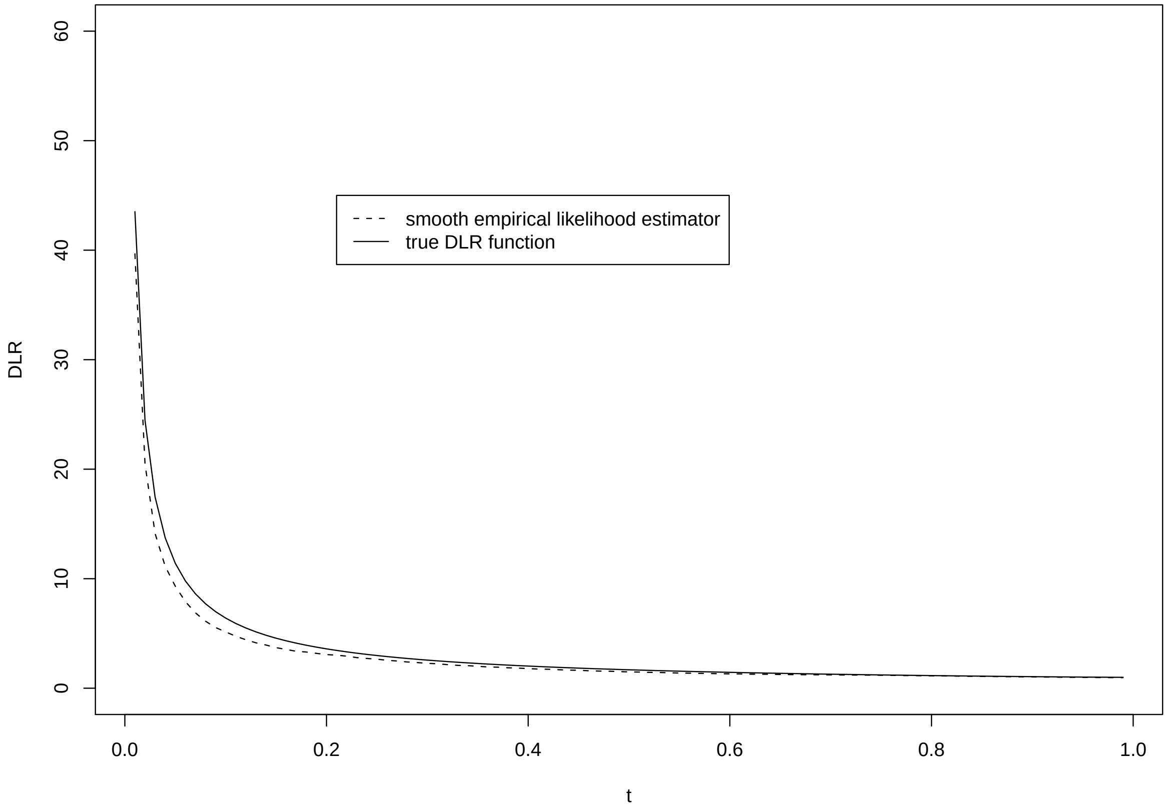

Figure 2.2: DLR for normal data with sample sizes $n_1 = 50, n_2 = 90$

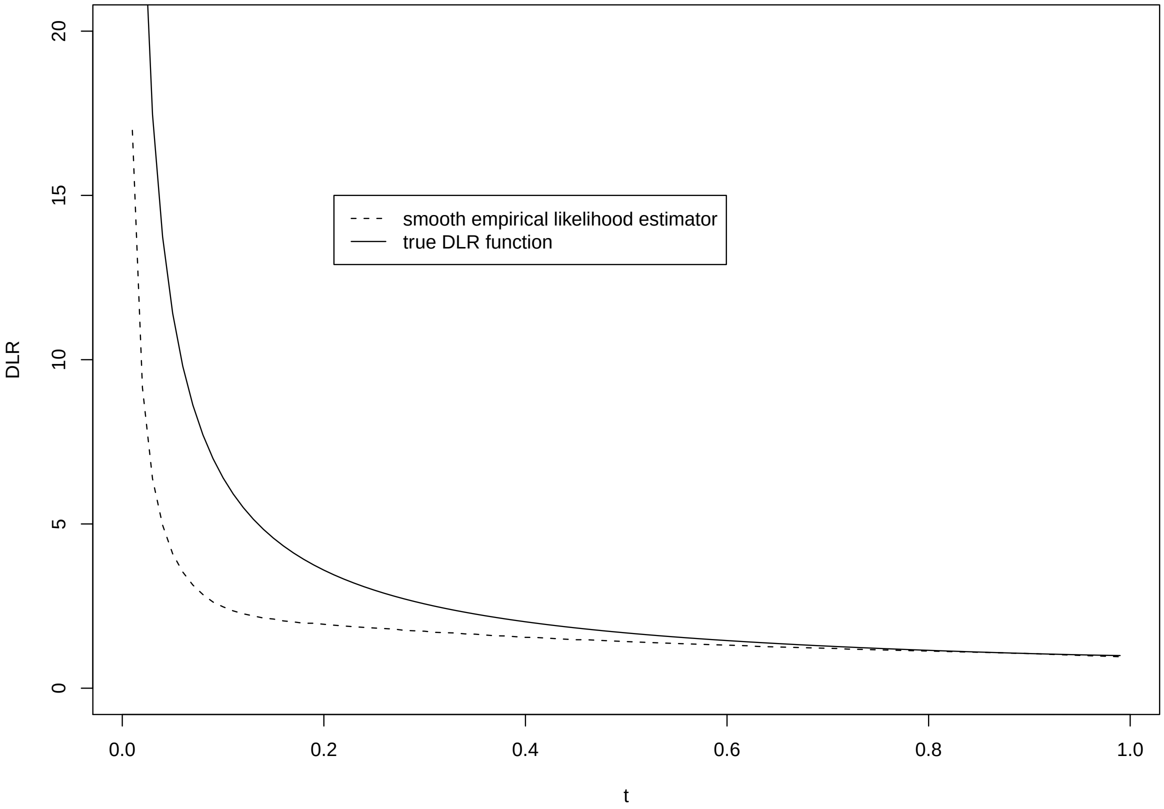

Figure 2.3: DLR for exponential data with sample sizes $n_1 = n_2 = 50$

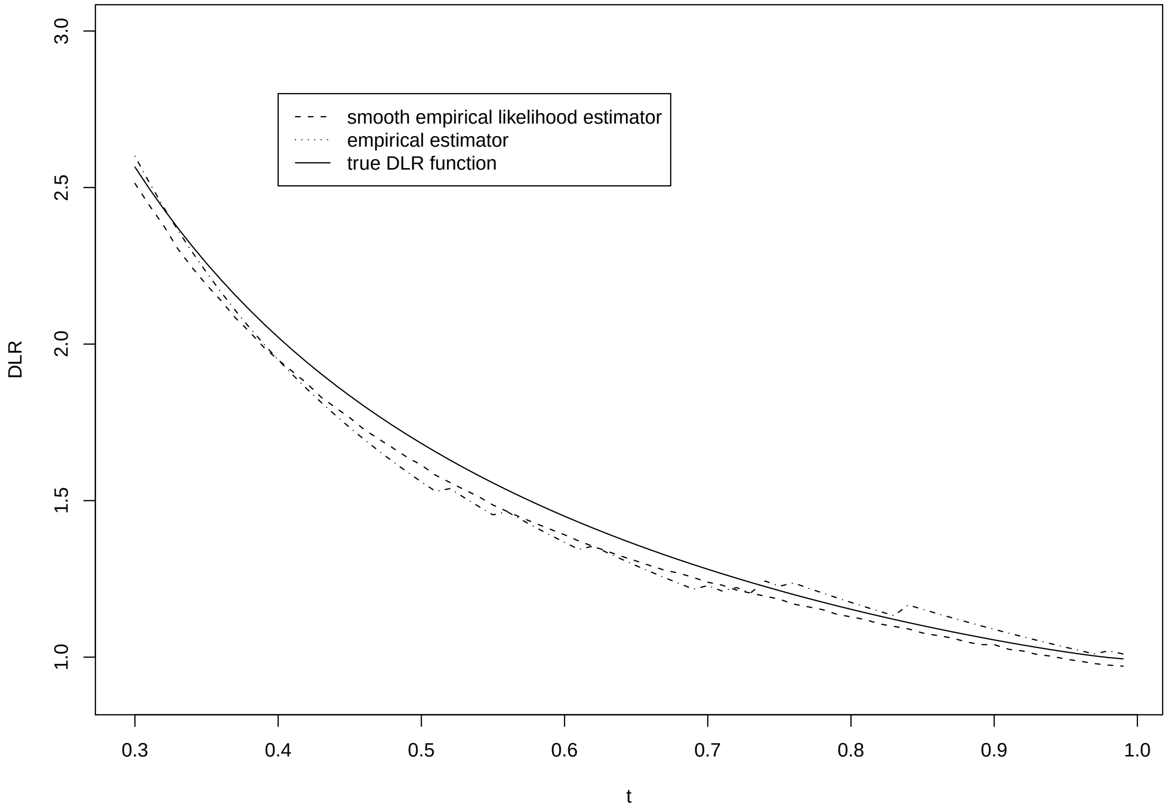

Figure 2.4: Comparing the SEL and EE estimators of the DLR function for $F_1 \sim N(6,2),\ F_2 \sim N(10,4)$ with $n_1 = n_2 = 50$

To compare the coverage accuracy of interval estimates based on the smooth empirical likelihood estimator with corresponding kernel and empirical ones, we conducted a Monte Carlo study using 10,000 pseudorandom samples for each scenario from $F_1 \sim N(6,2)$ and $F_2 \sim N(10,4)$ at a nominal confidence level of 95%. We used the same Gaussian kernel and bandwidth for both the kernel and the smooth empirical likelihood interval estimates.

Confidence intervals constructed from the empirical estimator of the DLR can be obtained as follows. The asymptotic variance of the empirical estimator is given by

$$V(t) = \frac{\theta(1-\theta t)}{n_2 t} + \left[\frac{f_2(\eta)}{f_1(\eta)}\right]^2 \frac{(1-t)}{n_1 t}$$

Replacing θ, η in the above formula by their empirical versions, and using kernel estimates for $f_i, i = 1, 2$, we can obtain a consistent estimator of $V(t)$, called $\hat{V}(t)$. Then the $100(1-\alpha)\%$ confidence interval corresponding to the empirical estimator is

$$\left(\hat{\theta}(t) - z_{\alpha/2}\sqrt{\hat{V}(t)},\ \ \hat{\theta}(t) + z_{\alpha/2}\sqrt{\hat{V}(t)}\right)$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of N(0,1). Likewise, if we replace θ, η, $f_i, i = 1, 2$ by their kernel versions, we can obtain an interval estimate based on the kernel estimator. By Theorem 2, there are no unknown quantities that arise in constructing a confidence interval based on the smooth empirical likelihood estimator. Therefore, we can use $A_{q_{1-\alpha}} = \{\theta : l_n(\theta) \leq q_{1-\alpha}\}$ to obtain a $100(1-\alpha)\%$ confidence interval based on the smooth empirical likelihood estimator. Table 2.2 summarizes the coverage accuracy of interval estimates corresponding to these three estimators at the nominal level of 95%.

Table 2.2: Estimated coverage probabilities for SEL, KE, EE

Percentage of estimated coverage accuracy and standard error of 95% confidence intervals for the smooth empirical likelihood (SEL), kernel (KE) and empirical (EE) estimators with different sample sizes for disease-free (n_1) and diseased (n_2)groups.

n_1	n_2	Method	$t = 0.1$	$t = 0.3$	$t = 0.5$	$t = 0.7$	$t = 0.9$
25	25	SEL	93.0	92.8	91.9	90.6	84.9
			(.26)	(.26)	(.27)	(.29)	(.35)
		KE	91.7	92.1	91.5	90.9	89.1
			(.28)	(.27)	(.28)	(.29)	(.31)
		EE	89.8	89.7	87.9	90.7	68.1
			(.30)	(.30)	(.33)	(.29)	(.47)
20	30	SEL	92.8	93.1	93.1	91.5	86.5
			(.26)	(.25)	(.25)	(.28)	(.34)
		KE	91.6	92.6	92.8	91.1	83.2
			(.28)	(.26)	(.26)	(.28)	(.37)
		EE	89.3	90.1	92.9	90.8	66.6
			(.31)	(.30)	(.26)	(.29)	(.47)
30	20	SEL	93.9	93.0	90.6	87.9	81.4
			(.24)	(.26)	(.29)	(.33)	(.39)
		KE	92.2	92.1	90.7	88.4	85.2
			(.27)	(.27)	(.29)	(.32)	(.36)
		EE	91.4	94.0	85.9	86.2	58.9
			(.28)	(.24)	(.35)	(.34)	(.49)
50	50	SEL	92.6	92.6	92.6	91.7	89.9
			(.26)	(.26)	(.26)	(.28)	(.30)
		KE	92.3	93.4	93.3	92.7	94.7
			(.27)	(.25)	(.25)	(.26)	(.22)
		EE	92.1	93.2	93.8	90.2	86.8
			(.27)	(.25)	(.24)	(.30)	(.34)
100	100	SEL	92.4	92.2	92.0	91.5	92.2
			(.26)	(.27)	(.27)	(.28)	(.27)
		KE	92.7	93.0	92.9	92.8	97.1
			(.26)	(.26)	(.26)	(.26)	(.17)
		EE	93.3	93.6	93.7	92.9	90.5
			(.25)	(.24)	(.24)	(.26)	(.29)

From Table 2.2 we notice that the SEL estimator has a higher coverage probability than either its kernel or empirical competitors in almost every situation when the sample from a patient group involves fewer than 50 individuals. This observation is particularly true whenever the false positive probability in the disease-free group is less than 0.7, which means it ought to apply in most practical situations involving diagnostic tests. However, if the sample sizes in the disease-free and diseased groups are larger than 50, it appears that the kernel-based estimated coverage probabilities are closest to the nominal value of 95%, although all three methods of interval estimation seem somewhat anti-conservative. Of course, the coverage accuracy of these methods depends on using the optimal bandwidth, which is still an open problem.

2.2.4 The CA 19-9 Diagnostic Test

We used the smooth empirical likelihood method to analyze data that were first published by Wieand *et al.* (1989) concerning CA 19-9 diagnostic test measurements in patients with pancreatic cancer (diseased) or pancreatitis (disease-free). Point estimates and 95% point-wise interval estimates for the positive DLR are displayed in Figure 2.5. The kernel functions we used were Gaussian with Gaussian bandwidths $h_1 = 0.355$ and $h_2 = 0.857$. From the relationship between the positive and negative DLRs we derived corresponding estimates for the negative DLR displayed in Figure 2.6.

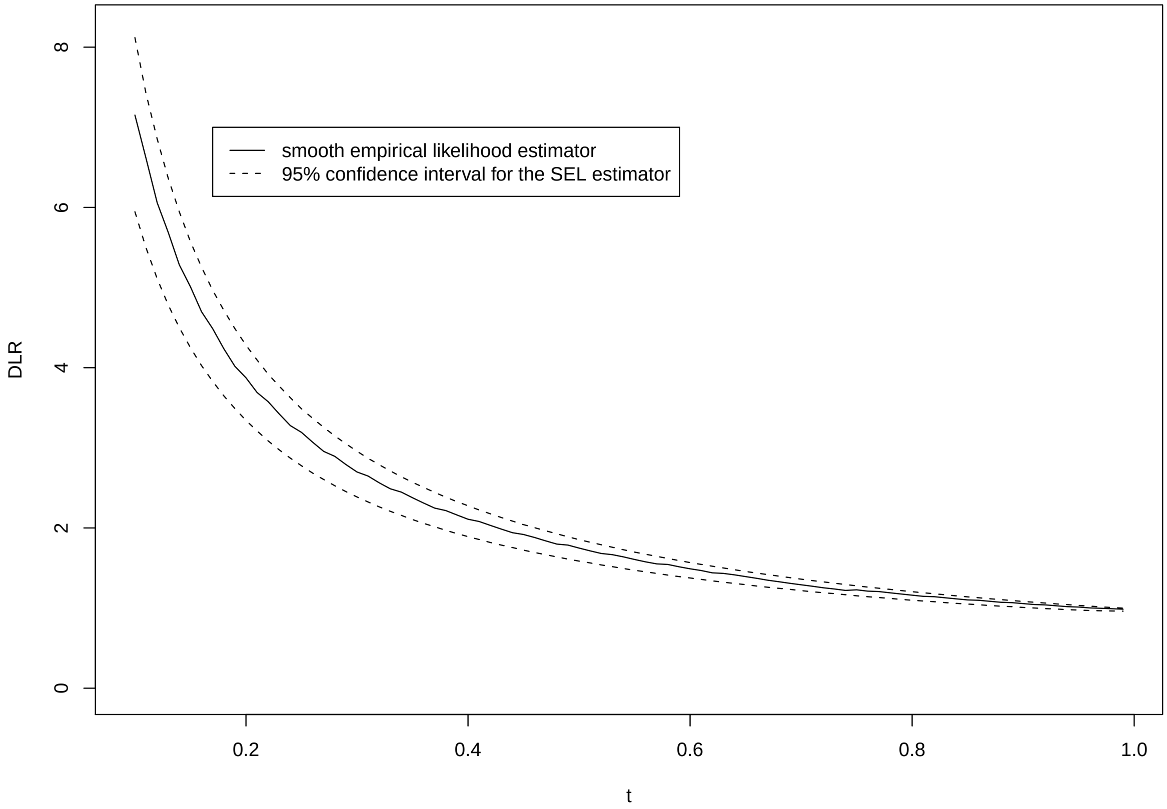

Figure 2.5: Positive DLR estimator for the CA 19-9 data

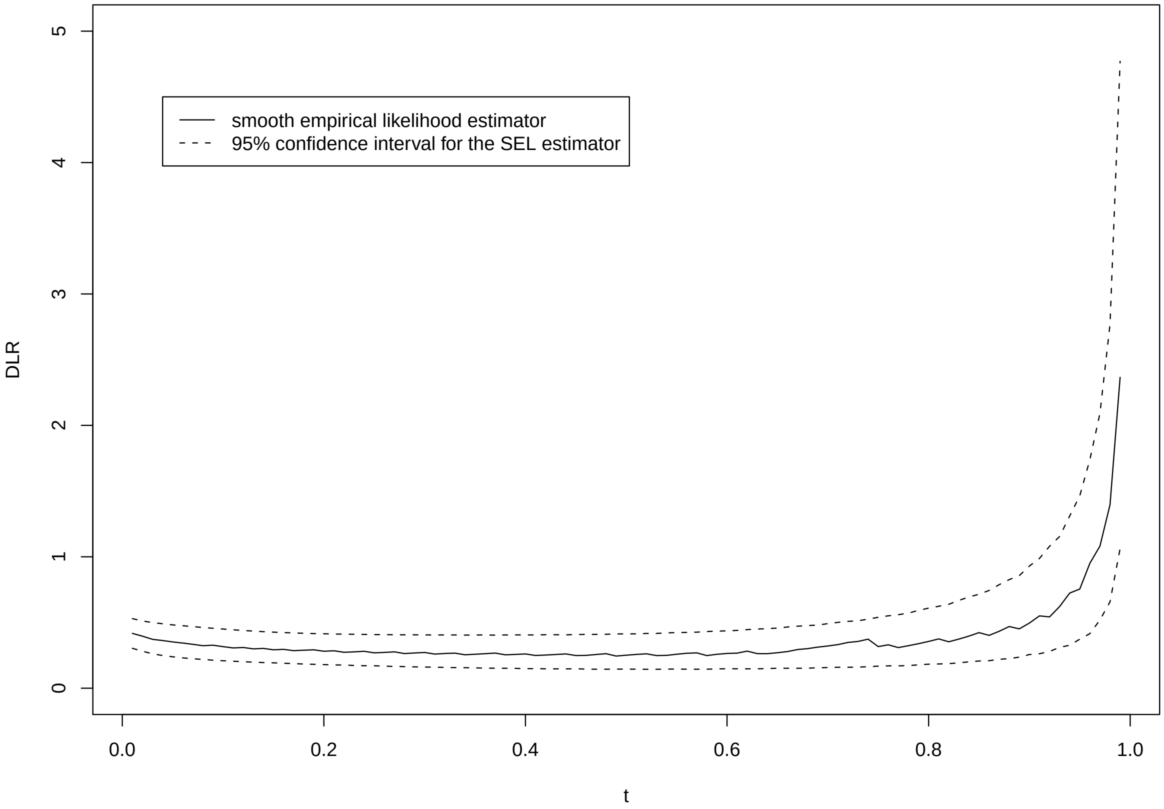

Figure 2.6: Negative DLR estimator for the CA 19-9 data

2.3 An Empirical Likelihood Method Using Nonsmooth Estimating Equations

Although the smooth empirical likelihood method can deliver a more efficient estimator of the DLR with superior coverage accuracy compared to its kernel and empirical competitors, its performance depends on selecting the proper bandwidth. This bandwidth selection problem may be difficult in some situations such as unequal sample sizes for the disease-free and diseased groups. To avoid this bandwidth selection problem, we next consider an empirical likelihood method involving nonsmooth estimating equations, based on indicator functions.

Let

$$\hat{F}_{1,p}(\eta) = \textstyle\sum_{i=1}^{n_1} p_i I(X_{1i} \leq \eta) \quad \text{and} \quad \hat{F}_{2,q}(\eta) = \textstyle\sum_{j=1}^{n_2} q_j I(X_{2j} \leq \eta).$$

Then the empirical likelihood of θ, the positive DLR, is

$$L(\theta) = \sup_{(p,q,\eta)} (\prod_{i=1}^{n_1} p_i)(\prod_{j=1}^{n_2} q_j), \tag{2.9}$$

subject to the constraints

$$\hat{F}_{1,p}(\eta) = 1 - t \quad \text{and} \quad \hat{F}_{2,q}(\eta) = 1 - \theta t, \quad 0 \leq t \leq 1, \ \theta > 0.$$

As before,

$$p_i = \frac{1}{n_1\{1 + \lambda_1 w_1(\eta, X_{1i})\}}, \ i = 1, ..., n_1,$$

$$q_j = \frac{1}{n_2\{1 + \lambda_2 w_2(\eta, X_{2j})\}}, \ j = 1, ..., n_2,$$

where

$$w_1(\eta, X_{1i}) = I(X_{1i} \leq \eta) - (1-t), \quad \text{and} \quad w_2(\eta, X_{2j}) = I(X_{2j} \leq \eta) - (1-\theta t).$$

Then

$$g_1(\eta, \theta) = \sum_{i=1}^{n_1} \frac{w_1(\eta, X_{1i})}{n_1\{1 + \lambda_1 w_1(\eta, X_{1i})\}} = 0, \tag{2.10}$$

$$g_2(\eta, \theta) = \sum_{j=1}^{n_2} \frac{w_2(\eta, X_{2j})}{n_2\{1 + \lambda_2 w_2(\eta, X_{2j})\}} = 0. \tag{2.11}$$

are the corresponding estimating equations for θ.

Note that for $j = 1, 2,\ g_j$ is a continuous function of λ_j, so we can make use of $L_*(\lambda_j)$ to obtain λ_j (see formula (2.7)).With the specific indicator functions, for given η, let $m_j(\eta) = \sum_{i=1}^{n_j} I(X_{ji} < \eta)$ for $j = 1, 2$. From equations (2.10) and (2.11) it follows that $\lambda_1 = \frac{1}{t}[1 - \frac{m_1(\eta)}{n_1(1-t)}]$ and $\lambda_2 = \frac{1}{\theta t}[1 - \frac{m_2(\eta)}{n_2(1-\theta t)}]$.

Let $\tilde{\eta} = \arg\max_\eta \min_\lambda \{\sum_{j=1}^2 L_*(\lambda_j, \theta, \eta)\}$. Then $\hat{\theta}_n = \arg\min_\theta l_n(\theta)$ is the empirical likelihood estimator of θ, where $l_n(\theta)$ is the profile empirical log-likelihood ratio

$$\begin{aligned} l_n(\theta) &= 2\sum_{j=1}^2 \sum_{i=1}^{n_j} \log\{1 + \tilde{\lambda}_j w_j(\tilde{\eta}, X_{ji})\} \\ &= 2\{m_1(\tilde{\eta}) \log \tfrac{m_1(\tilde{\eta})}{n_1(1-t)} + [n_1 - m_1(\tilde{\eta})] \log \tfrac{n_1 - m_1(\tilde{\eta})}{n_1 t}\} \\ &+2\{m_2(\tilde{\eta}) \log \tfrac{m_2(\tilde{\eta})}{n_2(1-\theta t)} + [n_2 - m_2(\tilde{\eta})] \log \tfrac{n_2 - m_2(\tilde{\eta})}{n_2 \theta t}\} \\ &= A_1 + A_2. \end{aligned} \tag{2.12}$$

Let $l(\theta) = 2\sum_{j=1}^2 \sum_{i=1}^{n_j} E\{\log(1 + \lambda_j w_j(\eta, X_{ji})\}$, where $\lambda_j(\eta)$ satisfies

$$E\{\frac{w_j(\eta, X_{ji})}{1 + \lambda_j w_j(\eta, X_{ji})}\} = 0,\ j = 1, 2. \tag{2.13}$$

For given θ_0, let η_0 satisfy (2.13); then $l(\theta_0) = 0$, since $\lambda(\eta_0) = 0$. Therefore $l(\theta)$, $l_n(\theta_0)$ satisfy the following conditions:

$$\text{(i)} \quad |l_n(\theta)| \leq o_p(1) + \inf_{\theta \in \Theta} |l_n(\theta)|$$

$$\text{(ii)} \quad l_n(\theta_0) = o_p(1)$$

$$\text{(iii)} \quad \sup_{\|\theta - \theta_0\| > \delta} \|l_n(\theta)\|^{-1} = O_p(1)$$

By theorem (3.1) of Pakes and Pollard (1989), $\tilde{\eta} \xrightarrow{p} \eta_0$ as $n_1, n_2 \to \infty$.

Now assume $\int X_1 dF_1(X_1) < \infty$ and $\int X_2 dF_2(X_2) < \infty$. By the WLLN, $m_1(\tilde{\eta})$ has an approximate normal distribution with mean $\{n_1 F_1(\tilde{\eta})\}$ and $F_1(\tilde{\eta}) \xrightarrow{p} F_1(\eta_0)$, so $m_1(\tilde{\eta}) \xrightarrow{D} N(n_1(1-t), n_1 t(1-t))$; likewise $m_2(\tilde{\eta}) \xrightarrow{D} N(n_2(1-\theta t), n_2 \theta t(1-\theta t))$.

Since A_1, A_2 are Wilks' statistics for binomial random variables with parameters $(n_1, 1-t)$, $(n_2, 1-\theta t)$, respectively, $l_n(\theta_0) \xrightarrow{D} \chi^2$ with $df = 2 - 1$ as $n_1, n_2 \to \infty$ since the parameter η is unknown and is estimated.

Unlike the situation where t, the false positive probability in the disease-free population is given, it is common in diagnostic testing to define a test result to be positive if $X_{ji} > \eta$, for given η. Then we define

$$\rho_+ = \frac{1 - F_2(\eta)}{1 - F_1(\eta)}, \quad \rho_- = \frac{F_2(\eta)}{F_1(\eta)}$$

In this case t is a nuisance parameter. From the estimating equations (2.10) and (2.11) we can obtain $\lambda_1 = \frac{1}{t}[1 - \frac{m_1(\eta)}{n_1(1-t)}]$, $\lambda_2 = \frac{1}{\theta t}[1 - \frac{m_2(\eta)}{n_2(1-\theta t)}]$, and $\tilde{t} =$

$\arg\max_t \min_\lambda\{\sum_{j=1}^2 L_*(\lambda_j,\theta,t)\}$. By the same argument we have $l_n(\theta_0) \xrightarrow{D} \chi_1^2$.

In order to compare the coverage accuracy of confidence intervals obtained from this empirical likelihood method with the usual normal approximation, we computed the variance of the empirical estimator $\hat{\rho}_+ = \frac{1-\hat{F}_2(\eta)}{1-\hat{F}_1(\eta)}$ by the delta method as follows:

$$\begin{aligned} Var(\hat{\rho}_+) &= \exp\{\log(1-F_2(\eta)) - \log(1-F_1(\eta))\}^2\{\text{Var}[\log((1-\hat{F}_2(\eta))] \\ &+\text{Var}[\log((1-\hat{F}_1(\eta))\} \\ &= \left[\tfrac{1-F_2(\eta)}{1-F_1(\eta)}\right]^2\{\tfrac{1}{[1-F_2(\eta)]^2}\text{Var}(1-\hat{F}_2(\eta)) + \tfrac{1}{[1-F_1(\eta)]^2}\text{Var}(1-\hat{F}_1(\eta))\} \\ &=\left[\tfrac{1-F_2(\eta)}{1-F_1(\eta)}\right]^2\{\tfrac{F_2(\eta)}{n_2[1-F_2(\eta)]} + \tfrac{F_1(\eta)}{n_1[1-F_1(\eta)]}\} \end{aligned}$$

We conducted a second simulation study to compare the estimated coverage probabilities of 95% confidence intervals using empirical likelihood and the usual normal approximation. For each sample size we generated 10,000 pseudorandom samples with $F_1 \sim N(6,2)$ and $F_2 \sim N(10,4)$. The simulation results are shown in Table 2.3. From the table we observe that the empirical likelihood method has smaller coverage error and therefore higher coverage accuracy than the corresponding normal approximation when $\eta = 4, 6, 8$. The coverage errors for both methods are relatively large when $\eta = 2$ because there are few observations that can be used to estimate ρ_+ in that region of the test measurement scale.

2.4 Conclusions

Empirical likelihood, as a nonparametric method of statistical inference, is an effective tool that can be used to pool information from different data sources to

Table 2.3: Percentage of estimated coverage probabilities and standard errors of non-smooth empirical likelihood (NEL) and the corresponding empirical estimator (EE)

n_1	n_2	Method	$\eta = 2$	$\eta = 4$	$\eta = 6$	$\eta = 8$
25	25	NEL	97.5	93.7	94.6	95.0
			(.16)	(.24)	(.23)	(.22)
		EE	99.8	93.6	94.5	90.3
			(.04)	(.24)	(.23)	(.30)
50	50	NEL	93.1	95.0	94.8	94.7
			(.25)	(.22)	(.22)	(.22)
		EE	98.5	95.9	95.3	91.5
			(.12)	(.20)	(.21)	(.28)
50	90	NEL	93.0	94.2	94.7	94.4
			(.26)	(.23)	(.22)	(.23)
		EE	95.0	94.1	94.4	93.2
			(.22)	(.24)	(.23)	(.25)
90	50	NEL	93.1	94.4	94.9	94.5
			(.25)	(.23)	(.22)	(.23)
		EE	97.6	95.1	95.4	94.4
			(.15)	(.22)	(.21)	(.23)
100	100	NEL	93.3	95.1	95.3	94.8
			(.25)	(.22)	(.21)	(.22)
		EE	96.8	94.7	95.1	94.3
			(.18)	(.22)	(.22)	(.23)

produce more accurate point and interval estimators. We employ the empirical likelihood method to incorporate information from samples of disease-free and diseased subjects to estimate diagnostic likelihood ratios, which are widely used in the medical and clinical literature.

For continuous-scale diagnostic tests, we combine the empirical likelihood method with kernel estimating equations to obtain a smooth empirical likelihood estimator that is more efficient than competing kernel and empirical estimators. Moreover this smooth empirical likelihood interval estimator has higher coverage accuracy in small sample settings than its kernel and empirical competitors. However, the smooth empirical likelihood method involves selecting a suitable bandwidth that may be a challenging problem in some situations. To avoid this bandwidth selection problem, we adopt an empirical likelihood method with non-smooth estimating equations to estimate DLRs.

The non-smooth empirical likelihood estimator of DLR is an optimization estimator, and under certain regularity conditions we show that the empirical log-likelihood ratio statistic converges to a chi-squared random variable. Our simulation study demonstrates that the non-smooth empirical likelihood estimator has smaller coverage errors, and therefore higher coverage accuracy in term of 95% confidence intervals than the usual normal approximation. By combining empirical likelihood with non-smooth estimating equations, we have extended the application of empirical likelihood to more general situations in which the estimating equations, such as those based on quantile or indicator functions, may not be smooth.

Chapter 3

Empirical Likelihood for Cumulative Hazard Ratio Estimation

3.1 Introduction

In medical studies that assess a treatment effect in terms of hazard ratios, the absence of proportionality can be problematic. To cope with nonproportional hazards in the Cox regression model, investigators usually assume that the treatment effect has some smooth functional form over time or perhaps is piece-wise constant. However, it is generally difficult to assess whether the functional form chosen for the treatment effect is correct. Moreover, study investigators may be more interested in the cumulative effect of treatment over time, rather than its instantaneous value. These considerations motivate us to propose a nonparametric estimator for the cumulative treatment effect under nonproportional hazards.

Several methods have been proposed in the literature for estimating the ratio of cumulative hazards in nonparametric settings. Kalbfleisch and Prentice (1981) estimated an average hazard ratio using a weight function. Schemper (1992) suggested a covariate-adjusted estimator of the average hazard ratio in the two populations via a weighted Cox model. Under a nonproportional hazards model, Xu and O'Quigley (2000) employed a weighted score equation to estimate the average regression effect. In 2008, Wei and Schaubel proposed an estimator of the ratio of baseline cumulative hazards in two populations under a stratified Cox model. The resulting estimator has an asymptotic normal distribution, but the normal approximation-based confidence region is not easy to construct. Moreover it is always symmetric, which may not be desirable in every situation, and the coverage probability of a $100(1-\alpha)\%$ interval estimator for the true cumulative hazard ratio when the sample size is small is also far below the nominal level (as shown in their simulation studies).

To overcome these limitations of the normal approximation, and improve the coverage accuracy of the corresponding interval estimates, we used empirical likelihood (EL) to derive an interval estimator for the ratio of covariate-adjusted cumulative hazards in two populations. Compared to a normal approximation, our EL-based confidence region for the cumulative hazard ratio has the the following advantages: (1) It is easier to construct since there is no need to compute a variance estimator; (2) It has superb coverage accuracy in small samples; (3) It is not necessarily symmetric, which enables it to better reflect the shape of the underlying distribution.

Many authors have investigated the use of EL in time-to-event settings. The pioneering contributions were due to Kaplan and Meier (1958) and Thomas and Grunkemeier (1975). Li (1995) and Murphy (1995) provided a theoretical foundation for applications of EL used in survival analysis. Li, Qin and Tiwari (1997) and

Hollander, McKeague and Yang (1997) derived EL-based confidence intervals for survival functions using truncated or right-censored data. Wang and Jing (2001) applied an adjusted EL to the estimation of a class of functionals of the survival function involving right-censored data. Pan and Zhou (2002) studied statistical behaviour of the EL ratio statistic for data that may be right censored when the parameter of interest is a linear functional of the cumulative hazard function. Li and van Keilegom (2002) extended the pioneering work of Thomas and Grunkemeier (1975) to the nonparametric regression setting, obtaining confidence intervals and bands for conditional survival and quantile functions. McKeague and Zhao (2002, 2005) constructed a simultaneous confidence band for the difference or ratio of two survival functions based on independent right-censored data.

As far as we are aware, no one has describe the use of EL to estimate the ratio of arbitrary baseline cumulative hazard functions in two populations, in the presence of covariate adjustments. To address this problem of estimation, we begin with the Poisson extension of the exact likelihood function for the cumulative hazard function introduced by Murphy (1995), since it can incorporate the Cox regression model directly, and thereby allow for covariate adjustment of the cumulative hazard ratio of interest even when the functional itself is not constant and therefore the two baseline cumulative hazards are not proportional. In what follows we outline such a nonparametric estimator, obtaining both point and interval estimates. The rest of the chapter is structured as follows. In Section 2, we describe the ratio of arbitrary cumulative hazards in two populations where the adjustment for other covariate information follows a stratified Cox regression model. In Section 3, we relax the requirements of the stratified regression model to include the possibility of group-specific adjustment for other covariate information. Simulation studies that investigate the performance of these EL-based estimators compared to the usual normal approximation are described in section 4. We then illustrate each of the proposed methods in section 5, using separate datasets concerning the survival experience of non-Hodgkin's lymphoma and ovarian cancer patients. The chapter concludes with some summary remarks.

3.2 Empirical Likelihood Estimation of a Covariate-adjusted Cumulative Hazard Ratio

Suppose that $T_{11},...,T_{1n_1}$ and $T_{21},...,T_{2n_2}$ are independent samples of event times from a well-defined, common origin for populations 1 and 2 with distribution functions F_1 and F_2, respectively. We refer to group 1 as the reference category, and assume that the cumulative hazard functions for the two groups are not proportional, i.e., the corresponding ratio is arbitrary, under any right-censoring mechanism. For $j = 1, 2$, let $C_{j1},...,C_{jn_j}$ be independent censoring times with corresponding distribution functions $G_j, j = 1, 2$, respectively. We assume that T and C are unconditionally independent. The observation time and observed event indicator are $X_{ji} = \min(T_{ji}, C_{ji})$ and $\delta_{ji} = I(X_{ji} \leq C_{ji})$. The function $N_{ji}(t) = \delta_{ji} I(X_{ji} < t)$ is the corresponding counting process; the risk indicator is $Y_{ji}(t) = I(X_{ji} \geq t)$. Thus, the observed data consist of $n = n_1 + n_2$ mutually independent vectors, each consisting of X_{ji}, δ_{ji} and Z_{ji}, a vector of subject-specific covariate information.

For group j, we assume that T_{ji} follows a Cox regression model with hazard function

$$\lambda_{ji}(t) = \lambda_{j0}(t)\exp(\beta_0^T Z_{ji}), \tag{3.1}$$

where $\lambda_{j0}(t)$ is an unspecified baseline hazard function, and β_0 is an unknown parameter vector. Under model (3.1), we assume that the hazards are proportional with respect to the adjustment for covariate information within each group but not across the groups, which is less restrictive. Note also that we assume the covariate vector is constant over time. Hence model (3.1) represents a stratified Cox regression model in which the two strata correspond to the two groups of interest.

Let $\hat{\beta}$ be the partial likelihood (Cox, 1975) estimator of β_0, which we obtain by solving the equation $U(\beta) = 0$, where

$$U(\beta) = \sum_{j=1}^{2}\sum_{i=1}^{n_j}\int_0^\infty \{Z_{ji} - \bar{Z}_j(t,\beta)\}dN_{ji}(t),$$

and

$$\bar{Z}_j(t,\beta) = \frac{\sum_{i=1}^{n_j} Y_{ji}(t) Z_{ji}\exp(\beta^T Z_{ji})}{\sum_{i=1}^{n_j} Y_{ji}(t)\exp(\beta^T Z_{ji})}.$$

We define the parameter of interest to be

$$\theta(t) = \frac{\Lambda_{20}(t)}{\Lambda_{10}(t)}, \tag{3.2}$$

where $\Lambda_{j0}(t) = \int_0^t \lambda_{j0}(s)ds$ is the baseline cumulative hazard function for group j. This ratio of the baseline cumulative hazards characterizes any discrepancy in aggregate response experience between the two groups over the interval $(0,\ t]$, after adjustment for other covariate information. In addition, if the two groups represent treatment levels, then equation (3.1) implies that $\theta(t)$ reflects the contrast effect of treatment between subjects whose covariate information is identical.

To simplify subsequent notation, we will suppress the time-dependence of $\theta(t)$ and just refer to θ. If we adopt the usual normal approximation, then

$$\hat{\theta} = \frac{\hat{\Lambda}_{20}(\hat{\beta}, t)}{\hat{\Lambda}_{10}(\hat{\beta}, t)}, \tag{3.3}$$

where $\hat{\Lambda}_{j0}(\hat{\beta}, t)$ is the Breslow (1972) estimator

$$\hat{\Lambda}_{j0}(\hat{\beta}, t) = \frac{1}{n_j} \sum_{i=1}^{n_j} \int_0^t \frac{dN_{ji}(s)}{S_j^0(s, \hat{\beta})} \qquad j = 1, 2. \tag{3.4}$$

Here, $S_j^0(t, \hat{\beta}) = n_j^{-1} \sum_{i=1}^{n_j} Y_{ji}(t) \exp(\hat{\beta}^T Z_{ji})$.

To derive an empirical likelihood estimator for $\theta(t)$ we first consider the two likelihood functions

$$EL(\Lambda_j) = \prod_{i=1}^{n_j} [\triangle\Lambda_j(X_{ji})]^{\delta_{ji}} \exp(-\Lambda_j(X_{ji})) \qquad j = 1, 2, \tag{3.5}$$

for the two cumulative hazards, $\Lambda_1(t)$ and $\Lambda_2(t)$, where $\triangle\Lambda_j(x) = \frac{\triangle F_j(x)}{1-F_j(x-)}$. Note the likelihood function specified in equation (3.5) is not the exact likelihood function but the Poisson extension of the likelihood; see Murphy (1995) for details.

Without loss of generality, we assume that $X_{j1} \leq X_{j2} \leq ... \leq X_{jn_j}$, for $j = 1, 2$. Let $w_{ji}^0 = dN_{ji}(X_{ji})/\{n_j S_j^0(X_{ji}, \hat{\beta})\}$ be the hazard increment for the Breslow estimator. To define empirical likelihood hazard increments $\{p_i\}$, $\{q_k\}$, for $i =$

$1, ..., n_1, k = 1, ..., n_2$, we force the last increase in the estimated cumulative hazard function to be the same as that of the Breslow increment, i.e., $p_{n_1} = w^0_{1n_1}$ and $q_{n_2} = w^0_{2n_2}$. This follows from the definition of $\triangle\Lambda_j(x) = \frac{\triangle F_j(x)}{1-F_j(x-)}$, which requires the last jump of a proper discrete cumulative hazard function to be 1. Correspondingly, the last observation for the Breslow dominated discrete cumulative hazard function has the same jump as the Breslow estimator.

Therefore, after covariate adjustment the empirical likelihood of θ is:

$$EL(\theta) = \sup_{(p,q,\eta)} \left(\prod_{i=1}^{n_1} [p_i \exp(\hat{\beta}^T Z_{1i})]^{\delta_{1i}} \exp\{-(\sum_{m=1}^{i} p_m) \cdot \exp(\hat{\beta}^T Z_{1i})\} \right) \cdot$$

$$\left(\prod_{k=1}^{n_2} [q_k \exp(\hat{\beta}^T Z_{2k})]^{\delta_{2k}} \exp\{-(\sum_{m=1}^{k} q_m) \cdot \exp(\hat{\beta}^T Z_{2k})\} \right), \tag{3.6}$$

subject to the following constraints:

$$\sum_{i=1}^{n_1-1} \delta_{1i} I(X_{1i} \leq t) \cdot p_i + \delta_{1n_1} I(X_{1n_1} \leq t) \cdot p_{n_1} = \eta, \tag{3.7}$$

$$\sum_{k=1}^{n_2-1} \delta_{2k} I(X_{2k} \leq t) \cdot q_k + \delta_{2n_2} I(X_{2n_2} \leq t) \cdot q_{n_2} = \eta \cdot \theta, \tag{3.8}$$

where $p_i > 0,\ q_k > 0,\ i = 1, ...n_1,\ k = 1, ..., n_2$, which satisfies the usual requirements for a hazard increment.

Although it is possible to estimate β and θ jointly, the focus of scientific interest is the ratio of the baseline cumulative hazard functions, and the values of β should not be associated with the value of θ. Instead, estimates of β should be evaluated independently, using relevant information collected in each group of subjects.

Therefore, here we have adopted a commonly-used estimator, the maximum partial likelihood estimator as the estimator of β. Using this fixed value of β, we can then derive an interval estimate of the ratio of baseline cumulative hazard functions that is our primary focus. This approach is reinforced by the results of Johansen (1983), who demonstrated that the Nelson-Aalen estimator is the profile estimator of the baseline cumulative hazard function when the vector of regression coefficients, β, is fixed. In the same paper Johansen also showed that to estimate β, we should maximize the familiar partial likelihood function of Cox (1975).

Using the Lagrange multipliers ξ_1 and ξ_2, we can represent the empirical log-likelihood of θ under constraints (3.7) and (3.8) by,

$$
\begin{aligned}
l(\theta) = &\sum_{i=1}^{n_1} \delta_{1i}[\log(p_i) + \hat{\beta}^T Z_{1i}] - \sum_{i=1}^{n_1}\{(\sum_{m=1}^{i} p_m) \cdot \exp(\hat{\beta}^T Z_{1i})\} + \\
&\sum_{k=1}^{n_2} \delta_{2k}[\log(q_k) + \hat{\beta}^T Z_{2k}] - \sum_{k=1}^{n_2}\{(\sum_{m=1}^{k} q_m) \cdot \exp(\hat{\beta}^T Z_{2k})\} + \\
&n_1\xi_1\{\sum_{i=1}^{n_1-1} \delta_{1i} I(X_{1i} \le t) \cdot p_i + \delta_{1n_1} I(X_{1n_1} \le t) \cdot p_{n_1} - \eta\} + \\
&n_2\xi_2\{\sum_{k=1}^{n_2-1} \delta_{2k} I(X_{2k} \le t) \cdot q_k + \delta_{2n_2} I(X_{2n_2} \le t) \cdot q_{n_2} - \eta \cdot \theta\}. \qquad (3.9)
\end{aligned}
$$

From the score equations $\partial l(\theta)/\partial p_i = 0$ and $\partial l(\theta)/\partial q_k = 0$ we obtain, for $i = 1, ..., n_1 - 1, k = 1, ..., n_2 - 1$,

$$
\begin{aligned}
p_i &= \frac{\delta_{1i}}{\sum_{m=i}^{n_1} \exp(\hat{\beta}^T Z_{1m}) + n_1 \xi_1 \delta_{1i} I(X_{1i} \le t)} \\
&= \frac{\delta_{1i}}{\sum_{m=i}^{n_1} \exp(\hat{\beta}^T Z_{1m})} \cdot \frac{1}{1 + n_1 \xi_1 \delta_{1i} I(X_{1i} \le t) / \sum_{m=i}^{n_1} \exp(\hat{\beta}^T Z_{1m})} \\
&= w_{1i}^0 \cdot \frac{1}{1 + n_1 \xi_1 \delta_{1i} I(X_{1i} \le t) / \sum_{m=i}^{n_1} \exp(\hat{\beta}^T Z_{1m})},
\end{aligned}
$$

$$
\begin{aligned}
q_k &= \frac{\delta_{2k}}{\sum_{m=k}^{n_2} \exp(\hat{\beta}^T Z_{2m}) + n_2 \xi_2 \delta_{2k} I(X_{2k} \le t)} \\
&= \frac{\delta_{2k}}{\sum_{m=k}^{n_2} \exp(\hat{\beta}^T Z_{2m})} \cdot \frac{1}{1 + n_2 \xi_2 \delta_{2k} I(X_{2k} \le t) / \sum_{m=k}^{n_2} \exp(\hat{\beta}^T Z_{2m})} \\
&= w_{2k}^0 \cdot \frac{1}{1 + n_2 \xi_2 \delta_{2k} I(X_{2k} \le t) / \sum_{m=k}^{n_2} \exp(\hat{\beta}^T Z_{2m})},
\end{aligned}
$$

where $w_{1i}^0 = \frac{\delta_{1i}}{\sum_{m=i}^{n_1} \exp(\hat{\beta}^T Z_{1m})}$, $w_{2k}^0 = \frac{\delta_{2k}}{\sum_{m=k}^{n_2} \exp(\hat{\beta}^T Z_{2m})}$ are the Breslow (1972) cumulative hazard increments, and ξ_1 and ξ_2 satisfy constraints (3.7) and (3.8). For any fixed value of θ, by substituting these expressions for p_i and q_k into constraints (3.7) and (3.8), we can obtain $\xi_1 = \xi_1(\eta)$, $\xi_2 = \xi_2(\eta)$. Then if we substitute ξ_1, ξ_2, p_i and q_k as functions of η into the log-likelihood function we obtain the profile log-likelihood function of (θ, η)

$$
\begin{aligned}
l(\theta, \eta) = \sum_{i=1}^{n_1} \delta_{1i} [\log(p_i) + \hat{\beta}^T Z_{1i}] - \sum_{i=1}^{n_1} \{ (\sum_{m=1}^{i} p_m) \cdot \exp(\hat{\beta}^T Z_{1i}) \} + \\
\sum_{k=1}^{n_2} \delta_{2k} [\log(q_j) + \hat{\beta}^T Z_{2k}] - \sum_{k=1}^{n_2} \{ (\sum_{m=1}^{k} q_m) \cdot \exp(\hat{\beta}^T Z_{2k}) \}. \qquad (3.10)
\end{aligned}
$$

Let $\hat{\eta} = \arg\max_\eta l(\theta, \eta)$, and $l_n(\theta) = l(\theta, \hat{\eta})$; then $l_n(\theta)$ is the empirical log-likelihood function of θ and $\hat{\theta} = \arg\max_\theta l_n(\theta)$ is the empirical likelihood estimator of the cumulative hazard ratio $\theta(t)$ after the covariate adjustment.

Without constraints (3.7) and (3.8), the log-likelihood function with covariate adjustment is maximized by the Breslow cumulative hazard increments w_{ji}^0, $i = 1, 2, ...n_j$, $j = 1, 2$, and is equal to

$$\begin{aligned} l_0 &= \sum_{i=1}^{n_1} \delta_{1i}[\log(w_{1i}^0) + \hat{\beta}^T Z_{1i}] - \sum_{i=1}^{n_1}\{\sum_{m=i}^{n_1} \exp(\hat{\beta}^T Z_{1m}) w_{1i}^0\} + \\ & \sum_{k=1}^{n_2} \delta_{2k}[\log(w_{2k}^0) + \hat{\beta}^T Z_{2k}] - \sum_{k=1}^{n_2}\{\sum_{m=k}^{n_2} \exp(\hat{\beta}^T Z_{2m}) w_{2k}^0\}. \end{aligned}$$

Therefore, the empirical log-likelihood ratio is $l_E(\theta) = l_n(\theta) - l_0$.

3.2.1 Asymptotic Properties

To study the limiting distribution of the profile empirical log-likelihood ratio of θ, we assume the following regularity conditions hold for subjects in group j, $j = 1, 2$

(C1) The observed data (X_j, δ_j, Z_j) are independent and identically distributed random vectors.

(C2) Z_{ji} is bounded for all $i = 1, ..., n_j$.

(C3) $\int_0^\tau \lambda_{j0}(s)ds < \infty$ for some prespecified time point τ.

(C4) $s_j^0(t, \beta)$, which is the limiting value of $S_j^0(t, \beta)$ as $n_j \to \infty$, is bounded away from 0 for $t \in [0, \tau]$ and β in a neighborhood of β_0, the true value of the regression parameter in model (3.1).

Let $h(x) = I(x \leq t)$, and write

$$A_{ji} = \frac{\delta_{ji} h(X_{ji})}{\sum_{m=i}^{n_j} \exp(\hat{\beta}^T Z_{jm})/n_j}.$$

Lemma 1. *Under regularity conditions (C1)-(C4), the solutions of constraint (3.7) for* ξ_1 *and (3.8) for* ξ_2 *satisfy*

$$\xi_1 = \frac{1/n_1 \sum_{i=1}^{n_1} A_{1i} - \hat{\eta}}{1/n_1 \sum_{i=1}^{n_1-1} A_{1i}^2} + o_p(n_1^{-1/2}),$$

$$\xi_2 = \frac{1/n_2 \sum_{i=1}^{n_2} A_{2i} - \theta\hat{\eta}}{1/n_2 \sum_{i=1}^{n_2-1} A_{2i}^2} + o_p(n_2^{-1/2}).$$

Therefore,

$$\sqrt{n_j}\xi_j \xrightarrow{D} N(0, [\sigma_j^2(h)]^{-1}),$$

where $\sigma_j^2(h) = \int \frac{h^2(x) d\Lambda_{j0}(x)}{s_j^0(x,\beta_0)(1-G_j(x))}$.

Proof. Apply Lemma 1 of Pan and Zhou (2002) to $\xi_j, j = 1, 2$. □

Theorem 1. *Under regularity conditions (C1)-(C4), the empirical log-likelihood ratio* $l_E(\theta)$ *satisfies* $-2l_E(\theta) \xrightarrow{D} \chi_1^2$.

Proof. Note that

$$\begin{aligned} p_i &= w_{1i}^0 \cdot \tfrac{1}{1+\xi_1 A_{1i}}, \quad && i = 1, ..., n_1 - 1 \\ &= w_{1n_1}^0, && i = n_1 \\ q_k &= w_{2k}^0 \cdot \tfrac{1}{1+\xi_2 A_{2k}}, && k = 1, ..., n_2 - 1 \\ &= w_{2n_2}^0, && k = n_2 \end{aligned}$$

Thus, we can rewrite $l_n(\theta)$ in terms of w_{ji}^0 as follows:

$$
\begin{aligned}
l_n(\theta) &= \sum_{i=1}^{n_1} \delta_{1i}[\log(p_i) + \hat{\beta}^T Z_{1i}] - \sum_{i=1}^{n_1}(\sum_{m=i}^{n_1} \exp\{\hat{\beta}^T Z_{1m}\}) \cdot p_i + \\
&\quad \sum_{k=1}^{n_2} \delta_{2k}[\log(q_k) + \hat{\beta}^T Z_{2k}] - \sum_{k=1}^{n_2}(\sum_{m=k}^{n_2} \exp\{\hat{\beta}^T Z_{2m}\}) \cdot q_k \\
&= \sum_{i=1}^{n_1-1} \delta_{1i} \log(\frac{w_{1i}^0}{1+\xi_1 A_{1i}}) + \sum_{i=1}^{n_1} \delta_{1i}\hat{\beta}^T Z_{1i} - \sum_{i=1}^{n_1-1}\{\sum_{m=i}^{n_1} \exp(\hat{\beta}^T Z_{1m}) \frac{w_{1i}^0}{1+\xi_1 A_{1i}}\} + \\
&\quad \delta_{1n_1} \log(w_{1n_1}^0) - w_{1n_1}^0 \exp(\hat{\beta}^T Z_{1n_1}) + \sum_{k=1}^{n_2-1} \delta_{2k} \log(\frac{w_{2k}^0}{1+\xi_2 A_{2k}}) + \sum_{k=1}^{n_2} \delta_{2k}\hat{\beta}^T Z_{2k} - \\
&\quad \sum_{k=1}^{n_2-1}\{\sum_{m=k}^{n_2} \exp(\hat{\beta}^T Z_{2m}) \frac{w_{2k}^0}{1+\xi_2 A_{2k}}\} + \delta_{2n_2} \log(w_{2n_2}^0) - w_{2n_2}^0 \exp(\hat{\beta}^T Z_{2n_2}).
\end{aligned}
$$

Without constraints (3.7) and (3.8), the maximized value of the log-likelihood function with covariate adjustment is:

$$
\begin{aligned}
l_0 &= \sum_{i=1}^{n_1} \delta_{1i}[\log(w_{1i}^0) + \hat{\beta}^T Z_{1i}] - \sum_{i=1}^{n_1}\{\sum_{m=i}^{n_1} \exp(\hat{\beta}^T Z_{1m}) w_{1i}^0\} + \\
&\quad \sum_{k=1}^{n_2} \delta_{2k}[\log(w_{2k}^0) + \hat{\beta}^T Z_{2k}] - \sum_{k=1}^{n_2}\{\sum_{m=k}^{n_2} \exp(\hat{\beta}^T Z_{2m}) w_{2k}^0\}.
\end{aligned}
$$

Therefore, the logarithmic profile empirical likelihood ratio for θ is

$$
\begin{aligned}
l_n(\theta) - l_0 &= \sum_{i=1}^{n_1-1} \delta_{1i} \log(\frac{1}{1+\xi_1 A_{1i}}) - \sum_{i=1}^{n_1-1}\{\sum_{m=i}^{n_1} \exp(\hat{\beta}^T Z_{1m})(\frac{w_{1i}^0}{1+\xi_1 A_{1i}} - w_{1i}^0)\} + \\
&\quad \sum_{k=1}^{n_2-1} \delta_{2k} \log(\frac{1}{1+\xi_2 A_{2k}}) - \sum_{k=1}^{n_2-1}\{\sum_{m=k}^{n_2} \exp(\hat{\beta}^T Z_{2m})(\frac{w_{2k}^0}{1+\xi_2 A_{2k}} - w_{2k}^0\}.
\end{aligned}
$$

Since

$$w_{ji}^0 = \frac{\delta_{ji}}{\sum_{m=i}^{n_j} \exp(\hat{\beta}^T Z_{jm})},$$

we can use the Taylor expansions of $\frac{1}{1+x} = 1 - x + x^2 + O(x^3)$ and $\log(1+x) = x - \frac{1}{2}x^2 + O(x^3)$ to obtain

$$\begin{aligned} l_n(\theta) - l_0 &= \sum_{i=1}^{n_1-1} \delta_{1i} \log(\frac{1}{1+\xi_1 A_{1i}}) + \sum_{i=1}^{n_1-1} \{\delta_{1i}\xi_1 A_{1i} - \delta_{1i}(\xi_1 A_{1i})^2 + O_p(|\xi_1 A_{1i}|^3)\} + \\ & \sum_{k=1}^{n_2-1} \delta_{2k} \log(\frac{1}{1+\xi_2 A_{2k}}) + \sum_{k=1}^{n_2-1} \{\delta_{2k}\xi_2 A_{2k} - \delta_{2k}(\xi_2 A_{2k})^2 + O_p(|\xi_2 A_{2k})|^3\} \\ &= -\sum_{i=1}^{n_1-1} [\frac{1}{2}\delta_{1i}(\xi_1 A_{1i})^2 + O_p(|\xi_1 A_{1i}|^3)] - \sum_{k=1}^{n_2-1} [\frac{1}{2}\delta_{2k}(\xi_2 A_{2k})^2 + O_p(|\xi_2 A_{2k}|^3)]. \end{aligned}$$

Since $\delta_{ji} A_{ji} = A_{ji}$, we have

$$-2(l_n(\theta) - l_0) = \sum_{i=1}^{n_1-1} \{\xi_1^2 A_{1i}^2 + O_p(|\xi_1 A_{1i}|^3)\} + \sum_{k=1}^{n_2-1} \{\xi_2^2 A_{2k}^2 + O_p(|\xi_2 A_{2k}|^3)\} \quad (3.11)$$

where

$$\begin{aligned} \sum_{i=1}^{n_j-1} O_p(|\xi_j A_{ji}|^3) &\leq O_p(|\xi_j|^3) O_p(\max|A_{ji}|) \sum_{i=1}^{n_j-1} A_{ji}^2 \\ &\leq O_p(n_j^{-\frac{1}{2}}) o_p(n_j^{\frac{1}{2}}) \cdot \frac{1}{n_j} \sum_{i=1}^{n_j-1} A_{ji}^2 \\ &= o_p(1) \qquad as \qquad n_j \to \infty. \end{aligned}$$

Also, as $n_j \to \infty$, $\hat{\beta} \xrightarrow{p} \beta_0$; therefore,

$$\frac{1}{n_j}\sum_{i=1}^{n_j-1} A_{ji}^2 = \frac{1}{n_j}\sum_{i=1}^{n_j} A_{ji}^2 = \int \frac{h^2(X_{ji})w_{ji}^0}{\sum_{k=1}^{n_j}\exp(\hat{\beta}^T Z_{jk})/n_j} \xrightarrow{p} \int \frac{h^2(x)d\Lambda_{j0}(x)}{s_j^0(x,\beta_0)(1-G_j(x))} < \infty.$$

And from Lemma 1,

$$\xi_1 = \frac{1/n_1 \sum_{i=1}^{n_1} A_{1i} - \hat{\eta}}{1/n_1 \sum_{i=1}^{n_1-1} A_{1i}^2} + o_p(n_1^{-1/2}),$$

$$\xi_2 = \frac{1/n_2 \sum_{i=1}^{n_2} A_{2i} - \theta\hat{\eta}}{1/n_2 \sum_{i=1}^{n_2-1} A_{2i}^2} + o_p(n_2^{-1/2}).$$

By Theorem (3.1) of Pakes and Pollard (1989) we have $\hat{\eta} \xrightarrow{p} \eta$. Therefore, by Slutsky's theorem, each term of expression (3.11) converges in distribution to a χ_1^2 random variable. However since we are profiling with respect to the variable η, the logarithmic profile empirical likelihood ratio satisfies $-2(l_n(\theta) - l_0) \xrightarrow{D} \chi^2$ with $2 - 1 = 1$ degree of freedom. □

3.3 Empirical Likelihood Estimation of a Group-specific Covariate-adjusted Cumulative Hazard Ratio

Instead of assuming that the covariate effects are the same for both groups of subjects, we now consider situations in which the covariate adjustments in each group are different. For example, patients with the same blood pressure level may experience differential effects on their respective times to response. For these situations we consider the model

$$\lambda_{ji}(t) = \lambda_{j0}(t)\exp(\beta_j^T Z_{ji}) \tag{3.12}$$

for $i = 1, 2, ..., n_j$ and $j = 1, 2$. Note that $\beta_1 \neq \beta_2$ so that the assumed model is no longer a stratified proportional hazards regression model but one with a different covariate adjustment within each group of subjects.

Let $\hat{\beta}_j$ be the regression estimator for this PH model in group j. Then the cumulative hazard ratio after covariate adjustment is

$$\theta(t) = \frac{\Lambda_{20}(t)}{\Lambda_{10}(t)}. \tag{3.13}$$

The usual estimator based on a normal approximation is

$$\hat{\theta}(t) = \frac{\hat{\Lambda}_{20}(\hat{\beta}_2, t)}{\hat{\Lambda}_{10}(\hat{\beta}_1, t)}, \tag{3.14}$$

where $\hat{\Lambda}_{j0}(\hat{\beta}_j, t)$ is the Breslow (1972) estimator.

If we replace the value of $\hat{\beta}$ associated with group j by $\hat{\beta}_j$ in expression (3.6), we obtain a profile empirical likelihood function for θ after covariate adjustment, which is

$$EL(\theta) = \sup_{(p,q,\eta)} \left(\prod_{i=1}^{n_1} [p_i \exp(\hat{\beta}_1^T Z_{1i})]^{\delta_{1i}} \exp\{-(\sum_{m=1}^{i} p_m) \cdot \exp(\hat{\beta}_1^T Z_{1i})\} \right) \cdot$$

$$\left(\prod_{k=1}^{n_2} [q_k \exp(\hat{\beta}_2^T Z_{2k})]^{\delta_{2k}} \exp\{-(\sum_{m=1}^{k} q_m) \cdot \exp(\hat{\beta}_2^T Z_{2k})\} \right), \tag{3.15}$$

subject to constraints (3.7) and (3.8) as well as $p_i > 0, q_k > 0, i = 1, ..., n_1, k =$

$1, ..., n_2$.

By simply adapting our previous results for $i = 1, ..., n_1 - 1, k = 1, ..., n_2 - 1$, we have

$$\begin{aligned} p_i &= \frac{\delta_{1i}}{\sum_{m=i}^{n_1} \exp(\hat{\beta}_1^T Z_{1m}) + n_1 \lambda_1 \delta_{1i} I(X_{1i} \leq t)} \\ &= \frac{\delta_{1i}}{\sum_{m=i}^{n_1} \exp(\hat{\beta}_1^T Z_{1m})} \cdot \frac{1}{1 + n_1 \lambda_1 \delta_{1i} I(X_{1i} \leq t) / \sum_{m=i}^{n_1} \exp(\hat{\beta}_1^T Z_{1m})} \\ &= w_{1i}^0 \cdot \frac{1}{1 + n_1 \lambda_1 \delta_{1i} I(X_{1i} \leq t) / \sum_{m=i}^{n_1} \exp(\hat{\beta}_1^T Z_{1m})}, \end{aligned}$$

$$\begin{aligned} q_k &= \frac{\delta_{2k}}{\sum_{m=k}^{n_2} \exp(\hat{\beta}_2^T Z_{2m}) + n_2 \lambda_2 \delta_{2k} I(X_{2k} \leq t)} \\ &= \frac{\delta_{2k}}{\sum_{m=k}^{n_2} \exp(\hat{\beta}_2^T Z_{2m})} \cdot \frac{1}{1 + n_2 \lambda_2 \delta_{2k} I(X_{2k} \leq t) / \sum_{m=k}^{n_2} \exp(\hat{\beta}_2^T Z_{2m})} \\ &= w_{2k}^0 \cdot \frac{1}{1 + n_2 \lambda_2 \delta_{2k} I(X_{2k} \leq t) / \sum_{m=k}^{n_2} \exp(\hat{\beta}_2^T Z_{2m})}. \end{aligned}$$

Substituting p_i and q_k into the empirical log-likelihood function for θ we obtain

$$l(\theta, \eta) = \sum_{i=1}^{n_1} \delta_{1i} [\log(p_i) + \hat{\beta}_1^T Z_{1i}] - \sum_{i=1}^{n_1} \{ \sum_{m=1}^{i} p_m \cdot \exp(\hat{\beta}_1^T Z_{1i}) \} +$$

$$\sum_{k=1}^{n_2} \delta_{2k} [\log(q_j) + \hat{\beta}_2^T Z_{2k}] - \sum_{k=1}^{n_2} \{ \sum_{m=1}^{k} q_m \cdot \exp(\hat{\beta}_2^T Z_{2k}) \}. \tag{3.16}$$

Let $l_n(\theta) = \max_\eta l(\theta, \eta)$; then $l_n(\theta)$ is the profile empirical log-likelihood function of θ, and $\hat{\eta} = \arg\max_\eta l(\theta, \eta)$. Clearly, the empirical likelihood estimator for the cumulative hazard ratio θ under group-specific covariate adjustment is $\hat{\theta} = \arg\max_\theta l_n(\theta)$.

As in the previous section, we define the empirical log-likelihood ratio $l_E(\theta) =$

$l_n(\theta) - l_0$, where

$$\begin{aligned} l_0 &= \sum_{i=1}^{n_1} \delta_{1i}[\log(w_{1i}^0) + \hat{\beta}_1^T Z_{1i}] - \sum_{i=1}^{n_1}\{(\sum_{m=i}^{n_1} \exp(\hat{\beta}_1^T Z_{1m}) w_{1i}^0\} + \\ & \sum_{k=1}^{n_2} \delta_{2k}[\log(w_{2k}^0) + \hat{\beta}_2^T Z_{2k}] - \sum_{k=1}^{n_2}\{(\sum_{m=k}^{n_2} \exp(\hat{\beta}_2^T Z_{2m}) w_{2k}^0)\}. \end{aligned}$$

Note that $w_{ji}^0 = dN_{ji}(X_{ji})/\{n_j S_j^0(X_{ji}, \hat{\beta}_j)\}$ for $i = 1, 2, ..., n_j$, and $j = 1, 2$.

To derive the asymptotic distribution of $l_n(\theta)$ we need to modify regularity condition (C4) to require

$(C4)'$ $s_j^0(t, \beta)$, which is the limiting value of $S_j^0(t, \beta)$ as $n_j \to \infty$, is bounded away from 0 for $t \in [0, \tau]$ and β in a neighborhood of β_j, the true value of the vector of regression coefficients in model (3.12).

Theorem 2. *Under regularity conditions $(C1) - (C3)$ and $(C4)'$, the empirical log-likelihood ratio $l_E(\theta)$ satisfies $-2l_E(\theta) \xrightarrow{D} \chi_1^2$.*

Proof. As we previously showed in the proof of Theorem 1, the analog of the first and second terms in the expansion of $l_E(\theta)$ each converge independently to a χ_1^2 random variable. However, since the parameter of interest is a scalar quantity for each fixed value of t, we estimated θ by profiling the joint empirical likelihood function with respect to the value of η. Thus, the logarithmic profile empirical likelihood ratio, $-2l_E(\theta) = -2(l_n(\theta) - l_0)$, has a limiting distribution that is χ_1^2 as $n_j \to \infty,\ j = 1, 2$. □

Using the empirical log-likelihood ratio statistic we can construct the 100(1 −

α)% confidence interval,

$$I_{1-\alpha} = \{\theta : -2l_E(\theta) \leq q_{1,(1-\alpha)}\}$$

for θ, where $q_{1,(1-\alpha)}$ denotes the $(1-\alpha)$-quantile of χ_1^2.

3.4 Simulation Study for the EL-based Estimator of the Cumulative Hazard Ratio Under the Stratified Model

Wei and Schaubel (2008) investigated the properties of the normal approximation for obtaining point estimates and point-wise interval estimates of the covariate-adjusted cumulative hazard ratio for treatment effect. However when the sample sizes are small, i.e., $n_1 + n_2 = 50$, the estimated coverage probability of their 95% confidence interval is no more than 92%. To compare the coverage accuracy of our empirical likelihood estimator with their normal approximation at a nominal level of 95%, we adopted the same simulation design that they described.

Let $T_{ji}, i = 1, ..., n_j, j = 1, 2$, be the event times. These are generated via the transformation

$$T_{ji} = \{-\log(U_{ji})/[\alpha_j \exp(\beta_0 Z_{ji})]\}^{1/\gamma_j}$$

where U_{ji} is a uniform $(0, 1)$ random variable, $\beta_0 = 0.5$, $Z_{ji} \sim \text{Bernoulli}(0.5)$. In this set-up, $\{T_{ji}\}$ follows a Weibull distribution with hazard function

$$\lambda_{ji}(t) = \alpha_j \gamma_j t^{\gamma_j - 1} \exp(\beta_0 Z_{ji}).$$

Therefore, within each of the two strata, the hazards that correspond to distinct values of Z are proportional. By choosing different values of $\gamma_j,\ j = 1, 2$, we ensure that the baseline hazard functions for the two groups will not be proportional. Let the censoring times $C_{ji} \sim \text{uniform}(2.5, 5)$. By varying the value of α_j we can adjust the proportion of censoring. For sample sizes $n = 50, 70, 100, 200, 500$, we used the Monte Carlo method to generate 1000 replicate samples, each involving a total of n observations. From each replicate sample we calculated the point-wise 95% confidence interval at the 75th percentile of the combined observation times in the two groups. The study results are summarized in Table 3.1 as estimated coverage probabilities for the resulting interval estimates.

The results in Table 3.1 show that our empirical likelihood estimator has an estimated coverage probability that is closer to the nominal value of 95% than the corresponding value for the normal approximation reported by Wei and Schaubel (2008), when the sample size is small. Wei and Schaubel (2008) reported that when the total sample size is 50, the estimated coverage probability of their normal approximation is no more than 92%. Clearly, the empirical likelihood estimator has an estimated coverage probability very close to the nominal level of 95%, even when a high proportion of the observations are right censored (say 40%). Also, unlike the symmetric interval estimates generated via the normal approximation, the confidence regions produced by the empirical likelihood method directly reflect

Table 3.1: Estimated coverage probabilities for adjusted cumulative hazard ratio interval estimates of treatment effect, at a nominal level of 95%. C% represents percent censored; C.P. represents coverage probability.

γ_1	γ_2	α_1	α_2	n_1	n_2	C%	C.P.
1.4	1.2	0.4	0.35	25	25	0%	94.6%
				30	40	0%	95.7%
				50	50	0%	94.4%
				100	100	0%	94.5%
				250	250	0%	94.8%
1.4	1.2	0.4	0.35	25	25	10%	95.1%
				30	40	10%	94.1%
				50	50	10%	94.4%
				100	100	10%	94.6%
				250	250	10%	95.0%
1	1.5	0.2	0.1	25	25	40%	93.7%
				30	40	40%	94.1%
				50	50	40%	94.9%
				100	100	40%	94.9%
				250	250	40%	95.0%

the shape of the data, which should be more appropriate in practice.

Since Wei and Schaubel (2008) also consider the log-transformation to improve the coverage probability when the sample is small, we compare it with our EL method in terms of coverage accuracy and average length of the estimated cumulative hazard ratio at the 75th percentile of the total observation time under a stratified model when the total sample size is 50. The results are given in Table 3.2. From these simulation results we find that the EL method has better coverage accuracy and a slightly wider confidence interval than the log transformation, and both of methods of estimation outperform the usual normal transformation.

Table 3.2: Estimated coverage probabilities and average lengths (in parentheses) for adjusted cumulative hazard ratio interval estimates of treatment effect under a stratified model, at a nominal level of 95%. C% represents percent censored; Log represents the logarithmic ratio; EL represents the empirical likelihood.

γ_1	γ_2	α_1	α_2	n_1	n_2	C%	Log	EL
1.4	1.2	0.4	0.35	25	25	0%	94.7%(1.062)	94.9%(1.091)
1.4	1.2	0.4	0.35	25	25	10%	94.6%(1.116)	94.9%(1.141)
1	1.5	0.2	0.1	25	25	40%	94.5%(1.811)	94.9%(1.842)

3.5 Simulation Study for the EL-based Estimator of the Group-Specific Cumulative Hazard Ratio

To investigate the coverage probability of the group-specific covariate adjustment method that we described in §3.3, we used the same simulation set-up as we described above, except that $\beta_1 = 0.5$ and $\beta_2 = 1.5$. We obtained the estimated coverage probabilities of 95% confidence intervals based on the empirical likelihood estimator with total sample sizes $50, 100, 200, 500$. The simulation results are summarized in Table 3.3.

Since the maximum partial likelihood estimator used in the group-specific adjustment specified in formula (3.12) is evaluated separately in each of the groups, we should anticipate some loss of efficiency compared to the results that we obtained when covariate adjustment is based on a stratified proportional hazards regression model. Therefore, it is not surprising that the estimated coverage probabilities for the empirical likelihood estimator of θ summarized in Table 3.3 are noticeably lower

Table 3.3: Estimated coverage probabilities for group-specific adjusted cumulative hazard ratio interval estimates of treatment effect, at a nominal level of 95%. C% represents percent censored; C.P. represents coverage probability.

γ_1	γ_2	α_1	α_2	n_1	n_2	C%	C.P.
1.4	1.2	0.4	0.35	25	25	0%	88.6%
				50	50	0%	89.1%
				100	100	0%	89.1%
				250	250	0%	89.4%
1.4	1.2	0.4	0.35	25	25	9%	87.6%
				50	50	9%	90.2%
				100	100	9%	89.6%
				250	250	9%	90.3%
1	1.5	0.115	0.1	25	25	40%	81.0%
				50	50	40%	82.2%
				100	100	40%	81.3%
				250	250	40%	82.4%

than the corresponding values that we report in Table 3.1. In addition, it appears that the statistical behaviour of our empirical likelihood estimator is more sensitive to the effects of right censoring. In particular, when right censoring of the data is severe, e.g., 40%, coverage errors increase markedly.

However, since empirical likelihood is Bartlett correctable, we can use the bootstrap method to derive a null distribution that provides better calibration for the empirical likelihood estimator with group-specific covariate adjustment.

3.5.1 Bartlett Correction

As one of key properties of empirical likelihood, Bartlett correction is a delicate second-order property, implying that a simple mean adjustment to the likelihood ratio can improve the approximation to the limiting chi-square distribution by one order of magnitude. Therefore, it can be used to enhance the coverage accuracy of likelihood-based confidence regions. In the context of testing hypotheses, Bartlett correction reduces the errors between the nominal and actual significant levels of an EL-based test.

Following the arguments in the previous sections, we have

$$P\{-2l_E(\theta) < z\} = P(\chi_1^2 < z) + O(n^{-1})$$

Using the Edgeworth expansion of the test statistic $-2l_E(\theta)$, we can obtain an adjustment a such that

$$P\{-2l_E(\theta) < (1 + an^{-1})z\} = P(\chi_1^2 < z) + O(n^{-2})$$

The exact formula for a can be very complex. However, we can use the bootstrap method suggested by Chen and Cui (2007) to obtain $\hat{a}$, an estimator of a, to improve the coverage accuracy of θ at a specified significance level of α.

To implement Bartlett correction in a general situation, the adjustment value a has to be estimated. Due to the complexity of the Edgeworth expansion, the formula for a can be lengthy; therefore, we adapt the following bootstrap estimator $\hat{\gamma} = 1 + \hat{a}n^{-1}$ to replace $(1 + an^{-1})$ in Bartlett correction.

Step 1: Generate a bootstrap resample $(X_i^*, Y_i^*)_{i=1}^n$ by sampling with replacement from the original sample $(X_i, Y_i)_{i=1}^n$ and compute $d^*(\hat{\theta}) = -2l_E^*(\hat{\theta})$, where $\hat{\theta}$ is the empirical likelihood estimator based on the original sample, and l^* is the logarithm of the empirical likelihood ratio based on the bootstrap sample.

Step 2: For a large integer B, repeat Step 1 B times and obtain $d_1^*(\hat{\theta}), ..., d_B^*(\hat{\theta})$. Then $\hat{\gamma} = 1/B \sum_{i=1}^{B} d_i^*(\hat{\theta})$.

Following standard bootstrap arguments, (see Hall (1992) for details), we have

$$E(\hat{\gamma}) = (1 + an^{-1})\{1 + O_p(n^{-1/2})\}.$$

Therefore, $\hat{\gamma}$ is a $\sqrt{n}-$consistent estimator of $(1+an^{-1})$. The corresponding critical region based on Bartlett correction is

$$I_{BC} = \{\theta : l(\theta) > \hat{\gamma} q_{1,(1-\alpha)}\}.$$

The above use of the bootstrap to estimate γ can be computationally intensive when B is large. Instead of using the bootstrap for $\hat{\gamma}$, one can use a bootstrap quantile to calibrate the logarithm of the empirical likelihood ratio directly. Let $\hat{q}_{b,(1-\alpha)}$ be the $([B(1-\alpha)]+1)$ ordered value of $-2l_E^*(\hat{\theta})_{i=1}^B$. Then a direct bootstrap critical region at a nominal level α is $I_B = \{\theta : -2l_E(\theta) > \hat{q}_{b,(1-\alpha)}\}$.

3.5.2 EL-based Bootstrap

The bootstrap method of inference was first introduced by Efron (1979) for complete data, and then by Efron (1981) and Reid (1981) for censored data. Using simulation studies, Efron (1982) showed that confidence intervals produced by the bootstrap method are more accurate than those based on the asymptotic distributions of parameter estimators. Akritas(1986) investigated the bootstrapping of the

Kaplan-Meier estimator and found that the bootstrap confidence band has more accurate coverage than the Hall-Wellner (HW) band, especially in small sample scenarios. In our problem setting involving a group-specific cumulative treatment effect, following Chen and Cui (2007), our EL-based bootstrap method approximates Bartlett correction under certain regularity conditions. Therefore, we should expect to reduce the coverage errors to $O(n^{-2})$, and correspondingly observe improved coverage accuracy for EL-based confidence intervals of the cumulative hazard ratio with covariate adjustments.

For the given sample $(X_{j1}, \delta_{j1}), \ldots, (X_{jn_j}, \delta_{jn_j}),\ j = 1, 2$, let $\hat{\theta}$ be the point estimate of the group-specific covariate-adjusted baseline cumulative hazard ratio, where $\hat{\beta}_j$ is the estimated regression parameter for group j. Following Efron (1981), we take a bootstrap sample $(X^*_{j1}, \delta^*_{j1}), \ldots, (X^*_{jn_j}, \delta^*_{jn_j})$ for each of the two groups by associating the probability mass n_j^{-1} with each of the observed pairs in a group, and then drawing a bootstrap sample of size n_j, with replacement, from the data for group j.

From each pair of bootstrap samples from the two groups, we calculate the corresponding value of the profile empirical log-likelihood ratio $l^*_E(\hat{\theta})$. Repeating this series of bootstrap sample calculations B times generates a bootstrap sampling distribution for the logarithmic profile empirical likelihood ratio. By extracting the $100(1-\alpha)\%$ quantile from this bootstrap distribution and using it as the critical value, we can generate a corresponding pointwise $100(1-\alpha)\%$ confidence interval for $\theta(t)$.

Simulation results for this bootstrap procedure are summarized in Table 3.4 in terms of estimated coverage probabilities for approximate 95% confidence intervals. Evidently, using critical values from the bootstrap sampling distribution leads to noticeable improvements in the coverage accuracy of our empirical likelihood-based procedure in this group-specific covariate adjustment problem setting. The estimated coverage probabilities are very close to the nominal value of 0.95 except when the proportion of censored observations is fairly substantial, e.g., roughly 40% of the combined sample size.

Table 3.4: Estimated coverage probabilities for group-specific adjusted cumulative hazard ratio interval estimates of treatment effect, at a nominal level of 95%. The interval estimates were obtained using bootstrap critical values for the logarithmic profile empirical likelihood ratio.

γ_1	γ_2	α_1	α_2	n_1	n_2	C%	C.P.
1.4	1.2	0.4	0.35	25	25	0%	94.7%
				50	50	0%	94.5%
				100	100	0%	94.6%
				250	250	0%	94.6%
1.4	1.2	0.4	0.35	25	25	9%	94.6%
				50	50	9%	94.7%
				100	100	9%	94.6%
				250	250	9%	94.8%
1	1.5	0.115	0.1	25	25	40%	92.0%
				50	50	40%	92.2%
				100	100	40%	92.6%
				250	250	40%	92.7%

We also compare the coverage accuracy and average length of the interval-estimated cumulative hazard ratio at the 75th percentile of the total observation time under a group-specific model using a normal approximation, a logarithmic transformation, the EL method directly, as well as the bootstrap method, when the total sample size is 50. These comparisons are summarized in Table 3.5.

Table 3.5: Estimated coverage probabilities and average length for adjusted cumulative hazard ratio interval estimates of treatment effect under a group-specific model, at a nominal level of 95%. C% represents percent censored; Log represents the logarithmic ratio; EL indicates the empirical likelihood method; EB represents empirical likelihood using the bootstrap procedure.

γ_1	γ_2	α_1	α_2	n_1	n_2	C%	Normal	Log	EL	EB
1.4	1.2	0.4	0.35	25	25	0%	84.7%(1.657)	86.6%(1.634)	87.9%(1.784)	94.7%(2.251)
1.4	1.2	0.4	0.35	25	25	10%	82.3%(1.712)	85.4%(1.691)	87.6%(1.803)	94.6%(2.345)
1	1.5	0.2	0.1	25	25	40%	76.3%(1.952)	78.3%(1.910)	81.0%(2.025)	92.0%(2.549)

The simulations show that the EL method outperforms both the normal approximation and the logarithmic transformation in terms of achieving a coverage probability that is closer to the nominal level of 95%. Since the sample size of each group is small, using the asymptotic critical value from χ_1^2 for the empirical log-likelihood ratio of θ does not work particularly well. However, the EL-bootstrap method reduces the coverage error significantly. Consequently, the coverage probability for our EL-based bootstrap method is much closer to the nominal level of 95%. Of course, the average lengths of the EL-based interval estimates are somewhat greater than the alternatives based on the usual normal approximations. However, this increase in average length should be expected, and represents the cost of superior coverage probability.

3.6 Applications

3.6.1 Non-Hodgkin's Lymphoma Data Analysis

To illustrate the use of our empirical likelihood approach in a practical problem, we analyzed data from Matthews and Farewell (2007) concerning 64 non-Hodgkin's lymphoma patients with different stages of disease at diagnosis to compare their survival experience. The data also include information about presenting symptoms and bulky disease as covariates, both of which are statistically significant with respect to patient survival in our stratified model (stratified by stage, stage IV vs Stage II or III disease). We chose to adopt a stratified model based on a likelihood ratio test, i.e., two times the difference of the maximized log-likelihood function between the stratified model and the non-stratified model with different regression parameters for the two groups, which is 2.6, much less than 3.84, the 95% quantile of χ_1^2.

To obtain the cumulative hazard ratio for Stage IV versus Stage II or III disease, adjusted for the effect of presenting symptoms and bulky disease, we first estimated the regression parameter for symptoms and bulky disease using the stratified model. The estimated regression coefficients were 1.11 for presenting symptoms and 1.80 for bulky disease, with corresponding estimated standard errors of 0.41 and 0.69. Then we used the empirical likelihood method outlined in section 2 to derive the point estimates and pointwise 95% confidence intervals for the covariate-adjusted cumulative baseline hazard ratio. The resulting estimates are displayed in Figure 3.1.

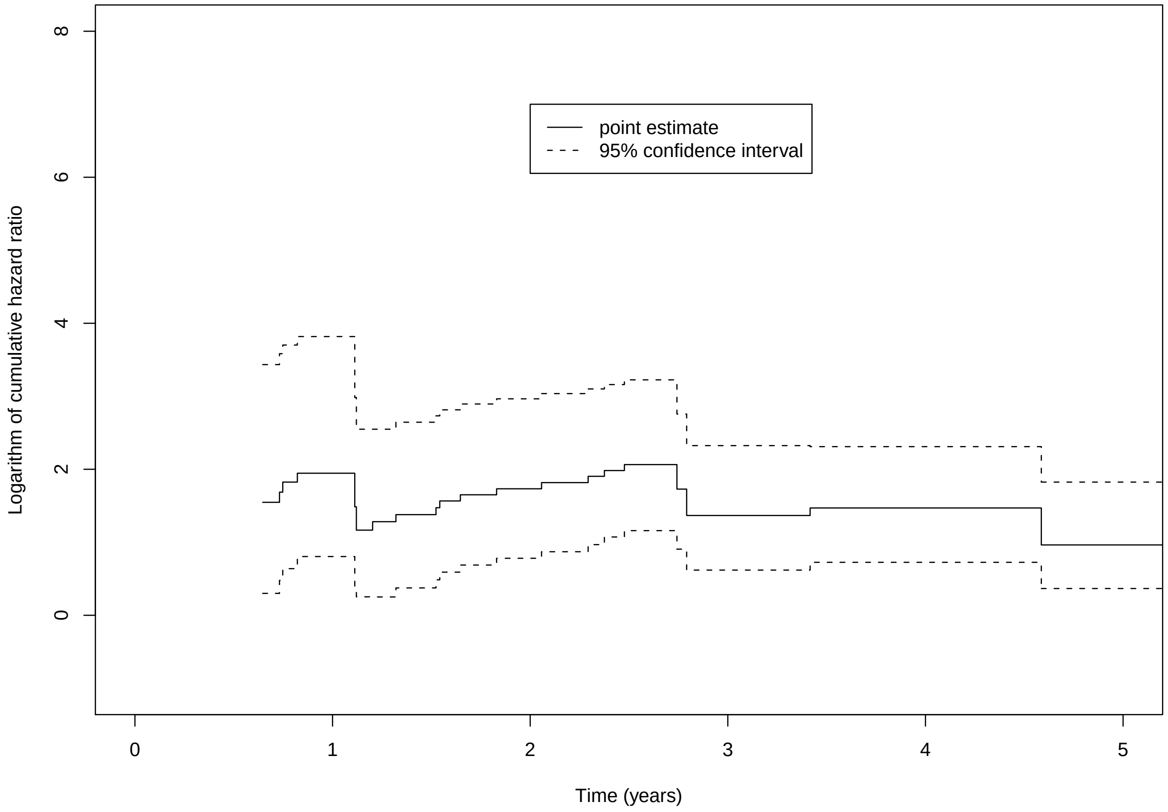

Figure 3.1: Estimated baseline cumulative hazard ratios of death for non-Hodgkin's patients with Stage IV compared to Stage II or III disease, adjusted for the effect of presenting symptoms and the presence of bulky disease

Prior to the eight-month mark, the estimated cumulative hazard ratio does not exist because observed deaths had not occurred in both groups of patients. Thereafter, the estimated ratio increases slightly as observed deaths occur among Stage IV patients. When two Stage III patients die at roughly the same time (406 and 409 days following diagnosis), the estimated cumulative hazard ratio declines. Overall, since many patients with very advanced (Stage IV) disease are observed to die, whereas Stage II and III patients give rise to right-censored observations, the estimated ratio tends to increase gradually with time, with only occasional declines observed as the time from diagnosis increases. From this plot we see that all the lower bounds of the pointwise interval estimates of the cumulative hazard ratio exceed 1 implying that the cumulative hazard function for patients with Stage IV disease is statistically greater than that for Stage II or III patients, after adjusting for the effect of presenting symptoms and bulky disease at diagnosis. We conclude that the adjusted risk of death for patients with Stage IV disease is greater than that for patients with Stage II or III disease at diagnosis. However that risk ratio does not appear to be constant, but increases or decreases over time, especially during the first three years following diagnosis with non-Hodgkin's lymphoma.

3.6.2 Ovarian Cancer Data Analysis

To illustrate the results outlined in section 3, we use data from an observational study of 146 ovarian cancer patients that were kindly provided by a Finnish researcher. Each patient had six covariates that were recorded at the beginning of follow-up — disease stage, grade, patient age, an indicator of residual tumor size, as well as the values of human chorionic gonadotropin beta (hcg) and ca125, a particular cancer antigen. After fitting separate proportional hazard regression models for each covariate measured, we found that both the amount of hcg and the logarithmic ca125 measurements affected patient survival, and their effects differed in the two groups of patients according to the residual tumor size when patient follow-up began. Since there was no residual tumor size measurement for one study subject, we divided the remaining 145 patients into two groups, 41 with at least a 1 cm residual tumor and 104 with little or no residual disease. Within each group, we adjusted the survival experience for the combined effects of hcg and logarithmic ca125 concentrations, and then estimated the ratio of the two baseline cumulative hazards for these ovarian cancer patients.

For patients with little or no residual tumor, the estimated regression coefficients for hcg and log ca125 were 0.242 and 0.487, with corresponding estimated standard errors of 0.087 and 0.137, respectively. Among the 41 patients with a residual tumor exceeding 1 cm in diameter, the estimated regression coefficients and estimated standard errors for these effects on patient survival were 0.052 (0.026) and 0.199 (0.134), respectively. The point and interval estimates of the resulting ratio are displayed in Figure 3.2. Since the pointwise 95% lower confidence bounds of this baseline cumulative hazard ratio all exceed 1, we conclude that the risk of death for patients in the group with a residual tumor after treatment greater than 1 cm is distinctly greater than that experienced by patients with little or no residual tumor, after adjustment for the differential effect of hcg and logarithmic ca125 in these two patient groups. This estimated risk ratio appears to be most elevated during the second year following treatment, and then gradually decreases to a long-term stable value of roughly 2.7 on the logarithmic scale.

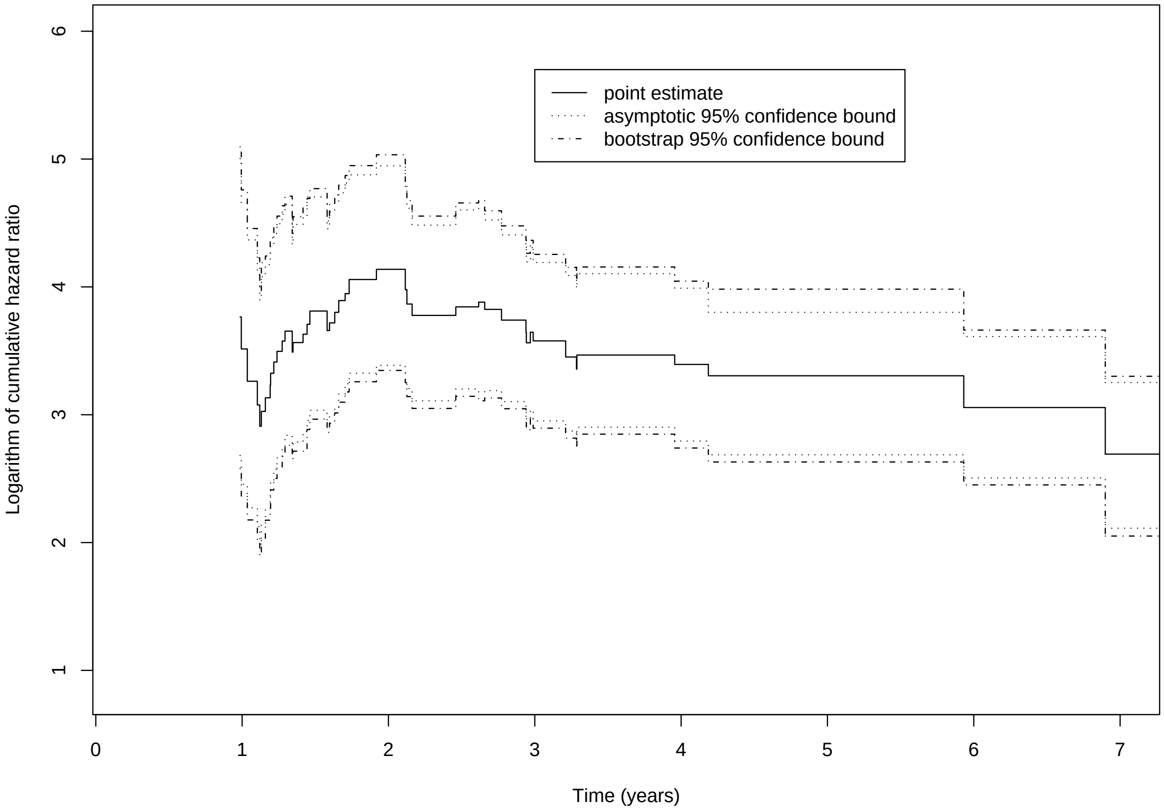

Figure 3.2: Logarithm of estimated baseline cumulative hazard ratios for ovarian cancer patients with residual tumors exceeding 1 cm compared to those with little or no residual tumor, adjusted for the differential effects of hcg and logarithmic ca125 on patient survival experience.

3.7 Discussion

Despite the widespread use of proportional hazards regression modelling of data from observational and randomized studies, the justification for doing so is frequently overlooked, or rarely mentioned in published reports. When the hazard functions for important subgroups are not proportional, careful investigators are forced to rely on alternative methods of summarizing the relevant data with respect to the focal interest of the study. Wei and Schaubel (2008) proposed use of the ratio of cumulative hazards, and described an estimator of this ratio which has asymptotic properties that derive from the usual normal approximation. In the discussion of their method of estimation, they outlined several excellent reasons for preferring a summary measure that is cumulative, rather than instantaneous, when the key hazard functions involved do not appear to satisfy the usual proportional hazards assumption. In large enough samples, their estimator should be adequate; however, in settings that involve fewer subjects, more reliable statistical tools would be desirable.

Using the tools of empirical likelihood, we have described methods for deriving point and interval estimators of the cumulative hazard ratio that appear to be better suited to those study settings involving non-proportional hazards and fewer subjects. If adjustment for confounding variables, by means of appropriate proportional hazard modelling, is required, our estimators, like those of Wei and Schaubel (2008), can accommodate the added computational complexity. This accommodation applies both in the case of adjustment via stratification, or via distinctly separate models in the two primary subgroups represented in the numerator and denominator of the hazard ratio. In each of these cases, theoretical calculations reveal that the asymptotic calibrating distribution for the derivation of pointwise interval estimates of the cumulative hazard ratio should be chi-squared with one degree of freedom. Simulation studies that rely on the original design of Wei and Schaubel (2008) provide persuasive evidence that when the number of subjects in each stratum or subgroup is no more than 100, our proposed estimators have estimated coverage probabilities that are noticeably closer to the nominal value of 0.95 than the corresponding estimated coverage probabilities reported by Wei and Schaubel (2008) in the same study setting. Moreover, our interval estimators are invariant under one-to-one transformation, range preserving, and their shape is wholly determined by the data, since they inherit these properties directly from the empirical log-likelihood function. In addition, no variance estimate is required, and the computations involved are straightforward to carry out. Yet another advantage of empirical likelihood is that by introducing a Bartlett correction, we can reduce the error rate for interval estimates from $O(n^{-1})$ to $O(n^{-2})$. Since this Bartlett correction can be approximated by using the bootstrap method for censored data (see Efron, 1981), we can employ bootstrap quantiles to calibrate the critical value for the asymptotic distribution of the empirical likelihood ratio function. With the bootstrap method we can effectively improve the coverage accuracy for the EL-based estimator with group-specific covariate adjustment.

In two separate examples discussed in §3.6, corresponding to observational studies of mortality in non-Hodgkin's lymphoma and ovarian cancer patients, we illustrated the use of our proposed methods. The resulting point and pointwise interval estimates of the cumulative hazard ratios that we report would contradict reliance on PH modelling with respect to disease stage in the former instance, and for the subgroups of ovarian cancer patients determined by the size of any residual tumor following primary treatment for their disease. In each instance, adjustment for the possible confounding effect of concomitant measurements collected at the beginning of the study period was warranted, and was incorporated into the estimators that we reported.

Both the estimator proposed by Wei and Schaubel (2008), and our empirical likelihood-based alternative, cannot be calculated prior to the larger of the smallest complete observation recorded in the two subgroups of study subjects involved in the cumulative hazard ratio. This restriction avoids any possibility that the denominator of the estimated ratio is 0, resulting in an estimated ratio that is undefined. Without making further assumptions that may be unwarranted, this particular restriction is clearly unavoidable.

Chapter 4

Simultaneous Confidence Bands for Survival and Ratio of Cumulative Hazard Functions

4.1 Introduction

In certain statistical inference problems involving ratio estimation, point estimates and point-wise confidence intervals may not be sufficient. For example, researchers may want to construct a confidence region for the ratio of interest, simultaneously, for all points in a domain $D \in \Re^p$. Such a goal is analogous, in the hypothesis testing context, to testing the null hypothesis that a ratio function, $R(t)$, is equal to $R_0(t)$, for all $t \in [a, b]$, a specified interval, at the overall significance level α. To achieve this goal, we need to construct a confidence band for the function of interest. However, instead of constructing such a confidence region by relying on asymptotic properties, which typically necessitate very large sample sizes to achieve reliable coverage accuracy, bootstrap calibration is commonly used as a basis for statistical inference.

The bootstrap method of inference was first introduced by Efron (1979) for complete data, and then by Efron (1981) and Reid (1981) for censored data. Using simulation studies, Efron (1982) showed that confidence intervals produced by the bootstrap method can be more accurate than those based on the asymptotic distributions of parameter estimators. For complete data, Li, Tiwari and Wells (1999) used a bootstrap percentile to construct a simultaneous confidence band for a vertical quantile comparison function. Claeskens and van Keilegom (2003) built bootstrap confidence bands for regression curves and their derivatives. By combining the bootstrap method with an empirical likelihood estimator, Hall and Owen (1993) developed empirical likelihood confidence bands for kernel estimates. Likewise, Claeskens *et al.* (2003) investigated a bootstrap confidence band for comparison distributions and ROC curves.

For survival data, it is commonly the case that an asymptotic global $(1-\alpha)$ confidence band is not well-behaved for small samples. In terms of the Nelson-Aalen estimator, although one can obtain confidence bands such as the equal precision band (EP), or the Hall-Wellner (HW) band, via the weak convergence of the Nelson-Aalen estimator, these bands perform badly even with sample sizes of 100-200; see Andersen *et al.* (1993). It is even harder to construct confidence bands for cumulative hazard ratios. McKeague and Zhao (2002) constructed a simultaneous confidence band for the ratio of two survival functions based on independent, right-censored data. In subsequent work (McKeague and Zhao, 2005) they described a method of estimating either the difference or ratio of two distribution functions that relies on empirical likelihood. However, since they used the exact nonparametric likelihood, it may be difficult to account for covariate adjustments in any estimated functionals of interest using their approach. Wei and Schaubel (2008) built a confidence band for a cumulative hazard ratio with covariate adjustments based on simulations of a limiting Gaussian process, but their estimator is not easy to implement because of the complexity of the associated variance formula.

An alternative approach involves using bootstrap methods. Following the bootstrap scheme proposed by Efron (1981), Akritas(1986) investigated the bootstrapping of the Kaplan-Meier estimator and found that the bootstrap confidence band has more accurate coverage than the Hall-Wellner (HW) band, especially in small sample scenarios. In this chapter, we first investigate using the bootstrap to estimate simultaneous confidence bands for the survival function when the exact likelihood is used. Then we adapt the bootstrap method to incorporate the Poisson extension of the likelihood function to derive a simultaneous confidence band for the ratio of cumulative hazard functions with covariate adjustment. Via a simulation study, we compare our EL-based bootstrap with several competitors in terms of coverage probabilities at the nominal level of 95%. We illustrate the method in the problem of estimating a cumulative treatment effect using the two observational studies concerning the survival of non-Hodgkin's and ovarian cancer patients that we described in the previous chapter.

4.2 Simultaneous Confidence Bands in the Statistical Literature

4.2.1 HW and EP Confidence Bands for a Survival Function

The HW and EP bands for a survival function are simultaneous confidence bands that are based on the asymptotic distribution of the Kaplan-Meier estimator, uniformly, over the time span of interest. We consider a continuous time interval $\mathcal{I} = [0, \tau]$ or $[0, \tau)$ for given stopping time τ, $0 < \tau < \infty$. Let (Ω, F) be a measurable space equipped with a filtration $(\mathcal{F}_t, t \in \mathcal{I})$. Then we can define a counting process $N = \{N(t), t \in \mathcal{I}\}$ on $(\Omega, \mathcal{F})$.

Let $T_1, ..., T_n$ be i.i.d. survival times with the survival function S, and let $C_1, \ldots, C_n$ be the i.i.d corresponding censoring times with survival function S_C, independent of the $T_i' s$. The observed data consist of $(X_1, \delta_1), \ldots, (X_n, \delta_n)$, where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$. Let $Y(t) = \Sigma_{i=1}^n I(X_i \geq t)$ be the number of individuals at risk just before time t; then $N(t) = \Sigma_{i=1}^n I(X_i \leq t, \delta_i = 1)$.

Definition 1. *Let Θ denote a connected, nonempty, random subset of the rectangle $[0, \tau] \times [0, 1)$, such that $\Theta \cap \{(t, p) : 0 \leq p \leq 1\}$ is nonempty for each $t \in [0, \tau]$. We call Θ a confidence band for S over the set $\mathcal{A} \in [0, \tau]$ with coverage probability $(1 - \alpha)$ if $P\{(t, S(t)) \in \Theta$ for all $t \in \mathcal{A}\} = 1 - \alpha$.*

For arbitrary right-censored data, we denote the true underlying survival function by S_0, and the corresponding Kaplan-Meier estimator, based on a sample of size n, by S_n . The HW and EP bands for $\mathcal{A}$, a finite interval, are constructed using the weak convergence of $n^{1/2}\{\hat{S}_n(t) - S_0(t)\}$, to a Gaussian process, for $t \in \mathcal{A}$. By transforming the Gaussian process to a Brownian bridge limit, we can obtain the confidence band

$$\hat{S}_n(t) \pm n^{-1/2}\hat{S}_n(t)K_{q,\alpha}(d_1, d_2)(1 + \hat{\sigma}^2(t))/q\{\frac{\hat{\sigma}^2(t)}{1 + \hat{\sigma}^2(t)}\}$$

for $\mathcal{A}$ on $[t_1,\ t_2]$, where $K_{q,\alpha}(d_1, d_2)$ is the $(1 - \alpha)$ quantile of the distribution of

$$\sup_{t\in(d_1,d_2)} |q(t)B^0(t)|.$$

$B^0(t)$ represents the standardized Brownian bridge, and the constants d_1, d_2 are approximated by

$$d_i = \frac{\hat{\sigma}^2(t_i)}{1 + \hat{\sigma}^2(t_i)}.$$

Here $\hat{\sigma}^2(t) = n \int_0^t \frac{I(Y(x)>0)}{Y(x)(Y(x)-dN(x))} dN(x)$ estimates the variance of the cumulative hazard function at time t.

For the HW band $q(t) = 1$, whereas for the EP band $q(t) = \{t(1-t)\}^{-1/2}$. Since the quantile $K_{q,\alpha}(d_1, d_2)$ can be obtained by simulating a standardized Brownian bridge, computing either band is not hard. However, each estimator has the drawback that it can give rise to values outside $[0, 1]$. Moreover, Bie *et al.* (1987) show that the coverage probabilities associated with either band are not satisfactory, even when the sample size is 100–200.

A natural remedy for these problems involves using a suitable transformation of the parameter of interest to improve the approximation of the limiting distribution.

The usual transformations to consider include the log-log and arcsine. However, as Bie (1987) observes, the HW band is too wide in the tails of the distribution, and the EP band is too wide in the middle of the distribution, even after applying transformations; therefore the two confidence bands are not very useful in practice.

4.2.2 EL-based Confidence Bands for a Survival Function

Hollander *et.al.* (1997) adapted empirical likelihood to obtain both the HW and EP-type confidence bands for the survival function as well as the cumulative hazard function. Compared to the usual HW and EP bands, these EL alternatives have higher coverage accuracy in small samples, while maintaining the range-preserving and transformation-respecting advantages enjoyed by the empirical likelihood method of inference.

Let $L(S)$ be the likelihood function for the survival function S; then

$$L(S) = \prod_u [S(X_i^-) - S(X_i)] \prod_c S(X_i), \tag{4.1}$$

where u, c represent the sets of uncensored and censored observations, respectively, in the sample.

Following Thomas and Grunkemeier(1975), the EL-based LR statistic is

$$R(p,t) = \frac{\sup\{L(S) : S(t) = p, S \in \Gamma\}}{L(S_n)},$$

where Γ denotes the family of all discrete survivor functions supported by the distinct, complete observations in the sample, $0 < p < 1$. For any fixed value of t, the point-wise asymptotic $100(1-\alpha)\%$ confidence interval for $S_0(t)$ first obtained by Thomas and Grunkemeier is

$$\{p : -2\log R(p,t) \le q_{1-\alpha}\},$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the chi-squared distribution with 1 degree of freedom.

To construct the EL-based HW and EP bands for S_0, we need the asymptotic properties of the LR statistic to hold uniformly. Let $\mathcal{L}(S,t) = -2\log R(S(t),t)$, and define $W(S,t)$ to be the signed root-log-LR statistic, that is,

$$W(S,t) = \operatorname{sgn}\{S_n(t) - S(t)\}\sqrt{\mathcal{L}(S,t)}.$$

Hollander *et al.* (1997) show that the process $\{\hat{\sigma}(t)W(S_0,t), t \in [0,\tau)\}$ converges in distribution to a Gaussian martingale with mean zero and variance $\sigma^2(t)$, where $\hat{\sigma}^2(t)$, as we have previously defined it, is a consistent estimator of $\sigma^2(t)$.

By transforming this limiting process to a Brownian bridge B^0, Hollander *et al.* (1997) show that

$$\sup_{t\in[0,\tau]} \left| \frac{\hat{\sigma}(t)W(S_0,t)}{1+\hat{\sigma}^2(t)} \right| \xrightarrow{D} \sup_{x\in[0,d]} |B^0(x)|,$$

where $B^0(x)$ denotes a standardized Brownian bridge, and d is a constant that can be approximated by $\hat{d} = \hat{\sigma}^2(\tau)/(1+\hat{\sigma}^2(\tau))$. Then the EL-based $100(1-\alpha)\%$ HW band for S_0 is

$$\begin{aligned}\mathcal{B}_{\text{HW}} &= \{S(t) : \left|\tfrac{\hat{\sigma}(t)W(S,t)}{1+\hat{\sigma}^2(t)}\right| \leq K_{q,\alpha}(d), t \in [0,\tau]\} \\ &= \{S(t) : |W(S(t),t)| \leq C(t), t \in [0,\tau]\} \\ &= \{S(t) : \mathcal{L}(S(t),t) \leq C^2(t), t \in [0,\tau]\}\end{aligned}$$

where $C(0) = 0$, and

$$C(t) = K_{q,\alpha}(\hat{d})\frac{1+\hat{\sigma}^2(t)}{\hat{\sigma}(t)}, \quad \text{for} \quad t > 0.$$

If $C^2(t)$ in this HW band is replaced by $e_\alpha^2(a,b)$, Hollander *et al.* (1997) showed that the EL-based $100(1-\alpha)\%$ EP band for S_0 is

$$\mathcal{B}_{\text{EP}} = \{S(t) : \mathcal{L}(S(t),t) \leq e_\alpha^2(a,b)\}$$

for all the values of t in the set $\{t : a \leq \frac{\hat{\sigma}^2(t)}{1+\hat{\sigma}^2(t)} \leq b\}$. Here $e_\alpha^2(a,b)$ is the upper α quantile of the distribution of $\sup_{x\in[a,b]} \frac{|B^0(x)|}{[x(1-x)]^{1/2}}$.

Although these EL-based confidence bands will be range-preserving and transformation-respecting, they may be as unattractive as the usual HW and EP bands due to excessive width. As a result, in the succeeding section we investigate using the bootstrap method to achieve our goal of deriving a confidence band for the true survivor function S_0.

4.3 A Bootstrap Confidence Band for the Survival Function

Since the HW and EP confidence bands may be too wide to be of practical use, an alternative way to obtain a confidence band for S_0 that has more attractive features is to rely on the bootstrap. This method of inference was first introduced by Efron (1979) for complete data, and then by Efron (1981) and Reid (1981) for censored data. Using simulation studies, Efron (1982) showed that the confidence intervals produced by the bootstrap method are more accurate than those based on asymptotic distributions of the estimators. Akritas (1986) investigated the bootstrapping of the Kaplan-Meier estimator and found that only Efron's resampling plan can lead to asymptotically correct confidence bands for the survival function, and the bootstrap confidence band has more accurate coverage than the HW band, especially in small sample scenarios.

Although direct bootstrapping of the Kaplan-Meier estimator can improve the coverage accuracy of confidence bands for the survivor function, this method cannot circumvent the range problem, i.e., producing values of the confidence band that lie outside the interval [0, 1]. Hollander *et al.* (1997) proposed EL-based confidence bands for a survivor function that are range-preserving, transformation-invariant, have greater coverage accuracy than the corresponding asymptotic confidence band, and a shape that is determined by the observed data. However, the disadvantage of this method is the width of the resulting confidence band, in the tail region when the EL-based band is of HW-type, and in the middle of the distribution when the EL-based band is an EP-type.

Therefore, to enhance the performance of these EL-based confidence regions for a survival function, we now consider using empirical likelihood together with the bootstrap method to calibrate the critical value of the sampling distribution better. Our goal in doing so is to reduce the width of the EL-based confidence band without sacrificing coverage accuracy or the other attractive features of this method of inference.

We assume random censorship, and only focus on the bootstrap method described in Efron (1981). Let $(X_i, \delta_i), \quad i = 1, ...n$, be the i.i.d. censored data, where $X_i = \min\{T_i, C_i\}$ is the observed survival time, and δ_i is the censoring indicator, which is 1 when $X_i = T_i$ and 0 when $X_i = C_i$; the variables T_i and C_i are the true survival time and censoring time respectively. For $t \in [0, \ \tau]$, where τ is a fixed stopping time, let $S_n(t)$ and $C_n(t)$ be the corresponding Kaplan-Meier estimators of $S_0(t)$ and $S_C(t)$. Following Efron (1981), we obtain an i.i.d. sample $T_i^*, \ldots, T_m^*$ with replacement from the observed complete response times, T_j, and a corresponding i.i.d. sample $C_1^*, \ldots, C_m^*$ with replacement from the observed censoring times C_j. Then $(X_i^*, \delta_i^*), \quad i = 1, ..., m$ is the bootstrap sample, where

$$X_i^* = \min\{T_i^*, C_i^*\}, \qquad \delta_i^* = I(X_i^* = T_i^*).$$

As shown by Efron(1981), this resampling plan is equivalent to taking a sample $(X_i^*, \delta_i^*), i = 1, ..., m$ with replacement from $(X_i, \delta_i), i = 1, ..., n$.

In order to adapt the empirical likelihood method to the bootstrap sample, we need a process of likelihood ratio statistics. According to Hollander *et al.* (1997), the likelihood ratio can be written as

$$\mathcal{L}(S,t) = \frac{nK^2}{\hat{\sigma}^2} + O_p(n^{-1/2}), \tag{4.2}$$

where $K(S,t) = \log S_n(t) - \log S(t)$. Let $\varphi(S)(t) = K(S,t)$. Applying the delta method, we can approximate $\varphi(S_0)$ by $d\varphi(S_0)(S_n - S_0)$, where $d\varphi$ is the Hadamard derivative of φ, with the expression

$$d\varphi(S) = \frac{-1}{S(t)}.$$

Following the well-known weak convergence result

$$\sqrt{n}(S_n - S_0) \xrightarrow{\mathcal{D}} -S_0 U$$

(see, Anderson *et al.* (1993), theorem. IV.3.2) and the delta method, we have $\sqrt{n}K(S_0,t) \xrightarrow{\mathcal{D}} U(t)$, where $U(t)$ is a Gaussian martingale with zero mean and variance function $\sigma^2(t)$. Further,

$$\mathcal{L}(S_0,t) = \left\{\frac{U(t)}{\hat{\sigma}(t)}\right\}^2 = \left\{\frac{U(t)}{1+\hat{\sigma}^2(t)}\right\}^2 \cdot A(t)^2$$

where $A(t) = \frac{1+\hat{\sigma}^2(t)}{\hat{\sigma}(t)}$. Since the process $\frac{U(t)}{1+\hat{\sigma}^2(t)}$ has the same distribution as the Brownian bridge $B^0\{\hat{\sigma}^2(t)/[1+\hat{\sigma}^2(t)]\}$, write $D(t) = \hat{\sigma}^2(t)/[1+\hat{\sigma}^2(t)]$. Therefore $\text{sgn}\{S_n(t) - S(t)\}\sqrt{\mathcal{L}(S_0,t)}/A(t) \xrightarrow{\mathcal{D}} B^0 \circ D(t)$.

From Gill *et al.* (1989) we require the following definition.

Definition 2. *Let B_1, B_2 be normed vector spaces and $\phi : B_1 \to B_2$ is a measurable* *function. Then ϕ is said to be weak continuous compact differentiable at $x = F$ if there exists a linear and continuous function $d\phi : B_1 \to B_2$ such that*

$$\frac{\phi(x_n + t_n h_n) - \phi(x_n)}{t_n} = d\phi(x) \cdot h,$$

as $x_n \xrightarrow{||\cdot||} x$, $h_n \xrightarrow{||\cdot||} h$, $t_n \to 0$ in $\Re$, Here $||\cdot||$ is the supremum norm, i.e, $||x|| = \sup_t |x(t)|$.

Now consider the likelihood ratio statistic $\mathcal{L}(S^*, t)$, from the bootstrap sample. Under the assumption of weak continuous compact differentiability, as defined above, Gill *et al.*(1989) shows that the bootstrap works if the delta method works, i.e., $n^{1/2}(\phi(S^*) - \phi(S_n))$ has the same limiting distribution as $n^{1/2}(\phi(S_n) - \phi(S_0))$, if the limiting distribution of the latter exists.

Theorem 1. *The bootstrap likelihood ratio statistic sgn$\{S_n^*(t) - S_n(t)\}\sqrt{\mathcal{L}^*(S_n, t)}/A(t) \xrightarrow{\mathcal{D}} B^0 \circ D(t)$.*

Proof. Let $\phi(S)(t) = K^*(S, t) = \log S_n^*(t) - \log S(t)$, where $S_n^*(t)$ is the Kaplan-Meier estimator obtained from the bootstrap sample. By Theorem 4 of Gill (1989), $n^{1/2}K^*(S_n, t) \xrightarrow{\mathcal{D}} d\phi(S_0)(t) \cdot \{-S_0(t)U(t)\} = U(t)$, where $U(t)$ is a Gaussian martingale with zero mean and variance function $\sigma^2(t)$. From equation (4.2) we have

$$\mathcal{L}^*(S_n, t) = \frac{nK^{*2}}{\hat{\sigma}^2} + O_p(n^{-1/2}). \tag{4.3}$$

If we replace the limiting distribution of $n^{1/2}K^*(S_n, t)$ by $U(t)$ and apply Slutsky's theorem we have $\text{sgn}\{S_n^*(t) - S_n(t)\}\sqrt{\mathcal{L}^*(S_n, t)}/A(t) \xrightarrow{\mathcal{D}} B^0 \circ D(t)$ as required. □

For each bootstrap sample we can obtain $\sup_{t\in[0,\tau]} \mathcal{L}^*(S_n,t)/A(t)^2$. By repeating the bootstrap process B times we can acquire the $100(1-\alpha)\%$ bootstrap quantile of these values. By taking the bootstrap quantile as the critical value, denoted by $(K^*_{q,\alpha})^2$ we can obtain an HW-type bootstrap confidence band for $S(t)$ when $t \in (0,\tau)$ i.e.,

$$\mathcal{B}^*_{\text{HW}} = \{S(t) : \mathcal{L}(S(t),t) \leq C^*(t)^2, t \in [0,\tau]\}$$

where $C^*(0) = 0$,

$$C^*(t) = K^*_{q,\alpha}(\hat{d})\frac{\hat{\sigma}(t)}{1+\hat{\sigma}^2(t)}, \quad \text{for} \quad t > 0.$$

Similarly for all values of t that satisfy $a \leq \frac{\hat{\sigma}^2(t)}{1+\hat{\sigma}^2(t)} \leq b$, if we use the $100\%(1-\alpha)$ bootstrap quantile of $\mathcal{L}^*(S_n,t)$ as the critical value, which is denoted by $e^*(a,b)^2$, we can obtain the EP type of bootstrap confidence band for the survival function in $(0,\tau]$.

$$\mathcal{B}^*_{\text{EP}} = \{S(t) : \mathcal{L}(S(t),t) \leq e^*(a,b)^2, \ t \in [0,\tau]\}.$$

Since the above bootstrap methods are based on suitable transformations to Brownian bridge, although they can reduce the coverage errors by using the bootstrap quantile of $K^*_{q,\alpha}(\hat{d})$ they still cannot solve the problem of excessive width in certain regions of the HW-type or EP-type bands. An alternative method involves simply using $(1-\alpha)$ bootstrap sample quantile of $\sup_{t\in[a,b]} \mathcal{L}^*(S_n,t)$ as the critical value to obtain the simultaneous confidence band for the survival function $S_0(t), t \in [a,b]$. The proof of the result follows from equation (4.3), since the RHS of that expression has a limiting distribution $U^2(t)/\sigma^2(t)$ uniformly for $t \in [a,b]$. Therefore $\sup_{t\in[a,b]} \mathcal{L}^*(S_n,t) \xrightarrow{D} \sup_{t\in[a,b]} U^2(t)/\sigma^2(t)$.

4.4 A Bootstrap Confidence Band for the Ratio of Cumulative Hazard Functions

Another key function in survival analysis is the cumulative hazard function

$$\Lambda(t) = -\int_0^t \frac{dS(s)}{S(s-)}, \quad t \in [0,\tau].$$

Since

$$S(t) = \prod_{s\leq t}(1-\Delta\Lambda(s)), \tag{4.4}$$

and $\Delta\Lambda(t) = -\Delta S(t)$, we can express the likelihood function of $S(t)$ — see equation (4.1) — in terms of the cumulative hazard function, and obtain

$$L(\Lambda) = \prod_{i=1}^{n} \Delta\Lambda(T_i)^{\delta_i}\left[\prod_{j:T_j<T_i}(1-\Delta\Lambda(T_j))\right]^{\delta_i} \cdot \left[\prod_{j:T_j\leq T_i}(1-\Delta\Lambda(T_j))\right]^{1-\delta_i}. \tag{4.5}$$

Since the Nelson-Aalen estimator, which maximizes $L(\Lambda)$, is a nonparametric maximum likelihood estimator of the true cumulative hazard function Λ_0, the likelihood ratio with respect to the cumulative hazard function A is

$$R(A,t) = \frac{\sup\{L(\Lambda) : \Lambda(t) = A, \Lambda \in \Upsilon\}}{L(\Lambda_n)},$$

where Λ_n is the Nelson-Aalen estimator based on a sample size of n. Then, as Hollander *et al.* (1997) show, an asymptotic $100(1-\alpha)\%$ confidence band for Λ_0 is given by

$$\mathcal{D} = \{A : -2\log R(A,t) \leq C^2(t), \quad t \in [0,\tau]\},$$

where $C^2(t)$ is defined as before. Define $\mathcal{L}(A,t) = -2\log R(A,t)$; Hollander *et al.* (1997) derive the EL-based HW band

$$\mathcal{D}_{HW} = \{A(t) : \mathcal{L}(A(t),t) \leq C(t)^2, t \in [0,\tau]\},$$

and EP band

$$\mathcal{D}_{EP} = \{A(t) : \mathcal{L}(A(t),t) \leq e(a,b)^2, \ t \in [0,\tau]\}.$$

Equation (4.4) indicates that $\Lambda(t)$ is a function of $S(t)$, i.e., $\Lambda(t) = \phi(S)$. Since the EL-based estimator is transformation-preserving, we can obtain an EL-based confidence band for Λ_0 simply by transforming an EL-based confidence band for S_0 or vice versa, i.e.,

$$\mathcal{D} = \{A(t) = \phi(S) : S \in \mathcal{B}\}.$$

Although the exact likelihood (4.5) is widely used to estimate functionals of the survival and cumulative hazard, this approach maybe more suitable for randomized clinical trials without covariate adjustments. Mckeague and Zhao (2005) used the exact form to construct simultaneous confidence bands for the difference and the ratio of two survival functions under an independent right-censoring mechanism. However, since they used the exact nonparametric likelihood, it may be difficult to account for covariate adjustments in any estimated functionals of interest using their approach. In observational studies with covariate adjustments, where Cox regression models are very commonly used, using the Poisson extension of the likelihood function for $\Lambda(t)$ (Murphy 1995) that can accommodate the Cox regression model directly, maybe a prudent choice to deal with the situation.

Murphy (1995) discussed the merits of the Poisson extension of the likelihood function for Λ, that is

$$L(\Lambda) = \prod_{i=1}^{n} \Delta\Lambda(T_i)^{\delta_i} \exp(-\Lambda(T_i)). \tag{4.6}$$

Using this Poisson extension of the likelihood function for Λ leads to the same nonparametric maximum likelihood estimators for the survivor function and the cumulative hazard function that maximize the exact version of $L(\Lambda)$ specified in equation (4.5), i.e, the Kaplan-Meier estimator of S_0, and the Nelson-Aalen estimator of Λ_0. Therefore, the Poisson extension specified in equation (4.6) can be used to obtain a corresponding likelihood ratio statistic and to derive the EL-based estimator.

A notable advantage of this Poisson extension is the fact that the only constraint which $\Delta\Lambda(t)$ must satisfy is that it has to be positive. This absence of any other restrictions can circumvent certain complications that can arise in some samples when the version of $L(\Lambda)$ specified in equation (4.5) is used to define a likelihood ratio statistic, and derive confidence intervals or bands for $\Lambda(t)$. In addition, as Pan and Zhou (2002) show, the difference between the Poisson extension and the exact version goes to zero in probability as $n \to \infty$, i.e, the Poisson extension of $L(\Lambda)$ is asymptotically equivalent to equation (4.5), and is the version that we will use to derive EL-based confidence bands in what follows.

Since the confidence bands that Mckeague and Zhao derived in 2005 are based on the limiting distribution of the empirical log-likelihood ratio, adopting their approach would necessitate large sample sizes to achieve any desired coverage accuracy. When the sample sizes are small, EL-type confidence bands for survival functions need bias adjustments to reduce coverage errors (see Hollander *et al.* 1997). However, Mckeague and Zhao (2005) did not consider such adjustments. To achieve the appropriate adjustments in estimating a confidence band for the ratio of the baseline cumulative hazard functions in a small sample scenario, we employ a bootstrap method to improve the coverage accuracy of the confidence bands.

Recall from chapter 3, the profile empirical log-likelihood ratio is $-2l_E(\theta(t))$; let $\mathcal{L}(\theta(t),t) = -2l_E(\theta(t))$. For each observed failure time point $t \in [a,b]$, generate a bootstrap sample and calculate the corresponding value of the profile empirical log-likelihood ratio $\mathcal{L}^*(\hat{\theta}(t),t)$, where $\hat{\theta}(t)$ is the point estimate at time t . Then take the supremum of $\mathcal{L}^*(\hat{\theta}(t),t)$ for $t \in [a,b]$, i.e., $\sup_{t\in[a,b]} \mathcal{L}^*(\hat{\theta}(t),t)$. Repeating this series of bootstrap sample calculations B times generates a bootstrap sampling distribution for the supremum of the logarithmic profile empirical likelihood ratio for all $t \in [a,b]$. By extracting the $100(1-\alpha)\%$ quantile from this bootstrap distribution and using it as the critical value, denoted by $q_{1-\alpha}$, we can generate a corresponding $100(1-\alpha)\%$ confidence band for $\theta(t), t \in [a,b]$, i.e.,

$$\mathcal{B} = \{\theta(t) : \mathcal{L}(\theta(t),t) \leq q_{1-\alpha}\}.$$

To justify our bootstrap approach, we first notice that from the proofs of theorems 1 and 2 in Chapter 3, the logarithmic empirical likelihood ratio is a function of $\xi_i, i = 1, 2$. Each Lagrange multiplier converges in distribution to a Gaussian martingale, and therefore it has a limiting distribution uniformly for $t \in [a, b]$; see Anderson *et al.* (1993). Then, based on the weak consistency of the bootstrap, as established in theorem 5 from Gill (1989), we can use the $100(1-\alpha)\%$ bootstrap quantile of the supremum of a finite set of logarithmic empirical likelihood ratios as the critical value to calibrate the confidence interval or band of $\theta(t)$ at overall level of confidence $1-\alpha$, for all $t \in [a, b]$.

Following Chen and Cui (2007), for a given t, this bootstrap method approximates Bartlett correction under certain regularity conditions. Consequently, we should expect to reduce the coverage errors to $O(n^{-2})$. As shown in our simulation of pointwise intervals for a given t in the Chapter 3, this bootstrap method noticeably reduces the coverage errors and correspondingly improves coverage accuracy of a confidence interval for the cumulative hazard ratio with covariate adjustments. Consequently, we should expect that using these quantiles from the bootstrap distribution of the supremum of logarithmic empirical likelihood ratios would have a similar effect on simultaneous confidence bands of the cumulative hazard ratio for all $t \in [a, b]$.

4.5 Simulation Study

To compare our bootstrap confidence band with the EL-type HW and EP bands proposed by Hollander *et al.* (1997), we generate the survival times and censoring times from $S_0 = \exp(1)$, and $S_C = \text{uniform}(\alpha)$ respectively. By choosing different values of α, we can adjust the proportion of censoring in our data. Here, we set $\alpha = \infty$, 3.72 and 1.595 to represent no censoring, 25% and 50% censoring. The stopping time τ is selected to guarantee that at least 10% of the total sample size of survival times is greater than or equal to τ. This is done to avoid instability in the estimated confidence bands for large τ. For the EL-type HW and EP bands, we need the asymptotic critical values at the nominal level of 95%; according to Hollander *et al.* (1997), these are 1.358 and 3.31 respectively. Based on 5000 samples we calculated the error rate, as the estimated probability that $S_0(t)$ falls outside the confidence band for some $t \in [0, \tau]$; the results are given in Table 4.1.

Table 4.1: Observed error rates and corresponding estimated standard errors (in parenthesis) of EL and bootstrap confidence bands for the survival function, $S(t)$, at a nominal confidence level of 95%.

θ	n	EL-HW	EL-EP	Boot-HW	Boot-EP	EL-boot
∞	25	4.0(.284)	7.3(.368)	4.2(.284)	7.0(.361)	4.8(.302)
	50	4.2(.277)	7.2(.366)	4.1(.280)	7.0(.361)	4.9(.305)
	100	5.3(.317)	6.8(.356)	5.2(.314)	6.5(.349)	5.1(.311)
	200	5.5(.322)	6.3(.344)	5.3(.317)	6.2(.341)	5.0(.308)
3.72	25	3.8(.270)	7.4(.370)	4.1(.280)	7.2(.366)	4.9(.305)
	50	4.3(.287)	7.3(.368)	4.4(.290)	6.9(.358)	4.8(.302)
	100	4.8(.302)	7.1(.363)	4.7(.299)	7.0(.361)	4.9(.305)
	200	4.9(.305)	6.5(.349)	5.0(.308)	6.3(.344)	5.0(.308)
1.595	25	3.1(.245)	7.2(.366)	3.5(.260)	6.8(.356)	5.0(.308)
	50	3.6(.263)	7.1(.363)	3.8(.270)	6.9(.358)	5.1(.311)
	100	3.8(.270)	7.2(.366)	4.0(.277)	6.7(.354)	4.9(.305)
	200	4.3(.287)	6.8(.356)	4.5(.293)	6.4(.346)	5.0(.308)

From the estimated error rates displayed in the table, we notice that both bootstrap-HW and bootstrap-EP have a coverage probability closer to the nominal level of 95% than the corresponding direct EL method of deriving HW or EP bands for the survival function in virtually all scenarios. But the improvement in the associated coverage accuracy is relatively limited. On the other hand, when the EL-bootstrap is used to estimate the confidence band for a survival function, we obtain an estimated coverage probability very close to the nominal level of 95%. This is due to an advantage of the EL-bootstrap, namely that it can approximate Bartlett correction, and correspondingly reduce the error rate and improve the coverage accuracy of the confidence band for the survival function, especially in small sample scenarios. In addition, by using the bootstrap quantile of the supremum of the logarithm of the empirical likelihood ratio statistic for all t in a fixed time interval, the EL-bootstrap avoids the excess width in the tails of the distribution that characterizes the HW-bands, as well as the corresponding excess in the middle part of the EP-bands.

4.6 Applications

We use two datasets concerning the survival experience of non-Hodgkin's lymphoma and ovarian cancer patients that we previously analyzed in Chapter 3 to illustrate how to compute a simultaneous confidence band for the cumulative treatment effect $\theta(t)$, for $t \in [a, b]$.

The 95% simultaneous confidence bands of baseline cumulative hazard ratios with covariate adjustment for these two datasets are displayed in Figures 4.1 and 4.2. From Figure 4.1 we see that although all the lower bounds of the pointwise interval estimates of the cumulative hazard ratio exceed 1, the 95% confidence band for the adjusted cumulative hazard ratio, displayed on a logarithmic scale, includes 1 from roughly 13 to 18 months following diagnosis. Apart from this brief anomaly, however, it seems reasonable to conclude that the adjusted risk of death for non-Hodgkin's lymphoma patients with stage IV disease is greater than the corresponding risk for patients with stage II or III disease at diagnosis.

For ovarian cancer patients, although the 95% simultaneous confidence bands of the baseline cumulative hazard ratio are wider than the corresponding pointwise confidence bounds, the lower band exceeds 1 uniformly, which confirms our previous conclusion based on the pointwise interval estimates. We conclude that the risk of death for patients in the group with a residual tumor after treatment greater than 1 cm is distinctly greater than that experienced by patients with little or no residual tumor, after adjustment for the differential effect of hcg and logarithmic ca125 in these two patient groups.

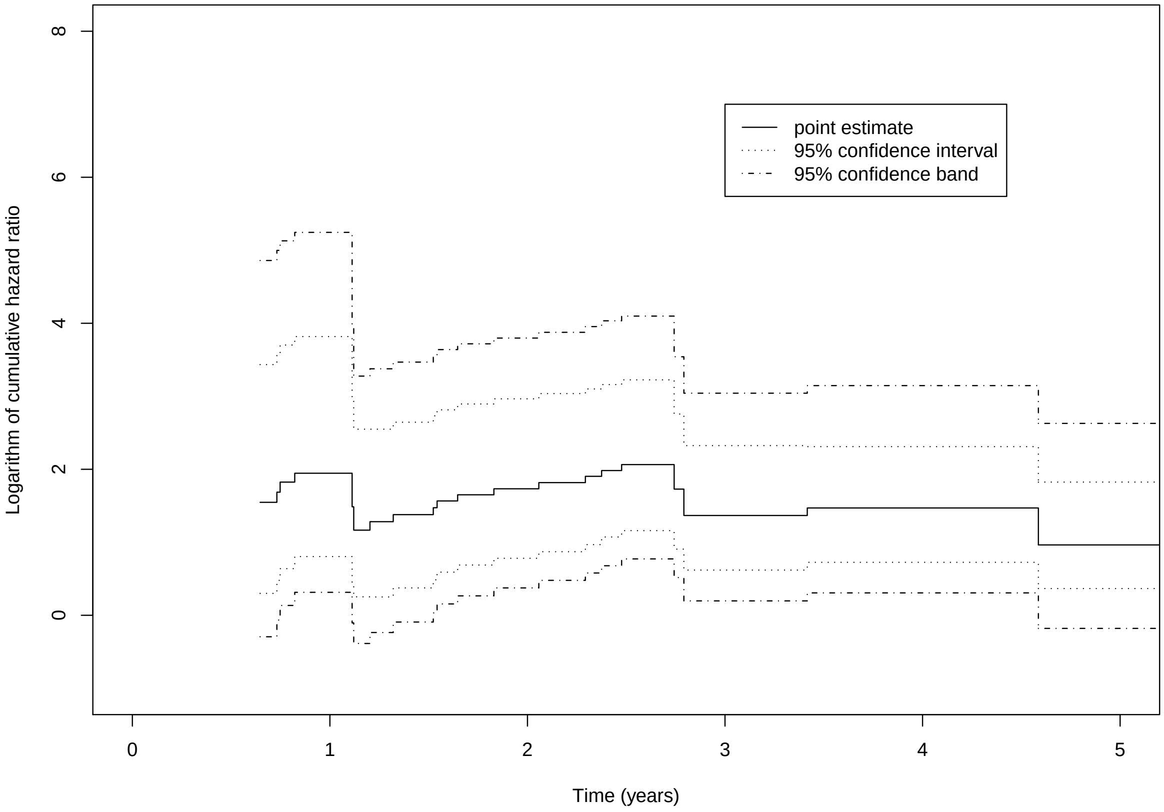

Figure 4.1: Estimated baseline cumulative hazard ratios of death for non-Hodgkin's patients with Stage IV compared to Stage II or III disease, adjusted for the effect of presenting symptoms and the presence of bulky disease

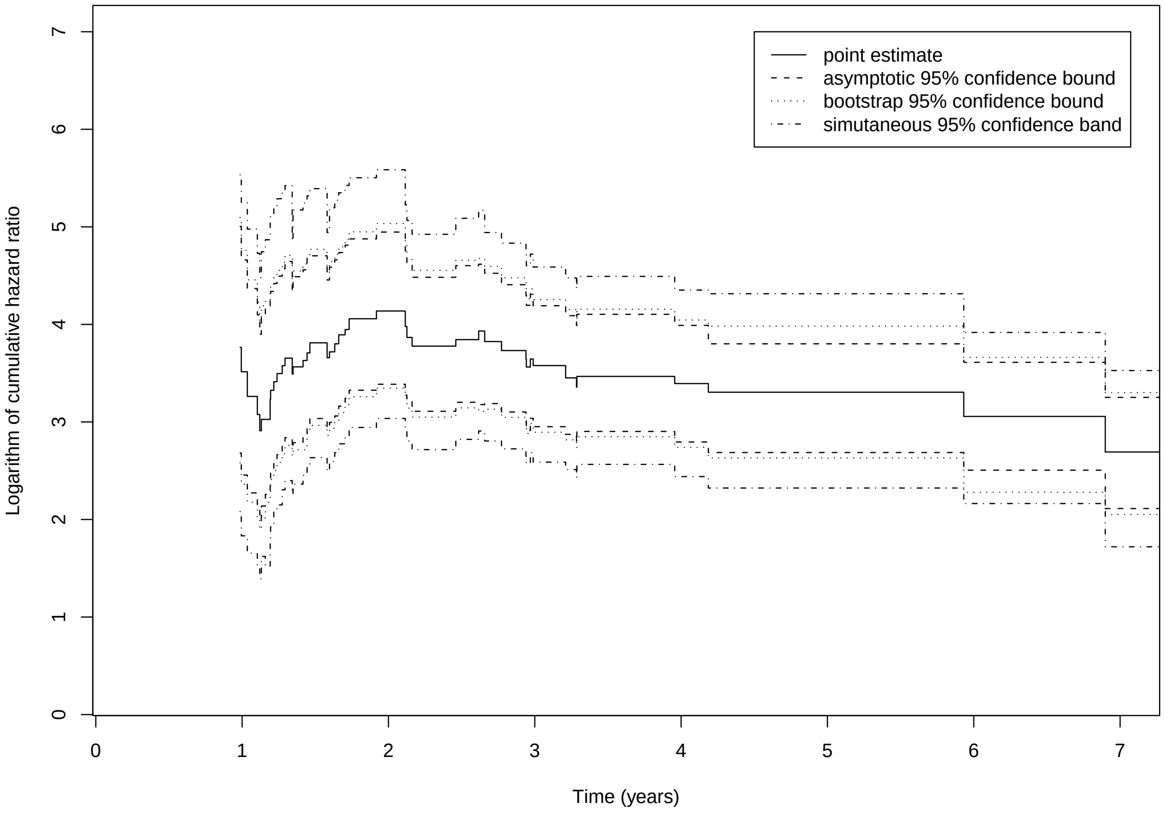

Figure 4.2: Estimated baseline cumulative hazard ratios for ovarian cancer patients with residual tumors exceeding 1 cm compared to those with little or no residual tumor, adjusted for the differential effects of hcg and logarithmic ca125 on patient survival experience.

4.7 Summary

To derive point and interval estimators for time-to-event data, the method of empirical likelihood has obvious appeal. As we mentioned previously, it is range-preserving and transformation-invariant. The resulting interval estimates are data-determined, may be asymmetric, and have better coverage probability for small samples. However, when we use this method and its asymptotic critical value to construct simultaneous confidence bands for survival or cumulative hazard functions in a small sample scenario, the estimated error rate is higher than expected and the width of the confidence bands is frequently excessive at some points in the interval of interest.

To overcome these limitations of the empirical likelihood approach, we propose using a bootstrap method to recalibrate the critical value of the sampling distribution of the sample log-likelihood ratios. The resulting bootstrap-based band has an estimated coverage probability closer to the nominal level of 95%. In addition, this approach seems to rectify some of the excesses associated with either HW or EP bands.

By extending this bootstrap method from the exact likelihood for the survival function to Murphy's Poisson extension for the cumulative hazard function, we were able to consider adjusted hazards that involve the Cox proportional hazards regression model. Using this extension, we showed how to derive simultaneous confidence bands for the ratio of two baseline cumulative hazards with covariate adjustment, and illustrated our method using the two datasets.

Chapter 5

Empirical Likelihood for Copula-based Estimation

5.1 Introduction

Copulas, which are functions that couple marginal one-dimensional distributions to obtain multivariate distribution functions, have been a particular focus of statistical research for modeling dependent data in recent years. By combining marginal distributions of any specified form via a suitable copula, statistical researchers are able to construct multivariate distributions and study the resulting dependence. Since an appealing feature of copula models is that the margins do not depend on the choice of the dependency structure, copula-based estimation can be divided into two stages, i.e., marginal and joint analysis. This two-stage approach can make the estimation simpler and possibly more reliable since the marginal distributions can be estimated using well-established tools for statistical inference.

This interest in copula models has prompted new developments in the analysis of multivariate survival data that consist of several possibly related failure times, e.g., the times to occurrence of a certain event such as the occurrence of a particular disease for siblings. Many authors have employed a copula approach to construct the joint survival function of such data or to measure the association among the related failure variables. Joe (1997) and Hougaard (2000) described how to combine appropriate marginal models with a suitable copula to form a valid and flexible multivariate distribution. Sun, Wang, and Sun (2006) suggested fitting a Clayton copula with nonparametric marginal distributions to estimate the association for bivariate interval-censored failure data. Bogaerts and Lesaffre (2008) used a one-parameter copula, where the association parameter can depend on covariates, to model the marginal distributions with an accelerated failure time model and a flexible error term.

While a copula is a good statistical tool, selecting a parametric copula is a non-trivial task that may lead to model misspecification because different copula families involve different correlation structures. This observation motivates us to use empirical likelihood to estimate a copula nonparametrically and, thereby, to obtain the joint distribution (survival) functions or the correlation parameter of interest. Section 5.2 outlines the details of our method. With this EL-based estimator of a copula, we can also derive a goodness-of-fit test for assessing a specific parametric copula model. The specifics of such a test can be found in Section 5.3. By means of a simulation study, we demonstrate the merits of our EL-based testing procedure.

5.2 Empirical Likelihood Estimator of Copulas for Complete Data

In medical studies, researchers may be interested in estimating the risk, in patients who suffer from kidney disease, that both organs fail prior to some fixed t that measures the time since diagnosis of the disease. Knowing this risk would enable clinicians to determine whether or not to recommend kidney transplantation.

Let (X, Y) be the times of organ failure after diagnosis for a subject's left and right kidneys, respectively. Let H be the joint distribution function of X, Y. For a certain time point, e.g., 5 years, researchers want to estimate $\Pr(X \leq 5, Y \leq 5)$. Since the event of one organ failure may be associated with that of the other organ in the same patient, we consider using the copula model and empirical likelihood to tackle this problem.

Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are independent and identically distributed random vectors with distribution function H. The copula of H is defined by $C(x, y) = H(F_1^{-1}(x), F_2^{-1}(y))$, where $F_1(x), F_2(y)$ denote the marginal distribution functions for $\{X_i\}$ and $\{Y_i\}$, respectively, and H is the joint distribution for $(X_i, Y_i), i = 1, ..., n$.

In order to construct an estimator for $C(x, y)$, which we denote by θ, we introduce link variables s, t such that $F_1(s) = x, F_2(t) = y$ for $0 < x, y < 1$, and $C(x, y) = H(s, t) = \theta$. Let $\hat{F}_X(s)$ and $\hat{F}_Y(t)$ be the estimators of F_1 and F_2, respectively. Define

$$\hat{F}_{X_i}(s) = I(X_i \leq s), \qquad \hat{F}_{Y_i}(t) = I(Y_i \leq t),$$

and

$$w_i(s,t) = \hat{F}_{X_i}(s)\hat{F}_{Y_i}(t) - \theta, \quad w_{1i}(s) = \hat{F}_{X_i}(s) - x, \quad w_{2i}(t) = \hat{F}_{Y_i}(t) - y.$$

By defining $p = (p_1, ..., p_n)$ to be a probability vector with $\sum_{i=1}^{n} p_i = 1$, we have the empirical likelihood for θ,

$$L(\theta) = \sup(\prod_{i=1}^{n} p_i)$$

subject to the following three constraints:

$$\sum_{i=1}^{n} p_i w_i(s,t) = 0, \sum_{i=1}^{n} p_i w_{1i}(s) = 0, \sum_{i=1}^{n} p_i w_{2i}(t) = 0.$$

Using Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3$, it follows that

$$p_i = \frac{1}{n\{1 + \lambda_1 w_i(s,t) + \lambda_2 w_{1i}(s) + \lambda_3 w_{2i}(t)\}}, \ i = 1, ..., n.$$

Therefore, the logarithmic empirical likelihood ratio statistic for θ is

$$l_0(\theta) = -2\sum_{i=1}^{n} \log(np_i) = 2\sum_{i=1}^{n} \log\{1 + \lambda_1 w_i(s,t) + \lambda_2 w_{1i}(s) + \lambda_3 w_{2i}(t)\},$$

where $\lambda_i, i = 1, 2, 3$ should satisfy the following equations:

$$\sum_{i=1}^{n} \frac{w_i(s,t)}{\{1 + \lambda_1 w_i(s,t) + \lambda_2 w_{1i}(s) + \lambda_3 w_{2i}(t)\}} = 0, \tag{5.1}$$

$$\sum_{i=1}^{n} \frac{w_{1i}(s)}{\{1+\lambda_1 w_i(s,t)+\lambda_2 w_{1i}(s)+\lambda_3 w_{2i}(t)\}} = 0, \tag{5.2}$$

$$\sum_{i=1}^{n} \frac{w_{2i}(t)}{\{1+\lambda_1 w_i(s,t)+\lambda_2 w_{1i}(s)+\lambda_3 w_{2i}(t)\}} = 0, \tag{5.3}$$

and constraints $1+\lambda_1 w_i(s,t)+\lambda_2 w_{1i}(s)+\lambda_3 w_{2i}(t) > 1/n$, which come from the probability requirements $p_i < 1$. Using the same method that we first adopted in section 2.2, we can derive the constrained solutions of $\lambda_i(s,t), i = 1, 2, 3$, where s, t are nuisance parameters. Since this copula function of θ does not depend on the marginal distributions F_1 and F_2, we can plug in the estimators of s and t that derive from the marginal distributions, i.e. $\hat{s} = \hat{F}_1^{-1}(x), \hat{t} = \hat{F}_2^{-1}(y)$, to obtain the empirical log-likelihood ratio statistic

$$l(\theta) = 2\sum_{i=1}^{n} \log\{1+\lambda_1 w_i(\hat{s},\hat{t})+\lambda_2 w_{1i}(\hat{s})+\lambda_3 w_{2i}(\hat{t})\}. \tag{5.4}$$

Therefore, the point estimator of θ is $\hat{\theta} = \arg\min_\theta l(\theta)$.

To obtain the interval estimators of θ, we need the asymptotic property of $l(\theta)$ that depends on the selected estimators of marginal distribution functions, $\hat{F}_X(s), \hat{F}_Y(t)$. Next, we discuss the use of empirical and kernel methods to incorporate these estimated marginal distributions into the empirical likelihood framework to estimate the copula of interest.

First, we consider the empirical estimators of $F_1(s), F_2(t)$. Recall that

$$\hat{F}_{X_i}(s) = I(X_i \leq s), \quad \hat{F}_{Y_i}(t) = I(Y_i \leq t).$$

Then

$$w_i(s,t) = I(X_i \leq s)I(Y_i \leq t) - \theta, \quad w_{1i}(s) = I(X_i \leq s) - x, \quad w_{2i}(t) = I(Y_i \leq t) - y.$$

Let

$$A = \sum_{i=1}^{n} I(X_i \leq s)I(Y_i \leq t), \quad B = \sum_{i=1}^{n} I(X_i \leq s)I(Y_i > t),$$

$$C = \sum_{i=1}^{n} I(X_i > s)I(Y_i \leq t), \quad D = \sum_{i=1}^{n} I(X_i > s)I(Y_i > t).$$

Define

$$\rho_1 = \Pr\{X \leq s, Y \leq t\}, \quad \rho_2 = \Pr\{X \leq s, Y > t\},$$

$$\rho_3 = \Pr\{X > s, Y \leq t\}, \quad \rho_4 = \Pr\{X > s, Y > t\}.$$

Therefore, $\{A/n, B/n, C/n, D/n\}$ follows a multinomial distribution with parameters $\{\rho_1, \rho_2, \rho_3, \rho_4\}$. Denote $\hat{P} = (A/n, B/n, C/n, D/n)^T$, $P = (p_1, \rho_2, \rho_3, \rho_4)^T$. According to central limit theory $\hat{P} \xrightarrow{D} N(P, \Sigma_0)$ as $n \to \infty$, where Σ_0 is the standard covariance matrix for the multinomial distribution, i.e., for $i < j$, the (i,j)-th element of Σ_0 is

$$\begin{array}{ll} n\rho_i(1-\rho_i) & \text{for} \quad i = j \\ -n\rho_i\rho_j & \text{for} \quad i \neq j. \end{array}$$

Let Q_1, Q_2 and Q_3 be the LHS of (5.1),(5.2) and (5.3), respectively, where $\tilde{Q} = (Q_1, Q_2, Q_3)^T$, and $\tilde{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$. Then $\tilde{\lambda}$ satisfies $\tilde{Q}(\tilde{\lambda}) = \tilde{0}$. Using a

Taylor expansion

$$\begin{aligned}
\tilde{Q}(\tilde{\lambda}) &\simeq \tilde{Q}(\tilde{0}) + \tfrac{\partial \tilde{Q}}{\partial \tilde{\lambda}}|_{\tilde{\lambda}=\tilde{0}}\tilde{\lambda} \\
&= \begin{pmatrix} \sum_{i=1}^n w_i \\ \sum_{i=1}^n w_{1i} \\ \sum_{i=1}^n w_{2i} \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^n w_i^2 & \sum_{i=1}^n w_i w_{1i} & \sum_{i=1}^n w_i w_{2i} \\ \sum_{i=1}^n w_i w_{1i} & \sum_{i=1}^n w_{1i}^2 & \sum_{i=1}^n w_{1i} w_{2i} \\ \sum_{i=1}^n w_i w_{2i} & \sum_{i=1}^n w_{1i} w_{2i} & \sum_{i=1}^n w_{2i}^2 \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} \\
&= \begin{pmatrix} A - n\theta \\ A + B - nx \\ A + C - ny \end{pmatrix} + \begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix},
\end{aligned}$$

where $V = (v_{ij}) =$

$$\begin{pmatrix} A - 2A\theta + n\theta^2 & A - Ax - A\theta - B\theta + n\theta x & A - Ay - A\theta - C\theta + n\theta y \\ A - Ax - A\theta - B\theta + n\theta x & A - 2Ax + B - 2Bx + nx^2 & A - Ay - Ax - By - Cx + nxy \\ A - Ay - A\theta - C\theta + n\theta y & A - Ay - Ax - By - Cx + nxy & A - 2Ay + C - 2Cy + ny^2 \end{pmatrix}$$

Therefore, we have $\tilde{\lambda} = -V^{-1}\tilde{Q}(\tilde{0})$. Using a Taylor expansion of the empirical log-likelihood,

$$\begin{aligned}
l_0(\theta) &= 2\sum_{i=1}^n \log\{1 + \lambda_1 w_i(s,t) + \lambda_2 w_{1i}(s) + \lambda_3 w_{2i}(t)\} \\
&\simeq \tilde{\lambda}^T V \tilde{\lambda} \\
&= \tilde{Q}(\tilde{0})^T V^{-1} \tilde{Q}(\tilde{0}).
\end{aligned}$$

As $n \to \infty$, $\tilde{Q}(\tilde{0}) = (A - n\theta, A + B - nx, A + C - ny)^T$ has a multinormal distribution with mean zero and covariance matrix

$$\Sigma = \begin{pmatrix} n\theta(1-\theta) & n\theta(1-x) & n\theta(1-y) \\ n\theta(1-x) & nx(1-x) & n\theta - nxy \\ n\theta(1-y) & n\theta - nxy & ny(1-y). \end{pmatrix}$$

And by the SLLN, $V \xrightarrow{a.s.} \Sigma$, as $n \to \infty$. Hence, $l_0(\theta) \xrightarrow{D} \chi^2_d$, where $d = 3$. Since the nuisance parameters s and t are unknown and need to be estimated from the marginal distributions, and clearly $\hat{s} \xrightarrow{a.s.} s$ and $\hat{t} \xrightarrow{a.s.} t$, using Slutsky's theorem we have $l(\theta) \xrightarrow{D} \chi^2$ with $df = 3 - 2 = 1$.

Now, we consider combining kernel estimators of the two marginal distributions with the empirical likelihood for the joint distribution to obtain an estimator of the copula. Let $k(x)$ be a kernel density function with distribution function $K(x) = \int_{-\infty}^{x} k(u)du$. Let $h > 0$ be a bandwidth. Define

$$w_i(s,t) = K\left(\frac{X_i - s}{h}\right)K\left(\frac{Y_i - t}{h}\right) - \theta,$$

$$w_{1i}(s) = K\left(\frac{X_i - s}{h}\right) - x, \quad w_{2i}(s,t) = K\left(\frac{Y_i - t}{h}\right) - y.$$

Let C_0 be the true copula, and $\theta_0 = C_0(x,y)$. According to Chen *et al.* (2009), the empirical log-likelihood ratio $l(\theta) \xrightarrow{D} \chi^2_1$ under the following regularity conditions:

(i) $F_1^{''}(s)$, $F_2^{''}(t)$, $\frac{\partial H(s,t)}{\partial s^2}$, $\frac{\partial H(s,t)}{\partial t^2}$, $\frac{\partial H(s,t)}{\partial s \partial t}$ are continuous at point $(s,t) = (s_0, t_0)$, where s_0 and t_0 satisfy $F_1(s_0) = x$ and $F_2(t_0) = y$;

(ii) $k(x)$ is a symmetric density with support $(-1, 1)$ and $k'(x)$ is bounded;

(iii) $nh^4 \to 0$ and $n^{-3/4}h^{-2}(\log n)^2 \to 0$ as $n \to \infty$.

Since $l(\theta)$ has an asymptotic χ_1^2 distribution in both of the situations where kernel or empirical estimators replace the two marginal distributions, the empirical likelihood-based confidence interval for θ is

$$I_{1-\alpha}(x, y) = \{\theta : l(\theta) \leq q_{1-\alpha}\},$$

where $q_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of χ_1^2.

5.3 An Empirical Likelihood-based Goodness-of-fit Test for Copulas

Many authors have investigated goodness-of-fit (gof) tests for copulas. For complete data, Genest *et al.* (2009) provided a critical review summarizing the procedures used to develop a gof test. They categorized gof tests into 3 classes: tests based on the empirical copula, tests based on Kendall's transform, as well as tests based on Rosenblatt's transform. Via a large-scale Monte Carlo study to assess the finite-sample properties of a selection of the proposed gof tests for various choices of dependence structures and degrees of association, they presented the power estimates for these proposed gof tests, In the spirit of their investigation, we now consider an EL-based gof test.

Since the empirical likelihood estimator of the copula $C(x, y)$, which we denote by θ, is estimated nonparametrically, and $l(\theta)$, for the true parameter θ, has an asymptotic χ_1^2 distribution, we can use $l(\theta)$ as a test statistic for the goodness of fit under null hypothesis $H_0 : C \in C_0$ for a certain class C_0 of copulas.

For given x, y, we can calculate $\theta = C_0(x, y)$ under the null hypothesis $H_0 : C \in C_0$. If $l(\theta) < q_{1-\alpha}$, the $(1-\alpha)$ quantile for χ_1^2, then we do not reject H_0 at a significant level of α; otherwise, we reject the null hypothesis at significance level α. Usually, we cannot directly obtain the true θ, since it may involve one or more parameters β , say, under the null hypothesis. However, if we replace β in the expression for $C_0(x, y)$ with its parametric estimator, we can obtain the parametric estimator of θ, which we denote by $\hat{\theta}$. Therefore, $l(\hat{\theta})$ is the test statistic for the goodness of fit under suitable regularity conditions.

5.3.1 Example

The data consist of 1000 two-dimensional replicates with sample sizes of $50, 100, 200$ and 500, respectively, generated from a Gumbel copula with parameter $\beta = 3$, i.e.

$$C_0(x, y, \beta) = \exp\{-[(-\log x)^\beta + (-\log y)^\beta]^{1/\beta}\}$$

For selected ordered pairs (x, y), where $(x, y) = (0.25, 0.25), (0.5, 0.5)$, and $(0.75, 0.75)$, respectively, we want to test $H_0 : C = C_0$ against the composite alternative $H_1, C \neq C_0$. First, we assume that β, the copula parameter under the null hypothesis, is known, and use the asymptotic distribution of $l(\theta)$, which is χ_1^2. The critical region is

$$\{\theta : l(\theta) > q_{1,(1-\alpha)}\}$$

Table 5.1: Percentage of the estimated reject rates for true θ and estimated θ (in parentheses) at a significant level of 5% by using empirical (EM) and kernel marginal (K_1, K_2) estimates with bandwidths $h_1 = 0.5n^{-1/3}$ and $h_2 = 1/3n^{-1/3}$, respectively.

n	(x, y)	EM	K_1	K_2
50	$(0.25, 0.25)$	3.7(4.4)	9.3(7.2)	6.9(5.2)
	$(0.50, 0.50)$	5.7(3.9)	6.7(8.0)	6.7(4.8)
	$(0.75, 0.75)$	2.7(4.8)	15.6(16.0)	9.3(9.1)
100	$(0.25, 0.25)$	4.0(3.6)	6.9(5.1)	5.3(3.1)
	$(0.50, 0.50)$	2.9(3.0)	6.0(4.2)	5.9(3.4)
	$(0.75, 0.75)$	6.5(5.8)	13.3(11.2)	8.8(6.5)
200	$(0.25, 0.25)$	4.4(4.8)	7.0(5.5)	5.8(4.4)
	$(0.50, 0.50)$	5.1(3.8)	6.6(5.0)	6.4(3.4)
	$(0.75, 0.75)$	4.4(4.1)	11.2(9.6)	6.8(6.0)
500	$(0.25, 0.25)$	5.4(3.3)	5.2(3.6)	4.8(3.6)
	$(0.50, 0.50)$	5.9(4.5)	5.3(4.0)	4.5(3.7)
	$(0.75, 0.75)$	5.4(4.2)	7.5(4.7)	5.5(4.4)

where $q_{1,(1-\alpha)}$ is the $(1-\alpha)$ quantile of χ^2_1, and α is the nominal size of the EL test. Therefore, we reject H_0 if $l(\theta) > q_{1,(1-\alpha)}$. When β is unknown, we assume H_0 is true and use a parametric estimator, $\hat{\beta}$, to replace β. Again, we reject H_0 if $l(\hat{\theta}) > q_{1,(1-\alpha)}$. We used both empirical and kernel estimators of the two marginal distributions of the copula. In the latter case we used the kernel function $k(x) = 3/4(1-x^2), |x| \leq 1$. The estimated rejection rates, based on 1000 replicates with each of the different sample sizes mentioned above at significance level $\alpha = 0.05$ are displayed in Table 5.1.

From Table 5.1, we observe that using empirical estimators of the two marginal distributions outperforms the approach involving kernel estimators in terms of having estimated rejection rates closer to the nominal significance level of 5% for roughly 75% of the scenarios that we considered. And when the kernel estimators of the marginal distribution are used, the type I errors can be noticeably reduced by selecting an appropriate bandwidth. For example, using bandwidth $h_2 = 1/3n^{-1/3}$, the kernel estimator K_2 yields lower type I errors than the alternative K_1 with bandwidth $h_1 = 1/2n^{-1/3}$. Unfortunately, bandwidth selection can be a challenging problem in practice. Chen *et al.* (2009) used cross-validation to select a suitable bandwidth but, as they pointed out, this method of selecting a bandwidth cannot guarantee the required coverage accuracy. Selecting the optimal bandwidth to ensure coverage accuracy in the context of using kernel estimators of the two marginal distributions to obtain EL-based estimates of copula functions is an open problem.

To illustrate visually the association induced by the Gumbel copula with parameter $\beta = 3$, we plot realizations from this copula with sample sizes $n = 50, 100, 200, 500$ in Figures 5.1-5.4, respectively. From these plots we see that the Gumbel copula exhibits strong right-tail dependence and relatively weak left-tail dependence when the sample sizes are greater than 200. However, when the sample size is less than 200, the dependence structure is less evident. Since the Gumbel copula belongs to the Archimedean family of copulas (see the Appendix to this chapter), we also show plots of samples from the other Archimedean copulas, namely the Clayton and Frank copulas, with sample sizes of 200, and parameters $\beta = 2$ and $\beta = 1.81$, respectively; see Figures 5.5, 5.6. These choices of copula parameters for the Clayton and Frank families guarantee they both have the same degree of association as the Gumbel copula family with $\beta = 3$, i.e., a value of 2/3 for Kendall's τ.

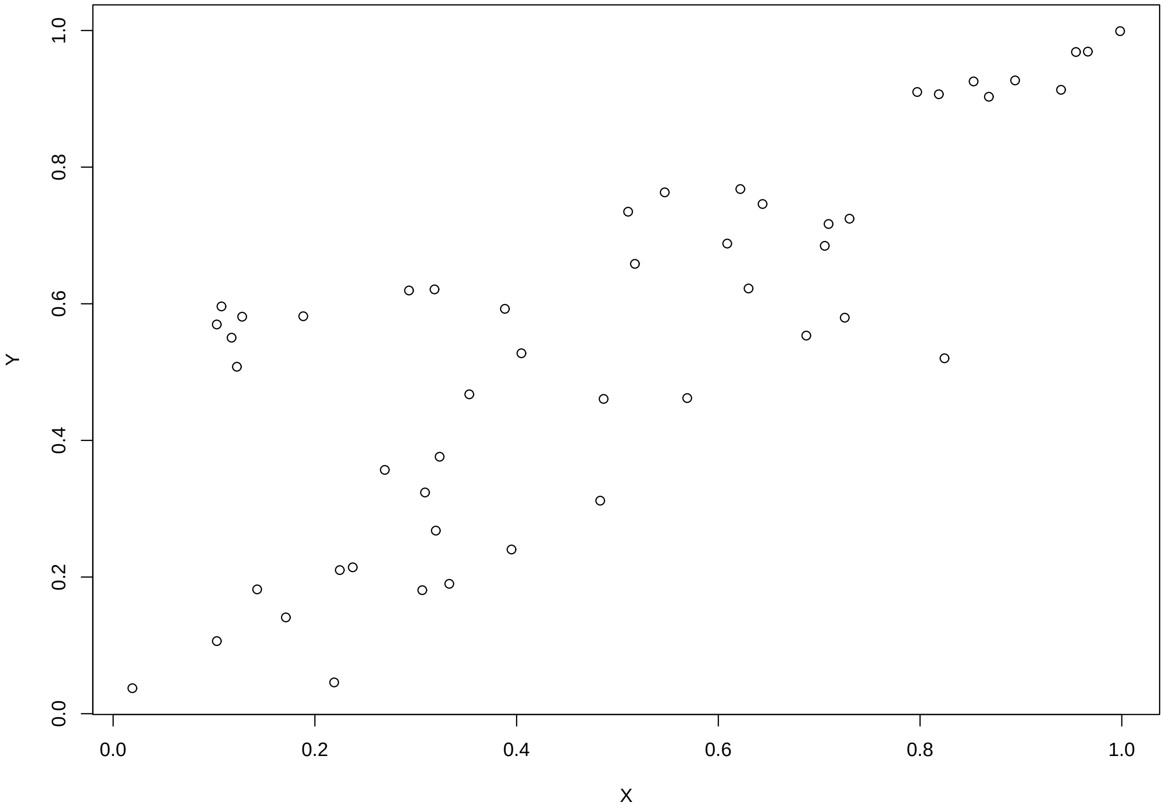

Figure 5.1: A sample of size 50 from a Gumbel copula with $\beta = 3$

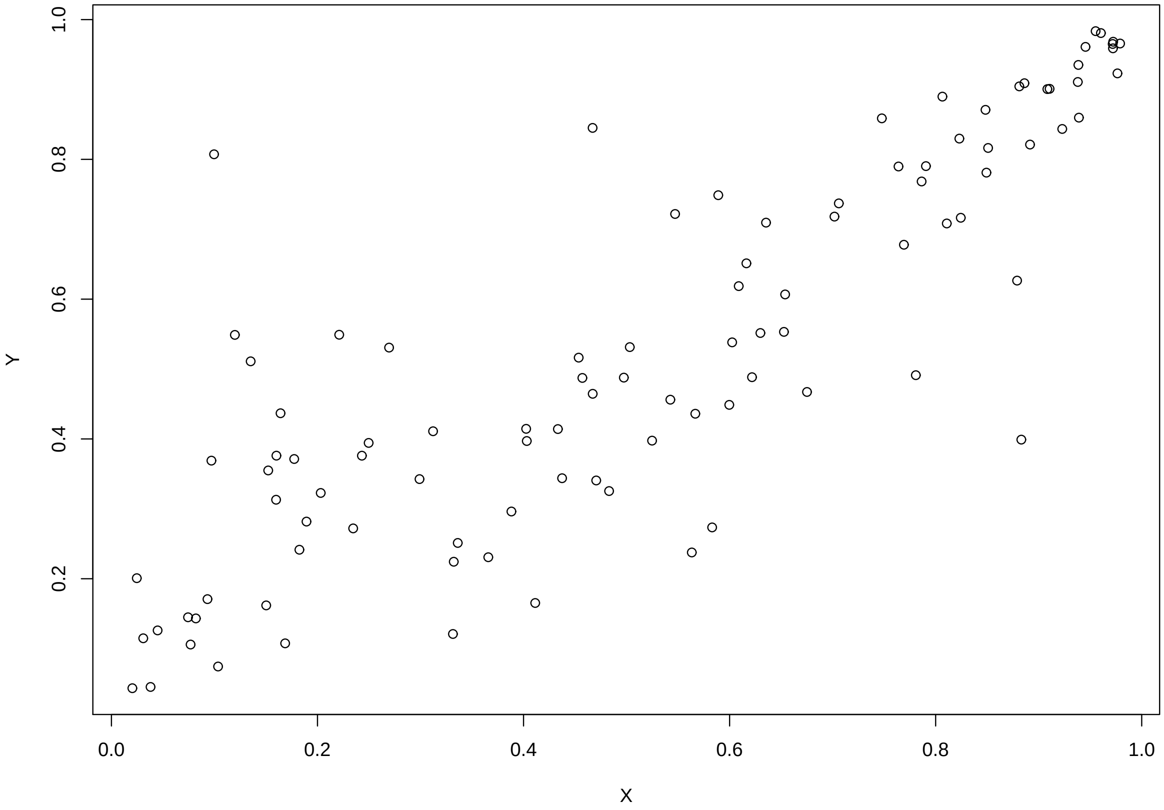

Figure 5.2: A sample of size 100 from a Gumbel copula with $\beta = 3$

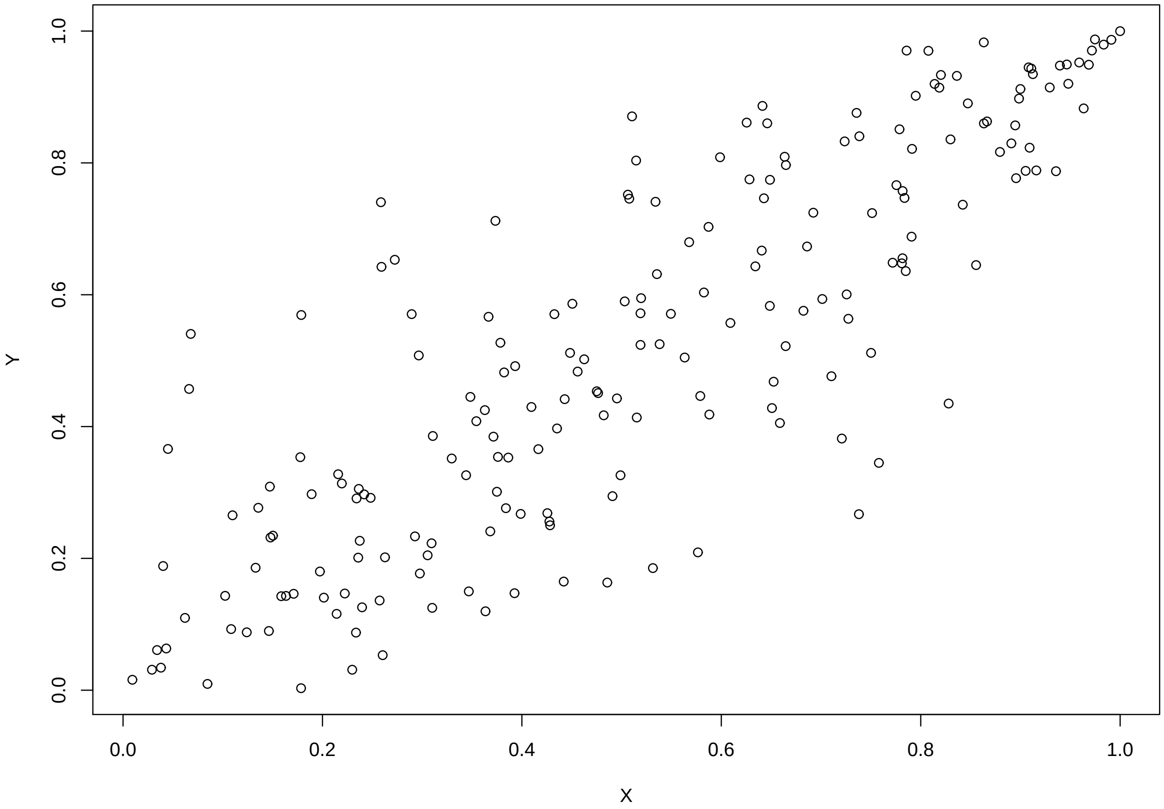

Figure 5.3: A sample of size 200 from a Gumbel copula with $\beta = 3$

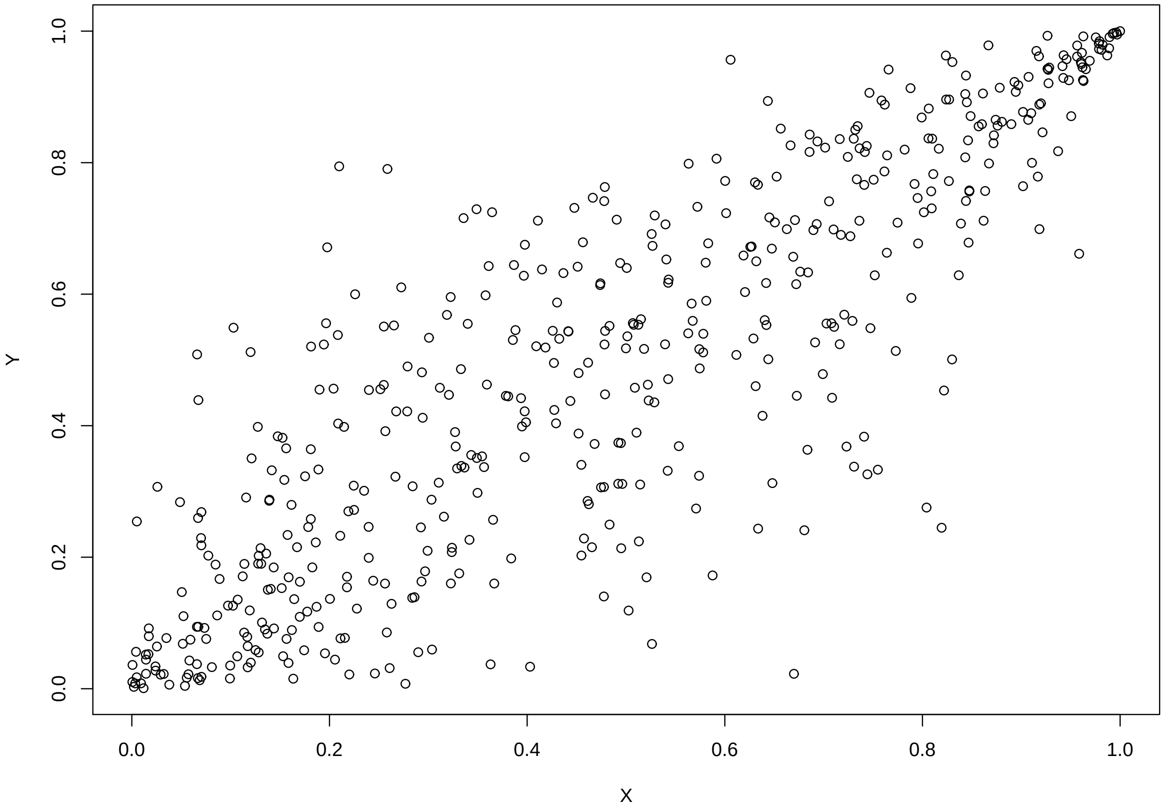

Figure 5.4: A sample of size 500 from a Gumbel copula with $\beta = 3$

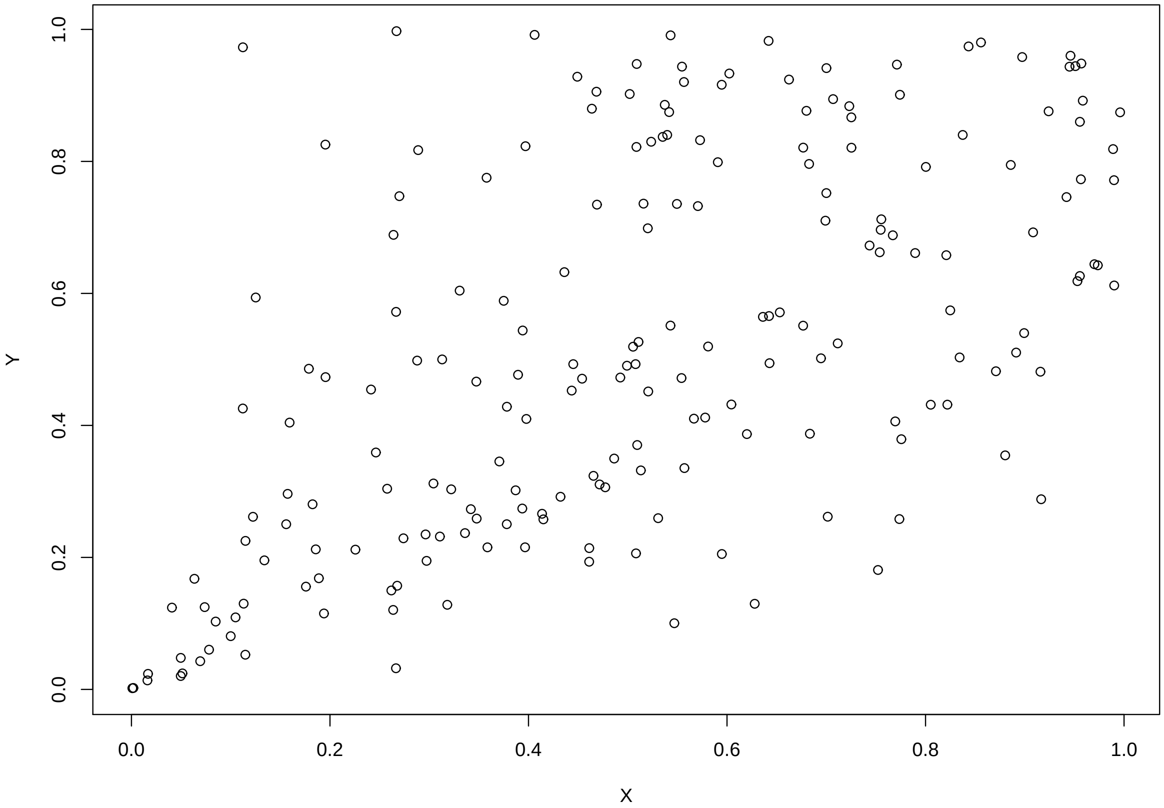

Figure 5.5: A sample of size 200 from a Clayton copula with $\beta = 2$

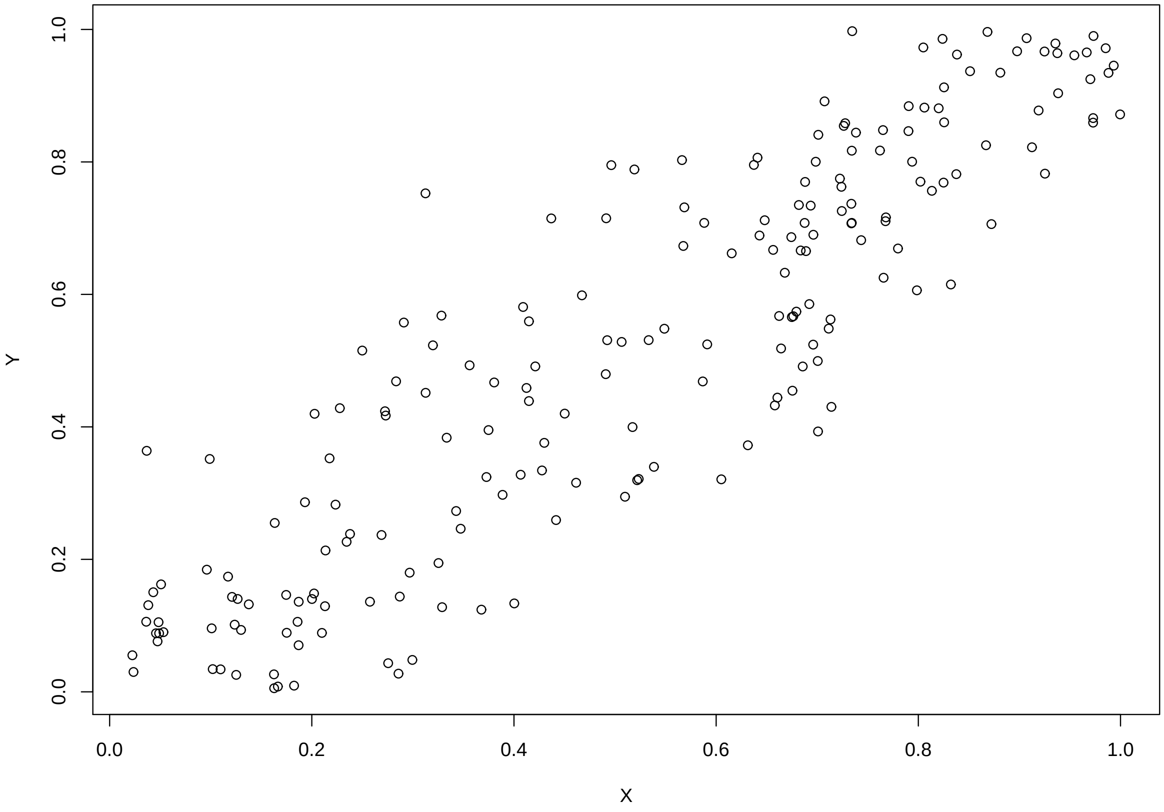

Figure 5.6: A sample of size 200 from a Frank copula with $\beta = 10.05$

5.4 Simulation Study and Data Analysis

Instead of using $l(\theta)$ or $l(\hat{\theta})$, a logarithmic likelihood ratio derived from the empirical copula as the test statistic, to test $H_0 : C \in C_0$ for a given pair, (x, y), researchers may want to test a null hypothesis, $H_0 : C \in C_0$ for all $(x, y) \in D_1 \times D_2 \subseteq [0, 1] \times [0, 1]$. We can use the bootstrap method for a simultaneous confidence band that we described in Chapter 4, either $K_n = \sup_{(x,y)\in D_1\times D_2} l(\theta)$ or $K_n = \sup_{(x,y)\in D_1\times D_2} l(\hat{\theta})$, as a test statistic. Denote the $1 - \alpha$ quantile of K_n under H_0 as the critical value $q_{1-\alpha}$. Then, reject H_0 at the overall significance level of α if $K_n > q_{1-\alpha}$.

In order to compare the power of our test statistic K_n with its empirical counterpart T_n proposed by Genest *et al.* (2009), we generated samples from various Archimedean copulas with sample sizes of 150 and $\tau = .25$. For each copula specified under H_0, we calculated the power of our EL-based test statistic, K_n, and the corresponding empirical test statistic, T_n, if the true copulas are the two other Archimedean copula families. The results are summarized in Table 5.2. We notice that K_n outperforms T_n in terms of having greater power to detect the copula families that are different from the null models in the Archimedean family of copulas. Another advantage of K_n is that it is transformation-invariant, which makes it easy to compute in practice without any of the complexity that other choices of transformations may involve.

We illustrate our method of copula estimation using data from Wieand *et al.* (1989) concerning CA 19-9 and CA 125 diagnostic test measurements in patients with pancreatic cancer (diseased) or pancreatitis (disease-free). For the subset of 90 patients with pancreatic cancer, we display a scatter plot of CA 19-9 and

Table 5.2: Percentage of estimated rejection rate of K_n and T_n, using a sample size of 150, for $\tau = .25$. EM indicates the empirical statistic proposed by Genest *et al.* (2009).

.

Copula under H_0	True copula	EL	EM
Gumbel	Gumbel	5.4(0.71)	4.4(0.65)
	Clayton	75.5(1.36)	62.4(1.53)
	Frank	16.3(1.17)	15.1(1.03)
Clayton	Gumbel	67.9(1.48)	61.2(1.54)
	Clayton	5.2(0.07)	5.4(.71)
	Frank	36.8(1.52)	32.7(1.48)
Frank	Clayton	38.3(1.54)	36.5(1.52)
	Gumbel	18.6(1.23)	18.3(1.22)
	Frank	4.6(0.66)	4.7(0.67)

CA 125 in Figure 5.7. The plot reveals that the relationship between CA 19-9 and CA 125 is clearly not linear, so using a Pearson correlation coefficient to characterize the association between the two biomarkers will not prove satisfactory. To estimate the association appropriately, we plot the empirical distributions of CA 19-9 and CA 125 in Figure 5.8, which suggests that a Frank copula may fit the data. After fitting a Frank copula, we obtain the estimated copula parameter $\hat{\beta} = .6177$ with a corresponding estimated Kendall's $\hat{\tau} = 0.0684$, as well as the test statistic $K_n = 9.721$. By bootstrapping the Frank copula with parameter $\beta = 0.6177$, we estimated the p-value of $K_n = 9.721$ as 0.787. Therefore, we do not reject the null hypothesis of a Frank copula with parameter $\beta = 0.6177$. We also calculated the standard error of $\hat{\beta}$ from the bootstrap samples, which is $\mathrm{SD}(\hat{\beta}) = 0.392$. Since the 95% confidence interval for β is $0.6177 \pm 1.96 * 0.392 = (-0.151, 1.386)$, and using the relationship between β and τ from a Frank copula family, we obtain the

[]

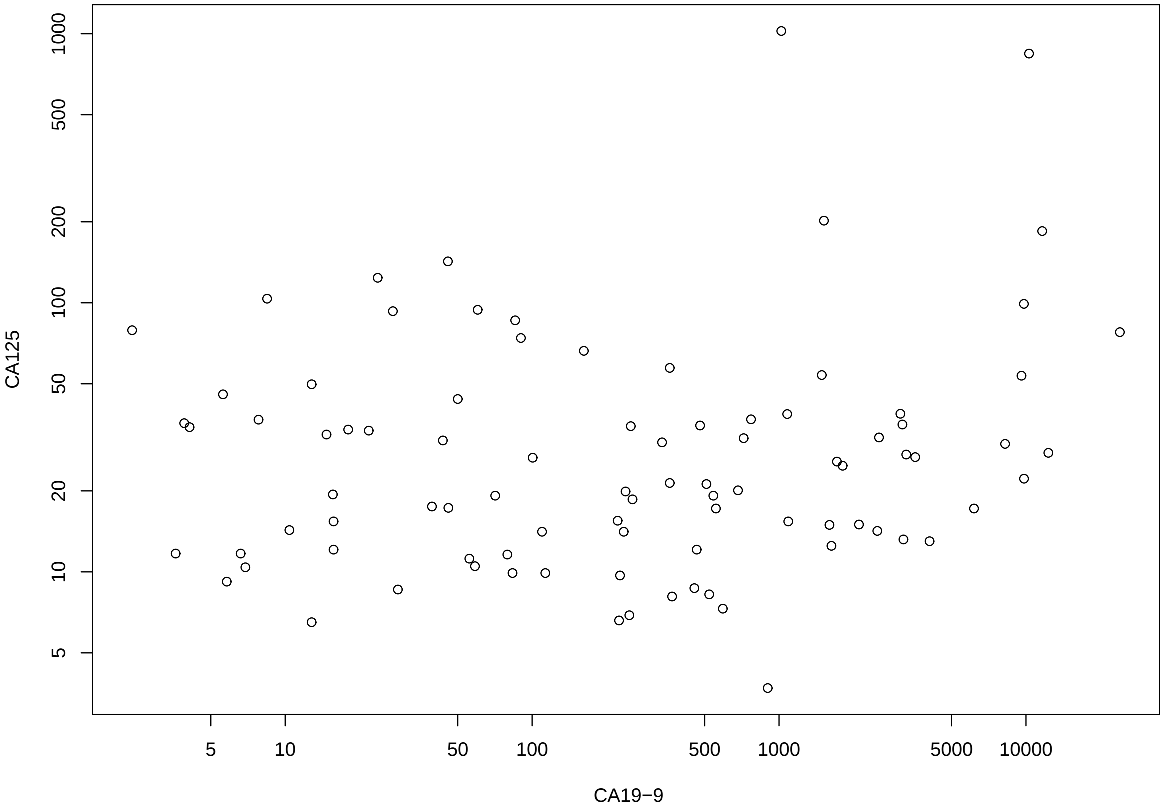

Figure 5.7: Scatter plot of the CA19-9 and CA125 biomarkers for 90 pancreatic cancer patients

corresponding interval is (-0.0167, 0.151)for τ, we conclude that there is no strong association between CA 19-9 and CA 125 measurements.

[]

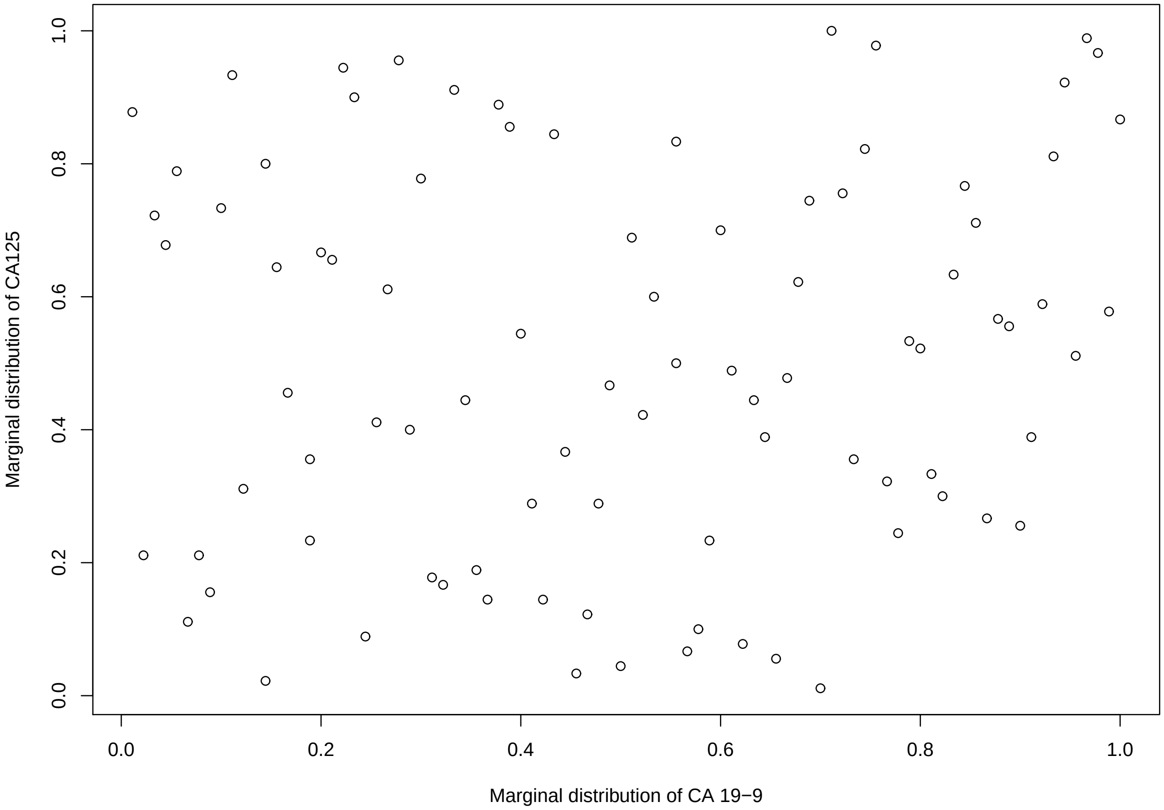

Figure 5.8: EL-based estimated association between the biomarkers CA19-9 and CA 125 in 90 pancreatic cancer patients

5.5 Appendix: Archimedean Family of Copulas

The Archimedean copulas are an important class of copulas, which are defined by

$$C(x, y) = \psi^{-1}(\psi(x) + \psi(y)).$$

where ψ is a generator of the copula. Because of the ease with which they can be constructed and the nice properties they possess, there are three Archimedean copulas in common use, the Gumbel, Clayton and Frank.

5.5.1 Gumbel Copula

The Gumbel copula (also referred to as Gumbel-Hougard copula) is an asymmetric Archimedean copula, exhibiting greater dependence in the positive tail than in the negative. This copula is given by:

$$C_\beta(x, y) = \exp\{-[(-\log x)^\beta + (-\log y)^\beta]^{1/\beta}\};$$

its generator is

$$\psi_\beta(t) = (-\ln(t))^\beta$$

where $\beta \in [1, \infty)$. The relationship between Kendall's τ and the Gumbel copula parameter β is given by:

$$\beta = \frac{1}{1 - \tau}.$$

5.5.2 Clayton Copula

The Clayton copula is an asymmetric Archimedean copula, exhibiting greater dependence in the negative tail than in the positive. This copula is given by:

$$C_\beta(x,y) = \max([x^{-\beta} + y^{-\beta} - 1]^{-1/\beta}, 0);$$

its generator is

$$\psi_\beta(t) = \frac{1}{\beta}(t^{-\beta} - 1)$$

where $\beta \in [-1, \infty)\backslash 0$. The relationship between Kendall's τ and the Clayton copula parameter β is given by:

$$\beta = \frac{2\tau}{1-\tau}.$$

5.5.3 Frank Copula

The Frank copula is a symmetric Archimedean copula given by:

$$C_\beta(x,y) = -\frac{1}{\beta}\ln\left(1 + \frac{(e^{-\beta x} - 1)(e^{-\beta y} - 1)}{(e^{-\beta} - 1)}\right);$$

its generator is

$$\psi_\beta(t) = \ln\left(\frac{e^{-\beta t} - 1}{e^{-\beta} - 1}\right)$$

where $\beta \in (-\infty, \infty)\backslash 0$. The relationship between Kendall's τ and the Frank copula parameter β is given by:

$$\frac{D_1(\beta) - 1}{\beta} = \frac{1-\tau}{4}$$

where $D_1(\beta) = \frac{1}{\beta}\int_0^\beta \frac{t}{e^t - 1} dt$ is a Debye function of the first kind.

Chapter 6

Future work

6.1 Copula estimation for censored data

In multivariate survival analysis that involves multiple possibly related events and right censoring , the commonly used dependency measures such as Pearson's correlation coefficient, Kendall's tau, as well as Spearman's rho cannot fully characterize the association structure between the times of occurrence of these events. In recent years, copulas have become a popular tool for modeling the dependence in a vector of continuous time-to-event random variables subject to censoring; see, for example, Chen and Bandeen-Roche (2005), Lakhal-Chaieb *et al.* (2008), and Lakhal-Chaieb (2010). Under a copula model, complex joint probabilities can be efficiently estimated.

Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are independent and identically distributed random vectors with joint survival function π. For $j = 1, 2$, let $C_{j1}, ..., C_{jn}$ be right-censoring times, with corresponding distribution functions G_j, which are independent of (X, Y). According to Sklar (1959), there exists a unique copula C such that the joint survival function of (X, Y) can be expressed as

$$\pi(s,t) = \Pr(X > s; Y > t) = C\{S_X(s), S_Y(t)\},$$

i.e.,

$$C(x,y) = \pi(S_X^{-1}(x), S_Y^{-1}(y)),$$

where S_X and S_Y are the marginal survival functions of X and Y, respectively, and $S_X(s) = x, S_Y(t) = y$ for $0 < x, y < 1$. Typically, copulas can often be indexed by some parameter β, which reflects the level of association between X and Y.

In order to construct an estimator of $C(x, y)$, which we denote by θ, let $\hat{S}_X(s)$ and $\hat{S}_Y(t)$ be the estimators for S_X and S_Y, respectively. Some natural choices might be the Kaplan-Meier estimators, or perhaps the Nelson-Aalen estimators. Let

$$w_i(s,t) = \hat{S}_{X_i}(s)\hat{S}_{Y_i}(t) - \theta, \quad w_{1i}(s) = \hat{S}_{X_i}(s) - x, \quad w_{2i}(t) = \hat{S}_{Y_i}(t) - y,$$

Following the same procedure that we described in §5.2, we obtain the empirical log-likelihood ratio of θ

$$l(\theta) = 2\sum_{i=1}^{n} \log\{1 + \lambda_1 w_i(s,t) + \lambda_2 w_{1i}(s) + \lambda_3 w_{2i}(t)\}. \tag{6.1}$$

where $\lambda_j, j = 1, 2, 3$, should satisfy equations (5.1) $-$ (5.3) and the n additional constraints $1+\lambda_1 w_i(s,t)+\lambda_2 w_{1i}(s)+\lambda_3 w_{2i}(t) > 1/n$. Using nuisance estimators $\hat{s}, \hat{t}$, derived from the marginal survival functions S_X, S_Y, respectively, the logarithmic likelihood ratio statistic based on the empirical log-likelihood of θ should be

$$l(\theta) = 2\sum_{i=1}^{n} \log\{1 + \lambda_1 w_i(\hat{s}, \hat{t}) + \lambda_2 w_{1i}(\hat{s}) + \lambda_3 w_{2i}(\hat{t})\}. \tag{6.2}$$

Therefore, the point estimator of θ is $\hat{\theta} = \arg\min_\theta l(\theta)$.

In order to obtain an interval estimate of $\theta = C(x, y)$ at a certain fixed point (x, y) of interest, we need the asymptotic distribution of $l(\theta)$, the logarithmic empirical likelihood ratio statistic of θ. Based on this asymptotic distribution, we should be able to construct a point-wise confidence interval for θ at a nominal confidence level of $100(1-\alpha)\%$. Alternatively, we could evaluate a goodness-of-fit test at each of these points of interest.

An alternative way to obtain the required confidence interval or a confidence band for θ is via bootstrapping. Using the $100(1-\alpha)\%$ sample quantile from the bootstrap sample distribution of $l(\hat{\theta})$ as a critical value, we could then derive the corresponding confidence interval at a nominal level of $1-\alpha$, where $\hat{\theta}$ is the point estimate of θ. By taking the supremum of $l(\hat{\theta})$ for $\{(x, y) \in [a, b] \times [c, d] \subseteq [0, 1] \times [0, 1]\}$ from each bootstrap sample, we would obtain the sample quantile at a nominal level of $100(1-\alpha)\%$ from this bootstrap sample distribution, and hence derive a corresponding confidence band for θ, for all $(x, y) \in [a, b] \times [c, d]$ at the overall confidence level of $1-\alpha$.

6.2 EL-based Estimation for a Marginal Survival Function Under Dependent Censoring

In medical studies involving the analysis of multiple events, situations may arise when the censoring mechanism is not independent of event times of interest. For example, to assess the possible benefits of radiation therapy in a cancer clinical trial where the time to death is a primary outcome, researchers may also be interested in the time to relapse or normal tissue toxicity (morbidity). Since morbidity can only occur before death, the time to morbidity is censored by death. Clearly the time to morbidity may be correlated with time to death, therefore we cannot use the Kaplan-Meier estimator, which assumes independent censoring, to estimate the survival function of morbidity time. Thus, we have to address the problem of estimating a survival function in presence of dependent censoring.

This semi-competing risks problem which is defined by one event being censored by another but not vice versa, was first introduced by Fine *et al.*(2001). These researchers proposed estimators of two marginal survival functions by using a parametric copula family to characterize the underlying dependency. Jiang *et al.* (2005) discussed some drawbacks of these estimators and suggested a self-consistent estimator of a copula model. Lakhal-Chaieb *et al.* (2008) used a general copula model, Archimedean copula, and a copula-graphic estimator, to estimate a marginal survival function subject to dependent censoring.

To obtain the estimator of a marginal survival function of interest in the semi-competing risk setting, they first assumed a parametric form of copula, and then estimate the copula parameter, which is independent of the marginal survival functions. Based on this association parameter and the relationship between the copula model and the marginal survival functions, they derived the marginal survival function of interest. However, since they estimated the association parameter first, it is hard to test if the selected copula model is valid. Also, when the copula model is misspecified, the resulting estimator of the survival function would be biased or invalid.

To deal with the model misspecification problem, we plan to use the empirical likelihood method to estimate the association parameter and marginal survival function nonparametrically. With this model and empirical likelihood ratio statistics, we should be able to carry out a goodness-of-fit test for the selected copula, and corresponding marginal survival estimator.

6.3 EL for Frailty Models

Clustered survival data are encountered in many scientific disciplines including human and veterinary medicine, biology, epidemiology, public health and demography. Frailty models provide a powerful tool for analyzing clustered survival data. In recent years a number of papers and a wide variety of frailty models have been investigated. L. Duchateau and P. Janssen (2008) gave a comprehensive introduction to frailty models in their book.

In order to use a frailty model to accommodate the dependency of clustered survival data, the distribution of the frailty terms must be specified. Duchateau and Janssen (2008) discussed several distributions for frailty terms, and identified the corresponding type of dependence that they induce on the event times in the cluster. However as these authors observe, the cluster dependency induced by certain frailty distributions, such as the lognormal, is hard to evaluate so that the set of parametric families that can be chosen for frailty terms remains limited in practice.

To overcome this limitation of the parametric frailty model, we plan to investigate the potential of the frailty model in a nonparametric setting, i.e., no parametric distribution for the frailty terms, by using empirical likelihood. By comparison with using a parametric frailty model, we would be able to study the efficiency of an estimator of interest and further identify the possible model misspecification problem involved in making parametric assumptions for frailty terms.

As an alternative to frailty models, copulas can also take the clustering of data into account, and in some situations these two models are equivalent. Comparing the two models in the same simulation setting may provide some insights into important practical aspects of model selection for clustered data.

References

[1] M.G. Akritas. Bootstrapping the Kaplan-Meier estimator. *Journal of the American Statistical Association*, 81:1032–1038, 1986.

[2] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1993.

[3] O. Bie, Ø . Borgan, and K. Liestø l. Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics*, 14:221–233, 1987.

[4] K. Bogaerts and E. Lesaffre. Modeling the association of bivariate interval-censored data using the copula approach. *Statistics in Medicine*, 27:6379–6392, 2008.

[5] N. Breslow. Discussion of Professor Cox's paper. *Journal of the Royal Statistical Society, Series B*, 34:216–217, 1972.

[6] J. Chen, L. Peng, and Y. Zhao. Empirical likelihood based confidence intervals for copulas. *Journal of Multivariate Analysis*, 100:137–151, 2009.

[7] M.C. Chen and K. Bandeen-Roche. A diagnostic for association in bivariate survival models. *Lifetime Data Analysis*, 11:245–264, 2005.

[8] S.X. Chen and H. Cui. On the second-order properties of empirical likelihood with moment restrictions. *Journal of Econometrics*, 141:492–516, 2007.

[9] G. Claeskens, B.-Y. Jing, L. Peng, and W. Zhou. Empirical likelihood confidence regions for comparison distribution and ROC curves. *Canadian Journal of Statistics*, 31:173–190, 2003.

[10] G. Claeskens and I. Van Keilegom. Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, 31:1852–1884, 2003.

[11] D.R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

[12] L. Duchateau and P. Janssen. *The Frailty Model.* Springer-Verlag, New York, 2008.

[13] B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[14] B. Efron. Censored data and the bootstrap. *Journal of the American Statistical Association*, 76:312–319, 1981.

[15] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia*, 38, 1982.

[16] J. Fan, H. Hung, and W.H. Wong. Geometric understanding of likelihood ratio statistics. *Journal of the American Statistical Association*, 95:836–841, 2000.

[17] J.P. Fine, H. Jiang, and R Chappell. On semicompeting risks. *Biometrika*, 88:907–919, 2001.

[18] C. Genest, B. Remillard, and D. Beaudoin. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics & Economics*, 44:199–213, 2009.

[19] R.D. Gill, J.A. Wellner, and J. Prästgaard. Non- and semi-parametric maximum likelihood estimators and the von mises method (part 1) [with discussion and reply]. *Scandinavian Journal of Statistics*, 16:97–128, 1989.

[20] P. Hall. *The Bootstrap and Edgeworth Expansions.* Springer, New York, 1992.

[21] P. Hall and A.B. Owen. Empirical likelihood confidence bands in density estimation. *Journal of Computational and Graphic Statistics*, 2:273–289, 1993.

[22] M. Hollander, I.W. McKeague, and J. Yang. Likelihood ratio-based confidence bands for survival functions. *Journal of the American Statistical Association*, 92:215–226, 1997.

[23] P. Hougaard. *Analysis of Multivariate Survival Data.* Springer-Verlag, New York, 2000.

[24] H. Jiang, J. P. Fine, R. Kosorok, and R. Chappell. Pseudo self-consistent estimation of a copula model with informative censoring. *Scandinavian Journal of Statistics*, 32:1–20, 2005.

[25] H. Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall, New York, 1997.

[26] S. Johansen. An extension of Cox's regression model. *International Statistical Review*, 51:165–174, 1983.

[27] J. D. Kalbfleisch and R. L. Prentice. Estimation of the average hazard ratio. *Biometrika*, 68:105–112, 1981.

[28] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[29] M.L. Lakhal-Chaieb. Copula inference under censoring. *Biometrika*, 97:505–512, 2010.

[30] M.L. Lakhal-Chaieb, L.-P. Rivest, and B. Abdous. Estimating survival and association in a semicompeting risks model. *Biometrics*, 64:180–188, 2008.

[31] W. Leisenring and M.S. Pepe. Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. *Biometrics*, 54:444–452, 1998.

[32] G. Li. On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statistics and Probability Letters*, 25:95–104, 1995a.

[33] G. Li. Nonparametric likelihood ratio estimation of probabilities for truncated data. *Journal of the American Statistical Association*, 90:997–1003, 1995b.

[34] G. Li, J. Qin, and R.C. Tiwari. Semiparametric likelihood ratio-based inferences for truncated data. *Journal of the American Statistical Association*, 20:236–245, 1997.

[35] G. Li, R. C. Tiwari, and M. T. Wells. Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. *Biometrika*, 86:487–502, 1999.

[36] G. Li and I. van Keilegom. Likelihood ratio confidence bands in non-parametric regression with censored data. *Scandinavian Journal of Statistics*, 29:547–562, 2002.

[37] C.J. Lloyd and Z. Yong. Kernel estimators of the ROC curve are better than empirical. *Statistics and Probability Letters*, 44:221–228, 1999.

[38] D.E. Matthews and V.T. Farewell. *Using and Understanding Medical Statistics. 4th Completely Revised and Enlarged Edition.* S Karger AG, Basel, 2007.

[39] I.W. McKeague and Y. Zhao. Simultaneous confidence bands for ratios of survival functions via empirical likelihood. *Statistics and Probability Letters*, 60:405–415, 2002.

[40] I.W. McKeague and Y. Zhao. Comparing distribution functions via empirical likelihood. *The International Journal of Biostatistics 1, Issue 1, Article 5.*, 2005.

[41] S.A. Murphy. Likelihood ratio-based confidence intervals in survival analysis. *Journal of the American Statistical Association*, 90:1399–1405, 1995.

[42] A.B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249, 1988.

[43] A.B. Owen. *Empirical Likelihood.* CHAPMAN and HALL/CRC, Boca Raton, 2001.

[44] A. Pakes and D. Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica*, 57:1027–1057, 1989.

[45] X.R. Pan and M. Zhou. Empirical likelihood ratio in terms of cumulative hazard function for censored data. *Journal of Multivariate Analysis*, 80:166–188, 2002.

[46] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *Annuals of Statisitcs*, 22:300–325, 1994.

[47] N. Reid. Estimating the median survival time. *Biometrika*, 68:601–608, 1981.

[48] M. Schemper. Cox analysis of survival data with nonproportional hazard functions. *The Statistician*, 41:455–465, 1992.

[49] S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53:683–690, 1991.

[50] A. Sklar. Fonctions de repartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–31, 1959.

[51] L. Sun, L.M. Wang, and J.G. Sun. Estimation of the association for bivariate interval-censored failure time data. *Scandinavian Journal of Statistics*, 33:637–649, 2006.

[52] D.R. Thomas and G.L. Grunkemeier. Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70:865–871, 1975.

[53] Q.H. Wang and B.Y. Jing. Empirical likelihood for a class of functionals of survival distribution with censored data. *Annals of the Institute of Statistical Mathematics*, 53:517–527, 2001.

[54] G. Wei and D.E. Schaubel. Estimating cumulative treatment effects in the presence of nonproportional hazards. *Biometrics*, 64:724–732, 2008.

[55] S. Wieand, M.H. Gail, B. James, and K James. Nonparametric procedures for comparing diagnostic tests with paired or unpaired data. *Biometrika*, 75:585–592, 1989.

[56] S.S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62, 1938.

[57] R. Xu and J. O'Quigley. Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1:423–439, 2000.

[58] W. Zhou and B.Y. Jing. Adjusted empirical likelihood method for the quantiles. *Annals of the Institute of Statistical Mathematics*, 55:689–703, 2003.