Spring 4-25-2011

Empirical Likelihood Confidence Intervals for ROC Curves with Missing Data

Yueheng An

EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR ROC CURVES WITH MISSING DATA

by

YUEHENG AN

Under the Direction of Dr. Yichuan Zhao

ABSTRACT

The receiver operating characteristic, or the ROC curve, is widely utilized to evaluate the diagnostic performance of a test, in other words, the accuracy of a test to discriminate normal cases from diseased cases. In the biomedical studies, we often meet with missing data, which the regular inference procedures cannot be applied to directly. In this thesis, the random hot deck imputation is used to obtain a 'complete' sample. Then empirical likelihood (EL) confidence intervals are constructed for ROC curves. The empirical log-likelihood ratio statistic is derived whose asymptotic distribution is proved to be a weighted chi-square distribution. The results of simulation study show that the EL confidence intervals perform well in terms of the coverage probability and the average length for various sample sizes and response rates.

INDEX WORDS: Confidence interval, Missing data, ROC curve, Empirical likelihood

EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR ROC CURVES WITH MISSING DATA

by

YUEHENG AN

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2011

EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS FOR ROC CURVES WITH MISSING DATA

by

YUEHENG AN

Committee Chair: Dr. Yichuan Zhao

Committee: Dr. Yuanhui Xiao

Dr. Ruiyan Luo

Dr. Xu Zhang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2011

ACKNOWLEDGEMENTS

I would like to gratefully and sincerely acknowledge those who have assisted me in my graduate study.

First and foremost, I would like to express my gratitude to my advisor, Dr. Yichuan Zhao, for all his guidance, patience and supports. I have learnt a lot from his classes, Monte Carlo Methods and Statistical Inference II, as well as his supervision on this thesis.

I would also like to thank my thesis committee, Dr. Yuanhui Xiao, Dr. Ruiyan Luo, and Dr. Xu Zhang for taking their precious time to read my thesis and providing me valuable suggestions.

Besides, I would like to acknowledge all the professors in the Department of Mathematics and Statistics at Georgia State University who taught me in my graduate study. They all assisted me to develop my statistical knowledge and skills.

Moreover, I must give a special thank to my parents and boyfriend who have always been there for me with unconditional love and encouragement. In addition, I need to thank my classmates Hanfang Yang, Meng Zhao, Ye Cui and Huayu Liu for their useful help and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 ROC Curve

The receiver operating characteristic, or the ROC curve simply, has been extensively used to evaluate the diagnostic tests and is currently the best-developed statistical tool for describing the performance of such tests. ROC curves provide a comprehensive and visually attractive way to summarize the accuracy of predictions. Generally speaking, the ROC curve is a graphical plot of the *sensitivity*, or true positives, versus $(1 - specificity)$, or false positives. It has been in use for years, which was first developed during World War II for signal detection. Its potential for medical diagnostic testing was recognized as early as 1960, although it was in the early 1980s that it became popular, especially in radiology (Pepe, 2003). Nowadays, ROC curves enjoy broader applications in medicine (Shapiro, 1999).

In a medical test resulting in a continuous measurement T, the disease is diagnosed if $T > t$, for a given threshold t. Let D denote the disease status with

$$D = \begin{cases} 1, & \text{diseased,} \\ 0, & \text{non-diseased} \end{cases}$$

and the corresponding true and false positive fractions at t be $TPF(t)$ and $FPF(t)$,

respectively, where $TPF(t) = Pr(T \geq t|D = 1)$, $FPF(t) = Pr(T \geq t|D = 0)$. The ROC curve is the entire set of possible true and false positive fractions attained by dichotomizing T with different thresholds (Pepe, 2003). That is, the ROC curve is

$$ROC(\cdot) = \{(FPF(t), TPF(t)),\, t \in (-\infty, \infty)\}.$$

When t increases, both $FPF(t)$ and $TPF(t)$ decrease. For extreme cases, when $t \to \infty$, we can get $\lim_{t\to\infty} TPF(t) = 0$ and $\lim_{t\to\infty} FPF(t) = 0$. On the other hand, when $t \to -\infty$, we have $\lim_{t\to-\infty} TPF(t) = 1$ and $\lim_{t\to-\infty} FPF(t) = 1$. Thus, the ROC curve is actually a monotone increasing function in the positive quadrant.

On the other hand, considering the results of a certain test in two populations, diseased against non-disease, we will rarely observe a perfect separation between the two groups.

Suppose the distribution function of T is F conditional on disease and G conditional on non-disease. The ROC curve is defined as the graph $(1 - G(t); 1 - F(t))$ for various values of the threshold t, or in other words, $sensitivity$ versus $(1-specificity)$, $power$ versus $size$ for a test with critical region $\{T > t\}$.

Now we consider a specific test in two populations, one with disease and the other without disease. At a fixed cut-off point or threshold t, the sensitivity and specificity are defined as $Se = Pr(X \geq t)$ and $Sp = Pr(Y < t)$, respectively. If $F(\cdot)$ is the distribution function of X and $G(\cdot)$ is the distribution function of Y , the sensitivity and specificity can then be written as $Se = 1 - F(t)$ and $Sp = G(t)$. Then the ROC curve is actually a plot of $1 - F(t)$ versus $1 - G(t)$, for $-\infty < t < \infty$. At a

given level $q = (1 - specificity)$, the ROC curve can be represented by

$$\Delta = 1 - F(G^{-1}(1-q)), \quad for\ 0 < q < 1,$$

where G^{-1} is the inverse function of G, i.e., $G^{-1}(q) = inf\{t : G(t) \geq q\}$.

1.2 Missing Data and Random Hot Deck Imputation

It is typically to assume that all the responses in the sample are available for statistical inferences. In practical applications, however, this may not be true. Some of the responses may not be obtained for many reasons such as that some sampled units are not willing to provide certain information, some of the investigators failed to gather the correct information, uncontrollable factors leaded loss of information and so forth. As a matter of fact, missing responses happen in a regular base in mail enquiries, medical studies, opinion polls, market research surveys and other scientific experiments (Wang and Rao, 2002).

Consider the following simple random samples of incomplete data associated with populations (x, δ_x) and (y, δ_y):

$$(x_i, \delta_{x_i}), i = 1, ..., m; \quad (y_j, \delta_{y_j}), j = 1, ..., n,$$

where

$$\delta_{x_i} = \begin{cases} 0 & \text{if } x_i \text{ is missing,} \\ 1 & \text{otherwise,} \end{cases}$$

$$\delta_{y_j} = \begin{cases} 0 & \text{if } y_j \text{ is missing,} \\ 1 & \text{otherwise.} \end{cases}$$

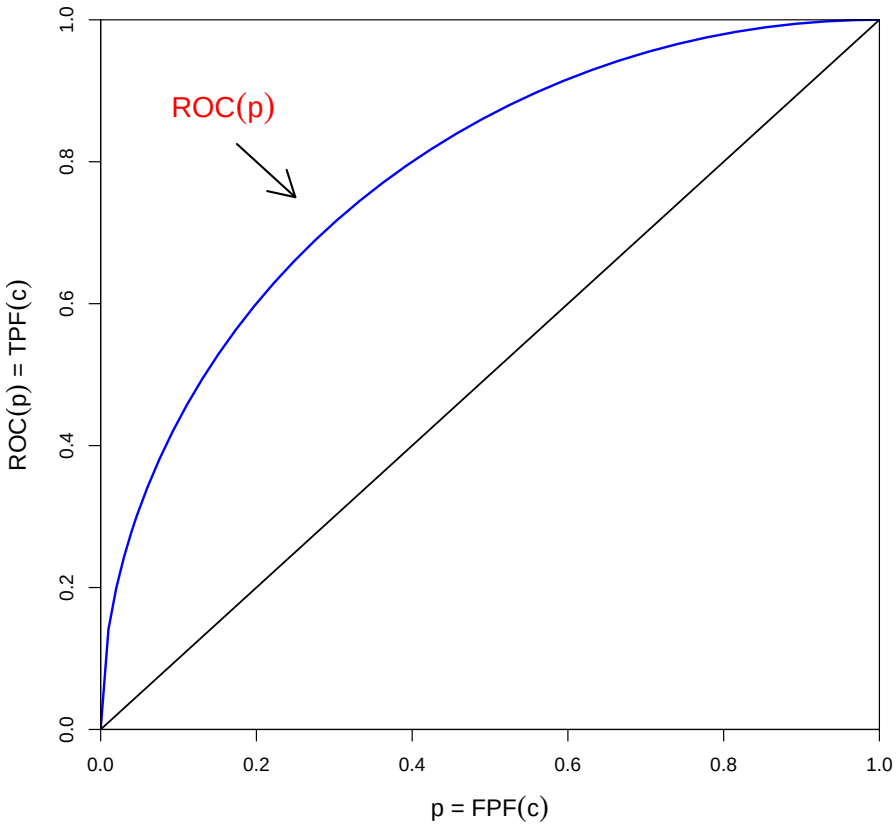

Figure 1.1. ROC Curve

Let $r_x = \sum_1^m \delta_{x_i}$, $r_y = \sum_1^n \delta_{y_j}$, $m_x = m - r_x$ and $m_y = n - r_y$. Denote the sets of respondents with respect to x and y as s_{rx} and s_{ry}, and the sets of non-respondents with respect to x and y as s_{m_x} and s_{m_y}, thus the means of respondents with respect to x and y as

$$\overline{x}_r = \frac{1}{r_x} \sum_{i \in s_{r_x}} x_i, \quad \overline{y}_r = \frac{1}{r_y} \sum_{i \in s_{r_y}} y_i.$$

Throughout this thesis, we assume that x, y are missing completely at random (M-CAR), i.e., $P(\delta_x = 1|x) = P_1(constant)$ and $P(\delta_y = 1|y) = P_2(constant)$. We also assume that (x, δ_x) and (y, δ_y) are independent.

Imputation-based procedure is one of the most common methods or treatments dealing with missing data (Little and Rubin, 2002). Standard statistics methods to the complete data are applied after imputation, that is, to impute a value for each missing datum. Deterministic imputation and random imputation are commonly used imputation methods.

Here random hot deck imputation method is utilized to impute the missing values, since the deterministic imputation is not proper in making inference for distribution functions.

Let x_i^* and y_j^* be the imputed values for the missing data with respect to x and y. Random hot deck imputation selects a simple random sample of size m_x with replacement from s_{r_x} and then uses the associated x-values as donors, that is, $x_i^* = x_k$ for some $k \in s_{r_x}$. Similarly, we obtain y_j^*. Let

$$x_{I,i} = \delta_{x_i} x_i + (1 - \delta_{x_i}) x_i^*, \quad y_{I,j} = \delta_{y_j} y_j + (1 - \delta_{y_j}) y_j^*,$$

$i = 1, ..., m$, $j = 1, ..., n$, which represent the 'complete' data after imputation.

In this thesis, we explore the asymptotic properties of the empirical likelihood

ratio statistic for Δ based on $x_{I,i}$, $i = 1, ..., m$, $y_{I,j}$, $j = 1, ..., n$. The results are used to build asymptotic confidence intervals for Δ.

1.3 Empirical Likelihood

Empirical likelihood, a nonparametric method of statistical inference, use likelihood methods without having to assume that the data come from a known family of distributions (Owen, 2001). Thus, empirical likelihood can be thought of as a likelihood without parametric assumptions, and as a bootstrap without resampling. Among many of the advantages over competitors, improvement of the confidence region and increase of accuracy in coverage are the most appealing features resulting from using auxiliary information and easy implementation.

For a random variable $X \in \Re$, the cumulative distribution function (CDF) is defined as

$$F(x) = Pr(X \leq x),$$

where $-\infty < x < \infty$.

Denote

$$F(x-) = Pr(X < x),$$

then

$$Pr(X = x) = F(x) - F(x-).$$

Let $X_1, ..., X_n$ be n independent samples follow the common CDF of F_0, the empirical likelihood cumulative distribution function (ECDF) of $X_1, ..., X_n$ is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}, \quad -\infty < x < \infty,$$

where $I_{X_i \leq x}$ is the indicator function.

For $X_1, ..., X_n$, X_i $i.i.d.$ $\sim F_0$, $i = 1, ..., n$, the nonparametric likelihood of the CDF F is

$$L(F) = \prod_{i=1}^{n} (F(x_i) - F(x_i-)).$$

For a distribution F, ratios of the nonparametric likelihood for hypothesis testings and confidence intervals is

$$\mathcal{R}(F) = \frac{L(F)}{L(F_n)}.$$

When F is continuous, the likelihood of F, $L(F) \equiv 0$. Owen (2001) also proves that given $F(x) \neq F_n(x)$,

$$L(F) < L(F_n),$$

which means the ECDF is the nonparametric maximum likelihood estimate (NPMLE) of F.

Suppose Γ is the space of a distribution function G defined on $[0, \infty)$. For $F \in \Gamma$, let $\theta = G(F)$, the profile likelihood ratio is defined as:

$$\mathcal{R}(\theta) = \sup\{\mathcal{R}(F) | G(F) = \theta, F \in \Gamma\}.$$

Then Owen (1990) gives a remarkable result similar to the Wilk's theorem (Owen, 1988, 1990, 2001), that is, the empirical likelihood ratio has a limiting chi-square distribution, as following:

$$-2 \log \mathcal{R}(\theta_0) \to \chi^2_m, \quad as\ n \to \infty,$$

where m is the degrees of freedom $d.f. = dimension(\theta)$, θ_0 is the true value of θ. Based on this fact, we can construct the confidence intervals for θ.

1.4 Structure

The rest of the thesis is organized as follows. In chapter 2, the empirical likelihood ratio statistic is constructed, the limiting distribution of the statistic is given, and the empirical likelihood based confidence interval for the cut-off points on the ROC curve is constructed. In chapter 3, we report that the results of a simulation study on the finite sample performance of empirical likelihood based confidence interval on Δ's. The conclusion is given in chapter 4, and all the technical derivations are provided in the appendices.

Chapter 2

INFERENCE PROCEDURE

In this chapter, we construct the empirical likelihood ratio statistic, develop the limiting distribution of the statistic, and give the empirical likelihood based confidence interval for the ROC curve Δ.

Take the bandwidth $a = a_m > 0$, $b = b_n > 0$ and the kernels K_1 and K_2, where $a_m \to 0$ as $m \to \infty$ and $b_n \to 0$ as $n \to \infty$. Define

$$F(t) = \int_{-\infty}^{t/a} K_1(u)du, \quad G(t) = \int_{-\infty}^{t/b} K_2(u)du.$$

Similar to Qin and Zhao (1997) and Chen and Hall (1993), the empirical likelihood function is defined as

$$\prod_{i=1}^{m} p_i \prod_{j=1}^{n} q_j, \tag{2.1}$$

where $p_i > 0$, i=1,...,m, $\sum_i p_i = 1$, and $q_j > 0$, j=1,...,n, $\sum_j q_j = 1$. Define the log-empirical likelihood ratio statistic:

$$R(\Delta) = \sup_{p_i, i=1,...,m, q_j, j=1,...,n, \theta} \{\sum_{i=1}^{m} log(mp_i) + \sum_{j=1}^{n} log(nq_j)\} = \sup_{\theta} R(\Delta, \theta), \tag{2.2}$$

where

$$R(\Delta,\theta)=\sup_{p_i,i=1,...,m,q_j,j=1,...,n,}\{\sum_{i=1}^{m}log(mp_i)+\sum_{j=1}^{n}log(nq_j)\}, \tag{2.3}$$

and p_i, q_j are subject to restrictions:

$$\sum_{i=1}^{m}p_i=1,\quad \sum_{i=1}^{m}p_iF(\theta-x_{I,i})=1-\Delta,\quad p_i>0, i=1,...,m, \tag{2.4}$$

and

$$\sum_{j=1}^{n}q_j=1,\quad \sum_{j=1}^{n}q_jG(\theta-y_{I,j})=1-q,\quad q_j>0, j=1,...,n. \tag{2.5}$$

Denote

$$\omega_1(x_{I,i},\theta,\Delta)=F(\theta-x_{I,i})-1+\Delta,\quad i=1,...,m,$$

$$\omega_2(y_{I,j},\theta,\Delta)=G(\theta-y_{I,j})-1+q,\quad j=1,...,n.$$

From Lagrange multipliers, we can show that

$$R(\Delta,\theta)=-\sum_{i=1}^{m}log\{1+\lambda_1(\theta)\omega_1(x_{I,i},\theta,\Delta)\}-\sum_{j=1}^{n}log\{1+\lambda_2(\theta)\omega_2(y_{I,j},\theta,\Delta)\}, \tag{2.6}$$

where $\lambda_j(\theta), j=1,2$, are determined by the following two equations:

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\omega_1(x_{I,i},\theta,\Delta)}{1+\lambda_1(\theta)\omega_1(x_{I,i},\theta,\Delta)}=0, \tag{2.7}$$

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\omega_2(y_{I,j},\theta,\Delta)}{1+\lambda_2(\theta)\omega_2(y_{I,j},\theta,\Delta)}=0. \tag{2.8}$$

Let $\partial\mathcal{R}(\theta,\Delta)/\partial\theta=0$.We can obtain the empirical likelihood equation:

$$\frac{1}{m}\lambda_1(\theta)\sum_{i=1}^{m}\frac{\alpha_1(x_{I,i},\theta,\Delta)}{1+\lambda_1(\theta)\omega_1(x_{I,i},\theta,\Delta)}+\frac{1}{m}\lambda_2(\theta)\sum_{j=1}^{n}\frac{\alpha_2(y_{I,j},\theta,\Delta)}{1+\lambda_2(\theta)\omega_2(y_{I,j},\theta,\Delta)}=0, \tag{2.9}$$

where $\alpha_1(x_{I,i},\theta,\Delta)=\dfrac{1}{a}K_1((\theta-x_{I,i})/a)$ and $\alpha_2(y_{I,j},\theta,\Delta)=\dfrac{1}{b}K_2((\theta-y_{I,j})/b)$.

Use θ_0 to denote the true value of θ, we have the following assumptions:

(i) $\theta_0\in\Omega$ and Ω is an open interval.

(ii) Denote $f(t)=\partial F(t)/\partial t$ and $g(t)=\partial G(t)/\partial t$. For some $t_0\geq 2$, suppose that $f^{(t_0-1)}(t)$ and $g^{(t_0-1)}(t)$ exist and are uniformly continuous and bounded in a neighborhood of θ_0. Assume that $f(\theta_0)g(\theta_0)>0$.

(iii) $n/m\to k\,(0<k<\infty)$ as $m,n\to\infty$.

(iv) For K_i's, $i=1,2$, are bounded and satisfy Lipschitz condition of order 1 and $K_i^{(2)}$ exists and is bounded. For $i=1,2$, assume that for some $C>0$,

$$\int_{|u|>C/a^{t_0}} K_1(u)du=O(a^{t_0}),$$

$$\int_{|u|>C/b^{t_0}} K_2(u)du=O(b^{t_0}),$$

$$\int |u^{t_0}K_i(u)|du<\infty,$$

$$\int u^j K_i(u)du=\begin{cases}1 & j=0,\\ 0 & 1\leq j\leq t_0-1.\end{cases}$$

(v) There exists $r\,(1/3<r<1/2)$ such that $n^r a^{t_0}\to 0$, $n^r b^{t_0}\to 0$, $n^r a\to\infty$, and $n^r b\to\infty$ as $m,n\to\infty$.

Theorem 1 gives the asymptotic distribution of the log-empirical likelihood ratio statistic. The proof of Theorem 1 is given in the Appendix A.

Theorem 1. *Suppose that assumptions (i) through (v) are satisfied, then there exists a root $\theta_{m,n}$ of equation (2.9) such that $\mathcal{R}(\Delta,\theta)$ attains its local maximum at $\theta_{m,n}$ and*

as $m, n \to \infty$,

$$\sqrt{m}(\theta_{m,n} - \theta_0) \xrightarrow{d}$$
$$N\left(0, q(1-q)\Delta(1-\Delta)\frac{\{q(1-q)(1-P_1+P_1^{-1})f^2(\theta_0) + k\Delta(1-\Delta)(1-P_2+P_2^{-1})g^2(\theta_0)\}}{c_0^2}\right),$$

$$-\mathcal{R}(\Delta, \theta_{m,n}) \xrightarrow{d} \frac{k\Delta(1-\Delta)(1-P_1+P_1^{-1})g^2(\theta_0) + q(1-q)(1-P_2+P_2^{-1})f^2(\theta_0)}{c_0}\chi_1^2,$$

where

$$c_0 = q(1-q)f^2(\theta_0) + k\Delta(1-\Delta)g^2(\theta_0).$$

It is interesting to notice that the empirical likelihood ratio under imputation is asymptotically distributed as a scaled chi-square variable. The reason for this deviation from the standard results is that the 'complete' data after imputation are dependent.

Denote

$$a_0(\Delta) = \{k\Delta(1-\Delta)(1-P_1+P_1^{-1})g^2(\theta_0) + q(1-q)(1-P_2+P_2^{-1})f^2(\theta_0)\}/c_0.$$

To construct a confidence interval for Δ using the above result, we need to get a consistent estimator of $a_0(\Delta)$. P_1 and P_2 can be consistently estimated by

$$\hat{P}_1 = \frac{1}{m}\sum_{i=1}^{m}\delta_{x_i}$$

and

$$\hat{P}_2 = \frac{1}{n}\sum_{j=1}^{n}\delta_{y_j},$$

respectively, and k is estimated by n/m. Similar to the proof of Lemma 2 in the Appendix A and the standard methods in nonparametric density estimation, it can

be shown that,

$$\hat{f}(\theta_0) = \frac{1}{ma}\sum_{i=1}^{m} K_1((\theta_{m,n} - x_{I,i})/a)$$

and

$$\hat{g}(\theta_0) = \frac{1}{nb}\sum_{j=1}^{n} K_2((\theta_{m,n} - y_{I,j})/b)$$

are consistent estimators of $f(\theta_0)$ and $g(\theta_0)$, respectively. In this case, we can get a consistent estimator $\hat{a}_0(\Delta)$ of $a_0(\Delta)$.

Let t_α satisfy that $P(\chi_1^2 < t_\alpha) = 1 - \alpha$. Thus, it follows from Theorem 1 that the empirical likelihood based confidence interval for Δ can be constructed as

$$\{\Delta : -2\hat{a}_0^{-1}(\Delta)\mathcal{R}(\Delta, \theta_{m,n}) \leq t_\alpha\},$$

where the asymptotically correct coverage probability is $1 - \alpha$.

We also notice that the result can apply to the complete data settings. In the complete data situation, $P_1 = P_2 = 1$. Thus we can see that the asymptotic distribution of the EL statistic is found to be a χ_1^2 distribution. The empirical likelihood based confidence interval for Δ for the complete data is constructed as

$$\{\Delta : -2R(\Delta, \theta_{m,n}) \leq t_\alpha\}.$$

Chapter 3

NUMERICAL STUDIES

3.1 Monte Carlo Simulation

Based on the results in the inference procedure, extensive simulation studies are conducted to explore the performance of the empirical likelihood confidence intervals for the ROC curve Δ, with different response rates and sample sizes.

In the simulation studies, the diseased population X is distributed as normal distribution with mean 1 and variance 1, while the non-diseased population Y follows the standard normal distribution. Random samples x and y are independently drawn from the population X and Y. The response rates for x and y are chosen as, $(p_1, p_2) = (0.7, 0.6)$, $(0.8, 0.7)$, $(0.9, 0.8)$, combined with the sample sizes for x and y of $(m, n) = (50, 50)$, $(75, 75)$, $(100, 100)$, $(200, 150)$. For a certain response rate and sample size, 1000 independent random samples of data $\{(x_i, \delta_{x_i}), i = 1, ..., m; (y_j, \delta_{y_j}), j = 1, ..., n\}$ are generated. Without loss of generality, the proposed empirical likelihood confidence intervals are constructed for the ROC curve at $q = 0.1$, 0.3, 0.5, $and\, 0.7$. The nominal level of the confidence intervals is $1 - \alpha = 95\%$.

Tables 1 to 4 reveal the following results:

1. For each response rate and sample size, the coverage probability is close to the nominal level 95%, and the average lengths of the confidence intervals are short.

2. In almost all the scenarios, as the respondence rates or the sample sizes increase,

the coverage probabilities get closer to 95%, and the average lengths of the intervals decreases respectively. This is reasonable since either bigger response rates or bigger sample sizes provide more information of the data under study.

Table 3.1. Empirical likelihood confidence intervals for the ROC curve at $q = 0.1\,(\Delta = 0.3891)$.

(p_1, p_2)	(c_1, c_2)	(m, n)	CP(%)	LE	RE	AL
(0.7, 0.6)	(1.3, 1.3)	(50, 50)	95.6	0.2205	0.6042	0.3837
		(75, 75)	94.7	0.2442	0.5724	0.3281
		(100, 100)	95.3	0.2570	0.5469	0.2899
		(200, 150)	95.3	0.2832	0.5075	0.2242
(0.8, 0.7)	(1.5, 1.5)	(50, 50)	93.5	0.1214	0.5276	0.4062
		(75, 75)	94.7	0.1335	0.5071	0.3736
		(100, 100)	94.5	0.1503	0.4999	0.3496
		(200, 150)	95.5	0.1701	0.4761	0.3060
(0.9, 0.8)	(1.2, 1.2)	(50, 50)	93.4	0.1274	0.5320	0.4046
		(75, 75)	94.7	0.1429	0.5097	0.3668
		(100, 100)	94.2	0.1584	0.4991	0.3407
		(200, 150)	94.9	0.1756	0.4735	0.2980

NOTE:
CP(%): coverage probability,
LE: the average left endpoint,
RE: the average right endpoint
AL: the average length of the interval.

Table 3.2. Empirical likelihood confidence intervals for the ROC curve at $q = 0.3\,(\Delta = 0.6828)$.

(p_1, p_2)	(c_1, c_2)	(m, n)	CP(%)	LE	RE	AL
(0.7, 0.6)	(1.3, 1.3)	(50, 50)	94.8	0.3031	0.8237	0.5206
		(75, 75)	96.2	0.3295	0.8055	0.4760
		(100, 100)	94.8	0.3374	0.8133	0.4758
		(200, 150)	94.8	0.3412	0.7897	0.4484
(0.8, 0.7)	(1.5, 1.5)	(50, 50)	94.4	0.3207	0.8092	0.4885
		(75, 75)	96.2	0.3359	0.8002	0.4643
		(100, 100)	94.8	0.3409	0.7960	0.4552
		(200, 150)	95.9	0.3414	0.7857	0.4443
(0.9, 0.8)	(1.2, 1.2)	(50, 50)	95.7	0.3271	0.8133	0.4862
		(75, 75)	95.0	0.3384	0.7986	0.4603
		(100, 100)	95.4	0.3409	0.7910	0.4501
		(200, 150)	94.6	0.3414	0.7836	0.4422

NOTE:
CP(%): coverage probability,
LE: the average left endpoint,
RE: the average right endpoint
AL: the average length of the interval.

Table 3.3. Empirical likelihood confidence intervals for the ROC curve at $q = 0.5\,(\Delta = 0.8413)$.

(p_1, p_2)	(c_1, c_2)	(m, n)	CP(%)	LE	RE	AL
(0.7, 0.6)	(1.3, 1.3)	(50, 50)	95.5	0.4167	0.9285	0.5118
		(75, 75)	95.2	0.4202	0.9166	0.4964
		(100, 100)	94.4	0.4207	0.9110	0.4903
		(200, 150)	95.0	0.4207	0.8986	0.4780
(0.8, 0.7)	(1.5, 1.5)	(50, 50)	94.5	0.4204	0.9213	0.5009
		(75, 75)	94.5	0.4207	0.9110	0.4903
		(100, 100)	95.3	0.4207	0.9037	0.4831
		(200, 150)	96.4	0.4207	0.8917	0.4710
(0.9, 0.8)	(1.2, 1.2)	(50, 50)	95.8	0.4205	0.9186	0.4981
		(75, 75)	94.1	0.4207	0.9092	0.4885
		(100, 100)	95.0	0.4207	0.9043	0.4836
		(200, 150)	95.1	0.4207	0.8933	0.4726

NOTE:
CP(%): coverage probability,
LE: the average left endpoint,
RE: the average right endpoint
AL: the average length of the interval.

Table 3.4. Empirical likelihood confidence intervals for the ROC curve at $q = 0.7\,(\Delta = 0.9363)$.

(p_1, p_2)	(c_1, c_2)	(m, n)	CP(%)	LE	RE	AL
(0.7, 0.6)	(1.3, 1.3)	(50, 50)	93.9	0.4679	0.9758	0.5079
		(75, 75)	95.2	0.4681	0.9697	0.5016
		(100, 100)	95.8	0.4681	0.9669	0.4988
		(200, 150)	94.5	0.4681	0.9627	0.4945
(0.8, 0.7)	(1.5, 1.5)	(50, 50)	93.3	0.4681	0.9703	0.5021
		(75, 75)	94.6	0.4681	0.9652	0.4971
		(100, 100)	96.0	0.4681	0.9628	0.4946
		(200, 150)	93.7	0.4681	0.9615	0.4933
(0.9, 0.8)	(1.2, 1.2)	(50, 50)	93.7	0.4681	0.9702	0.5021
		(75, 75)	95.0	0.4681	0.9654	0.4973
		(100, 100)	94.3	0.4681	0.9632	0.4951
		(200, 150)	94.9	0.4681	0.9611	0.4930

NOTE:
CP(%): coverage probability,
LE: the average left endpoint,
RE: the average right endpoint
AL: the average length of the interval.

Chapter 4

SUMMARY AND FUTURE WORK

4.1 Summary

In this thesis, a smoothed empirical likelihood method is proposed to construct the confidence intervals for ROC curves with missing data in both populations.

First, random hot deck imputation is applied to deal with data missing completely at random (MCAR). Then the empirical likelihood ratio statistic under imputation can be proved to converge to a weighted chi-square distribution asymptotically. Then the simulation studies evaluate the finite sample numerical performance of the inference. All coverage probabilities are close to the nominal level of 95%, and larger sample sizes lead to more accurate coverage probabilities, that is, closer to 95%, and smaller average length of the confidence intervals as well. Moreover, when the response rates are $P_1 = P_2 = 1$, which means both populations are complete, the asymptotic limit distribution is reduced to χ^2_1 distribution.

4.2 Future Work

In the future, we can continue the study in more than one way.

First, real data sets can be applied to testify the performance of the proposed method. Second, to obtain a more efficient confidence interval, we can try the boot-

strap method to explore better bandwidths for the kernel functions. Third, other than the hot deck imputation used in the this thesis, we can try other imputation methods, in order to utilize more information in the data.

In summary, the comparison of ROC curves can be further investigated in many different aspects.

REFERENCES

[1] Chen, J. and Rao, J. N. K., Asymptotic normality under two-phase sampling designs, *Statistica Sinica*, Vol. 17, pp. 1047-1064, 2007.

[2] Chen, S. X., On the accuracy of empirical likelihood confidence regions for linear regression model, *Annals of the Institute of Statistical Mathematics*, Vol. 45, pp. 621-637, 1993.

[3] Chen, S. X., Empirical likelihood confidence intervals for linear regression coefficients, *Journal of Multivariate Analysis*, Vol. 49, pp. 24-40, 1994.

[4] Chen, S. X. and Hall, P., Smoothed empirical likelihood confidence intervals for quantiles, *The Annals of Statistics*, Vol. 21, pp 1166-1181, 1993.

[5] Chen, S. X. and Qin, J., Empirical likelihood-based confidence intervals for data with possible zero observations, *Statistics & Probability Letters*, Vol. 65, pp. 29-37, 2003.

[6] Claeskens, G., Jing, B.-Y., Peng, L. and Zhou, W., Empirical likelihood confidence regions for comparison distributions and ROC curves, *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, Vol. 31, pp. 173-190, 2003.

[7] DiCiccio, T., Hall, P. and Romano, J., Empirical likelihood is Bartlettcorrectable, *The Annals of Statistics*, Vol. 19, pp. 1053-1061, 1991.

[8] Gastwirth, J. L. and Wang, J.-L., Control percentile test procedures for censored data, *Journal of Statistical Planning and Inference*, Vol. 18, pp. 267-276, 1988.

[9] Goddard, M. J. and Hinberg, I., Receiver operator characteristic (ROC) curves and non-normal data: An empirical study, *Statistics in Medicine*, Vol. 9, pp. 325-337, 1990.

[10] Hall, P. and La Scala, B., Methodology and algorithms of empirical likelihood, *International Statistical Review*, Vol. 58, pp. 109-127, 1990.

[11] Hartley, H. O. and Rao, J. N. K., A new estimation theory for sample surveys, *Biometrika*, Vol. 55, pp. 547-557, 1968.

[12] Hsieh, F. and Turnbull, B. W., Nonparametric and semiparametric estimation of the receiver operating characteristic curve, *The Annals of Statistics*, Vol. 24, pp. 25-40, 1996.

[13] Li, G., Tiwari, R. C. and Wells, M. T., Quantile comparison functions in twosample problems with application to comparisons of diagnostic markers, *Journal of the American Statistical Association*, Vol. 91, pp. 689-698, 1996.

[14] Liang, H. and Zhou, Y., Semiparametirc inference for ROC curves with censoring, *Scandinavian Journal of Statistics*, Vol. 35, pp. 212-227, 2008.

[15] Little, R. J. A. and Rubin, D. B., *Statistical Analysis with Missing Data*, Wiley & JohnSons, 2nd edn.

[16] Lloyd, C. J. , Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems, *Journal of the American Statistical Association*, Vol. 93, pp. 1356-1364, 1998.

[17] Owen, A.B., Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, Vol. 75, pp. 237-249, 1988.

[18] Owen, A.B., Empirical likelihood ratio confidence regions. The Annals of Statistics, *Biometrika*, Vol. 18, pp. 90-120, 1990.

[19] Owen, A.B., confidence regions, *Empirical likelihood*, Chapman & Hall Ltd, 2001.

[20] Pepe, M. S., The Statistical Evaluation of Medical Tests for Classification and Prediction, *Oxford: Oxford University Press.*

[21] Qin, J., Empirical likelihood ratio based confidence intervals for mixture proportions, *The Annals of Statistics*, Vol. 27, pp. 1368-1384, 1999.

[22] Qin, J. and Lawless, J., Empirical likelihood and general estimating equations, *The Annals of Statistics*, Vol. 22, pp. 300-325, 1994.

[23] Shapiro, D. E., The interpretation of diagnostic tests, *Statistical Methods in Medical Research*, Vol. 8, pp. 113-134, 1999.

[24] Su, H., Qin, Y. and Liang, H., Empirical Likelihood-Based Confidence Interval of ROC Curves, *Statistics in Biopharmaceutical Research*, Vol. 1, pp. 407-414, 2009.

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

[26] Tosteson, A. A. N. and Begg, C. B., A general regression methodology for roc curve estimation, *Medical Decision Making*, Vol. 8, pp. 204-215, 1988.

[27] Wang, Q. and Rao, J. N. K., Empirical likelihood-based inference under for missing response data, *The Annals of Statistics*, Vol. 30, pp. 896-924, 2002.

[28] Zhou, W. and Jing, B.-Y., Smoothed empirical likelihood confidence intervals for the difference of quantiles, *Statistica Sinica*, Vol. 13, pp. 83-95, 2003.

[29] Zhou, X.-H., McClish, D. K. and Obuchowski, N. A., *Statistical Methods in Diagnostic Medicine*, Wiley.

[30] Zhou, Y. and Liang, H., Empirical-likelihood-based semiparametric inference for the treatment effect in the two-sample problem with censoring, *Biometrika*, Vol. 92, pp. 271-282, 2005.

[31] Zou, K. H., Hall, W. J. and Shapiro, D. E. , Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests, *Statistics in Medicine*, Vol. 16, pp. 2143-2156, 1997.

[32] Zweig, M. and Campbell, G., Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, Vol. 39, pp. 561-577, 1993.

APPENDICES

Appendix A: Lemmas and Proofs

The following lemma of Chen and Rao (2007) will be used later.

Lemma 1. *Let U_n, V_n be two sequences of random variables and let $\mathcal{B}_n$ be a σ-algebra. Assume that*

1. *There exists $\sigma_{1n} > 0$ such that*

$$\sigma_{1n}^{-1} V_n \xrightarrow{d} N(0,1),$$

as $n \to \infty$, where V_n is $\mathcal{B}_n$ measurable.

2. *$E[U_n|\mathcal{B}_n] = 0$ and $Var(U_n|\mathcal{B}_n) = \sigma_{2n}^2$ such that*

$$\sup_t |P(\sigma_{2n}^{-1} U_n \leq t|\mathcal{B}_n) - \Phi(t)| = o_p(1),$$

where $\Phi(.)$ is the distribution function of the standard normal random variable.

3. *$\gamma_n^2 = \sigma_{1n}^2/\sigma_{2n}^2 = \gamma^2 + o_p(1)$*

Then, as $n \to \infty$,

$$\frac{U_n + V_n}{\sqrt{\sigma_{1n}^2 + \sigma_{2n}^2}} \xrightarrow{d} N(0,1).$$

To prove the main results, we need some additional lemmas.

Lemma 2. *Under the conditions of Theorem 1, as $m, n \to \infty$, we have*

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \omega_1(x_{I,i}, \theta_0, \Delta) \xrightarrow{d} N(0, \sigma_1^2),$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_2(x_{I,i}, \theta_0, \Delta) \xrightarrow{d} N(0, \sigma_2^2),$$

and

$$\frac{1}{m} \sum_{i=1}^{m} \omega_1^2(x_{I,i}, \theta_0, \Delta) = \Delta(1-\Delta) + o_p(1),$$

$$\frac{1}{n} \sum_{i=1}^{n} \omega_2^2(y_{I,j}, \theta_0, \Delta) = q(1-q) + o_p(1),$$

where

$$\sigma_1^2 = (1 - P_1 + P_1^{-1})\Delta(1-\Delta),\ \sigma_2^2 = (1 - P_2 + P_2^{-1})q(1-q).$$

Proof of Lemma 2. Let $\overline{\omega}_{1r} = \dfrac{1}{r_x} \sum_{i \in S_{rx}} \omega_1(x_i, \theta_0, \Delta)$ and $\mathcal{B}_m = \sigma((\delta_{xi}, x_i), i = 1, ..., m)$. Then

$$E(\omega_1(x_i^*, \theta_0, \Delta) | \mathcal{B}_m) = \overline{\omega}_{1r},$$

$$Var(\omega_1(x_i^*, \theta_0, \Delta) | \mathcal{B}_m) = \frac{1}{r_x} \sum_{i \in S_{rx}} \{\omega_1(x_i, \theta_0, \Delta) - \overline{\omega}_{1r}\}^2.$$

It follows that

$$\begin{aligned} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \omega_1(x_{I,i}, \theta_0, \Delta) &= \sqrt{m}\overline{\omega_{1r}} + \frac{1}{\sqrt{m}} \sum_{i \in S_{mx}} \{\omega_1(x_i^*, \theta_0, \Delta) - E(\omega_1(x_i^*, \theta_0, \Delta) | \mathcal{B}_m)\} \\ &= V_m + U_m. \end{aligned} \tag{1}$$

V_m is $\mathcal{B}_m$ measurable and

$$V_m = \sqrt{m}\frac{1}{r_x}\sum_{i\in S_{rx}}\{\omega_1(x_i,\theta_0,\Delta) - E\omega_1(x_i,\theta_0,\Delta)\} + \sqrt{m}E\omega_1(x_i,\theta_0,\Delta).$$

It can be shown that $E\omega_1(x_i,\theta_0,\Delta) = O(a^{t_0})$. Thus from Assumption (iii) and (v), it follows that $\sqrt{m}E\omega_1(x_i,\theta_0,\Delta) = o(1)$. Combining with the MCAR assumption and the Central Limit Theorem (CLT), it gives,

$$V_m \xrightarrow{d} N(0, P_1^{-1}\Delta(1-\Delta)).$$

From Berry-Esseen's Central Limit Theorem for independent random variables, we have

$$\sup_t |P(\sigma_{2m}^{-1}U_m \le t|\mathcal{B}_m) - \Phi(t)| = o_p(1),$$

where $\sigma_{2m}^2 = (1-P_1)E\omega_1(x,\theta_0,\Delta) = (1-P_1)\Delta(1-\Delta)$. Hence, from Lemma 1, we have

$$\frac{1}{\sqrt{m}}\sum_i \omega_1(x_{I,i},\theta_0,\Delta) \xrightarrow{d} N(0,\sigma_1^2).$$

On the other hand, denote the conditional probability given $\mathcal{B}_m$ as P^*. Then by the Law of Large Numbers and MCAR assumption,

$$\frac{1}{m_x}\sum_{i\in s_{mx}}\omega_1^2(x_i^*,\theta_0,\Delta) = \frac{1}{r_x}\sum_{i\in s_{rx}}\omega_1^2(x_i,\theta_0,\Delta) + o_{P^*}(1) = E\omega_1^2(x,\theta_0,\Delta) + o_p(1).$$

It follows that

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\omega_1^2(x_{I,i},\theta_0,\Delta) &= \frac{1}{m}\sum_{i=1}^{m}\{\delta_{xi}\omega_1^2(x_i,\theta_0,\Delta)+(1-\delta_{xi})\omega_1^2(x_i^*,\theta_0,\Delta)\} \\
&= P_1 E\omega_1^2(x,\theta_0,\Delta)+o_p(1)+\frac{m_x}{m}\frac{1}{m_x}\sum_{i\in s_{mx}}\omega_1^2(x_i^*,\theta_0,\Delta) \\
&= \frac{1}{m}\sum_{i=1}^{m}\{\delta_{xi}\omega_1^2(x_i,\theta_0,\Delta)+(1-\delta_{xi})\omega_1^2(x_i^*,\theta_0,\Delta)\} \\
&= P_1 E\omega_1^2(x,\theta_0,\Delta)+o_p(1)+(1-P_1)E\omega_1^2(x,\theta_0,\Delta)+o_p(1) \\
&= E\omega_1^2(x,\theta_0,\Delta)+o_p(1) \\
&= \Delta(1-\Delta)+o_p(1).
\end{aligned}$$

The rest of Lemma 2 can be proved similarly. So the proof of Lemma 2 is complete. □

Lemma 3. *Suppose that* $1/3<\eta<1/2$ *and the conditions of Theorem 1 are satisfied. Then, as* $m,n\to\infty$,

$$\lambda_1(\theta)=O_p(n^{-\eta}a^{-1}+a^{t_0}),$$

and

$$\lambda_2(\theta)=O_p(n^{-\eta}b^{-1}+b^{t_0}),$$

uniformly about $\theta\in\{\theta:|\theta-\theta_0|\le cn^{-\eta}\}$, *where* c *is a positive constant.*

Proof of Lemma 3. It can be shown that

$$|\omega_1(x_{I,i},\theta,\Delta)-\omega_1(x_{I,i},\theta_0,\Delta)|\le ca^{-1}n^{-\eta}$$

for some constant c as $\theta\in\{\theta:|\theta-\theta_0|\le cn^{-\eta}\}$. Combining with Lemma 2 we have

$$\frac{1}{m}\sum_{i=1}^{m}\omega_1(x_{I,i},\theta,\Delta)=O_p(n^{-\eta}a^{-1}+a^{t_0}),$$

$$\frac{1}{m}\sum_{i=1}^{m}\omega_1^2(x_{I,i},\theta,\Delta)=\Delta(1-\Delta)+o_p(1).$$

Denote $Z_m=\max\limits_{1\le i\le m}|\omega_1(x_{I,i},\theta,\Delta)|$. Then $Z_m\le c, a.s.$ Thus equation (2.7) gives that

$$\frac{|\lambda_1(\theta)|}{1+Z_m|\lambda_1(\theta)|}\{\Delta(1-\Delta)+O_P(1)\}=O_p(n^{-\eta}a^{-1}+a^{t_0}).$$

Therefore $\lambda_1(\theta)=O_p(n^{-\eta}a^{-1}+a^{t_0})$. The rest of Lemma 2 can be proved similarly. □

Lemma 4. *Suppose that* $1/3<\eta<1/2$ *and the conditions of Theorem 1 are satisfied. Then with probability tending to 1 there exists a root* $\theta_{m,n}$ *of equation* (2.9) *such that, as* $m,\ n\to\infty$,

$$|\theta_{m,n}-\theta_0|=O_p(n^{-\eta}),$$

and $R(\Delta,\theta)$ *attains its local maximum value at* $\theta_{m,n}$.

Proof of Lemma 4. Take $|\theta-\theta_0|=n^{-\eta}$. Denote $\overline{\omega}_{1j}(\theta)=\dfrac{1}{m}\sum\limits_{i=1}^{m}\omega_1^j(x_{I,i},\theta,\Delta)$ for $j=1,\ 2$ and

$$R_1(\Delta,\theta)=-\sum_{i=1}^{m}\log\{1+\lambda_1(\theta)\omega_1(x_{I,i},\theta,\Delta)\},$$

$$R_2(\Delta,\theta)=-\sum_{j=1}^{n}\log\{1+\lambda_2(\theta)\omega_2(y_{I,j},\theta,\Delta)\}.$$

From equation (2.7), we obtain that

$$\overline{\omega}_{11}(\theta)-\lambda_1(\theta)\overline{\omega}_{12}(\theta)+\frac{1}{m}\lambda_1^2(\theta)\sum_{i=1}^{m}\frac{\omega_1^3(x_{I,i},\theta,\Delta)}{1+\lambda_1\omega_1(x_{I,i},\theta,\Delta)}=0.$$

Using Lemma 3, we have

$$\lambda_1(\theta)=\overline{\omega}_{11}(\theta)\{\overline{\omega}_{12}(\theta)\}^{-1}+O_p\{\lambda_1^2(\theta)\}.$$

By the Taylor expansion we have

$$\begin{aligned}
-R_1(\Delta,\theta) &= \sum_{i=1}^{m}\lambda_1(\theta)\omega_1(x_{I,i},\theta,\Delta) - \frac{1}{2}\sum_{i=1}^{m}\lambda_1^2(\theta)\omega_1^2(x_{I,i},\theta,\Delta) + O_p\{m\lambda_1^3(\theta)\} \\
&= m\lambda_1(\theta)\overline{\omega}_{11}(\theta) - \frac{1}{2}m\lambda_1^2(\theta)\overline{\omega}_{12}(\theta) + O_p\{m\lambda_1^3(\theta)\} \\
&= \frac{m}{2}\{\overline{\omega}_{12}(\theta)\}^{-1}\{\overline{\omega}_{11}(\theta)\}^2 + O_p\{m\lambda_1^3(\theta)\} \\
&= \frac{m}{2}\{\overline{\omega}_{12}(\theta_0) + 0_p(1)\}^{-1}\{\overline{\omega}_{11}(\theta_0) + \overline{\gamma}_{11}(\theta_0)n^{-\eta} + O_p(n^{-2\eta})\}^2 + O_p\{m\lambda_1^3(\theta)\} \\
&= \frac{m}{2}\{q(1-q) + 0_p(1)\}^{-1}\{\overline{\omega}_{11}(\theta_0) + f(\theta_0)n^{-\eta} + o_p(n^{-\eta})\}^2 + o_p(mm^{-2\eta}),
\end{aligned}$$

where $\overline{\gamma}_{11}(\theta_0) = \frac{1}{ma}\sum_i K((\theta_0 - x_{I,i})/a)$. From Assumptions (iii), (v), Lemma 3 and its proof, it follows that $\overline{\omega}_{11}(\theta_0) = O_p(a^{t_0} + m^{-1/2}) = o_p(m^{-\eta})$. Thus,

$$-R_1(\Delta,\theta) = \frac{m}{2}\{q(1-q)\}^{-1} + f^2(\theta_0)n^{-2\eta} + o_p(mn^{-2\eta}).$$

On the other hand, from the above derivations, we can see that

$$-R_1(\Delta,\theta) = o_p(mn^{-2\eta}).$$

It follows that, when $|\theta - \theta_0| = n^{-\eta}$, with probability tending to 1,

$$-R_1(\Delta,\theta) > -R_1(\Delta,\theta_0).$$

Similarly,

$$-R_2(\Delta,\theta) > -R_2(\Delta,\theta_0).$$

Thus,

$$R(\Delta,\theta) < R(\Delta,\theta_0).$$

From the continuity of $R(\Delta, \theta)$, we have Lemma 4. □

Lemma 5. *Suppose that the conditions of Theorem 1 are satisfied and that $\theta_{m,n}$ is as in Lemma 4. Then, as $m, n \to \infty$,*

$$\begin{aligned}
&\sqrt{m}(\theta_{m,n} - \theta_0) \xrightarrow{d} N(0, (f^2(\theta_0)\sigma_1^2 + kg^2(\theta_0 + \Delta)\sigma_2^2)/c_0^2), \\
&\lambda_1(\theta_{m,n}) = -\frac{kg(\theta_0 + \Delta)}{f(\theta_0)}\lambda_2 9\theta_{m,n}) + o_p(n^{-1/2}), \\
&\sqrt{m}\lambda_2(\theta_{m,n}) \xrightarrow{d} N(0, \sigma^2),
\end{aligned}$$

where

$$\sigma^2 = \{q(1-q)\}^{-1} f^2(\theta_0) \frac{(1 - P_1 + P_1^{-1})g^2(\theta_0 + \Delta) + k^{-1}(1 - P_2 + P_2^{-1})f^2(\theta_0)}{c_0^2},$$

and σ_j^2, $j = 1, 2$ and c_0 are defined as in Lemma 2 and Theorem 1.

Proof of Lemma 5. Let $\lambda_1 = \lambda_1(\theta),\ \lambda_{E1} = \lambda_1(\theta_{m,n}),\ \lambda_2 = \lambda_2(\theta),\ \lambda_{E2} = \lambda_2(\theta_{m,n})$ and

$$Q_{1,m,n}(\theta, \lambda_1, \lambda_2) = \frac{1}{m}\sum_{i=1}^{m} \frac{\omega_1(x_{I,i}, \theta, \Delta)}{1 + \lambda_1\omega_1(x_{I,i}, \theta, \Delta)},$$

$$Q_{2,m,n}(\theta, \lambda_1, \lambda_2) = \frac{1}{n}\sum_{j=1}^{n} \frac{\omega_2(y_{I,j}, \theta, \Delta)}{1 + \lambda_2\omega_2(y_{I,j}, \theta, \Delta)},$$

$$Q_{3,m,n}(\theta, \lambda_1, \lambda_2) = \frac{\lambda_1}{ma}\sum_{i=1}^{m} \frac{K_1((\theta - x_{I,i})/a)}{1 + \lambda_1\omega_1(x_{I,i}, \theta, \Delta)} + \frac{\lambda_2}{mb}\sum_{j=1}^{n} \frac{K_2((\theta - y_{I,j})/b)}{1 + \lambda_2\omega_2(y_{I,j}, \theta, \Delta)}.$$

From Lemma 4, we have

$$Q_{i,m,n}(\theta_{m,n}, \lambda_{E1}, \lambda_{E2}) = 0,\ i = 1, 2, 3.$$

From the Taylor expansion and Lemma 3 and Lemma 4, we have

$$\begin{aligned}
0 &= Q_{i,m,n}(\theta_{m,n}, \lambda_{E1}, \lambda_{E2}) \\
&= Q_{i,m,n}(\theta_0, 0, 0) + \frac{\partial Q_{i,m,n}(\theta_0, 0, 0)}{\partial \theta}(\theta_{m,n} - \theta_0) \\
&+ \frac{\partial Q_{i,m,n}(\theta_0, 0, 0)}{\partial \lambda_1}\lambda_{E1} + \frac{\partial Q_{i,m,n}(\theta_0, 0, 0)}{\partial \lambda_2}\lambda_{E2} + o_p(\epsilon_n), \quad i = 1, 2, 3,
\end{aligned}$$

where $\epsilon_n = |\theta_{m,n} - \theta_0| + |\lambda_{E1}| + |\lambda_{E2}|$. Hence

$$\begin{aligned}
&Q_{i,m,n}(\theta_0, 0, 0) + \frac{\partial Q_{i,m,n}(\theta_0, 0, 0)}{\partial \theta}(\theta_{m,n} - \theta_0) \\
&+ \frac{\partial Q_{i,m,n}(\theta_0, 0, 0)}{\partial \lambda_1}\lambda_{E1} + \frac{\partial Q_{i,m,n}(\theta_0, 0, 0)}{\partial \lambda_2}\lambda_{E2} = o_p(\epsilon_n), \quad i = 1, 2, 3.
\end{aligned} \tag{2}$$

Similar to the proof of Lemma 2, it can be shown that

$$\begin{aligned}
\frac{\partial Q_{1,m,n}(\theta_0, 0, 0)}{\partial \theta} &= f(\theta_0) + o_p(1), \\
\frac{\partial Q_{1,m,n}(\theta_0, 0, 0)}{\partial \lambda_1} &= -\Delta(1 - \Delta) + o_p(1), \\
\frac{\partial Q_{1,m,n}(\theta_0, 0, 0)}{\partial \lambda_2} &= 0, \\
\frac{\partial Q_{2,m,n}(\theta_0, 0, 0)}{\partial \theta} &= g(\theta_0) + o_p(1), \\
\frac{\partial Q_{2,m,n}(\theta_0, 0, 0)}{\partial \lambda_1} &= 0, \\
\frac{\partial Q_{2,m,n}(\theta_0, 0, 0)}{\partial \lambda_2} &= -q(1 - q) + o_p(1), \\
\frac{\partial Q_{3,m,n}(\theta_0, 0, 0)}{\partial \theta} &= 0, \\
\frac{\partial Q_{3,m,n}(\theta_0, 0, 0)}{\partial \lambda_1} &= f(\theta_0) + o_p(1), \\
\frac{\partial Q_{3,m,n}(\theta_0, 0, 0)}{\partial \lambda_2} &= kg(\theta_0) + o_p(1).
\end{aligned}$$

Thus

$$\begin{pmatrix} \theta_{m,n} - \theta_0 \\ \lambda_{E1} \\ \lambda_{E2} \end{pmatrix} = S^{-1} \begin{pmatrix} -Q_{1,m,n}(\theta_0, 0, 0) \\ -Q_{2,m,n}(\theta_0, 0, 0) \\ 0 \end{pmatrix} + o_p(\epsilon_n),$$

where

$$S = \begin{pmatrix} f(\theta_0) & -\Delta(1-\Delta) & 0 \\ g(\theta_0) & 0 & -q(1-q) \\ 0 & f(\theta_0) & kg(\theta_0) \end{pmatrix}.$$

Combining with $\sqrt{n}Q_{j,m,n}(\theta_0, 0, 0) = O_p(1), j = 1, 2$, we have $\epsilon_n = O_p(n^{-1/2})$. It follows that

$$\theta_{m,n} - \theta_0 = -\frac{1}{c_0}\{q(1-q)f(\theta_0)Q_{1,m,n}(\theta_0, 0, 0) + k\Delta(1-\Delta)g(\theta_0)Q_{2,m,n}(\theta_0, 0, 0)\} + o_p(n^{-1/2}),$$

$$\lambda_{E1} = \frac{kg(\theta_0)}{c_0}\{g(\theta_0)Q_{1,m,n}(\theta_0, 0, 0) - f(\theta_0)Q_{2,m,n}(\theta_0, 0, 0)\} + o_p(n^{-1/2}),$$

$$\lambda_{E2} = -\frac{f(\theta_0)}{c_0}\{g(\theta_0)Q_{1,m,n}(\theta_0, 0, 0) - f(\theta_0)Q_{2,m,n}(\theta_0, 0, 0)\} + o_p(n^{-1/2}).$$

From Lemma 2, we have

$$\sqrt{m}\begin{pmatrix} Q_{1,m,n}(\theta_0, 0, 0) \\ Q_{2,m,n}(\theta_0, 0, 0) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & k^{-1}\sigma_2^2 \end{pmatrix}\right).$$

Thus Lemma 5 is proved. □

Proof of Theorem 1. Similar to the proof of Theorem 1 in Owen (1990), it can be

shown that

$$\begin{aligned}&-2R(\Delta,\theta_{m,n})\\=&m\lambda_1^2(\theta_{m,n})\times\frac{1}{m}\sum_{i=1}^{m}\omega_1^2(x_{I,i},\theta,\Delta)+n\lambda_2^2(\theta_{m,n})\times\frac{1}{n}\sum_{j=1}^{n}\omega_2^2(y_{I,j},\theta,\Delta)+o_p(1).\end{aligned}$$

Combining with Lemma 2 and Lemma 5, Theorem 1 is proved. □