Applications of Empirical Likelihood to Quantile Estimation and Longitudinal Data

by

Jien Chen

(Under the direction of Nicole A. Lazar)

Abstract

As a non-parametric method, Empirical Likelihood (EL) has been attracting serious attention from researchers in statistics, econometrics, engineering and biostatistics. By defining the estimation equations in EL appropriately, we can extend EL to various data settings, especially those in which parametric likelihoods are absent. In this dissertation, two applications of empirical likelihood are explored: quantile estimation and longitudinal data analysis. Quantile estimation for discrete data analysis has not been well studied. For a given $0 < p < 1$, the commonly used sample quantile may or may not be consistent for the pth quantile, depending on whether or not the underlying distribution has a plateau at the level of p. I propose an EL-based categorization procedure which not only helps determine the shape of the true distribution at level p, but also provides a way of formulating a new estimator that is consistent in any case. For non-Gaussian longitudinal data, generalized estimating equations (GEE) are a popular class of marginal models. While the GEE estimator is consistent and robust, it may suffer significant loss of efficiency if the working correlation structure is misspecified. I consider the use of EL to select working correlations for GEE models, for which parametric likelihoods are absent and quasi-likelihoods are difficult to construct.

INDEX WORDS: Empirical likelihood, Quantile estimation, Discrete distributions, Jittering, Bootstrap, Longitudinal data, Generalized estimating equations, Working correlation structure, Model selection, Information criteria

Applications of Empirical Likelihood to Quantile Estimation and Longitudinal Data

by

Jien Chen

B.S., University of Science and Technology of China, China, 1999

M.S., University of Science and Technology of China, China, 2002

M.S., The University of Georgia, U.S., 2004

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2008

Applications of Empirical Likelihood to Quantile Estimation and Longitudinal Data

by

Jien Chen

Approved:

Major Professor: Nicole A. Lazar

Committee: Gauri Datta
Daniel Hall
Lynne Seymour
Yehua Li

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2008

Dedication

To my mother, Qin and Eric

Acknowledgments

I sincerely thank my advisor Dr. Nicole Lazar for her insightful direction, constant encouragement and kind support throughout the entire duration of my dissertation research. She has guided me to an interesting research area, provided me tremendous help along the way, and constantly reassured me of my progress. Without her encouragement and support, I would never benefit from participating in the application for a National Science Foundation grant; this experience has turned out to be a successful and rewarding endeavor. I am especially grateful for her patience in helping me improve my writing skills. She has given me support that goes beyond the advisement needed for the completion of my dissertation. It has been a great pleasure to work with her, a mentor who truly cares for her students.

I am also thankful to my committee members, Dr. Gauri Datta, Dr. Daniel Hall, Dr. Lynne Seymour and Dr. Yehua Li, for their assistance with my dissertation and other help they have offered to me in the past five years. Specifically, I thank Dr. Gauri Datta for providing an interesting research question that I can continue to work on in the future, Dr. Daniel Hall for his constructive suggestions on my second research topic, Dr. Lynne Seymour for admitting me to this PhD program (so that I got the opportunity to become a statistician) and her advisement as the graduate coordinator, and Dr. Yehua Li for sharing his job-hunting experience.

My special thanks goes to Dr. Jaxk Reeves, who has given me invaluable opportunities to work on interesting statistical consulting projects and has contributed significantly to my career development.

Sincere appreciation is extended to the faculty, staff and students of the Department of Statistics, who have helped me in different ways and made my past five years so enjoyable.

I express my deepest gratitude to my mother for her unconditional love and long-term support. I am also deeply indebt to my husband, Qin Wang. Without his encouragement, I would not be able to pursue a PhD degree in the US. It is his love and support that has helped me overcome various difficulties.

Part of this work was funded by National Science Foundation Grant DMS-070192 (Dr. Nicole Lazar, Principal Investigator). Finally, I am so grateful to the UGA Graduate School for awarding me the Dissertation Completion Award that allowed me more time in the final year to concentrate on my research.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction to Empirical Likelihood

Empirical likelihood (EL), introduced by Owen (1988, 1990, 1991), is a nonparametric analog of the classical likelihood, and has been attracting serious attention from practitioners and researchers in statistics, econometrics, engineering and biostatistics. As a nonparametric method, EL is more robust than parametric likelihoods since it does not require the specification of a family of distributions for the data. On the other hand, EL carries many properties of parametric likelihood: EL determines the shape of confidence regions automatically; it can readily incorporate known constraints on parameters, and extend to biased sampling and censored data; it has favorable asymptotic power properties; it can be Bartlett corrected, providing accurate inferences. As such, this technique has the potential to yield powerful tests for various data settings.

1.1 Empirical Likelihood for the Mean

For a sample $X_1, \ldots, X_n$ from an unknown d-variate distribution F_0 having mean $\mu_0 \in \mathbb{R}^d$ ($d \geq 1$), the *empirical likelihood function* for a distribution F is

$$L(F) = \prod_{i=1}^{n} dF(X_i) = \prod_{i=1}^{n} w_i,$$

where $w_i = Pr(X = X_i)$, the probability mass placed on X_i by F. Note that this likelihood is nonzero only for distributions that put positive probability on each of the observed data points. Without any additional constraint on w_i, $L(F)$ is maximized by the empirical distribution function F_n which puts equal weight $1/n$ on each observation. Then the *empirical*

likelihood ratio for F is

$$\mathcal{R}(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} nw_i.$$

Suppose now one is interested in estimating a parameter, expressed as a functional $\theta(F_0)$; for simplicity, take $\theta(F_0) = \mu_0$. For estimating μ_0, the *profile empirical likelihood ratio* (ELR) for a candidate μ is

$$\mathcal{R}(\mu) = \sup\left\{\prod_{i=1}^{n} nw_i \mid w_i \geq 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i X_i = \mu\right\}.$$

Let $\mathcal{H}_n$ denote the convex hull of the sample points $X_1, \ldots, X_n$. For $\mu \notin \mathcal{H}_n$, the constraint $\sum_{i=1}^{n} w_i X_i = \mu$ cannot be satisfied by any valid set of w_i and $\mathcal{R}(\mu)$ is defined to be zero. For inference purpose, one only needs to consider $\mu \in \mathcal{H}_n$, for which a unique value of $\mathcal{R}(\mu)$ exists: $\prod_{i=1}^{n} nw_i$ is maximized subject to the constraints when $w_i = w_i(\mu) = [n(1 + \lambda^T(X_i - \mu))]^{-1}$, where $\lambda \in \mathbb{R}^d$ is a Lagrange multiplier that solves

$$\sum_{i=1}^{n} \frac{X_i - \mu}{1 + \lambda^T(X_i - \mu)} = 0.$$

Therefore, $\mathcal{R}(\mu)$ can be explicitly expressed as

$$\mathcal{R}(\mu) = \prod_{i=1}^{n} [1 + \lambda^T(X_i - \mu)]^{-1}.$$

Details can be found in Owen (2001). Since $\prod_{i=1}^{n} nw_i$ is maximized unconditionally by F_n, it follows that $\mathcal{R}(\mu)$ is maximized with respect to μ at $\hat{\mu} = \bar{X}$, *i.e.*, the *maximum empirical likelihood estimator* (MELE) of μ_0 coincides with the sample mean when no other condition is imposed.

The *empirical likelihood ratio test statistic* for μ is $-2\log\mathcal{R}(\mu)$. Empirical likelihood confidence regions are of the form $\{\mu \mid -2\log\mathcal{R}(\mu) \leq c_0\}$. Owen (1988, 1990) showed that, under mild conditions, the empirical likelihood test statistic inherits the basic asymptotic property of the parametric likelihood. In particular, if $X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbb{R}^d$ with mean μ_0 and finite variance covariance matrix V_0 of rank $q > 0$, then

$$-2\log\mathcal{R}(\mu_0) \xrightarrow{d} \chi^2_{(q)} \text{ as } n \to \infty, \tag{1.1}$$

where $q = d$ if V_0 has full rank. The coverage error of EL confidence intervals is of order $O(n^{-1})$. This approach also applies to general parameters $\theta(F)$ that are smooth functions of means.

1.2 Empirical Likelihood and Estimating Equations

Qin and Lawless (1994) extended EL to p-dimensional parameters θ defined via general r-dimensional estimating equations

$$E\{g(X, \theta)\} = 0, \tag{1.2}$$

where $g(X, \theta) = (\, g_1(X, \theta), \ldots, g_r(X, \theta)\,)^T$ and $r \geq p$. Many parameters can be formulated in this way. For μ considered above, $g(X, \mu) = X - \mu$; if θ stands for the median of a univariate continuous distribution, then $g(X, \theta) = 1(X \leq \theta) - 0.5$, where $1(X \leq \theta)$ is the indicator function. An estimating equation in the form of (1.2) translates straightforwardly into the constraint

$$\sum_{i=1}^{n} w_i g(X_i, \theta) = 0$$

in the definition of ELR for θ:

$$\mathcal{R}(\theta) = \sup \left\{ \prod_{i=1}^{n} n w_i \mid w_i \geq 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i g(X_i, \theta) = 0 \right\}. \tag{1.3}$$

For any θ, $Z_i(\theta) = g(X_i, \theta)$ are i.i.d. with common mean $\mu = E_{F_0}\{Z(\theta)\} = E_{F_0}\{g(X, \theta)\}$ provided that X_i are i.i.d., where $E_{F_0}(\cdot)$ emphasizes that the expectation is evaluated under the true distribution F_0. Under the assumption that $E_{F_0}\{Z(\theta_0)\} = 0_{r\times 1} := \mu_0$, it follows that $\mathcal{R}(\theta_0) = \mathcal{R}(\mu_0)$ since they are both the maximum of $\prod_{i=1}^{n} n w_i$ subject to the same constraints. Thus, Owen's result on the asymptotic distribution of $-2 \log \mathcal{R}(\mu_0)$ also applies to $-2 \log \mathcal{R}(\theta_0)$.

By allowing $r \geq p$, the framework of Qin and Lawless provides a general and flexible method of combining information about parameters. In many situations, pieces of information about F_0 can be formulated into equations which can be made into components

of $g(X, \theta)$. For example, suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. observations from a bivariate distribution with mean $(\theta_{x0}, \theta_{y0})$, and it is known that $\theta_{x0} = \theta_{y0} = \theta_0$, then the estimating function for a candidate θ can be taken to be

$$g(\,(X, Y),\, \theta\,) = (X - \theta, Y - \theta)^T\,, \tag{1.4}$$

and the MELE for θ_0 is

$$\hat{\theta} = \arg\max_{\theta} \mathcal{R}(\theta),$$

where $\mathcal{R}(\theta)$ is defined with estimating equation (1.4). In this case, while (1.2) holds for the true distribution F_0 with the true parameter θ_0, it will not generally hold for the empirical distribution F_n, since $(1/n)\sum_{i=1}^{n} g(\,(X_i, Y_i), \theta\,) = (0, 0)^T$ typically has no solution when the number of equations ($r = 2$) is more than the number of parameters ($p = 1$). The fact that (1.2) holds is a special feature of F_0 and constitutes important side information. Therefore, with additional information, the MELE of θ_0 in this example is not the same as the sample means $\bar{X}$, $\bar{Y}$ (the two separate sample means), or $(\bar{X} + \bar{Y})/2$ (the pooled sample mean).

Under mild assumptions, including that $E[g(X, \theta_0)g^T(X, \theta_0)]$ is positive definite and that $g(X, \theta)$ is smooth in θ, Qin and Lawless (1994) showed that the MELE $\hat{\theta}$ is consistent and asymptotically normal, i.e.,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V),$$

where

$$V = \left[E\left(\frac{\partial g}{\partial \theta^T}\right)^T (Egg^T)^{-1} E\left(\frac{\partial g}{\partial \theta^T}\right) \right]^{-1}.$$

One interesting property of EL is that it delivers sharper inference when more information is exploited (*i.e.*, more components are added to the estimating equation $g(X, \theta)$), in the sense that the asymptotic variance of the MELE will not increase and will typically become smaller.

Qin and Lawless also showed properties of empirical likelihood ratio statistics. For testing $H_0 : \theta = \theta_0$, the statistic

$$W_E(\theta_0) = -2\log\mathcal{R}(\theta_0) - [-2\log\mathcal{R}(\hat{\theta})]$$

is defined; if H_0 is true, then

$$W_E(\theta_0) \xrightarrow{d} \chi^2_{(p)} \text{ as } n \to \infty. \tag{1.5}$$

When $r = p$, $W_E(\theta_0)$ reduces to $-2\log\mathcal{R}(\theta_0)$, since $\hat{\theta}$ is identical to the solution of $\sum_{i=1}^{n}(1/n)g(X_i,\theta) = 0$ and $-2\log\mathcal{R}(\hat{\theta}) = 0$. Thus, it is seen that Owen's result on $-2\log\mathcal{R}(\mu_0)$ is a special case of (1.5), provided that $Var(X)$ is of full rank. When $r > p$, one may test the model (1.2) by using the ratio statistic

$$W_1 = -2\log\mathcal{R}(\hat{\theta}),$$

which is asymptotically $\chi^2_{(r-p)}$ if (1.2) is correct.

1.3 Applications and Development

Since its introduction, EL has been extended to a variety of contexts, including linear models (Owen, 1991), generalized linear models (Kolaczyk, 1994), density estimation (for example Chen, 1996), biased samples (for example Qin, 1993), survival data (for example Zhou, 2005), and time series (for example Kitamura, 1997). Moreover, DiCiccio *et al.* (1991) demonstrated that higher order properties of parametric likelihood, in particular Bartlett correctability, are also inherited by EL. This means that an empirical correction for scale reduces the order of coverage error from n^{-1} to n^{-2}. And Baggerly (1998) showed that EL is the only member of a rich family of alternative likelihoods to inherit these properties. Lazar (2003) extended the discussion of EL to the Bayesian setting. Other generalizations of EL include weighted empirical likelihood (Wu, 2004) and exponentially tilted EL (Schennach, 2005), which exploit the relationships between EL and other alternative likelihood structures (Efron, 1981). In sum, EL offers an attractive alternative to parametric likelihood analyses, retaining many desirable likelihood properties, but without the need to fully specify a parametric model.

1.4 Scope of this Dissertation

In this dissertation, two different applications of empirical likelihood are investigated: quantile estimation for discrete data, and longitudinal data analysis.

Quantile estimation via empirical likelihood presents an example of an estimating equation not smooth in the parameter, and hence the asymptotic properties of the MELE derived in Qin and Lawless (1994) do not apply. Some researchers have proposed smoothed versions of the estimating equation to improve the coverage of EL confidence intervals for a continuous distribution's quantile. Nevertheless, properties of the MELE of quantiles of a discrete distribution have not received much attention. The objective of the first part of the dissertation is to study the MELE of quantiles of a discrete distribution, and propose a way to improve the MELE.

A longitudinal data set consists of repeated measurements taken over time on a sample of subjects. The correlation of measurements clustered by subjects poses a challenge to the empirical likelihood method, as independence of observations is a required condition in the theory. Although one possible way to apply empirical likelihood to longitudinal data analysis is simply to ignore the within-subject correlation, this may not lead to efficient inference. How to improve the efficiency of the MELE in the context of longitudinal data is the focus of the second part of the dissertation.

Chapter 2

Quantile Estimation

2.1 Introduction

Quantiles, and in particular the quartiles, are important characteristics of a population that contain more information about the shape of the distribution than do moments. In practice, estimates of quantiles provide a more compact summary than do histograms or scatter plots. Therefore, quantile estimation is useful in a variety of problems.

For $0 < p < 1$ and an unknown distribution F_0, the pth quantile is defined by

$$\theta_p = F_0^{-1}(p), \tag{2.1}$$

where $F_0^{-1}(p) = \inf\{x \mid F_0(x) \geq p\}$. There is an abundant literature on estimation of θ_p under the basic assumption that F_0 is continuous and has a positive density at θ_p. The pth sample quantile

$$\tilde{\theta}_{pn} = F_n^{-1}(p),$$

where n denotes the sample size and F_n is the empirical distribution, is the conventional estimator whose asymptotic properties, in particular its consistency and asymptotic normality, are well studied and can be found in standard asymptotics (for example, Serfling 1980). Alternatives to sample quantiles have been proposed by Reiss (1980), Yang (1985) and others. Exact distribution-free confidence intervals for θ_p are discussed in David (1981). Asymptotic nonparametric confidence intervals for quantiles can be constructed using bootstrap (Efron 1979; Hall and Martin 1989; Ho and Lee 2005), or empirical likelihood (for example, Owen 1988).

All of the above quantile estimation methods are restricted by the continuity assumption about F_0 at θ_p, and hence cannot be applied to discrete distributions which typically consist of jumps and plateaus. Nevertheless, data from discrete distributions are prevalent in many social and behavioral science applications. For example, a health insurance company may be interested in the number of claims made by an individual from a certain age group during one year. Such a discrete distribution F_0 takes only discrete integer values, and hence its quantiles must be an integer where F_0 has a jump. Recall a theorem on the consistency of the sample quantile $\tilde{\theta}_{pn}$ (Serfling 1980, p. 75): If θ_p is the unique solution x of

$$F_0(x-) \leq p \leq F_0(x), \tag{2.2}$$

then $\tilde{\theta}_{pn} \xrightarrow{a.s.} \theta_p$. In other words, $\tilde{\theta}_{pn}$ is a consistent estimator for θ_p if (1) F_0 is a continuous distribution, or (2) F_0 is a discrete distribution but p is not a plateau of F_0. For a discrete distribution F_0, the sample quantile $\tilde{\theta}_{pn}$ is *not* consistent for θ_p if p is at a plateau of the distribution, *i.e.*, $F_0(\theta_p) = p$ *and* F_0 is flat in a right-neighborhood of θ_p (Serfling 1980, p. 74). In this case, the expected bias of $\tilde{\theta}_{pn}$ will not decrease as the sample size increases. For the health insurance example in which most observed values are small integers, $\tilde{\theta}_{pn}$ will be very misleading if $\tilde{\theta}_{pn}$ is off by only 1. Therefore, it is important to come up with a consistent quantile estimator for discrete distributions.

In contrast to the case where F_0 is continuous, the discrete case has not been well studied. Some recent exceptions include González-Barrios and Rueda (2001), and Machado and Santos Silva (2005). If p is at a plateau of F_0, then there is an interval $[a, b)$ or $[a, b]$ such that any x inside the interval satisfies $F_0(x) = p$, and thus $\theta_p = a$ by definition (2.1). González-Barrios and Rueda (2001) showed that there exist two random subsequences $\{n_k\}$ and $\{m_k\}$ of $\mathbb{N}$ such that $\tilde{\theta}_{pn_k}$ and $\tilde{\theta}_{pm_k}$ converge almost surely to a and b, respectively, as $k \to \infty$, and proposed an algorithm to find such subsequences. $\tilde{\theta}_{pn_k}$ is therefore the estimator for θ_p proposed by González-Barrios and Rueda. Their algorithm was shown to work well for a binomial distribution with only two possible outcomes, and for distribution functions having a plateau at the level of p *but* being strictly increasing to the left and right of the

plateau (as in Figure 2.1). However, for typical discrete distributions without any strictly

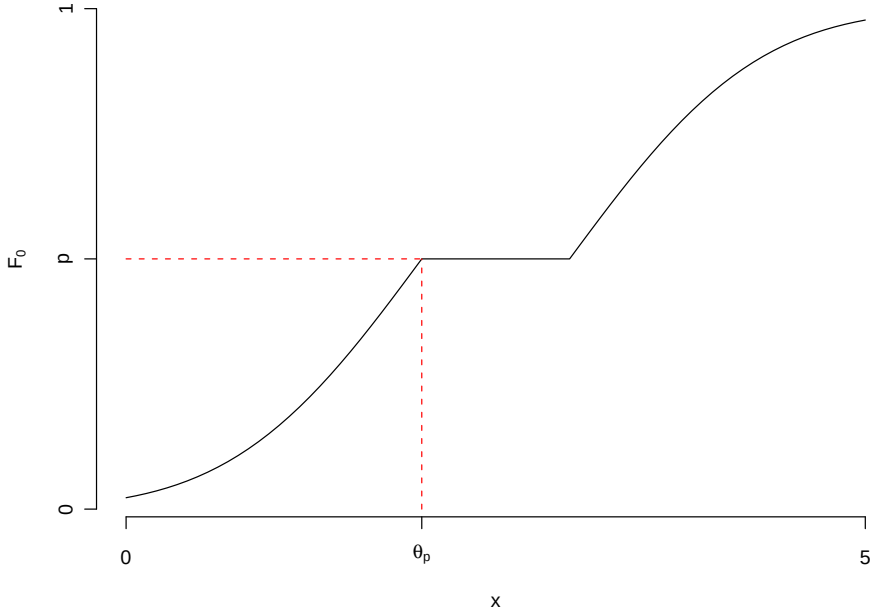

Figure 2.1: A typical distribution F_0, of which the pth quantile θ_p can be consistently estimated by the estimator proposed by González-Barrios and Rueda (2001). F_0 has a plateau at level p, and is strictly increasing to the left and right of the plateau.

increasing segment, such as the Poisson, we carried out a simulation study that showed that, if F_0 has a plateau at the level of p, the method of González-Barrios and Rueda fails to yield an estimator whose mean squared error is decreasing in n. Specifically, we generated random samples of sizes ranging from 200 (as noted by González-Barrios and Rueda, their algorithm requires that the sample size be fairly large) to 2000 from the Poisson distribution with mean 2; for each sample size n, we examined the mean and the variance of González-Barrios and Rueda's estimator using 1000 replicate samples. As Figure 2.2 shows, for $p = 0.406$, which is a plateau of Poisson(2), the variance of the estimator stabilizes at around 0.2 for all $n \geq 1000$, and the mean of the estimator does not approach the true quantile 1 as n increases.

Machado and Santos Silva (2005) considered quantile regression for counts data, and argued that quantile regression coefficients of covariates and hence conditional quantiles of

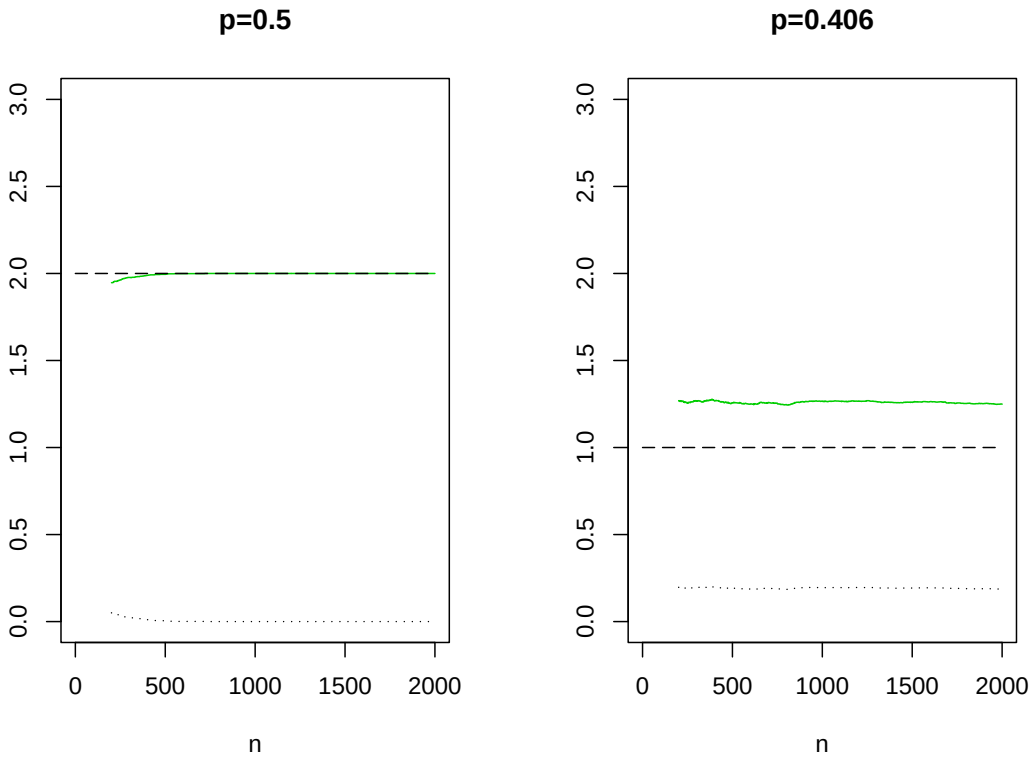

Figure 2.2: A simulation study testing the estimator proposed by González-Barrios and Rueda. 1000 samples were generated from Poisson(2) for $n \in \{200, \ldots, 2000\}$. Here, $p = 0.5$ is not a plateau of Poisson(2), but $p = 0.406$ is. In each panel, the dashed line indicates the true quantile, while the dotted line and the solid lines represent the variance and the mean of 1000 estimated quantiles based on 1000 replicate samples, respectively. In the left panel, the solid line for the mean essentially coincides with the dashed line for the true quantile for all large $n > 500$; while in the right panel, the solid line does not approach the dashed line as n increases.

the discrete response variable can be consistently estimated, provided that there is at least one continuous random covariate. Clearly, Machado and Santos Silva's approach does not apply to the setting where no covariate is involved and unconditional quantiles are of interest.

Considering the fact that the consistency of the sample quantile $\tilde{\theta}_{pn}$ depends on whether or not p, the order of interest, is at a plateau of the true distribution F_0, we can infer that a consistent estimator for θ_p, if it exists, must be based on correct (at least in the asymptotic sense) judgment on the shape of F_0 at level p. In the nonparametric setting, this judgment requires the information available from data, which can be summarized by empirical likelihood. Thus, our focus here is to explore quantile estimation for discrete distributions with the tool of empirical likelihood. The rest of the chapter is organized as follows. In Section 2.2, we first briefly review quantile estimation via empirical likelihood under the assumption that F_0 is continuous, and then provide an explicit form of the EL estimator and its consistency result. Estimation of θ_p in the discrete case is studied more elaborately in Section 2.3. Two data sets are analyzed in Section 2.4. Some practical issues are discussed in Section 2.5. Finally, Section 2.6 presents our conclusion.

2.2 Quantiles of Continuous Distributions

2.2.1 A Review of EL for Quantile Estimation

If F_0 is continuous and has a positive density at the pth quantile θ_p, then θ_p is a functional that can be defined as the root of the estimating equation

$$E[g(X, \theta_p)] = E[1(X \leq \theta_p) - p] = 0. \tag{2.3}$$

Suppose $\{X_1, \ldots, X_n\}$ is a random sample of size n from F_0. For $\theta \in [X_{(1)}, X_{(n)})$, the empirical likelihood ratio (ELR) of θ is defined to be

$$\mathcal{R}(\theta) = \sup \left\{ \prod_{i=1}^{n} n w_i \mid \sum_{i=1}^{n} w_i [1(X_i \leq \theta) - p] = 0,\ w_i \geq 0,\ \sum_{i=1}^{n} w_i = 1 \right\}. \tag{2.4}$$

The MELE for θ_p based on a sample of size n is

$$\hat{\theta}_{pn} = \arg\max_{\theta} \mathcal{R}(\theta). \tag{2.5}$$

The application of empirical likelihood to quantiles dates back to Owen (1988), who showed that $-2\log\mathcal{R}(\theta_p)$ is asymptotically calibrated by the $\chi^2_{(1)}$ distribution, and approximate confidence intervals for θ_p can be constructed accordingly. However, note that the estimating function $g(X, \theta_p) = 1(X \leq \theta_p) - p$ is not smooth in θ. This lack of smoothness raises difficulties in inference for quantiles using empirical likelihood: empirical likelihood cannot deliver confidence intervals with coverage accuracy better than $O(n^{-1/2})$, compared to $O(n^{-1})$ in other contexts; Qin and Lawless's consistency and asymptotic normality results about the MELE do not apply to quantiles.

The first problem has been addressed by some researchers. Chen and Hall (1993) smoothed $1(X \leq \theta)$ by a properly chosen kernel, and substituted the smoothed version for $1(X \leq \theta)$ in estimating equation (2.3). Alternatively, Adimari (1998), Zhou and Jing (2003a) used (2.3) to derive the explicit form of $-2\log\mathcal{R}(\theta)$ involving the unsmooth empirical distribution function F_n, and then replaced F_n by different smooth substitutes. These modifications have been shown to yield improved coverage accuracy to $O(n^{-1})$ of confidence intervals for θ_p. Yet, little attention has been paid to the second problem. Properties, in particular consistency, of the MELE for θ_p have not been explicitly discussed for continuous nor discrete distributions.

Since the first use of EL for quantile estimation by Owen (1988), EL has been extended by many others to related problems such as quantiles and conditional quantiles, both in the presence of auxiliary information (see Zhang, 1995, and Qin & Wu, 2001, respectively), and quantile differences (*e.g.*, the interquartile range; Zhou and Jing, 2003b). In spite of the importance of discrete data, all previous work on quantile estimation via empirical likelihood is under the continuity assumption about the underlying distribution.

Exploring the behavior of the MELE in the discrete case is our major interest. As we will see shortly, properties of $\hat{\theta}_{pn}$ in the continuous case help in determining its consistency

in the discrete case as well. Therefore, we begin with a brief study of $\hat{\theta}_{pn}$ in the continuous case.

2.2.2 MELE of Quantiles

As noted above, if the underlying distribution is continuous, then $\mathcal{R}(\theta)$ is defined by (2.4), and its explicit form can be derived by the method of Lagrange multipliers. Specifically,

$$\mathcal{R}(\theta) = \prod_{i=1}^{n} \{1 + \lambda[1(X_i \leq \theta) - p]\}^{-1}, \tag{2.6}$$

where λ solves

$$\sum_{i=1}^{n} \frac{1(X_i \leq \theta) - p}{\lambda[1(X_i \leq \theta) - p] + 1} = 0. \tag{2.7}$$

Equation (2.7) can be simplified by using the fact that $nF_n(\theta) = \sum_{i=1}^{n} 1(X_i \leq \theta)$. λ is then solved to be

$$\lambda = \frac{F_n(\theta) - p}{p(1-p)}. \tag{2.8}$$

Substituting (2.8) for λ in (2.6), we get

$$\mathcal{R}(\theta) = \left(\frac{p}{F_n(\theta)}\right)^{nF_n(\theta)} \left(\frac{1-p}{1-F_n(\theta)}\right)^{n-nF_n(\theta)}. \tag{2.9}$$

Note that the curve of $\mathcal{R}(\theta)$ is a right-continuous step function in θ, since $F_n(\theta)$ takes a jump at each observation and is right-continuous. This means that the MELE of θ_p is not a unique point but an interval. For uniqueness, we can redefine

$$\hat{\theta}_{pn} = \inf\left\{\theta' \mid \mathcal{R}(\theta') = \max_{\theta} \mathcal{R}(\theta)\right\}. \tag{2.10}$$

Also note that $\mathcal{R}(\theta)$ in equation (2.9) is a function of θ only through $F_n(\theta)$, and is maximized to 1 when $F_n(\theta) = p$. But the value p may not be attainable by $F_n(\theta)$, since $F_n(\theta)$ must take a value from $\{\frac{1}{n}, \ldots, \frac{n-1}{n}\}$. Therefore the maximum of $\mathcal{R}(\theta)$ is actually obtained when $F_n(\theta) = [np]/n$ or $F_n(\theta) = ([np]+1)/n$, and $\max_\theta \mathcal{R}(\theta)$ may not be exactly 1. Taking into account that $F_n(\theta) = i/n$ for all $\theta \in [X_{(i)}, X_{(i+1)})$ and definition (2.10), we

summarize the MELE of the pth quantile to be

$$\hat{\theta}_{pn} = \begin{cases} X_{np:n} & \text{if } np \in \mathbb{N} \\ X_{[np]:n} & \text{if } np \notin \mathbb{N} \text{ and } \mathcal{R}(X_{[np]:n}) \geq \mathcal{R}(X_{[np]+1:n}) \\ X_{[np]+1:n} & \text{if } np \notin \mathbb{N} \text{ and } \mathcal{R}(X_{[np]:n}) < \mathcal{R}(X_{[np]+1:n}) \end{cases}, \tag{2.11}$$

where $X_{i:n}$ is the extended form of $X_{(i)}$, with the explicit subscript n denoting the sample size, and $\mathcal{R}(X_{i:n}) = \left(\frac{p}{i/n}\right)^i \left(\frac{1-p}{1-i/n}\right)^{n-i}$.

2.2.3 Consistency of the MELE of a Quantile

It is worth noting that the MELE $\hat{\theta}_{pn}$ defined above is very close to the sample quantile

$$\tilde{\theta}_{pn} = F_n^{-1}(p) = \begin{cases} X_{np:n} & \text{if } np \in \mathbb{N} \\ X_{[np]+1:n} & \text{if } np \notin \mathbb{N} \end{cases}. \tag{2.12}$$

$\hat{\theta}_{pn}$ differs from $\tilde{\theta}_{pn}$ only when $np \notin \mathbb{N}$ *and* $\mathcal{R}(X_{[np]:n}) \geq \mathcal{R}(X_{[np]+1:n})$, and the difference is just the gap between two consecutive order statistics $X_{[np]:n}$ and $X_{[np]+1:n}$. This difference is insignificant for two reasons. First, it is caused by definitions, since both $\tilde{\theta}_{pn}$ and $\hat{\theta}_{pn}$ can be defined in multiple ways. $\hat{\theta}_{pn}$ could be defined to be exactly the same as $\tilde{\theta}_{pn}$, if one is willing to consistently follow the logic behind definition (2.12), namely, that we take $\hat{\theta}_{pn} = X_{np:n}$ if $p \in \{\frac{1}{n}, \ldots, \frac{n-1}{n}\}$, and take $\hat{\theta}_{pn} = X_{[np]+1:n}$ by using $([np]+1)/n$ in place of p if $p \notin \{\frac{1}{n}, \ldots, \frac{n-1}{n}\}$. Second, even if we keep the current definitions of $\tilde{\theta}_{pn}$ and $\hat{\theta}_{pn}$, the difference will converge to zero in probability as $n \to \infty$. This can be verified by applying Lemma 21.7 in van der Vaart (1998); see the proof of Result 2.2.1 (Appendix) for details.

We know from standard asymptotics (such as Serfling, 1980) that, for continuous distributions, $\tilde{\theta}_{pn}$ is consistent for θ_{pn}. By utilizing the close relationship between $\hat{\theta}_{pn}$ and $\tilde{\theta}_{pn}$, the following result about the consistency and asymptotic normality of the MELE $\hat{\theta}_{pn}$ can be proved.

Result 2.2.1. *If F_0 is continuous, for $0 < p < 1$, the MELE $\hat{\theta}_{pn}$ is consistent for θ_p, i.e. $\hat{\theta}_{pn} \xrightarrow{p} \theta_p$. If F_0 is twice differentiable at θ_p, then*

$$n^{1/2}\left(\hat{\theta}_{pn} - \theta_p\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{[F_0'(\theta_p)^2]}\right). \tag{2.13}$$

2.3 Quantiles of Discrete Distributions

2.3.1 A Jittering Method and MELE of Quantiles

If the underlying distribution is discrete, the equation $E[1(X \le \theta_p) - p] = 0$ does not hold for all $0 < p < 1$, and consequently it cannot be used directly to define the ELR for θ_p. Take the Poisson(2) distribution as an example. We know that $\Pr(X \le 1) = 0.406$ and $\Pr(X \le 2) = 0.677$, so the median $\theta_{0.5}$ is 2, but $E[1(X \le \theta_{0.5}) - 0.5] = 0.177$. Furthermore, there will be many ties in a sample, especially when the sample size is large. When F_0 is discrete and many ties are present, we may get an empty confidence region for θ_p if we use $\mathcal{R}(\theta)$ of the form (2.9) (Owen 2001, p.46). Because of ties, $F_n(\theta) = \sum_{i=1}^{n} 1(X_i \le \theta)/n$ skips many values in the set $\{\frac{1}{n}, \ldots, \frac{n-1}{n}\}$, possibly including all values corresponding to large $\mathcal{R}(\theta)$. In this case, at confidence level 95%, say, even $\max \mathcal{R}(\theta)$ may be too small for the MELE to be included in the confidence region, if the critical value from $\chi^2(1)$ is used. This possible situation is illustrated by Figure 2.3. The left panel of Figure 2.3 is the histogram of

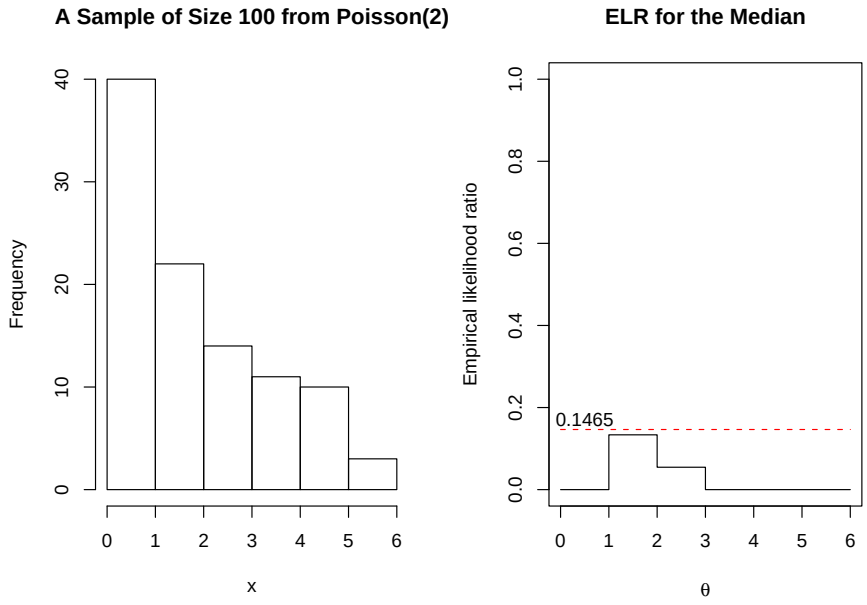

Figure 2.3: An empty confidence region. The left panel is the histogram of a random sample from Poisson(2). The right panel shows the empirical likelihood ratio curve for the median; the entire curve is below the threshold 0.1465.

a random sample from Poisson(2), which indicates many ties at the true median 2 (and every other value), and the right panel is the empirical likelihood ratio curve for the median. The maximum of $\mathcal{R}(\theta)$ for this sample is only 0.134 which is below $\exp(-\chi^2_{0.95}(1)/2) = 0.1465$, the threshold at the 95% level.

To circumvent these difficulties, we can proceed by adding a small jitter Z_i to each X_i, so that the transformed data $\{Y_i = X_i + Z_i,\ i = 1, 2, \ldots, n\}$ no longer have ties, and the order of $\{X_i\}$ is preserved in the sense that if $X_i < X_j$ then $Y_i < Y_j$. The method of jittering is discussed in Owen (2001) and Machado and Santos Silva (2005). We now use superscripts X and Y to distinguish functionals and statistics associated with the two sets of data. Without ties, $\{Y_i\}$ can be treated as if they came from a continuous distribution, so $\mathcal{R}^Y(\theta)$ has the form of (2.9) and is hence maximized to 1 or a value close to 1 at the MELE $\hat{\theta}^Y_{pn}$ determined by (2.11). Suppose that in the original sample $\{X_i, i = 1, \ldots, n\}$, there are $d < n$ distinct values, denoted by $x_1 < x_2 < \ldots < x_d$. For each $\theta = x_j$, $j = 1, 2, \ldots, d$, the empirical likelihood ratio for the pth quantile of X is defined by

$$\mathcal{R}^X(\theta)|_{\,\theta=x_j} = \mathcal{R}^X(x_j) = \max_{Y_i \in G_j} \left\{\mathcal{R}^Y(Y_i)\right\},$$

where G_j is the set of Y_is generated by the X_is having the same value x_j. For $\theta \notin \{x_1, x_2, \ldots, x_d\}$, $\mathcal{R}^X(\theta)$ is computed by (2.9), in the same way as in the continuous case. The MELE $\hat{\theta}^X_{pn}$ is obtained by transforming $\hat{\theta}^Y_{pn}$ back to the X scale.

Without loss of generality, we assume that the support of F_0^X is $\mathcal{S}_x = \{x_1, x_2, \ldots\}$, where $x_1 < x_2 < \ldots$ are consecutive integers. Let $0 < p_1,\ p_2,\ \ldots < 1$ be the weight that F_0^X puts on $x_1, x_2, \ldots$, respectively, where $\sum_{i=1}^{j} p_i = \Pr(X \le x_j) = P_j \to 1$ as j increases. The smallest nonzero gap in a sample $\{X_i\}$ from F_0^X is therefore at least 1, and it is convenient to use jitters $Z_i \overset{iid}{\sim} U(0, 1]$. Consequently, the probability mass that F_0^X puts on each integer point x_j is evenly spread over the interval $(x_j, x_{j+1}]$, and thus F_0^Y is continuous with a piece-wise constant slope that changes at each point x_j, *i.e.*,

$$\frac{\partial F_0^Y(y)}{\partial y}|_{y\in(x_j,x_{j+1})} = p_j, \tag{2.14}$$

and $F_0^Y(y)|_{y=x_j} = \sum_{i=1}^{j-1} p_i = P_{j-1}$. Figure 2.4 gives a simple example where X_i follows the Poisson(2) distribution. The left panel shows the cumulative distribution function of Poisson(2) (i.e., F_0^X) up to $x = 7$, and the right panel displays the corresponding CDF of the jittered variable $Y_i = X_i + Z_i$. Here, $x_j = j$, $P_j = F_0^X(j) = \Pr(X_i \leq j)$ and $F_0^Y(j) = P_{j-1}$ for $j = 0, 1, \ldots$. The inverse transformation from the Y scale to the X scale is defined to be $\lceil Y - 1 \rceil$, where $\lceil a \rceil$ is the ceiling function that returns the smallest integer greater than or equal to a. Therefore, $\hat{\theta}_{pn}^X = \left\lceil \hat{\theta}_{pn}^Y - 1 \right\rceil$.

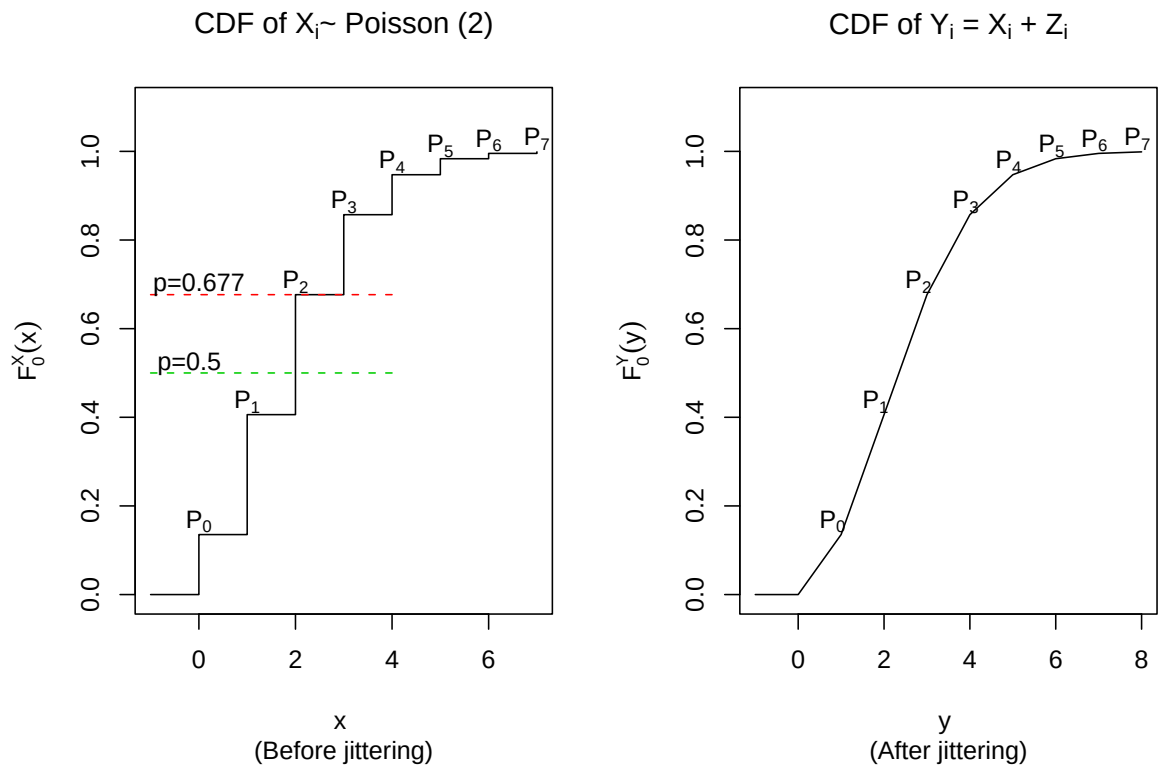

Figure 2.4: The jittering transformation. Jitters $Z_i \overset{iid}{\sim} U(0, 1]$ are used. Here, $P_j = \Pr(X_i \leq j)$; the right tails of the cumulative distributions F_0^X and F_0^Y are not shown. After jittering, the probability mass that F_0^X puts on each integer point j is evenly spread over the interval $(j, j + 1]$, so that F_0^Y is continuous with a piece-wise constant slope.

2.3.2 Consistency of the MELE

The jittering method considered in Section 2.3.1 is a convenient device that helps us examine the consistency of $\hat{\theta}_{pn}^X$ in the following two cases.

(i) First consider p such that $P_{k-1} < p < P_k$ for some $k \in \{1, 2, \ldots\}$, that is, p is not a plateau of F_0. In this case θ_p^X satisfies $F_0^X(\theta_p-) < p < F_0^X(\theta_p)$ and $E[1(X \le \theta_p) - p] > 0$. The median ($p = 0.5$) of Poisson(2) is an example, as shown in Figure 2.4. Here, $\theta_p^X = x_k$, $x_k < \theta_p^Y < x_{k+1}$, and $\theta_p^X = \lceil \theta_p^Y - 1 \rceil$. The inverse transformation $f(t) = \lceil t - 1 \rceil$ is continuous at $t = \theta_p^Y$. By the previous result for a continuous distribution, $\hat{\theta}_{pn}^Y \xrightarrow{p} \theta_p^Y$ and it follows that $\left\lceil \hat{\theta}_{pn}^Y - 1 \right\rceil \xrightarrow{p} \lceil \theta_p^Y - 1 \rceil$, *i.e.* $\hat{\theta}_{pn}^X \xrightarrow{p} \theta_p^X$.

(ii) Next consider $p = P_k$ for some $k \in \{1, 2, \ldots\}$, *i.e.*, p is a plateau of F_0. θ_p^X now satisfies $F_0^X(\theta_p) = p$ and $E[1(X \le \theta_p) - p] = 0$. The 0.677th quantile of Poisson(2) is an example of this case, as illustrated by Figure 2.4. Since $\theta_p^X = x_k$ and $\theta_p^Y = x_{k+1}$, it follows again that $\theta_p^X = \lceil \theta_p^Y - 1 \rceil$. Because θ_p^Y is an integer, the inverse transformation $f(t) = \lceil t - 1 \rceil$ is *discontinuous* at $t = \theta_p^Y = x_{k+1}$. Consequently, even though $\hat{\theta}_{pn}^Y \xrightarrow{p} \theta_p^Y$ still holds for $p = P_k$, $\left\lceil \hat{\theta}_{pn}^Y - 1 \right\rceil \xrightarrow{p} \lceil \theta_p^Y - 1 \rceil$ fails because of the discontinuity. In this case, the MELE $\hat{\theta}_{pn}^X$ is *not* a consistent estimator of θ_p^X.

Comment 2.3.1 The inconsistency in case (ii) can be further explained by the following result, which is derived from Theorem A of Serfling (1980, p. 77) and the relation between $\hat{\theta}_{pn}^Y$ and $\tilde{\theta}_{pn}^Y$. The proof is shown in the Appendix. Part (c) of this result says that, for large n, $\hat{\theta}_{pn}^Y > \theta_p^Y$ with probability roughly $\frac{1}{2}$. In the inconsistent case, $\theta_p^X = x_k$ and $\theta_p^Y = x_{k+1} = x_k + 1$; if $\hat{\theta}_{pn}^Y > \theta_p^Y = x_{k+1}$, then $\hat{\theta}_{pn}^X = \left\lceil \hat{\theta}_{pn}^Y - 1 \right\rceil \ge x_{k+1} = \theta_p^X + 1$. That is, $\hat{\theta}_{pn}^X \ge \theta_p^X + 1$ with an approximate probability of 0.5 when n is large.

Result 2.3.1. *Let $0 < p < 1$. Suppose that $F_0^Y(y)$ is continuous at $y = \theta_p^Y$.*

(a) If there exists $\frac{\partial F_0^Y(\theta_p^Y-)}{\partial y} > 0$, then for $t < 0$,

$$\lim_{n\to\infty} Pr\left(\frac{n^{1/2}\left(\hat{\theta}_{pn}^Y - \theta_p^Y\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le t \right) = \Phi(t),$$

(b) If there exists $\frac{\partial F_0^Y(\theta_p^Y+)}{\partial y} > 0$, then for $t > 0$,

$$\lim_{n\to\infty} Pr\left(\frac{n^{1/2}\left(\hat{\theta}_{pn}^Y - \theta_p^Y\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y+)}{\partial y}} \le t \right) = \Phi(t),$$

(c) In any case,

$$\lim_{n\to\infty} Pr\left(n^{1/2}\left(\hat{\theta}_{pn}^{Y}-\theta_{p}^{Y}\right)\leq 0\right)=\Phi(0)=\frac{1}{2}.$$

Comment 2.3.2 For the discrete distribution F_0^X, $\hat{\theta}_{pn}^X$ either hits the true parameter θ_p^X, or is off by at least 1. So a sufficient and necessary condition for the consistency of $\hat{\theta}_{pn}^X$ is $\lim_{n\to\infty}\Pr(\hat{\theta}_{pn}^X=\theta_p^X)=1$. In fact, Result 2.3.1 and the relationship between X and Y can be used to show that $\lim_{n\to\infty}\Pr(\hat{\theta}_{pn}^X=\theta_p^X)=1$ for p in case (i); Result 2.3.1. leads further to $\lim_{n\to\infty}\Pr(\hat{\theta}_{pn}^X=\theta_p^X)=0.5$ for p in case (ii). These are summarized in Result 2.3.2, which also points out that, as n increases, $\hat{\theta}_{pn}^X$ will jump only between θ_p^X and θ_p^X+1 with approximately equal probability. The proof of Result 2.3.2 is relatively simple and hence is omitted.

Result 2.3.2. *If p is such that $P_{k-1}<p<P_k$ for some integer $k\geq 1$, then*

$$\lim_{n\to\infty}\Pr(\hat{\theta}_{pn}^X=\theta_p^X)=1;$$

if p satisfies $p=P_k$ for some integer $k\geq 1$, then

$$\lim_{n\to\infty}\Pr(\hat{\theta}_{pn}^X=\theta_p^X)=\lim_{n\to\infty}\Pr(\hat{\theta}_{pn}^X=\theta_p^X+1)=0.5.$$

Comment 2.3.3 When the underlying distribution F_0^X is discrete, $\hat{\theta}_{pn}^X$ may or may not be consistent for θ_p^X, depending on the position of p. Consider an intermediate situation in which $P_{k-1}<p<P_k$, but p is very close to P_{k-1} or P_k. This is categorized as case (i), so $\hat{\theta}_{pn}^X$ is consistent but the rate at which convergence occurs will be very slow. It can be derived from Result 2.3.1 that

$$\Pr(\hat{\theta}_{pn}^X=\theta_p^X)\approx 1-\Phi\left(-\frac{(1-a_1)p_k\sqrt{n}}{\sqrt{p(1-p)}}\right)-\Phi\left(-\frac{a_1p_k\sqrt{n}}{\sqrt{p(1-p)}}\right), \tag{2.15}$$

where $a_1=(p-P_{k-1})/p_k$. For $\Pr(\hat{\theta}_{pn}^X=\theta_p^X)$ to converge to 1 as $n\to\infty$, $Z_1^*=\frac{(1-a_1)p_k\sqrt{n}}{\sqrt{p(1-p)}}$ and $Z_2^*=\frac{a_1p_k\sqrt{n}}{\sqrt{p(1-p)}}$ both need to increase as n increases. For fixed large values of Z_1^* and Z_2^*, n is proportional to $\max\{(1-a_1)^{-2},a_1^{-2}\}$. If p is close to P_{k-1} or P_k, then either a_1 or $(1-a_1)$ will be close to zero, and n will need to be very large for the last two terms in (2.15) to be approximately zero.

2.3.3 An EL-based Classification and a Consistent EL Estimator

The consistency of $\hat{\theta}_{pn}^X$ in the discrete case is not as straightforward as in the continuous case. If one knows in advance whether F_0^X is flat at a given level p, then one can tell whether the estimator (*e.g.*, the MELE or the sample quantile) is consistent or not. Unfortunately, this information is typically unavailable, and hence one may want to use a statistical procedure to categorize p into case (i) or case (ii). We now consider using ELRs associated with $\hat{\theta}_{pn}^Y$ and hence $\hat{\theta}_{pn}^X$ to categorize p. To begin, we suppose that $\hat{\theta}_{pn}^X = \lceil \hat{\theta}_{pn}^Y - 1 \rceil$ has the value x^*, and that $Y_{(L)}$ and $Y_{(U)}$ are the smallest and largest order statistics that satisfy $\lceil Y_{(i)} - 1 \rceil = x^*$. Note that the indices L and U are random variables. As shown in Figure 2.5, let $C_l = \mathcal{R}^Y(Y_{(L)})$ and $C_u = \mathcal{R}^Y(Y_{(U)})$. By (2.9), C_l and C_u can be explicitly written as

$$C_l = \left(\frac{p}{L/n}\right)^L \left(\frac{1-p}{1-L/n}\right)^{n-L} \quad \text{and} \quad C_u = \left(\frac{p}{U/n}\right)^U \left(\frac{1-p}{1-U/n}\right)^{n-U}. \tag{2.16}$$

Further, let

$$h_l = \Pr\{\, \chi^2(1) \le -2\log(C_l) \,\} \quad \text{and} \quad h_u = \Pr\{\, \chi^2(1) \le -2\log(C_u) \,\}. \tag{2.17}$$

Result 2.3.3 indicates that h_l and h_u behave differently for the two types of p.

Result 2.3.3. *Suppose that $\{X_i\}$ is a random sample of size n from a discrete distribution F_0^X, and $\{Y_i\}$ is the jittered sample with distribution F_0^Y. Let h_l and h_u be defined by (2.17). Then,*

(a) both $h_l \xrightarrow{p} 1$ and $h_u \xrightarrow{p} 1$ for p in case (i);

(b) $\max\{h_l, h_u\} \xrightarrow{p} 1$ but $\min\{h_l, h_u\} \xrightarrow{d} U(0,1)$ for p in case (ii).

A direct implication of Result 2.3.3 is that $\min(h_l, h_u)$ converges to 1 for p in case (i) but not for p in case (ii). Thus, we hope to categorize p by examining whether $\min(h_l, h_u)$ is close to 1. An immediate question would be "How close is defined to be close?" If $\min(h_l, h_u)$ were exactly 1 for p in case (i), we could simply check whether or not $\min(h_l, h_u)$ is 1, and the probability of misclassification would be zero. However, $\min(h_l, h_u)$ is random in either

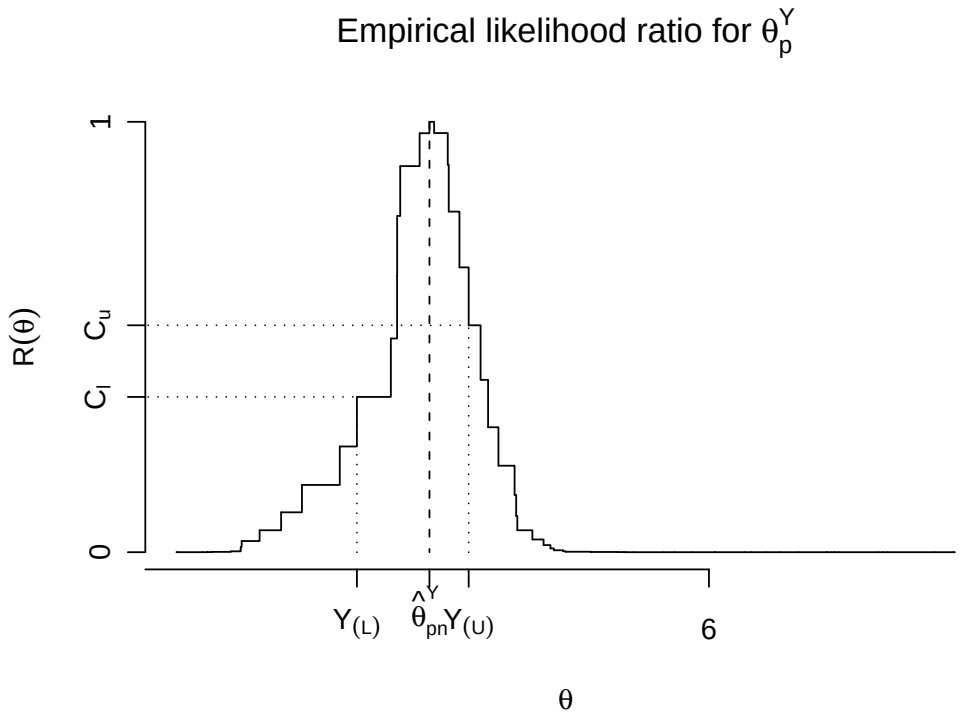

Figure 2.5: Two ELRs $C_l = \mathcal{R}^Y(Y_{(L)})$ and $C_u = \mathcal{R}^Y(Y_{(U)})$ associated with the MELE $\hat{\theta}^Y_{pn}$. Here, $Y_{(L)}$ and $Y_{(U)}$ are the smallest and largest order statistics satisfying $\lceil Y_{(i)} - 1 \rceil = \lceil \hat{\theta}^Y_{pn} - 1 \rceil$.

case, with its variability decreasing to zero for p in case (i). Therefore, we need to choose a classification criterion whose misclassification rate is acceptable for moderate samples and diminishes to zero as the sample size increases to infinity. In particular, we can use $\delta(n) \in (0,1)$, a function decreasing to zero in n, to categorize p according to

$$p \Longrightarrow \begin{cases} \text{case (i)}, & \text{if } 1 - \min(h_l, h_u) \leq \delta(n) \\ \text{case (ii)}, & \text{if } 1 - \min(h_l, h_u) > \delta(n) \end{cases} . \tag{2.18}$$

Possible candidates for $\delta(n)$ include n^{-1}, $n^{-1/2}$, $n^{-1/3}$, etc. The choice of $\delta(n)$ will be studied in Section 2.3.4.

As we have seen, whether or not the MELE $\hat{\theta}_{pn}^X$ is consistent for θ_p^X depends on whether F_0^X has a plateau at the level of p. Result 2.3.2 implies that, if the sample size is large enough, then $\hat{\theta}_{pn}^X$ will be either constant at θ_p^X for p in case (i), or vary only between θ_p^X and $\theta_p^X + 1$ with approximatively equal probability for p in case (ii). Thus, to obtain a consistent estimator, it is desirable to apply the categorization procedure proposed above to distinguish p in cases (i) and (ii), and then make an appropriate adjustment to $\hat{\theta}_{pn}^X$ if p is identified to be in case (ii). The only necessary adjustment is to subtract 1 from $\hat{\theta}_{pn}^X$ if $\hat{\theta}_{pn}^X = \theta_p^X + 1$, which will occur with large samples only if p is a plateau of F_0^X. Therefore, once a specific p is categorized into case (ii), we need to make a further judgment about whether $\hat{\theta}_{pn}^X = \theta_p^X + 1$. From the proof of Result 2.3.3 (Appendix), for p in case (ii), we find that

$$\begin{cases} h_l \geq h_u \text{ with probability tending to 1, if } \hat{\theta}_{pn}^X = \theta_p^X \\ h_l < h_u \text{ with probability tending to 1, if } \hat{\theta}_{pn}^X = \theta_p^X + 1 \end{cases} .$$

This inspires the following two-step judgment procedure:

- Step 1. Use rule (2.18) to categorize p.
- Step 2. If p is categorized into case (ii), then compare h_l with h_u: assume $\hat{\theta}_{pn}^X = \theta_p^X$ if $h_l \geq h_u$; otherwise, assume $\hat{\theta}_{pn}^X = \theta_p^X + 1$.

Now that an approach has been proposed for categorizing p and identifying the sub-case where $\hat{\theta}_{pn}^{X}$ is biased, a consistent EL estimator $\hat{\theta}_{pn}^{Xc}$ can be formulated as

$$\hat{\theta}_{pn}^{Xc} = \begin{cases} \hat{\theta}_{pn}^{X} - 1, & \text{if } 1 - \min(h_l, h_u) > \delta(n) \text{ and } h_l < h_u; \\ \hat{\theta}_{pn}^{X}, & \text{otherwise.} \end{cases} \tag{2.19}$$

2.3.4 Simulation Studies

In this section, we study the performance of the EL-based categorization procedure and the new estimator $\hat{\theta}_{pn}^{Xc}$ by simulation. In our simulation studies, n^{-1}, $n^{-1/2}$, $n^{-1/3}$, $n^{-1/4}$ and $n^{-1/5}$ are included as candidates for $\delta(n)$; of course these candidates are arbitrarily chosen, and other forms of $\delta(n)$ are possible. The choice of $\delta(n)$ is also discussed here.

In the first simulation study, for simplicity and without loss of generality, the discrete uniform distribution on $\{1, 2, \ldots, 10\}$ (*i.e,* F_0^X), which consists of plateaus at levels of $\{0.1, \ldots, 1.0\}$, is used to generate the data. Here, $p = 0.75$ and $p = 0.78$ are both examples of case (i), with $p = 0.75$ representing a value located exactly in the middle of two plateaus, and $p = 0.78$ representing a value which leans toward one of the two immediate plateaus; $p = 0.70$ is an example of case (ii). We repeat the simulation 1000 times, and count the numbers of correct ("C" in Table 2.1) and incorrect ("NC") categorizations for each combination of $\delta(n)$ and $n = 100, 500, 1000, 2000, 10000$. Note that "C" and "NC" also summarize the performance of Step 1 of the two-step judgment procedure. Entries "S0" and "S1" in Table 2.1 jointly show how Step 2 performs in making the second stage decision, given that p is already categorized as a plateau (or a step; denoted by "S") of F_0^X . "S0" represents the decision that $\hat{\theta}_{pn}^{X} = \theta_p^X$, with "0" meaning no bias, and "S1" stands for $\hat{\theta}_{pn}^{X} = \theta_p^X + 1$, with "1" meaning a bias of 1. Of course, for p in case (ii), only when the sample size is large enough, will $\hat{\theta}_{pn}^{X}$ take a value only from θ_p^X and $\theta_p^X + 1$; $\hat{\theta}_{pn}^{X}$ may take a value other than θ_p^X or $\theta_p^X + 1$ with a small or moderate sample. To reflect the discrepancy between $\hat{\theta}_{pn}^{X}$ being *determined* by Step 2 to be θ_p^X (or $\theta_p^X + 1$) and $\hat{\theta}_{pn}^{X}$ *actually* being θ_p^X (or $\theta_p^X + 1$), both "S0" and "S1" contain two numbers. Specifically, a/b under entry "S0" and c/d under "S1" mean that p is

categorized as a plateau of F_0^X at $b+d$ simulation runs, which equals "NC" for $p = 0.75$ and $p = 0.78$ in case (i) but equals "C" for $p = 0.7$ in case (ii); among the $b+d$ simulation runs, $h_l \geq h_u$ and hence $\hat{\theta}_{pn}^X$ is determined to be θ_p^X in b simulation runs, and $\hat{\theta}_{pn}^X$ is determined to be $\theta_p^X + 1$ in the other d runs; in fact, $\hat{\theta}_{pn}^X = \theta_p^X$ in a out of b runs, and $\hat{\theta}_{pn}^X = \theta_p^X + 1$ in c out of d runs. The information provided by "S0" and "S1" is further explained by the flow chart in Figure 2.6. Ideally, $a = b$ and $c = d$, which would mean that we have made 100% correct judgment on whether $\hat{\theta}_{pn}^X$ equals to θ_p^X or $\theta_p^X + 1$, given that p is already categorized as a plateau of F_0^X.

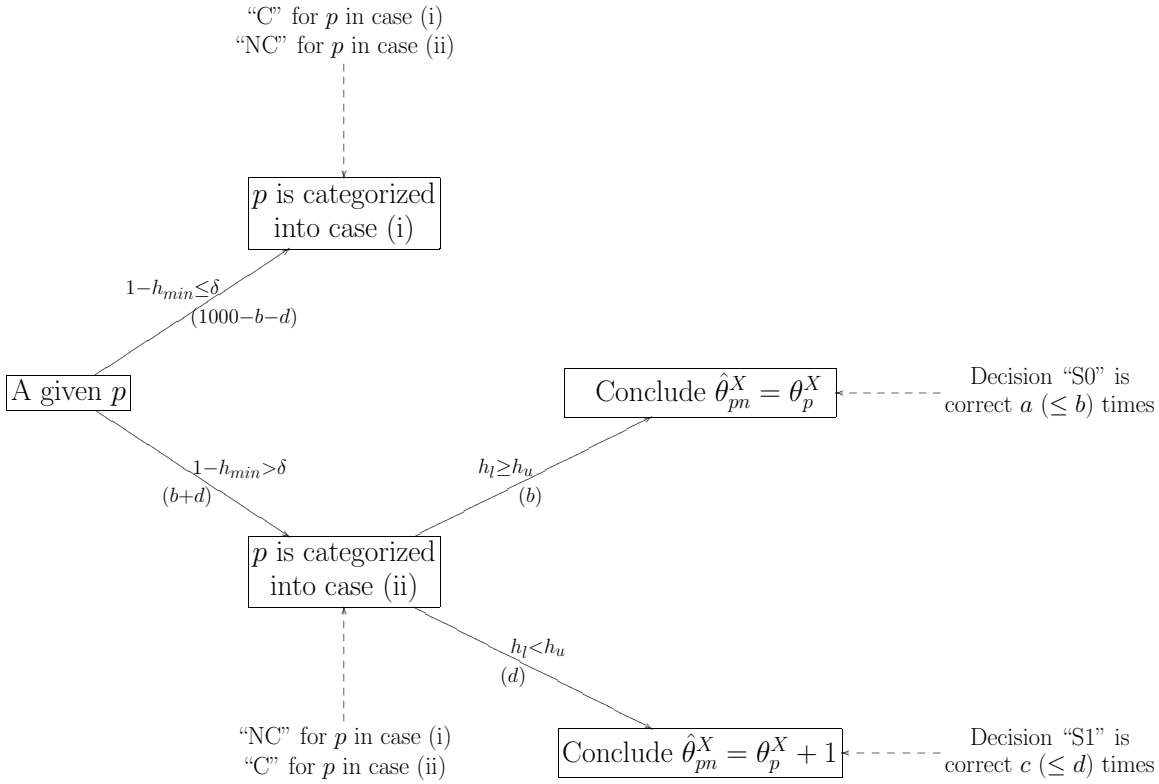

Figure 2.6: The two-step judgment procedure. A solid arrow "$\rightarrow$" points to a possible decision of Step 1 or 2; above the arrow is the condition leading to this particular decision; beneath the arrow is the number of simulation runs (out of 1000) in which this decision is made. A dashed arrow "$\dashrightarrow$" connects a decision and its corresponding notation used in Table 2.1.

The categorization procedure. As one can observe from the simulation results displayed in columns "C" and "NC" in Table 2.1, for any $\delta(n)$ and any p, the EL-based cate-

Table 2.1: Performance of the two-step judgment procedure

	$\delta(n) = n^{-1/5}$											
	$p = 0.75, \theta_p^X = 8$				$p = 0.7, \theta_p^X = 7$				$p = 0.78, \theta_p^X = 8$			
n	C	NC	S0	S1	C	NC	S0	S1	C	NC	S0	S1
100	247	753	228/371	67/382	752	248	303/372	278/380	229	771	341/381	215/390
500	865	135	52/57	4/78	708	292	345/345	363/363	540	460	334/334	125/126
1000	984	16	5/7	0/9	784	216	395/395	389/389	680	320	278/278	42/42
2000	1000	0	0/0	0/0	799	201	394/394	405/405	825	175	164/164	11/11
10000	1000	0	0/0	0/0	840	160	410/410	430/430	1000	0	0/0	1/0
	$\delta(n) = n^{-1/4}$											
	$p = 0.75, \theta_p^X = 8$				$p = 0.7, \theta_p^X = 7$				$p = 0.78, \theta_p^X = 8$			
n	C	NC	S0	S1	C	NC	S0	S1	C	NC	S0	S1
100	55	945	138/202	29/179	865	135	326/400	344/465	113	887	405/462	239/425
500	797	203	67/76	2/127	790	210	399/399	391/391	462	538	417/417	118/121
1000	977	23	6/6	0/17	822	178	412/412	410/410	604	396	345/345	51/51
2000	1000	0	0/0	0/0	881	119	439/439	442/442	790	210	197/197	13/13
10000	1000	0	0/0	0/0	916	84	460/460	456/456	1000	0	0/0	0/0
	$\delta(n) = n^{-1/3}$											
	$p = 0.75, \theta_p^X = 8$				$p = 0.7, \theta_p^X = 7$				$p = 0.78, \theta_p^X = 8$			
n	C	NC	S0	S1	C	NC	S0	S1	C	NC	S0	S1
100	52	948	314/469	92/479	953	47	366/456	350/497	75	925	422/481	235/444
500	703	297	110/116	3/181	897	103	476/476	420/421	338	662	540/540	110/112
1000	952	48	15/15	0/33	920	80	473/473	447/447	463	537	480/480	57/57
2000	1000	0	0/0	0/0	928	72	469/469	459/459	661	339	319/319	20/20
10000	1000	0	0/0	0/0	951	49	474/474	477/477	1000	0	0/0	0/0
	$\delta(n) = n^{-1/2}$											
	$p = 0.75, \theta_p^X = 8$				$p = 0.7, \theta_p^X = 7$				$p = 0.78, \theta_p^X = 8$			
100	14	986	307/448	95/538	998	2	349/471	375/527	5	995	469/516	238/479
500	426	574	248/257	3/317	963	37	484/484	477/479	148	852	689/689	136/163
1000	865	135	63/64	0/71	968	32	484/484	484/484	288	712	667/667	45/45
2000	993	7	2/2	0/5	984	16	498/498	486/486	466	534	522/522	12/12
10000	1000	0	0/0	0/0	994	6	494/494	500/500	995	5	5/5	0/0
	$\delta(n) = n^{-1}$											
	$p = 0.75, \theta_p^X = 8$				$p = 0.7, \theta_p^X = 7$				$p = 0.78, \theta_p^X = 8$			
100	0	1000	329/480	91/520	1000	0	399/491	370/509	0	1000	430/481	263/519
500	8	992	431/440	1/552	1000	0	506/509	481/491	2	998	820/820	129/178
1000	316	684	307/307	0/377	1000	0	495/495	505/505	41	959	904/904	51/55
2000	912	88	26/26	0/62	1000	0	508/508	492/492	83	917	907/907	10/10
10000	1000	0	0/0	0/0	1000	0	498/498	502/502	829	171	171/171	0/0

Step 1 of the judgment procedure in Section 2.3.3 is reflected by "C" and "NC", which record the numbers of correct and incorrect categorizations, respectively, over 1000 runs. Entries "S0" and "S1" jointly show the performance of Step 2. Here, "S" means p is already categorized as a step (plateau) of F_0^X; "S0" and "S1" stand for the two subcases where Step 2 leads to the decision that $\hat{\theta}_{pn}^X = \theta_p^X$ and that $\hat{\theta}_{pn}^X = \theta_p^X + 1$, respectively. In particular, a/b under entry "S0" and c/d under "S1" mean that $\hat{\theta}_{pn}^X$ was *judged* to be θ_p^X and $\theta_p^X + 1$ in b and d simulation runs, respectively, and $\hat{\theta}_{pn}^X$ was *actually* θ_p^X and $\theta_p^X + 1$ in a out of b runs and in c out of d runs, respectively.

gorization procedure performs better as the sample size n increases. With an appropriately chosen $\delta(n)$, this procedure can achieve a 100% correct categorization rate when $n = 10000$. In practice, it is unlikely that one will obtain a sample of size 10000. The main point of providing the simulation results with $n = 10000$ is to show that, asymptotically, any p can be classified into the correct category — a plateau or not a plateau. The categorization procedure has a stronger tendency to over-estimate plateaus with $p = 0.78$ than with $p = 0.75$. Since $p = 0.78$ is closer to 0.8 than to 0.7, it takes a larger sample size for the procedure to well differentiate $p = 0.78$ from the closest plateau $p = 0.8$.

The estimator $\hat{\theta}_{pn}^{Xc}$. With respect to the performance of Step 2 of the two-step judgment procedure, we observe that, for $p = 0.7$ and for all $\delta(n)$ candidates, a and c become very close to or the same as b and d, respectively, when the sample is 500 or larger. This confirms the judgment procedure, and also means that $\hat{\theta}_{pn}^{X}$ is correctly adjusted to form correct $\hat{\theta}_{pn}^{Xc}$ once $p = 0.7$ is identified as a plateau. For $p = 0.75$ and $p = 0.78$, "S0" and "S1" show how misclassifications distribute between the two sub-cases for finite samples, and this information is useful for examining the performance of the estimator $\hat{\theta}_{pn}^{Xc}$, which is the major concern here. Specifically, a/b under "S0" and c/d under "S1" also present proportions of $\hat{\theta}_{pn}^{X}$ that receive the correct treatment, after p is misclassified into the wrong category. For example, for the combination of $p = 0.78$, $\delta(n) = n^{-\frac{1}{5}}$ and $n = 500$, p is misclassified as a plateau in 460 simulation runs; among the 460 runs, $\hat{\theta}_{pn}^{X}$ is *actually* equal to θ_p^X in 334 out of 334 times when it is *judged* to be θ_p^X, and $\hat{\theta}_{pn}^{X}$ is *correctly judged* to be $\theta_p^X + 1$ and hence reduced by 1 in 125 out of the other 126 times; this means that in only 1 out of the 460 simulation runs, $\hat{\theta}_{pn}^{Xc}$ is not in fact correct. For $p = 0.78$, even though the misclassification rate converges to zero more slowly than for $p = 0.75$ and $p = 0.7$, the number of $\hat{\theta}_{pn}^{X}$ estimates that receive the wrong adjustment drops to zero very quickly. As shown in Table 2.2, as n increases, the mean squared error (MSE) of $\hat{\theta}_{pn}^{Xc}$ approaches zero for any p, indicating that $\hat{\theta}_{pn}^{Xc}$ is consistent. In particular, $\hat{\theta}_{pn}^{Xc}$ converges more quickly for $p = 0.78$ than for $p = 0.75$ and $p = 0.7$. Overall, the performance of $\hat{\theta}_{pn}^{Xc}$ is satisfactory when $n \geq 500$ and an appropriate $\delta(n)$ is used.

Table 2.2: Performance of the modified EL estimator $\hat{\theta}_{pn}^{Xc}$

		$p=0.75, \theta_p^X=8$		$p=0.7, \theta_p^X=7$		$p=0.78, \theta_p^X=8$	
$\delta(n)$	n	MEAN	MSE	MEAN	MSE	MEAN	MSE
$n^{-1/5}$	100	7.514	0.516	7.056	0.300	7.847	0.279
	500	7.921	0.079	7.150	0.150	8.015	0.017
	1000	7.989	0.011	7.107	0.107	8.002	0.002
	2000	8.000	0	7.094	0.094	8	0
	10000	8.000	0	7.078	0.078	8	0
$n^{-1/4}$	100	7.500	0.546	6.992	0.270	7.8	0.284
	*500	7.866	0.134	7.098	0.098	8.008	0.014
	1000	7.983	0.017	7.084	0.084	8.002	0.002
	2000	8.000	0	7.063	0.063	8	0
	10000	8.000	0	7.039	0.039	8	0
$n^{-1/3}$	100	7.447	0.581	6.938	0.272	7.749	0.291
	500	7.816	0.184	7.047	0.049	7.991	0.015
	*1000	7.967	0.033	7.036	0.036	8	0
	2000	8.000	0	7.030	0.030	8	0
	10000	8.000	0	7.022	0.022	8	0
$n^{-1/2}$	100	7.399	0.633	6.912	0.278	7.73	0.29
	500	7.677	0.323	7.016	0.020	7.974	0.028
	1000	7.928	0.072	7.010	0.010	8	0
	*2000	7.995	0.005	7.007	0.007	8	0
	10000	8.000	0	7.002	0.002	8	0
n^{-1}	100	7.404	0.634	6.903	0.237	7.706	0.322
	500	7.440	0.560	6.992	0.013	7.951	0.049
	1000	7.623	0.377	7.000	0	7.996	0.004
	2000	7.938	0.062	7.000	0	8.0	0
	*10000	8.000	0	7.000	0	8.0	0

Entries "MEAN" and "MSE" are the average and the mean squared error of $\hat{\theta}_{pn}^{Xc}$ over 1000 runs. Four choices of $\delta(n)$, *i.e.*, $n^{-1/4}$, $n^{-1/3}$, $n^{-1/2}$ and n^{-1}, have satisfactory and the most balanced performance for samples of sizes $n = 500$, 1000, 2000 and 10000, respectively. These compatible sample sizes are marked by $*$.

Choosing $\delta(n)$. Table 2.1 shows that a smaller $\delta(n)$ such as n^{-1} induces smaller misclassification rates for p in case (ii), while a larger $\delta(n)$ such as $n^{-1/5}$ performs better for p in case (i). This is not surprising, as the classification rule (2.18) will always put p into case (i) if $\delta(n) = 1$ (the upper limit) and into case (ii) if $\delta(n) = 0$ (the lower limit). Note that $n^{-1/5}$ and n^{-1} are the closest to 1 and 0, respectively, among all candidates in the simulation. Results in Table 2.1 are consistent with those in Table 2.2, from which one can observe that the MSE of $\hat{\theta}_{pn}^{Xc}$ shrinks to zero faster with a smaller $\delta(n)$ if p is in case (ii), but a larger $\delta(n)$ leads to a smaller MSE for a fixed sample size if p is in case (i). In this sense, there may not exist a $\delta(n)$ that is uniformly optimal for all p. Therefore, it is reasonable to choose a $\delta(n)$ that has balanced low MSE for p in the two cases, although there may be other criteria for choosing $\delta(n)$, as will be discussed in Section 2.5. According to the criterion of balanced performance, comparison among the five candidates of $\delta(n)$ is mixed in the sense that $n^{-1/5}$, $n^{-1/4}$, $n^{-1/3}$, $n^{-1/2}$ and n^{-1} have the best and most balanced MSEs for samples of sizes $n = 100$, 500,1000, 2000 and 10000, respectively. That is to say, the best choice of $\delta(n)$ depends on the sample size n. This is possibly due to the complexity of the underlying best $\delta(n)$, if it exists. If we require that the MSE be smaller than 0.15, then the sample size should be at least 500 for the uniform discrete distribution considered here. The performance of each of the five $\delta(n)$ candidates improves with large samples, with n^{-1} being the best if the sample size is as large as 10000.

Small and large values of p. We extend the above simulation study to $p = 0.05, 0.1, 0.9$ and 0.95 to investigate the performance of $\hat{\theta}_{pn}^{Xc}$ for p in the lower or upper tail of F_0^X. Simulation results for $p = 0.1$ and 0.9 are very similar to those for $p = 0.7$. Recall that, for p in case (ii), the proposed adjustment to $\hat{\theta}_{pn}^{X}$ is based on the fact that $\lim_{n\to\infty} \Pr(\hat{\theta}_{pn}^{X} = \theta_p \text{ or } \theta_p + 1) = 1$. How fast this probability converges to 1 is mainly determined by the probability mass F_0^X puts on θ_p and $\theta_p + 1$. Then it is natural to expect that the adjusted estimator $\hat{\theta}_{pn}^{Xc}$ behaves similarly for $p = 0.1, 0.7$ and 0.9, due to the uniformity of F_0^X. Results for $p = 0.05$ and $p = 0.95$ are actually much better than those for $p = 0.75$. For example,

when $p = 0.05$, $n = 100$ and $\delta(n) = n^{-1/5}$, the MSE of $\hat{\theta}^{Xc}_{pn}$ is only 0.002, which is also smaller than the MSE of $\tilde{\theta}_{pn}$, 0.018. Since $\theta^X_{0.05} = 1$ is at the left boundary of F^X_0, $\hat{\theta}^X_{0.05n}$ (or $\tilde{\theta}^X_{0.05n}$) cannot take any values below $\theta^X_{0.05n}$. In contrast, $\hat{\theta}^X_{0.75n}$ is likely to take any integer value between $\theta^X_{0.75n} - 2$ and $\theta^X_{0.05n} + 2$ if n is not very large. Thus, in 1000 replicates, $\hat{\theta}^X_{0.05n}$ takes fewer wrong values than $\hat{\theta}^X_{0.75n}$ for a fixed n, which explains why $\hat{\theta}^{Xc}_{pn}$ converges more quickly with $p = 0.05$ than with $p = 0.75$. A similar argument applies to $\hat{\theta}^{Xc}_{pn}$ with $p = 0.95$. When $\hat{\theta}^X_{0.05n}$ does take a wrong value (that must be greater than the true quantile), say $\theta^X_{0.05n} + 1$, $p = 0.05$ is often misclassified as a plateau and then $\hat{\theta}^X_{0.05n}$ is very likely to receive the correct adjustment so that the final estimate $\hat{\theta}^{Xc}_{pn}$ is often correct. This is why $\hat{\theta}^{Xc}_{pn}$ is also better than the sample quantile when $p = 0.05$.

To explore how $\hat{\theta}^{Xc}_{pn}$ behaves with small and large values of p for a more general discrete distribution, we also conduct simulations on the Poisson(2) and Poisson(20) distributions, both with unbounded upper tails. The former concentrates most of its probability mass on only a few points while the latter is much more scattered on a longer range of points. Table 2.3 displays the mean and MSE of $\hat{\theta}^{Xc}_{pn}$ for both distributions with $p = 0.1$ and $p = 0.9$, and also compares them to the performance of the sample quantile. When $p = 0.1$, $\hat{\theta}^{Xc}_{pn}$ converges quickly and outperforms the sample quantile $\tilde{\theta}^X_{pn}$ for both distributions. Note that $p = 0.1$ is very close to a plateau of the Poisson(20) distribution, so $\tilde{\theta}^X_{pn}$ (or $\hat{\theta}^X_{pn}$) converges very slowly (see Comment 2.3.3). When $p = 0.9$, the convergence of $\hat{\theta}^{Xc}_{pn}$ becomes slower, as the jumps between consecutive plateaus are smaller at the upper tails of both Poisson(2) and Poisson(20). Since Poisson(20) is more spread than Poisson(2), the convergence of $\hat{\theta}^{Xc}_{pn}$ is even slower for Poisson(20). It appears that $\hat{\theta}^{Xc}_{pn}$ is worse than $\tilde{\theta}^X_{pn}$ at the upper tail, if p is not a plateau of F^X_0. This is due to the trade-off between consistency and efficiency, which is more significant for p between two close plateaus. Typically the upper tail of F^X_0 is quite flat in the sense that it consists of smaller and smaller jumps, and hence $\hat{\theta}^X_{pn}$ varies among multiple values for n not very large. In this case, it is important to categorize p correctly

Table 2.3: $\hat{\theta}_{pn}^{Xc}$ for $p = 0.1$ and $p = 0.9$

		Poisson(2)				Poisson(20)			
		$p = 0.1, \theta_p^X = 0$		$p = 0.9, \theta_p^X = 4$		$p = 0.1, \theta_p^X = 14$		$p = 0.9, \theta_p^X = 26$	
$\delta(n)$	n	MEAN	MSE	MEAN	MSE	MEAN	MSE	MEAN	MSE
$n^{-1/5}$	100	0.012	0.012	3.556	0.444	13.840	0.536	25.176	1.368
	500	0	0	3.954	0.046	14.020	0.108	25.378	0.632
	1000	0	0	3.994	0.006	14.044	0.050	25.452	0.548
	2000	0	0	4	0	14.016	0.016	25.68	0.320
	10000	0	0	4	0	14.001	0.001	25.996	0.004
$n^{-1/4}$	100	0.013	0.013	3.444	0.558	13.877	0.523	25.207	1.337
	500	0	0	3.942	0.058	13.964	0.116	25.329	0.683
	1000	0	0	3.993	0.007	14.022	0.032	25.436	0.564
	2000	0	0	4	0	14.008	0.008	25.601	0.399
	10000	0	0	4	0	14.001	0.001	25.987	0.013
$n^{-1/3}$	100	0.005	0.005	3.471	0.529	13.823	0.531	25.205	1.287
	500	0	0	3.894	0.106	13.944	0.106	25.284	0.74
	1000	0	0	3.991	0.009	14.001	0.031	25.354	0.646
	2000	0	0	4	0	14.005	0.007	25.493	0.507
	10000	0	0	4	0	14	0	25.969	0.031
$n^{-1/2}$	100	0.003	0.003	3.374	0.626	13.798	0.534	25.178	1.354
	500	0	0	3.807	0.193	13.901	0.117	25.332	0.696
	1000	0	0	3.976	0.024	13.975	0.035	25.238	0.764
	2000	0	0	4	0	14	0	25.282	0.718
	10000	0	0	4	0	14	0	25.917	0.083
n^{-1}	100	0	0	3.333	0.667	13.837	0.533	25.205	1.339
	500	0	0	3.437	0.527	13.913	0.109	25.28	0.736
	1000	0	0	3.791	0.209	13.964	0.038	25.234	0.766
	2000	0	0	3.994	0.006	13.993	0.007	25.176	0.824
	10000	0	0	4	0	14	0	25.559	0.441
$\hat{\theta}_{pn}^X$	100	0.108	0.096	3.879	0.1485	14.316	0.579	25.723	0.763
	500	0.008	0.008	4	0	14.329	0.329	25.815	0.199
	1000	0	0	4	0	14.302	0.212	25.880	0.217
	2000	0	0	4	0	14.215	0.169	25.950	0.048
	10000	0	0	4	0	14.053	0.050	26	0

Entries "MEAN" and "MSE" are the average and the mean squared error of $\hat{\theta}_{pn}^{Xc}$ over 1000 runs. Note: for the Poisson(2) distribution, $\Pr(X \leq 0) = 0.135$, ..., $\Pr(X \leq 3) = 0.857$, and $\Pr(X \leq 4) = 0.947$; for the Poisson(20) distribution, $\Pr(X \leq 13) = 0.066$, $\Pr(X \leq 14) = 0.105$, ..., $\Pr(X \leq 25) = 0.887$, and $\Pr(X \leq 26) = 0.922$.

in order to avoid incorrect adjustment to $\hat{\theta}^X_{pn}$, but correct categorization of p requires the sample size to be very large.

2.3.5 Estimating the Probability of a Correct Estimate

When the underlying distribution is continuous, the maximum empirical likelihood estimator for a quantile (or the sample quantile estimator) is consistent in the sense that the estimator moves around the population quantile within a range shrinking gradually to zero as the sample size increases. In this case, a point estimate is often accompanied with a confidence interval at a desired confidence level. Now for F^X_0 discrete, the consistent quantile estimator $\hat{\theta}^{Xc}_{pn}$ hits θ^X_p with probability tending to 1 as $n \to \infty$. Unlike in the continuous case, constructing a confidence interval around $\hat{\theta}^{Xc}_{pn}$ for θ^X_p at a *given* confidence level is not feasible, because F^X_0 puts probability mass on discrete points. Therefore, we propose to estimate the probability of a correct estimate (PCE), *i.e.*, $\Pr(\hat{\theta}^{Xc}_{pn} = \theta^X_p)$. Note that, if we treat the point $\hat{\theta}^{Xc}_{pn}$ as a shrunken interval and write PCE as $\Pr(\{\hat{\theta}^{Xc}_{pn}\} \ni \theta^X_p)$, then PCE will have an interpretation similar to that of a confidence interval, namely the probability that $\{\hat{\theta}^{Xc}_{pn}\}$ covers the true parameter θ^X_p. PCE may also be used to measure how good an estimate $\hat{\theta}^{Xc}_{pn}$ is. If the PCE is high, then $\hat{\theta}^{Xc}_{pn}$ is very likely to estimate θ^X_p correctly.

To estimate PCE based on a single finite sample from a discrete distribution, one possible approach is to bootstrap. Applying the plug-in principle (Efron and Tibshirani 1993), we estimate $\Pr(\hat{\theta}^{Xc}_{pn} = \theta^X_p)$ by

$$\widehat{\Pr}(\hat{\theta}^{Xc*}_{pn} = \hat{\theta}^{Xc}_{pn}) = \frac{\sum_{b=1}^{B} 1(\hat{\theta}^{Xc*b}_{pn} = \hat{\theta}^{Xc}_{pn})}{B},$$

where $\hat{\theta}^{Xc*b}_{pn}$ is the estimate of θ^X_p based on (2.19) and the bth bootstrap sample $\mathbf{X}^{*b} = \{X^{*b}_i\}$, $\hat{\theta}^{Xc}_{pn}$ is obtained from the original sample $\{X_i\}$, and B is the number of bootstrap samples.

Table 2.4 shows the simulation results of bootstrapping PCE. The simulation setup is the same as in the first simulation study in Section 2.3.4. At each simulation run, a random sample is first drawn from F^X_0, $\hat{\theta}^{Xc}_{pn}$ is calculated, and then $\widehat{\Pr}(\hat{\theta}^{Xc*}_{pn} = \hat{\theta}^{Xc}_{pn})$ is obtained based on $B = 500$ bootstrap samples drawn from the original sample. Here, "mean" and "mse"

Table 2.4: Bootstrapping $\Pr(\hat{\theta}_{pn}^{Xc} = \theta_p^X)$

		$p = 0.75, \theta_p^X = 8$				$p = 0.7, \theta_p^X = 7$			
$\delta(n)$	n	true	est	mean	mse	true	est	mean	mse
$n^{-1/5}$	100	0.537	0.546	0.659	0.029	0.716	0.726	0.652	0.019
	500	0.907	0.906	0.855	0.023	0.869	0.876	0.838	0.021
	1000	0.990	0.992	0.951	0.009	0.886	0.881	0.839	0.023
	2000	1.000	1.000	0.996	0.000	0.902	0.903	0.847	0.024
	10000	1.000	1.000	1.000	0.000	0.923	0.925	0.867	0.023
$n^{-1/4}$	100	0.480	0.464	0.655	0.049	0.734	0.749	0.649	0.019
	*500	0.870	0.863	0.817	0.025	0.905	0.905	0.845	0.022
	1000	0.983	0.983	0.936	0.013	0.921	0.920	0.873	0.021
	2000	0.999	1.000	0.993	0.001	0.931	0.934	0.880	0.020
	10000	1.000	1.000	1.000	0.000	0.952	0.939	0.894	0.020
$n^{-1/3}$	100	0.444	0.472	0.666	0.062	0.747	0.729	0.654	0.020
	500	0.796	0.791	0.801	0.023	0.944	0.953	0.881	0.019
	*1000	0.967	0.966	0.915	0.016	0.960	0.959	0.909	0.016
	2000	0.999	1.000	0.987	0.002	0.966	0.964	0.914	0.016
	10000	1.000	1.000	1.000	0.000	0.978	0.982	0.933	0.013
$n^{-1/2}$	100	0.421	0.422	0.676	0.079	0.761	0.754	0.668	0.021
	500	0.670	0.679	0.768	0.031	0.977	0.974	0.903	0.018
	1000	0.911	0.920	0.861	0.022	0.987	0.986	0.945	0.011
	*2000	0.996	0.995	0.974	0.004	0.991	0.989	0.953	0.009
	10000	1.000	1.000	1.000	0.000	0.995	0.994	0.965	0.005
n^{-1}	100	0.419	0.443	0.679	0.082	0.763	0.763	0.673	0.020
	500	0.455	0.448	0.746	0.107	0.990	0.985	0.940	0.009
	1000	0.603	0.598	0.760	0.045	0.999	0.998	0.985	0.002
	2000	0.932	0.933	0.871	0.023	1.000	0.999	0.992	0.001
	*10000	1.000	1.000	1.000	0.000	1.000	1.000	0.996	0.000

The entry "true" is $\Pr(\hat{\theta}_{pn}^{Xc} = \theta_p^X)$ calculated using F_0^X; "est" is estimated PCE by $\#\{\hat{\theta}_{pn}^{Xc} = \theta_p^X\}/N$, where $N = 1000$ is the number of simulation runs; "mean" and "mse" are the average and the mean squared error of bootstrap estimates $\widehat{\Pr}(\hat{\theta}_{pn}^{Xc*} = \hat{\theta}_p^{Xc})$ over 1000 runs.

are the average and the mean squared error of $\widehat{\Pr}(\hat{\theta}_{pn}^{Xc*} = \hat{\theta}_p^{Xc})$ over 1000 runs. As one can observe, the bootstrap estimator is biased for a finite sample, and the speed at which the bias diminishes to zero varies with the choice of $\delta(n)$. The MSE approaches zero relatively fast when $\delta(n) = n^{-1/5}$ is chosen for $p = 0.75$ or $\delta(n) = n^{-1}$ is chosen for $p = 0.7$. When n is fixed and the type of p is unknown, $\delta(n) = n^{-1/5}, n^{-1/4}, n^{-1/3}, n^{-1/2}, n^{-1}$ perform best with sample sizes $n = 100$, 500, 1000, 2000 and 10000, respectively, which indicates that the results in Table 2.4 are consistent with those in Tables 2.1 and 2.2.

2.4 Applications

Table 2.5: Epileptic seizures ($n = 351$)

X	0	1	2	3	4	5	6	7	8
Frequency	126	80	59	42	24	8	5	4	3
$F_n(x)$	0.359	0.587	0.755	0.875	0.943	0.966	0.980	0.991	1.000
Estimate:	$\tilde{\theta}^X_{0.75n} = 2$		$\hat{\theta}^X_{0.75n} = 2$		$\hat{\theta}^{Xc}_{0.75n} = 2$				
		δ^{-1}	$\delta^{-1/2}$	$\delta^{-1/3}$	$\delta^{-1/4}$	$\delta^{-1/5}$			
Bootstrapped PCE:		0.989	0.999	1.000	1.000	1.000			

We present two examples to illustrate the application of the new estimator $\hat{\theta}_{pn}^{Xc}$. The first example is the epileptic seizure counts data discussed in Albert (1991), which exhibits significant overdispersion relative to a Poisson distribution. A patient with intractable epilepsy controlled by anti-convulsant drugs was observed for 351 days, and the patient's daily seizure counts were recorded. The observed counts are summarized in Table 2.5. The sample quantile $\tilde{\theta}^X_{0.75n}$ and the MELE $\hat{\theta}^X_{0.75n}$ are both 2. Note that 75.5% of the 351 counts are 2 or less, suggesting that the true distribution might have a plateau at level 0.75. The classification procedure (2.18) and $\hat{\theta}_{pn}^{Xc}$ are applied to the data for $p = 0.75$. As a result, $p = 0.75$ is classified into case (i) with any of the five choices of δ, and hence $\hat{\theta}^{Xc}_{0.75n}$ agrees with the sample quantile and the MELE in this example. All PCE estimates are very high, indicating that 2 is an accurate estimate of $\theta_{0.75}$.

Table 2.6: Counts of alpha-particles ($n = 2608$)

X	0	1	2	3	4	5	6
Frequency	57	203	383	525	532	408	273
$F_n(x)$	0.022	0.010	0.247	0.448	0.652	0.808	0.913
X	7	8	9	10	11	≥ 12	
Frequency	139	45	27	10	4	2	
$F_n(x)$	0.966	0.984	0.994	0.998	0.999	1.000	
Estimate:	$\tilde{\theta}^X_{0.25} = 3$		$\hat{\theta}^X_{0.25} = 3$		$\hat{\theta}^{Xc}_{0.25} = 2$		
		n^{-1}	$n^{-1/2}$	$n^{-1/3}$	$n^{-1/4}$		
Bootstrapped PCE:		0.999	0.979	0.920	0.870		

The second example is the classic data set from Rutherford and Geiger (1910), which consists of 2608 counts of scintillations caused by the radioactive decay of a quantity of the element polonium; all the counts were observed in 72-second intervals. Table 2.6 shows the count frequencies. Here, $\theta_{0.25}$ is of interest; both $\tilde{\theta}^X_{0.25n}$ and $\hat{\theta}^X_{0.25n}$ provide the same estimate, 3. However, $p = 0.25$ is classified into case (ii) by (2.18), and $\hat{\theta}^{Xc}_{0.75n} = 2$. Since the data are concentrated on only a few points and the sample size is large, $\delta = n^{-1}$ or $n^{-1/2}$ yield better estimates of PCE than other choices of δ. Note that the data agree excellently with the Poisson distribution with mean 3.87 (goodness of fit $\chi^2 = 12.99$ with degrees of freedom 11, p-value = 0.30), as shown in Figure 2.7. The 0.25th quantile of Poisson(3.87) is 2, which confirms that $\hat{\theta}^{Xc}_{0.75n} = 2$ is a better estimate of $\theta_{0.25}$.

2.5 Practical Issues

2.5.1 Sample Size

For the discrete uniform distribution considered above, in order for the proposed classification procedure and the quantile estimator $\hat{\theta}^{Xc}_{pn}$ to perform fairly well, the sample size n needs to be at least 500. We also examined the Poisson(2) and the Poisson(20) distributions. For Poisson(2), a balanced and acceptable correct-classification rate of 80% can be attained by

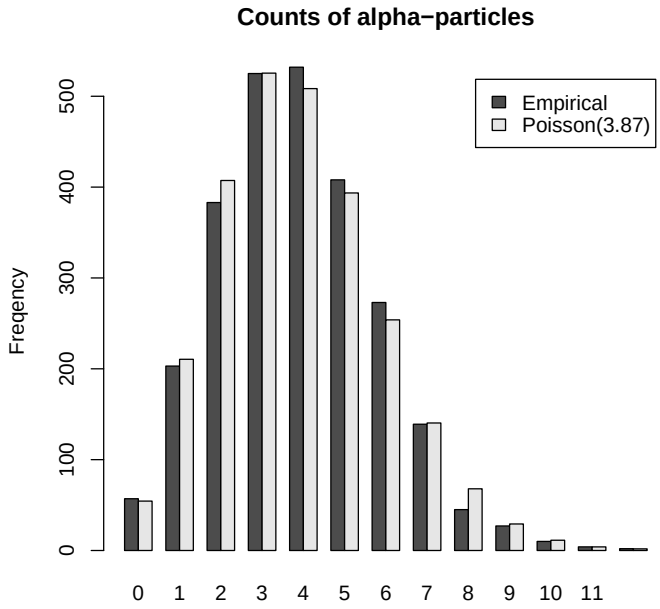

Figure 2.7: Counts of alpha-particles

a sample of size $n = 150$, while for Poisson(20), the sample size needs to be 1000 to achieve approximately the same performance. In any case, the sample size should be much larger than what is typically needed in other contexts. This high requirement of the sample size is mainly due to the discreteness of the underlying distribution. In the discrete case, it usually takes a fairly large sample for the MELE of a quantile (or the sample quantile) to reach a relatively stable status. A very large sample is necessary if the distribution is scattered or p is close to a plateau of F_0^X (as discussed in Comment 2.3.3). For this reason, Gonzále-Barrios and Rueda (2001) set the sample size to 1000, and Machado and Santos Silva (2005) used sample sizes of 500 and 2000 in their simulation studies. Perhaps, this is the price we have to pay for discrete data that are usually less tractable than continuous data. We can conjecture that, for a very scattered discrete distribution (i.e., the probability mass on each point is very small), $\hat{\theta}_{pn}^{Xc}$ performs poorly unless n is extremely large. In this seemingly problematic

situation, a quantile estimate off by 1 will not have much negative impact, and one can usually treat the data as if they were from a continuous distribution and use $\hat{\theta}_{pn}^{X}$ or $\tilde{\theta}_{pn}^{X}$ directly.

2.5.2 Choice of $\delta(n)$

The choice for $\delta(n)$ depends on the sample size, and consistently affects the performance of the EL-based classification procedure, $\hat{\theta}_{pn}^{Xc}$ and the bootstrap estimator of PCE. In our simulation example, $n^{-1/5}, n^{-1/4}, n^{-1/3}, n^{-1/2}$ and n^{-1} seem to be the best choice for $\delta(n)$ when $n = 100, 500, 1000, 2000, 10000$, respectively. However, the dependence of the best $\delta(n)$ on the sample size may vary with the underlying distribution. To explore this, we also compared Poisson(2) with Poisson(20). The performance of $\hat{\theta}_{pn}^{Xc}$ does not vary much for p in case (ii), even though F_0^X changes from the concentrated Poisson(2) to the scattered Poisson(20). If p is in case (i), for each $\delta(n)$ candidate, $\hat{\theta}_{pn}^{Xc}$ converges more quickly with a concentrated distribution than with a scattered distribution. The consequence is that the most balanced $\delta(n)$ corresponding to a particular sample size varies from distribution to distribution, meaning that there is no simple rule to choose $\delta(n)$ for the data at hand. For Poisson(2), $\delta(n) = n^{-1/4}$ and $n^{-1/2}$ are compatible with $n = 100$ and 500, respectively, and n^{-1} is the best when $n \geq 1000$. In contrast, for Poisson(20), $\delta(n) = n^{-1/5}$, $n^{-1/4}$ and $n^{-1/2}$ perform best with $n = 1000, 2000$ and 10000, respectively. Therefore, when the sample size is fixed, we can expect that a relatively small $\delta(n)$ is needed if there is evidence (*e.g.*, many ties in the data) that the underlying distribution is quite concentrated.

In practice, the chance that the p of interest is a plateau of F_0^X is generally very small. From this point of view, balanced performance of $\delta(n)$ for the two types of p is not important, and hence other criteria for choosing $\delta(n)$ may be considered. Suppose that the true PCE $\Pr(\hat{\theta}_{pn}^{Xc} = \theta_p^X)$ can be obtained for any $\delta(n)$. Then we can choose the $\delta(n)$ leading to the highest PCE. The logic is simply that the best $\delta(n)$ should yield an estimate that is most likely to be the true quantile. As one can observe from Tables 2.1, 2.2 and 2.4, for any fixed

n, if p is not a plateau of F_0^X, a larger $\delta(n)$ has a higher PCE and also corresponds to better performance of the classification procedure and the estimator $\hat{\theta}_{pn}^{Xc}$; if p is a plateau of F_0^X, a smaller $\delta(n)$ has a higher PCE and also corresponds to better performance of the procedure and the estimator. Consequently, the criterion of maximizing PCE automatically ensures the best performance of the classification procedure and $\hat{\theta}_{pn}^{Xc}$. However, the true PCE is not available since it requires detailed information about F_0^X. One may use bootstrap estimates of PCE in practice. Because bias and variation of the bootstrap estimator exist for finite samples and diminish to zero as $n \to \infty$, the chance that the best $\delta(n)$ will be chosen is not 100% for a real sample, but will increase as n increases.

Using the same discrete uniform distribution as above, we simulated the process of choosing $\delta(n)$ by the criterion of maximizing bootstrap estimates of PCE, and found that the best $\delta(n)$ for $p = 0.7$ was chosen 64% of the time when $n = 500$ and 80% of the time when $n = 1000$; similarly, the best $\delta(n)$ for $p = 0.75$ was chosen 60% of the time when $n = 500$ and 74% of the time when $n = 1000$. Since the two criteria for choosing $\delta(n)$ are not perfect, some care is needed when they are used in practice.

2.5.3 Jitters

When the data $\{X_i\}$ are drawn from a discrete distribution, jittering is useful for us to properly define $\mathcal{R}^X(\theta)$ and the MELE $\hat{\theta}_{pn}^X$. Another point of jittering is that it transforms $\{X_i\}$ to a new set of continuous data $\{Y_i\}$, so that we can study the consistency property of $\hat{\theta}_{pn}^X$ more easily using the results for $\hat{\theta}_{pn}^Y$. Under the assumption that the support of F_0^X consists of consecutive integer values, our choice of jitters $Z_i \overset{iid}{\sim} U(0, 1]$ is simple but not necessary. Random variables from any Beta distribution will serve the same purpose of constructing a continuous distribution F_0^Y, except that F_0^Y may not be piecewise linear. If the support of F_0^X contains unevenly spaced values, to ensure that the jittered variable Y has a continuous distribution, Z_i may be generated from distributions supported on different lengths of intervals, *i.e.*, Z_i may be non-identically distributed. The use of non-uniform

or non i.i.d. jitters will not affect the derivation of our results for $\hat{\theta}^X_{pn}$, although it may affect the shape of F^Y_0. $\hat{\theta}^X_{pn}$ and hence $\hat{\theta}^{Xc}_{pn}$ are actually invariant to the choice of jitters. Since jittering should preserve the order of the data (*i.e.*, $Y_{(i)}$ must be generated from $X_{(i)}$ even though $X_{(i)}$ may be tied with $X_{(i-1)}$ or $X_{(i+1)}$, and $Y_{(i)}$ recovers $X_{(i)}$ after the inverse transformation), $\hat{\theta}^X_{pn}$ must be $X_{(i^*)}$ with $i^* = np, [np]$ or $[np]+1$, no matter what specific jitters are used. In practice, to get estimates $\hat{\theta}^X_{pn}$ and $\hat{\theta}^{Xc}_{pn}$, one can skip the jittering step, and find i^* using (2.11) and then the smallest index L and the largest index U such that $X_{(L)} = X_{(i*)} = X_{(U)}$. For instance, if $n = 500$ and $p = 0.5$, then $\hat{\theta}^X_{pn} = X_{(250)}$; if we also know that $X_{(240)} < X_{(241)} = \ldots = X_{(250)} = \ldots = X_{(255)} < X_{(256)}$, say, then h_l and h_u can be calculated by plugging in 241 for L and 255 for U in (2.16); finally, $\hat{\theta}^{Xc}_{pn}$ can be determined based on the values of h_l and h_u.

2.6 Conclusion

The quantile estimator $\hat{\theta}^{Xc}_{pn}$ has been shown to be consistent for θ^X_p, when the underlying distribution is discrete. Although its consistency may cost significant amount of efficiency if p is not at a plateau of the true distribution, $\hat{\theta}^{Xc}_{pn}$ is particularly useful for consistently estimating quantiles θ^X_p for fairly concentrated discrete distributions. The invariance of $\hat{\theta}^X_{pn}$ to jittering contrasts to the dependence of Machado and Santos Silva's (2005) quantile regression coefficients on jitters, and makes practical quantile estimation very convenient. Compared to González-Barrios and Rueda's approach that involves calculating sample quantiles for a long sequence of sub-samples, the estimator $\hat{\theta}^{Xc}_{pn}$ is computationally easy to obtain once $\delta(n)$ is chosen; this advantage is even more prominent when the sample size is large. With a sample of size 1000, our estimator $\hat{\theta}^{Xc}_{pn}$ correctly estimates the quantile $\theta^X_{0.7}$ of the discrete uniform distribution at a rate of at least 0.92 (Table 2.1), which is much better than the performance of González-Barrios and Rueda's method. In our preliminary simulation study, the correct-estimation rate of their method reaches approximately 0.70 at $n = 1000$ but fails to go up for n up to 10000. Although we consider the performance of $\hat{\theta}^{Xc}_{pn}$ to be satisfactory,

it may be further improved by developing a sophisticated criterion for choosing $\delta(n)$. It is also of significance to improve the associated PCE estimator, as this not only provides more information about the data but also helps in choosing the best $\delta(n)$. A more accurate estimator for PCE might be obtained by modifying the simple bootstrap procedure, which could be a future research direction.

Chapter 3

Selection of Working Correlation Structure in GEE via Empirical Likelihood

3.1 Introduction

Longitudinal data consist of measurements taken repeatedly through time on a sample of subjects (e.g., human patients and animals) and a set of covariates for each subject. In contrast to cross-sectional studies in which the response of each individual is measured at a single occasion, longitudinal studies allow researchers to study changes over time and to evaluate treatments that influence these changes, and hence play a prominent role in medical, pharmaceutical and behavioral sciences.

Among various models for longitudinal data, the approach of generalized estimating equations (GEE) proposed by Liang and Zeger (1986) has become increasingly important and popular, due to its many attractive features. This approach is appropriate for longitudinal data in general, and is especially useful in dealing with Non-Gaussian longitudinal data that are often encountered in practice. It is not a likelihood-based method, and requires only a partial specification of the joint distribution of the repeated measurements within the same subject, and hence is more robust than a fully parametric model. Estimation in GEE models is computationally easier than in likelihood models, such as generalized linear mixed-effects models that usually require numerical or Monte Carlo integration.

While the GEE approach enjoys advantages of semi-parametric methods, it is also limited by its lack of a likelihood. It is known that likelihood methods are very effective in finding efficient estimators, constructing tests with good power properties and short confidence intervals (or small confidence regions), and selecting the best model from a pool of candidates. Combining the reliability of nonparametric methods with the flexibility and

effectiveness of likelihood approaches, empirical likelihood has the potential to add value to longitudinal data settings, in particular GEE models, as the semi-parametric character of the GEE approach matches well the philosophy of empirical likelihood.

This chapter explores the use of empirical likelihood to longitudinal data analysis with GEE models. The particular focus is on improving efficiency of the GEE estimator. First, a brief review of longitudinal data and the method of GEE is provided in Section 3.2. Next, Section 3.3 presents existing methods for improving the GEE estimator, including the quadratic inference function approach and model selection methods for GEE. After our investigation of these methods, it becomes clear that an effective way of improving the estimation efficiency within the GEE framework is to select among competing GEE models the one that assumes the correct working correlation structure for repeated measurements. Thus, our focus is further narrowed down to applying empirical likelihood to the problem of model selection in the context of GEE. Before proceeding to empirical likelihood-based estimation and model selection for GEE models, we review in Section 3.4 some existing work on empirical likelihood applied to regression models and to dependent data. We discuss in Section 3.5 using empirical likelihood to choose the best GEE model, in particular the best working correlation structure. Finally, Section 3.6 summarizes our conclusions and proposes some future directions.

3.2 Longitudinal Data and Generalized Estimating Equations

In a longitudinal data set, $Y_i = (Y_{i1}, \ldots, Y_{it_i})^T$ denotes the response vector of the ith subject, where Y_{ij} is the response observed at the jth time point on subject i, $i = 1, \ldots, n$ and $j = 1, \ldots, t_i$; $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^T$ is the $p \times 1$ vector of covariates associated with Y_{ij}, and $X_i = (X_{i1}, \ldots, X_{it_i})^T$ is the $t_i \times p$ design matrix for the ith subject. The vector X_{ij} may include two types of covariates: covariates whose values do not change throughout the duration of the study and those whose values change over time. Examples of the former

include sex and fixed experimental treatments, and examples of the latter include time since baseline and current smoking status.

Measurements taken on the same subject at different times form a cluster, and typically exhibit positive correlation which must be accounted for in the analysis. A number of models and methods have been developed to analyze longitudinal data. The type of the response variable is an important characteristic that dictates to some extent the choice of models. If the response variable is continuous and has an approximately symmetric distribution so that it is reasonable to assume normality, *marginal multivariate linear models* and *linear mixed-effects models* are common choices, depending on the research questions. In biomedical applications, the response variable is often binary or a count, for instance the presence or absence of some particular illness, and the number of epileptic seizures in a four-week interval. For non-Gaussian response variables, general models that are different extensions of the generalized linear model have been developed. Among those models (extensions), *generalized estimating equations* are a popular class of marginal models (marginalized over subject-specific random effects) .

The method of generalized estimating equations was proposed by Liang and Zeger (1986), and is denoted by GEE in the literature. The abbreviation GEE is not to be confused with general estimating equations as used in Qin and Lawless (1994) to define empirical likelihood for general parameters. Used to evaluate the *population-averaged* effects of treatments, GEE requires only a partial specification of the marginal distribution for the response, which is an appealing feature as there are few tractable multivariate distributions for non-Gaussian data. The GEE approach is based on the idea of the quasilikelihood for a generalized linear model and the concept of estimating equations. In particular, the following are assumed in GEE:

(1) The marginal mean of the response of subject i measured at time point j, $E(Y_{ij}) = \mu_{ij}$, depends upon the covariates X_{ij} through a known link function $h(\mu_{ij}) = \eta_{ij} = X_{ij}^T\beta$, where β is the p-dimensional parameter of interest.

(2) The marginal variance $\text{var}(Y_{ij})$ is assumed to depend on the marginal mean according to $\text{var}(Y_{ij}) = \phi v(\mu_{ij})$, where $v(\cdot)$ is a known variance function and ϕ is a possibly unknown scale parameter.

(3) A $t_i \times t_i$ *working* correlation matrix $R(\alpha)$ is assumed for the random vector Y_i, and $R(\alpha)$ is parameterized by $\alpha = (\alpha_1, \ldots, \alpha_s)^T$, a vector of nuisance parameters. The corresponding *working* covariance matrix is

$$V_i(\beta, \alpha, \phi) = A_i^{1/2}(\beta, \phi) R(\alpha) A_i^{1/2}(\beta, \phi), \tag{3.1}$$

where$A_i(\beta, \phi)$ is a diagonal matrix with $\text{var}(Y_{ij}) = \phi v(\mu_{ij})$, $j = 1, \ldots, t_i$, along the diagonal. Note that $V_i = \text{var}(Y_i)$ if $R(\alpha)$ is chosen correctly.

Then the generalized estimating equation for β is defined as

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta^T} \right)^T V_i^{-1}(\beta, \alpha, \phi)(Y_i - \mu_i) = 0, \tag{3.2}$$

where $\mu_i = (\mu_{i1}, \ldots, \mu_{it_i})^T$. The parameter α in the working correlation matrix is treated as a nuisance. The left-hand side of (3.2) can be written as a function of β alone given the data,

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta^T} \right)^T V_i^{-1}(\beta, \alpha, \phi)(Y_i - \mu_i) = \sum_{i=1}^{n} U_i[\, \beta, \hat{\alpha}(\beta, \hat{\phi}(\beta)\,)\,],$$

since ϕ in $V_i^{-1} = A_i^{-1/2}(\beta, \phi) R^{-1}(\alpha) A_i^{-1/2}(\beta, \phi)$ can be replaced by a $n^{1/2}$-consistent estimator $\hat{\phi}(\beta)$ given β, and α by a $n^{1/2}$-consistent estimator $\hat{\alpha}(\beta, \phi)$ given β and ϕ. Typically, method of moment estimators of ϕ and α are used. The solution to equation (3.2), $\hat{\beta}_G$, is the GEE estimator of the regression parameter β.

Liang and Zeger (1986) proved that, under mild regularity conditions, $\hat{\beta}_G$ is consistent and asymptotically normal:

$$n^{1/2}(\hat{\beta}_G - \beta) \xrightarrow{d} \mathcal{N}\,(0, V_G)\,, \tag{3.3}$$

where

$$V_G = \lim_{n \to \infty} n \left(\sum_{i=i}^{n} D_i^T V_i^{-1} D_i \right)^{-1} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} \text{cov}(Y_i) V_i^{-1} D_i \right) \left(\sum_{i=i}^{n} D_i^T V_i^{-1} D_i \right)^{-1}, \tag{3.4}$$

and $D_i = \partial\mu_i/\partial\beta^T$. Another important advantage of the GEE approach is that the consistency and asymptotic normality of $\hat{\beta}_G$ hold whether or not the working correlation R is correctly specified. Thus, valid inference for β can be made via the generalized Wald statistic that utilizes the asymptotic normality of $\hat{\beta}_G$, regardless of what working correlation is used. If R is correct so that $\text{var}(Y_i) = V_i$, then V_G will reduce to

$$V_G^* = \lim_{n\to\infty} \left(\frac{1}{n} \sum_{i=i}^{n} D_i^T V_i^{-1} D_i \right)^{-1}.$$

V_G^* is the optimal asymptotic variance that can be achieved by the class of estimating equations

$$\left\{ H^T \left(Y - \mu(\beta) \right) = 0 \mid H_{N\times p} \text{ does not involve } Y \right\}, \tag{3.5}$$

where $Y = (Y_1^T, \ldots, Y_n^T)^T$, $\mu = (\mu_1^T, \ldots, \mu_n^T)^T$, and $N = \sum_{i=1}^{n} t_i$ (McCullagh and Nelder, 1989, page 347-8; Heyde, 1997, page 25-6). Note that GEE is a member of class (3.5), since (3.2) can be rewritten as $D^T V^{-1}(Y - \mu) = 0$, where $D = (D_1^T, \ldots, D_n^T)^T$ and $V = \text{diag}\{V_1, \ldots, V_n\}$. The optimality is in the sense that, if $\tilde{\beta}$ is any estimator that solves an estimating equation in class (3.5), then $\text{var}[n^{1/2}(\tilde{\beta} - \beta)] - V_G^*$ is nonnegative definite, at least asymptotically. Thus, the full efficiency of $\hat{\beta}_G$ is gained when R is correct.

Since there is no associated likelihood or objective function in the GEE method, inference for β is limited to using the generalized Wald's statistic, and no goodness-of-fit statistic is available. It has been reported that Wald's statistic might behave poorly even for large samples (Moolgavkar & Venzon, 1986). Therefore, some modifications leading to alternative test statistics have been proposed. The generalized score statistic was considered by Rotnitsky & Jewell (1990) and Boos (1992). Rotnitsky & Jewell (1990) also proposed adjusted Wald and score statistics that are asymptotically equivalent to a weighted sum of independent chi-squared variables under the null. Hanfelt and Liang (1995) presented two ways of constructing approximate likelihood ratios, one based on quasi-likelihood and the other based on linear projection.

While it is appealing that the GEE estimator $\hat{\beta}_G$ is consistent even if the working correlation R differs from the true underlying structure, there is a price to pay in terms of efficiency. For example, the independence structure is seemingly the simplest working assumption that can be adopted in all cases, but, for time-varying covariates, the resulting efficiency of the GEE estimator may be as low as 60% compared to the GEE estimator obtained by using the correct correlation structure (Fitzmaurice, 1995). In general, it is more efficient to use an R closer to the true correlation. Then a practical question is "How to decide whether a working correlation structure is closer to the truth, compared to other alternatives?" The original GEE approach is not helpful in answering this question: working correlations are merely a device to provide consistent and asymptotically normal estimates for the regression parameters, and do not have associated standard errors, thus making Wald-type tests unavailable; likelihood-ratio tests cannot be used since GEE is not a likelihood method; score tests are not easy to use.

3.3 Methods for Improving the GEE Estimator

3.3.1 Quadratic Inference Function

Many authors have attempted to improve the efficiency of GEE estimators. Qu *et al.* (2000) applied the principle of generalized method of moments (GMM, Hansen, 1982) in the GEE framework to construct a quadratic inference function (QIF) for β. Their main idea is as follows. First, the inverse of the working correlation matrix $R^{-1}(\alpha)$ is assumed to be a linear combination of basis matrices in the form

$$R^{-1}(\alpha) = \sum_{l=1}^{m} a_l M_l, \tag{3.6}$$

where m and the set of matrices $\{M_l\}$ depend on the particular choice of $R(\alpha)$, and $\{a_l\}$ are functions of α. For instance, if $R(\alpha)_{m\times m}$ has an exchangeable structure (*i.e.*, its diagonal elements are all equal to 1 and off-diagonal elements are all equal to α), then its inverse can be represented as $R^{-1}(\alpha) = a_1 M_1 + a_2 M_2,$, with M_1 being the identity matrix I, M_2 being the matrix with 0 on the diagonal and 1 off the diagonal, $a_1 = -[(m-2)\alpha+1]/[(m-1)\alpha^2 -$

$(m-2)\alpha-1]$, and $a_2=\alpha/[(m-1)\alpha^2-(m-2)\alpha-1]$. Second, instead of the estimation of $\{a_l\}$ in 3.6, an extended estimating function is defined to be

$$g_e(Y_i,X_i,\beta)=\begin{pmatrix} \left(\frac{\partial\mu_i}{\partial\beta^T}\right)^T A_i^{-1/2}M_1A_i^{-1/2}(Y_i-\mu_i) \\ \vdots \\ \left(\frac{\partial\mu_i}{\partial\beta^T}\right)^T A_i^{-1/2}M_mA_i^{-1/2}(Y_i-\mu_i) \end{pmatrix}_{mp\times 1}, \tag{3.7}$$

and then the QIF

$$Q_n(\beta)=n\bar{g}_e^T(\beta)C_n^{-1}\bar{g}_e(\beta),$$

where $\bar{g}_e(\beta)=(1/n)\sum_{i=1}^n g_e(Y_i,X_i,\beta)$ and $C_n=(1/n)\sum_{i=1}^n g_e(Y_i,X_i,\beta)g_e^T(Y_i,X_i,\beta)$. Last, the QIF estimator is obtained by minimizing the objective function $Q_n(\beta)$, *i.e.*,

$$\hat{\beta}_Q=\arg\min_{\beta\in\mathbb{R}^p} Q_n(\beta).$$

Note that $\hat{\beta}_Q$ is the solution of

$$\frac{\partial Q_n(\beta)}{\partial\beta}=2n\frac{\partial\bar{g}_e^T(\beta)}{\partial\beta}C_n^{-1}\bar{g}_e(\beta)-n\bar{g}_e^T(\beta)C_n^{-1}\frac{\partial C_n}{\partial\beta}C_n^{-1}\bar{g}_e(\beta)=0_{p\times 1}. \tag{3.8}$$

Qu *et al.* pointed out that the second term of $\partial Q_n(\beta)/\partial\beta$ in (3.8) is $O_p(n^{-1})$, and hence solving (3.8) is asymptotically equivalent to solving

$$n\frac{\partial\bar{g}_e^T(\beta)}{\partial\beta}C_n^{-1}\bar{g}_e(\beta)=0_{p\times 1}. \tag{3.9}$$

They further argued that, by estimating equation theory, (3.9) is optimal among the class of estimating equations

$$\sum_{l=1}^m H_l\sum_{i=1}^N\left(\frac{\partial\mu_i}{\partial\beta^T}\right)^T A_i^{-1/2}M_lA_i^{-1/2}(Y_i-\mu_i)=0, \tag{3.10}$$

in the sense that the asymptotic variance of the solution to (3.9) reaches the minimum among all estimating equations in this class. Here, H_l are arbitrary $p\times p$ nonrandom matrices. If $H_l=a_lI$, then (3.10) becomes

$$\sum_{i=1}^n\left(\frac{\partial\mu_i}{\partial\beta^T}\right)^T A_i^{-1/2}(a_1M_1+\ldots+a_mM_m)A_i^{-1/2}(Y_i-\mu_i)=0,$$

which is exactly the GEE with the working correlation structure $R(\alpha)$. Since GEE is a member of the class in (3.10), Qu *et al.* concluded that, with the same incorrect choice of $R(\alpha)$, the QIF estimator $\hat{\beta}_Q$ is more efficient than the GEE estimator $\hat{\beta}_G$, and $\hat{\beta}_Q$ will be as efficient as $\hat{\beta}_G$ if $R(\alpha)$ is correct, as it is already known that the GEE estimator is fully efficient with the true correlation structure. Qu *et al.* also performed a simulation study to support this conclusion.

The QIF method sounds very appealing, since the estimator is claimed to have improved efficiency if $R(\alpha)$ is misspecified. The extended estimating equation in (3.7) is especially interesting to us, as it may be used to define empirical likelihood for β. Qin and Lawless (1994) showed that the MELE will have smaller asymptotic variance if additional information is formulated into added components of the estimating function. For these reasons, we investigate the QIF method in detail.

However, we get dramatically different results after replicating exactly the same simulation study as in Qu *et al.* (2000). In particular, the model is assumed to be

$$E(Y_{ij}) = \mu_{ij} = X_{ij}^T\beta \ \text{ and } \ Y_i = X_i\beta + \varepsilon_i,$$

where $X_{ij}^T = (X_{ij1}, X_{ij2})$ and $\beta = (\beta_1, \beta_2)^T = (1, 1)^T$. For $i = 1, \ldots, 20$ (i.e., $n = 20$) and $j = 1, \ldots, 10$ (i.e., $t_i \equiv t = 10$), X_{ij1} and X_{ij2} are generated independently from $N(0.1 \times j, 1)$, and ε_i is generated from $N_{10}(0, \Sigma_0(\alpha))$, with $\Sigma_0(\alpha) = R_0(\alpha)$, where $R_0(\alpha)$ has either an exchangeable correlation structure or an AR-1 structure, and $\alpha = 0.7$. To ensure correct results, much care is taken in our simulation procedure, which is written in R. For each simulated data set, the GEE estimate generated from our R program is found to agree with the one generated from the GENMOD procedure of SAS. Both the grid search algorithm and the Newton-Raphson algorithm are used to obtain QIF estimates, and the two sets of QIF estimates are identical. Table 3.1 shows our results obtained from 1000 replicate simulation runs, along with the results from Qu *et al.* (2000). Entries "GEE" and "QIF" are the mean squared errors (MSEs) of $\hat{\beta}_G$ and $\hat{\beta}_Q$, respectively; "SRE" is the simulated relative efficiency, *i.e.*, the ratio of $MSE(\hat{\beta}_G)$ to $MSE(\hat{\beta}_Q)$. If the relative efficiency is larger than 1, then it

verifies that $\hat{\beta}_Q$ is more efficient than $\hat{\beta}_G$. Our results indicate that $\hat{\beta}_Q$ is no better than $\hat{\beta}_G$ in any situation, which contradicts the conclusion of Qu *et al.* that $\hat{\beta}_Q$ has improved efficiency if $R(\alpha)$ is specified incorrectly. In particular, when the true correlation structure R_0 is exchangeable with $\alpha = 0.7$ but the working structure R is AR-1, our SRE is 0.86 in contrast to 2.07 reported by Qu *et al.*; when R_0 is actually AR-1 with $\alpha = 0.7$ but R is exchangeable, our SRE is 0.90 instead of 1.34 in Qu *et al.* (2000).

Table 3.1: Comparison between the GEE estimator and the QIF estimator ($n = 20$)

	Results reproduced						Results of Qu *et al.*	
	Working R						Working R	
True R_0	EX			AR-1			EX	AR
	GEE	QIF	SRE	GEE	QIF	SRE	SRE	SRE
EX	0.00437	0.00560	0.78	0.00296	0.00342	0.86	0.99	2.07
AR-1	0.00756	0.00839	0.90	0.00377	0.00518	0.73	1.34	0.98

Both the true exchangeable and the true AR-1 correlation structures are parameterized by $\alpha = 0.7$. Entries "GEE" and "QIF" are the mean squared errors (MSEs) of $\hat{\beta}_G$ and $\hat{\beta}_Q$, respectively, which are obtained from our 1000 simulation runs. "SRE" is the simulated ratio of $MSE(\hat{\beta}_G)$ to $MSE(\hat{\beta}_Q)$.

While the simulation procedure in Qu *et al.* (2000) is questionable, their argument is also flawed. First, as also indicated in Pilla and Loader (2006), the second term of $\partial Q_n(\beta)/\partial\beta$ in (3.8) is of order $O_p(1)$, rather than $O_p(n^{-1})$. This can be verified by checking that $\sqrt{n}\bar{g}_e(\hat{\beta}_Q) = O_p(1)$, $C_n \xrightarrow{p} Var[g_e(Y, X, \beta)]$ and

$$\frac{\partial C_n}{\partial \beta_k} \xrightarrow{p} E\left[g_e(Y, X, \beta)\frac{\partial g_e^T(Y, X, \beta)}{\partial \beta_k} + \frac{\partial g_e(Y, X, \beta)}{\partial \beta_k} g_e^T(Y, X, \beta)\right], \text{ for } k = 1, \ldots, p.$$

Thus for small samples, $\hat{\beta}_Q$ that solves (3.8) differs from the solution to the optimal estimating equation (3.9). It should be noted that the difference is ignorable if the sample size is very large, as the first term in $\partial Q_n(\beta)/\partial\beta$ has the order $O_p(\sqrt{n})$ and hence will become dominant for large n. Second, it should be emphasized that the so-called optimal estimating equation (3.9) is optimal only in the asymptotic sense. We now denote the class of (3.10) by $\mathcal{C}$. By the theory of estimating equations (for example, Heyde, 1997), the optimal estimating equation

in the class $\mathcal{C}$ is

$$nE\left[\frac{\partial g_e^T(Y,X,\beta)}{\partial\beta}\right][Var(g_e(Y,X,\beta))]^{-1}\,\bar{g}_e(\beta)=0_{p\times 1},$$

which is the limit of (3.9) as $n\to\infty$. The asymptotic optimality of (3.9) explains in part the poor performance of $\hat{\beta}_Q$ for small samples. Third (and most important), even though $\hat{\beta}_Q$ is asymptotically optimal in the class $\mathcal{C}$, it does not necessarily mean that $\hat{\beta}_Q$ is asymptotically superior to $\hat{\beta}_G$. To reach the conclusion that $\hat{\beta}_Q$ is indeed better than $\hat{\beta}_G$, one also needs to show that $\hat{\beta}_G$ is *not* also optimal in class $\mathcal{C}$.

In order to compare $\hat{\beta}_Q$ and $\hat{\beta}_G$ for large samples, we provide additional simulation results in Table 3.2. Here, the sample size n is increased to 200 and 1000, respectively. "OEF" represents the MSE of $\hat{\beta}_O$, where $\hat{\beta}_O$ denotes the solution to the asymptotic optimal estimating equation (3.9). Numerical results confirmed that $\hat{\beta}_Q$ and $\hat{\beta}_O$ are equivalent in the asymptotic sense. Also note that, when the sample size becomes large, the difference between the GEE estimator $\hat{\beta}_G$ and $\hat{\beta}_O$ also shrinks significantly, leading the relative efficiency to be very close to 1. One may conjecture based on these numerical results that $\hat{\beta}_G$ is also optimal in class $\mathcal{C}$, although this statement is still to be tested by rigorous theoretical proof. Recall that class $\mathcal{C}$ is defined by the set of $\{M_l, l=1,\ldots,m\}$, which is in turn determined by the choice of $R(\alpha)$. Therefore, the optimality is constrained by the working correlation assumption. If the optimality of this class is actually achieved when $H_l = a_l I$ for $l=1,\ldots,m$, then $\hat{\beta}_G$ is also optimal in class $\mathcal{C}$ and cannot be outperformed by $\hat{\beta}_Q$ under the same assumption about $R(\alpha)$. It is interesting to note that this hypothesis seems to be self-evident when $R(\alpha)=R_0$; in this situation, $\hat{\beta}_G$ is known to be fully efficient in the class (3.5) (as discussed in Section 3.2), and is also optimal in class $\mathcal{C}$ since $\mathcal{C}$ is a subclass of (3.5).

One may hope to improve the efficiency of $\hat{\beta}_Q$ by "expanding" the optimality of class $\mathcal{C}$, *i.e.*, by adding more matrices to the set $\{M_l, l=1,\ldots,m\}$, so that the true correlation structure can be accommodated or better approximated by $\sum_{l=1}^{m} a_l M_l$. However, there are other issues, such as how to decide m. Using more matrices may increase the chance of R_0^{-1} being included in the class $\sum_{l=1}^{m} a_l M_l$. On the other hand, using a very large set of

Table 3.2: The GEE estimator and QIF for large samples (1000 simulation runs)

		Working R							
	True R_0	EX				AR-1			
		GEE	QIF	OEF	SRE	GEE	QIF	OEF	SRE
$n = 200$	EX *	2.938	2.976	2.977	0.987	4.306	4.306	4.305	1.000
	AR-1 *	7.213	7.246	7.242	0.995	3.426	3.479	3.479	0.985
$n = 1000$	EX **	5.986	5.976	5.976	1.002	8.828	8.820	8.819	1.001
	AR-1 **	14.21	14.18	14.18	1.002	7.189	7.275	7.276	0.988

Entries "GEE", "QIF" and "OEF" are MSEs of $\hat{\beta}_G$, $\hat{\beta}_Q$ and $\hat{\beta}_O$ (the solution to the asymptotic optimal estimating equation (3.9)), respectively; the units of MSEs reported in rows marked by * and ** are 10^{-4} and 10^{-5}, respectively. "SRE" is the same as in Table 3.1.

$\{M_l, l = 1, \ldots, m\}$ may not make any sense. If m is large, the vector $g_e(Y, X, \beta)_{mp\times 1}$ will be very long, and it is impossible that each element of $g_e(Y, X, \beta)$ contains non- redundant information. In fact when $mp > t$, $Var[g_e(Y, X, \beta)]$ will not be of full rank and hence will not be strictly positive definite, which violates one basic assumption that is used to establish some asymptotic properties of $Q_n(\beta)$ and $\hat{\beta}_Q$ (Pilla and Loader, 2006). In summary, it is not convincing that the QIF method improves estimation in the GEE framework.

3.3.2 Selection of Working Correlation Matrix

An alternative way to improve the efficiency of the GEE estimator is to select an appropriate working correlation structure for the GEE model. To achieve this goal, Pan (2001) constructed a criterion called QIC which is a modification to AIC based on quasi-likelihood. Under the framework of the generalized linear model, if Y is a scalar response, then the (log) quasi-likelihood function (McCullagh and Nelder, 1989, page 325) is

$$Q(\mu, Y) = \int_Y^\mu \frac{Y - u}{V(u)} du,$$

which behaves like a log-likelihood function for μ under very mild assumptions. However, when Y denotes the t-component response vector of a subject and V is the working covariance matrix of Y defined with a general correlation structure $R(\alpha)$ as in (3.1), the integral

$$Q(\mu, Y, u(s)) = \int_{u(s)=Y}^{u(s)=\mu} (Y-\mu)^T \{V(u)\}^{-1} du(s)$$

along a smooth path $u(s)$ in $\mathbb{R}^t$ ordinarily depends on the particular path $u(s)$ chosen, and hence it does not make sense to use this function as a quasi-likelihood (McCullagh and Nelder, 1989, page 333-5). To avoid this difficulty, Pan (2001) opted for the independence structure $R = I$ to define the quasi-likelihood based on data $\{(Y_i, X_i), i = 1, \ldots, n\}$,

$$Q\left(\beta; I\right) = \sum_{i=1}^{n} \sum_{j=1}^{t_i} Q(\beta, (Y_{ij}, X_{ij})).$$

Analogous to the Kullback-Leibler distance which is used to derive AIC, a new discrepancy between the true model, indexed by the true parameter β^*, and a candidate model, indexed by β, was defined as

$$\Delta(\beta, \beta^*, I) = E[-2Q\left(\beta; I\right)]. \tag{3.11}$$

Further, Pan (2001) obtained an approximation of (3.11) using Taylor's expansion up to the second-order partial derivative, and reached the model selection criteria QIC by ignoring the first-order partial derivative term that is difficult to estimate. Specifically,

$$\text{QIC}(R) = -2Q(\hat{\beta}(R); I) + 2\text{tr}(\hat{\Omega}_I \hat{V}_R), \tag{3.12}$$

where $\hat{\Omega}_I$ is the negative Hessian of $Q(\hat{\beta}(R); I)$ under the independent correlation structure, and $\hat{V}_R$ is the sandwich covariance estimator (3.4) evaluated with the working correlation structure R. While QIC has the strength of not being based on a parametric likelihood, this criterion is not very powerful in choosing a working correlation structure due to the fact that $Q\left(\beta; I\right)$ does not contain any information about the within-subject correlation structure.

A nonparametric approach not involving any kind of likelihood was proposed by Pan and Connett (2002). They used resampling-based procedures (*i.e.*, bootstrap and cross-validation) to select the working correlation structure that minimizes the estimated predictive

mean squared error. Like the QIC criterion, this approach has also been shown by simulation to be effective to some extent. Since one of the several simulation settings in Pan and Connett (2002) is identical to that of Pan (2001), one can compare the two different approaches based on the simulation results reported in the two articles; there is some evidence, although not conclusive, that the QIC criterion has a higher chance to select the correct correlation structure.

3.4 Empirical Likelihood for Regression Models and for Dependent Data

In the context of longitudinal data, the primary interest is often in estimation and inference for the regression parameters, and responses measured on the same subject are dependent. So it is useful to review EL for regression models and dependent data.

Empirical likelihood has been applied to inference on regression parameters for independent data by Owen (1991) and Kolaczyk (1994) in the context of linear models and generalized linear models, respectively. Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ be the vector of covariates and response. In linear models, EL for the regression parameter β was defined with the estimating equation

$$E[g((Y,X),\beta)] = E[X(Y - X^T\beta)] = 0,$$

of which the solution minimizes the mean square prediction error $E((Y - X^T\beta)^2)$. Owen (1991) distinguished between linear models with random covariates and those with non-random covariates. In the former case, $g(X_i, Y_i, \beta)$, $i = 1, \ldots, n$, are i.i.d., and $-2\log\mathcal{R}(\beta_0) \to \chi^2_{(p)}$ follows from the results in Owen (1988, 1990) directly. In the latter case where Y_i are sampled independently given $X_i = x_i$, $g(X_i, Y_i, \beta)$ are independent but not identically distributed. Owen (1991) employed triangular array arguments to justify the asymptotic chi-squared distribution of $-2\log\mathcal{R}(\beta_0)$ for independent but not identically distributed data. Later, Kolaczyk (1994) used quasi-likelihood to derive the estimating function

$$g((Y,X),\beta)) = \frac{Y-\mu}{V(\mu)}\left(\frac{\partial\mu}{\partial\beta}\right)$$

for regression parameters in generalized linear models, and applied Owen's triangular array empirical likelihood theorem in this more general setting.

Although empirical likelihood was originally defined for independent observations, its extension to dependent data, especially time series, has also been studied by many others. Specifically, under the assumption of weak dependence, Kitamura (1997) preserved the dependence of the data by reorganizing the original time series $\{Y_t, t = 1, \ldots, N\}$ into a set of blocks $\{ B_i = (Y_{(i-1)L+1}, \ldots, Y_{(i-1)L+M}),\ i = 1, \ldots, n\}$, where $M \geq L$ and are properly chosen block length and distance between block starting points, respectively. To define ELR for θ, Kitamura used an estimating function $\phi(B_i, \theta)$ of observation blocks rather than an estimating function $g(Y_t, \theta)$ of individual observations. For simplicity, $\phi(B_i, \theta)$ was chosen to be $\phi(B_i, \theta) = \sum_{j=1}^{M} g(Y_{(i-1)L+j}, \theta)/M$. Kitamura also established the asymptotic distribution of ELR for weakly dependent time series.

Very recently, application of empirical likelihood to longitudinal data analysis has received attention from several authors. You *et al.* (2006) applied empirical likelihood to the semiparametric longitudinal partially linear regression model described in Zeger and Diggle (1994),

$$Y_{ij} = X_{ij}^T \beta + h(t_{ij}) + \varepsilon_i(t_{ij}) + e_{ij}, \tag{3.13}$$

where h is an arbitrary smooth function of time, $\{\varepsilon_i(t), i = 1, \ldots, n\}$ are independent replicates of a zero mean stationary process that accounts for the serial correlation within each cluster, and e_{ij} are i.i.d. measurement errors. In their approach, Y_{ij} and X_{ij} were first adjusted to $\hat{Y}_{ij}$ and $\hat{X}_{ij}$, respectively, using an estimator of $h(t_{ij})$ given β, so that $E(\hat{Y}_{ij} - \hat{X}_{ij}^T \beta) = o(1)$; to minimize $E[(\hat{Y}_{ij} - \hat{X}_{ij}^T \beta)^2]$ in the same way as for linear models, β is defined to satisfy

$$E[\hat{X}_{ij}(\hat{Y}_{ij} - \hat{X}_{ij}^T \beta)] = 0_{p \times 1}, \tag{3.14}$$

and then the estimating function for β

$$g((Y_i, X_i), \beta) = \sum_{j=1}^{t_i} \hat{X}_{ij}(\hat{Y}_{ij} - \hat{X}_{ij}^T \beta) \tag{3.15}$$

is derived. To accommodate the within-subject correlation, estimating function (3.15) is defined on a data cluster (or a block) (Y_i, X_i) rather than an individual observation (Y_{ij}, X_{ij}),

similar to the idea in Kitamura (1997). Therefore, an empirical likelihood ratio $\mathcal{R}(\beta)$ of the form in (1.3) will consist of n probability weights $\{w_i, i = 1, \ldots, n\}$ that are assigned to the n data clusters $\{(Y_i, X_i), i = 1, \ldots, n\}$. Since $g((Y_i, X_i), \beta)$ is simply the sum of $\hat{X}_{ij}(\hat{Y}_{ij} - \hat{X}_{ij}^T\beta)$ over all $j = 1, \ldots, t_i$, this formulation treats each observation within cluster i equally, and does not contain any information about the within-subject correlation.

Other recent works on the use of empirical likelihood for longitudinal data analysis include Xue and Zhu (2007), and Zhao and Jian (2007). Specifically, Xue and Zhu (2007) considered a varying coefficient model

$$Y_{ij} = X_i^T(t_{ij})\beta(t_{ij}) + \varepsilon_i(t_{ij}), \quad j = 1, \ldots, t_i, \tag{3.16}$$

where the coefficient $\beta(t)$ is assumed to be a vector of smooth functions of continuous time, and $\{\varepsilon_i(t)\}$ are the same as in (3.13); the estimating function

$$g((Y_i, X_i), \beta(t)) = \sum_{j=1}^{t_i} [Y_{ij} - X_i^T(t_{ij})\beta(t)]X_i(t_{ij})K_h(t - t_{ij})$$

was defined to yield a least-square estimator of $\beta(t)$ conditional on t, with a kernel function $K_h(\cdot)$ modeling the underlying density of t. Zhao and Jian (2007) used a prospective logistic model to analyze case-control longitudinal data, and defined empirical likelihood for β with

$$g((Y_i, X_i), \beta) = \sum_{j=1}^{t_i} I_{ij} X_{ij}^T \left[Y_{ij} - \frac{\exp(\eta_{ij}(\beta))}{1 + \exp(\eta_{ij}(\beta))} \right],$$

where I_{ij} is an indicator variable denoting whether each individual observation is included in the case-control sample, and $\eta_{ij}(\beta) = \text{logit}[E(Y_{ij}|X_{ij})]$.

Although all three existing articles focus on different models to analyze different types of longitudinal data, or to address different research issues, they share two common features. First, the estimating function for β, $g((Y_i, X_i), \beta)$, is the sum of t_i dependent subfunctions (*e.g.*, $\hat{X}_{ij}(\hat{Y}_{ij} - \hat{X}_{ij}^T\beta), j = 1, \ldots, t_i$, in (3.15)) that are derived from an equation that holds for each individual observation (*e.g.*, the equation in (3.14)); however, $g((Y_i, X_i), \beta)$ does not account for the within-subject correlation. Second, empirical likelihood is used as an alternative way to construct confidence intervals for the regression parameter. All of the

articles show that the empirical likelihood method has three advantages over the traditional method of normal approximation and/or the computational method of bootstrap: it leads to confidence intervals with better coverage properties; it does not require variance estimation for the estimator, which is rather complicated in nonparametric or semiparametric regression settings; it avoids intensive Monte Carlo simulations required by the bootstrap method.

To our knowledge, there is no published work that extends empirical likelihood to longitudinal data with the goal of improving efficiency of the estimator. Since non-Gaussian continuous, categorical and count longitudinal data are often encountered in practice, we are especially interested in improving point estimation for GEE models that are appropriate for longitudinal data in general when population-averaged effects of treatments are of interest.

3.5 Improving Estimation of GEE with Empirical Likelihood

In the GEE context, an inappropriate working correlation structure may significantly impair the efficiency of the estimator for β. It can be seen from Section 3.3.1 that, if the working correlation is misspecified, it is difficult (or infeasible) to obtain an estimator with improved efficiency by simply using an extended estimating function that contains no more information than the original GEE. According to Qin and Lawless (1994), an extended estimating function $g_e(Y, X, \beta)_{r\times 1}$, with $r > \dim(\beta) = p$, will result in a more efficient estimator if its r components are all functionally independent; that is, each component should contain non-redundant information. In the absence of additional information that can be used to effectively extend the original generalized estimating equation, a more plausible way to improve the GEE estimator is to select the working correlation structure most appropriate for the data at hand.

As discussed in Section 3.3.2, existing model selection methods for GEE models are not very powerful in choosing the correct correlation structure. One possible reason is that they are not based on a likelihood that contains information about the correlation among repeated measurements. While it is neither desirable nor easy to construct a parametric

likelihood for GEE models, empirical likelihood provides an attractive alternative. Thus, to achieve efficiency improvement, we explore applying empirical likelihood to selection of working correlation structures in GEE.

3.5.1 Empirical likelihood for parameters in GEE

To begin, we first need to define empirical likelihood for the regression parameter β in GEE. We have already seen that estimating equations form an important component in EL theory, combining information about parameters. As a special class of estimating equations, GEE fits naturally into empirical likelihood. Note that GEE is an extension of generalized linear models to longitudinal data. Thus, parallel to the path of extensions of linear models, namely "linear models → generalized linear models → GEE for longitudinal data", we study the potential path of extensions of EL: "EL for linear models → EL for GLM → EL for longitudinal data in the GEE framework". Here, by 'the GEE framework', we mean the original GEE and closely related variants that require the same moment assumptions about the underlying distribution of the response, so GEE extensions, such as Prentice's GEE and GEE2, that need more assumptions (Molenberghs & Verbeke, 2005) are not included in our current study.

It is quite natural to incorporate GEE into EL, so we can define the empirical likelihood ratio for the regression parameter β to be

$$\mathcal{R}(\beta) = \sup\left\{\prod_{i=1}^{n} nw_i \ : \ w_i \geq 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i g(Y_i, X_i, \beta) = 0\right\}, \tag{3.17}$$

with

$$g((Y_i, X_i), \beta) = \left(\frac{\partial \mu_i}{\partial \beta^T}\right)^T V_i^{-1}(\beta, \hat{\alpha}(\beta), \hat{\phi}(\beta))(Y_i - \mu_i), \tag{3.18}$$

where $V_i^{-1}(\beta, \hat{\alpha}(\beta), \hat{\phi}(\beta)) = A_i^{-1/2}(\mu_i, \hat{\phi}(\beta))R^{-1}(\hat{\alpha}(\beta))A_i^{-1/2}(\mu_i, \hat{\phi}(\beta))$. The definition in (3.17) is similar to Kitamura's definition for time series data in that empirical likelihood is applied to data clusters (or blocks) $(Y_i, X_i)_{t_i \times (p+1)}$. While blocks of time series data are not well defined and may overlap with one another (the block length and the size of overlap need to be selected; see Kitamura, 1997), the choice for blocks is obvious in

longitudinal data settings. Since data are clustered by subjects, each block simply consists of a cluster of responses and the corresponding covariates. Furthermore, time series blocks are not completely independent and a strong mixing-condition is needed in the derivation of theoretical results. However, longitudinal data blocks are independent of one another because of the independence between subjects. If we take the identity link $h(\mu_{ij}) = \mu_{ij}$, the variance function for normal data $v(\mu_{ij}) = 1$ and the independence working correlation structure $R(\alpha) = I$, then $\mathcal{R}(\beta)$ in (3.17) reduces to the form in You *et al.* (2006); see (3.15) in Section 3.4. Therefore the empirical likelihood ratio defined by You *et al.* is a special case of (3.17). Similarly, our empirical likelihood differs from the EL of Xue and Zhu (2007) and the EL of Zhao and Jian (2007) in that the within-subject correlation is accounted for by $\mathcal{R}(\beta)$ in (3.17), but is ignored in the other two empirical likelihood ratios. Although existing extensions of EL to longitudinal data have been shown to possess advantages over traditional inference methods, these approaches may not be entirely satisfactory due to the ignorance of the within-subject correlation.

The maximum empirical likelihood estimator for β is

$$\hat{\beta}_E = \arg\max_{\beta \in \mathbb{R}^p} \mathcal{R}(\beta).$$

In this just-determined case where the dimension of the parameter β and that of the estimating function $g((Y_i, X_i), \beta)$ in (3.18) are both p, the equation $\frac{1}{n}\sum_i^n g(Y_i, X_i, \beta) = 0$ has a unique solution which is nothing but $\hat{\beta}_G$. It follows that $\mathcal{R}(\beta)$ can achieve its unconditional maximum 1 with $w_i = 1/n$ $(i = 1, \ldots, n)$ when $\beta = \hat{\beta}_G$, *i.e.*, the MELE $\hat{\beta}_E$ is identical to $\hat{\beta}_G$. Consequently, the MELE $\hat{\beta}_E$ resulting from the empirical likelihood defined in (3.17) inherits all properties of the GEE estimator, namely that the estimator is consistent and asymptotically normal whether or not the working correlation structure $R(\alpha)$ is correct, but is generally not efficient if $R(\alpha)$ is misspecified. From this point of view, the empirical likelihood defined by You *et al.* (2006) cannot yield an optimal estimator unless the data within each cluster are independent.

When it comes to selecting the best working correlation matrix from a set of competing matrices using empirical likelihood, the first question to be answered is "How to compare empirical likelihoods of GEE models with different working correlation matrices?" As seen from above, $\mathcal{R}(\beta)$ in (3.17) is defined for a GEE model with a particular $R(\alpha)$, and hence is specific to the choice of $R(\alpha)$. That is, if there are M competing working correlation matrices $\{R_m, m = 1, \ldots, M\}$, then M different empirical likelihoods $\{\mathcal{R}^m(\beta), m = 1, \ldots, M\}$ can be defined by (3.17). For each R_m, the MELE $\hat{\beta}_E^m$ equals the corresponding GEE estimator $\hat{\beta}_G^m$, and $\max_\beta \mathcal{R}^m(\beta) = \mathcal{R}^m(\hat{\beta}_E^m) \equiv 1$ (or equivalently, $-2\log \mathcal{R}^m(\hat{\beta}_E^m) \equiv 0$). It is clear that comparing $\mathcal{R}^m(\hat{\beta}_E^m)$ for $m = 1, \ldots, M$ cannot achieve model comparison. In order to facilitate model comparison via empirical likelihood, we need to define a unified measure, in this case a unified empirical likelihood ratio, so that GEE models with different working correlation structures will "be assigned" different values.

We can proceed by embedding a set of competing working structures into a more general structure, of which the competing working structures are special cases. This idea is illustrated in the following example.

Example 3.1 As shown in (3.19), if $t_i = t = 3$ and we are to choose one from the independence (R_1), the exchangeable (R_2), and the AR-1 (R_3) correlation structures, a general structure could be the stationary structure (R_4). $R_4(\alpha_1, \alpha_2)$ reduces to R_1 if $\alpha_1 = \alpha_2 = 0$, to $R_2(\alpha)$ if $\alpha_1 = \alpha_2 = \alpha \neq 0$, and to $R_3(\alpha)$ if $\alpha_1 = \alpha \neq 0$ and $\alpha_2 = \alpha^2$.

$$R_1(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad R_2(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix},$$

Independence (IN) Exchangeable (EX)

$$R_3(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{pmatrix}, \quad R_4(\alpha_1, \alpha_2) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 \\ \alpha_1 & 1 & \alpha_1 \\ \alpha_2 & \alpha_1 & 1 \end{pmatrix},$$

AR-1 (AR) Stationary (ST) (3.19)

We denote the specified general structure by R_F, and call the GEE model with R_F the "full model". To make use of the constraints on α that are implied by the full model correlation structure $R_F = R_4$, instead of using $g(Y_i, X_i, \beta)$ in (3.18), we define a new estimating function for $\theta = (\beta^T, \alpha^T)^T$ to be

$$g^F((Y_i, X_i), \beta, \alpha_1, \ldots, \alpha_{t-1}) = \begin{pmatrix} \left(\frac{\partial \mu_i}{\partial \beta^T}\right)^T V_i^{-1}(\beta, \alpha_1, \ldots, \alpha_{t-1}, \hat{\phi}(\beta))(Y_i - \mu_i) \\ \sum_{j=1}^{t-1} e_{ij}(\beta) e_{i,j+1}(\beta) - \alpha_1 \hat{\phi}(\beta)(t - 1 - p/n) \\ \vdots \\ \sum_{j=1}^{1} e_{ij}(\beta) e_{i,j+t-1}(\beta) - \alpha_{t-1} \hat{\phi}(\beta)(1 - p/n) \end{pmatrix}_{(p+t-1)\times 1}, \tag{3.20}$$

where

$$e_{ij}(\beta) = (Y_{ij} - \mu_{ij}(\beta))/\sqrt{v(\mu_{ij}(\beta))}$$

(*i.e.*, the Pearson residual) and

$$\hat{\phi}(\beta) = \frac{1}{nt - p} \sum_{i=1}^{n} \sum_{j=1}^{t} e_{ij}^2$$

(*i.e.*, the method of moment estimator for ϕ). The second through the last components of $g^F(\cdot)$ are derived from the MOM estimator for the correlation coefficients in R_4,

$$\hat{\alpha}_k(\beta) = \frac{1}{(n(t-k) - p)\phi} \sum_{i=1}^{n} \sum_{j=1}^{t-k} e_{ij} e_{i,j+k},$$

which is also used in the original GEE that assumes R_4. Therefore, $g^F(\cdot)$ in (3.20) is equivalent to $g(\cdot)$ with $R = R_4$ in (3.18), in the sense that they yield the same $\hat{\beta}_G$ when used to form estimating equations for β. Then, we can define the full-model empirical likelihood ratio

$$\mathcal{R}^F(\beta, \alpha) = \sup \left\{ \prod_{i=1}^{n} n w_i \ : \ w_i \geq 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i g^F(Y_i, X_i, \beta, \alpha) = 0 \right\} \tag{3.21}$$

with $g^F(\cdot)$ in (3.20), where $\alpha = (\alpha_1, \ldots, \alpha_{t-1})^T$ and $t = 3$ in this example.

The purpose of defining a full model empirical likelihood ratio as in (3.21) is twofold: first, information about within-subject correlation can be built into $\mathcal{R}^F(\beta, \alpha)$ at the expense of a very weak assumption, such as the stationarity assumption of the underlying correlation structure; second, $\mathcal{R}^F(\beta, \alpha)$ serves as a unified measure that can be applied to each of the competing GEE models. To see the latter, let us return to the above example, where the full model correlation structure is parameterized by $\alpha = (\alpha_1, \alpha_2)$. Suppose $\mathcal{R}^F(\beta, \alpha_1, \alpha_2)$ is defined for the full model and we have four sets of GEE estimates based on the same data: $\hat{\beta}^{IN}$, $(\hat{\beta}^{EX}, \hat{\alpha}^{EX})$, $(\hat{\beta}^{AR}, \hat{\alpha}^{AR})$, and $(\hat{\beta}^{ST}, \hat{\alpha}_1^{ST}, \hat{\alpha}_2^{ST})$, each obtained with one of the four working correlation structures. Then these four GEE models can be evaluated by

$$\begin{aligned} \mathcal{R}_1 &= \mathcal{R}^F(\hat{\beta}^{IN}, 0, 0), \\ \mathcal{R}_2 &= \mathcal{R}^F(\hat{\beta}^{EX}, \hat{\alpha}^{EX}, \hat{\alpha}^{EX}), \\ \mathcal{R}_3 &= \mathcal{R}^F(\hat{\beta}^{AR}, \hat{\alpha}^{AR}, (\hat{\alpha}^{AR})^2), \\ \text{and} \qquad \mathcal{R}_4 &= \mathcal{R}^F(\hat{\beta}^{ST}, \hat{\alpha}_1^{ST}, \hat{\alpha}_2^{ST}), \end{aligned} \tag{3.22}$$

respectively. With the new definition (3.21), $\mathcal{R}_4 = \mathcal{R}^F(\hat{\beta}^{ST}, \hat{\alpha}_1^{ST}, \hat{\alpha}_2^{ST})$ is still equal to 0, since $(\hat{\beta}^{ST}, \hat{\alpha}_1^{ST}, \hat{\alpha}_2^{ST})$ also solves the equation $\frac{1}{n}\sum_i^n g^F(Y_i, X_i, \beta, \alpha_1, \alpha_2) = 0$; however, empirical likelihood ratios for the other GEE models (*i.e.*, $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{R}_3$) are no longer zero. Thus, we are now able to compare different GEE models using empirical likelihood.

3.5.2 Choosing a working correlation structure

First consider the simple situation where each competing GEE model has the same number of parameters (*i.e.*, $\dim(\theta)$). Note that θ includes the regression parameter β and the corre-

lation parameter α. The exchangeable structure is often used as an alternative to the AR-1 structure, and GEE models with both structures have an equal number of parameters. Intuitively, if the underlying correlation structure is AR-1, then $\hat{\theta}^{AR} = (\hat{\beta}^{AR}, \hat{\alpha}^{AR}, \ldots, (\hat{\alpha}^{AR})^{t-1})^T$ will be closer to the solution of

$$\frac{1}{n}\sum_{i}^{n} g^F(Y_i, X_i, \beta, \alpha_1, \ldots, \alpha_{t-1}) = 0,$$

and therefore $\mathcal{R}^F(\hat{\beta}^{AR}, \hat{\alpha}^{AR}, \ldots, (\hat{\alpha}^{AR})^{t-1})$ will be closer to 1. Thus we can select the working correlation structure that leads to the larger ELR.

We conduct a very simple simulation study to examine this idea. Specifically, we adopt the model

$$Y_{ij} = X_{ij}\beta + \varepsilon_{ij}, \quad (j = 1, \ldots, t, \text{ and } i = 1, \ldots, n), \tag{3.23}$$

where $t_i = t = 3$, $\beta = 2$, X_{ij} is a univariate random variable generated from $\mathcal{N}(0.1j, 1)$, and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^T$ are generated from a multivariate normal distribution $\mathcal{N}_3(0_{3\times 1}, \Sigma_0)$. That is, we assume the identity link $h(\mu_{ij}) = \mu_{ij}$. We let $\mathrm{Var}(\varepsilon_{ij}) = 1$, so $\Sigma_0 = R_0$, where the true correlation matrix R_0 has either the exchangeable (EX) structure or the AR-1 (AR) structure. The parameter in both correlation structures is specified to be $\alpha = 0.7$. Simulation results displayed in Table 3.3 indicate that the correct correlation structure is much more likely to be chosen, and the proportion of correct decisions increases when the sample size n becomes larger.

Table 3.3: Selecting a working correlation structure in the simple situation

		$n = 50$		$n = 100$	
		R		R	
		EX	AR	EX	AR
R_0	EX	920	80	921	79
	AR	204	796	68	932

R_0 and R stand for the true and working correlation structures, respectively. Each entry shows the number of times that a working structure is chosen over 1000 simulation runs.

In the more general situation where competing working correlation structures have different numbers of parameters, it is necessary to extend the above idea to a model selection criterion that takes the dimension of the model into account. In parametric settings, the Akaike information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwarz, 1978) are widely used for model selection. These criteria cannot be applied to GEEs due to the lack of parametric likelihood. However, we can modify AIC and BIC by substituting empirical likelihood for parametric likelihood, and get empirical likelihood versions of AIC and BIC:

$$EAIC(m) = -2\log\mathcal{R}^F(\hat{\theta}_G^m) + 2\dim(\theta^m), \tag{3.24}$$

$$EBIC(m) = -2\log\mathcal{R}^F(\hat{\theta}_G^m) + \dim(\theta^m)\log(n), \tag{3.25}$$

where m is the index for a candidate model parameterized by θ^m $(m = 1, \ldots, M)$, and $\hat{\theta}_G^m$ is the GEE estimate associated with the working correlation structure R_m. More specifically,

$$\hat{\theta}_G^m = \begin{pmatrix} \hat{\beta}_G^m \\ \hat{\alpha}_G^m \end{pmatrix},$$

where $\hat{\alpha}_G^m$ is the method of moment estimator of α given $\hat{\beta}_G^m$ and R_m.

In fact, Kolaczyk (1995) showed that a criterion analogous to AIC may be derived in the context of empirical likelihood for estimating equations. It is known that the Kullback-Leibler distance

$$K(\theta, \hat{\theta}) = \int \log\left(\frac{f(x|\theta)}{f(x|\hat{\theta})}\right) f(x|\theta)dx, \tag{3.26}$$

where $\hat{\theta}$ is an estimate of θ, is a measure of the discrepancy between the estimated and the true probability distributions. Suppose $\theta \in \mathbb{R}^L$, and let $\theta_k \in \mathbb{R}^k$ $(k < L)$ be a nested parameter of θ, *i.e.*, θ_k consists of k components of θ and assumes that the other $L - k$ components are zeros. Akaike (1973) showed that $2nE[K(\theta, \hat{\theta}_k^{MLE})]$, which describes the risk of modeling θ with $\hat{\theta}_k^{MLE}$, can be asymptotically estimated by

$$-2\sum_{i=1}^{n}\log\left(\frac{f(X_i|\hat{\theta}_k^{MLE})}{f(X_i|\hat{\theta}^{MLE})}\right) + 2k - L, \tag{3.27}$$

where $\hat{\theta}_k^{MLE}$ and $\hat{\theta}^{MLE}$ are the maximum likelihood estimates of θ_k and θ, respectively. Since it is assumed that competing models with different θ_k are nested in the common full model with $\theta \in \mathbb{R}^L$, one only needs to compute the part of (3.27) that is specific to θ_k, which leads to the AIC criterion

$$-2\sum_{i=1}^{n} \log f(X_i|\hat{\theta}_k^{MLE}) + 2k.$$

In the setting of Kolaczyk (1995), information about a regression parameter $\theta \in \mathbb{R}^L$ is summarized by an L-component estimating equation $E[g(X,\theta)] = 0_{L\times 1}$, via which the empirical likelihood ratio $\mathcal{R}(\theta)$ is defined; θ_k is assumed to be very near to θ, as in the derivation of AIC. Within the context of empirical likelihood, Kolaczyk defined an analog of the loss function in (3.26) to be

$$K_{EL}(\theta,\hat{\theta}) = \sum_{i=1}^{n} \log\left(\frac{w_i(\theta)}{w_i(\hat{\theta})}\right) w_i(\theta), \tag{3.28}$$

where $\{w_i(\theta), i = 1,\ldots,n\}$ and $\{w_i(\hat{\theta}), i = 1,\ldots,n\}$ are the two sets of probability weights associated with $\mathcal{R}(\theta)$ and $\mathcal{R}(\hat{\theta})$, respectively. Kolaczyk denoted the MELE of θ_k by $\tilde{\theta}_k$, and further showed that the statistic

$$-2\log\mathcal{R}(\tilde{\theta}_k) + 2k - L$$

is asymptotically an unbiased estimator of

$$2nE[K_{EL}(\theta,\tilde{\theta}_k)]. \tag{3.29}$$

Analogous to AIC, the EIC

$$EIC = -2\log\mathcal{R}(\tilde{\theta}_k) + 2k \tag{3.30}$$

is a statistic specific to θ_k, and can be used to compare the risks of approximating a larger model with parameter θ by smaller fitted submodels with different $\tilde{\theta}_k$. One advantage of EIC is that it requires no parametric assumption regarding the distribution of the data, but only unbiasedness of the estimating function. Although an information criterion statistic typically is used for choosing an optimal model among many choices, Kolaczyk (1995) focused only on the behavior of EIC as an *estimator* for (3.29), and left out the issue of model selection.

Our EAIC in (3.24) is a modified version of Kolaczyk's EIC, and is adapted to the context of GEEs. The most significant modification is that we replace the maximum empirical likelihood estimator in (3.30) by the GEE estimator $\hat{\theta}_G$. In Example 3.1, after the full model empirical likelihood ratio $\mathcal{R}^F(\beta, \alpha_1, \alpha_2)$ is defined with the stationary correlation structure $R_4(\alpha_1, \alpha_2)$, we plug four different GEE estimates into $\mathcal{R}^F(\beta, \alpha_1, \alpha_2)$ as in (3.22) to obtain empirical likelihood ratios for each of the four GEE models. An alternative approach, which is also used in Kolaczyk (1995), is to first obtain four different maximum empirical likelihood estimates with different constraints, and then use them in place of the GEE estimates. The full model MELE in Example 3.1 is

$$\tilde{\theta}^{ST} = \begin{pmatrix} \tilde{\beta}^{ST} \\ \tilde{\alpha}_1^{ST} \\ \tilde{\alpha}_2^{ST} \end{pmatrix}_{(p+2)\times 1} = \arg \max_{(\beta,\alpha_1,\alpha_2)^T} \mathcal{R}^F(\beta, \alpha_1, \alpha_2),$$

and is equal to the full model GEE estimate $\hat{\theta}_G^{ST}$ as they both solve the equation

$$\frac{1}{n} g^F((Y_i, X_i), \beta, \alpha_1, \alpha_2) = 0_{(p+2)\times 1}.$$

For other correlation structures that are nested in $R_F = R_4(\alpha_1, \alpha_2)$, maximum empirical likelihood estimates are different from the corresponding GEE estimates. For instance, the GEE estimate $\hat{\theta}_G^{EX}$ consists of $\hat{\beta}^{EX}$ that solves the GEE

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta^T} \right)^T [A_i^{1/2}(\beta, \hat{\phi}(\beta)) R_2(\hat{\alpha}(\beta)) A_i^{1/2}(\beta, \hat{\phi}(\beta))]^{-1} (Y_i - \mu_i) = 0,$$

and $\hat{\alpha}^{EX} = \hat{\alpha}(\hat{\beta}^{EX})$, whereas the MELE $\tilde{\theta}_G^{EX} = (\,(\tilde{\beta}^{EX})^T, \tilde{\alpha}^{EX})^T$ is obtained by maximizing $\mathcal{R}^F(\beta, \alpha_1, \alpha_2)$ with respect to $(\beta^T, \alpha_1, \alpha_2)^T$ under the additional constraint that $\alpha_1 = \alpha_2$. With the additional constraint, we can suppress α_2 and drop the subscript of α_1 in $\mathcal{R}^F(\cdot)$, and then express $\tilde{\theta}_G^{EX}$ more explicitly as

$$\tilde{\theta}_G^{EX} = \begin{pmatrix} \tilde{\beta}^{EX} \\ \tilde{\alpha}^{EX} \end{pmatrix}_{(p+1)\times 1} = \arg \max_{(\beta^T,\alpha)^T} \mathcal{R}^F(\beta, \alpha), \tag{3.31}$$

where the estimating function involved in $\mathcal{R}^F(\beta, \alpha)$ becomes

$$g^F(Y_i, X_i, \beta, \alpha) = \begin{pmatrix} \left(\frac{\partial \mu_i}{\partial \beta^T}\right)^T V_i^{-1}(\beta, \alpha, \hat{\phi}(\beta))(Y_i - \mu_i) \\ \sum_{j=1}^{2} e_{ij}(\beta) e_{i,j+1}(\beta) - \alpha\hat{\phi}(\beta)(2 - p/n) \\ \sum_{j=1}^{1} e_{ij}(\beta) e_{i,j+2}(\beta) - \alpha\hat{\phi}(\beta)(1 - p/n) \end{pmatrix}_{(p+2)\times 1} . \tag{3.32}$$

In this over-constrained case,

$$\dim(g^F(\cdot)) = p + 2 > \dim(\theta^{EX}) = p + 1,$$

with a difference of 1 caused by the additional constraint. If R_F is stationary with parameter $\alpha = (\alpha_1, \ldots, \alpha_{t-1})$, then the exchangeable structure imposes $t - 2$ additional constraints: $\alpha_1 = \alpha_2$, $\alpha_2 = \alpha_3$, ..., and $\alpha_{t-2} = \alpha_{t-1}$. Therefore the number of additional constraints increases with the cluster size t. Similarly, for other correlation structures embedded in R_F, maximization of the empirical likelihood ratio with respect to the parameter is subject to additional constraints. Thus, it can be seen that the approach of Kolaczyk (1995) involves searching for maximum likelihood estimates in the over-constrained case where $\dim(g^F(\cdot)) > \dim(\theta)$.

When $\dim(g^F(\cdot)) > \dim(\theta)$, the existence of a maximum empirical likelihood estimate $\tilde{\theta}$ is guaranteed only when 0 is inside the convex hull of the data, *i.e.*, $\mathcal{H}_n = ch\{g^F((Y_1, X_1), \theta), \ldots, g^F((Y_n, X_n), \theta)\}$. For each correlation structure embedded in R_F (submodel), $\Pr(0 \in \mathcal{H}_n) \to 1$ as $n \to \infty$ if

$$E[g^F((Y, X), \theta)] = 0 \tag{3.33}$$

holds (see Theorem 4.1 in Owen, 2001). It is conceivable that not all embedded correlation structures are correct, and hence (3.33) will not always hold. Even if an embedded correlation structure is correct and (3.33) holds, $\Pr(0 \in \mathcal{H}_n)$ is not 1 for finite samples. Therefore, 0 is not always inside $\mathcal{H}_n$ in practice. This problem becomes severe if the cluster size t is large so that some submodels become highly over-constrained. In the situation where 0 is near or outside $\mathcal{H}_n$, convergence to a valid solution is difficult or impossible. Variyath *et. al* (2007)

also discussed this issue. They proposed an adjusted empirical likelihood by introducing an artificial "observation"

$$g^F((Y_{n+1}, X_{n+1}), \theta) = -a_n \left(\frac{1}{n}\right) \sum_{i=1}^{n} g^F((Y_i, X_i), \theta),$$

where $a_n = o_p(n^{2/3})$ is a positive constant, so that 0 is always inside the enlarged convex hull $\mathcal{H}_{n+1} = ch\{g^F((Y_1, X_1), \theta), \ldots, g^F((Y_n, X_n), \theta), g^F((Y_{n+1}, X_{n+1}), \theta)\}$. However, this solution leaves users another practical issue, namely, choosing a_n. Although the adjusted empirical likelihood approach (with a properly chosen a_n) is shown to work well in selecting regression variables, it may not solve all problems encountered in the GEE setting. Unlike the regression parameter β, each element of the correlation parameter α is restricted between -1 and 1. There is some empirical evidence from our preliminary simulation study that maximization of $\mathcal{R}^F(\beta, \alpha)$ over subspaces constrained by embedded correlation structures may reach the boundary of the parameter space where one (or more) component of α is ± 1.

Our modification avoids the computational issues associated with Kolaczyk's (1995) approach, and makes the use of EAIC especially easy in practice. To use EAIC and EBIC in (3.24) and (3.25), one does not need to obtain maximum empirical likelihood estimates with over-constrained estimating equations, but only needs to evaluate $\mathcal{R}^F(\beta, \alpha)$ at GEE estimates for each correlation structure. To compute a GEE estimate $\hat{\theta}_G$, one iterates between a modified Fisher scoring for β and moment estimation of α and ϕ, and usually convergence can be reached. Due to the popularity of the GEE method, it has been implemented in many software packages, including SAS, S-Plus, Stata and SUDAAN. See Horton and Lipsitz (1999) for a review.

Not only is $-2 \log \mathcal{R}^F(\hat{\theta}_G^m)$ in (3.24) and (3.25) computationally convenient, but also it retains the asymptotic property of $-2 \log \mathcal{R}^F(\tilde{\theta}^m)$. Let $\dim(g^F) = r$ and $\dim(\theta^m) = q$. If R_m is correct and $E[g^F((Y, X), \theta^m)] = 0$, then by Corollary 4 in Qin and Lawless (1994), $-2 \log \mathcal{R}^F(\tilde{\theta}^m) \rightarrow \chi^2_{(r-q)}$. We find that $-2 \log \mathcal{R}^F(\hat{\theta}_G^m)$ also behaves like a $\chi^2_{(r-q)}$ random variable, given that the same conditions are met. Specifically, we simulate $-2 \log \mathcal{R}^F(\hat{\theta}_G^m)$ using the model in (3.23) with $t = 4$ and $p = 1$, and let $m = 1, 2, 3$ be the indices for the

independence, exchangeable and AR-1 working assumptions, respectively. Let $\dim(\beta) = p$, then $\dim(\theta^1) = p$, $\dim(\theta^2) = p + 1$ and $\dim(\theta^3) = p + 1$. The general correlation structure is assumed to be stationary, so $r = \dim(g^F) = p + t - 1 = p + 3$. Under each true correlation structure R_m $(m = 1, 2, 3)$, Figure 3.1 plots quantiles of $-2 \log \mathcal{R}^F(\hat{\theta}_G^m)$ against quantiles of $\chi^2_{(3)}$, $\chi^2_{(2)}$ and $\chi^2_{(2)}$, respectively. Empirical evidence in these Q-Q plots suggests that $-2 \log \mathcal{R}^F(\hat{\theta}_G^m) \to \chi^2_{(r-q)}$ if the working assumption R is correct. However, if R is not correct, our numerical results (not shown) indicate that the statistic $-2 \log \mathcal{R}^F(\hat{\theta}_G^m)$ diverges to infinity as $n \to \infty$, just as $-2 \log \mathcal{R}^F(\tilde{\theta}^m)$ does (Variyath *et al.*, 2007). Therefore, the asymptotic behavior of EAIC is the same as Kolaczyk's EIC.

Although EAIC and EBIC in (3.24) and (3.25) are proposed to select a working correlation matrix in GEE, the two criteria may also be used in other settings. To use EAIC and EBIC in the context of selecting regression variables as considered by Kolaczyk (1995), we use $\hat{\theta}_k$ (the solution to the k-dimensional submodel estimating equation) in (3.24) and (3.25), rather than the MELE

$$\tilde{\theta}_k = \arg \max_{\theta \in \mathbb{R}^k} \mathcal{R}(\theta),$$

where $\mathcal{R}(\theta)$ is defined with the L-dimensional $(k < L)$ full model estimating equation. The performance of EAIC and EBIC in variable selection is not examined here, as we focus only on selecting the working correlation in GEE.

3.5.3 Simulation Studies

In this section we perform extensive simulation studies to examine the reliability of using EAIC and EBIC to choose an optimal working correlation structure for GEE models. Although one could extend the use of Kolaczyk's (1995) EIC (which was designed to select regression variables) to this context, our preliminary simulation shows that computational issues of EIC are often encountered. Therefore we do not compare the performance of EIC with that of EAIC and EBIC.

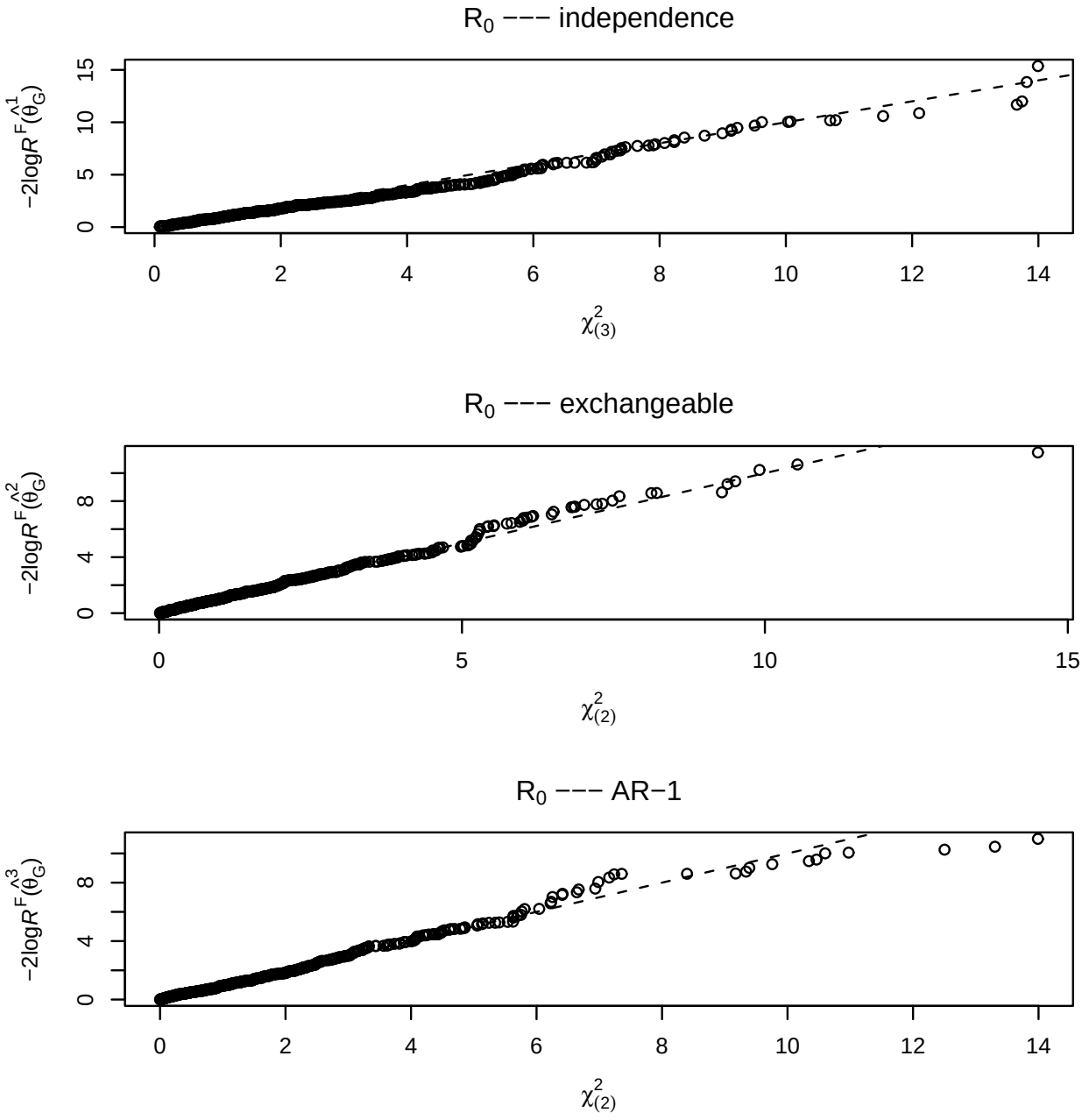

Figure 3.1: Q-Q plots of $-2\log \mathcal{R}^F(\hat{\theta}_G^m)$ and $\chi^2_{(r-q)}$ when the working correlation assumption is correct. $m = 1$ (the top panel), 2 (the middle panel), and 3 (the bottom panel) are indices for the independence, exchangeable and AR-1 working assumptions, respectively. $r = \dim(g^F) = p + 3$; $q = \dim(\theta^m)$ in the top, middle, bottom panels are p, $p+1$, and $p+1$, respectively, where $p = \dim(\beta)$; thus, $r - p$ equals 3 in the top panel, 2 in the middle and bottom panels. These plots are obtained from 1000 random samples of size 2000.

Simulation study 1: Gaussian longitudinal data

First, the model in (3.23) is used to generate Gaussian longitudinal data. For the true (R_0) and working (R) correlation matrices, we consider four structures: the independence (IN), exchangeable (EX), AR-1 (AR) and stationary (ST) structures. Since there is only one covariate and $t = 3$, $\dim(\theta^{IN}) = 1$, $\dim(\theta^{EX}) = 2$, $\dim(\theta^{AR}) = 2$, and $\dim(\theta^{ST}) = 3$. To examine the performance of EAIC and EBIC for moderately correlated longitudinal data, we now reduce α in the true exchangeable and AR-1 structures from 0.7 to 0.5, and parameterize the true stationary matrix with $\alpha = (0.5, 0.35)^T$. Under each true correlation specification, we record the numbers of times that different working correlation structures are chosen over 1000 simulation runs, as shown in Table 3.4.

Table 3.4: Model selection: Gaussian longitudinal data ($t = 3$)

		Working Correlation Structures R											
		$n = 50$				$n = 100$				$n = 200$			
		IN	EX	AR	ST	IN	EX	AR	ST	IN	EX	AR	ST
R_0 =IN	EAIC	688	121	127	75	702	103	121	74	730	93	115	62
	EBIC	864	45	68	23	932	30	32	6	961	18	19	2
R_0 =EX	EAIC	0	766	133	101	0	828	39	33	0	829	0	171
	EBIC	0	832	136	32	0	919	61	20	0	976	14	10
R_0 =AR	EAIC	0	122	754	124	0	55	820	155	0	1	816	183
	EBIC	0	124	840	36	0	69	909	22	0	15	966	19
R_0 =ST	EAIC	0	417	550	33	0	310	572	118	0	152	333	515
	EBIC	0	424	562	14	0	360	640	0	0	380	670	13

"IN", "EX", "AR" and "ST" stand for the independence, exchangeable, AR-1 and stationary structures, respectively. Here the cluster size is $t = 3$. The true exchangeable and AR-1 structures are parameterized by $\alpha = 0.5$, and the true stationary structure has $\alpha = (0.5, 0.35)^T$. Each entry shows the number of times that a working correlation structure is chosen over 1000 simulation runs.

It is seen from the simulation results displayed in Table 3.4 that, when the true correlation structure R_0 is independence, exchangeable or AR-1, both EAIC and EBIC are very effective in choosing the correct correlation structure, with EBIC being better than EAIC. The EAIC

criterion tends to choose the full-model correlation structure (ST) more often than EBIC does. This is not surprising, as it is well known that AIC is more likely to result in an over-parameterized model than BIC in parametric settings. However, when R_0 is stationary, frequencies of EAIC and EBIC choosing the correct structure drop significantly; EBIC rarely picks the stationary structure even when the sample size is as large as 200. In fact, EAIC and EBIC are not as useless as they appear in this case; even though the two criteria mistakenly choose a parsimonious correlation structure most of the time, they will not cause significant efficiency loss. Recall that in the GEE method, estimation of β is of major interest, while α is treated as a nuisance parameter. The purpose of choosing the correct correlation structure is to improve the efficiency of estimating β, since using a wrong working correlation may cause significant loss of efficiency. For example, Table 3.1 in Section 3.3.1 shows that, if R_0 is AR-1, the MSE of $\hat{\beta}_G^{EX}$ is as twice large as the MSE of $\hat{\beta}_G^{AR}$ ($0.00756/0.00377 \approx 2$). With the goal of improving efficiency in mind, we further examine the MSEs of $\hat{\beta}_G^{IN}$, $\hat{\beta}_G^{EX}$, $\hat{\beta}_G^{AR}$ and $\hat{\beta}_G^{ST}$ when R_0 is stationary, and find that $\hat{\beta}_G^{ST}$ is not necessarily more efficient than $\hat{\beta}_G^{EX}$ or $\hat{\beta}_G^{AR}$, especially when the sample size is not very large. This is because the stationary structure has more nuisance parameters and estimating them costs efficiency.

Simulation study 2: comparison with QIC for Gaussian longitudinal data

Next, we compare EAIC and EBIC to the QIC criterion proposed by Pan (2001). As in simulation study 1, we use the model in (3.23) to generate data, and let $t_i = t = 4$. When choosing a working correlation matrix, Pan (2001) considered only three candidates: IN, EX and AR. For this reason, we limit the pool of candidates to these three correlation structures. Note that, in our approach, the empirical likelihood ratio is always defined with a general correlation structure (*e.g.*, ST in this case) even if the general structure is not considered as a candidate for the working correlation matrix. Simulation results are presented in Table 3.5. For any choice of R_0, we see that both EAIC and EBIC perform much better than QIC. If R_0 is independence, QIC seems to be not effective at all in choosing the correct structure, while both EAIC and EBIC (especially EBIC) choose the right structure most of the time.

If R_0 is exchangeable or AR-1, QIC is effective to some extent and performs better with a larger sample size n; however, EAIC and EBIC are far better than QIC in any case, never choosing the incorrect IN structure. It is interesting to note that, if ST is excluded from consideration and if R_0 is EX or AR, the performance of EAIC and that of EBIC seem to be identical.

Table 3.5: Comparison of EAIC, EBIC and QIC (Gaussian response, $t = 4$)

		Working Correlation Structures R								
		$n = 50$			$n = 100$			$n = 200$		
		IN	EX	AR	IN	EX	AR	IN	EX	AR
R_0 =IN	EAIC	654	164	182	723	146	131	732	142	126
	EBIC	853	72	75	933	30	37	944	29	27
	QIC	254	378	368	240	375	385	264	379	357
R_0 =EX	EAIC	0	976	24	0	998	2	0	1000	0
	EBIC	0	976	24	0	998	2	0	1000	0
	QIC	190	579	231	136	645	219	132	687	181
R_0 =AR	EAIC	0	42	958	0	6	994	0	1000	0
	EBIC	0	42	958	0	6	994	0	1000	0
	QIC	166	240	594	120	206	674	119	197	684

The true exchangeable and AR-1 structures are parameterized by $\alpha = 0.5$, and the true stationary structure has $\alpha = (0.5, 0.35, 0.2)^T$. Each entry shows the number of times that a working correlation structure is chosen over 1000 simulation runs.

We now extend the above comparison to the situation where the full-model correlation structure ST is included in the pool of candidates. Simulation results in Table 3.6 show that QIC becomes ineffective for any choice of R_0 if the full-model correlation structure is also considered as a candidate. Since the only difference between the simulation setting of Table 3.6 and that of Table 3.4 is the cluster size, we can compare these two tables to see how the performance of EAIC and the performance of EBIC change with cluster size. With R_0 being independence, results in these two tables are similar. But if R_0 is EX, AR or ST, both EAIC and EBIC perform better for data with a larger cluster size. Intuitively, if the cluster

size increases, the pattern of within-subject correlation will become more prominent, and therefore will be more easily recognized.

Table 3.6: Continued comparison of EAIC, EBIC and QIC (Gaussian response, $t = 4$)

		Working Correlation Structures R											
		$n = 50$				$n = 100$				$n = 200$			
		IN	EX	AR	ST	IN	EX	AR	ST	IN	EX	AR	ST
R_0 =IN	EAIC	642	118	140	100	688	115	108	89	703	107	124	66
	EBIC	848	69	76	7	913	44	37	6	962	17	20	1
	QIC	202	185	186	427	189	189	179	443	210	169	206	415
R_0 =EX	EAIC	0	824	18	158	0	838	0	162	0	871	0	129
	EBIC	0	940	29	31	0	973	4	23	0	993	0	7
	QIC	160	326	211	303	150	313	142	395	103	366	164	367
R_0 =AR	EAIC	0	38	788	174	0	1	834	165	0	0	834	166
	EBIC	0	45	923	32	0	7	978	15	0	0	989	11
	QIC	136	210	299	355	108	194	300	398	82	176	348	394
R_0 =ST	EAIC	0	193	551	256	0	46	444	510	0	2	220	778
	EBIC	0	254	708	38	0	164	762	74	0	26	701	273
	QIC	137	238	247	378	139	247	239	375	131	219	229	421

The true exchangeable and AR-1 structures are parameterized by $\alpha = 0.5$, and the true stationary structure has $\alpha = (0.5, 0.35, 0.2)^T$.

Simulation study 3: comparison with QIC for binary longitudinal data

Since the GEE method has the strength of analyzing non-Gaussian longitudinal data, it is interesting to compare EAIC and EBIC to QIC for binary longitudinal data. Here we adopt the same model considered in Pan (2001) and Pan & Connett (2002):

$$\text{logit}(\mu_{ij}) = \beta_1 + \beta_2 X_{ij2} + \beta_3(j-1), \quad j = 1, 2, 3 \text{ and } i = 1, \ldots, n, \tag{3.34}$$

where X_{ij2} are i.i.d. Bernoulli with $\Pr(X_{ij2} = 1) = 1/2$, and $\beta_1 = 0.25 = -\beta_2 = -\beta_3$. This is also the model used in Fitzmaurice (1995) to show that the independence working assumption, if it is not correct, may result in 40% loss of efficiency compared to the correct correlation assumption. As in simulation study 2, we first include IN, EX and AR into the

pool of true and working correlation structures, for the sake of easy comparison with results from Pan (2001) and Pan & Connett (2002). Both the EX and AR matrices are parameterized with $\alpha = 0.5$ when used as the true correlation matrix. Results in Table 3.7 are similar to those in Table 3.5 for Gaussian longitudinal data. Both EAIC and EBIC are powerful in choosing the correct structure in all cases, and appear to be equivalent if R_0 is EX or AR. QIC is less powerful compared to EAIC and EBIC, and becomes ineffective if R_0 is the independence structure.

Table 3.7: Comparison of EAIC, EBIC and QIC (binary response, $t = 3$)

		Working Correlation Structures R								
		$n = 50$			$n = 100$			$n = 200$		
		IN	EX	AR	IN	EX	AR	IN	EX	AR
R_0 =IN	EAIC	734	147	119	754	126	120	776	121	103
	EBIC	913	46	41	941	29	30	950	29	21
	QIC	244	388	368	247	369	384	249	366	385
R_0 =EX	EAIC	4	862	134	0	945	55	0	991	9
	EBIC	4	862	134	0	945	55	0	991	9
	QIC	141	666	193	147	730	123	135	766	99
R_0 =AR	EAIC	0	157	843	0	69	931	0	15	985
	EBIC	0	157	843	0	69	931	0	15	985
	QIC	138	262	600	97	248	655	103	221	676

The true exchangeable and AR-1 correlation matrices are parameterized with $\alpha = 0.5$. These results are obtained from 1000 simulation runs.

Simulation results from Pan (2001) and Pan & Connett (2002) are reorganized into Table 3.8 so that we can make further comparison. Pan (2001) generated data only with the EX structure, and computed MLE and AIC since the true distribution was known. We note that Pan's (2001) results on the performance of QIC (see the first row of Table 3.8) are very close to those presented in the sixth row of Table 3.7, which confirms our simulation study. By comparing the second row of Table 3.8 to the fourth row of Table 3.7, we see that the performance of EAIC is comparable to that of AIC yet without specifying a full-parametric

model. Pan & Connett (2002) simulated three bootstrap-based criteria, denoted by BOOT, BOOT2 and BCV, which minimize the predictive mean squared error. It is seen from Table 3.8 that these three criteria are not nearly as effective as EAIC or EBIC.

Table 3.8: Simulation results from Pan (2001), and Pan & Connett(2002)

			Working Correlation Structures R					
			$n = 50$			$n = 100$		
			IN	EX	AR	IN	EX	AR
Pan (2001)	R_0 =EX	QIC	138	678	184	140	721	139
		AIC	0	836	164	0	946	54
Pan & Connett (2002)	R_0 =IN	BOOT	48	22	30	38	30	32
		BOOT2	35	31	34	30	34	36
		BCV	40	30	27	38	25	37
	R_0 =EX	BOOT	19	56	25	21	47	32
		BOOT2	17	52	31	18	48	34
		BCV	6	65	29	12	51	37
	R_0 =AR	BOOT	14	30	56	19	21	60
		BOOT2	12	33	55	18	23	59
		BCV	5	32	63	13	17	70

Results from Pan (2001) and Pan & Connett (2002) were obtained from 1000 and 100 simulation replicates, respectively.

Finally, we add the full-model correlation structure ST to this study, and present relevant results in Table 3.9; these are very similar to those in Table 3.6. QIC cannot well distinguish the correct correlation structure from others, and always tends to over-parameterize the working correlation matrix. Although QIC chooses ST more often when R_0 is actually ST, it does not actually lead to more efficient estimation of β because the full model itself may not be efficient. We denote by $\hat{\beta}_G^{EAIC}$, $\hat{\beta}_G^{EBIC}$ and $\hat{\beta}_G^{QIC}$ three estimates chosen by EAIC, EBIC and QIC, respectively, in a single simulation run. When R_0 and $n = 100$, based on over 1000 simulation runs we find that $MSE(\hat{\beta}_G^{EAIC}) = 0.0996$, $MSE(\hat{\beta}_G^{EBIC}) = 0.0998$, and

$MSE(\hat{\beta}_G^{QIC}) = 0.109$ (These are not shown in Table 3.9). In general, EAIC and EBIC are more powerful than QIC in terms of achieving the goal of efficiency improvement.

Table 3.9: Continued comparison of EAIC, EBIC and QIC (binary response, $t = 3$)

		Working Correlation Structures R											
		$n = 50$				$n = 100$				$n = 200$			
		IN	EX	AR	ST	IN	EX	AR	ST	IN	EX	AR	ST
	EAIC	727	105	104	64	732	99	122	47	741	111	100	48
R_0 =IN	EBIC	916	33	43	8	945	25	27	3	976	11	11	2
	QIC	194	184	209	413	175	190	201	434	182	175	197	446
	EAIC	0	737	118	145	0	822	44	134	0	838	4	158
R_0 =EX	EBIC	0	817	118	65	0	913	70	17	0	968	16	16
	QIC	153	369	124	345	147	383	112	358	125	386	73	416
	EAIC	0	154	726	120	0	50	815	135	0	1	844	155
R_0 =AR	EBIC	0	156	782	62	0	60	917	23	0	10	981	9
	QIC	120	216	309	355	107	171	308	414	103	185	333	379
	EAIC	1	437	520	42	0	370	525	105	0	138	367	495
R_0 =ST	EBIC	1	446	534	19	0	417	581	2	0	341	652	7
	QIC	115	294	218	373	133	275	184	408	111	268	158	463

The true exchangeable and AR-1 structures are parameterized by $\alpha = 0.5$, and the true stationary structure has $\alpha = (0.5, 0.35, 0.2)^T$. These results are based on 1000 simulation runs.

3.5.4 Example

To illustrate the use of empirical likelihood GEE models, we apply EAIC and EBIC to data arising from a clinical trial of 59 patients suffering from epileptic seizures. This study was carried out by Leppik *et al.* (1985), and the entire data set is presented in Thall and Vail (1990). A baseline count of the number of epileptic seizures in an 8-week period prior to randomization was obtained for each patient. Patients were then randomized to receive either a placebo or the drug progabide, in addition to standard therapy. Counts of epileptic seizures in each of four successive 2-week periods were reported. The goal of the study is to answer the question: "Does progabide reduce the seizure rate?"

The response variable is a count, suggesting a Poisson regression model. Thus we use the log link function $\log(\mu_{ij}) = X_{ij}^T\beta$, and assume the marginal variance has the form $\text{var}(Y_{ij}) = \phi v(\mu_{ij}) = \phi\mu_{ij}$. Overdispersion present in the data is accounted for by ϕ. Let y_{hij} be the response at time j ($j = 0, 1, 2, 3, 4$) for the ith subject in treatment group h ($h = 0$ for placebo and $h = 1$ for progabide). We consider a model for the log seizure rate that includes baseline seizure rate (y_{hi0}), computed as the logarithm of 1/4 the 8-week baseline count ($\log(y_{hi0}/4)$), treatment group, time, and the interaction between treatment and time. This is an ANCOVA model that can be written as

$$\log(\mu_{hij}) = \lambda_{hj} + \beta \log(y_{hi0}), \quad h = 1, 2;\ i = 1, \ldots, n_h;\ j = 1, 2, 3, 4.$$

We first fit the above model with GEE, using correlation structures IN, EX, AR and ST. Empirical likelihood ratio is defined with the general correlation structure ST. EAIC and EBIC are obtained with each of the four sets of GEE estimates. We also calculate QIC using (3.12). Table 3.10 shows the results. Both EAIC and EBIC choose the stationary structure, meaning that neither exchangeability nor AR-1 are sufficient to describe the correlation structure. In contrast, QIC chooses the independence structure; this choice is not reliable as seen from our simulation results. After choosing the GEE with a stationary working correlation structure, one can use this model to examine the interaction between treatment and time, and the two main effects. It turns out that these effects are not significant, meaning that the drug progabide cannot effectively reduce the seizure rate.

Table 3.10: Example: epileptic seizures

	Working Correlation Structures			
	IN	EX	AR	ST
EAIC	108.744	20.762	23.852	18.000
EBIC	127.442	39.459	42.550	36.698
QIC	-1176.468	-1175.557	-1168.187	-1171.905

3.6 Conclusions and Future Directions

3.6.1 Conclusions

In this chapter we have reviewed two classes of methods for improving the estimation of regression parameters in GEE models. One approach is to extend the generalized estimating equations by replacing the inverse of the working correlation matrix with a linear combination of known basis matrices. However, since the set of basis matrices originates from an assumption about the working correlation structure, this manipulation of estimating equations does not add any new (or change) information to the original GEE model, and hence cannot actually lead to an estimator more efficient than the GEE estimator. The claim that the quadratic inference function approach based on extended estimating functions is superior to the GEE method has been disproved by at least empirical evidence. A more plausible way to achieve efficiency improvement is model selection, in particular selecting the optimal working correlation structure for a GEE model. Existing model selection methods for GEE models cannot integrate any information about the underlying correlation structure, and therefore are not powerful in choosing the optimal working correlation structure.

To compare a set of candidate correlation structures, we embed them in a general correlation structure and define a unified empirical likelihood ratio with the general structure. Then the ELR of a GEE model with any candidate structure can be obtained using the associated GEE estimates of β and α. We get two model selection criteria — EAIC and EBIC — by simply substituting the log empirical likelihood ratio for the log likelihood ratio in AIC and BIC, respectively. Our approach is easy to use in practice, and avoids computational issues encountered by the EIC proposed by Kolaczyk (1995). EAIC and EBIC are compared to existing model selection methods for GEE, including QIC (Pan, 2001) and three bootstrap-based criteria of minimum predictive mean squared error (Pan and Connett, 2002). Simulation studies conducted under various scenarios suggest that EAIC and EBIC are much more powerful than existing methods. Just like their parametric counterparts, EAIC tends to over-parameterize the working correlation structure. Considering that

a correlation structure with many nuisance parameters, even if correct, may not improve the efficiency of estimating β when it can be approximated by a more parsimonious structure, we recommend using EBIC.

3.6.2 Future Work

An EL-based generalized information criterion

Although it is shown that EAIC is quite effective when used to select the working correlation structure in a GEE, there are potential ways to further improve its performance. One possible way is to develop an EL-based information criteria that has a solid theoretical basis and can be applied to estimating equations in general. AIC is derived as an estimator of the Kullback-Leibler information of the true model with respect to the fitted model, under the assumptions that (i) the parameter is estimated by maximizing likelihood and (ii) the specified parametric family of distributions contains the true model. Similarly, the derivation of Kolaczyk's (1995) EIC requires maximum empirical likelihood estimates and the assumption that the estimating equation model under consideration is correctly specified even though it may not be the most parsimonious one. The second assumption of AIC and of EIC imply that the model to be evaluated will produce a consistent estimator for the true parameter. Akaike (1973) argued that $-2\log$(likelihood ratio) evaluated at a bad estimate (resulting from an incorrect model) will be significantly larger than would be expected from the chi-square approximation, and hence the incorrect model will be automatically excluded from being considered as optimal. Therefore, the second assumption seems reasonable. However, it is not known whether this is the reason why AIC, EIC and EAIC are likely to favor a larger model when a more parsimonious model suffices. To evaluate parametric models, Konishi and Kitagawa (1996) proposed a generalized information criterion which also estimates Kullback-Leibler information but relaxes the above two assumptions of AIC; only a functional estimator, which includes MLE as a special case, is required by this criterion.

In the context of estimating equations, of which GEE is a special case, we may develop an EL-based generalized information criteria in the spirit of Konishi and Kitagawa (1996). Estimating sub-model parameters by maximizing the full model ELR with additional sub-model constraints may encounter problems such as those discussed in Section 3.5.2. Since it is more convenient to estimate a sub-model parameter by solving the sub-model estimating equation whose dimension is equal to that of the sub-model parameter, it is desirable to replace the MELE by this alternative estimate. Note that the solution of a sub-model estimating equation is an M-estimator, and M-estimators are a special class of functional estimators. In addition, we wish to relax the assumption in EIC that estimating equation models are correct and yield consistent estimators. Recall that when we choose GEE models with different working correlation structures, the correlation parameter α is considered as part of θ that parameterizes the entire model. If the working correlation is misspecified, $\hat{\alpha}_G$ cannot be consistent for the true α. For example, it is impossible that a fitted exchangeable correlation structure is consistent for the true AR-1 structure. Thus, it is not reasonable to assume that $\hat{\theta}_G$ obtained with any GEE model is consistent for the true θ, although $\hat{\beta}_G$ is always consistent for β. In fact, only the unstructured correlation is guaranteed to be correct.

In Konishi and Kitagawa (1996), a specified parametric family of distributions $\{F(x|\theta) : \theta \in \Theta\}$ may not contain the unknown true distribution $G(x|\theta_0)$. Let $f(x|\theta)$ and $g(x|\theta_0)$ be densities of F and G, respectively. Different from (3.26), the Kullback-Leibler information is now expressed as

$$\begin{aligned} K(g(\theta_0), f(\hat{\theta})) &= \int \log\left(\frac{g(x|\theta_0)}{f(x|\hat{\theta})}\right) g(x|\theta_0)dx \\ &= \text{an unknown constant} - \int \log(f(x|\hat{\theta})\, dG(x|\theta_0), \end{aligned}$$

where $\hat{\theta}$ is a (random) estimator. Minimizing $K(g(\theta_0), f(\hat{\theta}))$ is equivalent to maximizing

$$\eta(G) = \int \log(f(x|\hat{\theta})\, dG(x|\theta_0),$$

which needs to be estimated since $G(x|\theta_0)$ is unknown. A simple estimator of $\eta(G)$ is given by substituting the empirical distribution $\hat{G}$ for G:

$$\eta(\hat{G}) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\hat{\theta}),$$

which has a bias $b(G) = E[\eta(G) - \eta(\hat{G})]$. Then a generalized information criterion is defined to be

$$GIC = -2\sum_{i=1}^{n} \log f(X_i|\hat{\theta}) + 2nb(G) = -2n\{\eta(\hat{G}) - b(G)\}, \tag{3.35}$$

whose expectation is equal to $-2nE[\eta(G)]$. When used in practice, $b(G)$ is estimated by $\hat{b}(G)$. Konishi and Kitagawa (1996) showed that $b(G)$ and $\hat{b}(G)$ in GIC have an explicit form if $\hat{\theta}$ is an M-estimator, and GIC can be further simplified to AIC if AIC's assumptions are retained.

Following Kolacyzk (1995), we use the empirical analog of Kullback-Leibler information in (3.28), *i.e.*,

$$\begin{aligned} K_{EL}(\theta_0, \hat{\theta}) &= \sum_{i=1}^{n} \log\left(\frac{w_i(\theta_0)}{w_i(\hat{\theta})}\right) w_i(\theta_0) \\ &= \text{a constant given the data} - \sum_{i=1}^{n} \log(w_i(\hat{\theta}))w_i(\theta_0), \end{aligned}$$

where $\hat{\theta}$ is obtained by solving a sub-model estimating equation. Then we can define empirical likelihood versions of $\eta(G)$, $\eta(\hat{G})$ and $b(G)$ as

$$\begin{aligned} \eta_{EL}(\theta_0) &= \sum_{i=1}^{n} \log(w_i(\hat{\theta}))w_i(\theta_0), \\ \hat{\eta}_{EL}(\theta_0) &= \frac{1}{n}\sum_{i=1}^{n} \log(w_i(\hat{\theta})), \\ \text{and} \quad b_{EL}(\theta_0) &= \hat{\eta}_{EL}(\theta_0) - \eta_{EL}(\theta_0), \end{aligned}$$

respectively. An EL-based generalized criterion is of the form

$$-2\sum_{i=1}^{n} \log w_i(\hat{\theta}) + 2nb(\theta_0).$$

Thus developing an EL-based generalized criterion boils down the problem of estimating $2nb_{EL}(\theta_0)$. Since $\{w_i(\hat{\theta})\}$ in $2nb_{EL}(\theta_0)$ can be obtained in the process of computing $\log \mathcal{R}^F(\hat{\theta})$, the major task is to approximate $\{w_i(\theta_0)\}$. Note that

$$w_i(\theta_0) = \frac{1}{n}\frac{1}{1 + (\lambda(\theta_0))^T g^F(X_i, \theta_0)}.$$

A possible way to tackle the problem is to expand $w_i(\theta_0)$ around $\hat{\theta}_0$, the solution to the full-model estimating equation, as we know $w_i(\hat{\theta}_0) = 1/n$. Since the expansion will involve $(\theta_0 - \hat{\theta}_0)$, we may use the functional Taylor expansion (Konishi and Kitagawa, 1996, page 888)

$$\hat{\theta}_0 = \theta_0 + \frac{1}{n}\sum_{i=1}^{n} IF(X_i; G), + O_p(n^{-1})$$

where $IF(X_i; G)$ is the influence function of functional T at distribution G and has a more explicit form for the M-estimator $\hat{\theta}_0$ (Hampel *et al.*, 1986, page 20).

Whether an EL-based generalized criterion can be successfully developed in this way depends on (at least) the simplest form of $\hat{b}_{EL}(\theta_0)$ one can get eventually. If $\hat{b}_{EL}(\theta_0)$ has a simple form that facilitates practical use, then the criterion should be examined for its effectiveness in terms of selecting the optimal model.

Bayesian interpretation of EBIC

We get the empirical likelihood version of BIC in (3.25) by substituting empirical likelihood for parametric likelihood in the original BIC. Considering that empirical likelihood and parametric likelihoods share many common asymptotic properties, we may expect that EBIC has a Bayesian interpretation. Specifically, Lazar (2003) extended the discussion of EL to the Bayesian setting, showing that it is possible to replace the parametric likelihood in Bayes theorem with empirical likelihood, and to obtain valid posterior inferences. These posterior distributions are asymptotically normal and combine the prior and the data in the same way as parametric Bayes. Thus, another future direction of research is to provide a Bayesian justification of EBIC along the line of the original BIC.

Alternative estimating equations

To compare different working correlation structures using empirical likelihood, our idea is to combine the p-dimensional regression parameter β and the s-dimensional correlation parameter α of the specified general correlation structure into $\theta = (\beta^T, \alpha^T)^T$, and then define a $(p+s)$-dimensional estimating function $g^F(Y_i, X_i, \theta)$ for θ; the first p components of $g^F(Y_i, X_i, \theta)$ are simply the left-hand side of GEE, and the last s components incorporate constraints (information) on α that are implied by the general correlation structure. In particular, in Section 3.5.1 $g^F(Y_i, X_i, \theta)_{(p+s)\times 1}$ is defined as in (3.20), where the s components associated with α originate from the method of moment estimator used in GEE. It is natural and convenient to use $g^F(Y_i, X_i, \theta)$ in (3.20), since it is exactly equivalent to the corresponding GEE in terms of estimation of β and α. Despite this "advantage", there are potentially better options for devising $g^F(Y_i, X_i, \theta)$.

Liang, Zeger and Qaqish (1992) noted that the method of moment estimator of α used in GEE has low asymptotic efficiency relative to the maximum likelihood estimator. This is not surprising, as α is treated as a nuisance parameter in GEE, and the exact form of the method of moment estimator depends on the assumed form of $R(\alpha)$ (Liang and Zeger, 1986). Other authors have proposed different GEE extensions that estimate β and α simultaneously and achieve greater efficiency in estimating α. Thus an alternative to $g^F(Y_i, X_i, \theta)$ in (3.20) can be devised using such a GEE extension. We conjecture that, if an estimating equation yields a more precise $\hat{\alpha}$ for the full-model (general) correlation structure, then ELR defined with this estimating equation may lead to a more accurate evaluation of each sub-model correlation structure, and consequently the correct sub-model correlation structure may have a higher chance to be chosen as the optimal by an EL-based model selection criterion.

Among GEE extensions, the extended generalized estimating equations (EGEE) proposed by Hall and Severini (1998) requires no more assumptions than the original GEE approach and yields $\hat{\alpha}$ with improved efficiency without sacrificing consistency of $\hat{\beta}$ under a misspecified working covariance structure. In particular, an EGEE was derived from the derivatives of an

extended quasi-likelihood function $Q^+(\mu, \alpha, Y) = \sum_{i=1}^n Q_i^+(\mu_i, \alpha, Y_i)$, with

$$Q_i^+(\mu_i, \alpha, Y_i) = Q_i(\mu_i, Y_i) + f(\alpha, Y_i), \tag{3.36}$$

where $Q_i(\mu_i, Y_i)$ is the quasi-likelihood function, if it exists, and is given by

$$Q_i(\mu_i, Y_i) = \int_{u(s)=Y}^{u(s)=\mu} (Y_i - \mu_i)^T \{V_i(u)\}^{-1} du(s). \tag{3.37}$$

Here, $\partial Q_i^+(\mu_i, \alpha, Y_i)/\partial\beta^T = \partial Q_i(\mu_i, Y_i)/\partial\beta^T = D_i^T V_i^{-1}(Y_i - \mu_i)$. The form $f(\alpha, Y_i)$ is determined by setting $E[\partial Q_i^+(\mu_i, \alpha, Y_i)/\partial\alpha^T] = 0$. The extended estimating equation results from the first derivatives of Q^+ with respect to β and α, respectively, and is given by $\sum_{i=1}^n g_{egee}^F(Y_i, X_i, \beta, \alpha) = 0$, where

$$g_{egee}^F(Y_i, X_i, \beta, \alpha) = \begin{pmatrix} D_i^T V_i^{-1}(Y_i - \mu_i) \\ -(Y_i - \mu_i)^T \frac{\partial V_i^{-1}}{\partial \alpha_1}(Y_i - \mu_i) + \mathrm{tr}\left(V_i \frac{\partial V_i^{-1}}{\partial \alpha_1}\right) \\ \vdots \\ -(Y_i - \mu_i)^T \frac{\partial V_i^{-1}}{\partial \alpha_{s+1}}(Y_i - \mu_i) + \mathrm{tr}\left(V_i \frac{\partial V_i^{-1}}{\partial \alpha_{s+1}}\right) \end{pmatrix}_{(p+s)\times 1}, \tag{3.38}$$

and V_i is the working covariance of Y_i that assumes the general correlation structure. Although the quasi-likelihood Q and hence the extended quasi-likelihood Q^+ may not exist for all choices of the covariance matrix V_i, the first derivatives of Q^+ in (3.38) provide valid estimating functions for β and α. Along the line of selecting working correlation structures, we will study the use of $g_{egee}^F(Y_i, X_i, \beta, \alpha)$ as an alternative to $g^F(Y_i, X_i, \beta, \alpha)$ in (3.20).

Chapter 4

Summary

This dissertation applies empirical likelihood to two different problems: quantile estimation for discrete data, and selecting the working correlation structure in GEE. Although these two problems arise in unrelated settings, there are common features that bring them into the framework of empirical likelihood.

First, parametric assumption is not desired or difficult to make. Sample quantiles are often used in practice to provide a preliminary summary of the data with no parametric assumptions. Even if a parametric model is to be built based on the data, it is still useful to obtain nonparametric estimates of quantiles as they may be used to check the goodness of the proposed parametric model. For example, Q-Q plots are a widely used tool for diagnosing differences between a fitted parametric distribution and an empirical distribution. If data are drawn from a discrete distribution, sample quantiles are not necessary good estimates of population quantiles. In this situation, we use empirical likelihood to summarize information available from the data without any parametric assumptions, and then propose a nonparametric alternative to the sample quantile estimator. This new estimator has been shown to be consistent. When analyzing discrete or categorical longitudinal data, researchers often resort to the semiparametric approach of GEE, as there are relatively few tractable multivariate distributions that can be used to model this type of data. Since parametric likelihood are absent in GEE models, many powerful model selection methods based on likelihood cannot be applied to GEEs. Even though quasi-likelihood (Pan, 2001) and approximate likelihood based on quasi-likelihood (Hanfelt and Liang, 1995) have been proposed for GEEs, the usage of these likelihoods has been limited partly because exact quasi-likelihood may not exist. In Chapter 3, we construct empirical likelihood for GEEs, and show that EL versions of AIC

and BIC are more powerful than methods that are based on quasi-likelihood or resampling procedures.

Second, information or constraints on parameters can be expressed as estimating equations. It is seen from these two applications that empirical likelihood can be extended to various data settings as long as estimating equations are defined appropriately. In quantile estimation, the estimating equation $E[1(X \leq \theta_p) - p] = 0$ is not smooth in the parameter θ_p, and therefore Qin and Lawless' (1994) results on properties of the maximum empirical likelihood estimator do not apply. Nevertheless, we can study properties of the MELE for quantiles using its relationship with the sample quantile estimator. To improve the accuracy of interval estimation, one can smooth the estimating equation using kernel methods (*e.g.*, Chen and Hall, 1993). Normally, empirical likelihood requires that data be independent, and estimating equations are defined for independent individual observations. In the longitudinal data setting, observations made on the same subject (*i.e.*, within the same cluster) are correlated, but clusters are assumed to be independent of one another. Therefore, to construct a valid empirical likelihood ratio for longitudinal data, estimating equations need to be defined on clusters rather than individual observations. In the GEE approach, the regression parameter β is modeled as the solution of generalized estimating equations that are defined on clusters, so a GEE can be used directly as the estimating equation for empirical likelihood. However, to compare GEEs with different working correlation structures using empirical likelihood, it is necessary that the estimating equation contain constraints on the correlation parameter α. Thus, we form a new vector of estimating equation by adding equation components that explicitly estimate α to the original GEE.

In the literature, empirical likelihood has been extended to various data settings mostly as an alternative way to construct confidence intervals and hypotheses testing. It is shown in both theoretical and applied work on empirical likelihood that EL-based confidence intervals (or regions) have desirable properties, especially when compared to confidence intervals based on normal approximation (see for example, Chen et al., 2003). Nevertheless, we see from

the two applications here that the use of empirical likelihood is not limited to confidence intervals. In quantile estimation, we propose an EL-based categorization procedure which not only helps determine the shape of the true discrete distribution at level p, but also provides a way of formulating a consistent estimator. We show in the context of GEEs, empirical likelihood can be used in place of parametric likelihood to form model selection criteria. It is reasonable to expect that EL-based model selection criteria will be effective in selecting semiparametric models in the form of general estimating equations. In summary, empirical likelihood is a flexible and powerful method just as its parametric counterpart.

Although the majority of work on EL has focused on the case where data are continuous, we show that empirical likelihood is also very useful for analyzing discrete and categorical data that may be as simple as univariate data (such as in quantile estimation), or as complex as repeated measures with a number of covariates. Along the line of discrete and categorical data, another potential application of EL is analyzing contingency tables.

In recent years, there have been remarkable developments in methods for longitudinal data analysis, due to the prominent role of longitudinal data in the behavioral, health and medical sciences. In addition to commonly used models (*e.g.*, GEE and linear mixed effects models), there are various semiparametric models including the semiparametric partially linear regression model in (3.13) (Zeger and Diggle, 1994) and the varying coefficient model in (3.16) (for example, Xue and Zhu, 2007) mentioned in Section 3.4. Therefore, there may be various extensions of empirical likelihood to longitudinal data analysis, depending on the specific data type and the model for which EL is to be defined. We have focused on empirical likelihood for GEE, which accounts for the within-subject correlation via the working correlation matrix in GEE. Other authors have considered EL for models (3.13) and (3.16); their approaches ignore the within-subject correlation and hence might not be optimal. Notice that there is a subject-specific zero mean stochastic process $\varepsilon_i(t)$ in both (3.13) and (3.16), describing the within-subject serial correlation. Thus, we may conjecture that existing empirical likelihoods for models (3.13) and (3.16) might be further improved if $\varepsilon_i(t)$ is somehow

incorporated into the estimating equations. This problem might be generalized to a much broader topic, namely, how to construct a more informative empirical likelihood for longitudinal data (or more general, dependent data) based on basic model assumptions.

Bibliography

[1] Adimari, G. (1998). An empirical likelihood statistic for quantiles. *Journal of Statistical Computation and Simulation*, **60**, 85-95.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F. (eds.), *Second International Symposium on Information Theory*, pp. 267-281. Budapest: Akademiai Kiado.

[3] Boos, D.D. (1992). On generalized score tests. *The American Statistician*, **46**, 327-333.

[4] Chen, J., Chen, S.Y. and Rao, J.N.K. (2003). Empirical likelihood confidence intervals for the mean of a population containing may zero values. *The Canadian Journal of Statistics*, **31**, 53-68.

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

[6] David, H.A. (1981). *Order Statistics* (2nd ed.). New York: Wiley.

[7] Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *The Annals of Statistics*, **7**, 1-26.

[8] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

[9] Fitzmaurice, G.M. (1995). A caveat concerning independence estimating equations with multiple multivariate binary data. *Biometrics*, **51**, 309-517.

[10] González-Barrios, J.M. and Rueda, R. (2001). On convergence theorems for quantiles. *Communications in Statistics: Theory and Methods*, **30**, 943-955.

[11] Hall, D.B. and Severini, T.A. (1998). Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association*, **93**, 1365-1375.

[12] Hall, P. and Martin, M.A. (1989). A note on the accuracy of bootstrap percentile method confidence intervals for a quantile. *Statistics & Probability Letters*, **8**, 197-200.

[13] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics. The Approach Based on Influence Functions*. New York: Wiley.

[14] Hanfelt, J.J. and Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, **82**, 461-77.

[15] Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029-1054.

[16] Heyde, C.C. (1997), *Quasi-Likelihood And Its Application*. New York: Springer-Verlag.

[17] Ho, Y.H.S. and Lee, S.M.S. (2005). Smoothed bootstrap confidence intervals for population quantiles. *The Annals of Statistics*, **33**, 437-462.

[18] Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics*, **25**, 2084-2102.

[19] Kolaczyk, E.D. (1994). Empirical likelihood for generalized linear models. *Statistica Sinica*, **4**, 199-218.

[20] Kolaczyk, E.D. (1995). An information criterion for empirical likelihood with general estimating equations. Technical Report 417, Department of Statistics, The University of Chicago.

[21] Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika*, **83**, 875-890.

[22] Lazar, N.A. (2003). Bayesian empirical likelihood. *Biometrika*, **90**, 319-326.

[23] Leppik, I. E. (1985). A double-blind crossover evaluation of progabide in partial seizures. *Neurology*, **35**, 285.

[24] Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

[25] Liang, K.-Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society*, Ser. B, **45**, 3-40.

[26] Machado, J.A.F. and Santos Silva, J.M.C. (2005). Quantiles for counts. *Journal of the American Statistical Association*, **100**, 1226-1237

[27] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. New York: Chapman & Hall/CRC.

[28] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

[29] Moolgavkar, S.H. and Venzon, D.J. (1986). Confidence regions for case-control and survival studies with general relative risk functions. In *Modern Statistical Methods in Chronic Disease Epidemiology*, Ed. Moolgavkar, S.H. and Prentice, R.L., pp. 104-20. New York: John Wiley.

[30] Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.

[31] Owen, A.B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, **18**, 90-120.

[32] Owen, A.B. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, **19**, 1725-1747.

[33] Owen, A.B. (2001). *Empirical Likelihood*. New York: Chapman & Hall/CRC.

[34] Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120-125.

[35] Pan, W. and Connett, J.E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica*, **12**, 475-490.

[36] Pilla, R.S. and Loader, C. (2005). On large-sample estimation and testing via quadratic inference functions for correlated data. *E-print: arXiv:math.ST/0505360*.

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

[38] Qin, Y.S. and Wu, Y. (2001). An estimator of a conditional quantile in the presence of auxiliary information. *Journal of Statistical Planning and Inference*, **99**, 59-70.

[39] Qu, A., Lindsay, B. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, **87**, 823-836.

[40] Reiss, R.-D. (1980). Estimation of quantiles in certain nonparametric models. *The Annals of Statistics*, **8**, 87-105.

[41] Rotnitzky, A. and Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, **77**, 485-497.

[42] Schaumberger, N. (1988). A general form of the arithmetic-geometric mean inequality via the mean value theorem. *The College Mathematics Journal*, **19**, 172-173.

[43] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

[44] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc..

[45] Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657-671.

[46] van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[47] Variyath, A.M., Chen, J. and Abraham, B. (2007). Empirical likelihood based variable selection. Working Paper Series, 2007-06, Department of Statistics and Actuarial Science, University of Waterloo.

[48] Xue, L., and Zhu, L. (2007). Empirical likelihood for a varying coefficient model with longitudinal data. *Journal of the American Statistical Association*, **102**, 642-654.

[49] Yang, S.-S. (1985). A smooth nonparametric estimator of a quantile function. *Journal of the American Statistical Association*, **80**, 1004-1011.

[50] You, J., Chen, G. and Zhou, Y. (2006). Block empirical likelihood for longitudinal partially linear regression models. *The Canadian Journal of Statistics*, **34**, 79-96.

[51] Zeger, S.L. and Diggle, P.J. (1994). Semiparametric model for longitudinal data with application to CD4 cell numbers in HIV seroconvertiers. *Biometric*, **50**, 689-699.

[52] Zhang, B. (1995). M-estimation and quantile estimation in the presence of auxiliary information. *Journal of Statistical Planning and Inference*, **44**, 77-94.

[53] Zhao, Y. and Jian, W. (2007). Analysis of longitudinal data in the case-control studies via empirical likelihood. *Communications in Statistics — Simulation and Computation*, **36**, 565-578.

[54] Zhou, W. and Jing, B. (2003). Adjusted empirical likelihood method for quantiles. *Annals of the Institute of Statistical Mathematics*, **55**, 689-703.

Appendix A

Proofs

A.1 Proof of Result 2.2.1

Let $\{k_n\}$ be a sequence of positive integers that is determined by

$$k_n = \begin{cases} np & \text{if } \hat{\theta}_{pn} = X_{np:n} \\ [np] & \text{if } \hat{\theta}_{pn} = X_{[np]:n} \\ [np]+1 & \text{if } \hat{\theta}_{pn} = X_{[np]+1:n} \end{cases},$$

and therefore $\hat{\theta}_{pn} = X_{k_n:n}$. If $k_n = pn$, then $k_n/n = p$; if $k_n = [np]$, then $(np-1)/n < k_n/n \leq (np)/n$ and hence $k_n/n = p + o(1/\sqrt{n})$; if $k_n = [np]+1$, then $(np)/n < k_n/n \leq (np+1)/n$ and $k_n/n = p + o(1/\sqrt{n})$. Thus, the condition

$$k_n/n = p + o(1/\sqrt{n})$$

is satisfied in any case. Then by Lemma 21.7 in van der Vaart(1998), when the underlying distribution F_0 is continuous at θ_p, we have $\sqrt{n}\left(X_{k_n:n} - \tilde{\theta}_{pn}\right) \xrightarrow{p} 0$, *i.e.*

$$\sqrt{n}\left(\hat{\theta}_{pn} - \tilde{\theta}_{pn}\right) \xrightarrow{p} 0.$$

It is known that $\tilde{\theta}_{pn} \xrightarrow{a.s.} \theta_p$ for a continuous distribution. (See Serfling (1980), page 75). Then for any $\epsilon > 0$,

$$\begin{aligned} \Pr\left(\left|\hat{\theta}_{pn} - \theta_p\right| > \epsilon\right) &\leq \Pr\left(\left|\hat{\theta}_{pn} - \tilde{\theta}_{pn}\right| + \left|\tilde{\theta}_{pn} - \theta_p\right| > \epsilon\right) \\ &\leq \Pr\left(\left|\hat{\theta}_{pn} - \tilde{\theta}_{pn}\right| > \epsilon/2 \text{ or } \left|\tilde{\theta}_{pn} - \theta_p\right| > \epsilon/2\right) \\ &\leq \Pr\left(\left|\hat{\theta}_{pn} - \tilde{\theta}_{pn}\right| > \epsilon/2\right) + \Pr\left(\left|\tilde{\theta}_{pn} - \theta_p\right| > \epsilon/2\right) \\ &\to 0, \quad \text{as } n \to \infty. \end{aligned}$$

Therefore, the MELE $\hat{\theta}_{pn}$ is a consistent estimator for θ_p if F_0 is continuous. □

A.2 Proof of Result 2.3.1

First consider part (a). $\hat{\theta}^Y_{pn} \leq \tilde{\theta}^Y_{pn}$ by (2.11) and (2.12). As discussed in the proof of Result 2.2.1, $\sqrt{n}\left(\hat{\theta}^Y_{pn} - \tilde{\theta}^Y_{pn}\right) \xrightarrow{p} 0$. Now consider for $t < 0$, since $\hat{\theta}^Y_{pn} \leq \tilde{\theta}^Y_{pn}$,

$$\begin{aligned}
& \Pr \left(\frac{n^{1/2}\left(\hat{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y -)}{\partial y}} \leq t \right) \\
= & \Pr \left(\frac{n^{1/2}\left(\hat{\theta}^Y_{pn} - \tilde{\theta}^Y_{pn}\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y -)}{\partial y}} + \frac{n^{1/2}\left(\tilde{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y -)}{\partial y}} \leq t \right) \\
\geq & \Pr \left(\frac{n^{1/2}\left(\tilde{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y -)}{\partial y}} \leq t \right). \qquad \text{(A.1)}
\end{aligned}$$

By part (i) of Theorem A of Serfling (1980, page 77),

$$\lim_{n\to\infty} \Pr \left(\frac{n^{1/2}\left(\tilde{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y -)}{\partial y}} \leq t \right) = \Phi(t).$$

Taking the limits of both sides of inequality (A.1), we get

$$\lim_{n\to\infty} \Pr \left(\frac{n^{1/2}\left(\hat{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y -)}{\partial y}} \leq t \right) \geq \Phi(t). \qquad \text{(A.2)}$$

On the other hand, for any $\varepsilon > 0$ such that $t + \varepsilon < 0$,

$$\begin{aligned}
& \lim_{n\to\infty} \Pr\left(\frac{n^{1/2}\left(\hat{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le t \right) \\
= & \lim_{n\to\infty} \Pr\left(\frac{n^{1/2}\left(\hat{\theta}^Y_{pn} - \tilde{\theta}^Y_{pn}\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} + \frac{n^{1/2}\left(\tilde{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le t \right) \\
\le & \lim_{n\to\infty} \Pr\left(\frac{n^{1/2}\left(\hat{\theta}^Y_{pn} - \tilde{\theta}^Y_{pn}\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le -\varepsilon \text{ or } \frac{n^{1/2}\left(\tilde{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le t + \varepsilon \right) \\
\le & \lim_{n\to\infty} \Pr\left(\frac{n^{1/2}\left(\hat{\theta}^Y_{pn} - \tilde{\theta}^Y_{pn}\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le -\varepsilon \right) \\
& \quad + \lim_{n\to\infty} \Pr\left(\frac{n^{1/2}\left(\tilde{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le t + \varepsilon \right) \\
= & \lim_{n\to\infty} \Pr\left(\frac{n^{1/2}\left(\tilde{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le t + \varepsilon \right).
\end{aligned}$$

Letting $\varepsilon \to 0$, we have

$$\lim_{n\to\infty} \Pr\left(\frac{n^{1/2}\left(\hat{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le t \right) \le \Phi(t). \tag{A.3}$$

Combining (A.2) and (A.3), we get

$$\lim_{n\to\infty} \Pr\left(\frac{n^{1/2}\left(\hat{\theta}^Y_{pn} - \theta^Y_p\right)}{[p(1-p)]^{1/2} \Big/ \frac{\partial F_0^Y(\theta_p^Y-)}{\partial y}} \le t \right) = \Phi(t).$$

Parts (b) and (c) can be shown using similar arguments. □

A.3 Proof of Result 2.3.3

Suppose that F_0^X is supported on $\{x_1, x_2, \ldots\}$ with $x_1 < x_2 < \ldots$, and that $P_j = Pr(X \le x_j)$. We now introduce random variables L_j and U_j $(j = 1, 2, \ldots)$, which are the smallest and the largest values of l satisfying $X_{(l)} = x_j$ (or equivalently, the indices of the smallest and

largest order statistics of Y that are generated by $X = x_j$). For example, a random sample from an unknown Poisson distribution contains sorted values

$$\{0, 1, 1, 1, 2, 2, 2, 2, 3, 3, \ldots\}$$

and the jittered sample is

$$\{0.3, 1.2, 1.6, 1.9, 2.1, 2.3, 2.4, 2.8, 3.1, 3.5, \ldots\},$$

then

$$L_1 = 1,\ U_1 = 1,\ Y_{(L_1)} = 0.3,\ Y_{(U_1)} = 0.3,$$
$$L_2 = 2,\ U_2 = 4,\ Y_{(L_2)} = 1.2,\ Y_{(U_2)} = 1.9,$$

and so on. For any $j \geq 1$,

$$U_j = \#\{X_i \leq x_j\} \sim Binomial(n, P_j)$$

and $U_j/n \xrightarrow{p} P_j$; if we define $P_0 \equiv 0$ and $U_0 \equiv 0$, then

$$L_j = \begin{cases} U_{j-1}, & \text{with probability } q_{j0} = (P_k - P_{k-1})^n, \\ U_{j-1} + 1, & \text{with probability } q_{j1} = 1 - q_{j0}, \end{cases}$$

where q_{j0} is the probability that the value x_j does not appear in the sample $\{X_i, i = 1, \ldots, n\}$. It can be verified that $L_j/n \xrightarrow{p} P_{j-1}$.

Part (a)

Without loss of generality, we assume $P_{k-1} < p < P_k$ for some integer $k > 1$, so $\theta_p^X = x_k$. To show $h_l \xrightarrow{p} 1$ and $h_u \xrightarrow{p} 1$, it is equivalent to show $C_l = \mathcal{R}^Y(Y_{(L)}) \xrightarrow{p} 0$ and $C_u = \mathcal{R}^Y(Y_{(U)}) \xrightarrow{p} 0$.

By the definition of L and U (Section 2.3.3), the two events $[\hat{\theta}_{pn}^X = x_j]$ and $[L = L_j, U = U_j]$ are equivalent. Thus, it follows from $\lim_{n\to\infty} \Pr(\hat{\theta}_{pn}^X = x_k) = 1$ that $\lim_{n\to\infty} \Pr(L = L_k, U = U_k) = 1$, which implies $L/n - L_k/n \xrightarrow{p} 0$. Therefore, $L/n \xrightarrow{p} P_{k-1}$ since $L_k/n \xrightarrow{p} P_{k-1}$. From (2.9), we can write $C_l = [g(L/n)]^n$, where

$$g(t) = \left(\frac{p}{t}\right)^t \left(\frac{1-p}{1-t}\right)^{1-t}.$$

Due to the continuity of $g(\cdot)$, $g(L/n) \xrightarrow{p} g(P_{k-1})$. By a general form of the Arithmetic-Geometric Mean Inequality (see for example, Schaumberger 1988), for $0 < t < 1$,

$$g(t) \leq t \cdot \frac{p}{t} + (1-t) \cdot \frac{1-p}{1-t} = 1,$$

where the equality holds only when $t = p$. Therefore, $g(P_{k-1}) < 1$, and hence $g(L/n) < 1$ with probability tending to 1. We can further conclude that $C_l = [g(L/n)]^n \xrightarrow{p} 0$. Similar arguments can be used to show $U/n \xrightarrow{p} P_k$ and then $C_u \xrightarrow{p} 0$.

Part (b)

We now assume $p = P_k$ for some $k > 1$, so $\theta_p^X = x_k$ and $\theta_p^Y = x_{k+1}$ (see case (ii) in Section 2.3.2). Since $\lim_{n\to\infty} \Pr(\hat{\theta}_{pn}^X = x_k) = \lim_{n\to\infty} \Pr(\hat{\theta}_{pn}^X = x_k + 1) = 0.5$, we may consider the only two possible outcomes of $\hat{\theta}_{pn}^X$ for large n: x_k and x_{k+1}. If $\hat{\theta}_{pn}^X = x_k$, then

$$\begin{aligned} L = L_k, \quad & C_l = \mathcal{R}(Y_{(L_k)}) = [g(L_k/n)]^n, \\ U = U_k, \quad & C_u = \mathcal{R}(Y_{(U_k)}) = [g(U_k/n)]^n. \end{aligned}$$

If $\hat{\theta}_{pn}^X = x_k + 1$, then

$$\begin{aligned} L = L_{k+1}, \quad & C_l = \mathcal{R}(Y_{(L_{k+1})}) = [g(L_{k+1}/n)]^n, \\ U = U_{k+1}, \quad & C_u = \mathcal{R}(Y_{(U_{k+1})}) = [g(U_{k+1}/n)]^n. \end{aligned}$$

Note that $g(L_k/n) \xrightarrow{p} g(P_{k-1}) < 1$, $g(U_k/n) \xrightarrow{p} g(P_k) = g(p) = 1$, $g(L_{k+1}/n) \xrightarrow{p} g(P_k) = 1$ and $g(U_{k+1}/n) \xrightarrow{p} g(P_{k+1}) < 1$. It follows that $g(L_k/n) < g(U_k/n)$ and $g(L_{k+1}/n) > g(U_{k+1}/n)$, both with probability tending to 1. Thus for large n,

$$\hat{\theta}_{pn}^X = \begin{cases} x_k, \text{ then } C_l = [g(L_k/n)]^n < C_u = [g(U_k/n)]^n, \\ x_k + 1, \text{ then } C_l = [g(L_{k+1}/n)]^n > C_u = [g(U_{k+1}/n)]^n. \end{cases}$$

both with conditional probability (given $\hat{\theta}_{pn}^X = x_k$ or x_{k+1}) tending to 1. Now using arguments in part (a), one can verify that $\min(C_l, C_u) \xrightarrow{p} 0$, *i.e.*, $\max(h_l, h_u) = \Pr\{\chi^2(1) \leq -2\log(\min(C_l, C_u))\} \xrightarrow{p} 1$.

Next, consider $\min(h_l, h_u) = \Pr\{\chi^2(1) \leq -2\log(\max(C_l, C_u))\}$. For convenience, let

$$-2\log(\max(C_l, C_u)) = -2nf(W), \tag{A.4}$$

where $f(W) = W \log\left(\frac{p}{W}\right) + (1-W)\log\left(\frac{1-p}{1-W}\right)$ and

$$W = \begin{cases} U_k/n & \text{if } \hat{\theta}^X_{pn} = x_k \\ L_{k+1}/n = U_k/n + 1/n & \text{if } \hat{\theta}^X_{pn} = x_{k+1} \end{cases},$$

for n large. Recall that

$$-2\log(\mathcal{R}(\theta^Y_p)) = -2nf(F_n(\theta^Y_p)). \tag{A.5}$$

Now we show in three steps that (A.4) and (A.5) are asymptotically equivalent. First, when $p = P_k$, $U_k/n = \#\{X_i \le x_k = \theta^X_p\}/n = \#\{Y_i \le \theta^Y_p\}/n = F_n(\theta^Y_p)$ and hence $W - F_n(\theta^Y_p) = O(n^{-1})$. Second, by the Central Limit Theorem, $F_n(\theta^Y_p) = p + O_p(n^{-1/2})$. Third, we have the following expansion:

$$\begin{aligned} f(W) &= f(F_n(\theta^Y_p)) + f'(F_n(\theta^Y_p))(W - F_n(\theta^Y_p)) + O(n^{-2}) \\ &= f(F_n(\theta^Y_p)) + \left[f'(p) + f''(F_n(\theta^Y_p))(F_n(\theta^Y_p) - p) + o_p(n^{-1/2})\right](W - F_n(\theta^Y_p)) \\ &\quad + O(n^{-2}) \\ &= f(F_n(\theta^Y_p)) + o_p(n^{-1}). \end{aligned}$$

Therefore, $[-2\log(\max(C_l, C_u))] - [-2\log(\mathcal{R}(\theta^Y_p))] \xrightarrow{p} 0$. It is known from Owen (1988) that $-2\log(\mathcal{R}(\theta^Y_p)) \xrightarrow{d} \chi^2(1)$. Then by the Probability Integral Transformation Theorem,

$$\Pr\left(\chi^2(1) \le -2\log(\mathcal{R}(\theta^Y_p))\right) \xrightarrow{d} U(0,1). \tag{A.6}$$

Combining (A.6) with the asymptotic equivalence of $-2\log(\max(C_l, C_u))$ and $-2\log(\mathcal{R}(\theta^Y_p))$, we get

$$\min(h_l, h_u) = \Pr\left(\chi^2(1) \le -2\log(\max(C_l, C_u))\right) \xrightarrow{d} U(0,1).$$

□