Properties of Empirical and Adjusted Empirical Likelihoods

by

Yi Huang

B.Sc., University of Science and Technology of China, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2010

Abstract

Likelihood based statistical inferences have been advocated by generations of statisticians. As an alternative to the traditional parametric likelihood, empirical likelihood (EL) is appealing for its nonparametric setting and desirable asymptotic properties.

In this thesis, we first review and investigate the asymptotic and finite-sample properties of the empirical likelihood, particularly its implication to constructing confidence regions for population mean. We then study the properties of the adjusted empirical likelihood (AEL) proposed by Chen et al. (2008). The adjusted empirical likelihood was introduced to overcome the shortcomings of the empirical likelihood when it is applied to statistical models specified through general estimating equations. The adjusted empirical likelihood preserves the first order asymptotic properties of the empirical likelihood and its numerical problem is substantially simplified.

A major application of the empirical likelihood or adjusted empirical likelihood is the construction of confidence regions for the population mean. In addition, we discover that adjusted empirical likelihood, like empirical likelihood, has an important monotonicity property.

One major discovery of this thesis is that the adjusted empirical likelihood ratio statistic is always smaller than the empirical likelihood ratio statistic. It implies that the AEL-based confidence regions always contain the corresponding EL-based confidence regions and hence have higher coverage probability. This result has been observed in many empirical studies, and we prove it rigorously.

We also find that the original adjusted empirical likelihood as specified by Chen et al. (2008) has a bounded likelihood ratio statistic. This may result in confidence regions of infinite size, particularly when the sample size is small. We further investigate approaches to modify the adjusted empirical likelihood so that the resulting confidence regions of population mean are always bounded.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Foremost I would like to express my deep gratitude to my advisor, Professor Jiahua Chen. His inspiration, guidance, encouragement and insight helped me through the last two years.

I am grateful to Dr. Matías Salibián-Barrera for serving on my examining committee and for his valuable comments and suggestions.

Finally, I would like to give my heartful appreciation and gratitude to my parents for their support and encouragement. This thesis is dedicated to them.

Chapter 1

Introduction

Likelihood based statistical inferences have been advocated by generations of statisticians. Let us illustrate the likelihood approach through the problem of modeling the randomness of the wind speed which is an important covariate in weather forecasting. In meteorology, the Weibull distribution with shape and scale parameters is used to model the distribution of wind speed (Corotis et al., 1978; Lun and Lam, 2000); that is, we postulate that the distribution of the wind speed is Weibull with two unspecified parameter values. Suppose we are given a set of observations of the wind speed, and it is reasonable to assume that they are a random sample from a Weibull distribution. Based on this assumption, we may calculate the probability of obtaining the observed data, which is a function of these two parameters. This function of the parameters is called the likelihood function. The likelihood function is an effective means of summarizing the information about the unknown values of the parameters contained in the data: (1) the values that maximize the likelihood function are often used as point estimates of the unknown parameters, which are called the maximum likelihood estimates (MLEs); (2) the likelihood function can be used to perform statistical tests for hypothesis on the parameters, and to construct an confidence region for the parameters. These likelihood based statistical inferences possess many optimality properties under regularity conditions: (1) MLE is asymptotically efficient in many senses, and give intuitively best explanation of the data; (2) likelihood based statistical tests and confidence intervals or confidence regions have good asymptotic and small-sample properties; (3) likelihood is convenient for combining information from several data sources, and incorporating knowledge arising from outside of data, such as the domain and a prior distribution of the parameter(s).

Traditionally, the likelihood is defined through a pre-specified parametric model. However, the choice of the parametric model in some applications can be a difficult issue. In the previous example, the Weibull distribution is widely used to characterize the wind speed because it has been found to fit a wide collection of wind data in many empirical studies. If the true distribution of the wind speed cannot be fit well by a Weibull distribution, the optimality properties of the likelihood approach will be in question. In comparison, the nonparametric methods for statistical inference do not require specific parametric assumptions on the shape of the population distribution. Among these methods, the empirical likelihood (EL) approach proposed by Owen (1988, 1990) has gained increasing popularity. This approach retains a likelihood setting without activating a parametric assumption, and shares many desirable properties with the parametric likelihood.

In this thesis, we review and investigate the asymptotic and finite-sample properties of empirical likelihood. In Chapter 2, we first present a short summary of the properties of the parametric likelihood, followed by an introduction to the empirical likelihood. The profile empirical likelihood is then introduced for population mean and parameters defined through general estimating equations. We also discuss the numerical algorithm for computing the empirical likelihood and some asymptotic properties of the empirical likelihood. One of major successes of the empirical likelihood is its easiness to construct approximate confidence regions for parameter of interest. The EL-based confidence region possesses many advantages: it has a data-driven shape; it is invariant under parameter transformation; and it is range respecting.

On the other hand, the empirical likelihood method has a few shortcomings. The EL-based confidence regions often have lower than specified coverage probabilities, particularly when the sample size is small. This problem can be alleviated through Bartlett correction. However, the confident region of the population mean, for instance, is confined within the convex hull of the data. In some cases, even the convex hull of the data does not have large enough coverage probability of the population mean. In addition, the empirical likelihood is not defined at certain parameter values which may occur especially when they are defined through general estimating equations. Consequently, the empirical likelihood approach may fail to make a sensible inference in a particular application.

To overcome these shortcomings of the EL approach, Chen et al. (2008) proposed an adjusted empirical likelihood (AEL). The adjusted empirical likelihood is well defined on all parameter values defined through estimating equations. It shares the same desirable first-order asymptotic properties with the empirical likelihood. Its numerical computation is much simpler and faster. In Chapter 3, we first introduce and present some properties of the adjusted empirical likelihood. We further investigate some finite-sample properties of the adjusted empirical likelihood. One major discovery is that the adjusted empirical likelihood ratio statistic is always smaller than that of the empirical likelihood. Consequently, the AEL-based confidence regions always contain the corresponding EL-based confidence regions. Thus, it effectively rectifies the under-coverage problem suffered by the empirical likelihood when the sample size is not large. We also discovered that the adjusted empirical likelihood has a monotonicity property. Because of this, the AEL-based confidence region for population mean is star-shaped. It also enables us to design a simple algorithm for computing confidence regions of multivariate population mean. It also leaves us an open question whether the AEL-based confidence region for population mean is convex. In addition, we find the original recipe of the adjusted empirical likelihood given by Chen et al. (2008) results in bounded likelihood ratio statistic. As a result, the AEL-based confidence regions can be unbounded. However, this problem can be easily fixed. We propose one possible modification to the adjusted empirical likelihood so that the corresponding likelihood ratio statistic becomes unbounded.

In Chapter 4, We empirically examine the ability of the foregoing methods to statistical inference about population mean. Certain kinds of setting are considered to investigate the finite-sample performances of the foregoing methods.

Chapter 2

Empirical Likelihood

Empirical likelihood (EL) is a nonparametric analogue of the classical parametric likelihood. The empirical likelihood method is first formalized in the pioneering works of Owen (1988, 1990) for statistical inference on the population mean. Qin and Lawless (1994) generalize empirical likelihood to the case where parameters are defined through general estimating equations. We call this method "empirical" because the empirical distribution, which assigns equal point mass on the data point, plays a key role in the setting of this method.

The empirical likelihood method provides a versatile approach that may be applied to perform inference for a wide variety of parameters of interest, and has been employed in a number of different areas of statistics. Qin and Zhang (2007) apply the empirical likelihood method to make a constrained likelihood estimation of mean response in missing data problems. Chen et al. (2003) consider constructing EL-based confidence intervals for the mean of a population containing many zero values in the area of survey. Chen et al. (2002) design an EL-based algorithm to determine design weights in surveys that meet pre-specified range restrictions. Chen and Sitter (1999) develop a pseudo empirical likelihood approach to incorporating auxiliary information into estimates from complex surveys. Chen et al. (2009) examine the performance of EL-based confidence intervals for copulas. Chan and Ling (2006) develop an empirical likelihood ratio test for GARCH model in time series. Qin and Zhou (2005) propose an empirical likelihood approach for constructing confidence intervals for the area under the ROC curve. Nordman and Caragea (2008) present a spatial blockwise empirical likelihood method for estimating variogram model parameters in the analysis of spatial data on a grid. And many more varied topics.

In Section 2.1, we briefly review the parametric likelihood inferences. Section 2.2, 2.3 and 2.4 are contributed to summarizing the setting of empirical likelihood. In Section 2.5, some asymptotic properties of empirical likelihood are presented; we focus on results related to constructing approximate confidence regions for parameter of interest. We also present many finite-sample properties of the EL-based confidence region. In the end, we point out the limitations of the empirical likelihood method and the corresponding remedies in Section 2.6, which lead to the subject of Chapter 3.

2.1 Parametric Likelihood

Let $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ be a collection of probability density function with respect to some σ-finite measure, where $\theta \in \mathcal{R}^p$ is a parameter that uniquely determines the form of the density and Θ is a p-dimensional set of possible values for θ. Suppose a random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is generated from one distribution of this probability family. Given that $\mathbf{X} = \mathbf{x}$, the likelihood function of θ is defined as

$$L_n(\theta \,|\, \mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta).$$

The likelihood function is interpreted as the probability of obtaining the observed sample if the parameter value equals θ. Hence, the likelihood function provides a way to measure the plausibility of different parameter values. If we compare the values of the likelihood function at two parameter values θ_1 and θ_2 and find that

$$L_n(\theta_1 \,|\, \mathbf{x}) > L_n(\theta_2 \,|\, \mathbf{x}),$$

then the sample we observed is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$. That is, θ_1 is a more plausible value of θ than is θ_2.

Take the wind speed example, where the two-parameter Weibull distribution is postulated. The probability density function of the two-parameter

Weibull distribution is given by

$$f(x; k, \lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{x}{\lambda}\right)^{k}\right\}, \qquad x \geq 0,$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. In this example, we have $\theta = (k, \lambda)$ and $\Theta = (0, \infty) \times (0, \infty)$. If a collection of wind data is available in the form of n independent observations $\mathbf{x} = (x_1, x_2, \dots, x_n)$, the likelihood function will be

$$\begin{aligned} L_n(k, \lambda \,|\, \mathbf{x}) &= \prod_{i=1}^{n} f(x_i; k, \lambda) \\ &= \left(\frac{k}{\lambda}\right)^{n} \prod_{i=1}^{n} \left(\frac{x_i}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{x_i}{\lambda}\right)^{k}\right\} \\ &= \frac{k^n}{\lambda^{nk}} \exp\left\{-\frac{1}{\lambda^k}\sum_{i=1}^{n} x_i^k + (k-1)\sum_{i=1}^{n}\log x_i\right\}. \end{aligned}$$

In this example, we may be interested in answering the following question: given the observed sample, what value of θ is the most plausible? Let $\hat{\theta}_n(\mathbf{x})$ be the global maximum of $L_n(\theta \,|\, \mathbf{x})$:

$$\hat{\theta}_n(\mathbf{x}) = \operatorname*{argmax}_{\theta} L_n(\theta \,|\mathbf{x}).$$

We call $\hat{\theta}_n(\mathbf{x})$ the maximum likelihood estimate of θ. As a function of the random sample $\mathbf{X}$, $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X})$ is called the maximum likelihood estimator (MLE) of θ. Intuitively, the MLE is a reasonable choice for a point estimator: the observed sample is the most likely when the MLE is the parameter value.

The MLE possesses two important properties by its construction. Firstly, the MLE is range respect; the range of the MLE coincides with the range of the parameter. Secondly, the MLE is invariant under parameter transformation. Suppose a distribution family is indexed by a parameter θ, but the interest lies in finding an estimator of some function of θ, say $\eta(\theta)$. If $\hat{\theta}_n$ is the MLE of θ, then $\eta(\hat{\theta}_n)$ is the MLE of $\eta(\theta)$. The second property of MLE allows us to study a parameter that does not appear in the density function. In the wind speed example, we may be interested in the mean of the wind speed, say μ. Note that μ can be expressed as a function of k and λ: $\mu = \lambda\,\Gamma(1+k^{-1})$ where $\Gamma(\cdot)$ is the gamma function. Hence, if $\hat{k}$ and $\hat{\lambda}$ are the MLEs of k and λ respectively, then $\hat{\mu} = \hat{\lambda}\,\Gamma(1+\hat{k}^{-1})$ is the MLE of μ.

MLE possesses many nice asymptotic properties under some mild conditions on $f(x;\theta)$. Firstly, MLE is a consistent estimator of the parameter, i.e. MLE converges to the true parameter value almost surely as the sample size increases. Secondly, MLE is asymptotically efficient in the sense that its asymptotic variance equals the Cramér-Rao bound as the sample size tends to infinity.

In applications, we may prefer a guess of a region of parameter values to a guess of a single parameter value. We can imagine that those parameter values that are slightly "different" from the MLE are also good candidates of the true parameter value. The likelihood function can be used to quantify the "difference" between any parameter value and the MLE. According to the definition of the MLE, the likelihood ratio $R_n(\theta\,|\,\mathbf{x}) = L_n(\theta\,|\,\mathbf{x})/L_n(\hat{\theta}_n\,|\,\mathbf{x})$ is always less than 1. Thus, we may choose some constant $c \in (0,1)$ and claim that the true parameter value is likely contained in the following region of parameter values:

$$\mathcal{CR}_c = \{\theta : R_n(\theta\,|\,\mathbf{x}) \geq c\}. \tag{2.1}$$

The purpose of using a region estimator rather than a point estimator is to have some guarantee of capturing the true value of parameter. The certainty of this guarantee is quantified by the probability of $\mathcal{CR}_c$ covering the true parameter value, $\Pr\{\theta \in \mathcal{CR}_c\} = \Pr\{R_n(\theta\,|\,\mathbf{x}) \geq c\}$. With this in mind, we may choose the constant c such that $\mathcal{CR}_c$ has a pre-specified coverage probability. The guaranteed coverage probability is also called the confidence level of $\mathcal{CR}_c$. Thus, we need to know the distribution of $R_n(\theta\,|\,\mathbf{X})$. In general, the exact distribution of $R_n(\theta\,|\,\mathbf{X})$ is hard to determine. Wilks (1938) proves that under some wild conditions $-2\log R_n(\theta\,|\,\mathbf{X})$ converges to χ^2_p in distribution as $n \to \infty$, provided that the true parameter value is

θ. This asymptotic result is known as Wilks' theorem.

Using this χ^2 approximation, we may choose c in equation (2.1) to be $\exp\{-\chi_p^2(\alpha)/2\}$, where $\chi_p^2(\alpha)$ denotes the upper α quantile of χ_p^2, for small α. The resulting approximate $100\,(1-\alpha)\%$ confidence region for θ is

$$\mathcal{CR} = \big\{\theta : R_n(\theta\,|\,\mathbf{x}) \geq \exp\{-\chi_p^2(\alpha)/2\}\big\} = \{\theta : -2\,\log R_n(\theta\,|\,\mathbf{x}) \leq \chi_p^2(\alpha)\}.$$

Similar to the MLE, the foregoing confidence region is also range respect and invariant under parameter transformation to some degree. In the example of the wind speed, if $\mathcal{CR}$ is an approximate 95% confidence region for the parameter (k,λ) then $\mathcal{CR}' = \{\lambda\,\Gamma(1+k^{-1}) : \ (k,\lambda) \in \mathcal{CR}\}$ is an at least approximate 95% confidence region for the mean μ.

As widely recognized, the statistical inferences based on parametric likelihood has its own risk: If the true distribution deviates from the parametric distribution that we assume for the data, the foregoing nice properties of these inferences on the parameter of interest may be deprived of.

The difficulties in choosing a parametric family make many statisticians turn to nonparametric methods for statistical inferences. These nonparametric methods include the jackknife, the infinitesimal jackknife, the bootstrap method, and the empirical likelihood method. Each nonparametric method has its own advantages, but most of them are lack of a likelihood setting. The empirical likelihood method stands out since it combines the reliability of the nonparametric methods and the flexibility and effectiveness of the likelihood approach.

2.2 Definition of Empirical Likelihood

In this section, we present the setting of empirical likelihood.

Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) d-dimensional random vectors with unknown distribution F_0 for some $d \geq 1$.

The empirical likelihood of any distribution F is defined as

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

where $F(A)$ is $\Pr(X \in A)$ for $X \sim F$ and $A \subseteq \mathcal{R}^d$.

The definition of empirical likelihood is a direct analogue of parametric likelihood: the probability of observing the sample under the assumed distribution. The major difference between empirical likelihood and parametric likelihood is that the former is defined over a very broad range of distributions. That is, there are practically no restrictions on the shape of the distribution under consideration. The name "empirical likelihood" is adopted because the empirical distribution of the sample plays a key role in the setting of empirical likelihood. The empirical distribution is defined as

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

where δ_x denotes the distribution under which $\Pr(X = x) = 1$. The empirical likelihood is maximized at the empirical distribution.

Proposition 2.1. *Suppose $X_1, X_2, \ldots, X_n \in \mathcal{R}^d$ for some $d \geq 1$ are independent random vectors with a common distribution F_0 and F_n is the corresponding empirical distribution. For any distribution $F \neq F_n$, we have $L(F) < L(F_n)$.*

Proof. Let $p_i = F(\{X_i\})$ for $i = 1, 2, \ldots, n$. It is easy to see that $p_i \geq 0$ and $\sum_{i=1}^{n} p_i \leq 1$. Using a well-known fact that the arithmetic mean of a sequence of nonnegative numbers is always larger than or equal to its geometric mean, we have

$$L_n(F) = \prod_{i=1}^{n} p_i \leq \left(\frac{1}{n} \sum_{i=1}^{n} p_i \right)^n \leq n^{-n}. \tag{2.2}$$

The last equality in (2.2) holds if and only if all p_i's are equal and $\sum_{i=1}^{n} p_i = 1$. This inequality implies that $L(F)$ attains its maximum n^{-n} at $F =$

F_n. $\square$

By analogy with the definition of MLE under parametric model, we say that the empirical distribution F_n is the maximum empirical likelihood estimate (MELE) of the distribution F. In this spirit, the MELE of the population mean $\mu = \int x \,\mathrm{d}F(x)$ is $\hat{\mu}_n = \int x \,\mathrm{d}F_n(x) = \sum_{i=1}^n X_i/n = \bar{X}_n$, which is the sample mean.

The properties of $\bar{X}_n$ under some mild conditions have already been well studied: $\bar{X}_n$ is an unbiased and consistent estimator of μ; it has the smallest asymptotic variance among all the unbiased estimators of μ; it is asymptotic normal distributed; and so on. In this thesis, we mainly focus on the problem of constructing confidence regions for μ through the empirical likelihood. For this purpose, we introduce the profile empirical likelihood in the next section.

2.3 Profile Empirical Likelihood of the Population Mean

Let $X_1, X_2, \ldots, X_n$ be i.i.d. d-dimensional random vectors with unknown distribution F.

By analogy with the Wilks' theorem, we may also use the ratio of the empirical likelihood as a basis for constructing confidence regions. The empirical likelihood ratio for a distribution F is defined as

$$R_n(F) = \frac{L_n(F)}{L_n(F_n)}.$$

By Proposition 2.1, $R_n(F) \le 1$ and the equality holds if and only if $F = F_n$. Recall that the population mean is a functional of the population distribution. The likeliness of a specific value of μ can be inferred from this relationship. In the literature of empirical likelihood, we define the profile empirical likelihood of μ as

$$L_n(\mu) = \sup\left\{L_n(F) : \int x \, \mathrm{d}F(x) = \mu\right\}. \tag{2.3}$$

By analogy with the parametric likelihood ratio, we may define the profile empirical likelihood ratio function of μ as

$$R_n(\mu) = \frac{L_n(\mu)}{L_n(\bar{X}_n)} = \sup\left\{R_n(F) : \int x \, \mathrm{d}F(x) = \mu\right\}, \tag{2.4}$$

Yet without requiring the support of F being confined within the set of observed values of X, this profile empirical likelihood ratio for the population mean always equals 1. We illustrate this point as follows. For any given μ, let ε be a positive constant smaller than 1 and

$$x_\mu = \frac{1}{\varepsilon}\,\mu - \frac{1-\varepsilon}{\varepsilon}\,\bar{X}_n.$$

We construct a mixture distribution $F_{\mu,\varepsilon} = (1-\varepsilon)\,F_n + \varepsilon\,\delta_{x_\mu}$. Note that the mean of $(F_{\mu,\varepsilon})$ is μ and

$$R_n(F_{\mu,\varepsilon}) = \frac{L_n(F_{\mu,\varepsilon})}{L_n(F_n)} = \frac{[(1-\varepsilon)/n]^n}{(1/n)^n} = (1-\varepsilon)^n.$$

Hence for any pre-specified value of μ, $R_n(F_{\mu,\varepsilon})$ can be made arbitrarily close to 1 as long as ε is sufficiently small. As a result, $R_n(\mu)$ defined by equation (2.4) always equals 1 for any μ. Hence, this definition of the profile likelihood ratio function is not useful for constructing confidence regions.

The above problem can be easily solved by requiring the support of F being contained in the set of observed values of X. As proposed by Owen (1988), the empirical likelihood ratio will be profiled for a parameter over only the distributions with support on the data set. In other words, only distributions such that $p_i = F(\{X_i\}) > 0$ and $\sum_{i=1}^n p_i = 1$ will be considered. We denote such a distribution F as $F \ll F_n$. The definition of the profile empirical likelihood for μ in the literature is given by

$$\begin{aligned} L_n(\mu) &= \sup\left\{ L_n(F) :\ F \ll F_n, \int x \, \mathrm{d}F(x) = \mu \right\} \\ &= \sup\left\{ \prod_{i=1}^n p_i :\ p_i > 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \, X_i = \mu \right\}. \end{aligned} \tag{2.5}$$

Without further clarification, we refer to "profile empirical likelihood" as "empirical likelihood" from now on.

In this definition of $L_n(\mu)$, the sample mean $\bar{X}_n$ is its maximum point. We naturally define the profile empirical likelihood ratio for μ as

$$\begin{aligned} R_n(\mu) &= \frac{L_n(\mu)}{L_n(\bar{X}_n)} \\ &= \sup\left\{ \frac{L_n(F)}{L_n(F_n)} :\ F \ll F_n, \int x \, \mathrm{d}F(x) = \mu \right\} \\ &= \sup\left\{ \prod_{i=1}^n np_i : p_i > 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \, X_i = \mu \right\}. \end{aligned} \tag{2.6}$$

For the convenience of discussing asymptotic properties, we prefer working on the profile empirical likelihood ratio statistic defined as

$$W_n(\mu) = -2 \, \log R_n(\mu).$$

Because $R_n(\mu)$ is the maximum value of $\prod_{i=1}^n np_i$ subject to some constraints, $W_n(\mu)$ is the minimum value of $-2\sum_{i=1}^n \log(np_i)$ subject to the same constraints. We will refer to a set of weights $\{p_i\}_{i=1}^n$ that satisfy these constraints as sub-optimal weights for $W_n(\mu)$.

The second constraint may also be written as

$$\sum_{i=1}^n p_i(X_i - \mu) = 0. \tag{2.7}$$

The calculation of $R_n(\mu)$ and $W_n(\mu)$ at a given parameter value amounts to solving a constrained optimization problem. The Lagrange's method is well suited in this situation. Take $W_n(\mu)$ as an example. Let us define

$$
\begin{aligned}
H(p_1, p_2, \ldots, p_n; \lambda, \eta) = {} & -2 \sum_{i=1}^{n} \log(np_i) - n\lambda^T \left[\sum_{i=1}^{n} p_i (X_i - \mu) \right] \\
& + \eta \left(\sum_{i=1}^{n} p_i - 1 \right)
\end{aligned}
$$

with $\lambda \in \mathcal{R}^d$ and $\eta \in \mathcal{R}$ being the lagrange multipliers.

Setting the derivatives of H with respect to λ and η to zero, we recover the two equality constraints on p_i's. Differentiating H with respect to p_i and setting the derivatives equal to zero, we get

$$
0 = \frac{\partial H}{\partial p_i} = \frac{1}{p_i} - n\lambda^T (X_i - \mu) + \eta \tag{2.8}
$$

Multiplying the above equation by p_i and summing over i, with the help of two constraints, we get

$$
0 = \sum_{i=1}^{n} p_i \frac{\partial H}{\partial p_i} = n + \eta.
$$

It gives us $\eta = -n$. Substituting this result into equation (2.8) gives the optimal weights

$$
p_i = \frac{1}{n} \frac{1}{1 + \lambda^T (X_i - \mu)}, \qquad i = 1, \ldots, n. \tag{2.9}
$$

The value of λ can be computed through the constraint

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

Equivalently, we have

$$
\sum_{i=1}^{n} \frac{X_i - \mu}{1 + \lambda^T (X_i - \mu)} = 0. \tag{2.10}
$$

The above equation can be easily solved numerically. From now on, we will refer to the weights given by equation (2.9) as the optimal weights for $W_n(\mu)$.

Once the value of λ is obtained, we can compute $W_n(\mu)$ through

$$W_n(\mu) = 2\sum_{i=1}^{n} \log[1 + \lambda^T(X_i - \mu)]. \tag{2.11}$$

The primal constrained optimization problem for $W_n(\mu)$ must work on n variables $p_1, p_2, \ldots, p_n$. Equation (2.11) shows that $W_n(\mu)$ has a simple analytic expression, and the constrained optimization problem is reduced to finding an appropriate root λ to equation (2.10). This simple expression of $W_n(\mu)$ has two advantages. Firstly, this expression of $W_n(\mu)$ provides a feasible approach to calculate $W_n(\mu)$ numerically. Chen et al. (2002) propose a modified Newton's algorithm for finding the root to equation (2.10), whose algorithmic convergence is guaranteed when the solution exists. Secondly, this expression helps us study the asymptotic behavior of $W_n(\mu)$. In the investigation of the asymptotic properties of $W_n(\mu)$, the property of λ plays a key role.

2.4 Profile Empirical Likelihood for Parameters Defined Through General Estimating Equations

In additional to make inference on population means, empirical likelihood finds many applications to parameters defined in a nonparametric way. For instance, Owen (1991) applies empirical likelihood to make inference on the regression coefficients in linear models. In general, we can often define some parameters of interest through the so-called "general estimating equations". For a random variable $X \sim F$, a p-dimensional parameter θ can be defined as the solution to

$$\mathbf{E}_F[g(X;\theta)] = 0 \tag{2.12}$$

for some q-dimensional mapping $g(X;\theta)$ with $q \geq p$. The above system is known as the general estimating equation (GEE), and $g(x;\theta)$ is called the estimating function. When $g(x;\theta) = x - \theta$, the parameter θ is the mean of X. When $g(s;\theta) = \mathrm{I}(x \leq \theta) - \alpha$ for some $\alpha \in (0,1)$, θ is the α quantile of X.

The classic setting of GEE has $q = p$. Given a simple random sample $X_1, X_2, \ldots, X_n$, an estimator of θ, say $\hat{\theta}$, can be obtained as the solution to

$$\mathbf{E}_{F_n}[g(X;\theta)] = \frac{1}{n}\sum_{i=1}^{n} g(X_i;\theta) = 0. \tag{2.13}$$

Since F_n is the MELE of F, it implies that this estimator is the MELE of θ.

In econometrics applications, however, most interest attaches to the case of over-identification with $q > p$ (Imbens, 2002; Hansen, 1982; Hall, 2005). In this case, equation (2.13) may not have any solutions.

Empirical likelihood provides a natural approach to overcome this problem. Qin and Lawless (1994) develop a theory for the EL-based statistical inference for parameters defined through general estimating functions. They propose to define the profile empirical likelihood for θ as

$$L_n(\theta) = \max\left\{\prod_{i=1}^{n} p_i : \; p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \, g(X_i;\theta) = 0\right\}.$$

The corresponding profile empirical likelihood ratio for θ becomes

$$R_n(\theta) = \frac{L(\theta)}{L(F_n)} = \max\left\{\prod_{i=1}^{n} np_i : \; p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \, g(X_i;\theta) = 0,\right\}. \tag{2.14}$$

The likelihood ratio statistic is then given by

$$W_n(\theta) = -2 \log R_n(\theta).$$

Similar to the case of the population mean discussed in Section 2.3, $W_n(\theta)$

can be written as

$$W_n(\theta) = 2\sum_{i=1}^{n} \log[1 + \lambda^T g(X_i;\theta)] \tag{2.15}$$

with λ being the solution to

$$\sum_{i=1}^{n} \frac{g(X_i;\theta)}{1 + \lambda^T g(X_i;\theta)} = 0. \tag{2.16}$$

In the framework of general estimating equation, the MELE of θ, which is defined as the maximum point of $L_n(\theta)$, is not so trivial as that for population mean. One of the main contributions of Qin and Lawless (1994) is that they demonstrate the asymptotic normality of the MELE of θ under some regularity conditions on the estimating function. In addition, they justify the use of the empirical likelihood ratio statistic for testing or obtaining confidence regions for parameters in a completely analogous way to the parametric likelihood approach. But the main interest of this thesis lies in the statistical inference for population mean, so we will not explore this topic further here.

2.5 Asymptotic Properties of Empirical Likelihood and EL-Based Confidence Regions

The most impressive result in Owen (1988, 1990) is the following asymptotic limiting distribution of $W_n(\mu)$.

Theorem 2.2. *Let $X_1, X_2, \ldots, X_n$ be a simple random sample from some d-dimensional population X and $W_n(\mu)$ is the empirical likelihood ratio statistic for the population mean μ. If the variance-covariance matrix of X is positive definite and the true value of μ is μ_0, then*

$$W_n(\mu_0) \xrightarrow{d} \chi^2_d, \quad \text{as } n \to \infty. \tag{2.17}$$

Theorem 2.2 suggests an approximate $100(1-\alpha)\%$ confidence region for

μ in the form of

$$\mathcal{CR}_\alpha = \{\mu \,|\, W_n(\mu) \le \chi^2_d(\alpha)\},$$

where $\chi^2_d(\alpha)$ is the upper α quantile of χ^2_d. Hall and La Scala (1990) and Owen (2001) point out that the confidence region $\mathcal{CR}_\alpha$ is always convex. This is clearly a nice property to practitioners. We summarize their result as Proposition 2.3 with a simple proof.

Proposition 2.3. *Let $X_1, X_2, \ldots, X_n$ be a random sample from some population X and $W_n(\mu)$ be the empirical likelihood ratio statistic for the population mean. Suppose $\mu_1 \ne \mu_2$ and $\mu_1, \mu_2 \in \mathcal{CR}_\alpha$, and μ is a convex combination of μ_1 and μ_2. Then $\mu \in \mathcal{CR}_\alpha$.*

Proof. Let $\{p_i\}_{i=1}^n$ and $\{q_i\}_{i=1}^n$ be the optimal weights for $W_n(\mu_1)$ and $W_n(\mu_2)$, respectively. For any μ such that $\mu = \xi\,\mu_1 + (1-\xi)\,\mu_2$ for some $0 \le \xi \le 1$, it is easy to verify that $\{r_i = \xi\,p_i + (1-\xi)\,q_i\}_{i=1}^n$ are sub-optimal weights for $L_n(\mu)$. Note also that $\xi\,p_i + (1-\xi)\,q_i \ge p_i^\xi\,q_i^{1-\xi}$ for $i = 1, 2, \ldots, n$. Hence, we have

$$L_n(\mu) \ge \prod_{i=1}^n r_i = \prod_{i=1}^n [\xi\,p_i + (1-\xi)\,q_i] \ge \prod_{i=1}^n p_i^\xi\,q_i^{1-\xi} = [L_n(\mu_1)]^\xi\,[L_n(\mu_2)]^{1-\xi}.$$

It follows that

$$W_n(\mu) \le \xi\,W_n(\mu_1) + (1-\xi)\,W_n(\mu_2) \le \xi\,\chi^2_d(\alpha) + (1-\xi)\,\chi^2_d(\alpha) = \chi^2_d(\alpha).$$

By definition of $\mathcal{CR}_\alpha$, we conclude that $\mu \in \mathcal{CR}_\alpha$. This completes the proof. $\square$

Following this proposition, we can see that $W_n(\mu)$ has some kind of monotonicity property.

Corollary 2.4. *Assume the same conditions as in Proposition 2.3. Let $\mathbf{v}$ be a d-dimensional unit vector and consider the half line defined by $\bar{X}_n + t\,\mathbf{v}$ for $t > 0$. Then $W_n(\bar{X}_n + t\,\mathbf{v})$ is an increasing function of t.*

Proof. For any $0 < t_1 < t_2$, let $\mu_1 = \bar{X}_n + t_1\,\mathbf{v}$ and $\mu_2 = \bar{X}_n + t_2\,\mathbf{v}$. Note that μ_1 is a convex combination of μ_2 and $\bar{X}_n$.

Consider a confidence region for μ, $\mathcal{CR} = \{\mu : W_n(\mu) \le W_n(\mu_2)\}$. Note that μ_2 and $\bar{X}_n$ always fall inside of this region. By Proposition 2.3, μ_1 also belongs to $\mathcal{CR}$ and thus $W_n(\mu_1) \le W_n(\mu_2)$. □

Corollary 2.4 justifies a simple algorithm for numerically finding the boundary of the confidence region. We briefly describe this algorithm in the case of bivariate mean:

1. Choose a sufficiently dense sequence of angles from 0 to $2\,\pi$, for example, an arithmetic sequence from 0 to $2\,\pi$ with common difference $2\,\pi/M$ for some sufficiently large positive integer M.

2. Along each direction defined as a unit vector $\Phi_m = (\cos\phi_m, \sin\phi_m)^T$ with ϕ_m being an angle selected in Step 1, we search for a positive real number t_m such that $W_n(\bar{X}_n + t_m\,\Phi_m)$ is sufficiently close to the critical value determined by the χ^2 approximation. Since $W_n(\mu)$ is increasing along any direction starting from $\bar{X}_n$, as asserted in Corollary 2.4, a simple bisection algorithm is effective.

3. Let $\{\bar{X}_n + t_m\,\Phi_m, m = 1, 2, \ldots, M\}$ be the boundary points obtained in Step 2.

With this set of points, we not only can visualize the confidence regions through a two-dimensional plot, but can also calculate the approximate area of the confidence regions. Apparently, the approximation becomes better when M gets larger, but the exact accuracy is difficult to determine.

EL-based confidence region has many celebrating properties. We summarize some as follows:

1. EL-based confidence region has a data-driven shape. Figure 2.1 shows the boundary of the EL-based 95% confidence region for the bivariate mean based on a data set of 10 observations generated from a bivariate gamma distribution. It is seen that the shape of the confi-

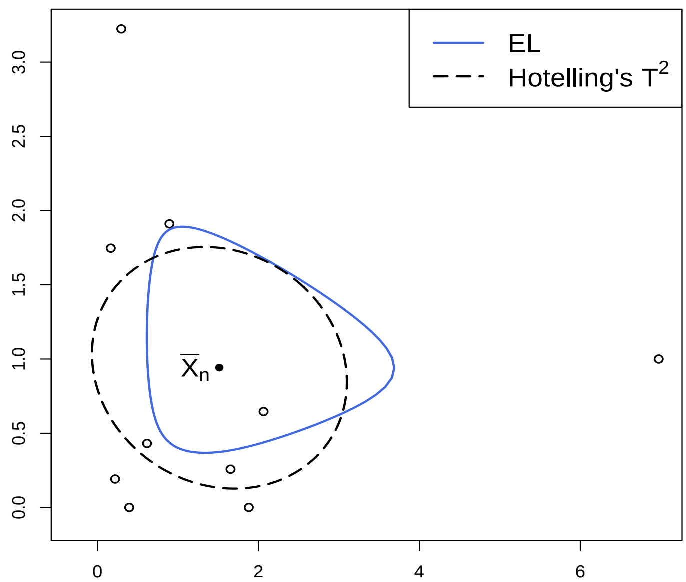


Figure 2.1: A two-dimensional example of EL-based confidence region

dence region based on the widely-used normal approximation is predetermined even before the data are available. On the contrary, the EL-based confidence region automatically reflects the emphasis on the data set. It is an appealing property to many practitioners since it upholds the principle of “letting the data speak.”

2. EL-based confidence region is range respecting and transformation invariant. For example, the confidence interval for the correlation always lies between −1 and 1.

3. The EL-based confidence region is Bartlett-correctable. In both parametric likelihood based and EL-based confidence regions, we select the critical value using the limiting distribution of the likelihood ratio statistic. Such approximations introduce error to the coverage accuracy of the resulting confidence regions. The actual coverage proba-

bility of the confidence region does not exactly agree with the nominal level. In the parametric setting, the coverage accuracy can be improved by the so-called Bartlett correction on the likelihood ratio statistic (Barndorff-Nielsen and Cox, 1984). As shown by Diciccio et al. (1991), the empirical likelihood ratio statistic is also Bartlett correctable. We will discuss it further in Section 2.6.1.

2.6 Limitations of Empirical Likelihood

While empirical likelihood has many nice properties as shown in Section 2.5, there are situations where its applications meet some practical obstacles. In this section, we discuss two related issues which lead to the adjusted empirical likelihood in Chapter 3.

2.6.1 Under-Coverage Problem

Theorem 2.2 suggests using the limiting distribution of $W_n(\mu)$ to calibrate the EL-based confidence region for μ. As expected, the coverage probability of the resulting confidence region does not exactly match the pre-specified confidence level. Diciccio et al. (1991) shows that

$$\Pr\{\mu_0 \in \mathcal{CR}_\alpha\} = \Pr\{W_n(\mu_0) \leq \chi^2_d(\alpha)\} = 1 - \alpha + O(n^{-1}). \tag{2.18}$$

Simulation results reveal that EL-based confidence regions suffer from the so-called "under-coverage" problem. That is, its coverage probability is lower than the nominal level particularly when the sample size is small or the population is skewed. Diciccio et al. (1991) prove that the empirical likelihood is Bartlett correctable; a simple correction on $W_n(\mu_0)$ can improve the approximating precision given in equation (2.18) from $O(n^{-1})$ to $O(n^{-2})$. Empirical studies reveal that the Bartlett correction significantly improves the coverage rate of the EL-based confidence regions.

The error in the approximation is partially accounted to the fact that the expectation of $W_n(\mu_0)$ does not match the expectation of the corresponding limiting distribution. Thus, the coverage accuracy may be improved by rescaling $W_n(\mu)$. Asymptotically, it is found that

$$\mathbf{E}[W_n(\mu)] = d\left[1 + \frac{b}{n} + O(n^{-2})\right]$$

with b being a constant depending on the first four moments of the population X. By applying the χ^2 approximation to $W_n(\mu)/(1 + b\,n^{-1})$, the resulting confidence region has higher precision; the coverage error is reduced from order n^{-1} to of order n^{-2}. More precisely,

$$\begin{aligned}\Pr\left\{\frac{W_n(\mu_0)}{1+b/n} \le \chi_d^2(\alpha)\right\} &= \Pr\left\{W_n(\mu_0) \le \chi_d^2(\alpha)\left(1 + \frac{b}{n}\right)\right\} \\ &= 1 - \alpha + O(n^{-2}). \end{aligned} \tag{2.19}$$

The asymptotic derivation of equation (2.19) is long and complex. The details can be found in Diciccio et al. (1991).

The constant b in equation (2.19) is called the Bartlett correction factor. The value of b depends on the first four moments of the population distribution. Its value must be estimated based on data in applications. Replacing b by a $\sqrt{n}$-consistent estimator in equation (2.19) will not affect the theoretical result.

The Bartlett correction is also applicable for EL-based confidence regions for parameters defined through general estimating equation (2.12). The Bartlett correction factor b is determined by the distribution of $g(X;\theta)$. Liu and Chen (2010) provide a detailed discussion on how to calculate the Bartlett correction factor in the framework of general estimating equation.

2.6.2 The No-Solution Problem

As described in Section 2.4, $W_n(\theta)$ equals the minimum value of $-2\sum_{i=1}^{n}\log(n\,p_i)$ over all sub-optimal weights $\{p_i\}_{i=1}^{n}$. Hence $W_n(\theta)$ is well defined if and only if there exists at least one set of sub-optimal weights. Let $\mathcal{CH}\{\cdots\}$ be the convex hull expanded by the set of points inside $\{\}$. Then, $W_n(\theta)$ is well defined when

$$0 \in \mathcal{CH}\{g(X_i;\theta), i = 1, 2, \ldots, n\}. \tag{2.20}$$

We take the population mean μ as an example. Condition (2.20) is satisfied if and only if $0 \in \mathcal{CH}\{X_i - \mu, i = 1, 2, \ldots, n\}$. Equivalently, we must have $\mu \in \mathcal{CH}\{X_i, i = 1, 2, \ldots, n\}$. When μ is one-dimensional, condition (2.20) can be further simplified to $X_{(1)} < \mu < X_{(n)}$. Let Θ_0 be the set of θ over which condition (2.20) is satisfied. We can see that Θ_0 is determined by the data. For complex estimating equations, it can be hard to specify the structure of Θ_0.

Owen (2001) proves that the true parameter value θ_0, defined as the unique solution to equation (2.12), is contained in Θ_0 almost surely as $n \to \infty$ under some regularity conditions on the estimating function. For θ is not close to θ_0 or when the sample size is small, it is very possible that $\theta \notin \Theta_0$ and thus equation (2.16) is not solvable. When $\theta \notin \Theta_0$, it is conventional to define $W_n(\theta) = \infty$. However, this setting has its own limitations. Firstly, for any two different parameter values $\theta_1, \theta_2 \notin \Theta_0$, we are unable to evaluate their relative plausibility based on $W_n(\theta)$. Secondly, using this setting implies that the confidence region is always a subset of Θ_0, which is determined by the data. This can be a problem when even Θ_0 itself does not achieve the desired confidence level especially when the sample size is small.

Aiming to solve the no-solution problem of empirical likelihood, Chen et al. (2008) propose an adjustment to the original empirical likelihood such that the resulting adjusted empirical likelihood (AEL) is well defined for all possible parameter values. Chapter 3 is contributed to summarizing the well-studied asymptotic properties of the adjusted empirical likelihood, and investigating the finite-sample properties of AEL-based confidence regions mainly in the case of population mean.

Chapter 3

Adjusted Empirical Likelihood

To overcome the obstacle caused by the no-solution problem in the application of empirical likelihood, Chen et al. (2008) propose an adjustment to empirical likelihood. The resulting adjusted empirical likelihood is attracting for its easy computation and desirable asymptotic properties. Recently, this method has found its applications in various areas. Zhu et al. (2009) incorporate the adjusted empirical likelihood and the exponentially tilted likelihood, and apply it to the analysis of morphometric measures in MRI studies. Liu and Yu (2010) propose a two-sample adjusted empirical likelihood approach to construct confidence regions for the difference of two population means. Variyath et al. (2010) introduce the information criteria under adjusted empirical likelihood to variable/model selection problems.

In this chapter, we first review the setting of the resulting adjusted empirical likelihood (AEL) and its asymptotic property. In addition, we present some new results on the finite-sample properties of the adjusted empirical likelihood.

3.1 Adjusted Empirical Likelihood and AEL-Based Confidence Regions

Let us start with a simple example. Suppose that we have a random sample of n bivariate observations, and we are interested in the population mean μ. Now consider a value of μ outside of $\mathcal{CH}\{X_i, i = 1, 2, \ldots, n\}$. Apparently, such a value of μ does not satisfy condition (2.20), and therefore $W_n(\mu)$

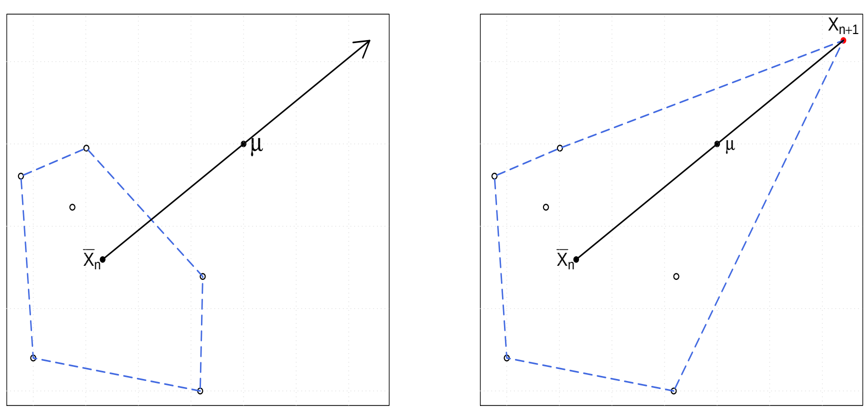
is not well defined. The idea of the adjustment proposed by Chen et al. (2008) is to add a pseudo observation X_{n+1} into the data set such that $\mu \in \mathcal{CH}\{X_i, i = 1, 2, \ldots, n+1\}$. More specifically, we may choose X_{n+1} as

$$X_{n+1} = \mu + a_n\,(\mu - \bar{X}_n), \tag{3.1}$$

for some positive constant a_n; or equivalently, we may write

$$X_{n+1} - \mu = -a_n\,(\bar{X}_n - \mu). \tag{3.2}$$

The rationale of adding such a pseudo point is illustrated in Figure 3.1. Note that $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ is always an interior point of $\mathcal{CH}\{X_i, i = 1, 2, \ldots, n\}$. Suppose μ be a parameter value outside of the convex hull of $\{X_1, X_2, \ldots, X_n\}$. Let us first draw a ray from $\bar{X}_n$ towards μ, and let X_{n+1} be a point on the further side of μ. Apparently, the constant a_n determines how far X_{n+1} should be placed. It is seen that μ is an interior point of the convex hull of $\{X_1, X_2, \ldots, X_{n+1}\}$.



(a) Plot of X_i's

(b) After adding a pseudo point

Figure 3.1: A two-dimensional example showing the position of the pseudo point.

This adjustment is generally applicable. For any given value of θ, let $g_i(\theta) = g(X_i;\theta)$ and $\bar{g}_n(\theta) = \sum_{i=1}^{n} g_i(\theta)/n$. And the pseudo observation is defined as

$$g_{n+1}(\theta) = -a_n\, \bar{g}_n(\theta). \tag{3.3}$$

By including this pseudo observation into the data set, the empirical likelihood ratio statistic for θ becomes

$$W_n^*(\theta; a_n) = -2\, \log R_n^*(\theta; a_n) \tag{3.4}$$

with

$$R_n^*(\theta) = (n+1)^{n+1}\, L_n^*(\theta; a_n),$$

and

$$L_n^*(\theta; a_n) = \max\left\{ \prod_{i=1}^{n+1} p_i : p_i > 0, \sum_{i=1}^{n+1} p_i = 1, \sum_{i=1}^{n+1} p_i\, g_i(\theta) = 0 \right\}.$$

We call a set of weights of $\{p_i\}_{i=1}^{n+1}$ sub-optimal for $W_n^*(\theta; a_n)$, $R_n^*(\theta; a_n)$ or $L_n^*(\theta; a_n)$ if they satisfy the above equality constraints.

Using the Lagrange's method, we can easily show that the optimal weights are given by

$$p_i = \frac{1}{n+1}\, \frac{1}{1 + \lambda^T g_i(\theta)}, \qquad i = 1, 2, \ldots, n+1,$$

where λ is the solution to

$$\sum_{i=1}^{n+1} \frac{g_i(\theta)}{1 + \lambda^T g_i(\theta)} = 0.$$

As a consequence, $W_n^*(\theta; a_n)$ can be expressed as

$$W_n^*(\theta; a_n) = 2 \sum_{i=1}^{n+1} \log[1 + \lambda^T g_i(\theta)].$$

Compared to the original empirical likelihood, adjusted empirical likelihood has many desirable properties. Firstly, adjusted empirical likelihood yields a sensible value of likelihood at any putative parameter value, and this allows us to evaluate the plausibility of any parameter value. On the contrary, the original empirical likelihood is well defined only over a data-dependent subset of the parameter space, and this subset is difficult to specify numerically when the estimating function $g(x;\theta)$ is complex.

Secondly, the first order asymptotic property of $W_n(\theta_0)$, where θ_0 is the true value of θ, is largely preserved for $W_n^*(\theta_0; a_n)$. For example, $W_n^*(\theta_0; a_n)$ has the same limiting distribution as that of $W_n(\theta_0)$ as long as $a_n = o_p(n^{2/3})$. Thus, the χ^2 calibration is still applicable to constructing confidence regions for parameter of interest. That is, $\mathcal{CR}_\alpha^* = \{\theta : W_n^*(\theta; a_n) \leq \chi_q^2(\alpha)\}$ remains an approximate $100(1-\alpha)\%$ confidence region for θ.

Thirdly, AEL-based confidence regions can achieve coverage precision of higher order with appropriately chosen a_n. The positive constant a_n in the definition of the pseudo point can be used as a tuning parameter which controls the level of adjustment. Recall that EL-based confidence regions have the under-coverage problem, and they are Bartlett correctable to achieve higher order precision. Apparently, tuning the size of a_n may achieve the same good. This is exactly what has been proposed in Liu and Chen (2010). They discover that when $a_n = b/2$ with b being the Bartlett correction factor, the coverage accuracy of AEL-based confidence regions is of order n^{-2}, which is the same as that of Bartlett-corrected EL-based confidence regions. The sign of b matters in the adjusted empirical likelihood. In the one-dimensional case, b is positive for any distribution. When the dimension is higher than 1, empirical studies (Liu and Chen, 2010) seem to support that b is positive, but theoretical justification is still needed.

Fourthly, although the original motivation of the adjusted empirical likelihood method is to handle the no-solution problem confronted with the EL method, empirical studies (Chen et al., 2008; Liu and Chen, 2010) reveal that AEL-based confidence regions have higher coverage rate than EL-based confidence regions. Note that this does not imply that the AEL-based confidence regions have more accurate coverage rates than the corresponding EL-based confidence regions, though it is often the case because EL-based confidence regions have the under-coverage problem. We will demonstrate this empirical discovery rigorously in Section 3.2.

3.2 Finite-Sample Properties of Adjusted Empirical Likelihood for Population Mean

We devote this section to the finite-sample properties of the adjusted empirical likelihood mainly in the case of population mean.

3.2.1 Monotonicity of $W_n^*(\mu; a_n)$ in μ

It is desirable that the confidence region of any parameter is convex. Empirical evidences seem to support that AEL-based confidence region for population mean is convex. This is yet to be confirmed theoretically. In this section, we prove that like $W_n(\mu)$, the adjusted likelihood ratio statistic $W_n^*(\mu; a_n)$ also has a monotonicity property. As mentioned earlier in Section 2.5, this property is critical for the numerical computation of multidimensional confidence regions.

Theorem 3.1. *Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from some population X, and $W_n^*(\mu; a_n)$ is the adjusted empirical likelihood ratio statistic for the population mean. For any d-dimensional unit vector $\mathbf{v}$, consider the half line $\bar{X}_n + t\,\mathbf{v}$ for $t \geq 0$. Then $W_n^*(\bar{X}_n + t\,\mathbf{v}; a_n)$ is an increasing function of t.*

Proof. For any $0 < t_1 < t_2$, let $\mu_1 = \bar{X}_n + t_1\,\mathbf{v}$ and $\mu_2 = \bar{X}_n + t_2\,\mathbf{v}$. Note that

$$W_n^*(\mu; a_n) = -2 \log R_n^*(\mu; a_n) = -2 \log L_n^*(\mu; a_n) - 2 \log(n+1)^{n+1},$$

and the logarithm transformation is monotone. It suffices to show $L_n^*(\mu_1; a_n) \geq L_n^*(\mu_2; a_n)$.

Let $\{p_i\}_{i=1}^{n+1}$ be the optimal weights for $L_n^*(\mu_2; a_n)$. If we can find a set of sub-optimal weights $\{q_i\}_{i=1}^{n+1}$ for $L_n^*(\mu_1; a_n)$ such that $\prod_{i=1}^{n+1} q_i \geq \prod_{i=1}^{n+1} p_i$, then the conclusion of the theorem follows since $L_n^*(\mu_1; a_n) \geq \prod_{i=1}^{n+1} q_i$.

For $i = 1, 2, \ldots, n$, we define

$$r_i = \frac{p_i}{1 - p_{n+1}}.$$

Then we have

$$\sum_{i=1}^{n} r_i\,(X_i - \mu_2) + \frac{p_{n+1}}{1 - p_{n+1}}\,(X_{n+1} - \mu_2) = 0.$$

Substituting $X_{n+1} - \mu = -a_n\,(\bar{X}_n - \mu)$ and letting $k = p_{n+1}\,a_n/(1 - p_{n+1})$, we have

$$\sum_{i=1}^{n} r_i\,(X_i - \mu_2) = k\,(\bar{X}_n - \mu_2) = (\bar{X}_n - \mu_2) + (k-1)\,(\bar{X}_n - \mu_2).$$

Define

$$\widetilde{L}_n(\phi) = \max\left\{\prod_{i=1}^{n} s_i : \ s_i > 0, \sum_{i=1}^{n} s_i = 1, \sum_{i=1}^{n} s_i\,(X_i - \mu_2) = \phi\right\}.$$

That is, $\widetilde{L}_n(\phi)$ is the profile empirical likelihood for $\phi = \mathbf{E}[X_1 - \mu_2]$. It is easy to verify that $\widetilde{L}_n(\phi)$ is maximized at $\bar{\phi}_n = \sum_{i=1}^{n}(X_i - \mu_2)/n = \bar{X}_n - \mu_2$, and $\{r_i\}_{i=1}^{n}$ is the optimizing weights for $\widetilde{L}_n(\phi_2)$ with $\phi_2 = (\bar{X}_n - \mu_2) + (k-1)\,(\bar{X}_n - \mu_2)$. Denote $\phi_1 = (\bar{X}_n - \mu_2) + (k-1)\,(\bar{X}_n - \mu_1)$. Note that μ_1 is a convex combination of μ_2 and $\bar{X}_n$. Hence ϕ_1 is also a convex combination of ϕ_2 and $\bar{\phi}_n$. By the monotonicity property of $\widetilde{L}_n(\phi)$ (Corollary 2.4), we have $\widetilde{L}_n(\phi_1) \geq \widetilde{L}_n(\phi_2)$.

Let $\{s_i\}_{i=1}^n$ be the optimal weights for $\widetilde{L}_n(\phi_1)$, and define $q_{n+1} = p_{n+1}$ and $q_i = (1 - p_{n+1})\, s_i$ for $i = 1, 2, \ldots, n$. We can easily verify that $\{q_i\}_{i=1}^{n+1}$ is a set of sub-optimal weights for $L_n^*(\mu_1)$. Hence,

$$
\begin{aligned}
L_n^*(\mu_1; a_n) &\geq \prod_{i=1}^{n+1} q_i \\
&= p_{n+1} \cdot (1 - p_{n+1})^n \prod_{i=1}^{n} s_i \\
&\geq p_{n+1} \cdot (1 - p_{n+1})^n \prod_{i=1}^{n} r_i \\
&= p_{n+1} \cdot (1 - p_{n+1})^n \prod_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} \\
&= \prod_{i=1}^{n+1} p_i \\
&= L_n^*(\mu_2; a_n).
\end{aligned}
$$

Hence, $W_n^*(\mu_1; a_n) \leq W_n^*(\mu_2; a_n)$. □

According to Theorem 3.1, we find that $\bar{X}_n$ is the minimum point of $W_n^*(\mu; a_n)$, and AEL-based confidence regions for the population mean are at least star-shaped with $\bar{X}_n$ being the center. In the case of the univariate mean, the AEL-based confidence regions are still intervals. This result guarantees that the bisection algorithm described in Section 2.5 also works in finding the boundary of the AEL-based confidence region.

Whether AEL-based confidence regions for population mean are convex or not is still not clear, though it seems to be the case in the simulation studies.

3.2.2 Monotonicity of $W_n^*(\theta; a_n)$ in a_n

Empirical studies in Chen et al. (2008) and Liu and Chen (2010) reveal that the AEL-based confidence regions have higher coverage rate than the corresponding EL-based confidence regions. Intuitively, the gain in coverage rate of AEL-based confidence regions may be explained by the way how the adjusted empirical likelihood ratio statistic is constructed. As argued by Hua (2009), for testing the null hypothesis $H_0 : \theta = \theta_0$, the pseudo point $g_{n+1}(\theta_0)$ is always placed at a position that is in favor of the null hypothesis. Thus, the adjusted empirical likelihood ratio statistic tends to favor the null hypothesis and deflates the type-I error. Consequently, the AEL-based confidence regions has higher coverage rate compared to the corresponding EL-based confidence regions. It turns out that this observation can be proved rigorously. Theorem 3.2 reveals the monotonicity of $W_n^*(\theta; a_n)$ in a_n, and an interesting relationship between adjusted empirical likelihood and empirical likelihood as a special case. It implies that the AEL-based confidence region strictly contains the corresponding EL-based confidence region.

Theorem 3.2. *Suppose $X_1, X_2, \ldots, X_n$ is a random sample from some population X, and $W_n(\theta)$ and $W_n^*(\theta; a_n)$ are the empirical likelihood ratio statistic and the adjusted empirical likelihood ratio statistic defined by equations* (2.15) *and* (3.4), *respectively. We adopt the conventional value ∞ for $W_n(\theta)$ when it is not well defined. Then we have*

(1) $W_n^*(\theta; a_n) = W_n(\theta)$ *if* $a_n = 0$.

(2) $W_n^*(\theta; a_n)$ *is a decreasing function of* a_n *on the closed interval* $[0, n]$.

Proof. (1) When $a_n = 0$, we have $g_{n+1}(\theta) = 0$. Hence, $W_n^*(\theta; 0)$ becomes

$$\begin{aligned} W_n^*(\theta; 0) &= 2 \sum_{i=1}^{n+1} \log[1 + \lambda^T g_i(\theta)] \\ &= 2 \sum_{i=1}^{n} \log[1 + \lambda^T g_i(\theta)], \end{aligned}$$

where λ is the solution to

$$\sum_{i=1}^{n+1} \frac{g_i(\theta)}{1 + \lambda^T g_i(\theta)} = \sum_{i=1}^{n} \frac{g_i(\theta)}{1 + \lambda^T g_i(\theta)} = 0.$$

It is clear that the expression of $W_n^*(\theta;0)$ and the equation to λ coincide with those in the definition of $W_n(\theta)$ (equation (2.15) and (2.16)). That is, we have $W_n^*(\theta;0)=W_n(\theta)$.

(2) We will only give the proof in the case of population mean μ for simplicity; the proof is the same for the case of general estimating equation. Without loss of generality, we also fix $\mu=0$ and assume $\bar{X}_n\neq 0$.

When $a_n=n$, it is easy to verify that weights $\{p_i=1/(n+1)\}_{i=1}^{n+1}$ are sub-optimal for $W_n^*(0;n)$ and thus are optimal for $W_n^*(0;n)$. Therefore, $W_n^*(0;n)=0\leq W_n^*(0;a_n)$ for any $a_n<n$. Next we will only consider $a_n\in[0,n)$.

Note that $W_n^*(0;a_n)$ can be expressed as

$$W_n^*(0;a_n)=2\sum_{i=1}^{n+1}\log(1+\lambda^T X_i),$$

where $X_{n+1}=-a_n\bar{X}_n$, and λ satisfies

$$\sum_{i=1}^{n+1}\frac{X_i}{1+\lambda^T X_i}=0. \tag{3.5}$$

The derivative of $W_n^*(0;a_n)$ with respect to a_n is

$$\begin{aligned}\frac{\mathrm{d}W^*(0;a_n)}{\mathrm{d}a_n}&=\sum_{i=1}^{n+1}\frac{\left(\frac{\mathrm{d}\lambda}{\mathrm{d}a_n}\right)^T X_i}{1+\lambda^T X_i}+\frac{\lambda^T\frac{\mathrm{d}X_{n+1}}{\mathrm{d}a_n}}{1+\lambda^T X_{n+1}}\\&=\left(\frac{\mathrm{d}\lambda}{\mathrm{d}a_n}\right)^T\sum_{i=1}^{n+1}\frac{X_i}{1+\lambda^T X_i}+\frac{\lambda^T(-\bar{X}_n)}{1+\lambda^T X_{n+1}}\\&=-\frac{\lambda^T\bar{X}_n}{1+\lambda^T X_{n+1}},\end{aligned}$$

where we substitute equation (3.5).

If the derivative of $W_n^*(0;a_n)$ is always negative, then we know $W_n^*(0;a_n)$ is a decreasing function of a_n. Our task is to prove that the derivative of $W_n^*(0;a_n)$ is negative for $a_n\in[0,n)$, or equivalently to prove $\lambda^T\bar{X}_n>0$.

Consider the following function

$$f(t) = \sum_{i=1}^{n+1} \frac{\lambda^T X_i}{1 + t \cdot \lambda^T X_i}.$$

Note that

$$f(0) = \sum_{i=1}^{n+1} \lambda^T X_i = \lambda^T \sum_{i=1}^{n} X_i + \lambda^T(-a_n \bar{X}_n) = (n - a_n)\, \lambda^T \bar{X}_n,$$

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

We also notice that the derivative of $f(t)$

$$\frac{\mathrm{d} f(t)}{\mathrm{d} t} = \sum_{i=1}^{n+1} \left[-\frac{\lambda^T X_i \cdot \lambda^T X_i}{(1 + t \cdot \lambda^T X_i)^2} \right] = -\sum_{i=1}^{n+1} \left(\frac{\lambda^T X_i}{1 + t \cdot \lambda^T X_i} \right)^2$$

is always negative, and thus $f(t)$ is a decreasing function of t. Therefore, we have $f(0) > f(1)$, that is $(n - a_n)\, \lambda^T \bar{X}_n > 0$. Since $a_n < n$, we find $\lambda^T \bar{X}_n > 0$. Consequently, $W_n^*(0; a_n)$ is a decreasing function of a_n, and it completes the proof. □

Figure 3.2 plots the EL and AEL likelihood ratio statistics based on an artificially generated data set. It clearly shows that $W_n^*(\mu; a_n) \leq W_n(\mu)$ for all μ. As a consequence, the AEL-based confidence interval for μ contains the corresponding EL-based confidence interval and hence the former has higher coverage probability. This conclusion is generally true for parameters defined through general estimating equation. It enhances the results in Liu and Chen (2010) that AEL-based confidence regions where $a_n = b/2$ with b being the Bartlett correction factor has not only higher coverage accuracy but also higher coverage probability.

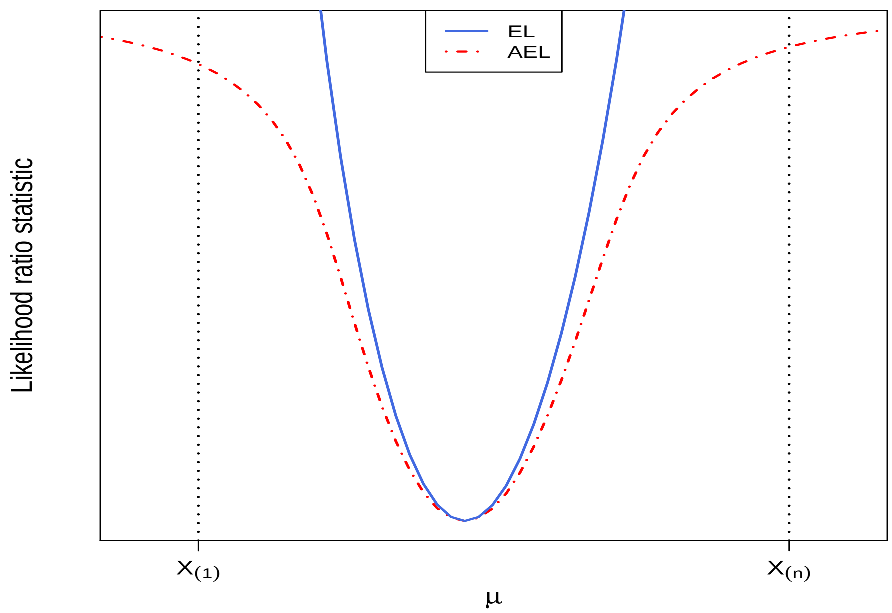


Figure 3.2: An example showing the effect of the pseudo point on the likelihood ratio statistic.

3.2.3 Boundedness of $W_n^*(\mu; a_n)$

Suppose $X_1, X_2, \ldots, X_n \in \mathcal{R}^d$ are independent random vectors from some population X, and $\theta \in \mathcal{R}^p$ is the parameter of interest defined through general estimating equation (2.12). The EL-based approximate $100(1-\alpha)\%$ confidence region for θ is defined as

$$\mathcal{CR}_\alpha = \{\theta : \; W_n(\theta) \le \chi_q^2(\alpha)\}$$

where $\chi_q^2(\alpha)$ is the upper α quantile of χ^2 distribution with q degrees of freedom. The AEL-based $100(1-\alpha)\%$ confidence region is defined in the same way except for replacing the foregoing $W_n(\theta)$ by $W_n^*(\theta; a_n)$.

In the case of population mean, since $W_n(\mu)$ is only well defined for μ in the convex hull of the sample, and $W_n(\mu)$ tends to infinity as μ approaches to the boundary of the convex hull, the EL-based confidence region for population mean is always of finite size. On the other hand, however, we find that $W_n^*(\theta; a_n)$ is bounded from above for any given n. Hence, AEL may give unbounded confidence region when the sample size is not large enough, or the confidence level $(1-\alpha)$ is too high. We state this result as follows.

Theorem 3.3. *Suppose we have a finite sample $X_1, X_2, \ldots, X_n$ and let $W_n^*(\theta; a_n)$ be the adjusted empirical likelihood ratio statistic defined in equation* (3.4)*. For any θ, we have*

$$W_n^*(\theta; a_n) \le -2\, n \log \left[\frac{(n+1)\, a_n}{n\,(1+a_n)}\right] - 2 \log \left[\frac{n+1}{1+a_n}\right].$$

Proof. Let

$$\begin{aligned} q_1 = q_2 = \cdots = q_n &= \frac{1}{n}\,\frac{a_n}{1+a_n}, \\ q_{n+1} &= \frac{1}{1+a_n}. \end{aligned}$$

It is clear that $q_i > 0$ and $\sum_{i=1}^{n+1} q_i = 1$. In addition, it is seen that

$$\begin{aligned} \sum_{i=1}^{n+1} q_i\, g_i(\theta) &= \frac{a_n}{1+a_n}\,\frac{1}{n} \sum_{i=1}^{n} g_i(\theta) + \frac{1}{1+a_n}\,[-a_n\, \bar{g}_n(\theta)] \\ &= \frac{a_n}{1+a_n}\,\bar{g}_n(\theta) - \frac{a_n}{1+a_n}\,\bar{g}_n(\theta) \\ &= 0. \end{aligned}$$

Hence, $\{q_i\}_{i=1}^{n+1}$ is a set of sub-optimal weights for $W_n^*(\theta; a_n)$. According to the definition of $W_n^*(\theta; a_n)$, we thus have

$$W_n^*(\theta; a_n) \le -2 \sum_{i=1}^{n+1} \log[(n+1)\, q_i] = -2\, n \log \left[\frac{(n+1)\, a_n}{n\,(1+a_n)}\right] - 2 \log \left[\frac{n+1}{1+a_n}\right].$$

It completes the proof. ∎

For the population mean, the next theorem shows that the upper bound in Theorem 3.3 is the supremum of $W_n^*(\mu; a_n)$.

Theorem 3.4. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. d-dimensional random vectors and μ be the population mean. Denote M as the upper bound in Theorem 3.3. For any d-dimensional unit vector $\mathbf{v}$, consider the half line $\bar{X}_n + t\,\mathbf{v}$ with $t > 0$. We have*

$$\lim_{t\to\infty} W_n^*(\bar{X}_n + t\,\mathbf{v}; a_n) = M.$$

Proof. We will present proof for the case when $d = 1$ and for the case when $d > 1$ separately.
Case 1: $d = 1$. We will only present the proof of the theorem in the case where $\mu \to -\infty$; the proof in the case where $\mu \to \infty$ is similar.

Let $\{p_i\}_{i=1}^{n+1}$ be the optimal weights for $W_n^*(\mu; a_n)$. We prove the result in three steps. Firstly, we demonstrate that $\lim\limits_{\mu\to-\infty} p_{n+1} = 1/(1+a_n)$. Secondly, we further show that $\lim\limits_{\mu\to-\infty} p_i = a_n/[n\,(1 + a_n)]$ for $i = 1, 2, \ldots, n$. In the final step, the conclusion of the proposition readily follows.
Step 1. Consider any μ such that $\mu < X_{(1)}$. Note that $\{p_i\}_{i=1}^{n+1}$ satisfy

$$\sum_{i=1}^{n+1} p_i\,(X_i - \mu_i) = 0.$$

From the above equation, we get

$$\sum_{i=1}^{n} p_i\,(X_i - \mu) = -p_{n+1}\,(X_{n+1} - \mu) = p_{n+1}\,a_n\,(\bar{X}_n - \mu).$$

Thus,

$$p_{n+1} = \sum_{i=1}^{n} \frac{p_i\,(X_i - \mu)}{a_n\,(\bar{X}_n - \mu)}.$$

Note that $0 < X_{(1)} - \mu \le X_i - \mu \le X_{(n)} - \mu$ for $i = 1, 2, \ldots, n$, and

$\sum_{i=1}^n p_i = 1 - p_{n+1}$. Hence

$$p_{n+1} \leq \sum_{i=1}^{n} \frac{p_i\left(X_{(n)}-\mu\right)}{a_n\left(\bar{X}_n-\mu\right)} = \frac{\sum_{i=1}^n p_i}{a_n} \frac{X_{(n)}-\mu}{\bar{X}_n-\mu} = \frac{1-p_{n+1}}{a_n} \frac{X_{(n)}-\mu}{\bar{X}_n-\mu}.$$

Since $p_{n+1} < 1$, we get an upper bound for p_{n+1} from the above equation:

$$p_{n+1} \leq \frac{X_{(n)}-\mu}{a_n\left(\bar{X}_n-\mu\right)+\left(X_{(n)}-\mu\right)}.$$

Similarly, we get a lower bound for p_{n+1}:

$$p_{n+1} \geq \frac{X_{(1)}-\mu}{a_n\left(\bar{X}_n-\mu\right)+\left(X_{(1)}-\mu\right)}.$$

Letting $\mu \to -\infty$, we get

$$\frac{1}{1+a_n} \leq \underline{\lim}_{\mu\to-\infty} p_{n+1} \leq \overline{\lim}_{\mu\to-\infty} p_{n+1} \leq \frac{1}{1+a_n}.$$

Hence,

$$\lim_{\mu\to-\infty} p_{n+1} = \frac{1}{1+a_n}. \tag{3.6}$$

Step 2. For $i = 1, 2, \ldots, n+1$, p_i can be expressed as

$$p_i = \frac{1}{n+1} \frac{1}{1+\lambda\left(X_i-\mu\right)} \tag{3.7}$$

for some λ. By equation (3.7) with $i = n+1$, we have

$$\lambda = -\frac{(n+1)-p_{n+1}^{-1}}{(n+1)(X_{n+1}-\mu)} = \frac{(n+1)-p_{n+1}^{-1}}{a_n\,(n+1)(\bar{X}_n-\mu)}.$$

For $i = 1, 2, \ldots, n$, substituting this expression of λ into equation (3.7) leads to

$$\begin{aligned} p_i &= \frac{1}{n+1}\left[1+\frac{(n+1)-p_{n+1}^{-1}}{(n+1)\,a_n}\,\frac{X_i-\mu}{\bar{X}_n-\mu}\right]^{-1} \\ &= \left[(n+1)+\frac{(n+1)-p_{n+1}^{-1}}{a_n}\,\frac{X_i-\mu}{\bar{X}_n-\mu}\right]^{-1}. \end{aligned}$$

Letting $\mu \to -\infty$ and using equation (3.6), we get

$$\lim_{\mu\to-\infty} p_i = \left[(n+1)+\frac{(n+1)-(1+a_n)}{a_n}\right]^{-1} = \frac{1}{n}\,\frac{a_n}{1+a_n}.$$

Step 3. Since $\{p_i\}_{i=1}^{n+1}$ are the optimal weights for $W_n^*(\mu; a_n)$, we have

$$W_n^*(\mu; a_n) = -2\sum_{i=1}^{n+1}\log[(n+1)\,p_i].$$

Consequently,

$$\begin{aligned} \lim_{\mu\to-\infty} W_n^*(\mu; a_n) &= \lim_{\mu\to-\infty} -2\sum_{i=1}^{n+1}\log[(n+1)\,p_i] \\ &= -2\,n\,\log\left[\frac{(n+1)\,a_n}{n\,(1+a_n)}\right] - 2\,\log\left[\frac{n+1}{1+a_n}\right], \end{aligned}$$

which is the conclusion.

Case 2: $d > 1$. For any d-dimensional unit vector $\mathbf{v}$ and $t > 0$, let $\{p_i\}_{i=1}^{n+1}$ be the optimal weights for $W_n^*(\bar{X}_n + t\,\mathbf{v}; a_n)$. Consider $Y_i = \mathbf{v}^T\,(X_i - \bar{X}_n)$ for $i = 1, 2, \ldots, n+1$. It is easy to verify that

$$\sum_{i=1}^{n+1} p_i\,(Y_i - t) = 0.$$

Define

$$\widetilde{R}_n^*(t; a_n) = \max\left\{\prod_{i=1}^{n+1}(n+1)\,p_i :\ p_i > 0, \sum_{i=1}^{n+1} p_i = 1, \sum_{i=1}^{n+1} p_i\,(Y_i - t) = 0\right\},$$

and $\widetilde{W}_n^*(t; a_n) = -2\log \widetilde{R}_n^*(t; a_n)$. Since $\{p_i\}_{i=1}^{n+1}$ are sub-optimal weights for $\widetilde{W}_n^*(t; a_n)$, we have

$$\widetilde{W}_n^*(t; a_n) \le W_n^*(\mu; a_n) \le M.$$

We have already proved that $\lim_{t\to\infty}\widetilde{W}_n^*(t; a_n) = M$. Consequently, we get

$$M = \varliminf_{t\to\infty}\widetilde{W}_n^*(t; a_n) \le \varliminf_{t\to\infty} W_n^*(\mu; a_n) \le \varlimsup_{t\to\infty} W_n^*(\mu; a_n) \le M.$$

Hence

$$\lim_{t\to\infty} W_n^*(\mu; a_n) = M.$$

This completes the proof. $\square$

Theorem 3.3 reveals that $W_n^*(\theta; a_n)$ is a bounded function of θ; Figure 3.3 shows the relationship between the upper bound and the sample size when $a_n = \log(n)/2$. When the sample size is small or the dimension is high, the upper bound of $W_n^*(\theta; a_n)$ is likely to be smaller than the upper α quantile of the χ^2 distribution. When this happens, the approximate confidence region based on the χ^2 calibration becomes the entire parameter space. Figure 3.4 shows the minimum sample size needed for the adjusted empirical likelihood method to give bounded confidence regions versus different degrees of freedom and confidence levels.

Even if the minimum sample size is attained in a particular situation, the adjusted empirical likelihood method may still produce unreasonably large confidence regions. For example, suppose we have a univariate sample of size 5 and we would like to construct the AEL-based 95% confidence interval for the population mean. As proposed by Chen et al. (2008), a_n is chosen to be $\log(5)/2 = 0.805$. For this choice of a_n, the upper bound of $W_n^*(\mu; a_n)$

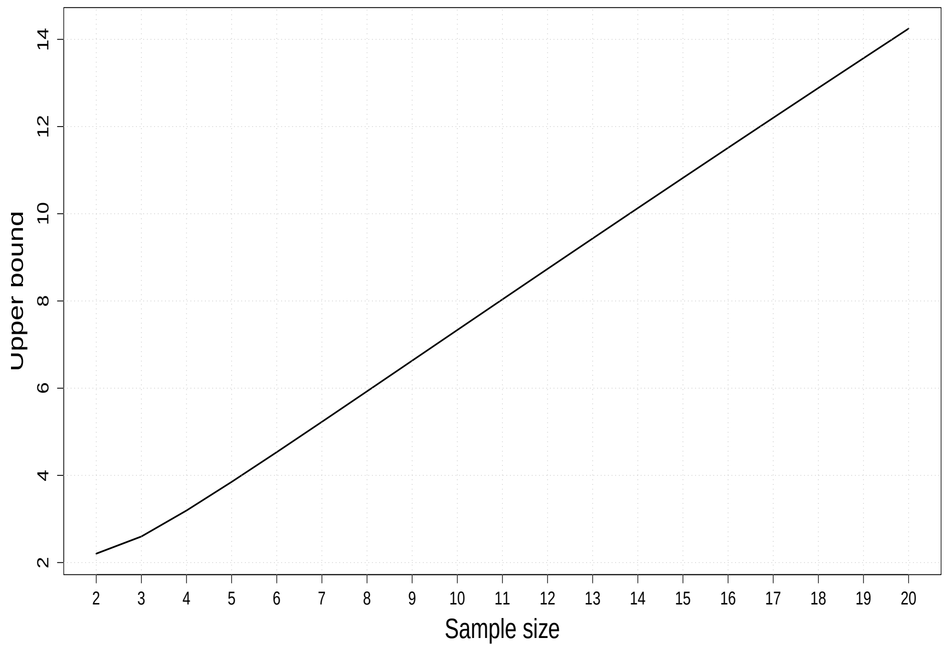


Figure 3.3: Plot of upper bound against sample size

is 3.851, which is only slightly larger than the upper 5% quantile of the χ^2_1 distribution, 3.841. Because of this, the resulting confidence interval is very long. We can imagine that the coverage rate is much higher than the nominal level 95%. Figure 3.5 illustrates this point with a data set of 5 points generated from $\mathrm{N}(0,1)$.

In the case of population mean, we may modify the adjusted empirical likelihood method so that the resulting $W_n^*(\mu; a_n)$ becomes unbounded from above. To motivate such a modification, let us once again look into Theorem 3.3. We view $W_n^*(\mu; a_n)$ as a function of a_n while regarding μ as a fixed constant satisfying $\mu \neq \bar{X}_n$. It is seen that $W_n^*(\mu; a_n)$ equals $W_n(\mu)$ when $a_n = 0$, and $W_n^*(\mu; a_n)$ is a decreasing function of a_n on the closed interval $[0, n]$. See Figure 3.6 for an illustration. Attempting to make $W_n^*(\mu; a_n)$

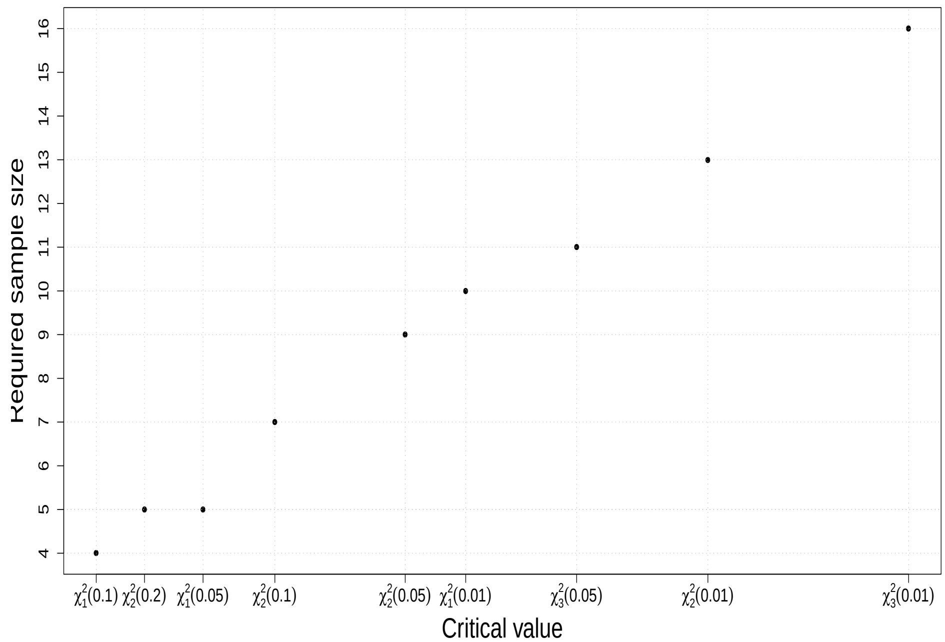


Figure 3.4: Plot of required sample size for various critical values

unbounded from above, we consider replacing the constant a_n by

$$a_n(\mu) = a_n \cdot \exp\left\{-\sqrt{(\bar{X}_n - \mu)^T \, S_n^{-1} \, (\bar{X}_n - \mu)}\right\}, \tag{3.8}$$

where S_n is the sample variance-covariance matrix. We assume that S_n is nonsingular.

The resulting $W_n^*(\mu; a_n(\mu))$ is always larger than $W_n^*(\mu; a_n)$ but smaller than $W_n(\mu)$ for any value of μ since $a_n(\mu)$ is always smaller than a_n but larger than 0. As μ deviates from $\bar{X}_n$, $a_n(\mu)$ tends to zero and thus $W_n^*(\mu; a_n(\mu))$ approaches $W(\mu)$. As a result, $W_n^*(\mu; a_n(\mu))$ is unbounded from above. Figure 3.7 visualizes the effect of $a_n(\mu)$ in the univariate case.

The modified adjusted empirical likelihood possesses two key advantages. Firstly, the modified adjusted empirical likelihood ratio statistic preserves the monotonicity of the adjusted empirical likelihood, and therefore the cor-

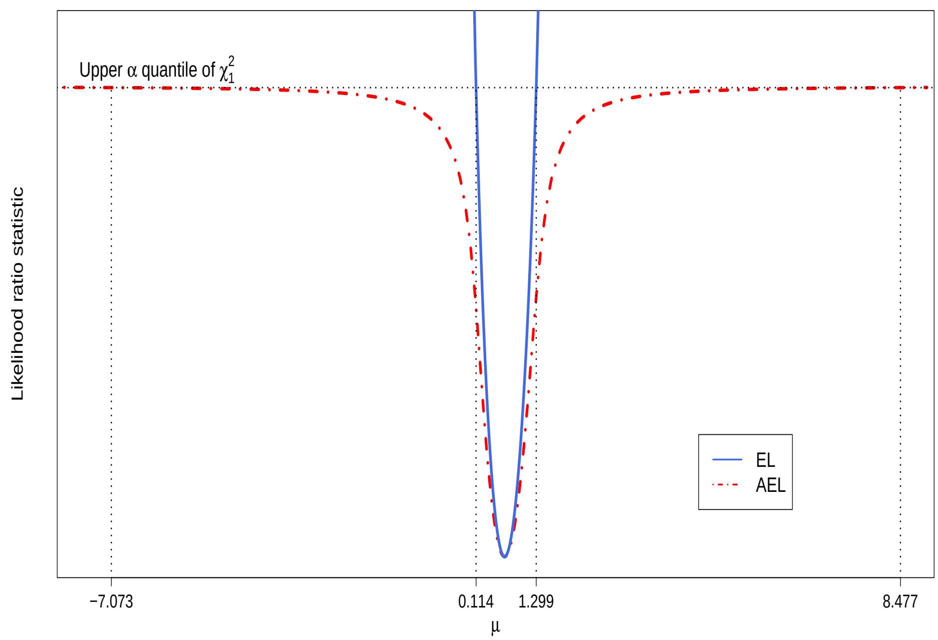


Figure 3.5: The EL-based and AEL-based 95% confidence intervals for the population mean

responding confidence region is star-shaped and bounded. Based on the foregoing discussion, we can imagine that the AEL-based confidence region contains the confidence region based on the modified adjusted empirical likelihood while the latter one contains the EL-based confidence region. Figure 3.8 shows the 95% approximate confidence regions based on the empirical likelihood method, the adjusted empirical likelihood method and the modified adjusted empirical likelihood method.

Secondly, it is seen that the multiplier in the definition of $a_n(\mu)$ converges to 1 of order $n^{-1/2}$ as $n \to \infty$ when μ equals the true value μ_0. This fact implies the modified adjusted empirical likelihood preserves both the first-order and second-order asymptotic properties of the original adjusted empirical likelihood.

Obviously, there are many choices for the multiplier in the definition of

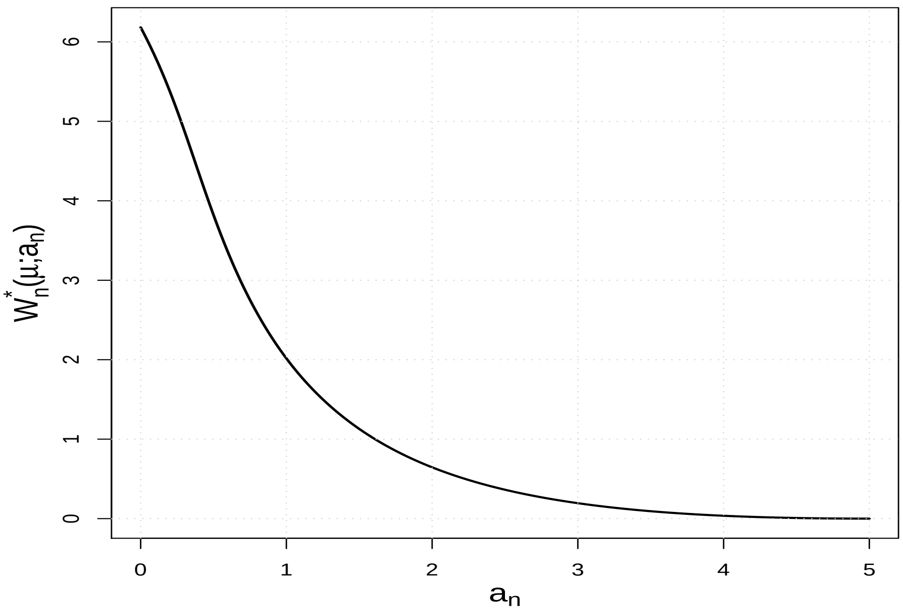


Figure 3.6: $W_n^*(\mu; a_n)$ as a function of a_n

$a_n(\mu)$. Based on the foregoing discussion, we may require the multiplier should satisfy:

(1) it decreases to 0 as μ deviates from $\bar{X}_n$; and

(2) it converges to 1 of order $n^{-1/2}$ as n increases.

If a multiplier satisfies these two conditions, the corresponding modified adjusted empirical likelihood not only gives bounded confidence regions, but also preserves the asymptotic and finite-sample properties of the original adjusted empirical likelihood presented in this thesis.

With these two conditions in mind, we can see that there are still many kinds of choice for the multiplier. The "optimal" choice of multiplier would be an interesting topic for future research.

In the literature, there is another variant of adjusted empirical likelihood that gives unbounded likelihood ratio statistic. Emerson and Owen (2009)

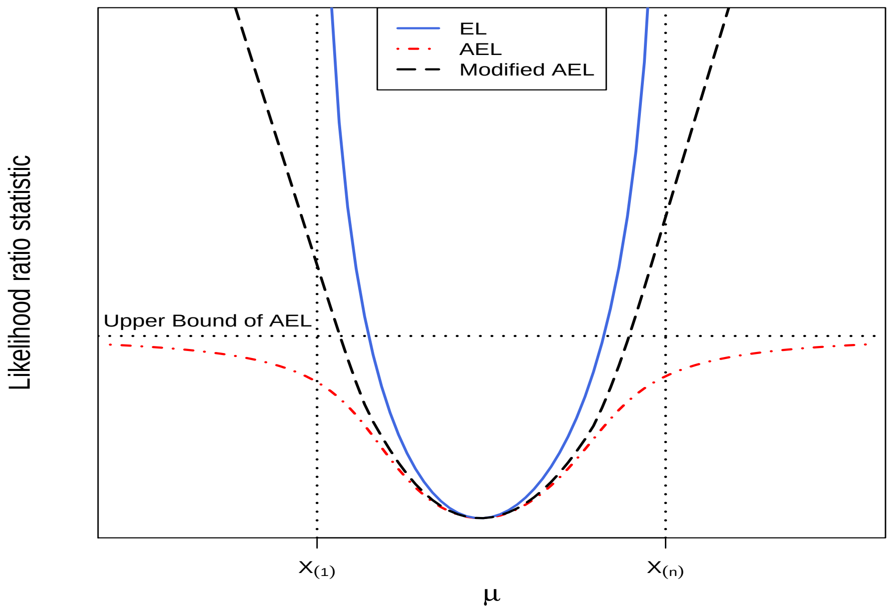


Figure 3.7: The effect of $a_n(\mu)$

also discover the boundedness of $W_n^*(\mu; a_n)$, and they consider a different way to modify the adjusted empirical likelihood so as to get an unbounded likelihood ratio statistic in the case of population mean. They propose adding two pseudo points to the original sample. More specifically, for any $\mu \neq \bar{X}_n$, the first pseudo point X_{n+1} is also added on the further side of μ but the distance between X_{n+1} and μ is a constant s, and the second X_{n+2} is added such that $\bar{X}_n$ is the midpoint of X_{n+1} and X_{n+2}. Then the likelihood ratio statistic is defined as

$$W_n^*(\mu) = -2 \max \left\{ \sum_{i=1}^{n+2} \log[(n+2)\, p_i] : p_i > 0, \sum_{i=1}^{n+2} p_i = 1, \sum_{i=1}^{n+2} p_i\, (X_i - \mu) = 0 \right\}.$$

The resulting method is called the balanced augmented empirical likelihood method. The method is called "balanced" because the sample mean of

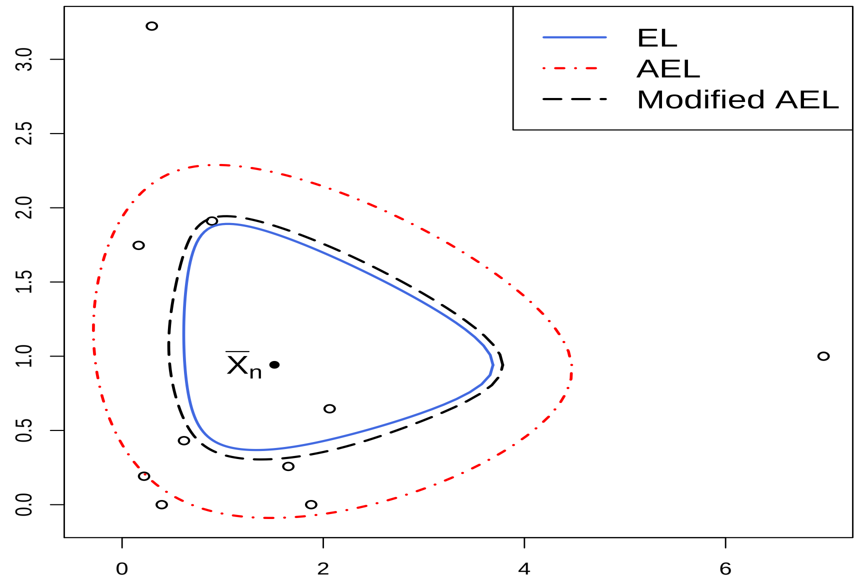


Figure 3.8: The 95% approximate confidence regions produced by EL, AEL and modified AEL

$\{X_1, X_2, \ldots, X_{n+2}\}$ is maintained at $\bar{X}_n$. They demonstrate this likelihood ratio statistic is unbounded from above, and establish the finite-sample relationship between this modified likelihood ratio method and the well-known Hotelling's T^2 test through the tuning parameter s. This topic is beyond the scope of the thesis; details can be found in Emerson and Owen (2009).

Chapter 4

Empirical Results

In this chapter, we conduct simulation studies on the finite-sample properties of the empirical likelihood method and its several variants. Particularly, we investigate the coverage probabilities and the sizes of a number of confidence regions for population mean.

Constructing confidence regions for population mean based on a simple random sample of n observations is a classical problem in statistical inference. The most widely-used method of constructing confidence regions for population mean is based on the Hotelling's T^2 statistic

$$T_n^2(\mu) = n\,(\bar{X}_n - \mu)^T S_n^{-1}(\bar{X}_n - \mu),$$

where $\bar{X}_n$ and S_n are the sample mean and the sample variance-covariance matrix, respectively. If the population distribution is multivariate normal of dimension d, then $(n-d)T^2(\mu_0)/[d\,(n-1)]$ is known to have an F distribution with d and $n-d$ degrees of freedom where μ_0 is the true parameter value. A $100(1-\alpha)\%$ confidence region for μ is given by

$$\mathcal{CR} = \left\{\mu : T_n^2(\mu) \le \frac{d\,(n-1)}{n-d}\,\mathrm{F}_{d,\,n-d}(\alpha)\right\},$$

where $\mathrm{F}_{d,\,n-d}(\alpha)$ denotes the upper α quantile of F distribution with d and $n-d$ degrees of freedom. When $d=1$, Hotelling's T^2 statistic becomes the square of the well-known Student's t statistic.

Many practitioners prefer using this kind of confidence region based on normal approximation because of its easy calculation and straightforward interpretation. Moreover, many numerical studies have found that the foregoing form of confidence region has surprisingly accurate coverage rate even when the population distribution is not normal and the sample size is small.

We investigate the coverage rates and the sizes of approximate 90% and 95% confidence intervals/regions in the cases of one-dimensional mean and two-dimensional mean. Seven methods are considered:

1. The Hotelling's T^2 method, denoted as T^2;
2. The original empirical likelihood method, denoted as EL;
3. The adjusted empirical likelihood method with $a_n = \log(n)/2$, denoted as AEL;
4. The modified adjusted empirical likelihood method with $a_n = \log(n)/2$, denoted as MAEL;
5. The Bartlett corrected empirical likelihood method, denoted as EL*;
6. The adjusted empirical likelihood method with $a_n = b/2$ where b is estimated by moments, denoted as AEL*;
7. The modified adjusted empirical likelihood method with $a_n = b/2$ where b is estimated by method of moments, denoted as MAEL*;

4.1 Confidence Intervals for One-Dimensional Mean

Three sample sizes ($n = 5, 10, 50$) are considered. For each sample size, we generated $1,000$ samples from each of the following four distributions:

1. Standard normal distribution, denoted as $\mathrm{N}(0, 1)$;
2. χ^2 distribution with 1 degree of freedom, denoted as χ_1^2;
3. Exponential distribution with rate 1, denoted as Exp(1); and
4. A normal mixture $0.1\,\mathrm{N}(-9, 1) + 0.9\,\mathrm{N}(1, 1)$, denoted as $0.1\,\mathrm{N}_1 + 0.9\,\mathrm{N}_2$.

For each sample, we calculated the 90% and 95% confidence intervals. Table 4.1 reports the coverage frequencies, and Table 4.2 gives the average lengths of the corresponding intervals.

Table 4.1: Coverage rates for one-dimensional mean

	Nominal level					
	0.9			0.95		
$\mathrm{N}(0,1)$	$n=5$	$n=10$	$n=50$	$n=5$	$n=10$	$n=50$
T^2	0.895	0.909	0.892	0.956	0.952	0.949
EL	0.757	0.858	0.889	0.815	0.913	0.935
AEL	0.881	0.910	0.897	1.000	0.954	0.945
MAEL	0.796	0.884	0.896	0.845	0.928	0.942
EL*	0.782	0.876	0.891	0.834	0.921	0.939
AEL*	0.797	0.880	0.891	0.855	0.923	0.939
MAEL*	0.779	0.870	0.889	0.831	0.917	0.937
χ_1^2						
T^2	0.785	0.809	0.898	0.840	0.857	0.940
EL	0.663	0.766	0.892	0.733	0.832	0.943
AEL	0.789	0.827	0.905	0.995	0.881	0.951
MAEL	0.709	0.804	0.904	0.765	0.856	0.947
EL*	0.691	0.784	0.901	0.757	0.850	0.946
AEL*	0.715	0.792	0.901	0.770	0.853	0.947
MAEL*	0.688	0.782	0.900	0.749	0.843	0.945
Exp(1)						
T^2	0.829	0.849	0.884	0.875	0.896	0.932
EL	0.712	0.799	0.878	0.765	0.869	0.934
AEL	0.829	0.864	0.892	0.999	0.913	0.945
MAEL	0.751	0.831	0.888	0.800	0.889	0.945
EL*	0.734	0.824	0.886	0.785	0.884	0.941
AEL*	0.749	0.828	0.886	0.813	0.887	0.941
MAEL*	0.722	0.819	0.884	0.773	0.878	0.940
$0.1\,\mathrm{N}_1 + 0.9\,\mathrm{N}_2$						
T^2	0.648	0.677	0.896	0.724	0.764	0.944
EL	0.474	0.625	0.909	0.534	0.660	0.952
AEL	0.636	0.660	0.923	0.999	0.718	0.959
MAEL	0.522	0.645	0.918	0.573	0.680	0.959
EL*	0.497	0.632	0.920	0.547	0.666	0.957
AEL*	0.516	0.637	0.921	0.584	0.670	0.957
MAEL*	0.495	0.632	0.917	0.545	0.665	0.957

Table 4.2: Confidence interval lengths for one-dimensional mean

	Nominal level					
	0.9			0.95		
N(0,1)	$n=5$	$n=10$	$n=50$	$n=5$	$n=10$	$n=50$
T^2	1.797	1.130	0.474	2.341	1.395	0.567
EL	1.178	0.966	0.465	1.371	1.150	0.556
AEL	1.729	1.135	0.485	18.201	1.398	0.581
MAEL	1.307	1.046	0.480	1.521	1.239	0.574
EL*	1.266	1.019	0.472	1.466	1.213	0.564
AEL*	1.328	1.031	0.472	1.596	1.232	0.565
MAEL*	1.240	1.002	0.470	1.438	1.189	0.562
χ^2_1						
T^2	2.252	1.497	0.660	2.933	1.847	0.791
EL	1.447	1.265	0.657	1.675	1.500	0.790
AEL	2.156	1.491	0.685	22.802	1.844	0.825
MAEL	1.614	1.370	0.678	1.882	1.621	0.815
EL*	1.547	1.339	0.675	1.782	1.587	0.813
AEL*	1.624	1.361	0.676	1.941	1.626	0.814
MAEL*	1.542	1.327	0.673	1.791	1.572	0.809
Exp(1)						
T^2	1.686	1.077	0.468	2.196	1.329	0.561
EL	1.091	0.913	0.464	1.266	1.084	0.557
AEL	1.617	1.075	0.484	17.075	1.327	0.581
MAEL	1.215	0.989	0.479	1.415	1.170	0.574
EL*	1.169	0.965	0.474	1.349	1.145	0.569
AEL*	1.226	0.979	0.474	1.467	1.169	0.569
MAEL*	1.150	0.949	0.472	1.333	1.125	0.566
$0.1\,\mathrm{N}_1 + 0.9\,\mathrm{N}_2$						
T^2	4.796	3.309	1.498	6.246	4.083	1.796
EL	3.076	2.801	1.468	3.562	3.327	1.756
AEL	4.591	3.302	1.532	48.563	4.088	1.834
MAEL	3.433	3.032	1.517	4.006	3.592	1.812
EL*	3.289	2.978	1.497	3.792	3.536	1.791
AEL*	3.456	3.032	1.498	4.138	3.633	1.792
MAEL*	3.234	2.921	1.491	3.739	3.460	1.782

4.2 Confidence Regions for Two-Dimensional Mean

We also consider constructing confidence regions for the population mean of the following bivariate distributions:

1. Standard normal distribution, $\mathrm{N}(0, I_2)$;
2. Distribution of (X_1, X_2) where $X_1 \sim \Gamma(U, 1)$ and $X_2 \sim \Gamma(U^{-1}, 1)$ with $U \sim \text{Uniform}(1.5, 2)$, denoted as Gamma-Gamma.
3. (X_1, X_2) is bivariate normal distributed with $\mathbf{Var}(X_1) = \mathbf{Var}(X_2) = 1$ and $\mathbf{Cov}(X_1, X_2) = \rho$ given $\rho \sim \text{Uniform}(0, 1)$, denoted as Normal-Uniform.

Two sample sizes ($n = 10, 50$) are considered. For each sample size, we generated $1,000$ samples from each of the above three distributions. Approximate 90% and 95% confidence regions were calculated. The coverage frequencies and average areas of the confidence regions based on various methods are summarized in Table 4.3 and 4.4. Note that the area of confidence region is calculated approximately.

4.3 Summary

Under the standard normal model, T^2 has very accurate coverage rates compared to its nonparametric alternatives except the AEL. It is because that T^2-based confidence interval/region achieves the nominal level in theory regardless of the sample size. The performance of the nonparametric methods gets better when the sample size increases. Under other distribution models, the performances of all methods in small sample cases are unsatisfactory. Especially in the mixture normal model, the coverage rates are dramatically lower than the nominal level.

On average, T^2-based confidence interval/region has larger size than its nonparametric alternatives. It makes sense since T^2-based confidence interval/region has higher coverage rate. Because of this trade-off between

Table 4.3: Coverage rates for two-dimensional mean

	Nominal level			
	0.9		0.95	
$N(0, I_2)$	$n = 10$	$n = 50$	$n = 10$	$n = 50$
T^2	0.904	0.909	0.952	0.958
EL	0.766	0.898	0.833	0.946
AEL	0.883	0.917	0.966	0.958
MAEL	0.813	0.914	0.861	0.956
EL*	0.801	0.907	0.852	0.952
AEL*	0.826	0.907	0.876	0.952
MAEL*	0.795	0.905	0.848	0.952
Gamma-Gamma				
T^2	0.827	0.877	0.878	0.930
EL	0.708	0.876	0.766	0.921
AEL	0.827	0.892	0.926	0.935
MAEL	0.736	0.887	0.795	0.929
EL*	0.742	0.887	0.797	0.928
AEL*	0.760	0.884	0.816	0.928
MAEL*	0.727	0.882	0.781	0.928
Normal-Uniform				
T^2	0.920	0.899	0.959	0.960
EL	0.752	0.890	0.831	0.944
AEL	0.897	0.907	0.977	0.958
MAEL	0.807	0.896	0.867	0.953
EL*	0.797	0.894	0.858	0.951
AEL*	0.822	0.894	0.882	0.951
MAEL*	0.790	0.894	0.849	0.948

coverage probability and the size of confidence region, it is difficult to say which method has better performance. How to evaluate the performance of certain kind of confidence region based on both the coverage probability and the region size is still an open question.

Surprisingly, the AEL keeps up with T^2 in terms of both coverage rate and confidence interval/region size for most cases. However, note that the

Table 4.4: Confidence region areas for two-dimensional mean

	Nominal level			
	0.9		0.95	
$N(0, I_2)$	$n = 10$	$n = 50$	$n = 10$	$n = 50$
T^2	1.963	0.304	2.812	0.401
EL	1.115	0.287	1.422	0.378
AEL	1.821	0.314	3.511	0.414
MAEL	1.293	0.306	1.644	0.401
EL*	1.260	0.299	1.600	0.392
AEL*	1.356	0.299	1.799	0.393
MAEL*	1.211	0.296	1.537	0.388
Gamma-Gamma				
T^2	1.872	0.316	2.681	0.417
EL	1.031	0.302	1.307	0.398
AEL	1.715	0.330	3.331	0.436
MAEL	1.194	0.321	1.519	0.422
EL*	1.161	0.318	1.466	0.419
AEL*	1.269	0.319	1.779	0.420
MAEL*	1.117	0.314	1.410	0.413
Normal-Uniform				
T^2	1.676	0.263	2.401	0.347
EL	0.954	0.251	1.218	0.330
AEL	1.558	0.274	3.001	0.362
MAEL	1.107	0.267	1.408	0.351
EL*	1.081	0.261	1.374	0.344
AEL*	1.168	0.262	1.568	0.345
MAEL*	1.038	0.259	1.319	0.340

AEL-based confidence interval has substantially higher than nominal coverage rate when the sample size is 5 and the nominal level is 95% in the univariate case. It is in accordance with the discussion in Section 3.2.3. In this situation, the upper bound of the adjusted empirical likelihood ratio statistic is only slightly larger than the critical value. It results in very long interval, which in turn possesses higher-than-expected coverage probability.

As expected, the MAEL is a compromise between the EL and the AEL. We observe that the performance of MAEL is more similar to that of EL than that of AEL especially in the bivariate case. It may be due to the fact that the multiplier in the definition of pseudo point in MAEL decreases to 0 very fast as μ deviates from $\bar{X}_n$; the decreasing is even faster as the dimension increases. It implies the difference between the MAEL and EL likelihood ratio statistics is smaller than that between the MAEL and AEL likelihood ratio statistics.

The EL*, AEL* and MAEL* have similar performance because they are known to be precise up to the same order n^{-2}.

Chapter 5

Conclusion

The main interest of this thesis lies in the finite-sample properties of adjusted empirical likelihood and its implication to constructing confidence regions for population mean. The monotonicity property of the adjusted empirical likelihood ratio statistic guarantees that AEL-based confidence regions for population mean are at least star-shaped. It is a desirable property for confidence regions because of its intuitive interpretation. We also discovered the connection between empirical likelihood and adjusted empirical likelihood as a special case of a more general conclusion, which justified the empirical observation that AEL-based confidence regions have higher coverage probability than the corresponding EL-based confidence regions. The boundedness of adjusted empirical likelihood ratio statistic reveals that constant level of adjustment may produce inappropriate confidence regions when the sample size is not large enough or the nominal confidence level is too high. We attempted to modify the level of adjustment so as to obtain an unbounded likelihood ratio statistic, and justified the proposed modification preserves both asymptotic and finite-sample properties of the original adjusted empirical likelihood.

As future research, the convexity of AEL-based confidence regions for population mean is of interest; current empirical studies support this proposition. On the other hand, the choice of $a_n(\mu)$ also remains an interesting topic. As discussed in Section 3.2.3, we may adjust the trade-off between the coverage probability and the size of confidence region for population mean. If we can come up with a sensible criterion to evaluate certain kind of confidence regions by taking both its coverage probability and its size into consideration, we may be able to find a family of $a_n(\mu)$ such that the resulting AEL-based confidence region for population mean has both asymptotic and finite-sample advantages.

Bibliography

Barndorff-Nielsen, O. E. and Cox, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *Journal of Royal Statistical Society, Series B*, 46(3):483–495.

Chan, N. H. and Ling, S. (2006). Empirical likelihood for GARCH models. *Econometric Theory*, 22:403–428.

Chen, J., Chen, S.-Y., and Rao, J. N. K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics*, 31(1):53–68.

Chen, J., Peng, L., and Zhao, Y. (2009). Empirical likelihood based confidence intervals for copulas. *Journal of Multivariate Analysis*, 100(1):137–151.

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9:385–406.

Chen, J., Sitter, R. R., and Wu, C. (2002). Using empirical likelihood method to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89(1):230–237.

Chen, J., Variyath, A. M., and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, 17(2):426–443.

Corotis, R. B., Sigl, A. B., and Klein, J. (1978). Probability models of wind velocity magnitude and persistence. *Solar Energy*, 20(6):483–493.

Diciccio, T., Hall, P., and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *The Annals of Statistics*, 19(2):1053–1061.

Emerson, S. C. and Owen, A. B. (2009). Calibration of the empirical likelihood method for a vector mean. *Electronic Journal of Statistics*, 3:1161–1192.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review*, 58(2):109–127.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hua, L. (2009). A report on an adjusted empirical likelihood. Manuscript, 16 pages.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20(4):493–506.

Liu, Y. and Chen, J. (2010). Adjusted empirical likelihood with high-order precision. *The Annals of Statistics*, 38(3):1341–1362.

Liu, Y. and Yu, C. W. (2010). Bartlett correctable two-sample adjusted empirical likelihood. *Journal of Multivariate Analysis*, 101(7):1701–1711.

Lun, I. Y. and Lam, J. C. (2000). A study of weibull parameters using long-term wind observations. *Renewable Energy*, 20(2):145–153.

Nordman, D. J. and Caragea, P. C. (2008). Point and interval estimation of variogram models using spatial empirical likelihood. *Journal of the American Statistical Association*, 103(481):350–361.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statstics*, 18(1):90–120.

Owen, A. B. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19(4):1725–1747.

Owen, A. B. (2001). *Empirical Likelihood.* Chapman & Hall/CRC.

Qin, G. and Zhou, X.-H. (2005). Empirical likelihood inference for the area under the ROC curve. *Biometrics*, 62(2):613–622.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B*, 69(1):101–122.

Variyath, A. M., Chen, J., and Abraham, B. (2010). Empirical likelihood based variable selection. *Journal of Statistical Planning and Inference*, 140(4):971–981.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Zhu, H., Zhou, H., Chen, J., Li, Y., Lieberman, J., and Styner, M. (2009). Adjusted exponentially tilted likelihood with applications to brain morphology. *Biometrics*, 65(3):919–927.