UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS & SOCIAL SCIENCES

School of Social Sciences

Some Aspects of Empirical Likelihood

by

Xing Wang

Thesis for the degree of Doctor of Philosophy

December 2008

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS & SOCIAL SCIENCES

SCHOOL OF SOCIAL SCIENCES

Doctor of Philosophy

Some Aspects of Empirical Likelihood

by

Xing Wang

Chapter 1 is a non technical introduction to the thesis.

In chapter 2, *Basics of Large Deviation Theory*, we illustrate the basic idea of large deviation theory and briefly review the history of its development. As a preparation, some of the important theorems which we will employ in the following chapters are also introduced.

In chapter 3, *Asymptotic Optimality of Empirical Likelihood Tests With Weakly Dependent Data*, we extend the result of Kitamura (2001) to stationary mixing data. The key thing in proving the large deviation optimality is that the empirical measure of the independently and identically distributed data will obey the large deviation principal (LDP) with rate function equal to the relative entropy, but in general the large deviation performance of empirical measure of dependent data is complicated. In this chapter we add S-mixing condition to the stationary process and we show that the rate function of the LDP of S-mixing process is indeed equal to the relative entropy, and then asymptotic optimality follows from the large deviation inequality.

In chapter 4, *Large Deviations of Empirical Likelihood with Nuisance Parameters*, we discuss the asymptotic efficiency of empirical likelihood in the presence of nuisance parameters combined with augmented moment conditions. We show that in the presence of nuisance parameters, the asymptotic efficiency of the empirical likelihood estimator of the parameter of interest will increase by adding more moment conditions, in the sense of the positive semidefiniteness of the difference of information matrices. As a by-product, we point out a necessary condition for the asymptotic efficiency to be increased when more moment condition are added. We also derive asymptotic lower bounds of the minimax risk functions for the estimator of the parameter of interest, and we show that the empirical likelihood estimator can achieve this bound.

In chapter 5, *Empirical Likelihood Estimation of Auction Models via Simulated Moment Conditions*, we apply empirical likelihood estimation to the simplest first-price sealed bid auction model with independent private values. Through estimation of the parameter in the distribution function of bidders' private values we consider a potential problem in the EL inference when the moment condition is not in an explicit form and hard to compute, or even not continuous in the parameter of interest. We deal with this issue following the method of simulated moment through importance sampling. We demonstrate the convergence of the empirical likelihood estimator from the simulated moment condition, and found that the asymptotic variance is larger than usual which is disturbed by simulation.

DECLARATION OF AUTHORSHIP

I, ... ,

declare that the thesis entitled

. . . .

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed: . . . .

Date:. . . . .

Acknowledgments

I would like to express most special thanks to my supervisor, Professor Grant Hillier, without whom this work could not be done. Grant brought me into the fantastic world of econometrics, and guided me through the whole process with many invaluable instructions and suggestions. He is not only a great mentor, but also a cornerstone of my career development. Also I would like to thank Professor Jean-Yves Pitarakis, who gave me many helpful comments, especially on some references of large deviation of stochastic processes.

Also, many thanks go to Gill Crumplin, Chris Thorn, Fatima Fauzel, and other members of the administrative support team. They are always ready to help and their efficient arrangement made the research environment smooth and comfortable.

Financial support from the Dorothy Hodgkin Postgraduate Scheme is gratefully acknowledged. Without these funding I could not have the chance to pursue my PhD.

My life in Southampton was made enjoyable by many friends, especially Yu Liu, who often cooked my favorite Sichuan food for me, and Corrado Giulietti, who helped me with computer stuff and played a lot of badminton with me.

The last, but far from the least, I wish to thank my family: my mother, my sister, and other relatives. Their support, understanding and encouragement have been always accompanying me. Particularly, I want to thank my wife, Bobo, you makes my life beautiful and promising.

TO

THE MEMORY OF MY FATHER

AN WEN, WANG

Contents

List of symbols and notation

$H\left(v\,\middle\|\mu\right)$	Kullback-Leibler distance between v and μ, where $v \ll \mu$
$v \ll \mu$	v is absolutely continuous with respect to μ
$ess\sup f$	the essential supremum of f
$M_1(\Sigma)$	space of probability measures on Σ
$B(\Sigma, R)$	space of all bounded, Borel measurable and real functions defined on Σ
$\\|\cdot\\|$	Euclidean norm
$A \subset\subset B$	A is a compact subset of B
Σ^n	n fold product space of Σ
P^n	n fold product measure of P
$x \vee y$	$\max\{x, y\}$
$x \wedge y$	$\min\{x, y\}$
I_A	Indicator function of set A
L^n	space of n integrable functions
$f^{(k)}$	k-th order derivative of the function f

Chapter 1

Introduction

The empirical likelihood (EL) method is a developing technique for estimation and inference, and it has attracted immense attention from both statisticians and econometricians in recent years. As an nonparametric analogue of parametric likelihood methods, EL is straightforward to be used to incorporate information from the observations, but without assuming a specific parametric distribution, and thus it is free of some forms of misspecification. It also shares many desirable statistical properties with ordinary likelihood methods. For instance, EL has been shown to be Bahadur optimal by Kitamura and Otsu (2005) in a minimax setting, while such optimality of the maximum likelihood estimator (MLE) is well known. Furthermore, EL has been found very convenient in dealing with moment condition models, and it is now being widely used as an important alternative to the generalized method of moments (GMM). Under some circumstances, EL has more desirable asymptotic properties than GMM. See Newey and Smith (2004).

This thesis makes contributions to several aspects of the EL method, particularly combined with the large deviation (LD) theory. Like standard asymptotic theory (SAT), LD also characterizes the limiting behavior of a sequences of random variables. The difference between them is as follows. SAT considers the *typical* behaviour of random variables, and checks if they will converge in probability or distribution to some fixed values or random variables when the sample size is large, often by applying a law of large numbers (LLN) or a central limit theorem (CLT). However, to some extent contrarily, LD characterizes the *deviant* behavior of random variables, which is sometimes called a *rare event* in the LD theory. Furthermore, LD theory focuses on a rate that the probability of the occurrence of a rare event vanishes. This accounts for the importance of LD in probability and statistical theory, because many standard inference problems involve an analysis of rare events. In this thesis we work with two examples of this: type I and type II errors in hypothesis testing, and the inaccuracy or the risk in estimation theory.

In Chapter 2 we briefly present a technical introduction to the theory of large deviations. We begin the illustration of the basic idea of LD with some examples in common probability and statistics problems, showing the exponential decay of the tail of the normal distribution. A very short history of the development of the LD theory and major contributors are also mentioned. We focus on the large deviations principle (LDP) which plays a central role in the following chapters. The LDP can be considered as a counterpart of CLT in SAT, since it provides asymptotic upper and lower bounds for sequence of probability measures. Moreover, the LD bounds are described by *rate functions*, which determine the speed of decay of the probability of rare events. Particularly, we introduce the famous Sanov theorem, which states that the empirical measure of a sequence of i.i.d. data satisfies the LDP, *i.e.*, the probability that the empirical measure lies in some subset of certain probability space (which can be treated as a rare event) will be bounded.

We also review two methods of how to prove that a sequence of probability measures satisfies the LDP. Firstly, sometimes it is convenient to show the existence of *weak* LDP, which is the LDP on compact sets. One can then extend the weak LDP to general sets by showing the sequence of probability is exponentially tight. Secondly, the contraction principle allows us to identify the LDP of a continuous function of probability measure family which satisfies the LDP.

Chapter 3 considers the application of LD theory in empirical likelihood based hypothesis testing. This is an extension of the work of Kitamura (2001) to weakly dependent data. In his paper Kitamura applies the Sanov theorem to show that the EL test of a set of moment conditions is optimal in Hoeffding sense, since the type II error of the EL test achieves the large deviation lower bound in an i.i.d. setting. The establishment of this optimality of EL test can be summarized as follows. Firstly, it can be shown that the EL test is to check if the empirical measure derived from the moment condition is 'close enough' to the true probability measure, and hence the rejection region of the EL test can be set by a value of the distance between the two measures. On the other hand, the Sanov theorem tells us that the empirical measure of i.i.d. observations obeys the LDP with rate function being the Kullback-Leibler distance, *i.e.*, the probability that the empirical measure lies in some certain area of the probability space is bounded by this rate function, and hence a large deviation lower bound of the asymptotic type II error can be established. Therefore, if we take the Kullback-Leibler distance as the distance between two probability measures, the optimality of EL test can be proved. Indeed, this framework is an application of the universal hypothesis problem in information theory, see Zeitouni and Gutman (1991) and Dembo and Zeitouni (1998).

Our contribution - in Chapter 3 - is to show that this type of optimality property of the EL test can also be obtained with dependent data. For this purpose it is necessary to add some restriction on the dependence of the data to make it satisfy the LDP with a suitable rate function which can be compared to the Kullback-Leibler distance. We adopt the S-mixing condition introduced by Bryc and Dembo (1996). The advantages of S-mixing are twofold. Firstly, it is a very weak assumption, and is implied by $\alpha-$mixing. Hence, the properties of EL test statistics derived by Kitamura (1997) and Smith (2004) under the $\alpha-$mixing condition are applicable under $S-$mixing as well. Secondly, we find that the rate function of the LDP of an $S-$mixing process equals the Kullback-Leibler distance under a certain assumption. Therefore, we can prove the optimality of EL test with an $S-$mixing process in a way similar to Kitamura (2001).

Chapter 4 considers the LD efficiency of EL estimation with nuisance parameters. Firstly, we present some standard asymptotic results of the EL method in the presence of nuisance parameters, particularly combined with augmented moment conditions. We find that the asymptotic efficiency of the estimator for the parameter of interest can be increased by additional nonorthogonal moment conditions, since it provides extra information. Secondly, we discuss the large deviation efficiency of the EL estimator for the parameter of interest, in the framework of Puhalskii and Spokoiny (1998), who find that if a family of probability measures characterized by some parameter satisfies the LDP, then there exists a asymptotic lower bound for the minimax risk of the estimator of the parameter. Following Kitamura and Otsu (2005), we show that the set of probability measures which satisfy the moment condition with nuisance parameters obeys the LDP with a particular rate function, and then a minimax risk bound can be determined by the likelihood ratio, the risk function, and the rate function. We find that the LD efficiency of the estimator of the parameter of interest can still achieve the lower bound.

In Chapter 5, we apply empirical likelihood to estimate the parameter of bidders' private values in auction models. At the beginning we describe the auction models in a game theoretical setting and briefly discuss the data generating processes of different auction models. We focus on the first price auction model with symmetric and independent private values, in which the winning bids can always be observed, and the distribution of the private values can be identified both parametrically and nonparametrically by the winning bids alone.

However, the moment condition derived from the Bayesian Nash equilibrium of the game theoretical model is highly nonlinear and not in a explicit form, and so is extremely hard to compute. Our contribution here is to suggest a method to deal with such moment conditions using empirical likelihood. We follow the method of simulated moment introduced by Pakes and Pollard (1989) and McFadden (1989) to simulate a new moment condition which is easy to handle. Particularly, we use importance sampling methods to do simulations, and we also find that this method can be applied when the moment condition is tractable, but is discrete in the parameter of interest, which is also the case considered recently by Parente and Smith (2008). We show the convergence of the empirical likelihood estimator from the simulated moment condition, and that the asymptotic variance is larger than usual by a factor due to the need for simulation, as one might imagine. A numerical simulation experiment is also conducted to illustrate the properties of the suggested procedure.

Chapter 2

Basics of Large Deviation Theory

In this preliminary chapter, we illustrate the basic ideas of large deviation theory, and briefly review the history of its development. As a preparation, some of the important theorems which we will employ in the following chapters are also introduced.

2.1 Introduction

Large deviation theory is now widely implemented in a variety of fields like mathematical statistics, engineering and physics, where we sometimes need to obtain detailed information on rare events. Rare events can be interesting and crucial, although they happen with relatively small probabilities. For example, in applications to queueing theory and communication systems, the rare event could represent an overload or breakdown of the system. In this case, large deviation methodology can lead to an efficient redesign of the system so that the overload or breakdown does not occur.

As for the limiting behavior of random variables, which is the main object of large deviation theory, actually we are familiar with some limit theorems such as the weak and strong law of large numbers, and the central limit theorem. These results depict the *typical* behavior of a random variable as converging to some other random variable or distribution. However, they tell little about the rate of convergence, or the *deviant* behavior at the tail of the distribution. Large deviation theory addresses just these two aspects. To have a first impression of large deviation theory in statistics, we begin with the following examples.

Example 1 (Dice Tossing) *Suppose we toss a dice 5 times, with a sequence of* *results* $(3, 3, 1, 6, 2)$ *.Then the empirical mean is* $\bar{x}_1 = (3 + 3 + 1 + 6 + 2)/5 = 3$*, and we can calculate the empirical distribution for the dice value* $(1, 2, 3, 4, 5, 6)$ *as*

$$\left(\frac{1}{5}, \frac{1}{5}, \frac{2}{5}, 0, 0, \frac{1}{5}\right).$$

However, we know that if the number of random throws are large enough, the theoretical distribution for the dice value should be

$$\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$$

and thus the mean value is $\bar{x} = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$*. Deviation from the theoretical mean value and distribution comes from insufficient number of throws, and large deviation theory tells us as the number of throws increases, deviations vanish at specific exponential rate. Figure 1 in the appendix shows that the tail of the distribution of the average value of dice decays exponentially as the number of throws increases, and the following example explores more theoretically of this issue.*

Example 2 (Tails of Normal Distribution) *Let* $x_{1,\ldots,}x_n$ *be a sequence of i.i.d., real-valued random variables drawn from standard normal distribution. Probably the most classical topic of probability theory is to study the behavior of the empirical mean:*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.1}$$

Since $\bar{x}$ *is again normally distributed, i.e,* $\bar{x} \sim N(0, 1/n)$*, it is easy to see for any interval* $A \subset \mathbb{R}$*,*

$$\Pr(\sqrt{n}\bar{x} \in A) \underset{n\to\infty}{\to} \frac{1}{\sqrt{2\pi}}\int_A e^{-x^2/2}dx. \tag{2.2}$$

This can be explained as $\bar{x}$ *takes the "typical" value of order* $1/\sqrt{n}$*, and this kind of convergence in* (2.2) *has been rigorously studied by the central limit theorem. An important and interesting problem is how frequently* $\bar{x}$ *takes some relatively large values, i.e,* $\bar{x}$ *exhibits "deviant" behavior? And people often want to know how "deviant" the behavior is. To see this, consider any* $\epsilon > 0$*,we have:*

$$\begin{aligned}\Pr(|\bar{x}| \geq \epsilon) &= 1 - \Pr(|\bar{x}| < \epsilon)\\ &= 1 - \tfrac{1}{\sqrt{2\pi}}\int_{-\epsilon\sqrt{n}}^{\epsilon\sqrt{n}} e^{-x^2/2}dx\end{aligned}$$

and this leads to:

$$\lim_{n\to\infty}\frac{1}{n}\log\Pr(|\bar{x}|\geq\epsilon)=-\frac{\epsilon^2}{2}. \tag{2.3}$$

Equation (2.3) *tells us that,* $\bar{x}$ *takes relatively large values with small probability of the order* $e^{-n\epsilon^2/2}$*. So a natural question will be, if some results similar to equation* (2.3) *can be obtained if* x_i *are not normally distributed? The answer is that in i.i.d. case the limit of* $\frac{1}{n}\log\Pr(|\bar{x}|\geq\epsilon)$ *always exists although its value depends on the distribution of* x_i*. Indeed, any probability measure of i.i.d. random variables is exponentially bounded by some rate function. This is just the content of the Sanov theorem, which we will discuss later on.*

The earliest idea of large deviations can be traced back to Laplace in the early 19th century, among his many contributions to probability and statistics. The first rigorous results concerning large deviations came from the Swedish mathematician Harald Cramer, who applied them to model the insurance business. Cramer gave a solution to his question for i.i.d. Gaussian random variables, where the rate function is expressed as a power series. However, the general abstract framework for the large deviation principle was proposed by Varadhan (1966, 1984), who may also have been the first to give such a terminology. Ventzell and Freidlin (1979) also make big contributions, describing their theory of small random perturbations of dynamic systems. A very incomplete list of mathematicians who have made important advances would include R. Ellis, A. Dembo and D. W. Strook. A systematic application of large deviations to statistical mechanics can be found in Ellis's work (1985), and Strook (1984) gave an introduction to the theory of large deviations together with a thorough treatment of the relation between empirical measure and analytical properties of Markov semigroups. A more comprehensive treatment of large deviation theory with application to statistics can be found in Dembo and Zeitouni (1998), and Deuschel and Strook (1989).

The remaining sections are organized as follows. Firstly we introduce the formal definition of the large deviation principle with related concepts. Section 2.3 provides some important theorems about large deviations which will be involved in our following chapters. More details of these well established results can be found in Dembo and Zeitouni (1998).

2.2 The Large Deviation Principle

2.2.1 Preliminaries

Let Σ be a topological space, so that open and closed subsets of it are well defined. Also denote the Borel σ-field on Σ as $\mathcal{A}$. Moreover, to avoid possible measurability problems, we assume all probability spaces to be complete and separable. The large deviation principle (henceforth LDP) is to characterize the limiting behavior of a sequence of probability measures $\{Q_n, n \geq 1\} \subseteq M_1(\Sigma)$ with respect to a rate function, where $M_1(\Sigma)$ represents the space of probability measures. Note that $M_1(\Sigma)$ includes discrete measures, such as empirical measure. Furthermore, we equip the measurable space $(\Sigma, \mathcal{A})$ with the τ-topology (strong topology) generated by the collection:

$$\left\{v \in M_1(\Sigma) : \left| \int_\Sigma f dv - x \right| < \varepsilon \right\} \tag{2.4}$$

where $x \in \mathbb{R}$, $\varepsilon > 0$ and $f \in B(\Sigma, \mathbb{R})$, the vector space of all bounded, real valued, Borel measurable functions on Σ. We equip f with the supremum norm.

Definition 1 *A function $I : \Sigma \to [0, \infty]$ is called a rate function if it is lower semicontinuous. If I is lower compact, i.e., the level set $\{x : I(x) \leq a\}$, $\forall a \in [0, +\infty)$ is compact, then I is called a good rate function.*[1]

Proposition 1 *A function f is lower compact if and only if for each decreasing sequence A_n of closed sets,*

$$\lim_{n\to\infty} \inf_{x\in A_n} f(x) = \inf_{x \in \cap_n A_n} f(x).$$

Proof. See, e.g., Puhalskii (2006). ■

Throughout, for any set Γ, let $\bar{\Gamma}$ denote the closure of Γ, Γ^o the interior of Γ, and Γ^c the complement of Γ. The infimum of a function over an empty set is set as ∞ (e.g., Dembo and Zeitouni (1998)).

Definition 2 (LDP) *A sequence of probability measures $\{Q_n, n \geq 1\}$ on Σ is*

[1] The reason that sometimes we want I to be a good rate function is that its infimum can be attained over closed sets.

said to satisfy the LDP with a rate function $I(x)$, if for all $\Gamma \in \mathcal{A}$,

$$\liminf_{n\to\infty} \frac{1}{n} \log Q_n(\Gamma) \geq -\inf_{x\in\Gamma^0} I(x) \tag{2.5}$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \log Q_n(\Gamma) \leq -\inf_{x\in\bar{\Gamma}} I(x). \tag{2.6}$$

We call the right hand side of (2.5) and (2.6) the large deviation lower and upper bound, respectively.

Remark 1 *Sometimes we say that a sequence of random variables satisfy the LDP if the corresponding sequence of distributions does.*

Remark 2 *From the definition it is straightforward that if*

$$\inf_{x\in\Gamma^0} I(x) = \inf_{x\in\bar{\Gamma}} I(x) = I_\Gamma, \tag{2.7}$$

then

$$\lim_{n\to\infty} \frac{1}{n} \log Q_n(\Gamma) = -I_\Gamma. \tag{2.8}$$

A set Γ satisfying (2.7) is called an I continuity set. The LDP implies a precise limit in (2.8) only for I continuity sets.

Remark 3 *The LDP is equivalent to stating that for any open set $A \subset \Sigma$, and any closed set $B \subset \Sigma$,*

$$-\inf_{x\in A} I(x) \leq \liminf_{n\to\infty} \frac{1}{n} \log Q_n(\Gamma) \leq \limsup_{n\to\infty} \frac{1}{n} \log Q_n(\Gamma) \leq -\inf_{x\in B} I(x). \tag{2.9}$$

Note that the upper bound trivially holds when $\inf_{x\in\bar{\Gamma}} I(x) = 0$, and the lower bound trivially holds when $\inf_{x\in\Gamma^0} I(x) = \infty$. And in practice, in proving the large deviation upper bound, we often prove it first for compact sets. So we have:

Definition 3 (Weak LDP) *A sequence of probability measures $\{Q_n, n \geq 1\}$ is said to satisfy the weak large deviation principle with rate function $I(x)$ if the upper bound in (2.2) holds only for compact sets $\Gamma \in \mathcal{B}$. Accordingly, the LDP in Definition 2 is referred as the full LDP.*

To strengthen the weak LDP to a full LDP we need to show that most of the probability mass is concentrated on compact sets. This motivates the following:

Definition 4 *A family of probability measures $\{Q_n\}$ on Σ is exponentially tight if for every $\alpha < \infty$, there exists a compact set $K_\alpha \subset \Sigma$ such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log Q_n \left(K_\alpha^c\right) < -\alpha. \tag{2.10}$$

If $\{Q_n\}$ is exponentially tight, then the large deviation upper bound for all compact sets implies the bound for all closed sets. This result is useful because it is often easier to prove upper bounds for compact sets by covering them by a finite class of sets, such as balls and half-spaces. Indeed we have the following theorem:

Theorem 1 *If an exponentially tight family of probability measures $\{Q_n, n \geq 1\}$ satisfies the weak LDP with a rate function $I(\cdot)$, then $\{Q_n, n \geq 1\}$ satisfies the (full) LDP with good rate function $I(\cdot)$.*

Proof. See Lemma 1.2.18 of Dembo and Zeitouni (1998). ■

The following result states that there is at most one rate function governing the large deviation of $\{Q_n, n \geq 1\}$.

Proposition 2 *Suppose $\{Q_n, n \geq 1\}$ satisfies the LDP with two rate functions, namely $I_1(x)$ and $I_2(x)$, then $I_1(x) = I_2(x)$ a.s.*

Proof. Let $B(\delta, r)$ denote a ball centered at δ with radius $r > 0$. If the non-increasing function $r \in (0, \infty) \longmapsto \inf_{B(\delta,r)} I_j(x)$, where $x \in B(\delta, r)$, and $j = 1, 2$, is continuous at r, we have $\inf_{B(\delta,r)} I_j(x) = \inf_{\bar{B}(\delta,r)} I_j(x)$ except for some countable number of r. Therefore, $B(\delta, r)$ is a I continuity set, and consequently

$$-\lim_{n\to\infty} \frac{1}{n} \log Q_n \left(B(\delta, r)\right) = \inf_{B(\delta,r)} I_1(x) = \inf_{B(\delta,r)} I_2(x) \tag{2.11}$$

for every r. Since a rate function is lower semicontinuous, we have $\lim_{r\to 0} \inf_{B(\delta,r)} I_j(x) = I(\delta)$ for all $\delta \in \Sigma$. Combined with (2.11), we have $I_1(x) = I_2(x)$ *a.s.* See e.g., Deuschel and Strook (1989). ■

2.2.2 Transformation of LDP

Given a large deviation principle on one space, it is often of interest to be able to construct a large deviation principle on another space. There are several results in this area:

Theorem 2 (Contraction Principle) *Let $\Sigma^{'}$ be a complete and separable metric space and $f : \Sigma \rightarrow \Sigma^{'}$ be a continuous function. If $\{Q_n\}$ obeys the LDP with rate function I, then the image-measure $\{Q_n \circ f^{-1}\}$ obeys the LDP on $X^{'}$ with rate function $I^{'}$, where $I^{'}(x^{'}) = I \circ f^{-1}(x^{'}) = \inf_{x \in f^{-1}(x^{'})} I(x)$.*

Proof. Let Γ be a closed subset of $\Sigma^{'}$. Since f is continuous, $f^{-1}(\Gamma)$ is a closed subset of Σ and hence $\Gamma = \bar{\Gamma}$. Therefore by the upper bound of the LDP for $\{Q_n\}$,

$$\begin{aligned}
\limsup_{n\rightarrow\infty} \frac{1}{n} \log\left(Q_n \circ f^{-1}\right)(\Gamma) &= \limsup_{n\rightarrow\infty} \frac{1}{n} \log Q_n\left(f^{-1}(\Gamma)\right) \\
&\leq -\inf_{x \in f^{-1}(\Gamma)} I(x) \\
&= -\inf_{x^{'}} \left\{ \inf \left\{ I(x) : f(x) = x^{'}, x \in \Sigma \right\} \right\} \\
&= -\inf_{x^{'}} I^{'}(x^{'}).
\end{aligned}$$

The lower bound can be proved if Γ is an open set by similar argument. See also Dembo and Zeitouni (1998). ■

The next theorem shows that if two random variables are exponentially close to each other, then they share the same LDP property.

Definition 5 *Two families of random variables X_n, Y_n, which both take values in Σ, are exponentially equivalent if for all $\varsigma > 0$,*

$$\lim_{n\rightarrow\infty} \frac{1}{n} \log P\left(d\left(X_n, Y_n\right) > \varsigma\right) = -\infty \tag{2.12}$$

where P is a probability measure on Σ, and $d(\cdot,\cdot)$ is a distance function defined on Σ.

Theorem 3 *If two families of random variables X_n, Y_n are exponentially equivalent, then one of them satisfies the LDP with good rate function $I(x)$ if and only if the other does as well.*

Proof. It suffices to show that the LDP for X_n implies the LDP for Y_n. Suppose that X_n satisfies the LDP with rate function $I(x)$. For any closed set $A \in \mathcal{A}$, let its closed δ neighborhood denoted by $A^{\delta} = \{x : \exists y \in A, d(x, y) \leq \delta\}$, then we have

$$P(Y_n \in A^{\delta}) \leq P\left(X_n \in A^{\delta}\right) + P\left(d(X_n, Y_n) > \delta\right) \tag{2.13}$$

using (2.6) and the LDP for X_n,

$$\begin{aligned}\limsup \frac{1}{n}\log P(Y_n &\in A) \\ &\le \limsup \left\{\frac{1}{n}\log P(X_n \in A^\delta) + \frac{1}{n}\log P\left(d(X_n, Y_n) > \delta\right)\right\} \\ &\le \limsup \frac{1}{n}\log P(X_n \in A^\delta) \vee \limsup \frac{1}{n}\log P\left(d(X_n, Y_n) > \delta\right) \\ &\le -I(A^\delta) \vee -\infty \\ &\le -I(A^\delta)\end{aligned}$$

Since I is a good rate function, $I(A^\delta) \uparrow I(A)$ as $\delta \downarrow 0$. And since δ is arbitrary, we have the upper bound:

$$\limsup \frac{1}{n}\log P(Y_n \in A) \le -I(A)$$

The lower bound can be proved in a similar way by considering an open set in (2.13) .See also Puhalskii (2006). ■

2.3 Sanov Theorem

Now we focus on the empirical measure of a sequence of random variables $\{x_i\}_{i=1}^n$ which is defined as:

$$\mu_n\left(A\right) = \frac{1}{n}\sum_{i=1}^{n} I_A(x_i) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}\left(A\right)$$

for all $A \subseteq \Sigma$, where $I_A(\cdot)$ is the indicator function for the set A, and δ_{x_i} denotes the probability mass at x_i. Since μ_n is again a random variable, it is interesting to research into the large deviation property of the sequence of empirical measures induced by increasing sample size, which plays a central role in our subsequent two chapters. Note that as mentioned earlier in this chapter, μ_n is a probability measure and $\mu_n \in M_1\left(\Sigma\right)$, the distribution of μ_n, say, P^n is then an element of the set $M_1\left(M_1\left(\Sigma\right)\right)$, and if μ_n satisfies the LDP, the argument of its rate function is also a probability measure. The following definition introduces such a rate function which is very important in large deviation theory.

Definition 6 *For two probability measures Q, $P \in M_1(\Sigma)$, the quantity*

$$H(Q\,|P) = \begin{cases} \int_\Sigma \frac{dQ}{dP} \log \frac{dQ}{dP} dP & \textit{if } Q \ll P \\ 0 & \textit{otherwise} \end{cases} \tag{2.14}$$

is called the relative entropy, or Kullback-Leibler distance between Q and P, where $Q \ll P$ means Q is absolutely continuous with respect to, or dominated by P, i.e., for some set $A \in \mathcal{A}$, $P(A) = 0$ implies $Q(A) = 0$.

The following theorem is about the large deviation property of the empirical measure μ_n of i.i.d. random variables in Σ.

Theorem 4 (Sanov) *Let $\{x_i\}_{i=1}^n$ be a sequence of i.i.d. random variables, and $\mu \in M_1(\Sigma)$ equipped with the τ-topology be the probability law of x_i. Then the sequence $\{P^n, n \geq 1\}$ satisfies the full LDP with the good, convex rate function $H(\cdot\,|\mu)$ defined as (2.14).*

Proof. See section 2.1.1 of Dembo and Zeitouni (1998) for a simple illustration in $\mathbb{R}^n$; see also Theorem 3.2.17 of Deuschel and Strook (1989) for a proof in more general Polish space. ■

Remark 4 *Shikimi (2002) extends this result to the kernel type empirical distribution:*

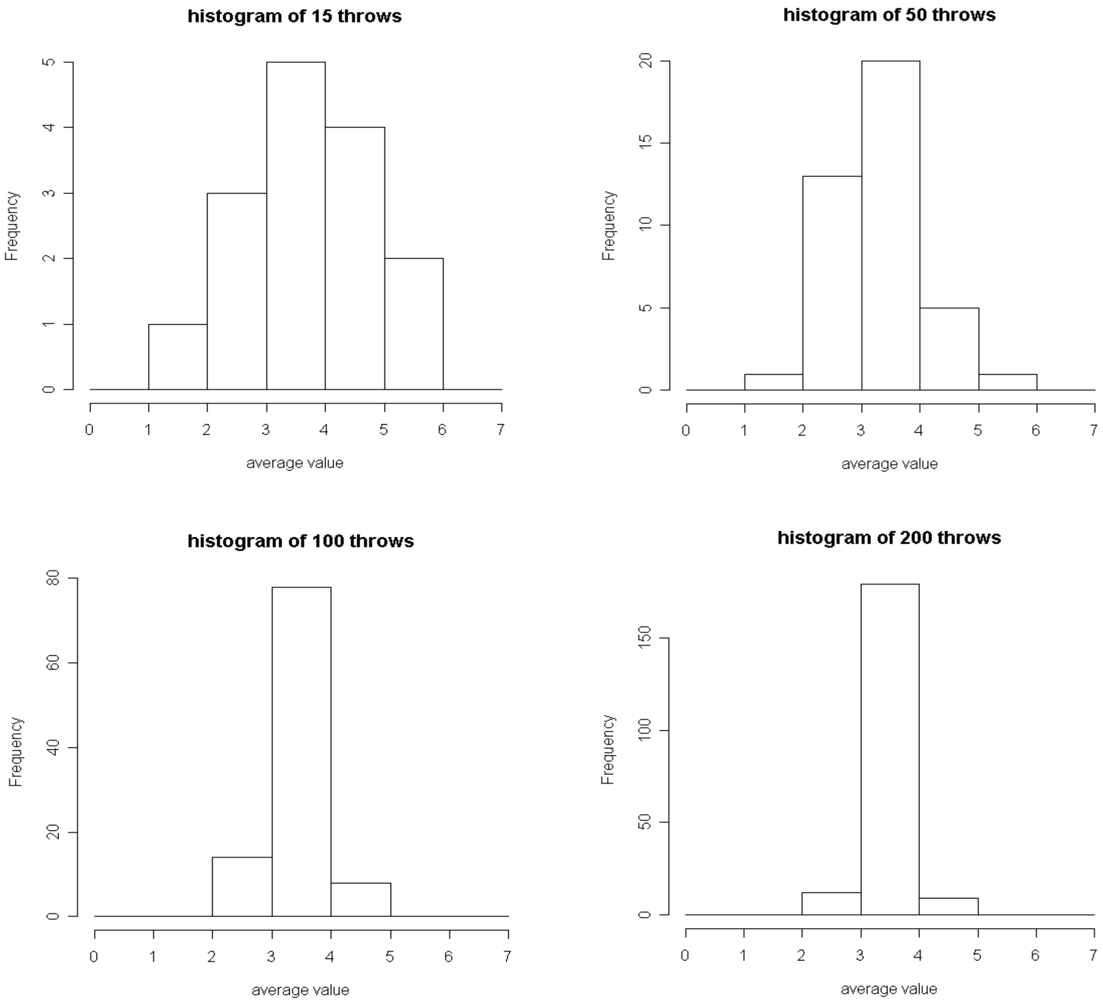
$$\tilde{\mu}_n = \frac{1}{n} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)$$

where $K(\cdot)$ is a kernel and h is the bandwidth.

We will see later on that Sanov theorem provides a very useful tool in analysing problems involving LDP in the simplest i.i.d. case, since the rate function, the Kullback-Leibler distance is convenient to be employed in many situations in statistics. Moreover, when dealing with non i.i.d. data where the Sanov theorem is not applicable, we always want to find some kind of analogue of Sanov theorem which characterizes the LDP of the data with a specific rate function. The next chapter presents such a situation and its important application in hypothesis testing.

2.4 Appendix

Figure1. histogram of average value of dice with increasing number of throws



Chapter 3

Asymptotic Optimality of Empirical Likelihood Tests With Weakly Dependent Data

Abstract

In this chapter we extend the result of Kitamura (2001) to stationary mixing data. Kitamura shows that empirical likelihood test of moment conditions is asymptotically optimal in the sense that the type II error of the EL test in i.i.d. context can achieve large deviation lower bound. The key thing in proving the large deviation optimality is that the empirical measure of the i.i.d. data will obey the large deviation principle with rate function equal to the relative entropy. However, in general the large deviation performance of the empirical measure for dependent data is more complicated. In this paper we impose an S-mixing condition (Bryc and Dembo, 1996) to the stationary process, and we show that the rate function derived by Bryc and Dembo is indeed equal to the relative entropy, and then asymptotic optimality follows from the large deviation inequality.

Key Words: weakly dependent, S-mixing, asymptotic relative efficiency.

3.1 Introduction

Literature on empirical likelihood (EL) method has been growing since being introduced by Owen (1988). In the past few years it has been found especially useful in inference in moment condition models as an alternative to the generalized method of moments (GMM). Therefore EL has a broad area of application, since in practice many economic implications are given in terms of moment conditions such as the Euler equation for instance. See the appendices of this chapter (section 3.7) for an introduction of how EL deals with moment condition models in an i.i.d. setting.

In this chapter we show that the asymptotic optimality of the EL test of moment conditions can be extended to the context of dependent data. Our work is an extension of the paper of Kitamura (2001), which proves that the type II error of the EL test achieves the large deviation lower bound in an i.i.d. setting. For a general introduction to EL, see the excellent monograph by Owen (2000). See also recent results by Newey and Smith (2004), which present desirable higher order properties of the EL estimator in finite samples.

There are various approaches to comparing the efficiency of tests with increasing sample size, by checking the asymptotic behaviour of type I and type II error probabilities. These methods include those of Pitman (1949), Chernoff (1952), Hoeffding (1965) and Bahadur (1967), which are briefly reviewed in section 3.4. Considering type I and type II errors as large deviation events, Kitamura (2001) follows Hoeffding's (1965) approach, which is a generalized Neyman-Pearson criterion, since a test will be called Hoeffding-optimal if it has the smallest large deviation type II error among all the tests with the same type I error.

Kitamura (2001) shows that the EL test of moment conditions is optimal in Hoeffding's sense if the observations are i.i.d. The methodology is as follows. Firstly, the EL test amounts to checking if the empirical measure derived from the moment condition, namely μ_n, is close enough to the true probability measure, and hence the rejection region of the EL test can be set by a value of the distance between μ_n and the true measure. On the other hand, Sanov's theorem tells us that μ_n obeys the large deviation principle (LDP) with rate function being the Kullback-Leibler distance, denoted by $H(\cdot|\cdot)$. That is, the probability that μ_n lies in some certain area of the probability space is bounded by $H(\cdot|\cdot)$, and hence a large deviation lower bound of the asymptotic type II error can be established. Therefore, if we take $H(\cdot|\cdot)$ as the distance between two probability measures, the optimality of EL test can be proved. Indeed, this framework is an application of the universal hypothesis problem in information theory, see Zeitouni and Gutman (1991) and section 7.1 of Dembo and Zeitouni (1998).

To extend Kitamura's (2001) result to the non i.i.d case, the first contribution we make in this chapter is to show the equivalence of the rate function of the LDP of $S-$mixing process and the Kullback-Leibler distance $H(\cdot|\cdot)$. It can be seen that $H(\cdot|\cdot)$ plays a critical role in proving the optimality of tests. When the sample is i.i.d, the famous Sanov theorem provides a quite straightforward tool to compare the large deviation probabilities. However, for the dependent case we need to impose some restrictions on the stochastic processes to make the effects of the dependence between observations vanish as the sample size goes to infinity, and to ensure the processes satisfy certain LDP. Mixing, first studied by Rosenblatt (1956), is such a condition that can ensure the processes satisfies some large sample properties such as the weak law of large numbers (WLLN) (e.g., van. der. Varrt (2001)), and central limit theorem (CLT) (e.g., Andrews (1983), Andrews and Pollard (1994)). Various mixing types such as α-mixing, ψ-mixing , $\phi-$mixing and $\beta-$mixing have been studied extensively in time series analysis. See Bradley (2005) for a comprehensive introduction to properties of different mixing conditions.

Also, the LDP has been proved valid for mixing stochastic processes as well. However, the rate function of the LDP for processes with different mixing conditions are not the same and, in general not equal to $H(\cdot|\cdot)$. See, e.g., chapters 5 and 6 of Deuschel and Strook (1989). In this chapter we adopt a certain mixing condition, called S-mixing, introduced by Bryc and Dembo (1996). The advantages of S-mixing are twofold. Firstly, it is a very weak assumption and is implied by $\alpha-$mixing as shown by Bryc and Dembo (1996). Hence the properties of EL test statistics derived by Kitamura (1997) and Smith (2004) under $\alpha-$mixing conditions carry over to $S-$mixing processes as well. Secondly, we find that the rate function of the LDP of $S-$mixing process equals the Kullback-Leibler distance $H(\cdot|\cdot)$, under assumption H-1 to be defined in section 3.2.2, and therefore we can prove the optimality of EL test with $S-$mixing process in a way similar to Dembo and Zeitouni (1998) and Kitamura (2001).

Before studying its optimality, we also review the methods of deriving the EL test statistic from the moment condition model with dependent data. It is not difficult to see that empirical likelihood will fail if the dependence of the data is ignored when we construct the EL estimator and test. See Kitamura (1997) for a simple example. So techniques to handle dependence are needed in empirical likelihood. Smith (2004) employs smoothed moment indicators instead of using the moment conditions directly. Since the smoothed moment indicators satisfy WLLN and CLT, a good asymptotic theory of empirical likelihood can then be developed. On the other hand, considering the similarity of empirical likelihood and the GMM, Kitamura(1997) uses blockwise resampling, which is similar to the GMM dealing with mixing dependent process (Hall and Horowitz (1996)). Reference to blocking techniques in bootstrapping can be found in Politis and Romano (1992).

This chapter is organized as follows. Section 3.2 presents some general results on mixing processes. Here we establish that the rate function of the LDP of an $S-$mixing process equals the Kullback-Leibler distance. In section 3.3 we review some methods to derive EL statistic with mixing data. Some criteria of comparing the relative asymptotic efficiency of tests are reviewed in section 3.4. In section 3.5 we prove the asymptotic optimality of EL test in Hoeffding's sense. Section 3.6 concludes.

3.2 Large Deviation of Weakly Dependent Data

To establish LDP for dependent data we need to put some restrictions on the degree of dependence. The importance of weak dependence conditions in probability theory is that it gives certain requirements under which some limiting properties of dependent processes will imitate their i.i.d. counterparts, such as laws of large numbers, a central limit theorem, and of course, the large deviation principle which we are working with. This section begins by introducing various conditions for weak dependence. With the terminology of dependent data, we will use *time series*, *stochastic process* or *process* interchangeably.

3.2.1 Weak Dependence and Mixing

M-dependence

Throughout $(\Sigma, \mathcal{A}, P)$ will denote a probability space, where Σ is a compact topological space, $\mathcal{A}$ is the associated $\sigma-$field and P is a probability measure. Let $\{X_t : t \in \mathbb{Z}\}$ be a stationary time series taking values in Σ, and $\mathcal{F}_a^b = \sigma\left(X_i : a \leq i \leq b\right)$ denote the $\sigma-$algebra generated by $\{X_i : a \leq i \leq b\}$. Dependence implies that X_{t+s} with $s > 1$ has *memory* from previous values X_{t+s-1}, $X_{t+s-2}, ...$, or in terms of probability theory, two arbitrary $\sigma-$algebras $\mathcal{F}_t^{t+s}$ and $\mathcal{F}_{t+m}^{t+n} \subset \mathcal{A}$, where $m > s, n > m$, are dependent, *i.e.*,

$$P\left(A \cap B\right) - P\left(A\right)P(B) \neq 0 \tag{3.1}$$

for any $A \in \mathcal{F}_t^{t+s}$ and $B \in \mathcal{F}_{t+m}^{t+n}$. This inequality can be considered as a condition for *strong* dependence, since it means that an arbitrary X_t will have memory from all past values. Therefore, if we want to weaken the condition and let X_t be finitely dependent or, weakly dependent, it is natural to require that it only has memory for a certain, say, m periods of time. This idea is generalized in the following definition.

Definition 7 *A time series $\{X_t : t \in \mathbb{Z}\}$ is said to be $m-$dependent* [1] *if the two $\sigma-$algebras $\mathcal{F}_{-\infty}^t$ and $\mathcal{F}_{t+m+1}^{+\infty}$ are independent, i.e.,*

$$P\left(A \cap B\right) - P\left(A\right)P(B) = 0 \tag{3.2}$$

[1] Note that as a special case, $0-$dependent means independent indeed.

for any $A \in \mathcal{F}_{-\infty}^{t}$, $B \in \mathcal{F}_{t+m+1}^{+\infty}$, *and at the same time, for any* $C \in \mathcal{F}_{t+r}^{+\infty}$ *where* $r < m+1$,

$$P(A \cap C) - P(A)P(C) \neq 0$$

Example 3 *The moving average process MA(q):* $X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$, *where* $\{\varepsilon_i\}$ *is a white noise process, is m-dependent with* $m = q$, *since for any t,* X_t *and* X_{t+q+1} *are independent and hence the* $\sigma-$*algebras* $\mathcal{F}_{-\infty}^{t}$ *and* $\mathcal{F}_{t+q+1}^{+\infty}$ *are independent.*

Mixing

Mixing conditions as a measure of weak dependence - a weaker form of (3.2) - were introduced by Rosenblatt (1956) - describing the tendency that two random variables will be approximately independent if they are separated far enough.

Definition 8 (Rosenblatt, 1956) *A strictly stationary time series* $\{X_t : t \in \mathbb{Z}\}$ *is called* $\alpha-$*mixing, or strong mixing, if when* $k \to \infty$, $k \in \mathbb{Z}$,

$$\alpha(k) = \sup_{A \in \mathcal{F}_{-\infty}^{t},\ B \in \mathcal{F}_{t+k}^{\infty}} |P(A \cap B) - P(A)P(B)| \to 0, \tag{3.3}$$

and $\alpha(k)$ *is called the* $\alpha-$*mixing coefficient.*

Note that if the process is $m-$dependent, $\alpha(k) = 0$ for $k > m$, so an $m-$dependent process is trivially strong mixing. Next we introduce some standard properties of $\alpha-$mixing which is useful in the following sections. These results can also be found in van. der. Varrt (2001) for instance.

Proposition 3 $\alpha(k)$ *is decreasing in k with range* $0 \leq \alpha(k) \leq 1/4 = \alpha(0)$.

Proof. The monotonicity comes from its definition since as k increases the $\sigma-$algebras are separated by longer distance. The range is obtained by noting that from Cauchy Schwartz inequality we have

$$\begin{aligned} P(A \cap B) - P(A)P(B) &= Cov(I_A, I_B) \\ &\leq \sqrt{Var(I_A)}\sqrt{Var(I_B)} \\ &= \sqrt{P(A) - [P(A)]^2}\sqrt{P(B) - [P(B)]^2} \\ &\leq \frac{1}{4}. \end{aligned}$$

To see the value of $\alpha(0)$, just let the two sets $A = B = \{X_t\}$, therefore $\alpha(0) = \sup|Var(I_A)| = 1/4$. ■

Proposition 4 *If $\{X_t : t \in \mathbb{Z}\}$ is strong mixing, then it is ergodic.*

Proof. Let D be any $P-$invariant set on Σ, *i.e.*, $I_{TD} = I_D$ where T is some group transformation. Since X_t is strong mixing, we have

$$\lim_{t\to\infty} P(P^{-t}D \cap D) = P(P^{-t}D)P(D) \tag{3.4}$$

where $P^{-t}D$ is the t-times iterated inverse image of D. Note that $P^{-t}D = D$ for any t, we have $\lim_{t\to\infty} P(P^{-t}D \cap D) = P(D)$. So (3.4) becomes $P(D) = [P(D)]^2$, which implies that $P(D) = 1$ or 0. Hence P is ergodic and so is X_t (see, e.g., Walters, 1982, for definition of ergodicity). ■

Proposition 5 *If a process $\{X_t : t \in \mathbb{Z}\}$ is $\alpha-$mixing with coefficient $\alpha(k)$, then the process $\{X_t^l : t \in \mathbb{Z}\}$, where $X_t^l = (X_t,\ X_{t+1},\ \dots\ ,\ X_{t+l-1})$ is an l-block of $\{X_t : t \in \mathbb{Z}\}$, is also $\alpha-$mixing with coefficient $\alpha^l(k) = \alpha(lk)$.*

Proof. The result is straightforward just considering the $\sigma-$algebra $\mathcal{F}_{t+k}^{\infty}$ in (3.3) replaced by $\mathcal{F}_{t+lk}^{\infty}$. ■

Proposition 6 *If $\{X_t : t \in \mathbb{Z}\}$ is $\alpha-$mixing, then for any real valued, monotonic and continuous function $f(\cdot)$, the process $\{f(X_t), t \in \mathbb{Z}\}$ is also $\alpha-$mixing.*

Proof. From the assumptions on $f(\cdot)$, we have $P(fA \cap fB) = P(A \cap B)$ and $P(fA) = P(A)$, $P(fB) = P(B)$, where $fA \equiv \{f(X) : X \in A\}$ and $fB \equiv \{f(X) : X \in B\}$. Hence the result follows. ■

Example 4 *An I.I.D sequence is strong mixing.*

Example 5 *Andrews (1983) shows that the stationary autoregressive process AR(1): $X_t = \theta X_{t-1} + \varepsilon_t$, where $|\theta| < 1$ and ε_t is Gaussian innovation, is strong mixing. Indeed, stationary ARMA(p,q): process $X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$ is strong mixing with Gaussian innovation ε_t, see, e.g., Dedecker et. al. (2007). However, if ε_t is binomial, X_t will not be strong mixing (e.g., Andrews (1984) and van. der. Varrt (2001)).*

Rosenblatt (1956) shows that a stationary $\alpha-$mixing process X_t with zero mean and finite variance satisfies the CLT. Also Durrett (1991) and Andrews and Pollard (1994) show that X_t obeys a functional CLT. Subsequent research has obtained some other mixing conditions which can guarantee a specific sort of CLT for stochastic process, such as $\beta-$mixing by Wolkonski and Rozanov (1959), $\phi-$mixing by Ibragimov (1962), and $\psi-$mixing by Blum *et.al.* (1963), among others. These mixing conditions are defined differently by mixing coefficients. Table I shows some commonly used different mixing coefficients with their ranges. Corresponding mixing processes are defined similarly to $\alpha-$mixing process. For instance, $\{X_t : t \in \mathbb{Z}\}$ is said to be $\phi-$mixing if $\phi(k) \rightarrow 0$ as $k \rightarrow +\infty$. For a complete introduction to various mixing processes, see, e.g., Bradley's (2005) survey.

TABLE 1 Mixing Conditions

Coeff.	Definition[2]	Range
$\alpha(k)$	$\sup \lvert P(A \cap B) - P(A)P(B) \rvert$	$\left[0, \frac{1}{4}\right]$
$\psi(k)$	$\sup \left\lvert \frac{P(A\cap B)-P(A)P(B)}{P(A)P(B)} \right\rvert$	$[0, \infty)$
$\phi(k)$	$\sup \lvert P(B \vert A) - P(B) \rvert$	$[0, 1]$
$\beta(k)$	$\sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \lvert P(A_i \cap B_j) - P(A_i)P(B_j) \rvert$	$[0, 1]$

It may be necessary to clarify some terminologies. When we mention *strong mixing* it refers to $\alpha-$mixing particularly. However, sometimes people would use *strong mixing conditions* (with plural) to call the four types of mixing mentioned above, since the other three conditions are all at least as strong as $\alpha-$mixing. Table 2 presents the well-known chain of implication of these four types of mixing, showing conditions for weak dependence from the strongest $m-$dependence to the weakest $\alpha-$mixing (see, e.g., Bryc and Dembo (1996) and Dedecker *et al. (*2007)).

TABLE 2 Transition of Weak Dependence Conditions

$m-$dependence						
$\Rightarrow$ $\nLeftarrow$						
$\psi-$mixing	$\Rightarrow$ $\nLeftarrow$	$\phi-$mixing	$\Rightarrow$ $\nLeftarrow$	$\beta-$mixing	$\Rightarrow$ $\nLeftarrow$	$\alpha-$mixing

[2] In these definitions the supremums are taken over all $t \in \mathbb{Z}$ and all the possible sets A and B in the $\sigma-$algebras $\mathcal{F}_{-\infty}^{t}$ and $\mathcal{F}_{t+k}^{\infty}$ respectively.

According to table 2, propositions 4-6 for the weakest $\alpha-$mixing process mentioned above are also valid for $m-$dependent, $\psi-$mixing, $\phi-$mixing and $\beta-$mixing processes. For instance, the process of l-block of a $\phi-$mixing process is $\phi-$mixing with coefficient $\phi(lk)$, according to Proposition 5.

It is also worth mentioning that there are other measures of weak dependence beyond mixing conditions. For example, Patrick *et al.* (2002) and Dedecker *et al.* (2007) mention *association* as a description of weak dependence in term of the covariance of functions of separated $\sigma-$algebras.

Definition 9 *The stochastic process $\{X_t : t \in \mathbb{Z}\}$ is associated if for any increasing real valued functions f and g,*

$$Cov\left[f\left(X_t, X_t \in A\right), g\left(X_t, X_t \in B\right)\right] \geq 0$$

where A and B are sets defined as in (3.1)

Also, a stochastic process which is weakly dependent in this sense is not necessarily mixing. For instance, Patrick *et al.* (2002) showed that the Bernoulli shift defined as

$$X_t = F\left(\eta_{t-j} : j \in \mathbb{Z}\right),$$

where $F : \mathbb{R}^{\mathbb{Z}} \to \Sigma$ is a measurable function and $\{\eta_t : t \in \mathbb{Z}\}$ is a sequence of real valued random variable , provides many examples of stochastic processes which are weakly dependent but not mixing.

3.2.2 S-Mixing and Large Deviations

We shall be interested in large deviation properties of the empirical measure μ_n of weakly dependent data. As many of the properties of I.I.D. processes, such as the CLT, have been proved to hold for strong mixing process, and hence for processes satisfying other strong mixing conditions mentioned above (e.g., see chapter 4 of Billingsley (1999)), it is reasonable to think that there exist some analogues of Sanov theorem for stochastic processes which are weakly dependent. Furthermore, we are particularly interested in the rate function of the LDP of the process, since it provides bounds on error probabilities.

There are some instances in the literature which provide different LD results for various mixing conditions. For example, Bryc (1992) shows that LDP holds for ψ-mixing processes, while the empirical measure of μ_n of a class of Doeblin chains, which are ϕ-mixing, fails to obey the LDP (Baxter *et al.* *(*1991)). Therefore it would be useful to find some mixing condition which can guarantee the LDP for empirical measures of general weakly dependent stochastic process. Bryc and Dembo (1996) found such a mixing condition, and called it S-mixing. The reasons we focus on $S-$mixing processes are twofold. Firstly, $S-$mixing is a fairly weak condition, since the $\alpha-$mixing condition suffices for $S-$mixing, so it is suitable for a quite general class of stochastic processes (see also Dembo and Zeitouni (1998)). Secondly, the rate function for the LDP of an $S-$mixing process, as Theorem 6 below shows, is equivalent to the relative entropy, as in the i.i.d case, which perfectly meets our needs.

Definition 10 (Bryc and Dembo, 1996) *A stationary process $\{X_t : t \in \mathbb{Z}\}$ is said to be $S-$mixing if, for any finite constant $C < \infty$, there exists a non-decreasing sequence $l(n) \in \mathbb{N}$ with*

$$\sum_{n=1}^{\infty} \frac{l(n)}{n(n+1)} < \infty \tag{3.5}$$

such that the $S-$mixing coefficient

$$S(n) \equiv \sup \left| P(A)P(B) - e^{l(n)} P\left(A \cap B\right) \right| \leq e^{-Cn} \tag{3.6}$$

where $A \in \mathcal{F}_0^{k_1}$, $B \in \mathcal{F}_{k_1+l(n)}^{k_1+k_2+l(n)}$, $k_1,\ k_2 \in \mathbb{Z}_+$.

Like other mixing coefficients we mentioned above, $S(n)$ is also a measure of dependence of separated random variables in the sequence, and the relationship of S-mixing and the other four types of mixing in Table 2 is indicated in the following proposition.

Proposition 7 *$\alpha-$mixing implies S-mixing.*

Proof. See proposition 2 of Bryc and Dembo (1996). ■

Hence according to the chain of implication in Table 2, S-mixing is the weakest mixing condition among the five. Also, Bryc and Dembo (1996) prove that $S-$mixing will hold if the process satisfies the following two conditions (H-1) and (H-2), which are sometimes called hypermixing conditions (see also section 5.4 of Deuschel and Strook (1989)).

Definition 11 *For any given integers $r \geqslant k \geqslant 2$, $l \geqslant 0$, a family of functions $\{f_i\}_{i=1}^k \in B(\Sigma, \mathbb{R})$ is called l-separated if there exist k disjoint intervals $J_1, ..., J_r$,* *such that dist* $(J_m, J_{m'}) \geq l$ *for 1*$\leq m < m' \leq r$ *and* f_m *is* J_m *measurable for any 1*$\leq m < n$.

Assumption 1 (H-1) *There exist* l, $\alpha < \infty$ *such that , for all* $k, r < \infty$, *and any l-separated functions* $f_i \in B(\Sigma, \mathbb{R})$,

$$E\left|\prod_{i=1}^{k} f_i\left(X_1, ...X_r\right)\right| \leq \prod_{i=1}^{k}\left(E\left[\left|f_i\left(X_1, ...X_r\right)\right|^{\alpha}\right]\right)^{1/\alpha}. \tag{3.7}$$

Assumption 2 (H-2) *There is some constant* l_0 *and functions.*$\beta(l) \geq 1, \omega(l) \geq 0$ *such that for all* $l > l_0$, *all* $r < \infty$, *and any two* $l-$*separated functions* f, $g \in B(\Sigma, \mathbb{R})$,

$$\begin{aligned} & \left|E\left[f\left(X_1, ...X_r\right)\right] E\left[g\left(X_1, ...X_r\right)\right] - E\left[f\left(X_1, ...X_r\right) g\left(X_1, ...X_r\right)\right]\right| \\ \leq \ & \omega(l)\left(E\left[\left|f\left(X_1, ...X_r\right)\right|^{\beta(l)}\right]\right)^{1/\beta(l)}\left(E\left[\left|g\left(X_1, ...X_r\right)\right|^{\beta(l)}\right]\right)^{1/\beta(l)} \end{aligned}$$

The following theorem is essential throughout this chapter, for with it we are able to evaluate large deviation probabilities of weakly dependent data. For a process $\{X_t : t \in \mathbb{Z}\}$, let Q be the underlying probability measure for the whole process and let Q_n denote the measure for a realization of x_t: $x_1, ..., x_n$ on Σ^n, *i.e.*, Q_n is the $n-$th marginal of Q and particularly, $Q_1 \in M_1(\Sigma)$ is the probability measure of a single realization.

Theorem 5 (Bryc and Dembo, 1996) *If a stationary process* $\{X_t : t \in \mathbb{Z}\}$ *is S-mixing, the empirical measure* μ_n *satisfies the LDP with respect to the* τ-*topology in* $M_1(\Sigma)$ *and this LDP is governed by the good rate function*

$$I(v) = \sup_{f \in B(\Sigma, \mathbb{R})}\left\{\int_{\Sigma} f dv - \Lambda(f)\right\} \tag{3.8}$$

i.e, for every set $\Gamma \subseteq M_1(\Sigma)$,

$$\lim_{n\to\infty}\inf \frac{1}{n} \log P^n\left(\mu_n \in \Gamma\right) \geq -\inf_{v \in \Gamma^0} I(v) \tag{3.9}$$

$$\lim_{n\to\infty}\sup \frac{1}{n} \log P^n\left(\mu_n \in \Gamma\right) \leq -\inf_{v \in \bar{\Gamma}} I(v) \tag{3.10}$$

where $P^n \in M_1(M_1(\Sigma))$ is the distribution of μ_n and

$$\Lambda(f) = \lim_{n\to\infty} \frac{1}{n}\Lambda_n(f) = \lim_{n\to\infty} \frac{1}{n}\log E_Q\left[\exp\left(\sum_{i=1}^{n} f(x_i)\right)\right] \tag{3.11}$$

And the limit exists for every $f \in B(\Sigma, \mathbb{R})$.

Proof. See theorem 1 of Bryc and Dembo (1996) or theorem 6.4.14 of Dembo and Zeitouni (1998). ■

Bryc and Dembo (1996) also point out that the result can also be extended to product measures: if the $S-$mixing condition holds for $\left\{X_t^l : t \in \mathbb{Z}\right\}$, then for each $r \in \mathbb{N}$, the process $\{(X_i, ..., X_{i+r-1})\}_{i=1}^n$ taking values in the product space Σ^r, is also $S-$mixing, according to Proposition 5. Hence the r-fold empirical measure:

$$\mu_{n,r} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i,...,x_{i+r-1}}$$

will also satisfy the LDP in $M_1(\Sigma^r)$ equipped with the τ-topology and with a convex rate function $\Lambda_r(\cdot)$ which is the Fenchel-Legendre transform of

$$\Lambda^{(r)}(f) = \lim_{n\to\infty} n^{-1}\log E\left[\exp\left(\sum_{i=1}^{n} f(x_i, ..., x_{i+r-1})\right)\right]$$

On the rate function of the LDP, Bryc and Dembo (1996) mentioned roughly in their paper that $I(v)$ in general will be less than specific Kullback-Leibler distance, which is different from the i.i.d. case, but they did not provide any proof. However, we find that $I(v)$ will be equal to the Kullback-Leibler distance if the $S-$mixing condition are combined with assumption (H-1) which is part of hypermixing condition. To show this result we firstly introduce the following lemma from Dembo and Zeitouni (1998).

Lemma 1 (Dembo and Zeitouni (1998)) *Given assumption (H-1), we have*

$$\Lambda(f) \leq \frac{1}{\gamma}\log\int_{\Sigma} e^{\gamma f(x)} dQ_1. \tag{3.12}$$

Proof. Since the limit in (3.11) $\Lambda(f)$ exists, we can set $n = ml$, where m, $l \in \mathbb{N}$, then according to Jensen's inequality,

$$\begin{aligned} E_Q\left[\exp\left(\sum_{i=1}^{ml} f(x_i)\right)\right] &= E_Q\left[\exp\left(l^{-1}\sum_{k=1}^{l} l\sum_{j=0}^{m-1} f(x_{k+jl})\right)\right] \\ &\leq l^{-1}\sum_{k=1}^{l} E_Q\left[\exp\left(l\sum_{j=0}^{m-1} f(x_{k+jl})\right)\right] \\ &\leq l^{-1}\sum_{k=1}^{l}\prod_{j=1}^{m-1}\left\{E_Q\left[\exp\left(\alpha l f(x_{k+jl})\right)\right]\right\}^{1/\alpha} \\ &= \left\{E_Q\left[\exp\left(\alpha l f(x_i)\right)\right]\right\}^{m/\alpha} \end{aligned}$$

where the second inequality follows from (3.7) by noting that $f(x_{k+jl})$ with k, $j \in \mathbb{N}$ are $l-$separated. The last equation comes from the stationary of X_t. So we have

$$\frac{1}{ml}\log E_Q\left[\exp\left(\sum_{i=1}^{ml} f(x_i)\right)\right] \leq \frac{1}{l\alpha}\log E_Q\left[\exp\left(alf(x_i)\right)\right]$$

Now let $l\alpha = \gamma$ and we get the result. ■

Assumption 3 *If $v \ll Q_1$, then the density dv/dQ_1 is bounded.*

With these results and conditions, now we can prove our main result which can be applied to prove the asymptotic optimality of the EL test.

Theorem 6 *If assumption (H-1) is satisfied, the rate function $I(v)$ in (3.8) of Theorem 5 satisfies:*

$$I(v) = H(v\,|Q_1)$$

Proof. Firstly we show that $I(v) \geq H(v\,|Q_1)$. From lemma 1 we have

$$\begin{aligned} I(v) &= \sup_{f\in B(\mathcal{X},\mathbb{R})}\left\{\int_\Sigma f dv - \Lambda(f)\right\} \\ &\geqslant \sup_{f\in B(\mathcal{X},\mathbb{R})}\left\{\int_\Sigma f dv - \frac{1}{\gamma}\log\int_\Sigma e^{\gamma f(x)}dQ_1\right\} \\ &\geqslant \int_\Sigma f dv - \frac{1}{\gamma}\log\int_\Sigma e^{\gamma f(x)}dQ_1 \qquad (3.13) \end{aligned}$$

The last inequality implies that v is absolute continuous with respect to Q_1. To see this, let $\Gamma \in \mathcal{A}$ satisfying $Q_1(\Gamma) = 0$. Because the inequality holds for any $f \in B(\mathcal{X}, \mathbb{R})$, we can take $f = \xi I_\Gamma$, where $\xi > 0$ and I_Γ is the indicator function of the set Γ. Note that $\int_\Sigma e^{\gamma \xi I_\Gamma} dQ_1 = 1$, so we have $I(v) \geqslant \xi v(\Gamma)$ for any $\xi > 0$. Since $I(v)$ is non-negative, we conclude $v(\Gamma) = 0$, *i.e.*, $v \ll Q_1$. Therefore the Radon-Nikodym derivative of v with respect to Q_1 exists, namely, $\varphi \equiv \frac{dv}{dQ_i}$. Hence $I(v) \geq H(v\,|Q_1)$ is implied by (3.13) if we take $f = \log \varphi$ with assumption 3.

On the other hand, by Jensen's inequality and the stationarity of $\{x_i\}$, we obtain

$$\begin{aligned} & \lim \frac{1}{n} \log E\left[\exp\left(\sum_{i=1}^n f(x_i)\right)\right] \\ \geq\ & \lim_{n\to\infty} \frac{1}{n} E\left[\log\exp\left(\sum_{i=1}^n f(x_i)\right)\right] \\ =\ & E\left[f(x)\right] \\ \geqslant\ & \int_\Sigma f(x)dv - H(v\,|Q_1) \\ \Rightarrow\ & H(v\,|Q_1) \geq \int_\Sigma f(x)dv - \Lambda(f) \end{aligned}$$

which completes the proof. ■

The importance of this theorem is that it links the rate function of the LDP of S-mixing process with the Kullback-Leibler distance. Therefore, we can apply Theorem 5 to analyse problems related to the LDP (such as test efficiency) of weakly dependent data, just in a similar way that we apply the Sanov theorem in i.i.d case. Before going on to discuss its asymptotic relative efficiency, in the next section we introduce the empirical likelihood test statistic obtained in the weakly dependent case.

3.3 Empirical Likelihood Inference of Weakly Dependent Data

3.3.1 Moment Condition Model

In this section we briefly review current methods of EL tests of moment conditions when the observations are weakly dependent. Let $\{x_i\}_{i=1}^n$ be a realization of a stationary $\alpha-$mixing (and hence ergodic and S-mixing) process $\{X_t : t \in \mathbb{Z}\}$ taking values on Σ. We are interested in applying EL to test the following moment condition:

$$E\left[g\left(x_i, \theta_0\right)\right] = \int_\Sigma g(x_i, \theta_0) dQ_1 = 0 \tag{3.14}$$

where the moment indicator g: $\mathbb{R}^d \times \Theta \to \mathbb{R}^m$ is continuous for all d-dimensional $x_i \in \Sigma$, and Q_i is the unknown distribution of x_i, *i,e.*, Q_i is the marginal of Q at x_i. Also $\theta_0 \in \Theta \in \mathbb{R}^p$ is the true parameter vector. We consider the over-identifying case where $m \geq p$. Furthermore, to avoid identification problem we assume that θ_0 uniquely solves (3.14). For notation, let

$$g_i(\theta) = g(x_i, \theta), \quad \Omega = E\left[g_i(\theta_0) g_i(\theta_0)'\right]$$

$$G = E\left[\frac{\partial g_i(\theta_0)}{\partial \theta}\right], \quad V = \left(G'\Omega^{-1}G\right)^{-1}.$$

When $\{x_i\}_{i=1}^n$ are i.i.d, well established results (e.g., Qin and Lawless (1994), Newey and Smith (2004)) show that, under mild regularity conditions the empirical likelihood test statistic for testing (3.14) is

$$\begin{aligned} W_1 &= \inf_{\theta\in\Theta} \sup_{\lambda\in\mathbb{R}^m} 2\sum_{i=1}^n \log\left(1 + \lambda' g(x_i, \theta)\right) && (3.15) \\ &\xrightarrow{d} \chi^2_{m-p} && (3.16) \end{aligned}$$

where λ is the Lagrangian multiplier vector. The validity of the convergence in distribution in (3.16) depends critically on the i.i.d. assumption on $\{x_i\}_{i=1}^n$, which implies that the weak law of large numbers and central limit theorem hold as:

$$\frac{1}{n}\sum_{i=1}^n g(x_i, \theta_0) \xrightarrow{p} E\left[g\left(x_i, \theta_0\right)\right] \tag{3.17}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i,\theta_0) \xrightarrow{d} N(0,\Omega) \tag{3.18}$$

However, as Kitamura (1997) shows, the convergence results with dependent processes will be different from (3.17) and (3.18), so the test statistic W_1 constructed by ignoring the dependence is not valid. A usual remedy to this problem is to remove (at least asymptotically) the dependent structure and make the data satisfy certain WLLN and CLT. In the following we introduce two techniques to remove the dependence, used respectively by Smith (2004) and Kitamura (1997), who both assume $\{X_t : t \in \mathbb{Z}\}$ is strong mixing.

3.3.2 Kernel Smoothing EL

Instead of using the moment indicator g directly, Smith (2004) suggests constructing a kernel smoothed moment indicator as

$$\tilde{g}_i(\theta) = \frac{1}{S_N}\sum_{s=i-N}^{i-1} k\left(\frac{s}{S_N}\right) g_{i-s}(x_i,\theta) \tag{3.19}$$

where S_N is a bandwidth and $k(\cdot)$ is a kernel. This method of kernel smoothing is similar to that used in heteroskedastic and autocorrelation consistent (HAC) covariance matrix estimation, see Andrews (1991). Smith (2004) put some restrictions on S_N and $k(\cdot)$ so that the EL estimator and test statistic derived from $\tilde{g}_i(\theta)$ can achieve desired asymptotic properties (see assumption 2.2 of Smith (2004)). Smith shows that $\tilde{g}_i(\theta)$ satisfies a uniform weak law of large numbers (UWLLN) and a central limit theorem:

$$\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\theta) - k_1 E[g(\theta)]\right\| = o_p(1) \tag{3.20}$$

$$\frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\theta) - E\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_i(\theta)\right]\right) \xrightarrow{d} N\left(0,\Omega(\theta)k_1^2\right)$$

where $k_i = \int_{-\infty}^{\infty} k(t)^i\, dt$ and $\Omega(\theta) = \lim_{n\to\infty} Var\left[n^{-1/2}\sum_{i=1}^{n} g(\theta)\right]$. Also, Smith's version of EL statistic for testing (3.14) is

$$W_2 = 2\frac{nk_1^2}{S_n k_2}\mathcal{R}\left(\hat{\theta},\hat{\lambda}\right) \tag{3.21}$$

where $\hat{\theta}$ and $\hat{\lambda}$ are the solutions to $\inf_{\theta\in\Theta}\sup_{\lambda\in\mathbb{R}^m}\mathcal{R}(\theta,\lambda)$ and

$$\mathcal{R}(\theta,\lambda)=\sum_{i=1}^{n}\log(1+\frac{k_1}{k_2}\lambda' g(x_i,\theta)) \tag{3.22}$$

Theorem 7 (Smith (2004)) *Under the following conditions:*

1) Θ is compact and $\theta_0\in int(\Theta)$.

2) For sufficiently small $\delta>0$ and $\eta>0$, $E\left[\sup_{\theta^*\in\Gamma(\theta,\delta)}\|g(x_i,\theta^*)\|^{2(1+\eta)}\right]<\infty$, for all $\theta\in\Theta$.

3) If a sequence $\{\theta_j\}_{j=1}^{\infty}$ converges to some $\theta\in\Theta$, then $g(x,\theta_j)\to g(x,\theta_j)$, a.s.

4) $Var\left(n^{-1/2}\sum_{i=1}^{n}g(x_i,\theta_0)\right)\to\Omega>0$

5) $E\left(\partial g(x,\theta_0)/\partial\theta'\right)$ is of full column rank.

We have

$$W_2\xrightarrow{d}\chi^2_{m-p}.$$

Proof. See theorem 4.1 of Smith (2004). ■

3.3.3 Blockwise EL

Kitamura (1997) uses block technique in EL which is widely applied in bootstrapping of time series (see, e.g., Hall and Horowitz (1996) as a classic example of blocking and boostrapping in GMM). The idea of blocking mixing stochastic processes comes from the intuition of strong mixing which implies that the dependence of the random variables will asymptotically vanish if they are separated far enough. Specifically, to make inference based on the moment condition (3.14), instead of using the observations $\{x_i\}_{i=1}^{n}$ directly, we firstly block them to create a new sequence of data. Let $M>1$ denote the block length and L be the separation between block starting points, then the i-th block of M consecutive data can be written as:

$$B_i=\left(x_{(i-1)L+1},...,x_{(i-1)L+M}\right),\quad i=1,...,T$$

where

$$T=\left[\frac{n-M}{L}+1\right].$$

Here $[\cdot]$ denotes the integer part of $\cdot$. Thus we separate the original sequence of observations into T blocks, and from proposition 5 we know that the new sequence of the T blocks is still strong mixing. Note that reasonable choices of L can be between 1 and M inclusive. When $L=M$, the T blocks do not overlap, and B_i's are asymptotically independent as $n \to \infty$, $M \to \infty$ (see, also Owen (2000)). However, some observations will be omitted if $L > M$. and if $L \doteq \alpha M$ with $\alpha < 1$, the dependencies do not become negligible, because there is a fixed fraction of overlap.

Thus we can construct a new moment indicator from the blocks as:

$$b(B_i, \theta) = \frac{1}{M} \sum_{j=1}^{M} g\left(x_{(i-1)L+j}, \theta\right)$$

Obviously $E\left[g(x_i, \theta_0)\right] = 0$ implies $E\left[b(B_i, \theta_0)\right] = 0$. Therefore the corresponding EL test statistic will be (Kitamura (1997) and Owen (2001)):

$$W_3 = \inf_{\theta \in \Theta} \sup_{\lambda \in \mathbb{R}^m} 2\left(\frac{n}{TM}\right) \sum_{i=1}^{T} \log(1 + \lambda' b(B_i, \theta)) \tag{3.23}$$

where the factor (n/TM) is needed to obtain the asymptotic chi-squared property of W_2, and it accounts for the effects of data overlapping between consecutive blocks.

Theorem 8 (Kitamura (1997)) *Under the same assumptions of theorem 7, as $n \to \infty$,*

$$W_3 \xrightarrow{d} \chi^2_{m-p}.$$

Proof. See theorem 3 of Kitamura (1997). ■

3.4 Asymptotic Relative Efficiency of Tests

Before we show the asymptotic optimality of EL test considered previously, in this section we review some criteria for comparing two tests asymptotically. Suppose we want to decide whether the random variable $\{x_i\}_{i=1}^n$ in the compact topological space Σ is from probability distribution Q_1 or alternatively from Q_2. A test T_A is conducted through a sequence of partitions $\Lambda(n) = (\Lambda_1(n), \Lambda_2(n))$ of Σ, with $\Lambda_2(n)$ being the critical region, and $\Lambda_1(n) \cap \Lambda_2(n) = \varnothing$, $\Lambda_1(n) \cup \Lambda_2(n) = \Sigma$. Often the partition is decided by a threshold value of the statistic T_A. Note that T_A is also a sequence which depends on the sample size n.

Let x denote the vector of observations and define

$$\alpha_n = Q_1(x \in \Lambda_2(n)), \qquad \beta_n = Q_2(x \in \Lambda_1(n))$$

where α_n and β_n are the type I and type II error respectively. Also α is called the *size* of the test, and $1-\beta_n$ is the *power*. Generally to improve the asymptotic performance of the test when n increase, we try to minimize β_n while holding α_n fixed at a low level, with requirement that $1 - \beta_n \geq \alpha_n$ which implies that the test is *unbiased*.

Consider another sequence of tests T_B with partitions $\bar{\Lambda}(n) = \big(\bar{\Lambda}_1(n), \bar{\Lambda}_2(n)\big)$. Pitman (1949) introduces the concept of asymptotic relative efficiency (ARE) to compare the quality of these test sequences. We will review Pitman's ARE criteria and several alternative approaches. These methods of comparison differ in the manner in which the Type I and Type II error probabilities vary with increasing sample size n, which is summarized in Table 3. In practice they are chosen according to both intuitive aspects and mathematical consideration to obtain the relative efficiency criterion.

TABLE 3

Type of ARE	Asy. Behavior of α_n	Asy. Behavior of β_n	Behavior of H_1
Pitman	$\alpha_n \to \alpha > 0$	$\beta_n \to \beta > 0$	$H_1 \to H_0$
Bahadur	$\alpha_n \to 0$	$\beta_n \to \beta > 0$	H_1 fixed
Chernoff	$\alpha_n \to 0$	$\beta_n \to 0$	H_1 fixed
Hoeffding	$\alpha_n \to 0$	$\beta_n \to 0$	H_1 fixed

3.4.1 The Pitman Criterion

The intuition of Pitman's approach is that two test sequences T_A and T_B are compared as α_n and β_n tend to positive limits α and β respectively. The asymptotic relative efficiency of T_A to T_B is defined as

$$ARE_{\left(T_A,T_B\right)} = \lim_{n\to\infty}\left(\frac{n_1}{n_2}\right)$$

where n_1 and n_2 are sample sizes such that T_A and T_B have the same power β, and the limit is taken as both n_1 and n_2 tend to infinity. So if $ARE_{\left(T_A,T_B\right)} < 1$, the test T_A is asymptotically more efficient than T_B, and vice versa.

It is important to notice that although the ratio n_1/n_2 will depend on the specific alternative generally, in asymptotic case this situation may be avoided. Indeed, the power with respect to a fixed alternative will be 1 with sufficiently large sample size. Consequently the asymptotic power will no longer provide a good criterion for ARE. So as Noether (1955) points out in Pitman's approach β_n should be evaluated at an alternative which converge to the null hypothesis. Specifically, if the distribution of x can be characterized by some parameters $\theta \in \Theta$ and the null hypothesis $Q_1 = Q\left(x;\theta_0\right)$, then we want to test

$$H_0 : \theta = \theta_0 \qquad \text{against} \qquad H_1 : \theta > \theta_0$$

where $\theta = \theta_n$ and $\lim_{n\to\infty}\theta_n = \theta_0$ at some reasonable rate. For two tests T_A and T_B, the ARE can be determined by the following theorem introduced by Pitman (1949) and Neother (1955).

Theorem 9 *Assume that :*

(i) There exist a function $\sigma_n\left(\theta\right)$ which is k times differentiable, a function $\rho_n\left(\theta\right)$, and a continuous, strictly increasing distribution function G, such that for the test T_A, the quantity $\left(T_A - \sigma_n\left(\theta\right)\right)/\rho_n\left(\theta\right)$ uniformly converges to G in $\left[\theta_0, \theta_0 + \delta\right]$ where $\delta > 0$.

(ii) $\sigma_n^{(1)}\left(\theta_0\right) = \sigma_n^{(2)}\left(\theta_0\right) = ... = \sigma_n^{(k-1)}\left(\theta_0\right) = 0 < \sigma_n^{(k)}\left(\theta_0\right)$ and

$$\rho_n\left(\theta\right) = c_A\frac{\sigma_n^{(k)}\left(\theta_0\right)}{n^q}$$

for some $q > 0$ and some constant c_A. Then

$$c_A = \lim_{n\to\infty} \frac{n^q \rho_n (\theta_0)}{\sigma_n^{(k)} (\theta_0)} \tag{3.24}$$

and for another test T_B which also satisfies (i) and (ii), we have

$$ARE_{\left(T_A, T_B\right)} = \left(\frac{c_B}{c_A}\right)^{1/q} \tag{3.25}$$

where c_B is obtained in the same way as c_A.

Proof. See Serfling (1980) or Rothe (1981). ■

Example 6 *Assume we have an i.i.d. sample $x_1, ..., x_n$ from a normal distribution $N(\mu, \sigma^2)$ with $\sigma^2 < \infty$. If we want to test*

$$H_0 : \theta = 0 \qquad \textit{versus} \qquad H_1 : \theta > 0.$$

We consider the following two test statistics:

$$T_A = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad T_B = \frac{\bar{x}}{S},$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$ is the sample variance.

For T_A, comparing to the conditions in the above theorem, we have $\sigma_n(\theta) = \theta$ with $k = 1$, $\rho_n(\theta) = \sqrt{\sigma^2/n}$, $q = \frac{1}{2}$ and G being the standard normal distribution. So $c_A = \lim\limits_{n\to\infty} \frac{n^{1/2}\sqrt{\sigma^2/n}}{1} = \sigma$. Likewise for T_B, the $t-$statistic, we can take $\sigma_n(\theta) = \theta/\sigma$ with $k = 1$, $\rho_n(\theta) = 1/n$, $q = 1/2$ and G being the standard normal distribution as well. So $c_B = \lim\limits_{n\to\infty} \frac{n^{1/2} 1/n}{1/\sigma} = \sigma$. Hence we have

$$ARE_{\left(T_A, T_B\right)} = \left(\frac{c_B}{c_A}\right)^{1/q} = 1$$

which implies that in this hypothesis testing problem the mean statistic and $t-$statistic are asymptotically equivalent in Pitman's sense.

3.4.2 Bahadur's Approach

The idea behind Bahadur's (1967) approach is as follows. Supposing that the alternative is true, a better test statistic should be the one which is more likely to reject the null hypothesis, or equivalently speaking, it should provide more evidence against the null. Consider the following hypothesis testing problem:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1 \tag{3.26}$$

where $\Theta_1 = \Theta \backslash \Theta_0$. Suppose a test rejects H_0 if $T_n > c$, where T_n is a test statistic and c is a constant. Define the $p-$value of T_n as

$$H_n(t) = \sup_{\theta \in \Theta_0} Q_{\theta_0}(T_n \geq t) = \sup_{\theta \in \Theta_0} [1 - F_{\theta_0}(t)]$$

where Q_{θ_0} is the probability measure characterized by θ_0 and F_{θ_0} is the distribution of the test statistic under θ_0. So $H_n(t)$, also called the 'level attained' by Bahadur, is the maximum probability that the test will produce a test statistic exceeding t, under all possible null hypothesis. Thus it represents a measure to which the test statistic tend to reject H_0. Specifically, the smaller the $H_n(t)$ is, the more likely the test will reject the null hypotheses. Bahadur (1967) suggests that for two sequences of tests T_A and T_B, if the alternative is true, T_A is more efficient than T_B if

$$H_{nA}(t) < H_{nB}(t)$$

or equivalently, $H_{nA}(t)$ goes to 0 at an exponential rate faster than $H_{nB}(t)$. Indeed, a test sequence T_n is said to have *slope* $c(\theta)$ if

$$-\frac{2}{n} \log H_n(t) \to c(\theta) \qquad a.s.$$

So in the nonnull case T_n tends to reject H_0 faster with larger $c(\theta)$, and T_A and T_B can be compared by the ratio (*Bahadur ARE*)

$$ARE_{(T_A, T_B)} = \frac{c_A(\theta)}{c_B(\theta)}.$$

Bahadur (1967) also proves a large deviation theorem which gives the lower bound for the exponential rate that the $p-$value decreases to zero.

Theorem 10 *For any real measurable test statistic T_n for the hypothesis problem* (3.26), *the $p-$value $H_n(t)$ satisfies*

$$\liminf_{n \to \infty} \frac{1}{n} \log H_n(t) \geq - \inf_{\theta \in \Theta_0} H(P_\theta \,| P_{\theta_0}) \qquad a.s.$$

where $H(\cdot\,|\cdot)$ is the Kullback-Leibler distance between Q_{θ_0} and Q_{θ_1}.

Proof. See Bahadur (1967) or Raghavachari (1970). ■

Example 7 (optimality of LR test) *Let $\{f(\cdot,\theta):\theta\in\Theta\}$ be a family of pdf's and $\{x_i\}_{i=1}^{n}$ be a sequence of i.i.d random variables with density $f(\cdot,\theta)$ where θ is unknown and $\theta\in\Theta$. Consider a simple hypothesis testing problem to decide whether $\theta=\theta_0$ or $\theta=\theta_1$, where θ_0, $\theta_1\in\Theta$. Define the likelihood ratio test statistic as*

$$T_n=\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(x_i,\theta_1)}{f(x_i,\theta_0)}$$

and accordingly the $p-$value: $H_n(t)=P_{\theta_0}(T_n\geq t)$. Suppose that $H(P_{\theta_0}|P_{\theta_1}))<\infty$, and that the sequence $\left\{\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(x_i,\theta_1)}{f(x_i,\theta_0)}\right\}$ satisfies the LDP, then when $\theta=\theta_1$ is true,

$$\liminf_{n\to\infty}\frac{1}{n}\log H_n(t)=-\inf_{\theta\in\Theta_0}H(P_\theta\,|P_{\theta_0})\qquad a.s.$$

See, e.g., Hsieh (1979) for a proof of this result. According to the theorem above, likelihood ratio test statistic is optimal in Bahadur's sense since it achieves the lower bound. See also Godambe (1960).

3.4.3 The Chernoff Criterion

Chernoff (1952) introduces a measure of ARE which is particularly useful to such tests and is based on the sum of i.i.d. observations $x_1,...,x_n$, which is a realization of random variable X. Let $S_n=\sum_{i=1}^{n}x_i$ be the test statistic for a simple hypothesis, and when $S_n>c_n$ we reject H_0 where c_n is the critical value. Examples of this kind of test include the mean test in Example 6, and also it is well-known that the LR statistic can be written in this form. Chernoff firstly shows that there is a bound for the probability of $|S_n|$ exceeding some value, as the following theorem shows.

Theorem 11 (Chernoff Bounds) *Suppose the distribution function of X is $P(x)$, with moment generating function $M(t)=E\left(e^{tx}\right)$. Define*

$$m(k)=\inf_{t}E\left[e^{t(x-k)}\right]=\inf_{t}e^{-tk}M(t).$$

If $E(X)>-\infty$ and $k\leq E(X)$, then

$$P(S_n\leq nk)\leq[m(k)]^n,$$

or if $E(X) < \infty$ and $k \geq E(X)$ then

$$P(S_n \geq nk) \leq [m(k)]^n .$$

Proof. The proof is based on Markov inequality. See Chernoff (1952). ■

Now suppose we want to test the null hypothesis that x is from distribution $Q(x) = Q_1$ against the alternative that $Q(x) = Q_2$. The test is to reject Q_1 if $S_n > nk$ for each k. So $\alpha_n = Q_1(S_n > nk)$ and $\beta_n = Q_1(S_n < nk)$. Chernoff (1952) argues that the traditional procedure of minimizing β_n for a fixed value of α_n might not be very appropriate as sample size goes to infinity. Instead he suggests that both type of error probability should decrease to zero as a reward of infinitely increasing sample size, and two tests can be compared with the rate of convergence of α_n and β_n converging to zero.[3] Specifically, we can minimize $\alpha_n + \lambda\beta_n$ for some constant λ, $0 < \lambda < \infty$, and Chernoff shows that if $\mu_1 = E_{P_1}(X) \leq \mu_2 = E_{P_2}(X)$, according to Theorem 11, the rate of exponential convergence of $\alpha_n + \lambda\beta_n$ to zero can be characterized by the *Chernoff index* which is defined as

$$\rho = \inf_{\mu_0 \leq k \leq \mu_1} \rho(k) \tag{3.27}$$

where

$$\rho(k) = \max_{i=1,2} m_i(k) \quad \text{and} \quad m_i(k) = \inf_t E_{P_i}\left[e^{t(x-k)}\right], \quad i = 1, 2.$$

Hence two tests T_A and T_B which are both based on sums of observations and have respective indices ρ_A and ρ_B defined according to (3.27), the asymptotic performance of T_A can be compared with that of T_B through the following ratio

$$ARE_{(T_A,T_B)} = \frac{\log \rho_A}{\log \rho_B}$$

and T_A is more efficient in Chernoff sense than T_B if $ARE_{(T_A,T_B)} > 1$.

Example 8 (optimality of LR test, continued) *Wilbert (1982) shows that for the hypothesis problem (3.26) within exponential family, the Chernoff indices*

[3] This is contrary to Pitman's approach, where two tests are compared when both type of errors are treated in an unbalanced way, *i.e.*, α_n and β_n tend to limits α and β which may be different.

of any test satisfy:

$$\limsup_{n\to\infty} -\frac{1}{n}\log\rho \geq \inf_{\theta\in\Theta_0} H(P_\theta\,|P_{\theta_0})\ ,\ \forall\ \theta\in\Theta_0.$$

And $\limsup_{n\to\infty} -\frac{1}{n}\log\rho^{LR}$ *achieves the lower bound where* ρ^{LR} *is the Chernoff index of the LR test. See also Brown (1971).*

3.4.4 Hoeffding's Approach

Hoeffding (1965) considered a similar method of comparison to Chernoff in the way that tests are compared as both types of error probability asymptotically approach 0 with fixed alternatives. Also like Bahadur and Chernoff, Hoeffding's approach relies on large deviation probabilities. Basically, in Hoeffding's sense a test T_A is asymptotically superior to another test T_B if:

$$\limsup_{n\to\infty}\frac{1}{n}\log\beta_n^A < \limsup_{n\to\infty}\frac{1}{n}\log\beta_n^B \tag{3.28}$$

when both of them satisfy:

$$\limsup_{n\to\infty}\frac{1}{n}\log\alpha_n \leq \eta \tag{3.29}$$

for some $\eta > 0$.

Within a multinomial model Hoeffding (1965) shows that the likelihood ratio (LR) test is asymptotically superior to the chi-squared test in the sense that as (3.28) describes, the exponential rate of the type II error of LR test approaching zero is higher than that of chi-squared test. Indeed, the LR test is optimal among all the test with the same $\frac{1}{n}\log\alpha_n$ because $\frac{1}{n}\log\beta_n^A$ of LR test can achieve the lower bound which is the Kullback-Leibler distance, as Sanov (1965) shows.

Hoeffding's classic work deals with where the sample space is a finite set. Zeitouni and Gutman (1991) extend the optimality to more general infinite spaces. Another of their important contributions is that we can always focus on such tests based only on the empirical measure μ_n.

Theorem 12 *If there is a test T which satisfy* (3.29)*, then there always exists a test A which depends on the observations only through the empirical measure μ_n such that*

$$\liminf_{n\to\infty}\frac{1}{n}\log\alpha_n^A \leq \liminf_{n\to\infty}\frac{1}{n}\log\alpha_n^T$$

and

$$\liminf_{n\to\infty}\frac{1}{n}\log\beta_n^A \leq \liminf_{n\to\infty}\frac{1}{n}\log\beta_n^T.$$

Proof. See lemma 3.5.3 of Dembo and Zeitouni (1998). ■

Based on the theorem, Zeintouni and Gutman (1991) suggest that when studying optimality of tests we can consider tests only depending on the empirical measure μ_n and they introduce the following theorem used by Kitamura (2001).

Theorem 13 *Consider* $(\Lambda_1(n), \Lambda_2(n))$ *is a partition of* $M_1(\Sigma)$ *induced by a test statistic* T*.If* T *is to reject the null hypothesis when*

$$H(\mu_n|Q_1) > \lambda$$

for some constant λ*, then the exponential convergence rate of type I error* α_n *is bounded above by* $-\lambda$*, i.e.,*

$$\limsup_{n\to\infty}\frac{1}{n}\log\alpha_n \leq -\lambda. \tag{3.30}$$

Proof. See Zeintouni and Gutman (1991). ■

(3.30) tells us that the size of the test T based on the empirical measure μ_n through $H(\mu_n|Q_1)$ is always bounded. So if we can show that T minimises the exponential convergence rate of type II error β_n : $\limsup_{n\to\infty}\frac{1}{n}\log\beta_n$, among all the tests satisfying (3.30), then it is optimal in Hoeffding's sense. So this theorem indicates a generalized version of the Neyman-Pearson criterion, which is extended to the large deviation context. This uniform optimality is also called a universal property in information theory, e.g., see Zeitouni and Gutman (1991) and Dembo and Zeitouni (1998). We will use this framework to discuss the optimality of the EL test in the next section.

3.5 Asymptotic Relative Efficiency of the EL Test

In this section, we show the asymptotic relative efficiency of EL test of moment conditions mentioned previously, in Heoffding's sense. The reason we consider this criterion is that the EL test can be considered as to compare the Kullback-Leibler distance between the empirical measure and the hypothetical probability measure. This idea falls into the framework provided by Theorem 13, within which we will show the following argument similar to Dembo and Zeitouni (1998) and Kitamura (2001).

3.5.1 The Test

Following the setup at the beginning of section 3.3, define

$$\mathcal{Q}(\theta) = \left\{ \mu \in M_1(\Sigma) : \int_{\Sigma} g(x_i, \theta_0) d\mu = 0 \right\}. \tag{3.31}$$

Let $\mathcal{Q} = \cup_{\theta \in \Theta} \mathcal{P}(\theta)$, thus $\mathcal{Q}$ is the set of probability measures which satisfy the moment condition over the parameter space. Hence the hypothesis testing problem can be written as:

$$H_0 : \mu_n \in \mathcal{Q} \qquad \text{versus} \qquad H_1 : \mu_n \notin \mathcal{Q} \tag{3.32}$$

where μ_n is the empirical measure of $\{x_i\}_{i=1}^n$. Intuitively, both of the empirical likelihood test statistic W_2 and W_3 which we have obtained in section 3.3 is to check whether the empirical measure μ_n which is constructed to be as close to the true probability measure as possible, is too far away from any of the measures in $\mathcal{Q}$ or not. Therefore, considering the Kullback-Leibler distance as a measure of distance between two probability measures (see appendix for a more detailed discussion of Kullback-Leibler distance), the EL test statistics W_2 and W_3 are indeed a result of the following minimizing problem:

$$\inf_{Q_1 \in \mathcal{Q}} H(\mu_n | Q_1). \tag{3.33}$$

Consequently, the empirical likelihood ratio test is to reject H_0 if:

$$\inf_{Q_1 \in \mathcal{Q}} H(\mu_n | Q_1) > c \tag{3.34}$$

for some threshold constant $c > 0$. This is to say, under the null hypothesis, the empirical measure $\mu_n \in \mathcal{Q}$, and therefore, if the distance between μ_n and any of the probability measures in $\mathcal{Q}$ is too large, then we shall reject the null hypothesis. It also tells that the test depends on the data only through μ_n (see, e.g., section 3.4 of Dembo and Zeitouni (1998)). Thus empirical likelihood test can be considered as a sequence of partitions $\Lambda(n) = (\Lambda_1(n), \Lambda_2(n))$ of $M_1(\Sigma)$ where $n = 1, 2....$ and

$$\Lambda_1(n) = \left\{ \mu \in M_1(\Sigma) : \inf_{Q_1 \in \mathcal{Q}} H(\mu|Q_1) < c \right\}, \quad \Lambda_2(n) = M_1(\Sigma) \backslash \Lambda_1 \qquad (3.35)$$

In the following we abbreviate $(\Lambda_1(n), \Lambda_2(n))$ as (Λ_1, Λ_2) for economy of notation, but its dependence on the sample size n should not be ignored. Since in general framework, pointwise bounds on error probabilities are not available (see Dembo and Zeitouni (1998) or Kitamura, 2001), we consider the δ-smoothing of the set Λ_2 :

$$\Lambda_2^\delta = \bigcup_{\mu \in \Lambda_2} B(\mu, \delta)$$

and

$$\Lambda_1^\delta = M_1(\Sigma) \backslash \Lambda_2^\delta$$

where $B(\mu, \delta)$ denotes an open ball of radius δ around μ, and the balls are taken in the Levy metric:

$$d(\mu_1, \mu_2) = \inf \{ \epsilon > 0 : \mu_1(A) \leq \mu_2(A) + \epsilon \quad \forall A \in \mathcal{A} \}$$

which is compatible with the weak, strong and uniform convergence of discrete probability measures (e.g., see Zeitouni and Gutman (1991)).

3.5.2 Optimality Argument

To directly apply large deviation property of μ_n in Theorem 5 and Theorem 6 to establish the optimality of the EL test, we firstly need some tightness and continuity conditions.

Assumption 4 ***a).*** $\sup_{\theta \in \Theta} \|g(x, \theta)\|$ *is bounded almost surely and thus it is a random variable under all* $Q_1 \in \mathcal{Q}$; ***b).*** *The functional* $\inf_{Q_1 \in \mathcal{Q}} H(\mu|Q_1)$ *is uniformly continuous in* $\mu \in M_1(\Sigma)$ *in the* $\tau-$*topology.*

Lemma 2 $\bar{\Lambda}_1 = \left\{ \mu \in M_1(\Sigma) : \inf_{Q_1 \in \mathcal{Q}} H(\mu|Q_1) \leq c \right\}.$

Proof. The argument is similar to lemma 2 of Zeitouni and Gutman (1991). Since $H(\mu|Q_1)$ is a lower-semicontinuous function, the set $\{\mu \in M_1(\Sigma) : \inf_{Q_1 \in \mathcal{Q}} H(\mu|Q_1) \leqslant c\}$ is closed. So we have:

$$\bar{\Lambda}_1 \subseteq \left\{ \mu \in M_1(\Sigma) : \inf_{Q_1 \in \mathcal{Q}} H(\mu|Q_1) \leqslant c \right\}.$$

To see the other direction, notice that assumption 4-b implies that $\mu \in \{\mu : \inf_{Q_1 \in \mathcal{Q}} H(\mu|Q_1) = c\}$ is a limit point of Λ. Hence the lemma follows. ■

Now we present our main theorem, which gives the optimality of EL test uniformly among all the tests with the same size in large deviation sense. Our result is new in that it extends Kitamura's (2001) result to a non i.i.d case.

Theorem 14 (optimality of EL test) *Let P_i^n with $i = 0, 1$ be the law of the empirical measure under the hypothesis H_0 and H_1 respectively. Then the empirical likelihood test (Λ_1, Λ_2) satisfies*

$$\limsup_{n \to \infty} \frac{1}{n} \log P_1^n \left\{ \mu_n \in \Lambda_2^\delta \right\} \leq -c. \tag{3.36}$$

Moreover, for any another test (Ω_1, Ω_2) which is also a partition of $M_1(\Sigma)$ and satisfies:

$$\limsup_{n \to \infty} \frac{1}{n} \log P_1^n \left\{ \mu_n \in \Omega_2^\delta \right\} \leq -c,$$

we have

$$\limsup_{n \to \infty} \frac{1}{n} \log P_2^n \left\{ \mu_n \in \Omega_1 \right\} \geqslant \limsup_{n \to \infty} \frac{1}{n} \log P_2^n \left\{ \mu_n \in \Lambda_1 \right\}. \tag{3.37}$$

Proof. The inequality (3.36) implies that the type I error of the EL test is bounded above by e^{-nc}. This boundness is straightforward from the LDP of the empirical measure μ_n indicated in theorem 5 and theorem 6 given assumption 4:

$$\limsup_{n \to \infty} \frac{1}{n} \log P_1^n \left\{ \mu_n \in \Lambda_2 \right\} \leq -\inf_{v \in \bar{\Lambda}_2} I(v) = -\underset{v \in \bar{\Lambda}_2}{H} (v|Q_1) \leqslant -c$$

The proof of (3.37) is similar to Zeitouni and Gutman (1991) and Kitamura (2001). Firstly we show that there exists some $n_0 \in \mathbb{N}$, such that $\Lambda_1 \subseteq \Omega(n)$ for all $n > n_0$ along the limit supremum. Suppose it is not so. Then there exists a subsequence n_k such that $\omega_{n_k} \in \bar{\Lambda}_1$ and $\omega_{n_k} \in \Omega_2^\delta$. Since the set $\bar{\Lambda}_1$ is compact due to lemma 2, there exists some $\omega \in \bar{\Lambda}_1$ such that $\omega_{n_k} \to \omega$. Note that $\omega_{n_k} \in \Omega_2^\delta$, thus $B(\omega_{n_k}, \delta) \subset \Omega_2^\delta(k)$ and $B(\omega, \delta/2) \subset \Omega_2^\delta(n_k)$ hold for infinitely many n_k. So,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log P_2^n \{\mu_n \in \Omega_1\} &\geqslant \liminf_{n\to\infty} \frac{1}{n_k} \log P_2^{n_k} \left\{\mu_{n_k} \in \Omega_2^\delta(n_k)\right\} \\
&\geqslant \liminf_{n\to\infty} \frac{1}{n} \log P_2^n \{\mu_n \in B(\omega, \delta/2)\} \\
&\geqslant -\inf_{v \in B(\omega,\delta/2)} I(v) \\
&= -\inf_{v \in B(\omega,\delta/2)} H(v|Q_1) \\
&\geqslant -c,
\end{aligned}$$

this contradicts (3.36) and thus $\Lambda_1^\delta \subseteq \Omega_1^\delta$ is verified. Consequently

$$\limsup_{n\to\infty} \frac{1}{n} \log P_2^n \{\mu_n \in \Omega_1\} \geqslant \limsup_{n\to\infty} \frac{1}{n} \log P_2^n \{\mu_n \in \Lambda_1\}.$$

■

It is worth mentioning that in our model the observations $\{x_i\}_{i=1}^n$ are discrete. As Zeitouni and Gutman (1991) show, to extend the results to continuous case it needs some modifications, mainly because lemma 2 will no longer hold and the smoothing of Λ_2 to Λ_2^δ will not be valid. To overcome these problems, Zeitouni and Gutman (1991) suggests restricting the test, *i.e.*, the partition (Λ_1, Λ_2) of $M_1(\Sigma)$ to be *regular*, namely,

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{n} \log P_2^n \left\{\mu_n \in \Lambda_2^\delta\right\} \geqslant \limsup_{n\to\infty} \frac{1}{n} \log P_2^n \{\mu_n \in \Lambda_2\}.$$

They also point out that this regularity condition is often satisfied.

3.6 Concluding Remarks

In this chapter, we have established asymptotic optimality of the empirical likelihood test with $S-$mixing processes, in Hoeffding's sense. And, as the examples 7 and 8 show, the LR test is optimal in both Bahadur and Chernoff sense, and it is reasonable to guess that the EL test is also Bahadur and Chernoff optimal, considering the similarity of EL and parametric likelihood. More importantly, because the $p-$value of tests in Bahadur's criterion is also bounded by the Kullback-Leibler distance as the case in Hoeffding's approach, which possibly provides us with a starting point from which to consider this issue.

Moreover, we restrict the data to be $S-$mixing and our results rely on the rate function of the LDP of $S-$mixing process. Although there are quite a lot econometric models which adopt $S-$mixing condition (such as ARMA model), so our results have broad application, it will be difficult to discuss the ARE of EL test with more general dependency. As shown in chapters 5 and 6 of Deuschel and Strook (1989) for instance, if a dependent process with other mixing rate (or possibly not mixing as we mentioned at the end of section 3.2.1) satisfies the LDP, it's rate function could be either larger or smaller than the Kullback-Leibler distance, or even extremely complicate to compare. Therefore if we want to establish the asymptotic optimality of EL test in more general circumstances, we might have to define the test by some other quantity which is more related to the specific rate function, other than the Kullback-Leibler distance.

3.7 Appendices

3.7.1 Moment Condition Models and the Empirical Likelihood Methodology

In the seminal paper of EL, Owen (1988) derives an EL confidence interval for the population mean of an i.i.d. sample. Since then the EL method has been extensively studied. Particularly, Qin and Lawless (1994) applies EL to inference of moment condition models, which attracts attention from econometricians. In this appendix we briefly review how EL deals with moment condition models in i.i.d. case, for simplicity.

Suppose that we have a random i.i.d. sample $\{x_i\}_{i=1}^{n}$ which satisfies the following moment condition:

$$E\left[g\left(x_i, \beta_0\right)\right] = 0, \tag{3.38}$$

where g is an $m \times 1$ real function, β_0 is a $p \times 1$ vector ($p < m$) of true parameter, and the expectation is taken with respect to the distribution of x_i. For a simple example, g can be x_i minus the population mean to be estimated, as the case considered by Owen (1988). To estimate β_0, usually a GMM estimator can be derived as

$$\hat{\beta}_{GMM} = \underset{\beta \in \mathcal{B}}{\arg\min}\, \hat{g}\left(\beta\right)' W_n \hat{g}\left(\beta\right), \tag{3.39}$$

where $\mathcal{B}$ is the parameter space, $\hat{g}\left(\beta\right)$ is the sample average of $g\left(x_i, \beta_0\right)$, and W_n is a positive semi-definite matrix and converges in probability to a positive definite matrix matrix W.

As an alternative of GMM, EL assigns a multinomial distribution $F\left(p_1, ..., p_n\right)$ to the i.i.d observation $\{x_i\}_{i=1}^{n}$, with p_i being the probability at x_i. Note that $p_i \geq 0$ and $\Sigma_{i=1}^{n} p_i = 1$. The empirical log-likelihood function is:

$$\log L\left(p_1, ..., p_n\right) = \sum_{i=1}^{n} \log p_i. \tag{3.40}$$

The idea of EL is that we maximize (3.40) subject to the moment restriction $\Sigma_{i=1}^{n} p_i g\left(x_i, \beta\right) = 0$. This can be done by setting up a Lagrangian:

$$\mathcal{L} = \sum_{i=1}^{n} \log p_i + \lambda\left(1 - \Sigma_{i=1}^{n} p_i\right) - n\mu' \sum_{i=1}^{n} p_i g\left(x_i, \beta\right),$$

where λ and μ are Lagrangian multipliers. The solution for p_i is

$$\hat{p}_i = \frac{1}{n\left(1 + \mu' g\left(x_i, \beta\right)\right)}.$$

At the same time, μ can be expressed as a function of β (Qin and Lawless (1994)), namely $\mu\left(\beta\right)$. Therefore, the maximized empirical log-likelihood function with moment restriction is

$$\begin{aligned} l\left(x;\beta\right) &= \sum_{i=1}^{n} \log \hat{p}_i \\ &= -n \log n - \sum_{i=1}^{n} \log\left(1 + \mu\left(\beta\right)' g\left(x_i, \beta\right)\right). \end{aligned}$$

An EL estimator $\hat{\beta}_{EL}$ for β_0 can be obtained by maximizing $l\left(x;\beta\right)$ with respect to β, or equivalently,

$$\hat{\beta}_{EL} = \arg \inf_{\beta\in\mathcal{B}} \max_{\mu\in\mathbb{R}^m} \sum_{i=1}^{n} \log\left(1 + \mu' g\left(x_i, \beta\right)\right). \tag{3.41}$$

Standard asymptotic properties of $\hat{\beta}_{EL}$, *i.e.*, asymptotic normality and consistency, have been proved by Qin and Lawless (1994). Furthermore, they also show that the moment condition model (3.38) can be tested by the following EL statistic:

$$W = \sum_{i=1}^{n} \log\left(1 + \mu(\hat{\beta})'(gx_i, \hat{\beta})\right)$$

which is asymptotically $\mathcal{X}^2_{m-p}$.

From the above procedure we can see that EL is a nonparametric analogue of maximum likelihood method. Without any parametric assumption, EL incorporates the information from the data directly and conveniently, and the EL estimator $\hat{\beta}_{EL}$ has a data driven confidence region (Owen 1988). Newey and Smith (2004) also show that higher order asymptotic properties of $\hat{\beta}_{EL}$, particularly compared to those of the GMM estimator $\hat{\beta}_{GMM}$. They find that $\hat{\beta}_{EL}$ is asymptotically less biased than $\hat{\beta}_{GMM}$ since EL does not need to estimate the weighting matrix W_n in (3.39), which is an important source of bias of $\hat{\beta}_{GMM}$. They also find that after bias correction, $\hat{\beta}_{EL}$ inherits the higher order properties of the maximum likelihood estimator (MLE). See also DiCiccio *et al* (1991) for the Bartlett-correctability of EL.

3.7.2 The Relative Entropy As a Measure of Distance

The Kullback-Leibler distance, or the relative entropy, was introduced by Kullback and Leibler (1951) and Kullback (1958). Now it is widely employed in probability theory and information theory. Let $F_1(x)$ and $F_2(x)$ be two probability measures on the measurable space $(\Sigma, \mathcal{A})$, with density $f_1(x)$ and $f_2(x)$ respectively.

Definition 12 *If $F_1(x) \ll F_2(x)$, i.e., $F_2(A) = 0 \Rightarrow F_1(A) = 0$, for $A \in \mathcal{A}$. Then the Kullback-Leibler distance of $F_1(x)$ and $F_2(x)$ is defined as*

$$H(F_1 | F_2) \equiv \int_\Sigma \log \frac{dF_1}{dF_2} dF_1 = \int_\Sigma f \log f dF_2,$$

where $f(x) = \frac{f_1(x)}{f_2(x)} > 0$.

Proposition 8 *$H(F_1 | F_2) \geq 0$, and the equality holds if and only if $F_1 = F_2$.*

Proof. It is obvious to see that $F_1 = F_2 \Rightarrow \log \frac{dF_1}{dF_2} = 0$. For the inequality, let $h = f \log f$ and expand h at $f = 1$:

$$h = f - 1 + \frac{1}{2}(f-1)^2 h''\left(\dot{f}(x)\right), \tag{3.42}$$

where $\dot{f}(x)$ is between $f(x)$ and 1. Note that $\int_\Sigma f dF_2 = \int_\Sigma f_1(x) dx = 1$, so by integrating both sides of (3.42) with respect to F_2 we obtain:

$$\int_\Sigma f \log f dF_2 = \int_\Sigma (f-1)^2 h''\left(\dot{f}(x)\right) dx \geq 0.$$

The inequality comes from $h''(t) = \frac{1}{t} > 0$ and $f(x) > 0$. ■

It is easy to see that $H(F_1 | F_2)$ is not symmetric and does not satisfy triangle inequality, and consequently is not a real metric, so in this sense $H(F_1 | F_2)$ is more often called *divergence* rather than distance. However, it still can be considered as some sort of measure of distance between probability measures and is especially useful in hypothesis testing problem. Suppose we have a random variable X taking values in Σ with observations $\{x_i\}_{i=1}^n$ and we want to test

$$H_1 : X \text{ is from } F_1(x) \qquad \text{versus} \qquad H_2 : X \text{ is from } F_2(x)$$

Starting from the conditional probability

$$P\left(H_i\left|x\right.\right) = \frac{P\left(H_i\right) f_i(x)}{P\left(H_1\right) f_1(x) + P\left(H_1\right) f_1(x)}, \quad i = 1, 2 \tag{3.43}$$

we have

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P\left(H_1\left|x\right.\right)}{P\left(H_2\left|x\right.\right)} - \log \frac{P\left(H_1\right)}{P\left(H_2\right)}. \tag{3.44}$$

The quantity on the left hand side, *i.e.*, the log-likelihood ratio, can be considered as the information contained in the observations for discrimination in favor of $F_1\left(x\right)$ against $F_2\left(x\right)$. Integrating $\log\left(f_1(x)/f_2(x)\right)$ with respect to $F_1\left(x\right)$ gives $H\left(F_1\left|F_2\right.\right)$, which provides the mean information over the entire sample space by which we can discriminate $F_1\left(x\right)$ and $F_2\left(x\right)$:

$$H\left(F_1\left|F_2\right.\right) = \int_\Sigma \log \frac{f_1(x)}{f_2(x)} dF_1 = \int_\Sigma \log \frac{P\left(H_1\left|x\right.\right)}{P\left(H_2\left|x\right.\right)} dF_1 - \log \frac{P\left(H_1\right)}{P\left(H_2\right)} \tag{3.45}$$

Example 9 (distance between two normal distributions) *Suppose two random variables $X_1 \sim N\left(0, \sigma_1^2\right)$, $X_2 \sim N\left(0, \sigma_2^2\right)$ and we want to test*

$$H_1 : X_1 \textit{ and } X_2 \textit{ are dependent with joint density } f(x_1, x_2)$$

against

$$H_1 : X_1 \textit{ and } X_2 \textit{ are independent with density } f(x_1) \textit{ and } f(x_2) \textit{ respectively.}$$

Since

$$\begin{aligned} f(x_1, x_2) &= (2\pi)^{-1}\left[\sigma_1^2\sigma_2^2\left(1-\rho^2\right)\right]^{-1/2} \\ &\quad \times \exp\left[-\left(2\left(1-\rho^2\right)\right)^{-1}\left(x^2/\sigma_1^2 - \left(2\rho x_1 x_2\right)/\sigma_1\sigma_2\right) + x_2^2/\sigma_2^2\right], \end{aligned}$$

where ρ is the correlation coefficient of X_1 and X_2, (3.45) *can be written as*

$$\begin{aligned} H\left(F_1\left|F_2\right.\right) &= \int_\Sigma \log \frac{P\left(H_1\left|x\right.\right)}{P\left(H_2\left|x\right.\right)} dF_1 - \log \frac{P\left(H_1\right)}{P\left(H_2\right)} \\ &= \int\int f(x_1, x_2) \log \frac{f(x_1, x_2)}{f(x_1) f(x_1)} dx_1 dx_2 \\ &= -\frac{1}{2}\log(1-\rho^2). \end{aligned} \tag{3.46}$$

The quantity in (3.46) *implies that the Kullback-Leibler distance between two* *normal distributions is a function of their correlation coefficient ρ only. If X_1 and X_2 are independent, the Kullback-Leibler distance or divergence is 0, while it will be ∞ if X_1 and X_2 are linearly dependent.*

Chapter 4

Large Deviations of Empirical Likelihood with Nuisance Parameters

Abstract

In this chapter we investigate the asymptotic efficiency of empirical likelihood in the presence of nuisance parameters combined with augmented moment conditions with i.i.d data, via both standard large sample theory and large deviations. We show that in the presence of nuisance parameters, the asymptotic efficiency of the empirical likelihood estimator of the parameter of interest will increase by adding more moment conditions, in the sense of the positive semidefiniteness of the difference of information matrices. As a by-product, we point out a necessary condition for the asymptotic efficiency to be increased when more moment conditions are added. Also, the asymptotic lower bound of the minimax risk function for the parameter of interest is derived.

Key words: empirical likelihood, minimax risks, nuisance parameter

4.1 Introduction

Likelihood inference may have some failings when estimating a parameter of interest in the presence of nuisance parameters. For example, Neyman and Scott (1948) treated this problem and found that maximum likelihood estimation could not be either consistent nor efficient in the presence of many nuisance parameters. As a nonparametric analogue of maximum likelihood, empirical likelihood has proved to have inherited many properties from ordinary parametric likelihood. However, Lazar and Mykland (1999) demonstrated through an Edgeworth expansion that the empirical likelihood ratio in the presence of nuisance parameters can not be corrected to $\mathcal{X}^2$ to the high order that ordinary likelihood achieves and it is no longer the dual likelihood statistic.

This chapter deals with empirical likelihood estimation in the presence of nuisance parameters, combined with selection of moment conditions. We show that in the presence of nuisance parameters, the asymptotic efficiency of the empirical likelihood estimator of the parameter of interest will increase by adding more moment conditions, in the sense of the positive semidefiniteness of the difference of information matrices.

There are quite a lot of examples in the literature which address the problem of inference with many instruments and moment conditions. It is well known that in over-identified models, the asymptotic variance of $\sqrt{n}(\tilde{\beta} - \beta)$ cannot decrease if a moment condition is dropped (Qin and Lawless (1994)). On the other hand, asymptotic properties of GMM estimator based on increasing number of moment conditions have been well established, see e.g., Newey (2001), Newey and Windmeijer (2005). They show that using many moment conditions can improve asymptotic efficiency. Koenker and Machado (1999) proves that in GMM estimation, whenever optimal instruments are not available, it can frequently be shown that adding over-identifying restrictions will increase asymptotic precision. In these cases, it should be noticed that the GMM estimator can be improved by adding more information on the data by augmenting moment conditions as a result of increasing the sample size. In our work, we discuss whether the asymptotic performance of our EL estimator for the parameter in the presence of nuisance parameter can be improved with more information by adding more moment conditions but with sample size fixed. If put it in another way, we want to check if the side effect of nuisance parameter can be counteracted by more moment conditions.

Particularly, we focus on a special case, where nuisance parameters only occur in some of the moment conditions. This case leads to an important result that the asymptotic efficiency can increase with added moment condition only if it is not orthogonal with the original moment conditions.

Furthermore, we investigate large deviation properties of the empirical likelihood inference of moment condition models in the presence of nuisance parameters. Puhalskii and Spokoiny (1998) established a unified framework dealing with statistical problems via large deviations. Within the framework of Puhalskii and Spokoiny (1998), we want to investigate whether the LD efficiency bound for the parameter of interest will remain valid in the presence of nuisance parameters, and then to investigate whether the empirical likelihood estimator and test can achieve the bound, as it does in the case of no nuisance parameters (Kitamura and Otsu (2005)).

The remaining of this chapter is organized as follows. In section 4.2 we derive EL estimator in the presence of nuisance parameters in standard asymptotic theory. We discuss conditions under which the asymptotic efficiency can be improved by more moment conditions. In section 4.3 we analyze the LD risk of EL estimator for the parameter of interest. Section 4.4 concludes.

4.2 Moment Condition with Nuisance Parameters

We present some standard asymptotic results on the EL estimator in the presence of nuisance parameters.

4.2.1 Model Setup and the Estimate of β

Consider a sequence of i.i.d. realizations $\{x_i\}_{i=1}^n$ of a random variable x from an unknown distribution F, with n being the sample size. Let θ be a p-dimensional vector of parameters in a compact parameter space $\Theta \subset \mathbb{R}^p$ associated with F. Suppose that for a true value of θ which is denoted as θ_0, $\{x_i\}_{i=1}^n$ satisfies the following moment condition

$$E\left[g\left(x_i;\theta\right)\right] = 0 \tag{4.1}$$

where g is a $m \times 1$ vector of real functions, and the expectation is taken with respect to F. We consider the over-identified case where $m \geqslant p$. Unlike Qin and Lawless (1994), we don't assume that the m functions of g are independent, since correlation between these functions plays an important role in the aspect of asymptotic efficiency, which we will discuss in the following section.

Now suppose the parameter θ can be decomposed as $\theta = (\beta^{'}, \phi^{'})^{'}$ with corresponding $\theta_0 = (\beta_0^{'}, \phi_0^{'})^{'}$, where $\beta \in \mathcal{B} \subset \mathbb{R}^q$, $\phi \in \Phi \subset \mathbb{R}^{p-q}$ and $\Theta = \mathcal{B}\times\Phi$. If we are only interest in β but not in ϕ, then ϕ is a nuisance parameter in the model, and we write the corresponding moment condition as

$$E\left[g\left(x_i;\beta,\phi\right)\right] = 0 \tag{4.2}$$

for the true value β_0 of β. The empirical likelihood ratio statistic for this model is

$$\mathcal{R}\left(\beta,\phi\right) = 2\sum_{i=1}^{n}\log\left(1+\lambda^{'}g\left(x_i;\beta,\phi\right)\right), \tag{4.3}$$

where λ is an $m \times 1$ vector of Lagrangian multipliers, which is a continuous differentiable function of $(\beta^{'}, \phi^{'})^{'}$ (see, e.g., Qin and Lawless (1994)), and is determined by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{g\left(x_i;\beta,\phi\right)}{1+\lambda^{'}g\left(x_i;\beta,\phi\right)} = 0. \tag{4.4}$$

To simplify notations, let

$$g(x_i;\theta) = g_i(\theta), \quad \hat{g}(\theta) = n^{-1}\Sigma_{i=1}^{n} g_i(\theta)$$

$$G_1 = E\left[\frac{\partial g(x;\theta_0)}{\partial\theta}\right], \quad \Omega_{11} = E\left[g(x;\theta_0)\, g(x;\theta_0)'\right].$$

Like ordinary parametric likelihood, empirical likelihood deals with nuisance parameter by profiling out ϕ (see, e.g., section 3.5 of Owen (2000)). Let $\tilde{\phi} = \tilde{\phi}(\beta)$ be the minimizer of $\mathcal{R}(\beta,\phi)$ with respect to ϕ. The profile log-empirical likelihood ratio for β is

$$\mathcal{R}(\beta) = \min_{\phi\in\Phi}\mathcal{R}(\beta,\phi) \tag{4.5}$$

and EL estimator for β is

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{B}}\mathcal{R}(\beta). \tag{4.6}$$

Assumption 5 *$\theta_0 = (\beta_0, \phi_0)$ solves $E[g(x;\theta)] = 0$ uniquely, or equivalently, both β_0 and ϕ_0 are strongly identified.*

Remark 5 *This condition combined with $m \geqslant p$ makes the parameter well identified. In the paper of Stock and Wright (2000), they considered the problem of weak identification of the parameter, by assuming that the subvector β of θ is completely identified, but ϕ is not, in the sense that the population moment function is steep in β around β_0 but is nearly flat in α. This idea provides us a framework to analysis problems mixed with nuisance parameters, weak identification and partial identification (Phillips (1989)). See also Guggenberger and Smith (2003).*

Assumption 6 ***a).*** *$\theta_0 \in int(\Theta)$;* ***b).*** *Ω_{11} is positive definite and nonsingular;* ***c).*** *$g(x,\theta)$ is twice continuously differentiable in a neighborhood of θ_0 and G_1 is of full rank p.* ***d).*** *$\|g(x,\theta)\|^3$, $\|\partial g(x,\theta)/\partial\theta\|$, and $\|\partial^2 g(x,\theta)/\partial\theta\partial\theta'\|$ are all bounded from above.*

We derive the properties of the EL estimator of β_0 in the next theorem.

Theorem 15 *Under assumption 1-2,*

$$\sqrt{n}\left(\tilde{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, V_\beta^1\right)$$

where

$$V_\beta^1 = \left\{ E\left[\frac{\partial g}{\partial \beta} + \frac{\partial g}{\partial \tilde{\phi}}\frac{\partial \tilde{\phi}}{\partial \beta}\right]' \tilde{\Omega}_{11}^{-1} E\left[\frac{\partial g}{\partial \beta} + \frac{\partial g}{\partial \tilde{\phi}}\frac{\partial \tilde{\phi}}{\partial \beta}\right] \right\}^{-1},$$

$$\tilde{\Omega}_{11} = E\left[g\left(\tilde{\phi}, \beta\right) g\left(\tilde{\phi}, \beta\right)'\right].$$

Proof. The proof is similar to Qin and Lawless (1994). Differentiate $\mathcal{R}(\beta)$ with respect to β and λ respectively gives:

$$\frac{\partial \mathcal{R}(\beta)}{\partial \beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\lambda' g_i\left(\beta, \tilde{\phi}\right)} \left(\frac{\partial g_i\left(\beta, \tilde{\phi}\right)}{\partial \beta} + \frac{\partial g_i\left(\beta, \tilde{\phi}\right)}{\partial \tilde{\phi}} \frac{\partial \tilde{\phi}}{\partial \beta} \right)' \lambda, \tag{4.7}$$

$$\frac{\partial \mathcal{R}(\beta)}{\partial \lambda} = \frac{1}{n}\sum_{i=1}^{n} \frac{g_i\left(\beta, \tilde{\phi}\right)}{1+\lambda' g_i\left(\beta, \tilde{\phi}\right)}. \tag{4.8}$$

Denote the right hand side of (4.7) and (4.8) as $Q_{1n}(\beta, \lambda)$ and $Q_{2n}(\beta, \lambda)$ respectively. Since $\hat{\beta}$ and $\hat{\lambda}$ maximize $\mathcal{R}(\beta)$, $Q_{1n}(\hat{\beta}, \hat{\lambda}) = Q_{2n}(\hat{\beta}, \hat{\lambda}) = 0$, and first order Taylor expansion around $(\beta_0, 0)$ gives:

$$\begin{aligned} 0 &= Q_{1n}(\hat{\beta}, \hat{\lambda}) \\ &= Q_{1n}(\beta_0, 0) + \frac{\partial Q_{1n}(\beta_0, 0)}{\partial \beta}(\tilde{\beta} - \beta_0) + \frac{\partial Q_{1n}(\beta_0, 0)}{\partial \lambda}\hat{\lambda} + o_p(\delta) \end{aligned}$$

$$\begin{aligned} 0 &= Q_{2n}(\hat{\beta}, \hat{\lambda}) \\ &= Q_{2n}(\beta_0, 0) + \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \beta}(\tilde{\beta} - \beta_0) + \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \lambda}\hat{\lambda} + o_p(\delta) \end{aligned}$$

where $\delta = \left\|\tilde{\beta} - \beta_0\right\| + \left\|\hat{\lambda}\right\|$. So $\tilde{\beta}$ and $\hat{\lambda}$ can be solved as:

$$\begin{aligned} \begin{bmatrix} \hat{\lambda} \\ \tilde{\beta} - \beta_0 \end{bmatrix} &= S_n^{-1} \begin{bmatrix} -Q_{1n}(\beta_0, 0) + o_p(\delta) \\ o_p(\delta) \end{bmatrix} \\ &= \begin{bmatrix} \left(I - S_{11}^{-1} E\left(\frac{\partial g}{\partial \beta} + \frac{\partial g}{\partial \tilde{\phi}}\frac{\partial \tilde{\phi}}{\partial \beta}\right) S_{22.1}^{-1} E\left(\frac{\partial g}{\partial \beta} + \frac{\partial g}{\partial \tilde{\phi}}\frac{\partial \tilde{\phi}}{\partial \beta}\right)' \right) S_{11}^{-1} Q_{1n}(\beta_0, 0) + o_p(1) \\ E\left(\frac{\partial g}{\partial \beta} + \frac{\partial g}{\partial \tilde{\phi}}\frac{\partial \tilde{\phi}}{\partial \beta}\right)' S_{11}^{-1} Q_{1n}(\beta_0, 0) + o_p(1) \end{bmatrix}, \end{aligned}$$

where

$$\begin{aligned} S_n &= \begin{bmatrix} \frac{\partial Q_{1n}}{\partial \mu'} & \frac{\partial Q_{1n}}{\partial \beta} \\ \frac{\partial Q_{2n}}{\partial \mu'} & 0 \end{bmatrix}_{(\beta_0,0)} \rightarrow \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & 0 \end{bmatrix} \\ &= \begin{bmatrix} -E(gg') & E\left(\frac{\partial g}{\partial \beta} + \frac{\partial g}{\partial \tilde{\phi}}\frac{\partial \tilde{\phi}}{\partial \beta}\right) \\ E\left(\frac{\partial g}{\partial \beta} + \frac{\partial g}{\partial \tilde{\phi}}\frac{\partial \tilde{\phi}}{\partial \beta}\right)' & 0 \end{bmatrix}. \end{aligned}$$

From lemma 1 of Qin and Lawless (1994) under Assumption 6 we have $\partial Q_{1n}(\beta_0, 0) = (1/n)\sum_{i=1}^{n} g(x_i, \theta) = O_p(n^{-1/2})$ and $\delta = O_p(n^{-1/2})$. So we obtain

$$\begin{aligned} \sqrt{n}\left(\tilde{\beta} - \beta_0\right) &= S_{22.1}^{-1} E\left(\frac{\partial g}{\partial \beta} + \frac{\partial g}{\partial \tilde{\phi}}\frac{\partial \tilde{\phi}}{\partial \beta}\right)' S_{11}^{-1} Q_{1n}(\beta_0, 0) + o_p(1) \\ &\xrightarrow{d} N\left(0, V_\beta^1\right). \end{aligned}$$

$\blacksquare$

Remark 6 ***a).*** *The structure of the asymptotic variance-covariance matrix V_β^1 is different from those in Stock and Wright (2000) and Guggenberger and Smith (2003), in which they decompose $E\left[\hat{g}(\theta)\right]$ as $E\left[\hat{g}(\theta)\right] = n^{-1/2} m_1(\theta) + m_2(\beta)$, where $m_1(\theta)$ involves both of the two parameters and $m_2(\beta)$ involves β and the true value of ϕ.*

b). *Lazar and Mykland (1999) consider higher order properties of $\hat{\beta}$ through Edgeworth expansion of $\mathcal{R}(\beta, \phi)$. They find that $\hat{\beta}$ may not achieve higher order accuracy which can be obtained by ordinary likelihood in the presence of nuisance parameters, also they show that the empirical likelihood ratio statistic does not admit Bartlett correction, unlike the case without nuisance parameters.*

4.2.2 More Moment Conditions

Now we focus on the asymptotic efficiency of $\hat{\beta}$ when there are more moment condition being added. Suppose based on model (4.1), we have the following new model by adding one more moment indicator $f(\cdot)$:

$$E\left[h\left(x_i; \beta_0, \phi_0\right)\right] = E\begin{bmatrix} g\left(x_i; \beta_0, \phi_0\right) \\ f(x_i, \beta_0, \phi_0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (4.9)$$

Similar to section 4.2.1, we define

$$\begin{aligned} G &\equiv E\left[\frac{\partial h\left(x;\theta_0\right)}{\partial\theta}\right], \qquad G_2 \equiv E\left[\frac{\partial f\left(x;\theta_0\right)}{\partial\theta}\right] \\ \Omega &\equiv E\left[h\left(\beta_0,\phi_0\right)h\left(\beta_0,\phi_0\right)^{'}\right] = \begin{bmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{bmatrix}. \end{aligned}$$

In this model, following the setup in the previous section, the parameter vector $\theta = (\beta^{'}, \phi^{'})'$ can be identified by (4.1) alone, and now we are interested in whether the covariance matrix V_β^1 can be improved with extra information given by f. Let the estimator of β based on both g and f denoted as $\tilde{\beta}$, and the corresponding covariance matrix as V_β^2. In general, well established results have shown that at least using f will not be harmful, *i.e.,* it will not increase the asymptotic variance of $\hat{\theta}$. And, nor will dropping f will decrease the asymptotic variance of the estimator, relative to that of the estimator based on both g and f. See, corollary 1 of Qin and Lawless (1994).

Remark 7 *A similar and relevant situation may be worth mention, which is described in Newey and Windmeijer (2005) and Han and Philips (2006), for instance. They assume that the number of moment conditions is increased with the sample size. Thus in this case extra information are provided by both extra data and extra moment conditions, while in our case only by the latter one with fixed sample size n. They also allow the moment conditions are weak, while we assume both g and f are strong as indicated in assumption 5. Estimation under many weak moment conditions is also discussed by Andrews and Stock (2005).*

Proposition 9 *The asymptotic efficiency of EL estimator of β can be increased by adding more moment conditions.*

Proof. Since we can always block the component of the vector of the moment function, for simplicity and without loss of generality, we assume that both g and f are of dimensional one.

For convenience let $E\left(\frac{\partial g}{\partial\beta} + \frac{\partial g}{\partial\tilde{\phi}}\frac{\partial\tilde{\phi}}{\partial\beta}\right) \equiv E_1$, $E\left(\frac{\partial f}{\partial\beta} + \frac{\partial f}{\partial\tilde{\phi}}\frac{\partial\tilde{\phi}}{\partial\beta}\right) \equiv E_2$.

The inverse of V_β^2, or the information matrix of β with both g and f is:

$$\begin{aligned} \mathcal{I}_\beta^2 &= E\left[\frac{\partial h}{\partial\beta}\right]'\left(E\left[hh'\right]\right)^{-1}E\left[\frac{\partial h}{\partial\beta}\right] \\ &= \begin{bmatrix}E_1 & E_2\end{bmatrix}\begin{bmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{bmatrix}^{-1}\begin{bmatrix}E_1\\ E_2\end{bmatrix}. \end{aligned} \tag{4.10}$$

Since without f, the information on β is

$$\begin{aligned}
\mathcal{I}_\beta^1 &= E_1 \left[E\left(gg'\right)\right]^{-1} E_1 \\
&= E_1' S_{11}^{-1} E_1,
\end{aligned}$$

we have:

$$\begin{aligned}
\mathcal{I}_\beta^2 - \mathcal{I}_\beta^1 &= E\left[\frac{\partial h}{\partial \beta}\right]' \left(E\left[hh'\right]\right)^{-1} E\left[\frac{\partial h}{\partial \beta}\right] - E_1' \left(E\left[gg'\right]\right)^{-1} E_1 \\
&= E_1' \left[S_{11}^{-1} S_{12} \Omega_{22} S_{21} S_{11}^{-1}\right] E_1 + E_2 \left[-\Omega_{22} S_{21} S_{11}^{-1}\right] E_1 \\
&\quad + E_1 \left[-S_{11}^{-1} S_{12} \Omega_{22}\right] E_2 + E_2 \Omega_{22} E_2 \\
&= \left(E_1' S_{11}^{-1} S_{12} - E_1\right) \Omega_{22} \left(E_1' S_{11}^{-1} S_{12} - E_2\right)',
\end{aligned}$$

which is positive semidefinite, providing $E\left(gg'\right)$ is p.d as Assumption 6 indicates. ■

Example 10 *Suppose we have a sequence of i.i.d observations of univariate random variable $x_1, ...x_n$. Let $E(x) = \mu$ and $var(x) = \sigma^2$. Thus we have the following two moment conditions:*

$$E\left[g(x;\beta)\right] = E(x-\mu) = 0, \tag{4.11}$$

$$E\left[f(x;\beta,\phi)\right] = E((x-\mu)^2 - \sigma^2) = 0. \tag{4.12}$$

And now we are only interested in the estimation of μ. The empirical likelihood estimator of μ is:

$$\hat{\mu} = \arg\min_\mu \sum_{i=1}^n \log\left(1 + t'\begin{pmatrix} x_i - \mu \\ (x_i-\mu)^2 - \hat{\sigma}^2 \end{pmatrix}\right),$$

and

$$\begin{aligned}
nVar(\hat{\mu}) &= \left(\begin{bmatrix} \frac{\partial g}{\partial \beta} & \frac{\partial f}{\partial \beta} \end{bmatrix} \begin{bmatrix} E(gg) & E(gf) \\ E(fg) & E(ff) \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial g}{\partial \beta} \\ \frac{\partial f}{\partial \beta} \end{bmatrix}\right)^{-1} \\
&= \Omega_{11}^{-1} \\
&= \sigma^2 - \frac{\left(E\left(x-\mu\right)^3\right)^2}{E\left(\left(x-\mu\right)^2 - \sigma^2\right)^2} \le \sigma^2.
\end{aligned} \tag{4.13}$$

Notice that without g_2, $nVar(\hat{\mu})$ equals σ^2 .

In the above example, we notice that $E\left(\frac{\partial f}{\partial \beta}\right) = 0$, and this feature simplifies the calculation dramatically. So we consider the following more special model, where g does not have nuisance parameter, but f has a nuisance parameter only, although it brings some information from the data.

$$E\left[h\left(x;\beta_0,\phi_0\right)\right] = E\begin{bmatrix} g\left(x;\beta_0\right) \\ f(x,\phi_0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{4.14}$$

The gradient vector of h in (4.14) is:

$$\frac{\partial h}{\partial \theta} = \begin{bmatrix} \frac{\partial g}{\partial \beta} & 0 \\ 0 & \frac{\partial f}{\partial \phi} \end{bmatrix},$$

the information on β is:

$$\begin{aligned} \mathcal{I}_\beta^2 &= E\left[\frac{\partial h}{\partial \beta}\right]' \left[E\left(hh'\right)\right]^{-1} E\left[\frac{\partial h}{\partial \beta}\right] \\ &= \begin{bmatrix} E\left(\frac{\partial g}{\partial \beta}\right) & 0 \end{bmatrix} \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}^{-1} \begin{bmatrix} E\left(\frac{\partial f}{\partial \beta}\right) \\ 0 \end{bmatrix} \\ &= E\left[\frac{\partial g}{\partial \beta}\right]' \left[\Omega_{11}^{-1}(I + \Omega_{12}(\Omega_{22} - \Omega_{12}\Omega_{11}^{-1}\Omega_{21})^{-1}\Omega_{21}\Omega_{11}^{-1}\right] E\left[\frac{\partial f}{\partial \beta}\right], \end{aligned}$$

where I is the corresponding identity matrix. Now we have

$$\mathcal{I}_\beta^2 - \mathcal{I}_\beta^1 = E\left[\frac{\partial f}{\partial \beta}\right]' V E\left[\frac{\partial f}{\partial \beta}\right] \tag{4.15}$$

where

$$V = \left[\Omega_{11}^{-1}\Omega_{12}(\Omega_{22} - \Omega_{12}\Omega_{11}^{-1}\Omega_{21})^{-1}\Omega_{21}\Omega_{11}^{-1}\right]. \tag{4.16}$$

By assumption $E(gg^{'})$ is positive semidefinite, so $(\Omega_{22} - \Omega_{12}\Omega_{11}^{-1}\Omega_{21})^{-1}$ is also p.s.d, and so is V. Thus we see that f provide extra information for β. However, if in (4.10), $E(gf) = \Omega_{12} = 0$, $V = 0$, so $\mathcal{I}_\beta^2 = \mathcal{I}_\beta^1$. So we have the following proposition.

Proposition 10 *Additional moment conditions which contains only nuisance parameters will provide extra information on the parameter of interest only if they are correlated to the original moment conditions.*

Remark 8 *Whether g and f are correlated is a testable condition. Since $E\left[g\left(x,\beta_0\right)\right] = E\left[f\left(x,\phi_0\right)\right] = 0$, to test the correlation of g and f it is equivalent to test the following additional moment condition*

$$E\left[\rho\left(x;\beta_0,\phi_0\right)\right] = E\left[g\left(x,\beta_0\right)f\left(x,\phi_0\right)\right] = 0 \tag{4.17}$$

and this can be done by standard EL test procedure.

4.3 Large Deviation Efficiency

In this section we use the same framework as in the previous section to analyse the large deviation efficiency of EL estimator with nuisance parameters and augmented moment conditions. Our work is similar to Kitamura and Otsu (2005), which shows that the minimax loss of EL estimator can achieve the large deviation lower bound in the framework of Puhalskii and Spokoiny (1998). Our result is new in that we incorporate nuisance parameter in the moment condition model. However, the large deviation efficiency in both cases depends on the LDP of statistical experiment, which is introduced below.

4.3.1 Preliminaries

Statistical Experiment

Following the terminology of statistical decision theory as in Blackwell (1953), Strasser (1985, 1996), LeCam (1986), and LeCam and Yang (2000), we call a family of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ a statistical experiment, where P_θ is on a $\sigma-$field $\mathcal{A}(\Sigma)$ of subsets of a set Σ. Let $\{\mathcal{P}_n, n \geq 1\}$ be a sequence of statistical experiments indexed by sample size n, where $\mathcal{P}_n = \{P_{n,\theta} : \theta \in \Theta\}$ and $P_{n,\theta}$ is the set of probability measures which satisfy the moment condition model (4.9) :

$$P_{n,\theta} \equiv \left\{ P \in M_1(\Sigma) : \int_\Sigma g(x,\theta)\, dP = 0 \right\},$$

where $M_1(\Sigma)$ is the space of all probability measures on Σ equipped with the Levy metric. We say $\{\mathcal{P}_n, n \geq 1\}$ is dominated by a probability measure P_0 if for all n and all $\theta \in \Theta$, $P_{n,\theta}$ is absolutely continuous with respect to P_0, i.e. $P_{n,\theta} \ll P_0$, and in this case we also denote $\{\mathcal{P}_n, n \geq 1\}$ as $\{\mathcal{P}_n, P_0, n \geq 1\}$. See, e.g, Puhalskii and Spokoiny (1998). Note that as our setup in section 4.2, $\theta = (\beta', \phi')'$, where $\beta \in \mathcal{B}$, $\phi \in \Phi$, and $\mathcal{B} \cup \Phi = \Theta$. In the following we will use θ and $(\beta', \phi')'$ interchangeably. Now we define:

$$Z_{n,\theta} = \left(\frac{dP_{n,\theta}}{dP_0} \right)^{1/n}, \quad \Xi_{n\theta} \equiv \log Z_{n,\theta} = \frac{1}{n} \log \frac{dP_{n,\theta}}{dP_0} \tag{4.18}$$

and let $Z_{n,\Theta} = \{Z_{n,\theta} \in \mathbb{R} : \theta \in \Theta\}$ and $\Xi_{n,\Theta} = \{\Xi_{n,\theta} \in \mathbb{R} : \theta \in \Theta\}$, so $Z_{n,\Theta}$ and $\Xi_{n,\Theta}$ are the process of likelihood ratio and the log-likelihood ratio respectively.

LDP of Statistical Experiment

Also, let $\mathcal{L}\left(\Xi_{n,\Theta}\left|P_0, n \geq 1\right.\right)$ denote the distribution of $\Xi_{n,\Theta}$ under P_0. Now we have the following definition.

Definition 13 *A sequence of dominated statistical experiments $\{\mathcal{P}_n, n \geq 1\}$ is said to obey the LDP if*

1. the sequence of distributions of likelihood ratio $\mathcal{L}\left(Z_{n,\Theta}\left|P_0, n \geq 1\right.\right)$ obeys the LDP with some rate function $I : \mathbb{R} \to [0, \infty]$.

2. the likelihood ratio $Z_{n,\theta}$ satisfies:

$$\lim_{M\to\infty} \limsup_{n\to\infty} E_n^{1/n}\left[\exp(n\Xi_{n\theta})\mathbf{1}\left(\Xi_{n\theta} > M\right)\right] = 0, \ \theta \in \Theta \tag{4.19}$$

where $E_n^{1/n}[\cdot] = \left(E_n[\cdot]\right)^{1/n}$.

Remark 9 (4.19) *is called the exponential tightness condition, and it is to ensure that lower bounds of minimax risks of the estimate of the parameter θ are independent of the choice of dominating measure P_0. See, Puhalskii and Spokoiny (1998) or LeCam and Yang (2000).*

Example 11 (Exponential Family) *Let $X_n = (X_{1,n}, ..., X_{n,n})$ be n i.i.d samples drawn from a exponential family with density of the standard form:*

$$f\left(X_{k,n}\right) = \exp\left\{\theta X_{k,n} + \kappa\left(\theta\right) + h\left(X_{k,n}\right)\right\}, \quad k = 1, ..., n$$

where $\kappa(\cdot)$ and $h(\cdot)$ are some real functions. For this model, $\Omega_n = \mathbb{R}^n$ and

$$P_{n,\theta} = \exp\left\{\theta \sum_{k=1}^{n} X_{k,n} + n\kappa\left(\theta\right) + \sum_{k=1}^{n} h\left(X_{k,n}\right)\right\}, \quad \theta \in \Theta \subset \mathbb{R}$$

If we take $P_{n,1}$ as the dominating measure, the corresponding log-likelihood ratio will be:

$$\begin{aligned}
\Xi_{n,\theta} &= \frac{1}{n}\log\frac{dP_{n,\theta}}{dP_{n,1}}(X_n) \\
&= (\theta - 1)\frac{1}{n}\sum_{k=1}^{n} X_{k,n} + \kappa\left(\theta\right) - \kappa\left(1\right) \\
&= (\theta - 1)\, Y_n + \kappa\left(\theta\right) - \kappa\left(1\right),
\end{aligned}$$

where

$$Y_n = \frac{1}{n}\sum_{k=1}^{n} X_{k,n}, \quad n \geq 1.$$

It is well known that $\{\mathcal{L}(Y_n\,|P_n), n \geq 1\}$, *the sequence of distributions of the empirical mean* Y_n, *satisfies the LDP on R with the rate function* $I^N(y) = y^2/2$, $y \in R$, *see Dembo and Zeitouni (1998) for instance. Hence the distribution of the log-likelihood ratio* $\mathcal{L}(\Xi_{nP}\,|P_n, n \geq 1)$ *satisfies the LDP by the contraction principle of the LDP.*

To check the first condition in definition 13, it is often convenient to use the following sufficient and necessary condition, see Varadhan (1984), Deuschel and Strook (1989) and also Puhalskii (1993, 2006).

Proposition 11 *A sequence of probability distributions* $\{Q_n, n \geq 1\}$ *obeys the LDP with a rate function I if and only if*

$$\lim_{n\to\infty}\left[\int_\Sigma (f(x))^n\, Q_n(dx)\right]^{1/n} = \sup_{x\in\Sigma} f(x)V(x) \tag{4.20}$$

holds for all nonnegative, bounded and continuous functions f on Σ, *where* Σ *is a metric space and* $V(x) = \exp(-I(x)) : \Sigma \to [0,1]$ *is called the deviability* [1] *of* $\{Q_n, n \geq 1\}$. *Moreover, if f is also nonnegative and lower semicontinuous,* (4.20) *implies*

$$\lim_{n\to\infty}\left[\int_\Sigma (f(x))^n\, Q_n(dx)\right]^{1/n} \geq \sup_{x\in\Sigma} f(x)V(x). \tag{4.21}$$

Proof. See page 493 of Puhalskii (1993). ■

Definition 14 *If* (4.20) *holds, we say* $\{Q_n, n \geqslant 1\}$ *converges to V in large deviation and denote this by* $Q_n \overset{l.d}{\to} V$. *Therefore, by Proposition 11,* $Q_n \overset{l.d}{\to} V$ *if and only if* $\{Q_n, n \geqslant 1\}$ *obeys the LDP with rate function* $I(x) = -\log V(x)$.

The next theorem of Puhalskii and Spokoiny (1998) and Kitamura and Otsu (2005) states that the statistical experiment of our moment condition model obeys the LDP.

Theorem 16 *Suppose* Σ *is a compact metric space, and the likelihood ratio* $\frac{dP_{n,\theta}}{dP_0}$

[1] Note that the range of the rate function $I(x)$ is $[0,\infty)$, and the mapping $I(x)$ to $V(x)$ is one to one.

is continuous and bounded from above, then the sequence of the dominated statistical experiments $\{\mathcal{P}_n, P_0, n \geq 1\}$ *obeys the LDP.*

Proof. The procedure of proof is firstly showing that the distribution of likelihood ratio $\frac{dP_{n,\theta}}{dP_0}$ satisfies the LDP, so the condition 1 in Definition 13 is verified. Secondly, it needs to show the likelihood ratio process is exponential tight, so it satisfies the second condition in Definition 13. See Puhalskii and Spokoiny (1998) or Kitamura and Otsu (2005). ■

4.3.2 Efficiency of Estimation

Minimax Risk Bound

In this section, we show that a large deviation efficiency bound of estimation of the parameter in the model (4.9) can be obtained by the LDP of the statistical experiment $\{\mathcal{P}_n, P_0, n \geq 1\}$. In terms of statistical decision theory, an estimator of the parameter $\theta_n : \Sigma \rightarrow \mathcal{D}$ is a decision in a decision space $\mathcal{D} \ni \theta_n$, and the efficiency of θ_n can be evaluated by a loss function $W : \Theta \times \mathcal{D} \rightarrow \bar{\mathbb{R}}^+$. We define the maximum logarithmic LD risk of the decision ρ_n in the experiment as

$$R(\theta_n) = \sup_{\theta \in \Theta} \sup_{P_{n,\theta} \in \mathcal{P}_n} \frac{1}{n} \log E_{n,P_{n,\theta}} [W_\theta (\theta_n)], \tag{4.22}$$

where $W_\theta(\theta_n)$ denotes the loss of θ_n as an estimator of the parameter θ.

Following LeCam and Yang (2000), we make the following assumptions to ensure the existence of $E_{n,P_\theta} [W_\beta (\theta_n)]$.

Assumption 7 ***a).*** $\inf_{\theta \in \Theta} W_\beta (\theta_n) > -\infty$. ***b).*** *the function* $W_\beta(\theta_n)$ *is measurable.*

Assumption 8 *The parameter space* Θ *is compact.*

An estimator θ_n^* will be called LD optimal if it minimises $R(\theta_n)$, and hence θ_n^* is a minimax estimator. See, e.g., Lehmann and Casella (1998). The reason we consider the minimax estimator, or the reason that we judge the estimator by its worst behavior along a sequence of alternatives converging to a fixed model, is that the uniformity has mathematical appeal because it excludes superefficient estimators, which exploit the weakness in a definition influenced only by pointwise limit behavior (see, e.g., Pollard (2003), and LeCam (1986)).

Bahadur (1960) shows that LD optimality of maximum likelihood estimates in the restricted setting of exponential families. For more general settings, Puhalskii and Spokoiny (1998) gives a framework through the following theorem for LD efficiency of estimates in a statistical experiment, which provides a asymptotic LD lower bound for appropriately defined risk functions if the experiment obeys the LDP. And in fact it is the motivation for introducing the concept of the LDP for sequence of statistical experiments.

Theorem 17 (Minimax LD Risk Bound) *Let θ_n be an estimator of θ in the dominated experiment $\{\mathcal{P}_n, P_0, n \geq 1\}$, which obeys the LDP with rate function $I(x)$. If θ_n is assessed by a level compact loss function W, then with assumption 7-8*

$$\liminf_{n\to\infty} \inf_{\theta_n \in \Theta} R(\theta_n) \geq R^*$$

where

$$R^* = \sup_{Z_{n,\theta}\in R_+} \inf_{\theta_n\in\Theta} \sup_{\theta\in\Theta} W_\theta(\theta_n) Z_{n,\theta} V(Z_{n,\theta}).$$

and $V(\cdot)$ is the deviability of the experiment.

Proof. See theorem 3.1 of Puhalskii and Spokoiny (1998). ■

Remark 10 *This result is indeed an LD analogue of LeCam's minimax theorem, which says that if a sequence of statistical experiments weakly converges, then there exists asymptotic lower bound for the risk of the estimator. See, e.g., LeCam and Yang (2000).*

Remark 11 *From the theorem we know that the minimax LD risk bound is determined by the loss function, the likelihood ratio and the rate function of the LDP of the sequence of experiments.*

Remark 12 *The existence of R^* requires the loss function to be level compact, see Puhalskii and Spokoiny (1998). In practice this condition is often satisfied. For example, the Bahadur type loss function which we will use in the following is level compact given assumption 8. See Kitamura and Otsu (2005).*

Let $b_n : \Sigma \to \mathcal{B}$ be an estimator for our parameter of interest β. The Bahadur-type loss function which we employ in this paper for estimation of β is given by

$$W_\beta(b_n) = \mathbf{1}(\|b_n - \beta\| > c), \; c > 0 \tag{4.23}$$

and we can evaluate the exponential rate of convergence of the LD error probability in estimating β by the following maximum risk function:

$$R(b_n) = \sup_{\beta \in \mathcal{B}} \sup_{P \in \mathcal{P}(\theta)} \frac{1}{n} \log E_{n,P_\theta} \left[W_\beta (b_n) \right]. \tag{4.24}$$

Puhalskii and Spokoiny's (1998) result shows that in a sequence of statistical experiments which obey the LDP, the minimax lower bound of estimates is just the supremum of the product of the rate function and loss function over the sample space, decision space and parameter space. In this framework, the following new theorem gives the lower bound of the minimax risk for our model with nuisance parameters.

Theorem 18 *Define*

$$\Gamma_P(\mu) = \int_\Sigma \log \frac{dP}{dP_0} d\mu$$

where $\mu \in M_1(\Sigma)$. For any estimator b_n of the true parameter β_0, we have

$$\liminf_{n \to \infty} \inf_{b_n \in \mathcal{B}} R(b_n) \geq R^*,$$

where

$$\begin{aligned} R^* &= \sup_{\mu \in M_1(\Sigma)} \inf_{b_n \in \mathcal{B}} \sup_{\beta \in \mathcal{B}} \sup_{P \in \mathcal{P}(\theta): \|b_n - \beta\| > c} \left(\Gamma_P(\mu) - H(\mu | P_0) \right) \\ &= \sup_{\mu \in M_1(\Sigma)} \inf_{b_n \in \mathcal{B}} \sup_{\beta \in \mathcal{B}} \sup_{P \in \mathcal{P}(\theta): \|b_n - \beta\| > c} - H(\mu | P_0). \end{aligned}$$

Proof. Firstly we show that

$$\liminf_{n \to \infty} \inf_{b_n \in \mathcal{B}} \sup_{\beta \in \mathcal{B}} \sup_{P \in \mathcal{P}(\theta)} E_{n,\theta}^{1/n} W_\beta^n (b_n) \geq \sup_{\mu \in M_1(\Sigma)} \inf_{b_n \in \mathcal{B}} \sup_{\beta \in \mathcal{B}} \sup_{P \in \mathcal{P}(\theta)} W_\beta (b_n) Z_\theta V_\Lambda (Z_\Lambda) \tag{4.25}$$

where Λ is some finite subset of $\mathcal{B}$.

By the definition of $Z_{n,\theta}$, we have

$$\begin{aligned}
\liminf_{n\to\infty}\sup_{\beta\in\mathcal{B}}\sup_{P\in\mathcal{P}(\theta)} E_{n,\theta}^{1/n} W_\beta^n\left(b_n\right) &= \liminf_{n\to\infty}\sup_{\beta\in\mathcal{B}}\sup_{P\in\mathcal{P}(\theta)} E_{n,P_0}^{1/n} W_\beta^n\left(b_n\right) Z_{n,\theta;\Lambda}^n \\
&\geq \liminf_{n\to\infty}\left(\frac{1}{|\Lambda|} E_{n,\Lambda} \textstyle\sum_{\theta\in\Theta} W_\beta^n\left(b_n\right) Z_{n,\theta;\Lambda}^n\right)^{1/n} \\
&\geq \liminf_{n\to\infty} E_{n,P_0}^{1/n} \sup_{\beta\in\mathcal{B}} W_\beta^n\left(b_n\right) Z_{n,\theta;\Lambda}^n \\
&\geq \liminf_{n\to\infty} E_{n,P_0}^{1/n} w^n\left(Z_{n,\theta;\Lambda}^n\right)
\end{aligned}$$

where

$$w\left(z_\Lambda\right)=\inf_{b_n\in\mathcal{B}}\sup_{\beta\in\mathcal{B}} W_\beta\left(b_n\right) Z_\theta, \quad Z_\Lambda=\left(Z_\theta, \theta\in\Lambda\right).$$

Considering $\mathcal{L}\left(Z_{n,*}\left|P_n, n\geq 1\right.\right)$ is large deviation converges to V_Λ, by (4.21) we have

$$\liminf_{n\to\infty} E_{n,\Lambda}^{1/n} w^n\left(Z_{n,*}\right) \geq \sup_{z_\Lambda\in R_+^\Lambda} w\left(Z_\Lambda\right) V_\Lambda\left(z_\Lambda\right)$$

which implies (4.25).

Since $w\left(Z_\Lambda\right)$ is nonnegative, continuous and homogeneous, by Lemma 2.5 of Puhalskii and Spokoiny (1998) , we can get

$$\sup_{z_\Lambda\in R_+^\Lambda}\inf_{b_n\in\mathcal{B}}\sup_{\beta\in\mathcal{B}} W_\beta\left(b_n\right) Z_\theta V_\Lambda\left(z_\Lambda\right)=\sup_{z_\Lambda\in R_+^\Theta}\inf_{b_n\in\mathcal{B}}\sup_{\beta\in\mathcal{B}} W_\beta\left(b_n\right) Z_\theta V_\Theta\left(Z_\Theta\right),$$

so combined with (4.25), we have

$$\liminf_{n\to\infty}\inf_{b_n\in\mathcal{B}}\sup_{\theta\in\Lambda} E_{n,\theta}^{1/n} W_\beta^n\left(b_n\right) \geq \sup_{z_\Lambda\in R_+^\Theta}\inf_{b_n\in\mathcal{B}}\sup_{\beta\in\mathcal{B}} W_\beta\left(b_n\right) z_\theta V_\Theta\left(z_\Theta\right).$$

Note that for every $Z_\Theta=\left(Z_\theta, \theta\in\Theta\right)$,

$$\sup_{\Lambda\in\Theta}\inf_{b_n\in\mathcal{B}}\sup_{\theta\in\Lambda} W_\beta\left(b_n\right) Z_\theta=\inf_{b_n\in\mathcal{B}}\sup_{\theta\in\Theta} W_\beta\left(b_n\right) Z_\theta$$

and the proof is completed by taking logs of both sides. ■

Remark 13 *Here we see that finally the efficiency bound turns out to be not dependent on the dominating measure.*

Now we consider the empirical likelihood estimator after profiling out the nuisance parameter ϕ

$$\hat{\beta} = \underset{b}{\arg}\inf_{b\in\mathcal{B}}\mathcal{R}\left(\beta\right)$$

where

$$\mathcal{R}\left(\beta\right) = \min_{\phi\in\Phi}\max_{\mu\in R^m}\sum_{i=1}^{n}\log\left(1+\mu^{'}\left(\beta,\phi\right)g\left(x_i;\beta,\phi\right)\right).$$

From theorem 18 we know that $\liminf\mathcal{R}\left(\hat{\beta}\right) \geq R^*$, we will check whether $\limsup R\left(\hat{\beta}\right)$ is smaller than R^*,i.e.,

$$\limsup R\left(\hat{\beta}\right) \leq R^*.$$

Firstly we present an important result which connects the EL estimation with the statistical experiment.

Lemma 3 *For each $\theta\in\Theta$, let $\mu_n \in M_1\left(\Sigma\right)$ be the empirical measure, we have*

$$\begin{aligned}
& \sup_{P\in\mathcal{P}(\theta)}\int_{\Sigma}\log\frac{dP}{dP_0}d\mu_n \\
= & -\max_{\mu\in R^m}\sum_{i=1}^{n}\log\left(1+\mu^{'}\left(\theta\right)g\left(x_i;\theta\right)\right) - \int_{\Sigma}\log\frac{dP_0}{d\mu_n}d\mu_n \\
\equiv & \ \mathcal{R}\left(\theta\right) - \tilde{L}.
\end{aligned}$$

Proof. See Borwein and Lewis (1993) or Kitamura and Otsu (2005). ■

Theorem 19 *Suppose $\hat{\beta}$ solves*

$$\inf_{\beta\in\mathcal{B}}\sup_{\beta\in\mathcal{B}:\|b_n-\beta\|>c}\mathcal{R}\left(\beta\right),$$

then

$$\lim_{n\to\infty}R\left(\hat{\beta}\right) = R^*.$$

Proof. Let $(a \wedge b)$ denote $\min(a, b)$. We have

$$\begin{aligned}
\exp\left(R\left(\hat{\beta}\right)\right) &= \sup_{P\in\mathcal{P}(\theta)} E^{1/n}_{nP_\theta}\left[1\left(\left\|\hat{\beta}-\beta\right\| > c\right)\right] = \sup_{P\in\mathcal{P}(\theta)} E^{1/n}_{nP_0}\left[1\left(\left\|\hat{\beta}-\beta\right\| > c\right) Z^n_{nP}\right] \\
&\leq \sup_{P\in\mathcal{P}(\theta)} E^{1/n}_{nP_0}\left[1\left(\left\|\hat{\beta}-\beta\right\| > c\right)\left(Z_{nP}\wedge e^M\right)^n\right] \\
&\quad + \sup_{P\in\mathcal{P}(\theta)} E^{1/n}_{nP_0}\left[1\left(\left\|\hat{\beta}-\beta\right\| > c\right) Z_{nP} 1(Z_{nP} > e^M)\right] \\
&\leq \sup_{P\in\mathcal{P}(\theta)} E^{1/n}_{nP_0}\left[1\left(\left\|\hat{\beta}-\beta\right\| > c\right)\left(Z_{nP}\wedge e^M\right)^n\right] \\
&\quad + \sup_{P\in\mathcal{P}(\theta)} E^{1/n}_{nP_0}\left[Z_{nP} 1(Z_{nP} > e^M)\right]
\end{aligned}$$

Since

$$\sup_{P\in\mathcal{P}(\theta)} E^{1/n}_{nP_0}\left[Z_{nP} 1(Z_{nP} > e^M)\right] \leq 0$$

we can just show

$$\sup_{P\in\mathcal{P}(\theta)} E^{1/n}_{nP_0}\left[1\left(\left\|\hat{\beta}-\beta\right\| > c\right)\left(Z_{nP}\wedge e^M\right)^n\right] \leq \exp\left(R^*\right)$$

Note that

$$\begin{aligned}
&\sup_{P\in\mathcal{P}(\theta)} E^{1/n}_{nP_0}\left[1\left(\left\|\hat{\beta}-\beta\right\| > c\right)\left(Z_{nP}\wedge e^M\right)^n\right] \\
\leq\; & E^{1/n}_{nP_0}\left[\left(\sup_{P\in\mathcal{P}(\theta)}\left\{1\left(\left\|\hat{\beta}-\beta\right\| > c\right) Z_{nP}\right\}\wedge e^M\right)^n\right] \\
=\; & E^{1/n}_{nP_0}\left[\left(\sup_{\beta\in\mathcal{B}:\|b-\beta\|>c} \sup_{P\in\mathcal{P}(\theta)}\left\{Z_{nP}\right\}\wedge e^M\right)^n\right] \\
\leq\; & E^{1/n}_{nP_0}\left[\left(\sup_{\beta\in\mathcal{B}:\left\|\hat{\beta}-\beta\right\|>c} \sup_{P\in\mathcal{P}(\theta)}\left\{\exp\int_{\mathcal{X}}\log\frac{dP}{dP_0}d\mu_n\right\}\wedge e^M\right)^n\right]
\end{aligned}$$

where the last inequality follows form Lemma 3 and the definition of $\hat{\beta}$. Thus, from the LD convergence of Z_{nP}

$$\begin{aligned}
& \lim_{n\to\infty} E_{nP_0}^{1/n}\left[\left(\sup_{\beta\in\mathcal{B}:\left\|\hat{\beta}-\beta\right\|>c}\sup_{P\in\mathcal{P}(\theta)}\{Z_{nP}\}\wedge e^{M}\right)^{n}\right] \\
\leq\ & \lim_{n\to\infty} E_{nP_0}^{1/n}\left[\left(\sup_{\beta\in\mathcal{B}:\left\|\hat{\beta}-\beta\right\|>c}\sup_{P\in\mathcal{P}(\theta)}\{Z_{nP}\}\right)^{n}\right] \\
=\ & \sup_{\mu\in M_1(\Sigma)}\sup_{\beta\in\mathcal{B}:\left\|\hat{\beta}-\beta\right\|>c}\sup_{P\in\mathcal{P}(\theta)}\{Z_{nP}\}\exp(-I(Q)) \\
=\ & R^{*}
\end{aligned}$$

■

The proofs of our main results, theorem 18 and 19, resemble those of theorem 3.1 and 4.1 in Puhalskii and Spokoiny's (1998), because the key thing is that as we mentioned at the beginning of this section, we need to show the LDP of the likelihood ratio process $\frac{dP_{n,\beta}}{dP_0}$ induced by the parameter of interest, hence the LDP of the statistical experiment. Intuitively, it is not hard to believe that if the likelihood ratio process $\frac{dP_{n,\theta}}{dP_0}$ satisfies the LDP, the process $\frac{dP_{n,\beta}}{dP_0}$ also satisfies the LDP, since β is a subvector of θ.

4.4 Concluding Remarks

In this chapter we have discussed the efficiency of the EL estimator in the presence of nuisance parameters, via both standard asymptotic method and large deviations. We are particularly interested in whether the asymptotic efficiency of the parameter of interest can be improved by adding more moment conditions. We found that a necessary condition for augmented moment condition to be useful to improve the asymptotic efficiency, is that it is correlated to the original moment condition. It is worth mentioning that here we incorporate more moment conditions with sample size being fixed, while researchers like Newey and Windmeijer (2005) and Han and Philips (2006) consider increasing the number of moment conditions brought by increasing sample size.

It would be interesting to extend our results to the non i.i.d case. We have shown some LDP results for weakly dependent data, and so it is not difficult to obtain a corresponding large deviation efficiency bound. But if we want to show that the EL estimator can achieve this bound it would be complicated since Lemma 3 depends on i.i.d assumption. Therefore some other results which connects the EL criterion and the likelihood ratio process may be needed.

Chapter 5

Empirical Likelihood Estimation of Auction Models via Simulated Moment Conditions

Abstract

In this chapter we apply empirical likelihood (EL) estimation to the simplest first-price sealed bid auction model with independent private values. Through estimation of the parameter in the distribution function of bidders' private value we consider a potential problem in the EL inference when the moment condition is not in an explicit form and is hard to compute, or even not continuous in the parameter of interest. We deal with this issue following the method of simulated moment (MSM) introduced by Pakes and Pollard (1989) and McFadden (1989), since in structural auction models the first moment of the optimal bid is highly nonlinear and thus intractable. Particularly we use the importance sampling method to simulate the moment condition, which is derived from the Bayesian Nash equilibrium in the game theoretical auction model. We demonstrate the convergence of the empirical likelihood estimator from the simulated moment condition, and find that the asymptotic variance is larger than usual, and is modified by simulation.

Key words: first-price auction, simulated moments, importance sampling

5.1 Introduction

We consider empirical likelihood estimation of a simple auction model in this chapter. Auctions are nowadays widely implemented as an efficient mechanism to allocate resources, to determine prices and to mitigate transaction costs. Although the use of auctions has a long history and can be traced back to Roman times, the theory of auctions in economics has flourished since Vickrey's (1961) seminal paper, showing how Pareto-optimal results can be achieved through auctions in imperfect markets. Particularly, the development of game theory has enhanced the research into auctions in the past decades. One of the most earliest and most general results is contributed by Milgrom and Weber (1982), who analyze English auctions [1], Dutch auctions [2], first price auctions [3]and second-price auctions [4] using a general game theoretical model and derive the equilibrium conditions in a setting of competitive and noncooperative bidding. Literature about other equilibrium strategies under different settings of auctions is also growing. For instance, Maskin and Riley (1984) consider the optimal auctions from the point of view of sellers who want to maximize the transaction price, assuming bidders are risk-averse. Swinkels (1998) discusses the strategy and conditions needed to make a large number of auctions to be efficient, in the sense that bidders who value the object the most will have the biggest possibility of winning it. For general reference and a recent survey of auction theory, Klemperer's (2004) book can be referred to.

One of the interesting features of auction models in empirical study is that they are fully structural [5], which means the model is derived from economic theory directly incorporating restrictions from the theory as assumptions of econometric models, and randomness enters the auction model naturally without adding stochastic error terms, unlike the usual econometric models. Based upon

[1] English auction: Aslo called ascending auction. The auctioneer begins the auction by announcing a starting price or reserve for the item on sale and then accepts increasingly higher bids from the bidders.

[2] Dutch auction: Also called descening auction. The auctioneer begins the auction by announcing a high price, then he lowers the price continuously until some bidder accept the price and the player win the object at that price.

[3] First-price (sealed bid) auction: All bidders submit their bids in an envelope simultaneously to the auctioneer. Bidder with the highest bid wins, paying a price equal to the exact amount that he or she bid.

[4] Second-price (sealed bid) auction: Also called Vickrey auction. All bidders submit their bids in an envelope simultaneously to the auctioneer. Bidder with the highest bid wins, paying a price equal to the exact amount of the *second highest* bid.

[5] Application of structural approach to auction model began from Paarsch (1992).

the structural approach, researchers are interested in testing economic theories implied by the auction model through both field and experimental data. Early examples include the work done by Hansen (1985) which aims to test the famous revenue equivalence theorem, which asserts that a seller will always get the same revenue from any allocation mechanism given some specific features of the bidders. Recently Haile, Hong and Shum (2003) conduct tests to examine whether the bidders have common value or private value on the objects. Paarsch and Robert (2003) generate laboratory data to test equilibrium behavior when bidders bid in discrete increments. See also Athey and Haile's (2005) introduction and a list of comprehensive references.

On the other hand, this empirical testing procedures rely on the distribution functions which characterize the bidders' valuation, or demand on the auctioned goods. As Milgrom and Weber (1982) have shown, in equilibrium the optimal bid is a function of bidders' private value, the number of bidders, the reservation price of the objects, and the optimal bid is monotonically increasing in bidders' private values, *i.e.*, the bidder with higher value will bid more. Based upon this result econometricians have attempted to estimate the distribution function of bidder's private value by the relationship between the optimal bid and private value, assuming the Bayesian Nash Equilibrium in the competitive bidding.

However, there are some difficulties in the structural analysis of the auction models. For the relevant variables, we can observe the number of bidders and the reservation price, but sometimes not all the bids can be observed [6]. Moreover, bidders' private values are latent, *i.e.*, the data generation process (DGP) which can be observed is incomplete. This leads to two problems. The first one is identification: can the distribution functions of the private value be recovered from the data in various types of auctions? Secondly, the equilibrium strategy is in a highly nonlinear form, so explicit calculation of the private value out of the strategy is unfeasible.[7]

[6] How many bids can be observed depends on the type of auctions. In an English auction, sometimes not all of the bids can be observed, since the potential bidders whose valuations of the object are lower than the current bids will not bid at all. In an Dutch auction, since the auction will end right after someone has made a bid, only the wining bid is observed. In first price and second price sealed bid auctions, all the bids can be observed.

[7] Apart from these two problems, some researchers like Albano and Jouneau (1998) who suggested a Bayesian approach to the first price auction model, pointed out that in the existing frequentist approach the auction models cannot indeed be fully structured indeed, since they need asymptotics and the proposed estimator is on a product space. Specifically, asymptotics on the number of bidders will involve an infinite number of bidders and if it is not the case, the econometric model will not be fully structured. However, the structural formula which

Athey and Haile (2002, 2005) have examined the condition for nonparametric identification in different model settings. One of their conclusions based on the equilibrium strategy found by Milgrom and Weber (1982) is that the distribution function can be identified even when only the winning bid is observed in the simplest independent private value (IPV) setting with symmetric bidders. This means that the bidder's private values are independently drawn from the same distribution function and each bidder only knows his/her own value. Indeed, Guerre *et.al* (2000) propose a two step nonparametric kernel estimator for the distribution function with optimal convergence rate, which does not need calculation of the equilibrium strategy.

In a parametric approach, early research by Paarsch (1992), and Donald and Paarsch (1993, 1996), also concludes that the parameter in the distribution function of private values can be identified through the winning bid within the IPV paradigm, assuming that the bidders's private value distribution is from some specific parametric family. Since the support of the bid distribution involved depends on the parameter of interest, the maximum likelihood estimator (MLE) from directly maximizing the likelihood function is not consistent. They therefore suggest pseudo maximum likelihood (PML) to estimate the parameter, but computation of likelihood function is challenging due to its high degree of non-linearity. To avoid calculating the density function, Laffont, Ossard and Vuong (1995), (hereafter Laffont *et al.* (1995)) consider simulating the first moment of the winning bid, and using non-linear least squares (NLLS) estimation which minimizes the sample analogue of the simulated moments.

Laffont *et al.* (1995)'s estimation method is an application of method of simulated moment (MSM) introduced by Pakes and Pollard (1989) (hereafter PP), McFadden (1989) and McFadden and Ruud (1994) (hereafter MR). In their paper they discuss the problem when the general method of moments (GMM) encounters moment conditions which cannot be handled as usual, specifically, when we estimate a parameter θ_0 based on the following population moment condition:

$$E\left[g\left(x,\theta_0\right)\right]=0.$$

GMM and also empirical likelihood (EL) will be difficult if the moment indicator

provides the optimal bidding strategy as a function of private values will converge uniformly to identity function when the number of bidders goes to infinity. On the other hand, if we do asymptotics on the number of repeated auctions instead of bidders, the equilibrium strategy in repeated auctions can be much more complex than the one in a static auction.

$g(x,\theta)$ is intractable and hard to compute, or even discontinuous in the parameter of interest, since both GMM and EL require explicit calculation of the sample analogue of the moment condition and existence of the derivative matrix of $g(x,\theta)$ with respect to θ. PP have established general asymptotic results of the estimator based upon simulation of the moment condition, confining the parameter space to some specific class. As an immediate application but in an independent piece of work, McFadden (1989) estimates a discrete response model by simulating the intractable response probability forming the moment condition. A similar case worth mentioning is that, the functional form of $g(x,\theta)$ itself is tractable but it contains some unknown function, say, $g(x,\theta) = g(x,h(x),\theta)$ where $h(x)$ is unknown. MR consider sample analogue based on simulation of $h(x)$, while Ai and Chen (2003) approximate $h(x)$ nonparametrically by sieve minimum distance (SMD) method. .

In this chapter, we apply empirical likelihood to estimate a first-price auction model under symmetric IPV assumption in a parametric setting. The moment condition we are based on, which is the same as Laffont *et al.* (1995) used, is from the expectation of the winning bid derived according to the equilibrium strategy. Also following Laffont *et al.* (1995), we simulate the intractable moment condition by importance sampling, which is used to evaluate the moment condition through observations from a different probability distribution and is easier to handle, rather than using $g(x,\theta)$ directly. General references about importance sampling technique can be found in Rubinstein (1981) and Hesterberg's thesis (1988), among others.

We notice that as McFadden (1989) points out, importance sampling can be used to smooth discrete moment conditions. So we extend our estimation method to more general case where the moment conditions may be either intractable or discrete. Similar to the results of PP, the proof of consistency of our EL estimator based on MSM requires only the continuity of the simulated moment condition, but not that of the original one. However, the proof of asymptotic normality dose require differentiability of the moment condition at the true parameter and the derivative matrix must be full rank.

This chapter is organized as follows. Firstly in section 5.2 we describe the game theoretical auction model. We review the derivation of equilibrium strategy and the conditions for identification in our symmetric IPV setup. In Section 5.3 we consider the empirical likelihood estimator using simulated moment condition by importance sampling, and asymptotic properties of the estimator will be established. Also extension from an auction model to a more general moment condition is mentioned. Section 5.4 provides experimental results. Section 5.5 concludes and proposes some extensions to further research.

5.2 First Price Auction with Symmetric IPV Bidders

5.2.1 The Game Theoretic Model

Consider a first price sealed bid auction as a noncooperative game. Some risk neutral [8] bidders which are indexed by $i = 1, ..., I$ with $I \geqslant 2$, bid for a single and indivisible object. Bidders submit their bids to the auctioneer simultaneously and the bidder with the highest bid wins, provided that her/his bid is no smaller than the reservation price u_0 set by the seller. Suppose that bidder i holds his own value [9] u_i on the object which is from a probability distribution $F(u)$ with a bounded support $suppU = [\underline{u},\bar{u}]$ where $0 \leq \underline{u} < \bar{u} < \infty$. Furthermore, each bidder knows the number I and the function $F(u)$ and he knows that the others know, etc. As shown by Riley and Samuelson (1981), by making the bid $b_i \equiv \beta(u_i)$ according to her/his private value, in the game bidder i want to maximize her/his expected utility $U_i \equiv U(u_i, b_i)$, i.e.,

$$E[U_i] = (u_i - \beta(u_i)) \times p_i \tag{5.1}$$

where p_i is the probability of the bidder i wins.

For the value u_i and the distribution function $F(u)$ we make the following assumptions which form the IPV model:

Assumption 9 (Symmetric and independent bidders) *All the I bidders' values are independently drawn from the same distribution $F(u)$.*

Symmetric bidders are not identical however, since their private signal will be different. Indeed, in a symmetric game heterogeneity across bidders is embodied in differences among private signals. The case of asymmetric bidders can arise from the fact that some of them are well informed, some of them may have collusion, or they have different sizes and locations that can affect their distribution of private signal.

Sometimes it is convenient to use order statistics for explaining independent private values drawn from $F(u)$. For the set of I private values $\{u_1, ...u_I\}$, let

[8] In general, models with risk averse or risk seeking bidders are nonidentifiable if no additional restrictions are given. See, e.g. Maskin and Riley (1984) or Campo *et al.* (2000). So for simplicity and concentration on estimation methods we only consider risk neutral bidders.

[9] Here 'value' can also be termed as 'utility'. See Athey and Haile (2002). Indeed, the bidder i will receive utility $u_i - p$ if he wins the object at price p.

$u^{(k:I)}$ denote the kth order statistic, and $u^{(I:I)} = \max\{u_1, ...u_I\}$. Correspondingly, $b^{(k:I)}$ will denote the kth order statistic of the I independent bids.

Assumption 10 (Private value) *Each bidder knows only her/his value but does not know others' values. Equivalently, each bidder does not know any information relevant to other bidders' utility.*

In contrast, if bidders have common value [10], a bidder's belief would be influenced by other bidders' information or signal other than her/his own. And by private value assumption we avoid the problem of 'winners curse', which means the winner will tend to overpay.[11].

Assumption 11 *The distribution function $F(u)$ is absolute continuous with density $f(u)$ with respect to Lebesgue measure and the expectation of the private value is finite, i.e., $E(u) = \int uf(u)du < \infty$.*

5.2.2 The Equilibrium Bid Function

Suppose now we have incomplete knowledge of $F(u)$ and we want to estimate it. [12]Since in an auction bidders' private values u cannot be observed, we have to obtain the relationship between u and b which we can observe. The optimal bidding strategy for bidder i is a result of symmetric Bayesian-Nash equilibrium (SBNE), obtained by Riley and Samuelson (1981). A bidding strategy $\beta(u_i)$ is a SBNE strategy if for all valuations, it is a best response for bidder i if for all bidders $j \neq i$ also use $\beta(u_i)$. Maskin and Riley (2000) show that if $\beta(u_i)$ is a

[10] To test whether bidders have private value or common value. Hailey etl. (2003) conducted a nonparametric test based on their finding that in a first price auction with private values the equilibrium optimal bid is invariant to the number of bidders I, while in a common value model it is strictly increasing in I.

[11] For a formal illustration of this result obtaied by conditional expectation, see e.g., McFadden's note: http://elsa.berkeley.edu/~mcfadden/eC103_f03/curse2.pdf

[12] An example of the significance of finding $F(u)$ in economic practice. From the point of the seller, his expected revenue R is

$$E(R) = N \int_{u_0}^{+\infty} \left(u\left(F(u)\right)^{N-1} - \int_{u_0}^{u} \left(F(x)\right)^{N-1} dx \right) f(u)du \tag{3}$$

A seller who wants to set a optimal reservation price p^* which can maximizes $E(R) + u_0 \left(F(u_0)\right)^N$ needs to know $F(u)$. According to Laffont and Maskin (1980), p^* should solve

$$p^* = u^0 + \frac{1 - F(p^*)}{f(p^*)}$$

bidder's best response then it is monotonic in valuations. Riley and Samuelson (1981) showed the unique symmetric Bayesian-Nash equilibrium which is our structural econometric model with which we estimate the distribution of the unobserved private values using the observed bids.

Theorem 20 (Riley and Samuelson,1981) *Suppose Assumption 9-10 hold. In a first price sealed bid auction with reservation price u_0, the optimal bidding strategy of a risk neutral bidder i with private value $u_i \geqslant p_0$ is*

$$b_i \equiv \beta(u_i) = u_i - \frac{1}{(F(u_i))^{I-1}} \int_{u_0}^{u_i} (F(\zeta))^{I-1} d\zeta. \tag{5.2}$$

Proof. Consider bidder i makes a bid of x. Note that x is also a function of u_i. The probability of bidder i wins with bid x is equal to the probability of all the other $I-1$ bidder's bids are smaller than x[13], i.e.,

$$p_i = \prod_{j=1,\ j\neq i}^{I} P(b_j < x) = [F(x)]^{I-1}, \tag{5.3}$$

since bidders are independent and symmetric. Now the bidder's problem (5.2) can be expressed as

$$\underset{x}{Max} \quad E[U_i] = [u_i - x(u_i)] \times [F(x)]^{I-1}. \tag{5.4}$$

The first order condition of (5.4) with respect to x is

$$\frac{\partial E(U_i)}{\partial x} = \left[-x'(u_i)\right] \times [F(x)]^{I-1} + (I-1)[u_i - x(u_i)][F(x)]^{I-2} f(x) = 0,$$

which implies that the optimal bid $\beta(u_i)$ satisfies the following differential equation:

$$\beta'(u_i) = (I-1)[u_i - b(u_i)] \frac{f(u_i)}{F(u_i)}. \tag{5.5}$$

Note that in equilibrium, if the bidder's private value is exactly the reservation price u_0, she/he will bid u_0, i.e., the boundary condition for (5.5) is $\beta(u_0) = u_0$. Since otherwise, if $\beta(u_0) > u_0$, the utility will be negative; and if $\beta(u_0) < u_0$, the object will remain unsold. Hence by integrating (5.5) with the boundary condition the result follows. ■

[13] We ignore the case of ties in the highest bids and private values.

Remarks.

1. In Dutch auction the equilibrium bid also satisfies (5.2), i.e., Dutch auction and first price sealed bid auction are strategically equivalent.[14]

2. The model (5.2) is only valid when bidder's private value is no less than the reservation price u_0. Otherwise if $u_i < u_0$, b_i can take any value strictly less than u_0, and the auctioned object remains unsold.

3. A special case of this relationship is that when I=1, i.e., there is only one bidder, the optimal strategy is to bid his own private value, given it is no smaller than the reservation price.

Since b_i is a function of u_i which is random, b_i is also a random variable from a probability distribution, say, $\Phi(\cdot)$ with density $\phi(\cdot)$, which is uniquely determined by (5.2). Note that b_i is strictly increasing in u_i on $[u_0, \bar{u}]$, i.e, a bidder will bid more if his private value is higher. Hence in a first price auction, a bid of particular interest, the winning bid b^w is a function of the highest private value $u^{(I:I)}$:

$$b^w = \beta(u^{(I:I)}) \tag{5.6}$$

where the density of $u^{(I:i)}$ is $f(u^{(I:i)}) = n\left[F(u)\right]^{n-1} f(u)$ (see e.g. David (1981)). The structural approach of auction models to estimate the latent private value distribution using observed bids is based on the relationship of $\Phi(b)$ and $F(u)$ through (5.2). For the distribution $\Phi(b)$ we have the following two results from the theorem, which have been mentioned by some authors without proof, see e.g., Guerre *et. al.* (2000).

Corollary 1 *Given assumption 11, the support of the distribution of the equilibrium strategy suppB*$=\left[\underline{b},\bar{b}\right]$ *is finite.*

Proof. From (5.2) we have

$$\begin{aligned} \bar{b} &= \int_{u_0}^{\bar{u}} u(I-1)f(u)F^{I-2}(u)\,du \\ &\leq (I-1)\int_{u_0}^{\bar{u}} uf(u)du < \infty, \end{aligned}$$

since $E(u)$ is finite according to assumption 11. ■

[14] Similarly, English and second price sealed bid auction are strategically equivalent. See, e.g., Milgrom and Weber (1982).

Corollary 2 *The private value u_i can also be written as a function of b_i and $G(b_i)$ as following:*

$$u_i = b_i + \frac{1}{I-1}\frac{\Phi(b_i)}{\phi(b_i)}.$$

Proof. Let $\beta^{-1}(b)$ denote the inverse function of the bidding strategy according to (5.2), so $u_i = \beta^{-1}(b_i)$. Note that

$$\Phi(b) = \Pr(B \leq b) = \Pr\left(U \leq \beta^{-1}(b)\right) = F\left(\beta^{-1}(b)\right) = F(u),$$

and hence, $\phi(b_i) = \partial\Phi(b_i)/\partial b_i = \partial F(u_i)/\partial b(u_i) = f(u_i)/b'(u_i)$. Now we have

$$\frac{\Phi(b_i)}{\phi(b_i)} = b'(u_i)\frac{F(u)}{f(u_i)}. \tag{5.7}$$

and the result follows by combining (5.7) and (5.5) ■

5.2.3 Nonparametric Identification

The identification problem in the structural auction model consists of whether the distribution of private values can be uniquely determined from observable data, including the number of bidders I, reservation price u_0, and bids, and therefore the first problem we often consider is data availability, which varies in different types of auctions. In an English auction, the reservation price u_0 is announced by the auctioneer but bids are only observed when they are called out by the bidders. Moreover, the last bid called out by a bidder only provides an upper bound on the private value of that bidder. In Dutch auctions, only one bid-the winning bid is observed, and the reservation price is not necessarily revealed. The number of potential bidders I is also unknown. In Vickrey auctions and first price auction, we can observe the number of bidders and all the bids.

Athey and Haile (2002) considered the nonparametric identification problem in various situations including both different assumptions on bidders' value and types of auctions. They found the following result in the simplest symmetric IPV case.

Theorem 21 (Athey and Haile (2002)) *For the simplest symmetric IPV case, $F(u)$ can be identified even when only the winning bid, or the the transaction price in first price auctions, is observed.*

Proof. The distribution of kth order statistic $u^{(k:I)}$ of independent samples of size I from distribution $F(u)$ is (see, e.g., David (1981))

$$F(u^{(k:I)}) = \frac{I!}{(I-k)!(k-1)!} \int_0^{F(u)} t^{k-1}(1-t)^{I-k} dt. \tag{5.8}$$

The right hand side of (5.8) is strictly increasing in $F(u)$, therefore if we know the distribution of $u^{(k:I)}$ for any k, including the Ith order statistic $u^{(I:I)}$, $F(u)$ is uniquely determined. Hence $F(u)$ can be identified by the wining bid which is a monotonic function of the highest value $u^{(I:I)}$. ■

Remarks.

1. This result applies not only to first price and Dutch auction, but also to second price and English auctions, although they have different equilibrium strategies.

2. To sketch how $F(u)$ can be nonparametrically estimated by the winning bid b^w following this theorem, we firstly estimate nonparametrically the distribution $\Phi(b^w)$ of b^w and its density $\phi(b^w)$ by the empirical distribution function and the kernel density, respectively:

$$\tilde{\Phi}(b^w) = \frac{1}{N}\sum_{n=1}^{N} 1\left(b_n^w \leq b\right), \quad \tilde{\phi}(b^w) = \frac{1}{Nh_b}\sum_{n=1}^{N} K_b\left(\frac{b-b_n^w}{h_b}\right),$$

where $K_b(\cdot)$ is some kernel, h_b is a bandwidth and N is the number of auction being repeated. Then we can construct pseudo highest private value by corollary 2:

$$\tilde{u}_n^{(I:I)} = b_n^w + \frac{1}{I-1}\frac{\tilde{G}(b^w)}{\tilde{g}(b^w)},$$

and then the density of $u_n^{(I:I)}$ can be estimated by

$$\tilde{f}\left(u_n^{(I:I)}\right) = \frac{1}{Nh_u}\sum_{n=1}^{N} K_u\left(\frac{u-\tilde{u}_n^{(I:I)}}{h_u}\right),$$

where $K_u(\cdot)$ is some kernel and h_u is a bandwidth. Guerre et.al (2000) established uniform consistency of $\tilde{f}\left(u_n^{(I:I)}\right)$ and show that it has the best uniform convergence rate for estimating the latent density of private values from observed bids.

However, the identification problem will be complicated if the symmetric IPV assumption is violated. Specifically, if bidders are not symmetric, the distributions of their private values cannot be identified by a single bid unless more observations about the auction and bidders are available. Also, identifiability depends on type of auctions if bidders value are not independent. For instance, in second price auctions $F(u)$ is not identified unless all bids are observed. See also Athey and Haile (2002) for a complete treatment in different settings.

5.2.4 Parametric Setting

More often in the analysis of field auction data, the researcher assumes that the random variable u come from some specific family of probability distributions. From now on, following Laffont and Vuong (1996) we suppose that we know the true distribution of the private value $F(u)$ takes the following form:

$$F(u) = F\left(u; \theta_0, z\right), \tag{5.9}$$

which involves an unknown parameter vector $\theta_0 \in \Theta \subset \mathbb{R}^p$. Accordingly, the density is $f(u) = f\left(u; \theta_0, z\right)$. In the function the vector z represents some observable variables which affect bidders' value. In practice z could be some features of the auctioned objects, e.g., the estimated oil reserve of an oil well in an auction for drilling rights.

Similar to theorem 21, the parameter θ_0 can also be identified using only the winning bid, given z and the number I are observed. Specifically, Donald and Paarsch (1996) showed that the density of the winning bid b^w at first price and Dutch auctions can be written in terms of $F(\cdot)$ and the inverse bid function $\beta^{-1}(\cdot)$ as

$$\phi\left(b^w; I, \theta, z\right) = \frac{I \times F\left(\beta^{-1}\left(b^w; \theta\right), \theta, z\right)}{(I-1)\int_{u_0}^{\beta^{-1}(b^w;\theta)} F\left(\zeta; \theta, z\right) d\zeta}, \tag{5.10}$$

and identification of θ_0 means there is no element in Θ other than θ_0 which solves (5.10), where b^w is expressed according to Theorem 20.

5.3 Estimation

Based on the result that $F(u;\theta_0,z)$ can be identified by the winning bid, we focus on b^w [15] and suppose we observed a sequence of identical auctions[16] indexed by $n=1,...,N$ and by *identical* we assume that an auction is independently repeated N times and the joint distribution $F(u_1,...,u_I)$ is the same across the N auctions. So now we have a sequence of i.i.d random variables $b_1^w,...,b_N^w$ with density as (5.10). Since we are focusing on the wining bid, for simplifying notation and without confusion, from now on we will sometimes drop the superscript of b_n^w, e.g. b_n will denote the winning bid in the nth auction.

Donald and Paarsch (1996) consider maximum likelihood (ML) for estimating θ_0. One of the problems they mentioned is that the support of the distribution of b^w depends on θ_0[17], and thus the standard assumptions of ML estimation are violated. Moreover, the calculation of the inverse strategy in the density of b^w is computationally complicated. Indeed $\beta^{-1}(\cdot)$ cannot be expressed explicitly and numerical methods is needed. They solved these problem by maximizing an approximated objective function subject to some binding constraints, making it uniformly convergent to the joint density which is maximized at θ_0.

Alternatively, instead of considering the density of b^w, Laffont *et al.* (1995) used the first moment of the winning bid and avoided exact computation of the inverse bidding strategy. Specifically, if $m(z,\theta_0)$ denotes the expectation of b^w, i.e. $E(b_n)=m(z_n,\theta_0)$, a nonlinear least squares (NLLS) estimator can be

[15] However, if all the bids are available, they can be helpful to estimate θ_0 as well. See Li and Vuong (1997) for their extension of the framework of Laffont *et al.* (1995) to estimation by all bids.

[16] Asymptotics on many auctions may cause some theoretical problems, however. Since bidders' strategies in repeated auctions will be very complex compared to strategies in a single auction, the model implied by (5.2) is not fully structured (Albano and Jouneau (1998)). To simplify this situation we make two more assumptions.

Assumption 12 *The number of bidders in each auction is fixed and known to the researcher.*

This rules out the problem of entry and the number I remains fixed in the model (5.2). For how to deal with different numbers of bidders in each auction, see Li (2005).

Assumption 13 *The private values of bidder i are independent across auctions.*

If a bidder's private value is not independent, i.e., it depends on private values and bids in previous auctions, the optimal bidding strategy will no longer be the model (5.2). See Birhchandani (1988) for the equilibrium solution in repeated auctions.

[17] See also Corollary 1, considering $F(u)$ replaced by $F(u;\theta,z)$.

obtained by minimizing the objective function

$$Q(\theta) = (1/N) \sum_{n=1}^{N} (b_n - \bar{m}(z_n, \theta))^2$$

with respect to θ, where $\bar{m}(z_n, \theta)$ is an unbiased simulated estimator of $m(z_n, \theta_0)$.[18]

Following this framework, we propose here an alternative estimator based on the simulated moment condition, but using, instead of NLLS as in Laffont *et al.* (1995), empirical likelihood methods to estimate θ_0. Let

$$g(x_n, \theta) = b_n - m(z_n, \theta), \tag{5.11}$$

where x_n denotes the vector of observable data including b_n and z_n. Then the i.i.d random variables $b_1, ..., b_N$ satisfy the following moment condition:

$$E[g(x_n, \theta_0)] = 0, \tag{5.12}$$

where the expectation is taken with respect to $F(u; \theta_0, z)$. The difficulty for both NLLS and other methods of estimation, is that $m(z_n, \theta)$, and hence $g(x_n, \theta_0)$, is not directly available.

5.3.1 Simulated Moment Condition

The Problem

Following the usual setup (e.g., Qin and Lawless (1994), Kitamura (2001), Newey and Smith (2004)), the EL estimator based on (5.12) is defined as

$$\hat{\theta} \equiv \arg\min_{\theta\in\Theta} \sup_{\lambda\in\mathbb{R}^p} \mathcal{R}(\theta, \lambda),$$

where

$$\mathcal{R}(\theta, \lambda) \equiv \sum_{n=1}^{N} \log\left(1 + \lambda' g(x_n, \theta)\right) \tag{5.13}$$

and λ is a vector of Lagrangian multipliers.

[18] Indeed, NLLS estimator from directly minimizing $Q(\theta)$ will be inconsistent, so LOV instead minimize

$$Q_N(\theta) = (1/N)\,\Sigma_1^N \left(b_n - \bar{X}(z_n, \theta)\right)^2 - (1/N)\,\Sigma_1^N (b_n - m(z_n, \theta))^2 - \Delta(\theta)$$

where $\Delta(\theta) = (1/N)\,\Sigma_1^N E\left[var\left(\bar{X}(z_n, \theta)\right)\right]$.

However, a problem in empirical likelihood estimation of θ by minimizing (5.13), as Laffont *et al.* (1995) also encountered in NLLS estimation, is that $g(x, \theta_0)$, in particular $m(z, \theta)$, is intractable and not in an explicit form so that we cannot calculate its sample analogue, nor we can get its derivative. In a different context, a similar case is also considered by Ai and Chen (2003) who use sieve method to estimate $g(\cdot)$ which may contains unknown functions. Another situation in which we cannot use the moment indicator $g(\cdot)$ directly, is that sometimes $g(\cdot)$ is not continuous in θ, but usual empirical likelihood estimation assumes that $g(\cdot)$ should be continuous and differentiable in the parameter of interest, so that we can demonstrate the consistency of EL estimator. (see, e.g., assumption 1 of Newey and Smith (2004)). Parente and Smith (2008) discuss another example of this non-smooth case, where $g(\cdot)$ is not even differentiable. To summarize these situations we list the following cases.

Case 1 *$g(\cdot)$ is discontinuous in θ.*

Example 12 *McFadden (1989) considered estimation of discrete response model. Suppose we have obtained the model like*

$$y_i = I(\beta x_i + \varepsilon_i > 0) \tag{5.14}$$

where $I(\cdot)$ is the indicator function and ε_i is i.i.d with density $p(\varepsilon)$. So we have the moment conditions $E[g(x, \beta)] \equiv E[y_i - I(\beta x_i + \varepsilon_i > 0)]$ and the GMM estimator $\hat{\beta}$ is based on the following sample analogue:

$$\hat{g}(x, \beta) = \frac{1}{N} \sum_{i=1}^{N} [y_i - I(\beta x_i + \varepsilon_i > 0)].$$

Problems arises because $\hat{g}(x, \beta)$ is not continuous in β.

Our auction model provides an example for the second case due to the high nonlinearity of the equilibrium strategy.

Case 2 *Computation of $g(\cdot)$ is infeasible.*

Pakes and Pollard (1989) considered simulating a good estimate $\tilde{g}(\cdot)$ of $g(\cdot)$ when the expectation of $g(x, \theta_0)$ is difficult to evaluate, including the case that $g(x, \theta)$ is nonsmooth or even discontinuous. Specifically, if we let $G_n(\theta)$ be a simulation of $E[g(x, \theta)]$ and $\tilde{\theta}$ be the GMM estimator based on $G_n(\theta)$, then the conditions under which $\tilde{\theta}$ converges to θ_0 are described in the following theorem.

Theorem 22 *$\tilde{\theta}$ converges in probability to θ_0 if*

a. $\left\| G_n\left(\tilde{\theta}\right)\right\| \leq \inf_{\theta\in\Theta} \left\|G_n\left(\theta\right)\right\| + o_p(1)$

b. $G_n\left(\theta_0\right) = o_p(1)$

c. $\sup_{\|\theta-\theta_0\|>\delta} \left\|G_n\left(\theta\right)\right\|^{-1} = O_p(1), \quad \forall\delta > 0.$

where $\|\cdot\|$ is some norm depending on θ.

Proof. See Pakes and Pollard (1989). ■

Remarks

The intuition for these conditions is to require the simulation $G_n\left(\cdot\right)$ be as close to $E\left[g\left(x,\theta\right)\right]$ as possible. Specifically,

a. $G_n\left(\cdot\right)$ evaluated at the estimator $\tilde{\theta}$ cannot be much bigger than the smallest value of $G_n\left(\theta\right)$ in Θ.

b. $G_n\left(\cdot\right)$ evaluated at the true parameter θ_0 cannot be much bigger than zero.

c. $G_n\left(\cdot\right)$ evaluated outside some neighborhood of θ_0 should be large.

Based on this approach, Laffont *et al.* (1995) show that the optimal bid b^w can be written as the expectation of the maximum of the second highest bid and the reservation price conditional on the highest private value, which provides a way to simulate the first moment of b^w.

Proposition 12 (Laffont *et al.* (1995)) *Given the number of bidders I, the reservation price u_0, and $F\left(\cdot\right)$, then for $u \geq u_0$, the optimal bidding strategy $\beta\left(u\right)$ can be expressed as*

$$\beta\left(u\right) = E\left[\max\left(u^{(I-1)}, u^0\right)|u^{(I:i)} = u\right]. \tag{5.15}$$

Proof. The result is obtained by combining Milgrom and Weber (1982) Theorem 14 and the equilibrium strategy (5.2), noticing that the conditional cdf. of $u^{((I-1):i)}$ given $u^{(I:i)} = u$ is $\left[F\left(u\right)/F\left(u^{(I:i)}\right)\right]^{I-1}$. ■

Taking expectation of (5.15) with respect to $u^{(I:i)}$, we have

$$E\left(b^w\right) = E\left[\max\left(u^{(I-1)}, u^0\right)\right]. \tag{5.16}$$

(5.16) can be viewed as an integral with respect to the density of $u^{((I-1):i)}$, which is a function of $u_1, ..., u_N$, independently drawn from $F\left(\cdot\right)$. So (5.16) can be written as

$$E\left(b^{w}\right)=\int_{v_1}...\int_{v_n}\max\left(u_{(N-1)},p^{0}\right)f\left(u_1\right)...f\left(u_n\right)du_1...du_n. \tag{5.17}$$

Since in the above integral, the private values u_i is not observable, $E\left(b^{w}\right)$ cannot be obtained directly. Following Laffont *et al.* (1995), we use importance sampling methods to get an estimator of $E\left(b^{w}\right)$, through sampling u_i from another distribution. In the next section we give a brief introduction to the importance sampling method.

Importance sampling

Importance sampling is a simulation method which is useful to estimate an integral about a probability distribution from a different distribution. Suppose we want to evaluate the integral

$$E_p\left[g(x)\right]=\int_D g(x)p(x)dx$$

where $g(x)$ is a function of x and $p(x)$ is the density of x. If it is difficult to sample from $p(x)$ [19], we can choose another probability distribution $Q(x)$ with density $q(x)$, which is called the importance function [20]and has the same support as $p(x)$, and transform $E_p\left[g(x)\right]$ as

$$E_p\left[g(x)\right]=\int g(x)\frac{p(x)}{q(x)}q(x)dx=E_q\left[g(x)w(x)\right], \tag{5.18}$$

where $w(x)=p(x)/q(x)$ is called the importance weight (also inverse likelihood ratio). Note that $w\left(x\right)$ is always positive, $E_q\left[w(x)\right]=1$, and this weight function reflects the important regions of the sampling space. A special case is that $q(x)=p(x)$, when $w\left(x\right)=1$.

(5.18) motivates an unbiased estimator for $E_p\left[g(x)\right]$ by sampling S independent values from $Q(x)$ and calculating

$$\frac{1}{S}\sum_{s=1}^{S}g(x_{ns})w(x_{ns}) \tag{5.19}$$

[19] e.g., Owen and Zhou (2000) considered the case that $g(x)$ is a *spiky* function, which means that the variance of $g(x)$ is may depend on a subset of D having relatively small probability under sampling from $q(x)$.

[20] $q(x)$ is also called the *importance sampling density*, *proposal density*, or *instrumental density* as we use it as an instrument to obtain information about the integral.

as simulated value of $g(x)w(x)$. Hence $E_p\left[g(x)\right]$ can be estimated by

$$\tilde{E}_p\left[g(x)\right] = \frac{1}{NS}\sum_{n=1}^{N}\sum_{s=1}^{S} g(x_{ns})w(x_{ns}). \tag{5.20}$$

Note that $g(x)w(x)$ is an unbiased estimator of $E_p\left[g(x)\right]$ by construction, with expectation taken with respect to $q(x)$. It is interesting to check the expectation of $g(x)w(x)$ with respect to $p(x)$. Generally it will depend on the choice of $q(x)$, but in some circumstances this expectation can be bounded by a function that does not depend on the choice of $q(x)$. The following result will be useful later:

Proposition 13 *Assume that $g(x)$ is nonnegative and the importance weight $w(x) = p(x)/q(x)$ is infinitely integrable, i.e., $E_p^{\infty}\left[w(x)\right] < M$, where M is finite, then $E_p\left[g(x)w(x)\right]$ is also bounded, in particular*

$$E_p\left[g(x)w(x)\right] \leq E_p\left[g(x)\right] M. \tag{5.21}$$

Proof. The result is directly from the Hölder inequality:

$$\begin{aligned} E_p\left[g(x)w(x)\right] &= \int g(x)\frac{p(x)}{q(x)}p(x)dx \\ &\leq \left(\int g(x)p(x)dx\right)\left\|w(x)\right\|_{\infty} \\ &\leq E_p\left[g(x)\right] M, \end{aligned}$$

where $\left\|\cdot\right\|_{\infty}$ denotes the norm in L^{∞} space. ■

The following example shows importance sampling can also smooth discontinuous moment conditions.

Example 13 *Let $u = \beta x_i + \varepsilon_i$ and $p\left(u\left|x, \beta\right.\right)$ be the conditional density of u given x, β, through change of variables we have*

$$\begin{aligned} E\left[I(\beta x_i + \varepsilon_i > 0)\right] &= \int I(\beta x_i + \varepsilon_i > 0)p\left(\varepsilon\right)d\varepsilon \\ &= \int I(u > 0)p\left(u\left|x, \beta\right.\right)du \\ &= \int \frac{I(u > 0)p\left(u\left|x, \beta\right.\right)}{q(u)}q(u)du, \end{aligned}$$

where $q(u)$ is an arbitrary non zero density with a support which contains that *of u. Now we can draw S independent samples $u_1, ...u_S$ from $q(u)$ and construct*

$$\tilde{E}\left[I(\beta x_i + \varepsilon_i > 0)\right] \equiv \frac{1}{S}\sum_{s=1}^{S} \frac{p\left(u_s \left| x, \beta\right.\right)}{q(u_s)}.$$

Note $\tilde{E}\left[I(\beta x_i + \varepsilon_i > 0)\right]$ is an unbiased estimator of $E\left[I(\beta x_i + \varepsilon_i > 0)\right]$ since

$$\begin{aligned} E\left[\tilde{E}\left[I(\beta x_i + \varepsilon_i > 0)\right]\right] &= E\left[\frac{1}{S}\sum_{s=1}^{S} \frac{I(u_s > 0) p\left(u_s \left| x, \beta\right.\right)}{q(u_s)}\right] \\ &= \int \frac{I(u > 0) p\left(u \left| x, \beta\right.\right)}{q(u)} q(u) du \\ &= E\left[I(\beta x_i + \varepsilon_i > 0)\right]. \end{aligned}$$

The critical thing in $\tilde{E}\left[I(\beta x_i + \varepsilon_i > 0)\right]$ is that it is continuous in β. Thus through importance sampling we have obtained a smooth approximation of the discrete moment conditions.

5.3.2 Large Sample Theory

Following Laffont *et al.* (1995), we simulate the expectation of the optimal bid $E\left(b^w\right)$ according to (5.17):

$$\begin{aligned} E\left(b^w\right) &= \int_{u_1} ... \int_{u_n} \max\left(u_{(I-1)}, u_0\right) \frac{f\left(u_1\right)...f\left(u_n\right)}{q\left(u_1\right)...q\left(u_n\right)} \\ &\quad \times q\left(u_1\right)...q\left(u_n\right) du_1...du_n. \end{aligned} \tag{5.22}$$

For each $n = 1, ..., N$, we draw S independent samples from another distribution $q(\cdot)$, each of size I, denoted $u_n^{s1}, ..., u_n^{sI}$, $s = 1, ..., S$. Hence by using (5.22) we can construct a estimator for $m\left(z_n, \theta\right)$ for each n :

$$\tilde{m}\left(z_n, \theta\right) = \frac{1}{S}\sum_{s=1}^{S} \max\left(u_n^{s(I-1)}, u_0\right) \frac{f\left(u_n^{s1}\right)...f\left(u_n^{sI}\right)}{q\left(u_n^{s1}\right)...q\left(u_n^{sI}\right)}. \tag{5.23}$$

Remarks.

a. *$\tilde{m}\left(z_n, \theta\right)$ is by construction an unbiased estimator for $m\left(z_n, \theta\right)$ even for $S = 1$, $E_q\left[\tilde{m}\left(z_n, \theta\right)\right] = m\left(z_n, \theta\right)$.*

b. *The simulations, hence the simulated moments $\tilde{m}\left(z_n, \theta\right)$, are not conditional on the observation b^w.*

c. *Let* $\kappa(\theta) \equiv \max\left(u_n^{s(I-1)}, u_0\right) f\left(u_n^{s1}\right)...f\left(u_n^{sI}\right)$ *, then* $var_q\left[\tilde{m}(z_n,\theta)\right] = \frac{1}{S}\sigma_\kappa^2$*, where*

$$\begin{aligned}
\sigma_\kappa^2 &= E_q\left[\frac{\kappa(\theta)}{q\left(u_n^{s1}\right)...q\left(u_n^{sI}\right)} - m(z_n,\theta)\right]^2 \qquad (5.24)\\
&= \int\left(\max\left(u_n^{s(I-1)}, u_0\right)\frac{f\left(u_n^{s1}\right)...f\left(u_n^{sI}\right)}{q\left(u_n^{s1}\right)...q\left(u_n^{sI}\right)} - m(z,\theta)\right)^2 q du\\
&= \int\left[\max\left(u_n^{s(I-1)}, u_0\right)\right]^2\frac{f^2}{q^2}q du - 2m\int\max\left(u_n^{s(I-1)}, u_0\right)\frac{f}{q}q du + m^2\int q du\\
&= \int\left[\max\left(u_n^{s(I-1)}, u_0\right)\right]^2\frac{f^2}{q}du - \left[m(z,\theta)\right]^2. \qquad (5.25)
\end{aligned}$$

Therefore, if $q = m(z,\theta)^{-1}\kappa(\theta)$*, then the variance will be zero, although this is not realistic since* θ *is as yet unknown. However, we can choose* $q(x)$ *which is of roughly the same shape as* $\kappa(\theta)$ *, i.e.,* $q(x)$ *is proportional to* $\kappa(\theta)$*:* $q(x) \propto \kappa(\theta)$*, so that the variance of the estimator can be as small as possible. (e.g., see Rubinstein (1981), Owen and Zhou (2000)).*

Now let

$$\tilde{g}(x_n,\theta) = b_n - \tilde{m}(z_n,\theta), \qquad (5.26)$$

$$\tilde{g}(\theta) \equiv \frac{1}{N}\sum_{n=1}^{N}\tilde{g}(x_n,\theta), \qquad (5.27)$$

$$\tilde{G} \equiv E\left[\frac{\partial\tilde{g}(x_n,\theta_0)}{\partial\theta}\right], \qquad (5.28)$$

and

$$\tilde{\Omega} \equiv E\left[\tilde{g}(x_n,\theta_0)'\,\tilde{g}(x_n,\theta_0)\right], \qquad (5.29)$$

and let their counterparts from $g(x_n,\theta)$ be defined analogously, and denoted without accent above, *e.g.*, $g(\theta) \equiv \frac{1}{N}\Sigma_{n=1}^{N}g(x_n,\theta)$. To apply the results of theorem 22 we define the empirical likelihood estimator $\tilde{\theta}$ as the solution to the following problem:

$$\tilde{\mathcal{R}}(\tilde{\theta},\tilde{\gamma}) \leq \min_{\theta}\sup_{\gamma\in\mathbb{R}^p}\tilde{\mathcal{R}}(\theta,\gamma) + o_p(N^{-1}), \qquad (5.30)$$

where

$$\tilde{\mathcal{R}}(\theta,\gamma) = \frac{1}{N}\sum_{n=1}^{N}\log(1+\gamma'\tilde{g}(x_n,\theta))$$

and γ is a vector of Lagrangian multipliers which is a function of θ implicitly defined through

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\tilde{g}\left(x_n,\theta\right)}{1+\gamma'\tilde{g}\left(x_n,\theta\right)}=0,$$

e.g., see Qin and Lawless (1994).

For the general asymptotic properties of empirical likelihood estimator, we make the following regularity assumption.

Assumption 14 ***a.*** *$\theta_0 \in int\left(\Theta\right)$, and Θ is a compact subset of $\mathbb{R}^p$.*

b. $E\left[\sup_{\theta\in\Theta}\left\|g\left(x,\theta\right)\right\|^{\epsilon}\right]<\infty$, $\forall\epsilon>2.$

c. $\Omega\equiv E\left[g\left(x_n,\theta_0\right)'g\left(x_n,\theta_0\right)\right]$ *is nonsingular.*

Furthermore, we need a smoothing condition for uniform convergence. Let the *simulation residual process* defined as

$$\omega\left(\theta\right)=\sqrt{N}\left(\tilde{g}\left(\theta\right)-E_q\left[\tilde{g}\left(x,\theta\right)\right]\right). \tag{5.31}$$

Assumption 15 *The process $\omega\left(\theta\right)$ is stochastically equicontinuous*[21] *at θ_0, i.e., for any $\epsilon>0$, there exists a neighborhood U of θ_0, which satisfies*

$$\sup_{\theta\in U}\left|\omega\left(\theta\right)-\omega\left(\theta_0\right)\right|\leq\varepsilon \qquad a.s$$

Although we have mentioned that θ_0 can be identified by our model under symmetric IPV setting, we make the following more specific assumption about identification of θ_0 through $g\left(\cdot\right)$.

Assumption 16 *For any $\delta>0$, $\sup_{\left\|\theta-\theta_0\right\|>\delta}\left\|g\left(\theta\right)\right\|^{-1}=O_p(N^{-1})$.*

The following theorem demonstrates the consistency of $\tilde{\theta}$, by checking similar conditions given in theorem 22.

Theorem 23 *Given assumption 14-16, we have the following results:*

1. $\sup_{\left\|\theta-\theta_0\right\|>\delta}\left\|\tilde{g}\left(\theta\right)\right\|^{-1}=O_p(N^{-1})$.

2. $\tilde{g}\left(\theta_0\right)=o_p(1)$

3. $\tilde{g}(\tilde{\theta})=o_p(1)$

[21] Detailed illustration of stochastically equicontinuous and uniform convergence can be found in e.g., Pollard (1984) chapter 7 or Newey (1991). Indeed, to make $\omega\left(\theta\right)$ be stochastically equicontinuous, we can choose a importance function $q(x)$ such that $\tilde{g}\left(x_n,\theta\right)$ is probably Lipschitz. See, e.g., lemma 3 of McFadden and Ruud (1994).

4. $\tilde{\mathcal{R}}(\theta_0, \bar{\gamma}) = O_p\left(N^{-1/2}\right)$, *where* $\bar{\gamma} = \arg\sup_\gamma \tilde{\mathcal{R}}(\theta_0, \gamma)$.
and then $\tilde{\theta}$ *converges in probability to* θ_0.

Proof. The first result is to say that $\tilde{g}(\theta)$ is big outside some neighborhood of θ_0, which is from the identification of θ_0. To see this, note that from triangle inequality we have

$$\begin{aligned} \sup_{\|\theta-\theta_0\|>\delta} \|\tilde{g}(\theta)\| &= \sup_{\|\theta-\theta_0\|>\delta} \|-g(\theta) - (\tilde{g}(\theta) - g(\theta))\| \\ &\geq \sup_{\|\theta-\theta_0\|>\delta} \|g(\theta)\| - \sup_{\|\theta-\theta_0\|>\delta} \|\tilde{g}(\theta) - g(\theta)\| \\ &\geq \sup_{\|\theta-\theta_0\|>\delta} \|g(\theta)\| - \sup_{\theta} \|\tilde{g}(\theta) - g(\theta)\|, \end{aligned}$$

given the assumption 15 of stochastic equicontinuity, $\sup_\theta \|\tilde{g}(\theta) - g(\theta)\| = o_p(1)$, and with assumption 16 we have $\sup_{\|\theta-\theta_0\|>\delta} \|\tilde{g}(\theta)\|^{-1} = O_p(N^{-1})$.

Secondly we follow the way of McFadden (1989), McFadden and Ruud (1994), where $\sqrt{N}\tilde{g}(\theta)$ is decomposed as

$$\sqrt{N}\tilde{g}(\theta) = A_N + [\omega(\theta) - \omega(\theta_0)] + B_N(\theta) + C_N(\theta) \tag{5.32}$$

where

$$A_N \equiv g(z, \theta_0) + \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \left(\tilde{g}(x_n, \theta_0) - E_q[\tilde{g}(x_n, \theta_0)]\right),$$

$$C_N(\theta) \equiv \frac{1}{\sqrt{N}} \sum_{n=1}^{N} g(x_n, \theta) - g(x_n, \theta_0),$$

$$B_N(\theta) \equiv \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \left(E_q[\tilde{g}(x_n, \theta)] - g(x_n, \theta)\right).$$

McFadden and Ruud (1994) have shown that $A_N = o_p(N^{1/2})$, with i.i.d assumption on the observations and simulations. Also note that $C_N(\theta_0) = o_p(N^{1/2})$, and $B_N(\theta_0) = 0$, so we have$\sqrt{N}\tilde{g}(\theta_0) = o_p(N^{1/2})+o_p(N^{1/2})$ and hence $\tilde{g}(\theta_0) = o_p(1)$.

To see the third results, a second order Taylor expansion of $\tilde{\mathcal{R}}(\theta, \gamma)$ around $\gamma = 0$ gives

$$\tilde{\mathcal{R}}(\theta, \gamma) = \gamma'(\tilde{g}(\theta)) - \frac{1}{2}\gamma' \left[\frac{1}{N} \sum_{n=1}^{N} \frac{\tilde{g}(x_n, \theta)\tilde{g}(x_n, \theta)'}{\left(1 + \dot{\gamma}'\tilde{g}(x_n, \theta)\right)^2}\right] \gamma, \tag{5.33}$$

where $\dot{\gamma}$ lies between 0 and γ. According to Lemma A1 and A2 of Newey and

Smith (2004) we have $\bar{\gamma} = O_p\left(N^{-1/2}\right)$ and $\frac{1}{\left(1+\dot{\gamma}'\tilde{g}(x_n,\theta)\right)^2} \leq -1/2$. Thus from (5.33) and result 1 we have

$$\begin{aligned}
\tilde{\mathcal{R}}\left(\theta_0,\bar{\gamma}\right) &\leq O_p\left(N^{-1/2}\right)o_p(1) + O_p\left(N^{-1}\right)\left(\frac{1}{N}\sum_{n=1}^{N}\tilde{g}\left(x_n,\theta\right)'\tilde{g}\left(x_n,\theta\right)\right) \\
&= o_p(N^{-1/2}) + O_p\left(N^{-1}\right) \\
&= O_p\left(N^{-1/2}\right).
\end{aligned}$$

Now from the definition of $\tilde{\theta}$ we have

$$\begin{aligned}
\tilde{\mathcal{R}}(\tilde{\theta},\tilde{\gamma}) &= O_p\left(N^{-1/2}\right)\tilde{g}\left(\tilde{\theta}\right) + O_p\left(N^{-1}\right) \qquad (5.34) \\
&\leq \min_{\theta}\sup_{\gamma\in\mathbb{R}^p}\tilde{\mathcal{R}}\left(\theta,\gamma\right) + o_p(N^{-1}) \\
&\leq \tilde{\mathcal{R}}\left(\theta_0,\bar{\gamma}\right) + o_p(N^{-1}) \\
&= O_p\left(N^{-1/2}\right).
\end{aligned}$$

Solving $\tilde{g}(\tilde{\theta})$ out of (5.34) gives

$$\left\|\tilde{g}(\tilde{\theta})\right\| = o_p(1). \qquad (5.35)$$

Then the following argument is similar to Pakes and Pollard (1989). By result 1 we have just proved, for arbitrary $\delta > 0$, there exists a bounded, positive constant M such that $\sup_{\|\theta-\theta_0\|>\delta}\left\|\tilde{g}\left(\theta\right)\right\|^{-1} < M$. On the other hand, since $\left\|\tilde{g}(\tilde{\theta})\right\|$ is $o_p(1)$, for N large enough $\left\|\tilde{g}(\tilde{\theta})\right\|^{-1} > M$ with probability approaching one. Hence

$$\sup_{\|\theta-\theta_0\|>\delta}\left\|\tilde{g}\left(\theta\right)\right\|^{-1} < M < \left\|\tilde{g}(\tilde{\theta})\right\|^{-1},$$

which implies $\tilde{\theta}$ must be within the neighborhood of θ_0 of radius δ, by noting that $\tilde{g}(\theta)$ is continuous. The convergence follows since δ can be arbitrary small. ■

Remarks

1. The consistency of $\tilde{\theta}$ does not depend on the choice of number of simulations S, although S does affect the asymptotic efficiency of $\tilde{\theta}$.

2. The result also holds if $\tilde{m}(z_n, \theta)$ is a biased estimator for $m(z_n, \theta)$ if

$$\sup_{\Theta} N^{1/2} |B| = o(1)$$

where $B \equiv E_q[\tilde{m}(z_n, \theta)] - m(z_n, \theta)$ is the simulation bias. See, e.g., McFadden (1989), who uses smoothed kernel simulator, which is biased.

Before we consider asymptotic normality of $\tilde{\theta}$, we look at the variance covariance matrix of $\tilde{\theta}$ with respect to f. Let $\tilde{\Sigma} \equiv var_f[\tilde{g}(x_n, \theta_0)]$.

Lemma 4 (Decomposition of Covariance Matrix) *Given i.i.d observations b_n and simulations for u_n,*

$$\sqrt{N}\tilde{g}(x_n, \theta_0) \xrightarrow{d} N\left(0, \tilde{\Sigma}\right) \tag{5.36}$$

with

$$\tilde{\Sigma} = \Sigma_m + \Sigma_S = \Sigma_m + \frac{1}{S}\Sigma_m,$$

where

$$\Sigma_m = var[g(x, \theta_0)], \quad \Sigma_S = E_p(var_q[\tilde{g}(x, \theta_0)]).$$

Proof. By the law of total variance we have

$$\begin{aligned} & Var[\tilde{g}(x_n, \theta)|u] \\ = \; & E(var_q[\tilde{g}(x, \theta_0)]|u) + var[E[g(x, \theta_0)]|u] \\ = \; & \Sigma_m + \frac{1}{S}\Sigma_m, \end{aligned}$$

where the second equality follows the law of iterated expectations and the fact that the estimator through importance sampling simulation is unbiased.

Note that under the i.i.d assumption and with the Lindberg-Levy central limit theorem we have $\sqrt{N}g(x_n, \theta_0) \xrightarrow{p} N(0, \Sigma_m)$. So

$$\begin{aligned} \sqrt{N}\tilde{g}(x_n, \theta_0) &= \sqrt{N}(b_n - m(x_n, \theta_0)) - \sqrt{N}(\tilde{m}(x_n, \theta_0) - m(x_n, \theta_0)) \\ &\xrightarrow{d} N\left(0, \left(\Sigma_m + \frac{1}{S}\Sigma_m\right)\right). \end{aligned}$$

■

Assumption 17 *$g(x, \theta)$ is differentiable at θ_0 and $G = E[\partial g(x, \theta_0)/\partial\theta]$ is of full rank.*

Theorem 24 *Given assumption 14-17,* $\sqrt{n}\left(\tilde{\theta}-\theta_0\right) \xrightarrow{d} N(0,V)$, *where*

$$V=\left(G'\tilde{\Sigma}^{-1}G\right)^{-1}.$$

Proof. First we show that $\sqrt{n}\left(\tilde{\theta}-\theta_0\right)$ is stochastically bounded. Since $\tilde{g}(\tilde{\theta})=o_p(1)$, hence $C_N\left(\hat{\theta}\right)=Op(1)$ and by expanding $C_N\left(\tilde{\theta}\right)$ we have

$$\begin{aligned} C_N\left(\tilde{\theta}\right) &= \sqrt{n}\left(\tilde{\theta}-\theta_0\right)\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial m\left(x_n,\theta_0\right)}{\partial\theta}+O\left(\tilde{\theta}-\theta_0\right)\right)\right) \\ &= O_p(1). \end{aligned}$$

With Assumption 17 and $\tilde{\theta}\xrightarrow{p}\theta_0$, we have $\sqrt{n}\left(\tilde{\theta}-\theta_0\right)=O_p(1)$.

Now we define

$$\left(\dot{\theta},\dot{\gamma}\right)=\arg\min_{\theta}\sup_{\gamma\in\mathbb{R}^p}\tilde{\mathcal{R}}\left(\theta,\gamma\right). \tag{5.37}$$

Let $G_n\left(\theta\right)=\partial\tilde{m}_n\left(\theta\right)/\partial\theta, G\left(\theta_0\right)=\frac{1}{N}\sum_{n=1}^{N}G_n\left(\theta_0\right), \tilde{\Omega}_n=\frac{1}{N}\sum_{n=1}^{N}\tilde{m}_n\left(\theta_0\right)\tilde{m}_n\left(\theta_0\right)'$. Expand the first order condition for the saddlepoint problem of (5.37) around θ_0 and $\gamma_0=0$:

$$\begin{aligned} \frac{\partial\tilde{\mathcal{R}}\left(\theta,\gamma\right)}{\partial\theta} &= 0=-\sum_{n=1}^{N}\frac{G_n\left(\dot{\theta}\right)'\gamma}{1+\dot{\gamma}'\left(b_n-\tilde{m}_n\left(\dot{\theta}\right)\right)} \\ &\simeq \frac{1}{N}\sum_{n=1}^{N}G_n\left(\theta_0\right)'\dot{\gamma}, \end{aligned} \tag{5.38}$$

$$\begin{aligned} \frac{\partial\tilde{\mathcal{R}}\left(\theta,\gamma\right)}{\partial\gamma} &= 0=-\sum_{n=1}^{N}\frac{\tilde{m}_n\left(\dot{\theta}\right)'}{1+\dot{\gamma}'\left(b_n-\tilde{m}_n\left(\dot{\theta}\right)\right)} \\ &\simeq -\tilde{g}\left(\theta_0\right)-\frac{1}{N}\sum_{n=1}^{N}G_n\left(\theta_0\right)\left(\theta-\dot{\theta}\right)+\tilde{\Omega}_n\dot{\gamma}, \end{aligned} \tag{5.39}$$

(5.38) and (5.39) imply

$$\sqrt{N}\left(\dot{\theta}-\theta\right)\simeq-\left(G\left(\theta_0\right)'\tilde{\Omega}_n^{-1}G\left(\theta_0\right)\right)^{-1}G\left(\theta_0\right)\tilde{\Omega}_n^{-1}\sqrt{N}\tilde{g}\left(\theta_0\right).$$

Note that from Lemma 4 we have

$$\sqrt{N}\tilde{g}\left(\theta_0\right) \xrightarrow{d} N\left(0, \tilde{\Sigma}\right).$$

Also from i.i.d assumption and unconditional simulation,

$$\frac{1}{N}\sum_{n=1}^{N} G_n\left(\theta_0\right) \xrightarrow{p} E\left[G_n\left(\theta_0\right)\right] = G.$$

So $\sqrt{n}\left(\dot{\theta} - \theta_0\right) \rightarrow N(0, V)$. Next we show $\tilde{\theta}$ and $\dot{\theta}$ are asymptotically equivalent. The definition of $\tilde{\theta}$ implies:

$$\tilde{\mathcal{R}}(\tilde{\theta}, \tilde{\gamma}) \leq \tilde{\mathcal{R}}(\dot{\theta}, \dot{\gamma}) + o_p\left(N^{-1}\right) \leq \tilde{\mathcal{R}}(\dot{\theta}, \tilde{\gamma}) + o_p\left(N^{-1}\right).$$

Then with the similar expansion as (5.34) we have

$$\begin{aligned} & O_p\left(N^{-1/2}\right)\tilde{g}\left(\tilde{\theta}\right) + O_p\left(N^{-1}\right) \\ \leq\; & O_p\left(N^{-1/2}\right)\tilde{g}\left(\dot{\theta}\right) + O_p\left(N^{-1}\right) + o_p\left(N^{-1}\right) \\ \Rightarrow\; & \tilde{g}\left(\tilde{\theta}\right) - \tilde{g}\left(\dot{\theta}\right) = O_p\left(N^{-1/2}\right). \end{aligned}$$

So $\tilde{g}(\tilde{\theta}) - \tilde{g}(\dot{\theta}) = o_p(1)$. Thus according to the continuity of $\tilde{g}$ we have $\tilde{\theta} = \dot{\theta} + o_p(1)$. ■

Remarks

It turns out that the asymptotic variance-covariance matrix of $\tilde{\theta}$ does not depend on the choice of importance function $q(\cdot)$, but on the number of simulations S. This is the case which MR called unconditional simulation. As S goes to infinity the disturbance of simulation vanishes, and thus $\tilde{\theta}$ is asymptotically equivalent to usual EL estimators.

These asymptotic results follows closely that of McFadden and Ruud (1994). In their paper they also get a consistent GMM estimator for θ_0 based on general simulations. The covariance matrix of their estimator is larger than usual GMM estimator due to simulations, which is slightly different from the covariance matrix of our EL estimator. However, both of our proofs aim to show that, the simulated moment indicator evaluated at the true parameter and at the estimator satisfies similar conditions indicated in the proof of theorem 3.1 of Pakes and Pollard (1989).

5.4 Numerical Results

In this section, simulation results are presented to check the performance of EL estimator. For the data generation processes, throughout the experiment, let the private value be the exponential distribution, *i.e.*,

$$F(u;\theta) = 1 - \exp\left(-u\theta\right), \quad f(u;\theta) = \theta \exp\left(-u\theta\right)$$

where $\theta = \frac{1}{4}$. We take the reservation price $u_0 = 0$, the number of bidders $I = 2$, and the number of independent repetition of auctions $N = 100$. Moreover, for each n, the observed bids (optimal bids in equilibrium) can be calculated through $F(u)$ according to (5.2) in Theorem 20:

$$b_i = u_i - \frac{u_i + 4\exp(-\frac{u_i}{4}) - 4}{1 - \exp(-\frac{u_i}{4})}.$$

And for each n we take $b_n = \max_{i=1,2} b_i$ as the winning bid. Then we simulate $m(z_n, \theta)$ through importance sampling indicated by (5.23), and the simulated moment indicator is $b_n - \tilde{m}(z_n, \theta)$. Specifically, we choose two importance functions $q(x) = \frac{1}{8}\exp(-\frac{\theta}{8})$, and $q(x) = \frac{1}{12}\exp(-\frac{\theta}{12})$ to compare. Also to check performance of the asymptotic variance according to the number of simulations S, we calculate $\hat{\theta}$ under $S = 300, S = 500, S = 1000$ respectively. At last, for the above procedure of DGP we repeat 500 times.

For computation of the empirical likelihood estimator we use Bruce Hansen's package,[22] whose algorithm is to separately evaluate the *inner loop* and the *outer loop*, *i.e.*, firstly to compute the log value of the profile likelihood at each θ, and then maximize it over θ. Also for comparison, we calculate 2-step GMM estimator from the simulated moments.

As the results in table 4 and 5 showing, the variance is decreasing as S increasing, *i.e.*, the randomness from the simulation will be counteracted by the number of simulations.

[22] See, http://www.ssc.wisc.edu/~bhansen/progs/progs_gmm.html

Table 4: $q(x) = \frac{1}{8}\exp(-\frac{\theta}{8})$

	$S = 300$	$S = 500$	$S = 1000$
θ_{EL}			
MEAN	0.2473	0.2545	0.2521
MEDIAN	0.2416	0.2561	0.2507
SD	0.0129	0.0112	0.0084
θ_{2-GMM}			
MEAN	0.2512	0.2388	0.2446
MEDIAN	0.2471	0.2413	0.2427
SD	0.0140	0.0121	0.0103

Table 5: $q(x) = \frac{1}{12}\exp(-\frac{\theta}{12})$

	$S = 300$	$S = 500$	$S = 1000$
θ_{EL}			
MEAN	0.2371	0.2530	0.2437
MEDIAN	0.2390	0.2519	0.2468
SD	0.0147	0.0122	0.0116
θ_{2-GMM}			
MEAN	0.2613	0.2452	0.2429
MEDIAN	0.2570	0.2490	0.2458
SD	0.0152	0.0141	0.0136

5.5 Concluding Remarks

We have presented EL estimation of first price auction models under symmetric IPV setting, as an example showing how to deal with moment condition which is intractable in empirical likelihood. Based on simulated first moment of the winning bid by importance sampling, our estimator for the parameter of the distribution of private values has the usual asymptotic properties such as consistency and asymptotic normality, but it is different in that the covariance matrix is larger with additional part $(1/S)\,\Sigma_m$, which represents the randomness from simulation.

We also mentioned that simulation by importance sampling can be used to smooth moment condition with discreteness in parameter. This is a different way from Parente and Smith (2008) approach. Rather than simulating the moment indicator, they put different assumption on it to ensure the EL estimator to have standard first order asymptotic properties.

It is important to note that these asymptotic results of our estimator rely heavily on i.i.d assumptions on observations and simulations, and for time series model our EL estimator may fail since the general conditions for uniform convergence and the law of large numbers will not be satisfied. So if we want to use EL by simulating moment conditions with dependent data through importance sampling, more assumptions on stochastic convergence (e.g., see Pollard (1984) and chapter 4 of Billingsley (1999)) should be added, and the choice of importance function should also be carefully considered, to make the simulated moments satisfy certain conditions. These are the directions of our further research.

Bibliography

[1] Ai. C., and X., Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions.," *Econometrica*, Vol.71, 1795-1844.

[2] Albano, G. L., and F. Jouneau (1998): "A Bayesian Approach to the Econometrics of First-price Auctions", *working paper.*

[3] Andrews, D. W. K. (1983): "First Order Autoregressive Processes and Strong Mixing", *Cowles Foundation Discussion Papers*, NO. 664.

[4] Andrews, D. W. K. (1984): "Non-Strong Mixing Autoregressive Processes", *Journal of Applied Probability*, 21, 930-934.

[5] Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica*, 59, 817-858.

[6] Andrews, D. W. K. and D. Pollard (1994): "An Introduction to Functional Central Limit Theorems for Dependent Stochastic Processes", *International Statistical Review*, 62, 119–132.

[7] Andrews, D. W. K. and J. H .Stock (2005): "Inference with Weak Instruments", *Cowles Foundation Discussion Papers*, NO. 1530.

[8] Athey, S. and P. A. Haile (2002): "Identification of Standard Auction Models", *Econometrica*, 70, 2107-2140.

[9] Athey, S. and P. A. Haile (2005): "Nonparametric Approaches to Auctions ", *Handbook of Econometrics*, Vol 6.

[10] Bahadur, R. (1960): "On the Asymptotic Efficiency of Tests and Estimators", *Sankhya*, 22, 229-252.

[11] Bahadur, R. (1967): “Rates of Convergence of Estimates and Test Statistics”, *Annals of Mathematical Statistics*, 38, 303–324.

[12] Baxter, R. J., Jain. N. C. and S. R. S. Varadhan (1991): “Some Familiar Examples For Which the Large Deviation principle Does Not Hold”, *Communications of Pure and Applied Mathematics*, 34, 911-923.

[13] Billingsley, P (1999): *Convergence of Probability Measures*, John Wiley and Sons.

[14] Birhchandani, G. (1988): “Reputation in Repeated Second-Price Auctions”, *Journal of Economic Theory*, 46, 97-119.

[15] Blackwell, D. (1953): “Equivalent Comparisons of Experiments”, *Annals of Mathematical Statistics*, 24, 265-272.

[16] Blum, J. R., Hanson, D. L., and L. H. Koopmans (1963) “On the Strong Law of Large Numbers For a Class of Stochastic Processes”, *Z. Wahrsch. verw. Gebiete*, 2, 1-11.

[17] Bradley, R. C. (2005): “Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions”, *Probability Surveys*, Vol. 2, 107-144.

[18] Brown. L.D. (1971): “Non-local Asymptotic Optimality of Appropriate Likelihood Ratio Test”, *Annals of Mathematical Statistics*, 42, 1206–1240.

[19] Bryc, W. (1992): “On the Large Deviation principle for Stationary Weakly Dependent Random Fields”, *Annals of Probability*, 20, 1004-1030.

[20] Bryc, W. and A., Dembo (1996): “Large Deviations and Strong Mixing”, *Annales de l’I. H.P* , section B, NO.4, 549-569.

[21] Campo, S., Guerre, E., Perrigne I and Q. Vuong (2000): “Semiparametric Estimation of First-Price Auctions with Risk Averse Bidders”, *working paper.*

[22] Chernoff, H. (1952): “A Measure of Asymptotic Efficiency For Tests of A Hypothesis Based On Sums of Observations”, *Annals of Mathematical Statistics*, 23, 493-507.

[23] David, H. A. (1981): *Order Statistics*, John Wiley and Sons.

[24] Dedecker, J., Doukhan, P., Lang G., Jose, R., Louhichi, S. and Prieur, C. (2007): *Weak Dependence*, Springer, New York.

[25] Dembo, A., and O., Zeitouni (1998): *Large Deviations Techniques and Applications*. Second Edition. Springer, New York.

[26] Deuschel, J. D. and D. W. Strook (1989): *Large Deviations*. Boston: Academic Press.

[27] DiCiccio, T., Hall, P. and J. Romano (1991): "Empirical likelihood is Bartlett-correctable", *Annals of Statistics*, 19, 1053-1061.

[28] Donald, G. S. S. and H. J. Paarsch. (1993): "Piecewise Pseudo-Maximum Likelihood Estimation in Empirical Models of Auctions", *International Economic Review*, Vol. 34, 121-148.

[29] Donald, G. S. S. and H. J. Paarsch. (1996): "Identification, Estimation, and Testing in Parametric Empirical Models of Auctions within the Independent Private Values Paradigm", *Econometric Theory*, 12, 517-567.

[30] Durrett, R. (1991): *Probability: Theory and Examples*. Duxbury Press.

[31] Ellis, R. S. (1985): *Entropy, Large Deviations and Statistical Mechanics*. Springer-Verlag, New York.

[32] Godambe, V. P. (1960): "An Optimal Property of Regular Maximum Likelihood Estimation", *Annals of Mathematical Statistics*, 31, 1208-1211.

[33] Guggenberger. P. and R. J. Smith (2003): "Generalized Empirical Likelihood Estimators and Tests under Partial, Weak and Strong Identification", *CEMMAP Working Paper*, CWP08/03.

[34] Guerre, M., I. Perrigne and Q. Vuong (2000): "Optimal Nonparametric Estimation of First-Price Auctions", *Econometrica*, Vol.68, 525-574.

[35] Haile, P. A., Hong, H. and M. Shum (2003): "Nonparametric Tests for Common Values In First-Price Sealed-Bid Auctions", *working paper*.

[36] Hall, P. and J. L. Horowitz (1996): "Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators", *Econometrica*, 64, 891-916.

[37] Han, C. and P. C. B. Phillips (2006): "GMM with Many Moment Conditions", *Econometrica*, 74, 147-192.

[38] Hansen, R.G (1985): "Empirical Testing of Auction Theory", *The American Economic Review*, Vol. 75, No.2, 156-159.

[39] Hesterberg, T. C. (1988): *Advances in Importance Sampling*, PhD dissertation, Stanford University.

[40] Hoeffding, W (1965): "Asymptotically Optimal Tests for Multinomial Distributions", *Annals of Mathematical Statistics*, 36, 369-408.

[41] Hsieh, H.K. (1979): "On Asymptotic Optimality of Likelihood Ratio Tests for Multivariate Normal Distributions", *Annals of Statistics*, 7, 592-598.

[42] Ibragimov, I. A (1962): "Some Limit Theorems for Stationary Processes", *Theory of Probability and Its Applications*, 7, 349-382.

[43] Kitamura, Y. (1997): "Empirical Likelihood with Weakly Dependent Processes", *Annals of Statistics*, 25, 2084-2102.

[44] Kitamura, Y. (2001): "Asymptotic Optimality of Empirical Likelihood For testing Moment Restrictions", *Econometrica*, 69, 1661-1672.

[45] Kitamura and Otsu (2005): "Minimax Estimation and Testing for Moment Condition Models via Large Deviations". *Draft.*

[46] Klemperer, P (2004): *Auctions: Theory and Practice* (The Toulouse Lectures in Economics), Princeton University Press, 2004.

[47] Koenker, R. and Machado, J. A. F. (1999): "GMM inference when the number of moment conditions is large", *Journal of Econometrics*, 93, 327-344.

[48] Kullback, S. (1958): *Information Theory and Statistics*, Dover Publications, New York.

[49] Kullback, S and R. A. Leibler (1951): "On information and sufficiency", *Annals of Mathematical Statistics*, 22, 79-86.

[50] Laffont, J. J. and E. Maskin (1980): "Optimal Reservation Price in the Vickrey Auction", *Economics Letters*, 6, 309-313.

[51] Laffont, J. J., H. Ossard and Q. Vuong (1995), "Econometrics of First-Price Auctions", *Econometrica*, 63, 953-980.

[52] Laffont, J. J., and Q. Vuong (1996): "Structural Analysis of Auction Data", *American Economic Review*, Papers and Proceedings, 86, 414—420.

[53] Lazar, N. A. and P. A. Mykland (1999): "Empirical Likelihood in the Presence of Nuisance Parameters", *Biometrika*, 86, 203-211

[54] LeCam, L. (1986): *Asymptotic Methods in Statistical Decision Theory.* Springer Verlag, New York.

[55] LeCam, L. and G. L. Yang (2000): *Asymptotics in Statistics: Some Basic Concepts.* Springer Verlag, New York.

[56] Lehmann, E. L. and G. Casella (1998): *Theory of Point Estimation.* Springer-Verlag, New York.

[57] Li, T. (2005): "Econometrics of first-price auctions with entry and binding reservation prices," *Journal of Econometrics*, 126,173-200.

[58] Li, T and Q. Vuong (1997): "Using All Bids in Estimation of First-Price Auctions," *Economic Letters*, 55, 321-325.

[59] Maskin, E., and J. Riley (1984): "Optimal Auctions with Risk-Averse Buyers", *Econometrica*, Vol.52, No. 6, 1473-1518.

[60] McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration", *Econometrica*, 57, 995-1026.

[61] McFadden, D. and Ruud, P. (1994): "Estimation by Simulation", *Review of Economics and Statistics*, 76, No.4, 591-608.

[62] Milgrom, P. and R. Webber (1982): "A Theory of Auctions and Competitive Bidding", *Econometrica*, 50, 1089-1122.

[63] Newey, W. K (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity", *Econometrica*, Vol.59, 1161-1167.

[64] Newey, W. K (2001): "Choosing the Number of Instruments", *Econometrica*, 69, 1161-1191.

[65] Newey, W. K. and F. Windmeijer (2005): “GMM with Many Weak Moment Conditions”, *CEMMAP Working Paper*, CWP18/05.

[66] Newey, W. K. and Smith, R.J. (2004): “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators”, *Econometrica*, 72, 219-255.

[67] Neyman, J. and E. L. Scott (1948): “Consistent Estimates Based on Partially Consistent Observations”, *Econometrica*, 16, 1-32.

[68] Noether, G. E. (1955): “On a Theorem of Pitman”, *Annals of Mathematical Statistics*, 26, 64-68.

[69] Owen, A. (1988): “Empirical Likelihood Ratio Confidence Intervals for a Single Functional”, *Biometrika*, 75, 237–249.

[70] Owen, A. (2000): *Empirical Likelihood*, London: Chapman & Hall, 2001.

[71] Owen, A and Y. Zhou (2000): “Safe and Effective Importance Sampling”, *Journal of the American Statistical Association*, Vol.95, No. 449, 135-143.

[72] Paarsch, J. H. (1992): “Deciding Between the Common and Private Value Paradigms in Empirical Models of Auctions” *Journal of Econometrics*, 51, 191-215.

[73] Paarsch, J. H and J. Robert (2003): “Testing Equilibrium Behaviour At First-Price, Sealed-Bid Auctions With Discrete Bid Increments”, *working paper.*

[74] Pakes, A., and D. Pollard (1989): “Simulation and the Asymptotics of Optimization Estimators”, *Econometrica*, 57, 1027-1057.

[75] Parente. P and R. J. Smith (2008): “GEL Methods for Nonsmooth Moment Indicators”, *CEMMAP Working Paper*, CWP19/08.

[76] Patrick, A. N., P. Buhlmann, P. and P.Doukhan (2002): “Weak Dependence Beyond Mixing and Asymptotics for Nonparametric Regression”, *Annals of Statistics*, 30, 397-430.

[77] Phillips, P. C. B. (1989): “Partially Identified Econometric Models”, *Econometric Theory,* 5, 181-240.

[78] Pitman, E. J. G (1949): *Lecture Notes on Nonparametric Statistical Inference.* Columbia University.

[79] Politis, D. N. and J. P. Romano (1992): "A General Resampling Scheme for Triangular Arrays of $\alpha-$mixing Random Variables with Application to the Problem of Spectral Density Estimation", *Annals of Statistics*, 20, 1985-2007.

[80] Pollard, D (1984): *Convergence of Stochastic Processes*, Springer-Verlag, New York.

[81] Pollard, D. (2003): *Asymptopia.* Draft Book.

[82] Puhalskii, A (1993): "On the Theory of Large Deviations", *Theory of Probability and Its Applications*, 38, 490–497.

[83] Puhalskii, A. (2006): *Large Deviation for Stochastic Processes*, Lecture Notes for LMS/EPRSC short course, Heriot-Watt University, Edingburgh.

[84] Puhalskii, A. and V. Spokoiny (1998): "On Large Deviation Efficiency in Statistical Inference", *Bernoulli*, 4, 203–272.

[85] Qin. J and J. Lawless (1994): "Empirical Likelihood and General Estimating equations", *Annals of Statistics*, 22, 300-325.

[86] Raghavachari, M. (1970): "On a Theorem of Bahadur on the Rate of Convergence of Test Statistics", *Annals of Mathematical Statistics*, 4, 1695–1699.

[87] Riley, J. G. and W. F. Samuelson (1981): "Optimal Auctions", *American Economic Review*, Vol. 71, 381-392.

[88] Rosenblatt, M. (1956): "A Central Limit Theorem and A Strong Mixing Condition", *Proceedings of the National Academy of Sciences*, 42, 43-47.

[89] Rothe, G (1981): "Some Properties of the Asymptotic Relative Pitman Efficiency", *Annals of Statistics*, 9, 663-669.

[90] Rubinstein, R. Y (1981): *Simulation and the Monte Carlo Method.* new York: Wiley.

[91] Rubinstein, R. Y and D. P. Kroese (2008): *Simulation and the Monte Carlo Method*, John Wiley and Sons.

[92] Sanov, I. N (1965): "On the Probability of Large Deviations of Random Variables", *Selected Translations of Mathematical Statistics and Probability I (English Translation)*, 213-244.

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

[94] Shikimi, T. (2002): "Large Deviations for Kernel-Type Empirical Distributions", *Statistics and Probability Letters*, 59, 23-28.

[95] Smith, R. J. (2004): "GEL Criteria for Moment Condition Models", *Cemmap working paper*, CWP19/04.

[96] Stock, J. H. and J. H. Wright (2000): "GMM with Weak Identification", *Econometrica*, 68,1055–1096.

[97] Strasser, H. (1985): *Mathematical Theory of Statistics : Statistical Experiments and Asymptotic Decision Theory.* Walter de Gruyter.

[98] Strasser, H. (1996): "Asymptotic Efficiency of Estimates for Models with Incidental Nuisance Parameters", *Annals of Statistics*, 24, 879-901.

[99] Strook, D. W. (1984): *An Introduction to the Theory of Large Deviations.* Springer-Verlag, Berlin.

[100] Swinkels, J. M. (1998): "Efficiency of Large Private Value Auctions", *Econometrica*, 69, 37-68.

[101] van. der. Varrt (2001): *Time Series.* Lecture Notes, University of Amsterdam.

[102] Varadhan, S. R. S. (1966): "Asymptotic Probabilities and Differential Equations", *Communications On Pure & Applied Mathematics*, 19, 261-286.

[103] Varadhan, S. R. S. (1984): *Large Deviations and Applications.* SIAM, Philadelphia.

[104] Ventzell, A. D. and M. I. Freidlin (1979): *Fluctuations in Dynamic Systems Caused by Small Random Perturbations.* Nauka, Moscow.

[105] Vickrey, M (1961): "Counterspeculation, Auctions, and Competitive Sealed Tenders", *Journal of Finance*, Vol. 16, 8-37.

[106] Walters, P. (1982): *An Introduction to Ergodic Theory.* Springer-Verlag, New York.

[107] Wilbert, C. M. K (1982): "Chernoff Efficiency and Deficiency", *Annals of Statistics*, 10, 583-594.

[108] Wolkonski, V. A. and Y. A. Rozanov (1959): "Some Limit Theorems for Random Functions", *Theory of Probability and Its Applications*, 4, 178-197.

[109] Zeitouni, O., and M. Gutman (1991): "On Universal Hypothesis Testing via Large Deviations", *IEEE Transactions on Information Theory*, 37, 285-290.