Asymptotic Optimality of Empirical Likelihood for Selecting Moment Restrictions

Taisuke Otsu*†

Yale University

Preliminary Draft

December 2005

Abstract

This paper studies large deviation optimal properties of the empirical likelihood sequential testing (ELST) procedures for selecting moment restrictions. Since moment selection problems have discrete parameter spaces, the Pitman-type local alternative approach is not very helpful. By the theory of large deviations, we analyze convergence rates of error probabilities under fixed distributions. We propose three optimal properties of the ELST procedures: (i) the generalized Neyman-Pearson optimality, (ii) the overestimation error optimality, and (iii) the minimax misclassification error optimality.

*E-mail address: taisuke.otsu@yale.edu.

†The author would like to thank Yuichi Kitamura for helpful comments.

1 Introduction

This paper studies large deviation optimal properties of some empirical likelihood-based procedures for selecting moment restrictions. Our problem is to select all correct moment restrictions from a fixed number of candidate moment restrictions, which possibly include incorrect moments. In a generalized method of moments (GMM) framework, Andrews (1999) developed several moment selection procedures, such as GMM-AIC, GMM-BIC, and GMM sequential testing procedures. Hong, Preston and Shum (2003) proposed generalized empirical likelihood (GEL) analogues of GMM-based moment selection procedures.[1] These papers derived the consistency of some moment selection procedures, i.e., the probability of choosing all correct moments converges to one. The purpose of the present paper is to provide another theoretical framework to evaluate moment selection procedures.

To this end, we focus on large deviation properties of moment selection procedures, i.e., convergence rates of error probabilities under fixed distributions. We particularly show some optimal properties of the empirical likelihood sequential testing (ELST) selection procedures proposed by Hong, Preston and Shum (2003). We split the moment selection problem into two estimation problems: (a) estimation of the number of correct moments, and (b) estimation of the set of correct moments given the number of correct moments. In practice, the implementation of the above selection procedures solves these problems at the same time. For the problem (a), we face two kinds of errors: overestimation and underestimation for the number of correct moments. Overestimation (resp. underestimation) causes inconsistency (resp. inefficiency) for the parameter estimator. For the problem (b), we face misclassification error caused by choosing an incorrect set of moments. Note that these estimation problems have discrete parameter spaces, and the conventional Pitman-type local alternative approach is not very useful. Instead, by using large deviation theory, we analyze the convergence rates of the above error probabilities under some fixed distributions. The large deviation properties can provide additional criteria to compare moment selection procedures beyond consistency.

We propose three optimal properties of the ELST procedures: (i) the generalized Neyman-Pearson optimality, (ii) the overestimation error optimality, and (iii) the minimax misclassification error optimality. (i) and (ii) are about estimation of the number of correct moments. (i) treats the convergence rates of the underestimation and overestimation probabilities as if those are the type I and type II error probabilities in hypothesis testing, i.e., under some restriction on the convergence rate of the underestimation probability, the ELST procedures attain the

[1] Andrews and Lu (2001) and Hong, Preston and Shum (2003) investigated model selection procedures. This paper focuses on the moment selection problem.

optimal convergence rate of the overestimation probability. (ii) considers a general class of moment selection procedures which contain several existing procedures, and says that the optimal convergence rate of the overestimation probability is attained by the ELST procedures. (iii) is about estimation of the set of correct moments given the number of correct moments, and says that the ELST procedures attain the lower bound of the worst convergence rate of the misclassification probability.

Recently, several alternatives to the GMM have been developed, such as continuous updating GMM (Hansen, Heaton and Yaron (1996)), empirical likelihood (Owen (1988) and Qin and Lawless (1994)), and exponential tilting (Kitamura and Stutzer (1997) and Imbens, Spady and Johnson (1998)). Newey and Smith (2004) proposed GEL which contains these alternatives as special cases and showed desirable higher order properties of the GEL estimator. Compared to higher order analysis which focuses on higher order local properties, large deviation analysis focuses on the first order but global properties of statistical decisions. Kitamura (2001) showed the generalized Neyman-Pearson optimality of the empirical likelihood overidentifying restriction test.

In the context of information theory, there are several applications of large deviation theory to model selection problems, such as Merhav, Gutman and Ziv (1989), Finesso, Liu and Narayan (1996) (for Markov chain order estimation), Merhav (1989), Chambaz (2006) (for parametric model selection), Khudanpur and Narayan (2002), and Gassiat and Boucheron (2003) (for hidden Markov order estimation). This paper extends these results to the moment selection problem. Compared to the previous results, where finite sample spaces and parametric models are assumed, several technical difficulties arise in our continuous sample space and semiparametric setup.

This paper is organized as follows. Section 2 introduces our basic setup. Section 3 presents main results. Section 4 contains simulation results. Section 5 concludes.

2 Setup

2.1 Moment Selection Problem

We consider the moment selection problem of Andrews (1999). Let $\{x_i\}_{i=1}^n$ be an iid sequence of $d \times 1$ random vectors drawn from an unknown probability measure P^o and $\theta \in \Theta \subset \mathbb{R}^p$ be a $p \times 1$ vector of unknown parameters. Suppose that we have an $M \times 1$ vector of candidate moment functions $g : \mathbb{R}^d \times \Theta \to \mathbb{R}^M$, where $M \in \mathbb{N}$ is a known constant satisfying $p < M < \infty$. If we assume that all moment restrictions by g are correct, then the model is written as

$$E\left[g\left(x_i, \theta^o\right)\right] = \int g\left(x, \theta^o\right) dP^o = 0 \tag{1}$$

for some $\theta^o \in \Theta$, and we can estimate the model by GMM or GEL. This paper considers the case where some moment restrictions are incorrect and (1) does not hold in some elements. To avoid inconsistent estimators for θ^o, we need to choose correct moments from the set of M moment functions. Also, to avoid inefficient estimators, we need to choose *all* correct moments.

Our notation closely follows that of Andrews (1999). Let $c = (c_1, \ldots, c_M)' \in \mathbb{R}^M$ be a moment selection vector for $g = (g_1, \ldots, g_M)'$, that is

$$c_j = \begin{cases} 0, & \text{if } g_j \text{ is not selected} \\ 1, & \text{if } g_j \text{ is selected} \end{cases},$$

for $j = 1, \ldots, M$. The space for c is

$$C = \left\{ c \in \mathbb{R}^M : c_j = 0 \text{ or } 1 \text{ for } j = 1, \ldots, M \right\}.$$

Let $|c| = \sum_{j=1}^M c_j$ be the number of moments selected by c, and $g^c(x, \theta)$ be the $|c| \times 1$ vector of selected moments by c. Define $c^o(\theta) = (c_1^o(\theta), \ldots, c_M^o(\theta))' \in \mathbb{R}^M$ as

$$c_j^o(\theta) = \begin{cases} 0, & \text{if} \quad E[g_j(x_i, \theta)] \neq 0 \\ 1, & \text{if} \quad E[g_j(x_i, \theta)] = 0 \end{cases},$$

for $j = 1, \ldots, M$ and $\theta \in \Theta$, i.e., $c^o(\theta)$ indicates the set of correct moments at $\theta \in \Theta$. Let

$$\begin{aligned} \mathcal{Z}^o &= \{c \in C : c = c^o(\theta) \text{ for some } \theta \in \Theta\}, \\ \mathcal{MZ}^o &= \{c \in \mathcal{Z}^o : |c| \geq |c'| \text{ for all } c' \in \mathcal{Z}^o\}, \end{aligned}$$

i.e., $\mathcal{Z}^o$ is the set of valid selection vectors $c^o(\theta)$ at some $\theta \in \Theta$, and $\mathcal{MZ}^o$ is the set of valid selection vectors which maximize the number of selected moments in $\mathcal{Z}^o$. Similar to Andrews (1999), we impose the following identification condition.

Assumption 2.1. *$\mathcal{MZ}^o$ contains a single element c^o.*

We call c^o the "true selection vector." Our problem is to estimate c^o in the discrete parameter space C. It is clear that $\mathcal{MZ}^o \subseteq \mathcal{Z}^o$. If $c \notin \mathcal{Z}^o$, then c selects some incorrect moments. Thus, the GMM or GEL estimator by $g^c(x_i, \theta)$ is inconsistent for θ^o. If $c \in \mathcal{Z}^o \setminus \mathcal{MZ}^o$, then c selects correct but relatively small number of moments comparing to $|c'|$ for $c' \in \mathcal{MZ}^o$. Thus, the GMM or GEL estimator by $g^c(x_i, \theta)$ is less efficient than the estimator by $g^{c'}(x_i, \theta)$. Note that Assumption 2.1 implies $|c^o| > p$, i.e., the true model selected by c^o is overidentified (see Andrews (1999, p. 548)). To obtain consistent estimators for both c^o and θ^o, we need to add an identification condition for θ^o, i.e., $E\left[g^{c^o}(x_i, \theta)\right] = 0$ has a unique solution $\theta^o \in \Theta$. However, to discuss the properties of estimators only for c^o, the identification condition of θ^o is unnecessary.

2.2 Example: Choice of Instruments

A typical example of the moment selection problem is the choice of valid instruments for linear instrumental variable (IV) regression models. Let $\{y_i, w_i\}_{i=1}^n$ be a sequence of iid data, and $\{z_i\}_{i=1}^n$ be an $M \times 1$ iid sequence of candidate IVs. If all instruments z_i are valid, the moment restrictions of the linear IV regression model are written as

$$E\left[g\left(x_i, \theta^o\right)\right] = E\left[z_i\left(y_i - w_i'\theta^o\right)\right] = 0, \tag{2}$$

for $i = 1, \ldots, n$, where $x_i = \left(y_i, w_i', z_i'\right)'$ and $\theta \in \mathbb{R}^p$. We consider the case where some instruments z_i are invalid and (2) does not hold in some elements. Let $z_i = \left(z_{1i}', z_{2i}'\right)'$ with $z_{1i} \in \mathbb{R}^{M_1}$ and $z_{2i} \in \mathbb{R}^{M_2}$. Assume that

$$\begin{aligned} E\left[z_{1i}\left(y_i - w_i'\theta\right)\right] &= 0 \quad \text{for some } \theta \in \Theta, \\ E\left[z_{2i}\left(y_i - w_i'\theta\right)\right] &\neq 0 \quad \text{for all } \theta \in \Theta, \end{aligned}$$

i.e., z_{1i} are valid instruments and z_{2i} are invalid ones. If we include some elements of z_{2i} in IV estimation, the IV estimator becomes inconsistent. If we employ a strict subset of z_{1i} as instruments, the IV estimator is consistent but less efficient than the IV estimator by all elements of z_{1i}. In this case, the true moment selection vector c^o selects $g^{c^o}\left(x_i, \theta\right) = z_{1i}\left(y_i - w_i'\theta\right)$.

2.3 Empirical Likelihood-Based Moment Selection Procedures

Empirical likelihood is non/semi-parametric likelihood constructed from the moment restrictions $E\left[g\left(x_i, \theta\right)\right] = 0$, that is

$$L\left(\theta\right) = \sup_{\{p_i\}_{i=1}^n} \left\{ \prod_{i=1}^n p_i \;\middle|\; p_i > 0, \;\; \sum_{i=1}^n p_i = 1, \;\; \sum_{i=1}^n p_i g\left(x_i, \theta\right) = 0 \right\}. \tag{3}$$

Without the moment restrictions $\sum_{i=1}^n p_i g\left(x_i, \theta\right) = 0$, unconstrained empirical likelihood is obtained as

$$L^u = \sup_{\{p_i\}_{i=1}^n} \left\{ \prod_{i=1}^n p_i \;\middle|\; p_i > 0, \;\; \sum_{i=1}^n p_i = 1 \right\} = n^{-n}.$$

We consider testing the overidentifying restriction:

$$H_0 : E\left[g\left(x, \theta\right)\right] = 0 \text{ for some } \theta \in \Theta, \quad H_1 : E\left[g\left(x, \theta\right)\right] \neq 0 \text{ for all } \theta \in \Theta. \tag{4}$$

As a test statistic, we can employ the empirical likelihood ratio:

$$L = -2\left(\sup_{\theta \in \Theta} \log L\left(\theta\right) - \log L^u\right) = \inf_{\theta \in \Theta} \max_{\gamma \in \mathbb{R}^M} 2 \sum_{i=1}^n \log\left(1 + \gamma' g\left(x_i, \theta\right)\right). \tag{5}$$

If the conditioning set of (3) is empty, set $L = \infty$. Qin and Lawless (1994) showed that under some regularity conditions $L \xrightarrow{d} \chi^2_{M-p}$ if H_0 is true.

We now introduce an information theoretic interpretation of empirical likelihood. Let $\mathcal{M}$ be the space of probability measures on the Borel σ-field $\left(\mathbb{R}^d, \mathcal{B}^d\right)$. Define

$$\mathcal{P}(\theta) = \left\{ P \in \mathcal{M} : \int g(x,\theta)\, dP = 0 \right\}, \quad \mathcal{P} = \bigcup_{\theta \in \Theta} \mathcal{P}(\theta).$$

Using this notation, (4) is written as $H_0 : P \in \mathcal{P}$ and $H_1 : P \notin \mathcal{P}$. The relative entropy (or Kullback-Leibler information criterion) for measures P and Q is defined as

$$\begin{aligned} I(P\|Q) &= \int \log\left(\frac{dP}{dQ}\right) dP \quad \text{if } P \ll Q \\ &= \infty \quad \text{otherwise.} \end{aligned}$$

Let μ_n be the empirical measure of $\{x_i\}_{i=1}^n$. From e.g. Borwein and Lewis (1993), it is known that (3) is equivalent to

$$\inf_{P \in \mathcal{P}} I(\mu_n \| P). \tag{6}$$

Thus, the empirical likelihood ratio test of H_0 reduces to the decision rule:

$$\text{reject } H_0 \text{ if } \inf_{P \in \mathcal{P}} I(\mu_n \| P) > C, \tag{7}$$

for some constant $C \in (0, \infty)$. Kitamura (2001) investigated the large deviation optimality of (7). This paper extends the above testing framework to the moment selection problem.

We introduce the following notation:

$$\begin{aligned} \mathcal{P}^c(\theta) &= \left\{ P \in \mathcal{M} : \int g^c(x,\theta)\, dP = 0 \right\}, \\ \mathcal{P}^c &= \bigcup_{\theta \in \Theta} \mathcal{P}^c(\theta), \quad \mathcal{P}_m = \bigcup_{c \in C_m} \mathcal{P}^c, \quad C_m = \{c \in C : |c| = m\}. \end{aligned}$$

$\mathcal{P}^c$ is a set of measures which satisfy $E\left[g^c(x_i,\theta)\right] = 0$ at some $\theta \in \Theta$, and $\mathcal{P}_m$ is a set of measures which satisfy some m moments. Note that $\mathcal{P}_M \subseteq \mathcal{P}_{M-1} \subseteq \cdots \subseteq \mathcal{P}_{p+1} \subseteq \mathcal{P}_p = \mathcal{M}$. Let $\mathcal{P}_{M+1}$ be the empty set by convention. We consider the following empirical likelihood-based moment selection procedures by Hong, Preston and Shum (2003).

Definition 2.1 (Empirical likelihood sequential testing procedures).

(i) Estimation of the number of correct moments: *Find*

$$\begin{aligned} \hat{m}_d &= \max\left\{ j : \inf_{P \in \mathcal{P}_j} I(\mu_n \| P) \le \eta_{j,n} \right\}, \qquad (8) \\ &= \max\left\{ j : \min_{c \in C_j} \inf_{P \in \mathcal{P}^c} I(\mu_n \| P) \le \eta_{j,n} \right\} \quad \textit{(downward testing)} \end{aligned}$$

i.e., start by $j = M$ and carry out the empirical likelihood ratio test for $H_0^j : P \in \mathcal{P}_j$ with progressively smaller j until H_0^j is accepted. Or compute

$$\begin{aligned} \hat{m}_u &= \min\left\{ j : \inf_{P \in \mathcal{P}_j} I(\mu_n \| P) > \eta_{j,n} \right\} - 1, \\ &= \min\left\{ j : \min_{c \in C_j} \inf_{P \in \mathcal{P}^c} I(\mu_n \| P) > \eta_{j,n} \right\} - 1 \quad \textit{(upward testing)} \end{aligned} \tag{9}$$

i.e., start by $j = p + 1$ and carry out the empirical likelihood ratio test for $H_0^j : P \in \mathcal{P}_j$ with progressively larger j until H_0^j is rejected.

(ii) Estimation of the selection vector: *Given $\hat{m} = \hat{m}_d$ or $\hat{m}_u$, find*

$$\hat{c} = \arg\min_{c \in C_{\hat{m}}} \left\{ \inf_{P \in \mathcal{P}^c} I(\mu_n \| P) \right\}, \tag{10}$$

i.e., find the selection vector which minimizes the empirical likelihood ratio $\inf_{P \in \mathcal{P}^c} I(\mu_n \| P)$ *in $C_{\hat{m}}$.*

Since $\inf_{P \in \mathcal{P}_{\hat{m}}} I(\mu_n \| P) = \min_{c \in C_{\hat{m}}} \{\inf_{P \in \mathcal{P}^c} I(\mu_n \| P)\}$, the second step is redundant in practice. $\hat{c}$ is obtained as a by-product of the implementation of $\hat{m}$. We analyze theoretical properties of $\hat{m}$ and $\hat{c}$ separately. Under some regularity conditions, Hong, Preston and Shum (2003) derived the consistency of $\hat{m}$ and $\hat{c}$. This paper studies large deviation properties of $\hat{m}$ and $\hat{c}$.

2.4 Large Deviations

We consider the following large deviation error probabilities for $\hat{m}$ and $\hat{c}$:

$$\begin{array}{rlll} P^n(\hat{m} < m) & \text{for each } P \in \mathcal{P}_m \setminus \mathcal{P}_{m+1} \text{ or } \mathcal{P}_m & \text{(underestimation)} & \\ P^n(\hat{m} > m) & \text{for each } P \in \mathcal{P}_m \setminus \mathcal{P}_{m+1} & \text{(overestimation)} & (11) \\ P^n(\hat{c} \neq c) & \text{for each } P \in \mathcal{P}^c & \text{(misclassification)} & \end{array}$$

where P^n is the n-fold product measure of P. If $\hat{m}$ and $\hat{c}$ are consistent, these probabilities converge to zero under fixed P. Since the spaces for m and c are discrete, the Pitman-type local alternative approach is not useful. Instead, by using large deviation theory, we analyze the convergence rates of the error probabilities in (11). Since $\hat{m}$ and $\hat{c}$ are defined as decisions based on the empirical measure μ_n (see Definition 2.1), the large deviation properties of $\hat{m}$ and $\hat{c}$ can be analyzed by those of μ_n. For our purpose, Sanov's theorem (e.g., Deuschel and Stroock (1989, Theorem 3.2.17)) is helpful.

Theorem 2.1 (Sanov). *Let Σ be the Polish space (i.e., complete separable metric space), $\mathcal{M}(\Sigma)$ be the space of measures on Σ endowed with the Lévy metric,*[2] *and $P \in \mathcal{M}(\Sigma)$. Then*

$$\limsup_{n\to\infty} \frac{1}{n} \log P^n (\mu_n \in \mathcal{G}) \leq - \inf_{Q\in\mathcal{G}} I(Q\|P),$$

for all closed sets $\mathcal{G} \subset \mathcal{M}(\Sigma)$, and

$$\liminf_{n\to\infty} \frac{1}{n} \log P^n (\mu_n \in \mathcal{H}) \geq - \inf_{Q\in\mathcal{H}} I(Q\|P),$$

for all open sets $\mathcal{H} \subset \mathcal{M}(\Sigma)$.

Sanov's theorem says that the large deviation probabilities of μ_n are characterized by the relative entropy I. Another powerful large deviation result is Stein's lemma, which is first mentioned in Stein's unpublished work. We use a practical version of the lemma proposed by Bahadur, Zabell and Gupta (1980, Theorem 2.1).

Lemma 2.1 (Stein). *Let P and Q be probability measures on the Borel σ-field $\left(\mathbb{R}^d, \mathcal{B}^d\right)$, and $\{A_n\}_{n\in\mathbb{N}}$ be a sequence of measurable sets. Then $\liminf_{n\to\infty} Q^n(A_n) > 0$ implies that*

$$\liminf_{n\to\infty} \frac{1}{n} \log P^n (A_n) \geq -I(Q\|P).$$

Stein's lemma provides lower bounds of large deviation probabilities in quite general setups. Note that the lower bound does not depend on $\{A_n\}_{n\in\mathbb{N}}$. We apply this lemma to derive a lower bound of the convergence rate of the overestimation probability in (11).

3 Main Results

3.1 Generalized Neyman-Pearson Optimality

We first derive the generalized Neyman-Pearson optimality of the ELST procedure for estimating the number of correct moments. In the original Neyman-Pearson framework, we minimize the type II error of a test under some restriction of the type I error. In the generalized Neyman-Pearson framework, we replace those errors with the large deviation analogues: minimize the

[2]The Lévy metric of P_1 and P_2 is defined as

$$\rho(P_1, P_2) \equiv \inf\left\{\epsilon > 0 : F_1(x - \epsilon\mathbf{e}) - \epsilon \leq F_2(x) \leq F_1(x + \epsilon\mathbf{e}) + \epsilon \text{ for all } x \in R^d\right\},$$

where F_1 and F_2 are the distribution functions of measures P_1 and P_2, respectively, and $\mathbf{e} \equiv (1, \ldots, 1)'$. The Lévy metric is compatible with the weak topology on $\mathcal{M}$ (e.g., Dembo and Zeitouni (1998, D.2)).

convergence rate of the type II error under some restriction of the convergence rate of type I error. This idea was originally proposed by Hoeffding (1965). Kitamura (2001) showed the generalized Neyman-Pearson optimality of empirical likelihood for testing overidentifying restrictions. We extend the generalized Neyman-Pearson framework to the moment selection problem. Consider a class of estimators that satisfy

$$\sup_{P\in\mathcal{P}_m} \limsup_{n\to\infty} \frac{1}{n} \log P^n \left(\tilde{m} < m\right) \leq -\eta, \tag{12}$$

for each $m = p+1, \ldots, M$, where $\eta_m \in (0, \infty)$ is a given constant. Among such estimators, an estimator is called the generalized Neyman-Pearson optimal if it minimizes

$$\limsup_{n\to\infty} \frac{1}{n} \log P^n \left(\tilde{m} > m\right), \tag{13}$$

uniformly over all $P \in \mathcal{M}$. If we apply the terminology of hypothesis testing, the underestimation and overestimation probabilities are treated as if those are the type I and type II error probabilities, respectively. Under the restriction of the convergence rate of the underestimation probability in (12), we minimize the convergence rate the overestimation probability in (13). We show that the ELST procedures $\hat{m} = \hat{m}_d$ or $\hat{m}_u$ with a fixed critical value $\eta_{j,n} = \eta$ has the generalized Neyman-Pearson optimality.

Let $\bar{A}$ denote the complement of a set A. We introduce the following partitions of the space of measures $\mathcal{M}$:

$$\begin{aligned} \Lambda_j &= \left\{ Q \in \mathcal{M} : \inf_{P\in\mathcal{P}_j} I\left(Q\|P\right) \leq \eta \right\} \quad \text{(acceptance region for } H_0^j : P \in \mathcal{P}_j) \\ \overline{\Lambda_j} &= \left\{ Q \in \mathcal{M} : \inf_{P\in\mathcal{P}_j} I\left(Q\|P\right) > \eta \right\} \quad \text{(rejection region)} \end{aligned} \tag{14}$$

for $j = p+1, \ldots, M$. Let Λ_{M+1} be the empty set by convention. Note that $\Lambda_M \subseteq \Lambda_{M-1} \subseteq \cdots \subseteq \Lambda_{p+1}$. From (14), the underestimation and overestimation probabilities of $\hat{m}_d$ and $\hat{m}_u$ are written as

$$\begin{aligned} P^n\left(\hat{m}_d < m\right) &= P^n\left(\mu_n \in \bigcap_{m\leq j\leq M} \overline{\Lambda_j}\right) = P^n\left(\mu_n \in \overline{\Lambda_m}\right), \\ P^n\left(\hat{m}_d > m\right) &= P^n\left(\mu_n \in \bigcup_{m+1\leq j\leq M} \Lambda_j\right) = P^n\left(\mu_n \in \Lambda_{m+1}\right), \\ P^n\left(\hat{m}_u < m\right) &= P^n\left(\mu_n \in \bigcup_{p+1\leq j\leq m} \overline{\Lambda_j}\right) = P^n\left(\mu_n \in \overline{\Lambda_m}\right), \\ P^n\left(\hat{m}_u > m\right) &= P^n\left(\mu_n \in \bigcap_{p+1\leq j\leq m+1} \Lambda_j\right) = P^n\left(\mu_n \in \Lambda_{m+1}\right), \end{aligned} \tag{15}$$

for each $m = p+1, \ldots, M$. Therefore, in this setup, $\hat{m}_d$ and $\hat{m}_u$ have the same error probabilities. Also note that the underestimation and overestimation probabilities are written by the large deviation probabilities of the empirical measure μ_n. Let $\|a\|$ be the Euclidean norm of a vector a. Similarly as Kitamura (2001), we introduce the following assumptions.

Assumption 3.1. *Assume that*

(i) $\{x_i\}_{i=1}^n$ *is an iid sequence,*

(ii) $P\{\sup_{\theta\in\Theta} \|g(x,\theta)\| = \infty\} = 0$ *for all* $P \in \mathcal{P}_{p+1}$,

(ii) *at each* $\theta \in \Theta$, $g(x,\cdot)$ *is continuous for all* $x \in \mathbb{R}^d$.

All assumptions are very mild. See Kitamura (2001, p.1664) for comments on the assumptions. Let $\tilde{m}$ be an alternative estimator, which is characterized by partitions $\left\{\Omega_m(n), \overline{\Omega_m(n)}\right\}_{m=p+1}^M$ of $\mathcal{M}$ such that

$$P^n(\tilde{m} < m) = P^n\left(\mu_n \in \overline{\Omega_m(n)}\right), \quad P^n(\tilde{m} > m) = P^n(\mu_n \in \Omega_{m+1}(n)),$$

for each $m = p+1, \ldots, M$. Let $\Omega_{M+1}(n)$ be the empty set by convention. Note that $\tilde{m}$ contains the GMM and GEL-based moment selection procedures as special cases. Let $B(\mu,\delta)$ be an open ball with radius $\delta \in (0,\infty)$ around μ, $A^\delta = \cup_{\mu\in A} B(\mu,\delta)$ be a δ-blowup (or smoothing) of a set A. The generalized Neyman-Pearson optimality of the ELST procedures $\hat{m}$ with the fixed critical value η is obtained as follows.

Theorem 3.1 (Generalized Neyman-Pearson optimality). *Suppose that Assumptions 2.1 and 3.1 hold. Then the ELST procedure* $\hat{m} = \hat{m}_d$ *or* $\hat{m}_u$ *with the critical value* $\eta_{j,n} = \eta$ *satisfies:*

(i)

$$\sup_{P\in\mathcal{P}_m} \limsup_{n\to\infty} \frac{1}{n} \log P^n(\hat{m} < m) \leq -\eta, \tag{16}$$

for each $m = p+1, \ldots M$;

(ii) *for every alternative estimator* $\tilde{m}$, *which satisfies*

$$\sup_{P\in\mathcal{P}_m} \limsup_{n\to\infty} \frac{1}{n} \log P^n(\tilde{m} < m) \leq \sup_{P\in\mathcal{P}_m} \limsup_{n\to\infty} \frac{1}{n} \log P^n\left(\mu_n \in \left(\overline{\Omega_m(n)}\right)^\delta\right) \leq -\eta, \tag{17}$$

for each $m = p+1, \ldots M$ *at some* $\delta \in (0,\infty)$, *we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log Q^n(\hat{m} > m) \leq \limsup_{n\to\infty} \frac{1}{n} \log Q^n(\tilde{m} > m), \tag{18}$$

for each $m = p+1, \ldots M$ *and each* $Q \in \mathcal{M}$.

All proofs are contained in the Appendix. Remarks on the theorem follows.

Remark 3.1. Theorem 3.1 (i) says that the convergence rate of the underestimation probability of $\hat{m}$ is exponentially fast for all $P \in \mathcal{P}_m$ $(m = p+1, \dots M)$ and the rate is determined by the critical value η. This result off course implies $\sup_{P \in \mathcal{P}_m \setminus \mathcal{P}_{m+1}} \limsup_{n\to\infty} \frac{1}{n} \log P^n (\hat{m} < m) \leq -\eta$.

Remark 3.2. Theorem 3.1 (ii) says that if the convergence rate of the underestimation probability of $\tilde{m}$ is controlled as in (17), then the convergence rate of the overestimation probability of $\hat{m}$ is always smaller than that of $\tilde{m}$ uniformly over *all* $Q \in \mathcal{M}$.

Remark 3.3. In the context of universal hypothesis testing in information theory, this kind of result is called the generalized Neyman-Pearson δ-optimality (Zeitouni and Gutman (1991)). Since Sanov's theorem has a rough nature, we need to introduce the δ-blowup in (17). In particular, we need an open set to apply Theorem 2.1 (ii).

Remark 3.4. This theorem can be extended to the case where the critical value is fixed for n but depends on j (i.e., $\eta_{j,n} = \eta_j$). As far as the set inclusion relationships $\Lambda_M \subseteq \Lambda_{M-1} \subseteq \cdots \subseteq \Lambda_{p+1}$ are satisfied, Theorem 3.1 holds by replacing η with η_m. Even if the relationships $\Lambda_M \subseteq \Lambda_{M-1} \subseteq \cdots \subseteq \Lambda_{p+1}$ do not hold, we can still derive Theorem 3.1 (i) by replacing η with η_m. For Theorem 3.1 (ii), although we can derive that $\limsup_{n\to\infty} \frac{1}{n} \log Q^n (\mu_n \in \Lambda_{m+1}) \leq \limsup_{n\to\infty} \frac{1}{n} \log Q^n (\tilde{m} > m)$ for each $m = p+1, \dots, M$ and all $Q \in \mathcal{M}$, this result is not sufficient to derive Theorem 3.1 (ii) unless (15) holds.

Remark 3.5. Theorem 3.1 (ii) cannot exclude the possibility that the overestimation probability converges to one under some $Q \in \mathcal{M}$, i.e., the both terms in (18) converge to zero. In order to ensure the consistency of $\hat{m}$, we need to consider decreasing critical values $(\eta_{j,n} \to 0)$.

Remark 3.6. We conjecture that the other empirical likelihood-based moment selection criteria (e.g., AIC-type criterion) also satisfy the generalized Neyman-Pearson optimality. However, since the criterion-based procedures require to evaluate the criterion for all combinations of $c \in C$, they are computationally more expensive than the ELST procedures.

3.2 Overestimation Error Optimality

In this subsection, we analyze the existing moment selection procedures. Hong, Preston and Shum (2003, Proposition 2) showed that in order to ensure consistency of the ELST procedures $\hat{m}$, the critical values $\{\eta_{j,n}\}_{j,n}$ need to satisfy $\lim_{n\to\infty} \eta_{j,n} = 0$ and $\lim_{n\to\infty} n\eta_{j,n} = \infty$ for each $j = p+1, \dots, M$.[3] Intuitively, by letting $\eta_{j,n}$ be decreasing to zero, $\hat{m}$ will tend to take smaller

[3]In Hong, Preston and Shum (2003), $n\eta_n$ is defined as the critical value. Note that $\frac{L}{2n} = \inf_{P \in \mathcal{P}} I(\mu_n \| P)$.

values than the case of fixed critical values, and thus we can eliminate the cases where the over-estimation probability goes to one (see Remark 3.5). However, as a cost of obtaining consistency, the convergence rate of the underestimation probability becomes slower. We can expect that the convergence rate of the underestimation probability of $\hat{m}$ under decreasing critical values are typically non-exponential. For example, Chambaz (2006), Finesso, Liu and Narayan (1996), and Gassiat and Boucheron (2003) showed such non-exponential convergence phenomena of consistent selection procedures for the order estimation problems in parametric models, Markov chains, and hidden Markov models, respectively.

In order to analyze large deviation properties of $\hat{m}$ with decreasing critical values and the other existing moment selection procedures, we focus on the convergence rate of the overestimation probability. First, we first derive the optimal (fastest) convergence rate of overestimation probabilities for a broad class estimators of the number of correct moments. Then, we show that such an optimal rate is attained by $\hat{m}$ with decreasing critical values. Consider the following class of estimators for the number of correct moments.

Definition 3.1. *An estimator (or procedure) $\tilde{m}$ for the number of correct moments is called regular if for all $P \in \mathcal{P}_m \setminus \mathcal{P}_{m+1}$ and all $m = p+1, \ldots, M$,*

$$\liminf_{n\to\infty} P^n(\tilde{m} < m) \leq \alpha, \tag{19}$$

holds for some $\alpha \in [0, 1)$.

Recall that $\mathcal{P}_{M+1}$ is the empty set by convention. We can show that under some regularity conditions several existing estimators are regular, such as $\hat{m}$ with decreasing critical values and the GMM and GEL-based estimators. Also note that this class contains not only consistent estimators ($\alpha = 0$, e.g., BIC-type criteria) but also inconsistent ones for underestimation ($0 < \alpha < 1$, e.g., AIC-type criteria). In the class of regular estimators, we compare the convergence rates of the overestimation probabilities. The main result of this subsection is summarized as follows. Let $\Lambda_{m,\eta} = \{Q \in \mathcal{M} : \inf_{P\in\mathcal{P}_m} I(Q\|P) \leq \eta\}$.

Theorem 3.2 (Overestimation error optimality). (i) *Suppose that Assumptions 2.1 and 3.1 hold and $\tilde{m}$ is regular. Then*

$$\liminf_{n\to\infty} \frac{1}{n} \log P^n(\tilde{m} > m) \geq -\inf_{Q\in\mathcal{P}_{m+1}} I(Q\|P), \tag{20}$$

for each $P \in \mathcal{M} \setminus \mathcal{P}_{m+1}$ and each $m = p+1, \ldots, M$.

(ii) *Suppose that Assumptions 2.1 and 3.1 hold, $\hat{m}$ is regular, and the critical values satisfy $\lim_{n\to\infty} \eta_{j,n} = 0$ for each $j = p+1, \ldots, M$ and $\Lambda_{M,\eta_{M,n}} \subseteq \Lambda_{M-1,\eta_{M-1,n}} \subseteq \cdots \subseteq \Lambda_{p+1,\eta_{p+1,n}}$*

for all $n \in \mathbb{N}$. *Then*

$$\limsup_{n\to\infty} \frac{1}{n} \log P^n (\hat{m} > m) = - \inf_{Q \in \Lambda_{m+1,0}} I (Q \| P), \tag{21}$$

for each $P \in \left\{ P \in \mathcal{M} \setminus \mathcal{P}_{m+1} : \inf_{Q \in \mathcal{P}_{m+1}} I (Q \| P) < \infty \right\}$ *and each* $m = p+1, \ldots, M$.

(iii) *Suppose that Assumptions 2.1 and 3.1 hold,* $\hat{m}$ *is regular, and the critical values satisfy* $\lim_{n\to\infty} \eta_{j,n} = 0$ *for each* $j = p+1, \ldots, M$ *and* $\Lambda_{M,\eta_{M,n}} \subseteq \Lambda_{M-1,\eta_{M-1,n}} \subseteq \cdots \subseteq \Lambda_{p+1,\eta_{p+1,n}}$ *for all* $n \in \mathbb{N}$. *Moreover, assume that if* $Q \notin \mathcal{P}_{j+1}$, *then* $\inf_{P \in \mathcal{P}_j} I (Q \| P) < \inf_{P \in \mathcal{P}_{j+1}} I (Q \| P)$ *for each* $j = p+1, \ldots, M-1$. *Then*

$$\limsup_{n\to\infty} \frac{1}{n} \log P^n (\hat{m} > m) = - \inf_{Q \in \mathcal{P}_{m+1}} I (Q \| P), \tag{22}$$

for all $P \in \left\{ P \in \mathcal{M} \setminus \mathcal{P}_{m+1} : \inf_{Q \in \mathcal{P}_{m+1}} I (Q \| P) < \infty \right\}$ *and all* $m = p+1, \ldots, M$.

Remark 3.7. Theorem 3.2 (i) says that in the class of regular estimators the convergence rate of the overestimation probability is bounded from below by $-\inf_{Q \in \mathcal{P}_{m+1}} I (Q \| P)$. If $P \in \mathcal{M} \setminus \mathcal{P}_{m+1}$ satisfies $\inf_{Q \in \mathcal{P}_{m+1}} I (Q \| P) < \infty$, this bound becomes non-trivial. Note that α in (19) has no effect on the lower bound. In other words, even if we employ a inconsistent estimators with $\alpha > 0$ (e.g., AIC-type criterion), we cannot refine the optimal convergence rate of the overestimation probability.

Remark 3.8. Theorem 3.2 (ii) provides the convergence rate of the overestimation probability of $\hat{m}$ in a general setup. Theorem 3.2 (iii) says that under the additional assumption which ensures $\Lambda_{m+1,0} = \mathcal{P}_{m+1}$, $\hat{m}$ attains the lower bound in (20), i.e., $\hat{m}$ has the overestimation error optimality.

Remark 3.9. Compared to (18) in Theorem 3.1 (ii), (22) holds only for $P \in \left\{ P \in \mathcal{M} \setminus \mathcal{P}_{m+1} : \inf_{Q \in \mathcal{P}_{m+1}} I (Q \| P) < \infty \right\}$. However, we always have a non-trivial convergence rate in this set.

3.3 Minimax Misclassification Optimality

We now consider the optimality of empirical likelihood for estimating the true moment selection vector. For simplicity, we assume that the true number of moments m^o is known. Thus, the ELST-based estimator for true selection vector is:

$$\hat{c} = \arg \min_{c \in C_{m^o}} \left\{ \inf_{P \in \mathcal{P}^c} I (\mu_n \| P) \right\}, \tag{23}$$

i.e., choose the selection vector which minimizes the empirical likelihood ratio. Since the parameter space C_{m^o} is discrete, it is also reasonable to analyze the large deviation properties of $\hat{c}$ as well as the case of $\hat{m}$. Let

$$\Lambda_{m^o}^c = \left\{ \mu \in \mathcal{M} : \inf_{P \in \mathcal{P}^c} I(\mu||P) = \min_{c' \in C_{m^o}} \inf_{P \in \mathcal{P}^{c'}} I(\mu||P) \right\},$$

be a subset of $\mathcal{M}$, where c is selected. Note that $\hat{c}$ is defined by the partition $\{\Lambda_{m^o}^c\}_{c \in C_{m^o}}$. We particularly focus on the misclassification probability of $\hat{c}$, i.e., $P^n(\hat{c} \neq c)$ under $P \in \mathcal{P}^c$. By using $\Lambda_{m^o}^c$, the misclassification probability is written as

$$P^n(\hat{c} \neq c) = P^n\left(\mu_n \in \overline{\Lambda_{m^o}^c}\right), \tag{24}$$

for each $P \in \mathcal{P}^c$. If $\hat{c}$ is consistent, (24) converges to zero. Since (24) is written by μ_n, Sanov's theorem is useful to analyze the large deviation property. However, (24) is too ambitious as an optimality criterion for $\hat{c}$. As Choirat and Seri (2002) indicated, the globally optimal estimator, which attains the lower bound of $\liminf_{n\to\infty} \frac{1}{n} \log P^n(\hat{c} \neq c)$ for each $P \in \mathcal{P}^c$, does not exist in general. Therefore, we consider the maximum of the large deviation misclassification probability:

$$\max_{c \in C_{m^o}} \sup_{P \in \mathcal{P}^c} \limsup_{n\to\infty} \frac{1}{n} \log P^n(\hat{c} \neq c). \tag{25}$$

Let $\{\Omega_{m^o}^c\}_{c \in C_{m^o}}$ be an alternative partition of $\mathcal{M}$, and $\tilde{c}$ be an alternative estimators defined as

$$\tilde{c} = c \quad \text{if} \quad \mu_n \in \Omega_{m^o}^c, \tag{26}$$

for all $c \in C_{m^o}$. An estimator $\tilde{c}$ is called regular if

$$\lim_{\delta \to 0} \max_{c \in C_{m^o}} \sup_{P \in \mathcal{P}^c} \limsup_{n\to\infty} \frac{1}{n} \log P\left(\mu_n \in \left(\overline{\Omega_{m^o}^c}\right)^\delta\right) = \max_{c \in C_{m^o}} \sup_{P \in \mathcal{P}^c} \limsup_{n\to\infty} \frac{1}{n} \log P\left(\mu_n \in \left(\overline{\Omega_{m^o}^c}\right)\right).$$

See Zeitouni and Gutman (1991) for a discussion of this condition. The minimax optimality of $\hat{c}$ is obtained as follows.

Theorem 3.3 (Optimality of $\hat{c}$). *Suppose that Assumptions 2.1 and 3.1 hold, and $\tilde{c}$ is regular. Then*

$$\max_{c \in C_{m^o}} \sup_{P \in \mathcal{P}^c} \limsup_{n\to\infty} \frac{1}{n} \log P(\hat{c} \neq c) \leq \max_{c \in C_{m^o}} \sup_{P \in \mathcal{P}^c} \limsup_{n\to\infty} \frac{1}{n} \log P(\hat{c} \neq c).$$

Remark 3.10. Theorem 3.3 says that the worst (slowest) convergence rate of the misclassification probability is minimized by $\hat{c}$. The regularity assumption for $\tilde{c}$ is required due to a rough nature of Sanov's theorem.

Remark 3.11. Kitamura and Otsu (2005) analyzed the limit of the maximum large deviation probability (i.e., $\lim_{n\to\infty} \max_{c \in C_{m^o}} \sup_{P \in \mathcal{P}^c} \frac{1}{n} \log P(\hat{c} \neq c)$). We conjecture that $\hat{c}$ is also minimax optimal in the sense of Kitamura and Otsu (2005).

4 Simulation

To be written. Use the setup of Hong, Preston and Shum (2003).

5 Conclusion

This paper proposes optimal large deviation properties of the empirical likelihood sequential testing procedures for selecting all correct moment restrictions. We derive three optimal properties: (i) the generalized Neyman-Pearson optimality, (ii) the overestimation error optimality, and (iii) the minimax misclassification error optimality. We find that empirical likelihood optimally controls the large deviation error probabilities and is more preferable than the other GEL objective functions including GMM. Although it is outside the scope of this paper, we can expect that similar optimal properties hold for the other empirical likelihood-based selection procedures (e.g., BIC-type criterion). It is interesting to extend our approach to the other model selection problems (e.g., lag selection) or discrete parameter estimation problems (e.g., change point estimation).

A Mathematical Appendix

A.1 Proof of Theorem 3.1

A.1.1 Proof of (i)

From (15), it is sufficient to show that

$$\sup_{P\in\mathcal{P}_m} \limsup_{n\to\infty} \frac{1}{n} \log P^n \left(\hat{m} < m\right) = \sup_{P\in\mathcal{P}_m} \limsup_{n\to\infty} \frac{1}{n} \log P^n \left(\mu_n \in \overline{\Lambda_m}\right) \leq -\eta, \tag{27}$$

for each $m = p+1, \ldots, M$. A similar argument as the proof of Kitamura (2001, Theorem 2) yields (27).

A.1.2 Proof of (ii)

Pick any $m = p+1, \ldots M$. We first show that there exists $n_0 \in \mathbb{N}$ such that

$$\Lambda_{m+1} \subseteq \Omega_{m+1}(n) \tag{28}$$

holds for all $n > n_0$. Suppose (28) does not hold. Then there exists an infinite sequence of measures $\{\xi_\ell\}_{\ell\in\mathbb{N}}$ such that

$$\xi_\ell \in \Lambda_{m+1} \quad \text{and} \quad \xi_\ell \in \overline{\Omega_{m+1}(n_\ell)}.$$

Since the set $\Lambda_{m+1} = \left\{Q \in \mathcal{M} : \inf_{P \in \mathcal{P}_{m+1}} I(Q\|P) \leq \eta\right\}$ is compact in the weak topology (Deuschel and Stroock (1989, Ch. 3.2)), there exists a subsequence $\{\ell_k\}_{k\in\mathbb{N}}$ such that ξ_{ℓ_k} converges to a measure $\xi \in \Lambda_{m+1}$. For such a ξ, we can take an open ball $B(\xi, \delta/2)$ such that $B(\xi, \delta/2) \subset \left(\overline{\Omega_{m+1}(n_{\ell'})}\right)^{\delta}$ holds for some subsequence $\{n_{\ell'}\}_{\ell'\in\mathbb{N}}$. Also, there exists $\ell^* \in \mathbb{N}$ such that $\xi_{\ell^*} \in B(\xi, \delta/2)$. Since $\xi_{\ell^*} \in \Lambda_{m+1}$, we have

$$\inf_{P\in\mathcal{P}_{m+1}} I(\xi_{\ell^*}\|P) < \eta. \tag{29}$$

Therefore,

$$\begin{aligned}
& \sup_{P\in\mathcal{P}_{m+1}} \limsup_{n\to\infty} \frac{1}{n} \log P^n\left(\mu_n \in \left(\overline{\Omega_{m+1}(n_{\ell'})}\right)^{\delta}\right) \\
\geq & \sup_{P\in\mathcal{P}_{m+1}} \liminf_{\ell'\to\infty} \frac{1}{n_{\ell'}} \log P^{n_{\ell'}}\left(\mu_{n_{\ell'}} \in \left(\overline{\Omega_{m+1}(n_{\ell'})}\right)^{\delta}\right) \\
\geq & \sup_{P\in\mathcal{P}_{m+1}} \liminf_{n\to\infty} \frac{1}{n} \log P^n\left(\mu_n \in B(\xi, \delta/2)\right) \\
\geq & \sup_{P\in\mathcal{P}_{m+1}} \left[-\inf_{Q\in B(\xi,\delta/2)} I(Q\|P)\right] \\
\geq & \sup_{P\in\mathcal{P}_{m+1}} \left[-I(\xi_{\ell^*}\|P)\right] \\
> & -\eta,
\end{aligned} \tag{30}$$

where the second inequality follows from $B(\xi, \delta/2) \subset \left(\overline{\Omega_{m+1}(n_{\ell'})}\right)^{\delta}$, the third inequality follows from Sanov's theorem (the second part of Theorem 2.1), the fourth inequality follows from $\xi_{\ell^*} \in B(\xi, \delta/2)$, and the last inequality follows from (29). However, since (17) requires

$$\sup_{P\in\mathcal{P}_{m+1}} \limsup_{n\to\infty} \frac{1}{n} \log P^n\left(\mu_n \in \left(\overline{\Omega_{m+1}(n)}\right)^{\delta}\right) \leq -\eta,$$

we have a contradiction. Therefore, there exists $n_0 \in \mathbb{N}$ such that (28) holds for all $n > n_0$, and we have

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log Q^n(\hat{m} > m) &= \limsup_{n\to\infty} \frac{1}{n} \log Q^n(\mu_n \in \Lambda_{m+1}) \\
&\leq \limsup_{n\to\infty} \frac{1}{n} \log P^n(\mu_n \in \Omega_{m+1}(n)) = \limsup_{n\to\infty} \frac{1}{n} \log P^n(\tilde{m} > m),
\end{aligned}$$

for all $Q \in \mathcal{M}$. The conclusion is obtained.

A.2 Proof of Theorem 3.2

Proof of (i). This result is shown by applying Stein's lemma (Lemma 2.1). Let

$$\begin{aligned} A_n &= \{\tilde{m} \geq m+1\}, \\ P &\in \left\{ P \in \mathcal{M} \setminus \mathcal{P}_{m+1} : \inf_{Q \in \mathcal{P}_{m+1}} I(Q\|P) < \infty \right\}, \\ Q &\in \mathcal{P}_{m+1} \setminus \mathcal{P}_{m+2}. \end{aligned}$$

Pick any $m = p+1, \ldots M$. Since $\tilde{m}$ is regular, we have

$$\liminf_{n\to\infty} Q^n(A_n) = \liminf_{n\to\infty} Q^n(\tilde{m} \geq m+1) = 1 - \liminf_{n\to\infty} Q^n(\tilde{m} < m+1) \geq 1 - \alpha > 0,$$

for each $Q \in \mathcal{P}_{m+1} \setminus \mathcal{P}_{m+2}$. Since the assumption of Lemma 2.1 is satisfied, we have

$$\liminf_{n\to\infty} \frac{1}{n} \log P^n(\tilde{m} > m) \geq \liminf_{n\to\infty} \frac{1}{n} \log P^n(\tilde{m} \geq m+1) \geq -I(Q\|P), \tag{31}$$

for each $P \in \mathcal{M} \setminus \mathcal{P}_{m+1}$ and all $Q \in \mathcal{P}_{m+1} \setminus \mathcal{P}_{m+2}$. Similarly, we have

$$\liminf_{n\to\infty} Q^n(A_n) \geq \liminf_{n\to\infty} Q^n(\tilde{m} \geq m+j) = 1 - \liminf_{n\to\infty} Q^n(\tilde{m} < m+j) > 0,$$

for each $Q \in \mathcal{P}_{m+j} \setminus \mathcal{P}_{m+j+1}$ and each $j = 1, \ldots, M-m-1$. Thus, Lemma 2.1 yields

$$\liminf_{n\to\infty} \frac{1}{n} \log P^n(\tilde{m} > m) \geq \liminf_{n\to\infty} \frac{1}{n} \log P^n(\tilde{m} \geq m+j) \geq -I(Q\|P), \tag{32}$$

for each $P \in \mathcal{M} \setminus \mathcal{P}_{m+1}$ and each $Q \in \mathcal{P}_{m+j} \setminus \mathcal{P}_{m+j+1}$. Since $\mathcal{P}_{m+1} = \bigcup_{j=1}^{M-m} (\mathcal{P}_{m+j} \setminus \mathcal{P}_{m+j+1})$, (31) and (32) yield the conclusion.

Proof of (ii). Define

$$\Lambda_{j,\eta_{j,n}} = \left\{ Q \in \mathcal{M} : \inf_{P \in \mathcal{P}_j} I(Q\|P) \leq \eta_{j,n} \right\}, \quad \Lambda_{j,\eta} = \left\{ Q \in \mathcal{M} : \inf_{P \in \mathcal{P}_j} I(Q\|P) \leq \eta \right\}.$$

Pick any $m = p+1, \ldots, M$ and any $P^* \in \left\{ P \in \mathcal{M} \setminus \mathcal{P}_{m+1} : \inf_{Q \in \mathcal{P}_{m+1}} I(Q\|P) < \infty \right\}$. For all $\eta \in (0, \infty)$, the set $\Lambda_{m+1,\eta}$ is compact in the weak topology (see Deuschel and Stroock (1989, Ch. 3.2)). Therefore, Sanov's theorem (Theorem 2.1) implies that

$$\limsup_{n\to\infty} \frac{1}{n} \log P^{*n}\left(\mu_n \in \Lambda_{m,\eta_{m,n}}\right) \leq \limsup_{n\to\infty} \frac{1}{n} \log P^{*n}(\mu_n \in \Lambda_{m+1,\eta}) \leq -\inf_{Q \in \Lambda_{m+1,\eta}} I(Q\|P^*), \tag{33}$$

for all $\eta \in (0, \infty)$. Since the sequence of the sets $\{\Lambda_{m+1,\eta}\}$ is non-increasing as $\eta \downarrow 0$, the sequence $\left\{\inf_{Q \in \Lambda_{m+1,\eta}} I(Q\|P^*)\right\}$ is non-decreasing as $\eta \downarrow 0$. Also, from $\mathcal{P}_{m+1} \subseteq \Lambda_{m+1,0}$, $\left\{\inf_{Q \in \Lambda_{m+1,\eta}} I(Q\|P^*)\right\}$ is bounded above by $\inf_{Q \in \Lambda_{m+1,0}} I(Q\|P^*) \leq \inf_{Q \in \mathcal{P}_{m+1}} I(Q\|P^*) < \infty$.

Thus, the sequence $\left\{\inf_{Q\in\Lambda_{m+1,\eta}} I\left(Q\|P^*\right)\right\}$ converges as $\eta\downarrow 0$ to some limit $\bar{I}\leq \inf_{Q\in\Lambda_{m+1,0}} I\left(Q\|P^*\right)$, and we can take a decreasing sequence of positive numbers $\{\eta_\ell\}_{\ell\in\mathbb{N}}$ such that $\left\{\inf_{Q\in\Lambda_{m+1,\eta_\ell}} I\left(Q\|P^*\right)\right\}$ increases to $\bar{I}$ as $\ell\to\infty$.

Since Λ_{m+1,η_ℓ} is compact and $I\left(Q\|P\right)$ is lower semicontinuous for $Q\in\Lambda_{m+1,\eta_\ell}$, the infimum $\inf_{Q\in\Lambda_{m+1,\eta_\ell}} I\left(Q\|P^*\right)$ is attained on the compact set Λ_{m+1,η_ℓ}, i.e., there exists $Q_\ell\in\Lambda_{m+1,\eta_\ell}$ such that $I\left(Q_\ell\|P^*\right)=\inf_{Q\in\Lambda_{m+1,\eta_\ell}} I\left(Q\|P^*\right)$.

Now, consider the sequence of the sets $\left\{\mathrm{cl}\left(\{Q_\ell:\ell\geq\ell'\}\right)\right\}_{\ell'\in\mathbb{N}}$. Since $\mathrm{cl}\left(\{Q_\ell:\ell\geq\ell'\}\right)$ is closed and $\mathrm{cl}\left(\{Q_\ell:\ell\geq\ell'\}\right)\subseteq\Lambda_{m+1,\eta_{\ell'}}$ for all $\ell'\in\mathbb{N}$, $\mathrm{cl}\left(\{Q_\ell:\ell\geq\ell'\}\right)$ is compact for all $\ell'\in\mathbb{N}$. Therefore, since $\left\{\mathrm{cl}\left(\{Q_\ell:\ell\geq\ell'\}\right)\right\}_{\ell'\in\mathbb{N}}$ is a non-increasing sequence of non-empty compact sets, the Heine-Borel theorem implies that $\bigcap_{\ell'=1}^{\infty}\mathrm{cl}\left(\{Q_\ell:\ell\geq\ell'\}\right)$ is non-empty.

Pick $\bar{Q}\in\bigcap_{\ell'=1}^{\infty}\mathrm{cl}\left(\{Q_\ell:\ell\geq\ell'\}\right)$. Since $I\left(Q_\ell\|P^*\right)\leq\bar{I}$ for all $\ell\in\mathbb{N}$, we have $\{Q_\ell:\ell\geq\ell'\}\subseteq\left\{Q\in\mathcal{M}:I\left(Q\|P^*\right)\leq\bar{I}\right\}$. Thus, from $\mathrm{cl}\left(\{Q_\ell:\ell\geq\ell'\}\right)\subseteq\left\{Q\in\mathcal{M}:I\left(Q\|P^*\right)\leq\bar{I}\right\}$, we have $I\left(\bar{Q}\|P^*\right)\leq\bar{I}$. On the other hand, since $\bar{Q}\in\Lambda_{m+1,\eta_\ell}$ for all $\ell\in\mathbb{N}$, we have $\bar{Q}\in\Lambda_{m+1,0}$ and thus $\bar{I}\leq\inf_{Q\in\Lambda_{m+1,0}} I\left(Q\|P^*\right)\leq I\left(\bar{Q}\|P^*\right)$. Collecting these results,

$$I\left(\bar{Q}\|P^*\right)\leq\bar{I}\leq\inf_{Q\in\Lambda_{m+1,0}} I\left(Q\|P^*\right)\leq I\left(\bar{Q}\|P^*\right),$$

i.e., $\left\{\inf_{Q\in\Lambda_{m+1,\eta_\ell}} I\left(Q\|P^*\right)\right\}$ increases to $\bar{I}=\inf_{Q\in\Lambda_{m+1,0}} I\left(Q\|P^*\right)$ as $\ell\to\infty$.

Therefore, from (33), we have

$$\limsup_{n\to\infty}\frac{1}{n}\log P^n\left(\hat{m}>m\right)=\limsup_{n\to\infty}\frac{1}{n}\log P^n\left(\mu_n\in\Lambda_{m,\eta_{m,n}}\right)\leq-\inf_{Q\in\Lambda_{m+1,0}} I\left(Q\|P\right), \tag{34}$$

for all $P\in\left\{P\in\mathcal{M}\setminus\mathcal{P}_{m+1}:\inf_{Q\in\mathcal{P}_{m+1}} I\left(Q\|P\right)<\infty\right\}$ and all $m=p+1,\ldots,M$.

Proof of (iii). Since $\hat{m}$ is regular, Part (i) implies that

$$\liminf_{n\to\infty}\frac{1}{n}\log P^n\left(\hat{m}>m\right)\geq-\inf_{Q\in\mathcal{P}_{m+1}} I\left(Q\|P\right),$$

for each $P\in\left\{P\in\mathcal{M}\setminus\mathcal{P}_{m+1}:\inf_{Q\in\mathcal{P}_{m+1}} I\left(Q\|P\right)<\infty\right\}$ and each $m=p+1,\ldots,M$. Thus, from Part (ii), it is sufficient to show that

$$\Lambda_{m+1,0}\subseteq\mathcal{P}_{m+1}. \tag{35}$$

Suppose otherwise. Then there exists $\tilde{Q}\in\Lambda_{m+1,0}$ such that $\tilde{Q}\notin\mathcal{P}_{m+1}$. From $\tilde{Q}\in\Lambda_{m+1,0}$, we have $\inf_{P\in\mathcal{P}_{m+1}} I\left(\tilde{Q}\|P\right)=0$. However, from the additional assumption for Part (iii), we have $\inf_{P\in\mathcal{P}_m} I\left(\tilde{Q}\|P\right)<\inf_{P\in\mathcal{P}_{m+1}} I\left(\tilde{Q}\|P\right)=0$, and this is a contradiction. Therefore, (35) holds true, and we have

$$\limsup_{n\to\infty}\frac{1}{n}\log P^n\left(\hat{m}>m\right)\leq-\inf_{Q\in\Lambda_{m+1,0}} I\left(Q\|P\right)\leq-\inf_{Q\in\mathcal{P}_{m+1}} I\left(Q\|P\right),$$

for each $P\in\left\{P\in\mathcal{M}\setminus\mathcal{P}_{m+1}:\inf_{Q\in\mathcal{P}_{m+1}} I\left(Q\|P\right)<\infty\right\}$ and each $m=p+1,\ldots,M$.

A.3 Proof of Theorem 3.3

Sanov's theorem yields that

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log P^n\left(\hat{c}\neq c\right) &= \limsup_{n\to\infty}\frac{1}{n}\log P^n\left(\mu_n\in\overline{\Lambda^c_{m^o}}\right)\\ &\leq \limsup_{n\to\infty}\frac{1}{n}\log P^n\left(\mu_n\in \mathrm{cl}\left(\overline{\Lambda^c_{m^o}}\right)\right)\\ &\leq -\inf_{Q\in\mathrm{cl}\left(\overline{\Lambda^c_{m^o}}\right)} I\left(Q||P\right),\end{aligned}$$

for each $P\in\mathcal{P}^c$. Therefore,

$$\limsup_{n\to\infty}\frac{1}{n}\log P^n\left(\hat{c}\neq c\right)\leq \max_{c\in C_{m^o}}\sup_{P\in\mathcal{P}^c}\left[-\inf_{Q\in\mathrm{cl}\left(\overline{\Lambda^c_{m^o}}\right)} I\left(Q||P\right)\right]. \tag{36}$$

Pick any partition $\{\Omega^c_{m^o}\}_{c\in C_{m^o}}$. By applying Sanov's theorem again,

$$\liminf_{n\to\infty}\frac{1}{n}\log P^n\left(\mu_n\in\left(\overline{\Omega^c_{m^o}}\right)^\delta\right)\geq -\inf_{Q\in\left(\overline{\Omega^c_{m^o}}\right)^\delta} I\left(Q||P\right),$$

for each $\delta\in(0,\infty)$. Therefore,

$$\max_{c\in C_{m^o}}\sup_{P\in\mathcal{P}^c}\liminf_{n\to\infty}\frac{1}{n}\log P^n\left(\mu_n\in\left(\overline{\Omega^c_{m^o}}\right)^\delta\right)\geq \max_{c\in C_{m^o}}\sup_{P\in\mathcal{P}^c}\left[-\inf_{Q\in\left(\overline{\Omega^c_{m^o}}\right)^\delta} I\left(Q||P\right)\right], \tag{37}$$

for each $\delta\in(0,\infty)$. The definition of $\Lambda^c_{m^o}$ yields that

$$\max_{c\in C_{m^o}}\sup_{P\in\mathcal{P}^c}\left[-\inf_{Q\in\mathrm{cl}\left(\overline{\Lambda^c_{m^o}}\right)} I\left(Q||P\right)\right]\leq \max_{c\in C_{m^o}}\sup_{P\in\mathcal{P}^c}\left[-\inf_{Q\in\left(\overline{\Omega^c_{m^o}}\right)^\delta} I\left(Q||P\right)\right], \tag{38}$$

for each $\delta\in(0,\infty)$. Combining (36), (37), and (38),

$$\limsup_{n\to\infty}\frac{1}{n}\log P^n\left(\hat{c}\neq c\right)\leq \max_{c\in C_{m^o}}\sup_{P\in\mathcal{P}^c}\liminf_{n\to\infty}\frac{1}{n}\log P^n\left(\mu_n\in\left(\overline{\Omega^c_{m^o}}\right)^\delta\right),$$

for each $\delta\in(0,\infty)$. Therefore, the regularity of $\tilde{c}$ yields the conclusion.

References

[1] Andrews, D. W. K. (1999) Consistent moment selection procedures for generalized method of moments estimation, *Econometrica*, 67, 543-564.

[2] Andrews, D. W. K. and B. Lu (2001) Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models, *Journal of Econometrics*, 101, 123-164.

[3] Borwein, J. and A. Lewis (1993) Partially-finite programming in L_1 and the existence of maximum entropy estimates, *SIAM Journal of Optimization*, 3, 248-267.

[4] Bahadur, R., Zabell, S. and J. Gupta (1980) Large deviations, tests, and estimates, in *Asymptotic Theory of Statistical Tests and Estimation*, ed. by I. M. Chaterabarli, pp. 33-64, Academic Press, New York.

[5] Chaganty, N. R. and R. L. Karandikar (1996) Some properties of the Kullback-Leibler number, *Sankhyā*, 58, 69-80.

[6] Chambaz, A. (2006) Testing the order of a model, forthcoming in *Annals of Statistics*.

[7] Choirat, C. and R. Seri (2001) Estimation in discrete parameter models, Working paper.

[8] Dembo, A. and O. Zeitouni (1998) *Large Deviations Techniques and Applications*, Springer, second edition.

[9] Deuschel, J. D. and D. W. Stroock (1989) *Large Deviations*, Academic Press.

[10] Finesso, L., Liu, C. C. and P. Narayan (1996) The optimal error exponent for Markov order estimation, *IEEE Transactions on Information Theory*, 42, 1488-1497.

[11] Gassiat, E. and S. Boucheron (2003) Optimal error exponents in hidden Markov models order estimation, *IEEE Transactions on Information Theory*, 49, 964-980.

[12] Hansen, L. P., Heaton, J. and A. Yaron (1996) Finite-sample properties of some alternative GMM estimators, *Journal of Business and Economic Statistics*, 14, 262-280.

[13] Hoeffding, W. (1965) Asymptotically optimal tests for multinomial distributions (with Discussion), *Annals of Mathematical Statistics*, 36, 369-408.

[14] Hong, H., Preston, B. and M. Shum (2003) Generalized empirical likelihood-based model selection criteria for moment condition models, *Econometric Theory*, 19, 923-943.

[15] Imbens, G. W., Spady, R. H., and P. Johnson (1998) Information theoretic approaches to inference in moment condition models, *Econometrica*, 66, 333-357.

[16] Khudanpur, S. and P. Narayan (2002) Order estimation for a special class of hidden Markov sources and binary renewal processes, *IEEE Transactions on Information Theory*, 48, 1704-1713.

[17] Kitamura, Y. (2001) Asymptotic optimality of empirical likelihood for testing moment restrictions, *Econometrica*, 69, 1661-1672.

[18] Kitamura, Y. and T. Otsu (2005) Minimax estimation and testing for moment condition models via large deviations, Working paper.

[19] Kitamura, Y. and M. Stutzer (1997) An information theoretic alternative to generalized method of moments estimation, *Econometrica*, 65, 861-874.

[20] Merhav, N. (1989) The estimation of the model order in exponential families, *IEEE Transactions on Information Theory*, 35, 1109-1114.

[21] Merhav, N., Gutman, M. and J. Ziv (1989) On the estimation of the order of a Markov chain and universal data compression, IEEE Transactions on Information Theory, 35, 1014-1019.

[22] Newey, W. K. and R. J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica*, 72, 219-256.

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

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

[25] Zeitouni, O. and M. Gutman (1991) On universal hypotheses testing via large deviations, *IEEE Transactions on Information Theory*, 37, 285-290.