AN EMPIRICAL LIKELIHOOD-BASED LOCAL ESTIMATION[1]

Zhengyuan Gao

This paper considers a modified empirical likelihood and uses it to estimate models with imperfect moment restrictions. Consistency, local normality and asymptotic optimality of the new estimator only require the local identifiability and the linear-quadratic representation. In imperfect situations, classical sandwich algorithms give bias estimations while the new estimator preserves its asymptotic properties. We consider two imperfect cases that are unbounded moment restrictions and weakly identified moment restrictions.

1. INTRODUCTION

A fully specified parametric model is often too restrictive for economic applications and can be very sensitive to misspecifications. It is desirable therefore to relax the parametric assumptions on the stochastic part of the model even if one wants to preserve the deterministic part which might be driven by economic theory. Empirical Likelihood (EL Owen, 1988) is a likelihood inference approach based on the empirical measure that can incorporate known constraints on parameters derived from economic theory. In this way we obtain a model that is close to the data yet respects the economic constraints without fully specifying the model parametrically.

Many estimation and inference procedures are based on moment constraints. Ordinary Least Square (OLS), Instrumental Variables (IV), and Generalized Method of Moments (GMM) can all be formulated as moment based methods. The estimated parameters of these methods optimize objective functions in terms of different metrics or discrepancies subject to the moment constraints. For example, GMM optimizes a quadratic form using a weighting matrix and EL optimizes an information criterion by choosing weights of individual observations optimally. If the weighting matrix of GMM is chosen optimally, it shares the same first order properties with EL and a number of other moment based methods, see Newey and Smith (2004).

EL belongs to a general technique where the optimal value minimizes the relative divergence between specified models and empirical data. This relative divergence can be measured using different criterion functions and is sometimes referred to as "empirical discrepancy". Examples include f-divergence (Csiszar, 1984), Cressie-Read (Baggerly, 1998) and GEL (Smith, 1997). The qualities and properties of inference procedures depend on the choice of discrepancy measure. Baggerly (1998) proves that EL is the only Bartlett correctable member in Cressie-Read family and Owen (2001) shows that relative entropy divergence is the only member in the f-divergence class that can attribute zero weight for ill-behaved observations such as outliers. So called dirty data are of considerable concern in practice, since they are prevalent and can have far reaching consequences for the economic conclusions drawn. On the other hand, methods that are not sensitive to outliers and the like tend to be much less precise when the data are generated by the model specified. The choice of divergence is important for efficiency and robustness and there is an implicit tradeoff. From a mathematical point of view, this relates the topological structure induced by the divergence. A "weaker" topological structure would induce more robust inference but requires a large sample to obtain the same level of accuracy. A "stronger" topology recognizes differences more easily and generates more efficient inference in general.

Cowles Foundation, Yale University. Economics Department, University of Amsterdam. zhengyuan.gao@yale.edu

[1]This is a preliminary version. Comments are welcomed.

To address the dilemma, Schennach (2007) suggests a two-step inference method. The method balances efficiency and robustness by switching the empirical discrepancy between Kullback-Leibler and Shannon entropy. Kitamura, Otsu, and Evdokimov (2009) propose to use Hellinger's distance to overcome the risk of making an extreme choice.

The term "local" in nonparametric and semi-parametric inference procedures usually refers to local smoothing techniques that convolve or average the locations of a limited number of data points for each individual observation. The smoothed function reduces the influence of individual data points. Some of the econometrics literature that uses "local" in this sense can be found in Brown and Newey (2002); Donald, Imbens, and Newey (2003); Kitamura, Tripathi, and Ahn (2004); Parente and Smith (????); Smith (2005), to name but a few.

The term local in our context refers to a local parameter that we introduce below. We consider a neighborhood in which we take two parameter values to construct an empirical log-likelihood ratio function used to obtain a smooth approximation. This technique often appears in the evaluation of local power of test statistics and statistical experiments, see Le Cam and Yang (2000); van der Vaart (1998).

This paper considers a one-step inference algorithm within the EL framework. The algorithm induces a local type estimator which preserves the EL properties. The algorithm uses a stable solver in the optimization step rather than changes in the choice of discrepancy to eliminate the effects of peculiar points.

We assume there are d moments available to estimate a k dimensional parameter vector θ and $k \leq d$. Let $m(\theta, x)$ be a d dimensional vector function of parameter θ and random variable x and define $M(\theta) = \mathbb{E}[m(\theta, x)]$, where the expectation $\mathbb{E}[\cdot]$ is taken w.r.t the probability measure P_{θ_0} for x. We call $\mathbb{E}[m(\theta, x)] = 0$ a moment restriction function or orthogonality condition. For given observation x_i we will write $m_i(\theta) = m(\theta, x_i)$ and we use $\rightsquigarrow$ to denote the weak convergence.

2. LE CAM TYPE LOCALIZED EMPIRICAL LIKELIHOOD

2.1. *Empirical Likelihood*

Empirical Likelihood is a likelihood inference approach without many parametric assumptions. The major advantage of EL or more generally GEL over GMM is that in the over-identification case the divergence does not depend on an unstable operator. The GMM criterion function is the estimated weighting matrix projects $\sum m_i(\theta)/n$ into a matrix norm $\| \sum m_i(\theta)/n \|^2_{W_n}$ where W_n, in the over-identified case, is usually unknown. Newey and Smith (2004) show the direct affect from the higher order bias term in Generalized EL estimation does not grow with increasing the number of moment restriction functions, but it grows without bound in GMM case.

EL can be regarded as a semi-parametric technique for the analysis of a finite dimensional parameter θ and an increasing number of probabilities. It simultaneously finds the optimal θ and the optimal weights satisfying the required constraints. The EL criterion is:

$$\sup \left\{ \sum_{i=1}^{n} \log np_i \mid p_i \geq 0, \quad \sum_{i}^{n} p_i = 1, \sum_{i=1}^{n} p_i m_i(\theta) = 0 \right\}.$$

For a given value of θ, an explicit expression for the optimal p_i's can be derived using the method of Lagrange:

$$\tilde{p}_{i,\theta} = \frac{1}{n} \frac{1}{1 + \lambda^T m_i(\theta)},$$

where $\tilde{p}_{i,\theta}$ is the implied probability of $m_i(\theta)$ and λ is a $d \times 1$ vector given as the solution of:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{m_i(\theta)}{1+\lambda^T m_i(\theta)} = 0. \tag{2.1}$$

We define the log-likelihood ratio of implied probabilities for two parameter values as:

$$\Lambda_n(\theta_1,\theta_2) = \sum_{i}^{n}\log(\tilde{p}_{i,\theta_1}/\tilde{p}_{i,\theta_2})$$

and the log-likelihood ratio of implied probability for θ and unconstrained empirical density, i.e. $1/n$, as

$$\Lambda_n(\theta) = \sum_{i}^{n}\log n\tilde{p}_{i,\theta}.$$

We also use $\lambda_n(\theta)$ when we want to stress the fact that it is a pointwise solution of equation (2.1) with n observations. The constraint $0 \le \tilde{p}_i \le 1$ implies that the inequality $1+\lambda^T m_i(\theta) \ge 1/n$ holds. The vector function $\lambda(\theta)$ must be located in a convex and closed set $\Gamma_\theta = \{\lambda : 1+\lambda^T m_i(\theta) \ge 1/n,\ i = 1,\dots,n\}$.

When there is a unique θ_0 such that $\mathbb{E}[m(\theta)] = 0$ then we say that the parameter θ is globally identified in Θ by the moment function $m(\cdot)$. EL uses a divergence criterion which preserves the identification property. In addition, EL improves the efficiency as compared to GMM.

Most estimation procedures are global approaches and assume the underlying parameters can be globally found. Global optimization procedure is used for problems with a small number of variables, where computing time is not critical, and the possibility of finding the true global solution is very high. The complexity of global optimization methods grows exponentially with the problem sizes on k and d. In the following sections, we show that global identification assumption may either not hold or not guarantee the problem is identifiable under the standard weakly convergence rate of the estimator. Therefore, we suggest a local approach try to overcome these drawbacks. The local method is able to obtain the normal asymptotic results under very weak conditions when the global method breaks down. Such weak conditions allow some imperfect situations in raw data sets.

CONDITION 1 *(i) $M(\theta)$ exists for all $\theta \in \Theta$ and has a unique zero at $\theta = \theta_0$.*

(ii) θ_0 is a well-separated point in $M(\theta)$ such that $\sup_{\theta:d(\theta,\theta_0)\ge\varepsilon} M(\theta) < M(\theta_0)$.

(iii) $m(x,\theta)$ is continuous in θ, $\lim_{\theta'\to\theta}\|m(x,\theta)-m(x,\theta')\| = 0$.

(iv) Let ∞ be the one-point compactification of Θ, then there exists a continuous function $b(\theta)$ bounded away from zero, such that

(1) $\sup_{\theta\in\Theta}\|m(x,\theta)\|/b(\theta)$ is integrable,

(2) $\liminf_{\theta\to\infty}\|M(\theta)\|/b(\theta) \ge 1$, and

(3) $\mathbb{E}[\limsup_{\theta\to\infty}\|m(x,\theta)-M(\theta)\|/b(\theta)] < 1$.

(v) $\mathbb{E}\left[m(x,\theta_0)m(x,\theta_0)^T\right]$ exists and has full rank.

(i) ensures the model is identified. (ii) is a local separable condition. (iii) is used to obtain the continuity of the Lagrangian multiplier. (iv) is an envelop assumption; we use it to obtain some dominated convergence results. The one-point (Alexandroff) compactification allows us to let θ approach any boundary point of Θ, even if Θ is not compact and may extend indefinitely. The

usual proof of EL consistency (Qin and Lawless, 1994) requires the existence of twice continuous derivative of $m(\theta)$ and that the derivative is full ranked. Condition 1 is less restrictive because we want to allow for less regular cases where the usual "Delta approach" does not work due to the non-differentiable or when the true θ_0 is not well separated. See Huber (1981) for a definition of "well separated" which we are going to adapt for use in the context of weak identification. Condition 1 (i)-(iv) are the standard M-estimator conditions in Huber (1981) and are very weak in the context of parametric models.

THEOREM 2.1 *If condition 2.1 holds, then every sequence T_n satisfying*

$$T_n := \arg\sup_{\theta\in\Theta} \sum_{i=1}^{n} \log n\tilde{p}_i(\theta)$$

will converge to θ_0 almost surely.

REMARK *Kitamura and Stutzer (1997); Kitamura, Tripathi, and Ahn (2004) relax the assumptions in Qin and Lawless (1994) and obtain consistency of the estimator based on Wald's approach (Wald, 1949). Newey and Smith (2004) assume the differentiability of Lagrangian multiplier rather than that of $m(x,\theta)$. Schennach (2007) gives another consistency proof for a non-differentiable objective function and avoids applications of a Taylor expansion. The differentiability of moment constraint, however, is assumed in order to obtain a valid approximation for the Lagrangian $\lambda(\theta)$. In this paper, the assumptions are similar to the standard M-estimator conditions in Huber (1981), thus the differentiablity assumption is not required.*

2.2. *Local Empirical Likelihood and its Properties*

Classic asymptotic theory in statistics relies heavily on linear quadratic approximations to the logarithms of likelihood ratios, or on the criterion function of M- or Z- estimations. To make the linear quadratic approximations valid, a crucial step is to make sure the objective function is smooth enough. Global smoothness can be hard to guarantee especially in social science context. In addition to smoothness, many computational methods use the Hessian matrix for the linear quadratic approximation. This matrix can be difficult to evaluate especially in regions that are either extremely flat or very erratic.

The poor-behavior of methods that simply use the original moment constraints inspired people to think of alternative local estimation methods. Examples include Kitamura, Tripathi, and Ahn (2004), **?** and Kitamura, Otsu, and Evdokimov (2009). Here we propose a new local estimator that is asymptotically optimal in a local neighborhood and with a limiting normal distribution. It does not require a globally smooth objective function and the optimization only depends on local values of the EL. It is more computationally efficient than calculating the second order derivative of the objective function.

Local in our context refers to a neighborhood of θ_0 such that $\{\theta : |\theta - \theta_0| \leq \delta_n\tau\}$ with a positive $\delta_n \to 0$ when $n \to \infty$ and we consider cases where it is difficult to distinguish $P_{\theta_0,n}$ and $P_{\theta,n}$. We therefore define a local parameter $\tau = \delta_n^{-1}(\theta - \theta_0)$. The linear-quadratic approximations to the log-likelihood ratios can possibly be with other minimum contrast estimators, but such constructions only lead to asymptotically sufficient estimates, in the sense of Le Cam, when the contrast function mimics the properties of log-likelihood function, at least locally.

We use following condition to ensure the existence of a valid Taylor expansion for the log-likelihood ratio at θ_0 for every x.

CONDITION 2 (*i*) $\partial m(x,\theta)/\partial\theta < \infty$ *for any* x.
(*ii*) *The rank of* $\mathbb{E}[\partial m(x,\theta)/\partial\theta]|_{\theta_0}$ *equals* $\dim(\theta_0)$.
(*iii*) $\mathbb{E}[m_i(\theta)]^2 < \infty$.
(*iv*) $\mathbb{E}[\lambda_n m_i(\theta)]^2 < \infty$.

PROPOSITION 1 *Under conditions 1 and 2, the log-likelihood ratio between* $\tilde{p}_{\theta_0}$ *and* $\tilde{p}_{\theta_0+\delta_n\tau}$ *can be approximated by:*

$$\begin{aligned} 2\sum_{i=1}^{n}\log\frac{\tilde{p}_{\theta_0+\delta_n\tau}}{\tilde{p}_{\theta_0}}(x_i) = {} & \delta_n\tau^T\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1}\sum_{i=1}^{n}m_i(\theta_0) \\ & + \frac{1}{2}\delta_n^2\tau^T\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta^T}\tau \\ & + o_p(1). \end{aligned} \tag{2.2}$$

The result is intuitive as it mimics the standard Local Asymptotic Normal (LAN) for parametric models, see e.g. van der Vaart (1998, pp 104). The first term is τ times a random vector and the second term involves its variance. With the additional normality assumption on the average of $m_i(\theta_0)$ that is often made in the econometric literature, and assuming $\delta_n = n^{-1/2}$ we will of course have:

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1}m_i(\theta_0) \\ & \qquad \rightsquigarrow \mathcal{N}\left(0, \mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}\right). \end{aligned}$$

The expansion (2.2) is obtained simply by Taylor expansion and the result therefore does not apply to the nonstandard applications we are interested in. Asymptotic normality of the EL estimator is established by equation (2.2) with additional conditions on the continuity or the boundness of second derivative of the moment restriction functions, e.g. Qin and Lawless (1994), Newey and Smith (2004) or Kitamura, Tripathi, and Ahn (2004).

An alternative way of deducing this asymptotic normality is via Differentiable Quadratic Mean (DQM). This entails the existence of a vector of measurable functions $S_{\theta_0,n}$ such that

$$\int\left[\tilde{p}_{\theta_0+\delta_n\tau}^{1/2} - \tilde{p}_{\theta_0}^{1/2} - \frac{1}{2}\delta_n\tau^T S_{\theta_0,n}\tilde{p}_{\theta_0}^{1/2}\right]^2 d\mu = o(\|\delta_n\|^2), \tag{2.3}$$

where $\delta_n \to 0$ and μ is a dominating measure of $\tilde{p}_\theta$ such that $d\tilde{P}_\theta = \tilde{p}_\theta d\mu$. See e.g. van der Vaart (1998, p 65). The DQM implies the following condition which does not require the pointwise definition of the derivative of $m(\theta,x)$ therefore it is less restrictive than Condition 2.

CONDITION 3 *The square root of the implied probability* $\theta \mapsto \sqrt{\tilde{p}_\theta}$ *is differentiable at* θ_0.

Note the relation between the derivatives of the square root density and the score function (when it exists):

$$2\frac{1}{\sqrt{\tilde{p}_\theta}}\frac{\partial}{\partial\theta}\sqrt{\tilde{p}_\theta} = \frac{\partial}{\partial\theta}\log\tilde{p}_\theta.$$

If the Taylor expansion of the square root of $\tilde{p}_\theta$ is valid and the remainder term is negligible in $L^2(\mu)$ norm, $S_{\theta,n}$ can be considered as the score function of the implied probability $\tilde{p}_\theta$ at θ_0. However, the implied probability includes the term $m(\theta, x)$ which is not always differentiable in nonstandard cases that we want to consider. Now we relax Condition 3 to allow for non-differentiability.

CONDITION 4 *For any θ, there is a random vector $S_{\theta,n}$ and a random matrix $K_{\theta,n}$ such that*

$$\sum_{i=1}^{n} \log \frac{\tilde{p}_{\theta+\delta_n\tau_n}}{\tilde{p}_\theta}(x_i) - \left[\tau_n^T S_{\theta,n} - \frac{1}{2}\tau_n^T K_{\theta,n}\tau_n\right] \tag{2.4}$$

tends to zero in $\tilde{P}_\theta$ probability for any bounded sequence $\{\tau_n\}$.

This condition is to restrict the log-likelihood ratio to a so-called Local Asymptotic Quadratic (LAQ) family. It is weaker than previous condition 2 or 3. It only assumes that log-likelihood ratios of implied probabilities can be approximated by a linear-quadratic expression.

PROPOSITION 2 *For the implied probability $\tilde{p}_\theta$, condition 2 implies Condition 3. Condition 3 implies Condition 4. The converse implications do not hold.*

PROPOSITION 3 *The matrices $K_{\theta,n}$ are almost surely positive definite. Any cluster point K_θ of $K_{\theta,n}$ in $P_{\theta,n}$-law is invertible.*

In order to obtain an estimator in this setting, which allows for very weak conditions on $m(\theta, x)$, that belongs to the LAQ family, we use a Le Cam type estimator based on a δ_n-sparse (discretization of the) parameter space (see Le Cam and Yang (2000, p 125)). This requires a sequence of subsets $\Theta_n \subset \Theta$ satisfying the conditions (i) that for any $\theta \in \Theta$ and any $b \in \mathbb{R}^+$, the ball $B(\theta, b\delta_n)$ contains a finite number of elements of Θ_n, independent of n, and (ii) that there exist a $c \in \mathbb{R}^+$ such that any $\theta \in \Theta$ is within a distance $c\delta_n$ of a point of Θ_n. If we think of Θ_n as nodes of a grid with a mesh that gets finer as n increases, then (i) says that the grid does not get too fine too fast and (ii) says that the mesh refines fast enough to have nodes close to any point in the original space Θ.

DEFINITION 1 *Given conditions 1 and 4, we define the following Le Cam type local EL estimator in 5 steps:*

Step 1. Find an auxiliary estimate θ_n^* using a δ_n-consistent estimator [1] *and restricted such that it lies in Θ_n (a δ_n-sparse discretization of Θ).*

Step 2. Construct a matrix $K_n = \{K_{n,i,j}\}$, $i, j = 1, 2, \ldots, k$, where

$$\begin{aligned} K_{n,i,j} = &- \{\Lambda_n[\theta_n^* + \delta_n(u_i + u_j), \theta_n^*] \\ &-\Lambda_n[\theta_n^* + \delta_n u_i, \theta_n^*] - \Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*]\} \end{aligned}$$

and $\{u_1, \ldots, u_k\}$ a basis of $\mathbb{R}^k$. By proposition 3, K_n is invertible. [2]

[1] In our implementation, the auxiliary estimate used is the GMM estimate. But the result should hold for an arbitrary choice of auxiliary estimate given that the number of trails is large.

[2] Note that log-likelihood ratio $\Lambda_n[\theta_1, \theta_2]$ contains the implied probability $\tilde{p}_\theta$. While $\tilde{p}_\theta$ depends on the Lagrangian multiplier $\lambda(\theta)$. In principle, one has to solve the nonlinear equation (2.1) to obtain λ^*. Solving such an equation not only affects the computational speed but also gives an unreliable because the numerical inversion is ill-conditioned and the unsmooth moment restrictions. With the help of local parameters, we propose an implementation which

Step 3. Construct the linear term:

$$u_j^T S_n = \Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*] + \frac{1}{2} K_{n,j,j}.$$

Since all the right hand side values are known, S_n can be computed and is a proper statistics.

Step 4. Construct the adjusted estimator:

$$T_n = \theta_n^* + \delta_n K_n^{-1} S_n.$$

Step 5. Return the value of $\sum_i \log n\tilde{p}_{T_n}(x_i)$ and confirm that it is indeed larger than the likelihood $\sum_i \log n\tilde{p}_{\theta_n^*}(x_i)$ based on the initial estimator.

The method of construction was originally proposed by Le Cam (1974) to give a smooth local approximation to the log-likelihood ratio function. He supposed that there is no special interest in the likelihood function at particular points. The advantage of the construction is that the quadratic term does not depend very much on the particular estimator used to obtain the value of θ_n^* and is only determined by the local behavior in a local neighborhood of this point. Le Cam and Yang (2000) show the relation of the auxiliary estimate $\theta_n^*$to the Bayesian Gaussian prior. We will use this relationship in our proof.

THEOREM 2.2 *Given Condition 1 and Condition 4, T_n, S_n and K_n have the following properties:*

(i) $K_n^{-1}S_n$ and K_n converge pointwise to $K_{\theta,n}^{-1}S_{\theta,n}$ and $K_{\theta,n}$ respectively in $\tilde{P}_{\theta,n}$-law.

(ii) $\delta_n^{-1}(T_n - \theta)$ is bounded in $\tilde{P}_{\theta,n}$ -law.

(iii) if Equation (2.3) holds and the moment restrictions are just-identifying, the sequence of models $\{\tilde{P}_\theta : \theta \in \Theta\}$ is LAN and

$$\delta_n^{-1}(T_n - \theta_0) \rightsquigarrow \mathcal{N}(0, \Omega)$$

where $\Omega = \mathbb{E}\frac{\partial m(x,\theta_0)}{\partial \theta}^T \left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1} \mathbb{E}\frac{\partial m(x,\theta_0)}{\partial \theta}$.

The LAN theory is useful in showing that many statistical models can be approximated by Gaussian models. In the parametric likelihood framework, when the original model P_θ is smooth in the parameters, i.e. DQM, the local parameter $\tau = \delta_n^{-1}(\theta - \theta_0)$ can be used to construct a log likelihood ratio based on $P_{\theta_0+\tau\delta_n}$ that is $\mathcal{N}(\tau, I_{\theta_0}^{-1})$. Here we use LAN in a moment based setting without further parametric assumptions. Once LAN is established, asymptotic optimality of estimators and of tests can be expressed in terms of LAN properties.

---

avoids computing λ^*. Rewrite the denominator of $\tilde{p}_{\theta_1}/\tilde{p}_{\theta_2} = (1 + \lambda_1^T m(\theta_1))/(1 + \lambda_2^T m(\theta_2))$ as:

$$\frac{1}{1 + \lambda_2^T m(\theta_2)} = \frac{1}{1 + \lambda_1^T m(\theta_1) + \lambda_1^T \left[(\lambda_2/\lambda_1)^T m(\theta_2) - m(\theta_1)\right]}$$

When θ_1 and θ_2 differ only slightly, λ_2 and λ_1 share the same order, λ_2/λ_1 close to one and $m(\theta_1)$ is difficult to distinguish from $m(\theta_2)$ so we can set the value inside the square bracket equal to a small deterministic number c. Series expansions at zero for $1/(\lambda_1 m(\theta_1) + c)$ and $1/\lambda_1 m(\theta_1)$ shows that:

$$\frac{\lambda_1 \sum_{i=0}^{\infty}(-m(\theta_1))^i}{\lambda_1 \sum_{i=0}^{\infty}(m(\theta_1) + c)^i}.$$

Thus optimizing λ can be ignored in our algorithm process. The likelihood ratio in our implementation is $(1 + m(\theta_1))/(1 - m(\theta_1))$.

There are other articles that utilize local information in a EL framework. Donald, Imbens, and Newey (2003) propose resampling data from a local EL estimated distribution. Kitamura, Tripathi, and Ahn (2004) consider another localized EL based on conditional moment restrictions and use them to re-construct a smooth global profile likelihood function. Smith (2005) extends moment smoothing to GEL. Parente and Smith (????) apply kernel methods to implied probabilities because of non-smooth moment indicators. These methods construct smooth objective functions, implicitly or explicitly. In contrast, our purpose is to discretize the parameter space and then construct local log-likelihood ratios as local objective functions.

Theorem 2.2 gives an asymptotic result on the weak convergence of the estimator. In the theorem, the limit distribution is based on a kind of Cramer-Rao type lower bound and is essentially a pointwise result. In order to obtain a result in a neighborhood rather than at a single point, we will now state and prove a minimax type theorem on the risk of any estimator.

Before giving the theorem, we need to introduce the technical concept of δ_n-regularity which expresses the desirable requirement that a small change in the parameter should not change the distribution of estimator too much. Let T_n be an estimating sequence. If the difference between the distributions of $\delta_n^{-1}(T_n - \theta_0 - \delta_n\tau)$ and $\delta_n^{-1}(T_n - \theta_0)$ when $P_{\theta_0+\delta_n\tau,n}$ and $P_{\theta_0,n}$ hold respectively, tends to zero, then T_n is called δ_n-regular at the point θ_0.

THEOREM 2.3 *Given Condition 1 and Condition 4 and letting W be a non-negative bowl shaped loss function. If T_n isδ_n-regular on all Θ, then for any estimator sequence Z_n of τ, one has*

$$\lim_{b\to\infty}\lim_{c\to\infty}\liminf_{n}\sup_{|\tau|\le c}\mathbb{E}_{\theta+\delta_n\tau}\left[\min(b, W(Z_n-\tau))\right] \ge \mathbb{E}[W(\xi)]$$

where ξ has a mixed Gaussian distribution $K^{-1/2}\times\mathcal{N}(0,I)$ with $\mathcal{N}(0,I)$ is independent of K . The lower bound is achieved by $Z_n = \delta_n^{-1}(T_n-\theta)$.

A loss function is "bowl-shaped" if the sublevel sets $\{u : W(u) \le a\}$ are convex and symmetric around the origin. The value b is used to construct a bounded function $\min(b, W(Z_n-\tau))$. We let c go to infinity in order to cover a general case. The expectation on the LHS is taken w.r.t. the distribution induced by $P_{\theta+\delta_n\tau}$. The RHS expectation is taken w.r.t. on ξ over K and a standard normal variable.

The theorem can be interpreted as follows. The δ_n-regularity condition implies that if $K_{n,\theta} \rightsquigarrow K$ for $\tilde{P}_\theta$, then $K_{n,\theta}$ will still converge weakly to K in a neighborhood of θ, i.e. for $\tilde{P}_{\theta+\delta_n\tau}$. By the LAQ assumption, within the local neighborhood, the distribution $K_{n,\theta}$ for $\tilde{P}_\theta$ converges to a limited unconditional distribution of K. Marginal distribution of $K_{n,\theta}$ goes to K thus the joint distribution of $K_{n,\theta}$ and τ will converge to a mixed normal distribution. Given a standard normal variable $\xi \sim \mathcal{N}(0,I)$, the convoluted distribution of $K^{-1/2}\xi$ will be lower bound of the limited distribution of $K_n^{-1/2}S_n$. The term $K_n^{-1/2}S_n$ is related to $Z_n - \tau$, which constructs the measurement of Bayes risk by the expected loss function $\mathbb{E}[W(\cdot)]$. Especially, the lower bound of the risk can be obtained by letting Γ goes to infinity.

This is local asymptotic minimax theorem. It is based on the minimax criterion and gives a lower bound for the maximum risk over a small neighborhood of the parameter θ. Because the local EL can achieve this lower bound, it is asymptotic optimal estimation.

3. APPLICATIONS

In the previous section we develop moment based semi-parametric methods into the LAN and LAQ framework and we construct a local Le Cam type estimator. We obtain the pointwise con-

vergence and asymptotic distribution and establish a lower bound for the local minimax risk. Our conditions are weaker than standard assumptions for moment based methods or the ones used in a full parametric setting. We will now apply these results to two non-regular problems where the standard assumptions are too strong since they are violated in these cases.

3.1. *Robustness*

The robustness we consider here is essentially Huber's idea that an estimator is insensitive to perturbation of the model. This includes insensitivity to outliers if we think of a outliers as being generated by a different model with a small probability. When the assumed structure of the model is in-correct or the DGP is wrongly specified, one can detect, in principle, the mis-specification by various testing procedures. This kind of testing and the deletion of potential outliers, however, directly affects the inference procedures. It should in principle condition on the outcome of the test and take explicitly into account the statistical properties of deleting outliers. In this section we do not consider testing for misspecification and cleaning the data in advance, but analyse an inference procedure that explicitly takes into account that the model can, to a certain extend, be misspecified.

The sensitivity of EL estimation results from the unboundness moment constraints. We borrow Huber's setting to illustrate this problem for moment constraints. Consider an estimation as a statistical functional T such that

$$\int m(\hat{\theta}, x) dP_n = \int m(T(P_n), x) dP_n = 0.$$

The functional T could be the EL estimator if it is defined as:

$$T(P_n) := \arg\sup_{\theta \in \Theta} \sum_{i=1}^{n} \log n\tilde{p}_i(\theta),$$

but it could be any other estimation procedure that makes the above equality hold. A natural robustness requirement on a statistical functional is the boundedness of its influence function. The influence function of a given statistical functional $T(\cdot)$ is:

$$IF(x, T, P_n) := \lim_{\epsilon \to 0} \frac{T((1-\epsilon)P_n + \epsilon\Delta_x) - T(P_n)}{\epsilon}$$

for all xs that make the limit exist, where Δ_x is the probability measure giving mass 1 to x. An alternative way is to think of T as a linear functional which is continuous w.r.t a weak(-star) topology, namely the map

$$P \mapsto \int \psi dP = T(P)$$

from the space of all probability measures on the sample space to $\mathbb{R}$ is continuous whenever ψ is bounded and continuous. If ψ is not bounded, a single error can completely upset $T(P_n)$; if ψ is not continuous, a mass on these discontinuity points may cause a significant change for $T(P_n)$ with even a small change in subsample of $\mathcal{X}$. Note that the influence function ask for stronger condition. Because $IF(x, T, P_n)$ is defined by the functional derivative on Δ_x direction while in the linear functional it implies bounded and continuity.

For example, Ronchetti and Trojani (2001) show that the influence function of exactly identified GMM is $[\mathbb{E}\partial m_i(T(P_\theta))/\partial T]^{-1} m_i(T(P_\theta))$ at a single observation x_i. It is obvious that the influence function may be unbounded if $m(\theta)$ is unbounded. An unbounded influence function implies an unbounded asymptotic bias of a statistic at single-point contaminations of the model. Suppose the moment restriction is a linear function $\sum Z^T(Y - X\beta)/n$, the unboundness says that $\sup_i z_i y_i = \infty$, thus the IV estimate is constructed by $\psi_i = (S_{XZ}^T S_{ZZ}^{-1} S_{XZ})^{-1} S_{XZ}^T S_{ZZ}^{-1}(yz_i)$ which is unbound, where S_{XZ} is the sample average of XZ. So unboundness assumption makes GMM estimator steep in some contamination directions.

White (1982) shows that Maximum Likelihood defined in the Kullback-Leibler divergence is robust. EL inherits a lot of properties from MEL, so one may expect it is also robust. However, Schennach (2007) gives a counterexample. Suppose the outliers of sample space $\mathcal{X}$ give $\sup_{x\in\mathcal{X}} m_i(\theta) = \infty$ so that $\inf_\theta \sum m_i(\theta)/n \neq 0$ for any $\theta \in \Theta$ but $\mathbb{E}[\|m(\theta, x)\|^2] < \infty$. The λ associating with these outliers' moment restriction functions will give strong penalties so that their λs stay close to zero independently of the θ's value. The implied density $\tilde{p}_i$ of each outlier' moment restriction function equals to $1/n$. When the sample size is very small which means that the effects of relative weights on the outliers are strong, the criterion function $n\sum \log \tilde{p}_i(\theta)$ will be quite flat on θ. Therefore the intrinsically misspecified EL estimator will have a slower convergent rate, though it is consistent.

Although EL does not have an analytical solution, its statistical functional is a solution of the moment restriction function:

$$\int m(T(\tilde{P}), x) d\tilde{P} = 0. \tag{3.1}$$

The moment restriction function $m(T(\tilde{P}), x)$ as in the previous discussion is discontinuous on those peculiar points and (3.1) is non-differentiable on these points as well, thus EL is not a robust estimation[3]. Surprisingly, a similar approach, Empirical Tilted (ET) (Imbens, Spady, and Johnson, 1998; Kitamura and Stutzer, 1997), is robust under this unboundness assumption. The intrinsic feature that ET preserves the weak(-star) continuous statistical functional is that ET's implied probability can automatically eliminate those peculiar points so that ψ preserves the continuity. We use the strategy of Schennach (2007, Theorem 8) to briefly illustrate this feature. Suppose $m_{i,a}$ and $m_{i,b}$ are subvectors of $m_i(\theta)$ and they are mutually independent such that $\mathbb{E}m_a(\theta, x)m_b(\theta, x) = \mathbb{E}m_a(\theta, x)\mathbb{E}m_b(\theta, x)$. If the values of $m_{i,b}$ on some peculiar points do not dependent on θ, there is

$$\begin{aligned}\mathbb{E}_{\tilde{p}_{ET}} m_a(x) &= \sum m_{i,a} \frac{\exp(\lambda(\theta)^T m_i(\theta, x))}{\mathbb{E}\exp(\lambda(\theta)^T m(\theta, x))} \\ &\to \frac{\mathbb{E}m_a(x)\exp(\lambda(\theta)^T m_a(x))}{\mathbb{E}\exp(\lambda(\theta)^T m_a(x))}\end{aligned}$$

with ET's implied probability $\tilde{p}_{ET} = \exp(\lambda(\theta)^T m_i(\theta, x))/\mathbb{E}\exp(\lambda(\theta)^T m(\theta, x))$. Thsu in (3.1) $P_{n,ET}$ assigns no mass on those peculiar points. In other words, we can informally think that ET's implied probability clears away the effect of "bad observations" and preserves the well-defined part of the moment restriction function. Schennach (2007) introduces this two-step correction approach and calls it Exponential Tilted Empirical Likelihood estimation (ETEL).

So it is clear that the lack of robustness in EL approaches because of the non-robust moment restriction functions not the EL procedure itself. If one can eliminate the outlier's influences, one will keep the robustness of EL. We propose that local EL, as an alternative way to ETEL, can

[3] Schennach (2007) uses the influence function to indicate the non-robustness property.

prevent such misbehaviors via a similar trick as ET. Note that the auxiliary estimator in local EL θ_n^* in the selected range of order δ_n does admit a good quadratic approximation of the log-likelihood ratios. This requirement regularizes the matrix in quadratic term to be semi-positive definite. Remember that the outliers occurrence will make the log-likelihood very flat over Θ, thus the difference between two log-likelihoods with a small change on θ will be close to zero. In such a case, there is no linear-quadratic formula that can fit the log likelihood ratio approximation. Actually, the construction of local EL already eliminates those unhealthy points. In step 2 of the local EL's construction, we let

$$\begin{aligned} K_{n,i,j} = &- \{\Lambda_n[\theta_n^* + \delta_n(u_i + u_j), \theta_n^*] - \Lambda_n[\theta_n^* + \delta_n u_i, \theta_n^*] \\ &-\Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*]\} \end{aligned}$$

If the peculiarities $\{x\}_k$ drive the moment restrictions function on u_i direction unbounded, there is

$$\begin{aligned} &\left\{\log \frac{\tilde{p}_{\theta_n^*+\delta_n(u_i+u_j)}}{\tilde{p}_{\theta_n^*}}(\{x\}_k) - \log \frac{\tilde{p}_{\theta_n^*+\delta_n u_i}}{\tilde{p}_{\theta_n^*}}(\{x\}_k) - \log \frac{\tilde{p}_{\theta_n^*+\delta_n u_j}}{\tilde{p}_{\theta_n^*}}(\{x\}_k)\right\} \\ =&\left\{\log \frac{\tilde{p}_{\theta_n^*+\delta_n u_j}}{\tilde{p}_{\theta_n^*}}(\{x\}_k) - \log \frac{\tilde{p}_{\theta_n^*}}{\tilde{p}_{\theta_n^*}}(\{x\}_k) - \log \frac{\tilde{p}_{\theta_n^*+\delta_n u_j}}{\tilde{p}_{\theta_n^*}}(\{x\}_k)\right\} = 0. \end{aligned}$$

The second step in above equation simply use the property that outliers are independent of specific values of θ. Therefore, the peculiar set $\{x\}_k$ plays no role in the quadratic construction. Figure 1 shows the residual distributions of four estimate methods for a simulated model under $n = 500$. The concrete discussion about the simulation procedure will be given in the following section. From the figure, local EL has the most concentrated bell shape distribution because the quadratic term in the objective function give strong penalties for large residuals. Although the quadratic term drop out the observations far from sample means, the local EL method captures fat tail behavior in small sample simulations by the linear term. The linear term performs equivalently for the observations while its shrinking local parameter mitigates the intrinsic mis-specification effect. We conclude that local EL is a robust estimation.

REMARK *One may ask why we should use robust estimation rather than give a mis-specification test or clean the data first. Essentially, there is no "mis-specification" in the model, although unboundness causes the constraints leave from zero. Blindly applying mis-specification test will reject a correct model. An insightful argument is given by Huber (1981).*

> Even if the original batch of observations consists of normal observations interspersed with some gross errors, the cleaned data will not be normal, and the situation is even worse when the original batch derives from a genuine non-normal distribution, instead of from a gross-error framework. Therefore the classical normal theory is not applicable to cleaned samples, and the actual performance of such a two-step (clean and test) procedure may be more difficult to work out than that of a straight robust procedure. -(Huber, 1981, Chapter 1)

REMARK *Schennach (2007) shows that the ETEL estimation is robust and almost as efficient as EL. However, it is still less efficient in the higher order than Local EL or EL. Moreover, the procedure of ETEL changes the divergence criterion in the intermediate step*[4]*. Despite both EL and ET base on the same convex hull, the different divergence criterion functions give different appearances*

[4]In first step, the criterion function is to minimize the log-likelihood ratio over empirical entropy while in the second step the criterion function is to minimize the log-likelihood ratio over sample average.

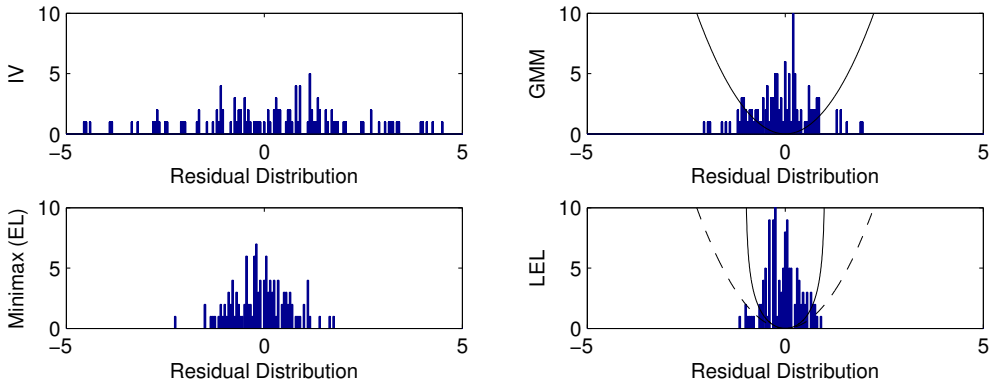

FIGURE 1.— A stronger penality makes LEL's residuals more concentrated.

of the re-shaped contour of the implied probability sequence $\{\tilde{p}_1, \ldots, \tilde{p}_n\}$. *The complement of the overlapping contour region is assigned zero measure by ETEL. This action will discard not only the weights of outliers but also some informative weights used to capture the fat tail feature of sample distributions. Ronchetti and Trojani (2001) construct a Huber-type GMM estimators based on a bounded self-standardized norm of the given orthogonality function. They also show that imposing this robustness correction has an impact on the power of the mis-specification test.*

3.2. *Weak Identification*

In GMM literature, people focus on an intermediate situation between strong identification, namely a unique solution of $\mathbb{E}m(\theta) = 0$ for $\theta \in \Theta$, and unidentification of the interested parameters, namely, there are more than one solution for $\mathbb{E}m(\theta) = 0$. The case is called weak identification. The idea dates back to Hansen, Heaton, and Yaron (1996); Stock and Wright (2000); Stock, Wright, and Yogo (2002). To construct this problem, suppose that θ is partitioned into two parts, the "good" part and the "bad" part. The "bad" part may have two possibilities: *1.* the objective function is not convex, namely the optimization procedure does not have a unique solution or *2.* the objective function becomes flat around the region of the true parameters but its graph is still convex[5], for instance, an objective function has higher full rank derivatives but its Jacobian and Hessian are both zero at the underlying θ_0. Generally speaking, when an objective function for some sub-parameters is very flat so that its value has insignificant or none variation in a set of these parameters, the model is weakly identification. Therefore the general regularization assumptions cannot guarantee consistent GMM-type estimators.

For econometricians, case *1* is an un-identification case for the point estimation while case *2* is a weak identified issue. The objective function in *2* is local convex over the neighborhood of the true parameters from the "bad" part, thus the estimated parameter from "bad" part is still able to be estimated consistently but with the convergent rate less than that of the "good" parameters, namely the parameters are local identifiable. If one drops this assumption on local convexity, then the restrictions of the "bad" part are non-informative so that it is meaningless to put these parameters into the model. So both cases *1* and *2* induce inconsistent estimators and people can use testing method to detect these issues, but in case *2* it is also possible to rescue the consistency. We will

[5]It means the set $\cup_{\theta \in B(\theta_0, \varepsilon)}(\theta, M(\theta))$ is a convex set for a subcover of θ_0.

give a formal argument about this procedure later.

Weak identification idea comes from the poor finite sample behavior of the approximations to the statistics, see Hansen, Heaton, and Yaron (1996). This issue induces a lots of works on bias correction and constructing pivotal testing under the weak identification, to name few of them, Guggenberger and Smith (2005, 2008); Kleibergen (2002, 2005). These paper concern on distilling the statistics of "good" parameters. Usually, these testings form a Wald type quadratic formula based on $L^2(\Pi)$ matrix distance such that

$$\|x - x_c\|_{\Pi}^2 = (x - x_c)^T \Pi^{-1} (x - x_c),$$

where Π determines how far the ellipsoid extends in every direction from x_c. The eigenvalues of Π concentrate on the statistics only with "good" parameters[6]. These tests depend on prior knowledge of the number of weakly identified or unidentified parameters. Kleibergen (2005) points out that the Jacobian of the moment function of GMM estimation has degeneration possibility in the weak identification case. He corrects the Jacobian estimator via accounting for the non-singular high order term of an approximation so that the statistic is asymptotically uncorrelated with these nuisance parameters. These testings exploit the crucial feature of the weak identification problem, singularity. However, because classical test statistics depend on non-singular information matrices, people are reluctant to extend the singularity condition to the Hessian matrix. We will show that local EL can handle this problem.

In the following section, we show that the weak identification can be interpreted as an issue of different convergent rates of the parameters. We first describe the setting of weak identification in Stock and Wright (2000) (SW henceforward) and then go to a setting where the local EL can be implemented.

CONDITION *[SW] (i) Partition θ as $\theta = (\alpha, \beta)$ where α is weakly identified and β is strongly identified. (ii) Let*

$$\begin{aligned}\mathbb{E}[\textstyle\sum m_i(\alpha,\beta)/n] =& \mathbb{E}[\textstyle\sum m_i(\alpha_0,\beta_0)/n] \\ &+ \overbrace{\Big(\mathbb{E}[\textstyle\sum m_i(\alpha,\beta)/n] - \mathbb{E}[\textstyle\sum m_i(\alpha_0,\beta)/n]\Big)}^{\tilde{m}_1(\theta)} \\ &+ \underbrace{\Big(\mathbb{E}[\textstyle\sum m_i(\alpha_0,\beta)/n] - \mathbb{E}[\textstyle\sum m_i(\alpha_0,\beta_0)/n]\Big)}_{m_2(\beta)}.\end{aligned}$$

(iii) $\tilde{m}_1(\theta) = n^{-1/2} m_1(\theta) + o(n^{-1/2})$, $m_1(\theta)$ is bounded. (iv) $\tilde{m}_1(\theta) \to m_1(\theta)$ uniformly in $\theta \in \Theta$, $m_1(\theta_0) = 0$, and $m_1(\theta)$ is continuous in θ. (v) $m_2(\beta_0) = 0$, $m_2(\beta) \neq 0$ for $\beta \neq \beta_0$, $\partial m_2(\beta)/\partial \beta'$ is continuous and full column rank. (vi) $n^{-1/2} \sum_i^n [m_i(\theta) - \mathbb{E} m_i(\theta)]$ is a Gaussian process on $\theta \in \Theta$, specially when $\theta = \theta_0$, is a normal distribution with zero mean and variance Ω.

In SW's condition *(iii)*, $m_1(\theta)$ is assumed to be bounded, so it will be at least $O(1)$ or $O(n^{-\delta})$ for $\delta > 0$ or $o(1)$. The choice of $n^{-1/2}$, according to Stock and Wright (2000), is to yield tractable asymptotic approximations to the distributions of estimator and test statistics. Namely, to avoid

[6]For example, Π in Kleibergen (2002) is a product of the score function of concentrated likelihoods of "good" parameters. Π in LM test Guggenberger and Smith (2005) is a sandwich term of the objective function's derivative, which is assumed to be non-singular on the "good" parameters.

the concentration of a statistics at the rate $n^{-1/2}$, SW give the weakly identified term a certain rate so that $n^{1/2}(\sum m_i(\alpha)/n - \mathbb{E}m_i(\alpha))$ will not vanish. Usually, the weakly convergence proof requires the condition of well-separation of parameters[7], namely in this case α_0 should be separated from the other α. One can, like SW, multiply a term $n^{1/2}$ to keep non-degeneration of the weak identified terms, however we use an alternative setting in order to local EL feasible to estimate the weakly identified models.

We try to follow the idea of SW and introduce a substitution of condition SW (iii) to describe a similar phenomena. Since (vi) in condition SW implies

$$\mathbb{E}^* \sup_{d(\alpha,\alpha_0)<\varepsilon} \sqrt{n}(\sum m_i(\alpha)/n - \mathbb{E}m_i(\alpha_0)) \leq C\varepsilon^{K'} \tag{3.2}$$

with $K^{'} = 1$. Consider $\sqrt{n}(\sum m_i(\alpha)/n - \mathbb{E}m_i(\alpha_0))$ as an empirical process indexed by α, Kolmogorov's continuity criterion[8] requests $K' = 1$ to preserve the Gaussian process properties. Next, we consider $\tilde{m}_1(\theta)$ such that

$$\sup_{d(\alpha,\alpha_0)<\varepsilon} \mathbb{E}[\sum m_i(\alpha)/n - \sum m_i(\alpha_0)/n] \leq -C\varepsilon^{K} \tag{3.3}$$

with $K \geq 2$. To see the argument $K \geq 2$, we give a Taylor expansion for $\tilde{m}_1(\theta)$ at the point of maximum α_0:

$$\begin{aligned}\mathbb{E}\left[\frac{\sum m_i(\alpha)}{n} - \frac{\sum m_i(\alpha_0)}{n}\right] = \frac{1}{2}(\alpha-\alpha_0)^T \\ \times \frac{\partial^2}{\partial\alpha^2}\,\mathbb{E}\left[\frac{\sum m_i(\alpha)}{n} - \frac{\sum m_i(\alpha_0)}{n}\right]\Bigg|_{\alpha_0} \\ \times (\alpha-\alpha_0) + o(\|\alpha-\alpha_0\|^2).\end{aligned}$$

The first derivative at α_0 vanishes. When the second derivative matrix is nonsingular, namely fully identified case, it is obvious that LHS in (3.3) is controlled by $\|\alpha-\alpha_0\|^2$ so that the order of K is two. The weakly identified $\tilde{m}_1(\alpha)$ is very flat round α_0, thus its second derivative matrix is singular. When the second derivative matrix is singular, we should use higher order terms to express the equation, so that K will be larger than two. For example, if third derivative matrix is non-singular, K can be set to three.

With (3.2) and (3.1), the rate of convergence theorem[9] shows that $n^{1/(2K-2K')}d(\hat{\alpha},\alpha) = O_p(1)$. $n^{1/(2K-2)}$ is the rate to control the convergent speed of α and is less than the convergent rate of $\hat{\beta}$ which is $n^{1/2}$. Therefore, we adapt the assumption to:

CONDITION *[SW(iii)'] $\sup_{d(\theta,\theta_0)<\varepsilon} m(\theta) \leq -C\varepsilon^K$, where C is a constant, the vector $K = (K_1,\ldots,K_k)$ is the order such that $\partial^{K_i} m(\theta_i)/\partial\theta_i^{K_i}$ is non-singular and that for any $l < K$ the matrix $\partial^l m(\theta_0)/\partial\theta^l$ is singular. A subset α of θ has a sub-vector of K with values either larger than two or $\emptyset$.*

[7] For example Lemma 3.2.1 in van der Vaart and Wellner (1996).

[8] The continuous Gaussian process $\mathbb{G}$ and its continuous covariance kernel (Assumption B in Stock and Wright (2000)) imply that $\mathbb{G}$ is equicontinuous w.r.t its sample path (Oodaira, 1973, Theorem 1). Since $\sqrt{n}(\sum m_i(\alpha)/n - \mathbb{E}m_i(\alpha_0)) = \mathbb{P}_\alpha$ is the modification of $\mathbb{G}$, Kolmogorov's continuity criterion implies $\mathbb{E}\|\mathbb{P}_s - \mathbb{P}_t\| \leq C|s-t|$.

[9] Theorem 5.52 in van der Vaart (1998). It says that if both (3.2) and (3.1) satisfied and the estimator is local identifiable, then $n^{1/(2K-2K')}d(\hat{\alpha},\alpha) = O_p(1)$ holds.

When $K = 2$ the parameters are strongly identified, when $K > 2$ the condition illustrates the curvature of the objective function on α is zero so that the parameters are nearly flat around α_0. For the unidentified case, α is just flat on a neighborhood of α_0, therefore all the higher order terms in the expansion will equal to zero and there is no K existent.

This condition allows for different convergent rates amongst θ. We give a simple example on second derivative of the moment restriction. Let $M_2 = \partial^2 m(\theta)/\partial\theta\theta^T$, θ is a $k \times 1$ vector and M_2 is $k \times k$ symmetric matrix with real eigenvalues. Choose the eigenvectors of such matrix to form an orthonormal basis of $\mathbb{R}^k$. Let κ be the eigenvalues and $\mathbf{u}_i$ be the corresponding eigenvectors. Because $M_2\mathbf{u}_i = \mathbf{u}_i\kappa_i$. For the matrix $\mathbf{U} = [\mathbf{u}_1, \dots \mathbf{u}_k]^T$, we have $M_2\mathbf{U} = \mathbf{UD}$, where $\mathbf{D}$ is the diagonal matrix with the eigenvalues on the diagonal. Since $\mathbf{U}$ is an orthogonal matrix, there is $M_2 = \mathbf{UDU}^T = \sum_{i=1}^k \kappa_i \mathbf{u}_i \mathbf{u}_i^T$. Let $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_k$, when $i > r$, $\kappa_i = 0$. If the third derivative is full ranked, $\{\kappa_i\}_{i>r}$ corresponds to the "poor" parameters α with $K = 3$. Therefore, α is weakly identified and has a slower convergent rate $n^{1/3}$ than strongly identified β.

REMARK *The condition seemingly does not account for a popular case, so-called weak instruments' problems in linear simultaneous models. Weak instruments problem is close to the situation that curvature of $m(\theta)$ is nearly singular but still has full-ranked positive semi-definite matrix and there is no higher order expansion terms. The usual setting in weak instrument problems*[10] *is:*

$$y = X^T\beta + \varepsilon$$
$$X = Z^T\Pi + e,$$

with $\mathbb{E}[X^T\varepsilon] \neq 0$ and both $\mathbb{E}[Z^T e]$ and $\mathbb{E}[Z^T\varepsilon]$ are zero. Z is a weak instrument means that $\mathbb{E}[ZX]$ is very small, namely Z has little explanatory power for X. The orthogonal equation is $\sum[Z(y - X^T\beta)]/n = 0$. The usual least square estimation or GMM estimation with a deterministic weighting matrix on this linear orthogonal equation have no third order derivative w.r.t β and the second derivative is a matrix of Z^TX which is nearly singular. The orthogonal equation plays as a solo constraint for the model's optimization, however, this constraint is almost slack everywhere on the parameter space. This is a serious ill-posed problem. In order to rescue the estimation, one has to put additional constraints or assume a more restrictive parameter space. Conversely if one agrees that "weak IV" problem is a mis-specification issue such that one incorrectly use simple linear regression to interpret a non-linear phenomena, then our weak identification phenomena could explain what happens in this "weak instrument" problem.

However, these singular matrices make the optimization algorithm unhealthy. The usual gradient method and Newton-Raphson method cannot give an informative searching direction. Simplex methods for finding a local optimum solution, like Nelder-Mead method or Powell's method, are alternatives. But when the number of variables/parameters increases, the edge of each face of the simplex grows as a polynomial number which makes the optimum searching extremely difficult. A Newton-Raphson type algorithm should be preferable. Actually, local EL bases on a discretization trick coupled with a Newton-Raphson approximation. Therefore, we can apply local EL to solve the weak identification problem. To illustrate this is feasible, we just apply previous results. Local EL can control the shrinkage rate δ_n in the local neighborhoods. If δ_n coincides with the convergent rate of weakly identified parameters, by corollary 2, the differentiability of moment restriction functions induces the LAQ condition. Theorem 2.2 indicates that K_θ is positive definite.

[10] The symbols used here are temporary. The aim is to accord with huge amounts of weak instruments' literature.

4. SIMULATION RESULTS

We consider a standard structural model, with $d \times m$ explanatory matrix X, instrument matrix Z and uncertainty matrix U, $m \times 1$ parameter vector β and $d \times 1$ i.i.d random vector ε:

$$Y = X\beta + \varepsilon$$
$$X = Z + U$$

The random vector and uncertainty matrix are both generated by standard normal distributions with the same random seed so that ε and U are correlated. Z is a deterministic matrix independent with ε and U. We assume Y and X are deterministic.

We compare four different estimate methods, Least Square, GMM, Minimax, and local EL, while minimax is the commonly used in EL-related estimations. All estimations are implemented on a convex optimization solver `SDPT3` [11] which requires the objective function to be strictly convex. The moment restriction function is $Z^T(Y - X\beta)$. For GMM (or Generalized IV), the weighting matrix is the optimal $\mathbb{E}\|Z^T(Y - X\beta)\|^2$. Since the distributions of U and ε are unknown, we use the sample counterpart $\mathbb{E}U^TU = P$:

$$\begin{aligned}\min_\beta \mathbb{E}\|Z^T(Y - X\beta)\|^2 =& \mathbb{E}(Z^TY - Z^TZ\beta - Z^TU\beta)^T(Z^TY - Z^TZ\beta + Z^TU\beta) \\ =& (Z^TY - Z^TZ\beta)^T(Z^TY - Z^TZ\beta)^T + \beta^TZ^TPZ\beta.\end{aligned}$$

The EL-related estimates are implemented by a standard minimax optimization:

$$\min_\beta \max_{\lambda \in \Lambda} \mathbb{E}\log(1 + \lambda^TZ^T(Y - X\beta)).$$

Instead of solving the equation (2.1) to construct the set Λ, we simply look for the worst case results of the moment restriction functions. So the inner loop operation becomes

$$\max\left\{Z_i(Y_i - Z_i\beta - U_i\beta),\ Z_i(Y_i - Z_i\beta + U_i\beta)\right\}_{i\in\{1,\ldots,d\}}.$$

This is a duality of optimizing λ, since the maximal Lagrangian multiplier associate with the worst cast constraint[12]. One should notice that if estimating the implied probability is of the main interest, Lagrangian multiplier is the most crucial element. However, in this simple example, our interest is on estimating parameters.

In Figure 3, the lines with squares, dots, and cross represent GMM (or GIV), Minimax and local EL respectively. The straight line is the least square estimate without considering endogenity issues. From top left to bottom right, the figures respectively indicate the cases: $d = 20, m = 10$; $d = 50, m = 20$; $d = 100, m = 20$; $d = 300, m = 20$. It is clear that for $U \geq 0$, local EL achieves the lowest residual among the four methods in these cases. While for $U < 0$, under small number of parameters case, local EL has similar performances as GMM and is not as good as minimax method. When d grows larger, except LS, the methods have close outcomes. Figure **??** shows the residual distributions under $d = 20, m = 10$; $d = 50, m = 20$; $d = 100, m = 20$. It seems that the residuals of local EL are more concentrated than the others. Minimax and GMM are also more concentrated than LS but the residuals spread evenly around the mean. Local EL has less wide spread residuals comparing to GMM and Minimax and the bell shape is more significant around the mean.

[11] This is a Matlab supported solver for semidefinite-quadratic-linear programming problems. It is free and available for downloading via `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`.

[12] This simplification may be not as tractable in highly nonlinear constraints as in this setting.

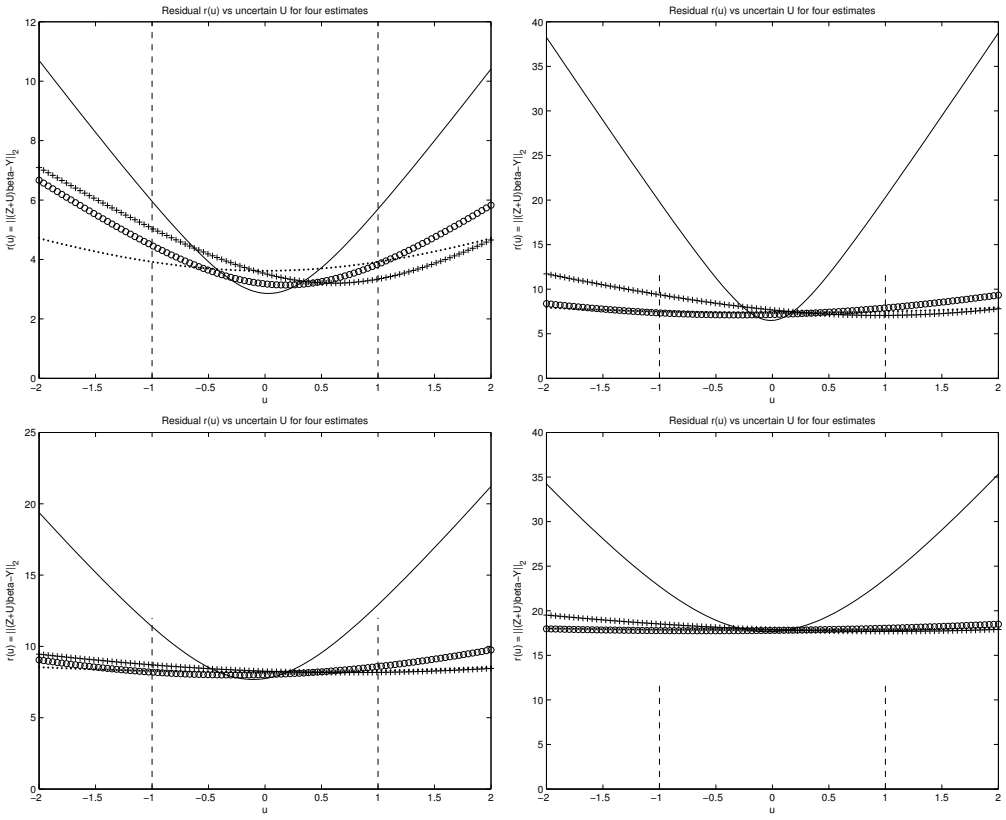

FIGURE 2.— The residual $\|Y - X\hat{\beta}\|^2$ as a function of the uncertain matrix U.

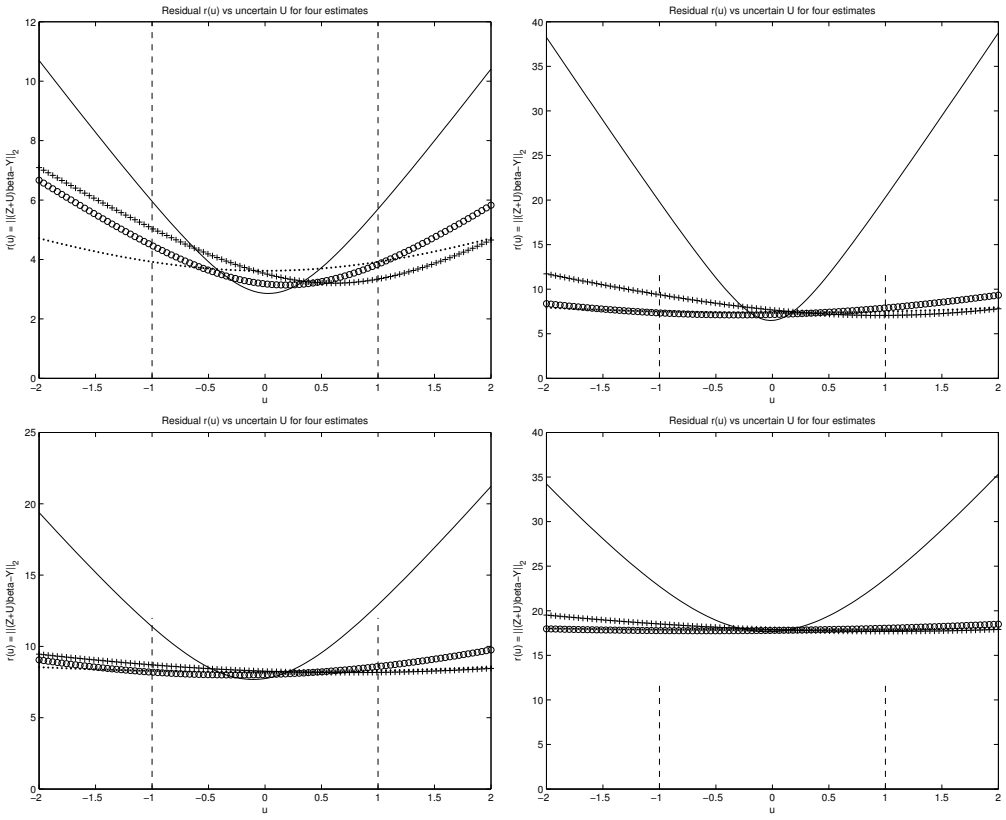

FIGURE 3.— The residual $\|Y - X\hat{\beta}\|^2$ as a function of the uncertain matrix U.

5. CONCLUSION

This paper considers a modified empirical likelihood and uses it to estimate models with imperfect constraints. Consistency, local normality and asymptotic optimality of the local method only require the local identifiability and the linear-quadratic representation. In imperfect situations, classical sandwich algorithms give bias estimations while the new estimator preserves its asymptotic properties. We consider two imperfect cases as applications, unbounded moment restrictions and weakly identified moment restrictions. Finally, a simple simulation shows some satisfied outcomes which coincide with our expectations.

APPENDIX

PROOF OF THEOREM 2.1: The EL objective function with Lagrange multipliers is

$$\mathcal{L} = \sum_{i=1}^{n} \log(np_i) - n\lambda^T \sum_{i=1}^{n} p_i m_i(\theta) + \gamma \left(\sum_{i=1}^{n} p_i - 1 \right),$$

where λ and γ are Lagrange multipliers. The partial derivative of $\mathcal{L}$ w.r.t. p_i gives $\gamma = n$. So we have implied probability $\tilde{p}_i = 1/(n + n\lambda^T m_i(\theta))$. By implicit function theorem, the partial derivative of $\sum_{i=1}^{n} \log \tilde{p}_i$ w.r.t λ gives a function of θ such that

$$\begin{aligned} &\frac{\partial \sum \log \tilde{p}_i}{\partial \lambda} = 0, \\ &\Longrightarrow \frac{1}{n} \sum_{i=1}^{n} \frac{m_i(\theta)}{1 + \lambda_n(\theta)^T m_i(\theta)} = \sum_{i=1}^{n} \tilde{p}_{\theta,i} m_i(\theta) \end{aligned} \tag{5.1}$$

where $\lambda_n(\theta)$ is unique for fix n and θ. Note that $(\partial \sum \log \tilde{p}_i / \partial \lambda)(\theta) = 0$ for $\forall \theta \in \Theta$, hence $(\partial \sum \log \tilde{p}_i / \partial \lambda)(\theta)$ is continuous on θ. By the continuity of $m(x, \theta)$, we have $\lambda_n(\theta)$ is also continuous on θ. The proof of the uniqueness of limited $\lambda_n(\theta)$ is straightforward, because the set $\Gamma(\theta) = \lim_{n\to\infty} \cap_{i=1,\ldots,n} \{\lambda | 1 + \lambda m_i(x, \theta) > 1/n\}$ is convex if it is not vanish, the function of $\log p$ is strictly concave on λ, so $\lambda(\theta)$ uniquely exists. Equation (5.1) can be re-written as

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \left[1 - \frac{\lambda_n(\theta)^T m_i(\theta)}{1 + \lambda_n(\theta)^T m_i(\theta)} \right] m_i(\theta) = 0 \\ \Longrightarrow \frac{1}{n} \sum_{i=1}^{n} m_i(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{m_i(\theta) \lambda_n(\theta)^T m_i(\theta)}{1 + \lambda_n(\theta)^T m_i(\theta)} \\ = \left[\sum_{i=1}^{n} \tilde{g}_{\theta,i} m_i(\theta) m_i(\theta)^T \right] \lambda_n(\theta). \end{aligned}$$

Assumption 1 (iii) shows that $m_i(\theta) m_i(\theta)^T$ is positive definite, let $\mathbf{c}$ is the maximum eigenvalues of $\mathbb{E}(m(\theta) m(\theta)^T)$ and v is the corresponding eigenvector. The convex combination of $m_i(\theta) m_i(\theta)^T$ is bounded by $v^T \mathbf{c} v$. Let $K = \|v^T \mathbf{c} v\|$. According to Assumption 1 (iv), $m_i(\theta)$ has an envelop function $b(\theta)$ such that $\liminf_\theta |m(x, \theta)|/b(\theta) \geq 1$, $\lim_{n\to\infty} \lambda_n(\theta)/b'(\theta) \geq 1$ for any θ where $b'(\theta) = b(\theta)/K$.

The optimization problem becomes

$$\theta = \arg\max \sum_{i=1}^{n} \log n\tilde{p}_{\theta,i}/n = \sum_{i=1}^{n} \Lambda_n(\theta)/n.$$

Let's first consider the existence of the limitation.

$$\lim_{n\to\infty} \mathbb{E}\frac{1}{n}\frac{n}{1+\lambda_n(\theta)^T m(x,\theta)} = \mathbb{E}\lim_{n\to\infty}\frac{1}{1+\lambda_n(\theta)^T m(x,\theta)} = \mathbb{E}\frac{1}{1+\lambda(\theta)^T m(x,\theta)}. \tag{5.2}$$

The above equation is obtained by dominated convergence theorem, since $[1+\lambda_n(\theta)^T m(x,\theta)]^{-1}$ is bounded and the limit of $\lambda_n(\theta)$ exists. Next we consider the continuity of $\lambda(\theta)^T m(x,\theta)$. Given n, we know that $\lambda_n(\theta)^T m(x,\theta)$ is continuous. The envelop functions $b'(\theta)$ and $b(\theta)$ are integrable and continuous (Assumption 1), $\lambda(\theta)^T m(x,\theta)$ is bounded by a continuous function. The continuity of $m(x,\theta)$ implies $\lambda(\theta)$ is also continuous.

Now choose a large enough compact set $\Theta_c \subset \Theta$ such that given δ

$$\mathbb{E}[v(x)] = \sup_{\theta\neq\Theta_C} \frac{|\sum \Lambda_n(\theta)/n - \mathbb{E}\Lambda(\theta)|}{b^2(\theta)} \leq 1-3\delta$$

and $\inf_{\theta\in\Theta_C}|\Lambda(\theta)|/b(\theta) \geq 1-\delta$, by Assumption 1. The strong LLN implies

$$\sup_{\theta\neq\Theta_C} \frac{\left|n^{-1}\sum[\Lambda_n(\theta)-\mathbb{E}\Lambda(\theta)]\right|}{b^2(\theta)} \leq \frac{1}{n}\sum v(x_i) \leq 1-2\delta;$$

therefore

$$\begin{aligned}\left|\frac{1}{n}\sum[\Lambda_n(\theta)-\mathbb{E}\Lambda(\theta)]\right| &\leq (1-2\delta)b(\theta)\\ &\leq \frac{1-2\delta}{1-\delta}|\mathbb{E}\ell_\theta| \leq (1-\delta)\mathbb{E}\ell_\theta\end{aligned}$$

for $\forall\theta \neq \Theta_C$. This equation implies, for any δ

$$\left|\frac{1}{n}\sum\Lambda_n(\theta)\right| \geq \delta|\mathbb{E}\Lambda(\theta)| \geq \delta(1-\delta)b_0.$$

The last inequality indicates that the any sequence T_n satisfying $\sum\Lambda_n(T_n)/n \to 0$ will locate in the compact set ultimately.

Now the rest of the proof follows Wald's method. The continuous $\lambda(\theta)m(x,\theta)$ in any x gives

$$\lim_{\theta'\to\theta}|\Lambda(\theta')(x)-\Lambda(\theta)(x)| = 0.$$

For any decreasing sequence U_l of open neighborhoods around θ of diameter l converging to zero, $\sup_{\theta'\in U_l}\Lambda(\theta')$ decrease for every l. So when l goes to 0, $\sup_{\theta'\in U_l}\Lambda(\theta')$ converges to $\Lambda(\theta)$ almost surely. Monotone convergence theorem implies

$$\lim_{l\to 0}\mathbb{E}\sup_{\theta'\in U_l}\Lambda(\theta')(x) = \mathbb{E}\lim_{l\to 0}\sup_{\theta'\in U_l}\Lambda(\theta')(x) = \mathbb{E}\Lambda(\theta)(x). \tag{5.3}$$

Let the open neighborhoods around θ_0 denote Θ_0 For any $\theta\in\Theta_C\backslash\Theta_0$, there exists open neighborhood U_θ around θ. The compactness of U_θ gives a finite subcover $U_{\theta_1},\ldots,U_{\theta_s}$ with sufficient small

diameter. For sufficient large n, there is

$$\sup_{\theta\in U_{\theta_1},\ldots,U_{\theta_s}} \frac{1}{n}\sum \Lambda_n(\theta) \leq$$
$$\sup_{1\leq j\leq s}\frac{1}{n}\sum \sup_{\theta\in U_{\theta_j}} \Lambda_n(\theta) \longrightarrow \sup_{1\leq j\leq s} \mathbb{E} \sup_{\theta\in U_{\theta_j}} \Lambda_n(\theta) < 0$$

by (5.3) and LLN. If $T_n \subset U_\theta$, then $\sup_{\theta\in U_\theta}\sum\Lambda_n(\theta)/n$ is at least $\sum\Lambda_n(T_n)/n$, which by definition of T_n such that $\sum\Lambda_n(T_n)/n + o_p(1) = 0$ by LLN. Thus

$$\{T_n \subset U_\theta\} \subset \left\{\sup_{\theta\in U_\theta}\frac{1}{n}\sum\Lambda_n(T_n) \geq \mathbb{E}\Lambda(\theta_0) - o_p(1)\right\}.$$

The probability of the event on the right side converges to zero as $n\to\infty$. *Q.E.D.*

Proof of Proposition 1: The proof is based on Taylor expansions. Note that

$$m(x,\theta_0+\delta_n\tau) = m(x,\theta_0) + \delta_n\tau[\partial m(x,\theta)/\partial\theta]|_{\theta_0} + o_p(\delta_n^2). \tag{5.4}$$

Let $\theta\in\{\theta|\|\theta-\theta_0\|\leq\tau\delta_n\}$. The result

$$\lambda_n(\theta) = \left(\sum_{i=1}^n [m_i(\theta)m_i(\theta)^T]/n\right)^{-1}\sum_{i=1}^n m_i(\theta)/n + o_p(n^{-1/2})$$

holds uniformly for θ in a neighborhood of θ_0, see the proofs in Qin and Lawless (1994, Lemma 1) or Owen (2001, Theorem 2.2). For the empirical log-likelihood at θ, by noting that $\lambda_n^T m_i$ is close to zero and using a second order approximation for $\log(1+\lambda_n^T m_i)$,we obtain:

$$\sum_{i=1}^n \log\tilde{p}_\theta = \sum_{i=1}^n\left[\lambda_n(\theta)^T m_i(\theta) - \frac{1}{2}\left(\lambda_n(\theta)^T m_i(\theta)m_i(\theta)^T\lambda_n(\theta)\right)\right] - n\log n + o_p(1).$$

The remainder term is based on bounding $\sum_{i=1}^n\left(\lambda_n^T m_i\right)^3$ for which Owen (1990) showed in Lemma 3 that it is of order $o_p(1)$. Note that his γ_i is our $\lambda_n^T m_i(\theta)$. Note that

$$\lambda_n(\theta)^T m_i(\theta) = \left(\sum_{i=1}^n\frac{m_i(\theta)}{n}\right)^T\left[\sum_{i=1}^n\frac{1}{n}\left(m_i(\theta)m_i(\theta)^T\right)\right]^{-1} m_i(\theta)$$

and after summation equals the squared term:

$$\sum_{i=1}^n\lambda(\theta)^T m_i(\theta)m_i(\theta)^T\lambda_n(\theta) =$$
$$\left(\sum_{i=1}^n\frac{m_i(\theta)}{n}\right)^T\left[\sum_{i=1}^n\frac{1}{n}\left(m_i(\theta)m_i(\theta)^T\right)\right]^{-1}\left(\sum_{i=1}^n\frac{m_i(\theta)}{n}\right).$$

So adding these two terms we obtain:

$$\sum_{i=1}^n\log\tilde{p}_\theta = \frac{1}{2}\left(\sum_{i=1}^n\frac{m_i(\theta)}{n}\right)^T\left[\sum_{i=1}^n\frac{1}{n}\left(m_i(\theta)m_i(\theta)^T\right)\right]^{-1}\left(\sum_{i=1}^n\frac{m_i(\theta)}{n}\right) - n\log n + o_p(1).$$

It implies:

$$2\sum_{i=1}^{n}\log\frac{\tilde{p}_{\theta_0+\delta_n\tau}}{\tilde{p}_{\theta_0}}(x_i)=\left\{\left(\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0+\delta_n\tau)\right)^T\left(\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0+\delta_n\tau)m_i(\theta_0+\delta_n\tau)^T]\right)^{-1}\sum_{i=1}^{n}m_i(\theta_0+\delta_n\tau\right.$$
$$\left.-\left(\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)\right)^T\left(\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0)m_i(\theta_0)^T]\right)^{-1}\sum_{i=1}^{n}m_i(\theta_0)+o_p(1).\right.$$

It follows from the approximation of λ above. Using equation (5.4) we can further simplify the terms involving $\theta+\delta_n\tau$. We obtain for the middle term :

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0+\delta_n\tau)m_i(\theta_0+\delta_n\tau)^T]=&\frac{1}{n}\sum_{i=1}^{n}\left[[m_i(\theta_0)m_i(\theta_0)^T]+\delta_n\tau\frac{\partial m_i(\theta_0)}{\partial\theta}^T m_i(\theta_0)\right.\\&\left.+\frac{(\delta_n\tau)^2}{4}\frac{\partial m_i(\theta_0)}{\partial\theta}^T\frac{\partial m_i(\theta_0)}{\partial\theta}+o_p(\delta_n^3)\right]\\=&\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0)m_i(\theta_0)^T]\\&+\frac{1}{n}\delta_nO_p(n^{1/2})+o_p(\delta_n^2)+o_p(\delta_n^3).\end{aligned}$$

With the big bracket becomes

$$\begin{aligned}&n\left[\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+\frac{1}{n}\sum_{i=1}^{n}\delta_n\tau\frac{\partial m_i(\theta_0)}{\partial\theta}\right]^T\left(\frac{1}{n}\sum_{i=1}^{n}[m_i(\theta_0)m_i(\theta_0)^T]\right)^{-1}\\&\times\left[\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+\frac{1}{n}\sum_{i=1}^{n}\delta_n\tau\frac{\partial m_i(\theta_0)}{\partial\theta}\right]\\=&\,n\left[\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+\delta_n\tau\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}+\delta_nO(n^{-1/2}(\log\log n)^{1/2})\right]^T\\&\times\left(\mathbb{E}\left(m(x,\theta_0)m(x,\theta_0)^T\right)\right)^{-1}\\&\times\left[\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+\delta_n\tau\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}+\delta_nO(n^{-1/2}(\log\log n)^{1/2})\right]\\=&\,2\delta_n\tau^T\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}\left(\mathbb{E}\left(m(x,\theta_0)m(x,\theta_0)^T\right)\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+\\&\delta_n^2\tau^T\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}\left(\mathbb{E}\left(m(x,\theta_0)m(x,\theta_0)^T\right)\right)^{-1}\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}\tau+\\&\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)\left(\mathbb{E}\left(m(x,\theta_0)m(x,\theta_0)^T\right)\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}m_i(\theta_0)+o_p(\delta_n^3)\end{aligned}$$

where $O(n^{-1/2}(\log\log n)^{1/2})$ is used to bound the difference of the sample average and the expec-

tation of a random vector. Thus the local EL is

$$2\sum_{i=1}^{n}\log\frac{\tilde{p}_{\theta_0+\delta_n\tau}}{\tilde{p}_{\theta_0}}(x_i)=\tau^T\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1}\delta_n\sum_{i=1}^{n}m_i(\theta_0)$$
$$+\frac{1}{2}(\delta_n)^2\tau^T\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(x,\theta_0)}{\partial\theta}\tau+o_p(1).$$

Note that $O(n^{-1/2}(\log\log n)^{1/2})\times\delta_n\sum_{i=1}^{n}m_i(\theta_0)=o_p(1)$ and $\lim_{n\to\infty}A_n\sum_i[m_i(\theta_0+\delta_n\tau)-m_i(\theta_0)]/n=o_p(1)$ with $A_n=\sum m_i(\theta_0)\left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1}$ by the continuity of $m_i(\theta)$. *Q.E.D.*

Proof of Corollary 2: Let $\mathbb{E}Y_i^2(\theta)=\mathbb{E}[\lambda_n(\theta)m_i(\theta)]^2<\infty$, we have $\sum_{i=1}^{n}\Pr(Y_i^2(\theta)>n)<\infty$. The Borel-Cantelli lemma implies only finite number of n satisfies $\Pr(|Y_i(\theta)|>n^{1/2})>1-\varepsilon$, for small enough ε. Therefore only finite number of $\max_{1\le i\le n}|Y_i(\theta)|$ satisfies $\max_{1\le i\le n}|Y_i(\theta)|>n^{1/2}$. We can introduce a constant A such that

$$\limsup_{n\to\infty}\max_{1\le i\le n}|Y_i(\theta)|n^{-1/2}\le A$$

holds with probability 1. Since A is arbitrary, we know $\max_{1\le i\le n}|Y_i(\theta)|=o(n^{1/2})$.

Equation (5.1) implies

$$\frac{1}{n}\sum_{i=1}^{n}\frac{m_i(\theta)}{1+Y_i(\theta)}=\frac{1}{n}\sum_{i=1}^{n}\frac{m_i(\theta+\tau)}{1+Y_i(\theta+\tau)}=0.$$

With some modifications, we have

$$0=\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{[1+Y_i(\theta+\tau)]m_i(\theta)-[1+Y_i(\theta)]m_i(\theta+\tau)}{[1+Y_i(\theta+\tau)][1+Y_i(\theta)]}\right\} \tag{5.5}$$
$$\Longrightarrow\frac{1}{n}\sum_{i=1}^{n}\frac{m_i(\theta)-m_i(\theta+\tau)}{[1+Y_i(\theta+\tau)][1+Y_i(\theta)]}.$$
$$=\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(\theta)m_i(\theta+\tau)-Y_i(\theta+\tau)m_i(\theta)}{[1+Y_i(\theta+\tau)][1+Y_i(\theta)]}$$

Multiply both side of (5.5) with τ and set $\tau\to0$ does not violate the identical equation.

$$\frac{1}{n}\lim_{\tau\to0}\sum_{i=1}^{n}\frac{1}{\tau}\frac{m_i(\theta)-m_i(\theta+\tau)}{[1+Y_i(\theta+\tau)][1+Y_i(\theta)]}=$$
$$\frac{1}{n}\lim_{\tau\to0}\sum_{i=1}^{n}\frac{1}{\tau}\frac{Y_i(\theta)m_i(\theta+\tau)-Y_i(\theta+\tau)m_i(\theta)}{[1+Y_i(\theta+\tau)][1+Y_i(\theta)]} \tag{5.6}$$

$[1+Y_i(\theta+\tau)][1+Y_i(\theta)]$ is bounded by previous argument $\max_{1\le i\le n}|Y_i(\theta)|=o(n^{1/2})$ and $m_i(\theta)-m_i(\theta+\tau)$ is bounded by condition 2. Dominated convergence theorem implies that the limit operator from LHS of (5.6) can be taken inside such that

$$\frac{1}{n}\sum_{i=1}^{n}\lim_{\tau\to0}\frac{1}{\tau}\frac{m_i(\theta)-m_i(\theta+\tau)}{[1+Y_i(\theta+\tau)][1+Y_i(\theta)]}=\frac{1}{n}\sum_{i=1}^{n}\lim_{\tau\to0}\frac{1}{\tau}\frac{m_i(\theta)-m_i(\theta+\tau)}{[1+o(n^{1/2})]}$$
$$=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m_i(\theta)}{\partial\theta}\times C_i<\infty.$$

$n^{-1}\sum_{i=1}^{n} m_i(\theta)/[1+Y_i(\theta)] = 0$ for all θ, this equation is differentiable at θ, thus

$$\frac{1}{n}\sum_{i=1}^{n}\lim_{\tau\to 0}\frac{1}{\tau}\left\{\frac{[1+Y_i(\theta+\tau)]m_i(\theta)-[1+Y_i(\theta)]m_i(\theta+\tau)}{[1+Y_i(\theta+\tau)][1+Y_i(\theta)]}\right\}$$

exists. This induces that the limit of the term in LHS of (5.6) exists. Because $\lim_{\tau\to 0} Y_i(\theta)m_i(\theta+\tau)/\tau$ is simply $Y_i(\theta)\partial m_i(\theta)/\partial\theta$, so we can deduce $\partial Y_i(\theta)/\partial\theta$ exists. It is straightforward to see that $\partial \log\tilde{p}_\theta/\partial\theta$ is differentiable.

Therefore $\sqrt{\tilde{p}_\theta}$ is continuously differentiable. Since $\tilde{p}_\theta$ is continuous, Lemma 7.6 in van der Vaart (1998) indicates that equation (2.3) is valid.

We already induces equation (2.3) holds.

$$\left\|\tilde{p}_{\theta+\tau}^{1/2}-\tilde{p}_\theta^{1/2}\right\|^2 \longrightarrow \left\|\frac{1}{2}\tau^T S_{\theta,n}\tilde{p}_\theta^{1/2}\right\|^2, \tag{5.7}$$

when $\tau\to 0$. Replace the local parameter with shrinking neighborhood δ_n, a sequence of n. Then given $\tau\to 0$ and $\left\|\tau^T S_{\theta,n}\tilde{p}_\theta^{1/2}\right\|$ is a real-value, we have

$$\int\left[\delta_n^{-1}\left(\tilde{p}_{\theta+\delta_n\tau}^{1/2}-\tilde{p}_\theta^{1/2}\right)\right]^2 d\mu = O(1).$$

Since δ_n^{-1} is an increasing sequence, $\tilde{p}_{\theta+\delta_n\tau}^{1/2}-\tilde{p}_\theta^{1/2}$ will converge to zero in $L^2(\mu)$. In other words, $\tilde{p}_{\theta+\delta_n\tau}^{1/2}$ converges to $\tilde{p}_\theta^{1/2}$ in L^2 norm. Theorem 7.2 in van der Vaart (1998) shows that for a bounded sequence τ_n,

$$\log\prod_{i=1}^{n}\frac{\tilde{p}_{\theta+\delta_n\tau_n}}{\tilde{p}_\theta}(x_i) = \tau_n^T S_{\theta,n} - \frac{1}{2}\tau_n^T K_\theta\tau_n + o_{\tilde{p}_\theta}(1). \tag{5.8}$$

Namely, the differences between $\Lambda_n(\theta+\delta_n\tau_n,\theta)$ and a linear-quadratic equation tend to zero in $\tilde{p}_\theta$ probability. *Q.E.D.*

Proof of Proposition 3: To prove K_θ is invertible, we need to prove K_θ is almost surely positive definite. If K_θ is postive semidefinite but not invertible, then there must be a non-zero vector τ such that $K_\theta\tau = 0$. Therefore $\tau^T K_\theta\tau = 0$ which contradicts our assumption about K_θ being positive definite.

$K_{\theta,n}$ and $S_{\theta,n}$ form a relatively compact sequence. Thus a subsequence $(K_{\theta,n_k}, S_{\theta,n_k}) \to (K_\theta, S_\theta)$. Le Cam's first lemma implies

$$\mathbb{E}\exp\left[\tau^T S_\theta - \frac{1}{2}\tau^T K_\theta\tau\right] = 1. \tag{5.9}$$

Because (5.9) holds for all τ, we can use a symmetrized method to simplify (5.9). For certain value τ and $-\tau$, we have

$$\mathbb{E}\left\{\exp\left[\tau^T S_\theta - \frac{1}{2}\tau^T K_\theta\tau\right] + \exp\left[-\tau^T S_\theta - \frac{1}{2}\tau^T K_\theta\tau\right]\right\} = 2.$$

By $\cosh \tau^T S_\theta = (\exp \tau^T S_\theta + \exp(-\tau^T S_\theta))/2$, we have $\mathbb{E}[(\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau/2)] = 1$. Assume some τ_i give $\tau^T K_\theta \tau$ negative values, then

$$\begin{aligned}(5.10)\quad &\mathbb{E}\left[\mathbb{I}_{\{\tau^T K_\theta \tau > 0\}} (\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau/2)\right] \\ &\leq \mathbb{E}\left[(\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau/2)\right] = 1\end{aligned}$$

where $\mathbb{I}_{\{\cdot\}}$ is an indicator function. However, since $\exp(-\tau^T K_\theta \tau/2) > 1$ when $\tau^T K_\theta \tau$ is negative and $(\cosh \tau^T S_\theta) > 0$, we know above equation can not be valid unless the set $\{\tau^T K_\theta \tau > 0\}$ is null. Therefore, for all τ, K_θ is positive definite. *Q.E.D.*

Proof of Theorem 2.2: (i) When θ is given, by condition 4

$$\begin{aligned}(5.11)\quad \Lambda_n(\theta + \delta_n \tau_n, \theta) =& \tau_n^T S_{\theta,n} - \frac{1}{2} \tau_n^T K_{\theta,n} \tau_n + o_{\tilde{p}_\theta}(1) \\ =& -\frac{1}{2}\Big[(K_{\theta,n}^{-1} S_{\theta,n} - \tau_n^T)^T K_{\theta,n} (K_{\theta,n}^{-1} S_{\theta,n} - \tau_n^T)) \\ & -(S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n})\Big] + o_{\tilde{p}_\theta}(1).\end{aligned}$$

Similarly,

$$\begin{aligned}(5.12)\quad \Lambda_n(\theta + \delta_n \tau_n, \theta) =& -\frac{1}{2}\left[(\delta_n(T_n - \theta) - \tau_n^T)^T K_n (\delta_n(T_n - \theta) - \tau_n^T)) \right. \\ & \left. -(\delta_n(T_n - \theta))^T K_n (\delta_n(T_n - \theta))\right].\end{aligned}$$

The difference between (5.11) and (5.12) tends to zero in probability $\tilde{G}_{\theta,n}$. The four quadratic terms and non-negativity of K_n and $K_{\theta,n}$ indicate that each quadratic term must be non-negative. The difference between $\tau_n^T K_{\theta,n} \tau_n - 2\tau_n^T S_{\theta,n}$ and $\tau_n^T K_n \tau_n - 2\tau_n^T \delta_n(T_n - \theta) K_n$ must converge to zero in probability, otherwise the arbitrary choosing of τ_n will make the equality invalid.

$$\tau_n^T K_n \tau_n - 2\tau_n^T \delta_n(T_n - \theta) K_n \sim \tau_n^T K_{\theta,n} \tau_n - 2\tau_n^T S_{\theta,n}$$

implies K_n converges to $K_{\theta,n}$ in probability. Therefore, $\delta_n(T_n - \theta)$ must converge to $K_{\theta,n}^{-1} S_{\theta,n}$.

(ii) By proposition 3, we know that clustering points K_θ of $K_{\theta,n}$ is invertible. Since $\delta_n(T_n - \theta)$ converges to $K_{\theta,n}^{-1} S_{\theta,n}$, the limit of $\delta_n(T_n - \theta)$ is $K_\theta^{-1} S_{\theta,n}$ which implies $\delta_n(T_n - \theta)$ is bounded in probability $\tilde{G}_{\theta,n}$.

(iii) By corollary 2, we know the condition DQM can imply 4, thus intuitively the linear-quadratic equation (**??**) may coincide with S_n and K_n. The log-likelihood process can be rewritten as a centered log-likelihood process $X_n(\cdot)$ plus a shift item $b_n(\cdot)$:

$$\begin{aligned}\Lambda_n(\eta, \theta)(x) =& \overbrace{\frac{1}{n}\sum_{i=1}^n \log \frac{\tilde{g}_\eta}{\tilde{g}_\theta}(x_i) - \int \log \frac{\tilde{g}_\eta}{\tilde{g}_\theta}(x) dG_\theta}^{X_n(\eta)} \\ & + \underbrace{\int \log \frac{\tilde{g}_\eta}{\tilde{g}_\theta}(x) dG_\theta}_{b_n(\eta)} + o_p(1).\end{aligned}$$

Let $\eta = \theta + \delta_n \tau$ and $\delta_n = n^{-1/2}$. Given λ value in the constraint of equation (**??**), $\log(\tilde{g}_\eta / \tilde{g}_\theta)(x_i)$ in $X_n(\eta)$ can be replaced by a linear quadratic formula w.r.t. τ, namely $\log(\tilde{g}_\eta / \tilde{g}_\theta)$ belongs to a smooth enough function class. Therefore the process $\eta \mapsto X_n(\eta)$ is an empirical process and $\sqrt{n} X_n(\eta) \rightsquigarrow X(\eta)$ by Donsker theorem (see van der Vaart (Example 19.9 1998)) where $X(\eta)$ is a Gaussian process. Note that $X(\eta)$ has mean $\int X(\eta) dG_\theta = 0$ and covariance kernel $K_\theta = \mathbb{E}_\theta X^2(\eta)$ under distribution G_θ. Le Cam's first lemma (contiguity) implies $\mathbb{E}_\theta \exp[X(\eta) + b(\eta)] = 1$ with the expectation taken under G_θ and $b_n(\eta) \to b(\eta)$. Gaussian properties give $b(\eta) = -(1/2)\mathbb{E}_\theta X^2(\eta)$. Therefore the log-likelihood process By proposition 1 and equation (**??**), we can show that

$$\begin{aligned} X(\eta) =& \tau^T S_\theta \\ b(\eta) =& -\frac{1}{2} \tau^T K_{\theta_0} \tau, \end{aligned}$$

and when $\theta = \theta_0$

$$\begin{aligned} X(\eta) =& \tau \mathbb{E} \frac{\partial m(x, \theta_0)}{\partial \theta}^T \left(\mathbb{E} m(x, \theta_0) m(x, \theta_0)^T \right)^{-1} \delta_n m_i(\theta_0) \\ b(\eta) =& -\frac{1}{2} \tau \mathbb{E} \frac{\partial m(x, \theta_0)}{\partial \theta}^T \left(\mathbb{E} m(x, \theta_0) m(x, \theta_0)^T \right)^{-1} \mathbb{E} \frac{\partial m(x, \theta_0)}{\partial \theta} \tau. \end{aligned}$$

It means that K_θ varying slowly enough as θ varies, one can locally approximate $(T_n - \theta)^T K_\theta (T_n - \theta)$ by Gaussian shift experiments. *Q.E.D.*

Proof of Theorem 2.3: The proof follows the strategies of van der Vaart (Proposition 8.6 1998) and Le Cam and Yang (Theorem 6.1 1990). The difficulty comes from the conditional expectation over local parameter τ. In Local EL the approximation is evaluated for each discretized θ_n^*. This implicitly imposes a priori assumption that true parameter concentrates on set $\{\eta : |\eta - \theta| \leq \delta_n \tau\}$ in every grid. We use a Bayesian argument for "local prior measures".

$b \wedge W(Z_n - \tau)$ gives a bounded function. We can consider the expectation as $b \wedge \mathbb{E}[W(Z_n - \tau) | \theta + \delta_n \tau]$. By the LAN property of T_n from Theorem 2.2(iii), it is straightforward to set the prior measures with Gaussian densities. Since both "prior" and "posterior" concentrate around θ_0, the updating information only occurs for covariance matrix. Let τ be the gaussian centered at 0 with inverse covariance Γ. The conjugate property indicates the posterior of τ can be written in terms of LEL representation:

$$Z_n = \delta_n^{-1}(T_n' - \theta) = (K_n + \Gamma)^{-1} K_n \delta_n^{-1}(T_n - \theta),$$

especially when $\Gamma = 0$, $Z_n = \delta_n^{-1}(T_n - \theta)$. Note that by Anderson's lemma $\mathbb{E}[W(Z_n - \tau) | \theta + \delta_n \tau] \geq \mathbb{E}[W Z_n | \theta + \delta_n \tau]$ for bounded W. Since $K_n \delta_n^{-1}(T_n - \theta) \sim \mathcal{N}(0, I)$, the lower bound of $\mathbb{E}[W(Z_n - \tau) | \theta + \delta_n \tau]$ is

$$\mathbb{E}\left\{ W\left[(K_n + \Gamma)^{-1/2} \times \mathcal{N}(0, I) \right] | K_n \right\}.$$

K_n is independent with $\mathcal{N}(0, I)$. With the condition $K_n \rightsquigarrow K_\theta$ in $\tilde{P}_\theta$ law, the limit becomes $\mathbb{E}\left\{ W\left[(K_\theta + \Gamma)^{-1/2} \times \mathcal{N}(0, I) \right] \right\}$.

When c is very large, the probability of normal prior $|\tau| > c$ is small enough thus

$$\liminf_n \sup_{|\tau| \leq c} \mathbb{E}\left\{ W\left[(K_n + \Gamma)^{-1/2} \times \mathcal{N}(0, I) \right] \right\} \geq \mathbb{E}\left\{ W\left[(K_\theta + \Gamma)^{-1/2} \times \mathcal{N}(0, I) \right] \right\} - \Delta$$

for small enough Δ. Let Γ go to zero, $Z_n = \delta_n^{-1}(T_n - \theta)$ obtains the lower bound $\mathbb{E}[W(K_\theta^{-1/2}) \times \mathcal{N}(0, I)]$. If $W = 1$ and $K_\theta = K_{\theta_0}$, by Theorem 2.2(iii) we achieve the efficient bound of semi-parametric estimators. *Q.E.D.*

REFERENCES

Baggerly, K. A. (1998): "Empirical Likelihood as a Goodness-of-Fit Measure," *Biometrika*, 85(3), 535–547.

Brown, B. W., and W. K. Newey (2002): "Generalized Method of Moments, Efficient Bootstrapping, and Improved Inference," *Journal of Business & Economic Statistics*, 20(4), 507–17.

Csiszar, I. (1984): "Sanov Property, Generalized I-Projection and a Conditional Limit Theorem," *The Annals of Probability*, 12(3), 768–793.

Donald, S., G. W. Imbens, and W. Newey (2003): "Empirical likelihood estimation and consistent tests with conditional moment restrictions," *Econometrica*, 117(1), 55–93.

Guggenberger, P., and R. J. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests under Partial, Weak, and Strong Identification," *Econometric Theory*, 21, 667–709.

——— (2008): "Generalized empirical likelihood tests in time series models with potential identification failure," *Journal of Econometrics*, 142(1), 134–161.

Hansen, L. P., J. Heaton, and A. Yaron (1996): "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 14, 262–280.

Huber, P. (1981): *Robust Statistics*. Willy, New York.

Imbens, G. W., R. H. Spady, and P. Johnson (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66(2), 333–357.

Kitamura, Y., T. Otsu, and K. Evdokimov (2009): "Robustness, Infinitesimal, Neighborhoods, and Moment Restrictions," *Working Paper*.

Kitamura, Y., and M. Stutzer (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65(4), 861–874.

Kitamura, Y., G. Tripathi, and H. Ahn (2004): "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models," *Econometrica*, 72(6), 1667–1714.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70(5), 1781–1803.

——— (2005): "Testing Parameters in GMM without Assuming That They Are Identified," *Econometrica*, 73(4), 1103–1123.

Le Cam, L. (1974): *Notes on Asymptotic Methods in Statistical Decision Theory*. Universit de Montral, Centre de Recherches Mathmatiques.

Le Cam, L., and G. Yang (1990): *Asymptotics in Statistics: Some Basic Concepts (Springer Series in Statistics)*. Springer-Verlag, New York.

——— (2000): *Asymptotics in Statistics: Some Basic Concepts Second Edition(Springer Series in Statistics)*. Springer-Verlag, New York.

Newey, W., and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72(1), 219–255.

Oodaira, H. (1973): "The Law of the Iterated Logarithm for Gaussian Processes," *Annals of Probability*, 1(6), 954–967.

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

——— (1990): "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18(1), 90–120.

——— (2001): *Empirical Likelihood*. Chapman & Hall/CRC, Florida.

Parente, P., and R. J. Smith (????): "GEL Methods for Non-Smooth Moment Indicators," *forthcoming in Econometric Theory*.

Qin, J., and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22(1), 300–325.

Ronchetti, E., and F. Trojani (2001): "Robust inference with GMM estimators," *Journal of Econometrics*, 101(1), 37–69.

Schennach, S. M. (2007): "Point estimation with exponentially tilted empirical likelihood," *The Annals of Statistics*, 35(2), 634–672.

Smith, R. (2005): "Local GEL methods for conditional moment restrictions," Discussion Paper CWP15/05.

Smith, R. J. (1997): "Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation," *The Economic Journal*, 107(441), 503–519.

Stock, J. H., and J. H. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68(5), 1055–1096.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20(4), 518–529.

van der Vaart, A. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag, New York.

Wald, A. (1949): "Note on the consistency of the maximum likelihood estimate," *Annals of Mathematical Statistics*, 20(4), 595–601.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50(1), 1–25.