Essays on empirical likelihood in economics

Gao, Z.

Publication date
2012

2

AN EMPIRICAL LIKELIHOOD BASED LOCAL ESTIMATION

2.1 INTRODUCTION

To estimate an economic model, the model is often represented in terms of a family of probability measures $\mathcal{E}_\theta = \{P_\theta; \theta \in \Theta\}$ depending on a parameter θ in $\Theta \in \mathbb{R}^d$. By adjusting the value of θ one can choose which P_θ best fits the data. There is a literature within econometrics that considers how to attain a suitable measure P_θ by comparing a specified *moment condition function*

$$\int m(x,\theta) dP_\theta(x) = \mathbb{E}_\theta[m(X,\theta)],$$

a $k \times 1$ vector with $k \geq d$, with its *sample counterpart*

$$\int m(x,\theta) dP_n(x) = \frac{1}{n} \sum_{i=1}^{n} m(X_i,\theta),$$

where P_n is the empirical distribution. Note that although P_θ is indexed by θ, the distribution of $m(X,\theta)$ does not necessarily fully depend on θ. The notation P_θ should be interpreted as a pseudo measure of $m(X,\theta)$ and the specification of this measure depends on the value of θ. In this chapter, we assume the random variable X_i to be i.i.d. A particular correspondence between $\mathcal{E}_\theta$ and $m(X,\theta)$ is established by Empirical Likelihood (EL) (Qin and Lawless, 1994; Kitamura and Stutzer, 1997). EL has been embedded into some more general problems, e.g. Csiszar (1984); Smith (1997); Baggerly (1998); Newey and Smith (2004). The aims of these methods are similar: to optimize a criterion function of θ, for example a likelihood ratio, subject to constraints based on $m(X,\theta)$.

The choice of criterion functions matters for the efficiency and robustness of an estimator. To balance the tradeoff between these two objectives, Schennach (2007) suggests a two-step inference method by switching the empirical discrepancy between

Kullback-Leibler and likelihood ratio. Kitamura et al. (2009) suggest using Hellinger's distance as the criterion. In this chapter, we will focus on a representation of the classical likelihood ratio. Classical likelihood ratio does have problems of maintaining robustness, but several ways of re-constructing the likelihood have been proposed that remedy this problem. This chapter will consider a "localization" technique of representing the likelihood ratio function of $m(X,\theta)$ when it has poor behavior over some critical points. The method was first suggested in statistics for parametric Maximum Likelihood Estimation by Le Cam even earlier than his 1974 published notes (Le Cam, 1974).

The "local" here is the analog of "differential". If one fixes a particular θ_0 in Θ and investigates what happens to the likelihood ratio function with parameter sequences of the form $\theta = \theta_0 + \delta_n\tau$, with $\delta_n \to 0$ as n goes to infinity, then δ_n yields a sort of differentiation rate just as the differentiation rate in basic calculus, and then the whole localization problem can be analyzed as a kind of differentiability problem. The term τ is called local parameter since it is an index for local features. This technique often appears in the evaluation of local power of test statistics and statistical experiments, see van der Vaart (1998) and Le Cam and Yang (2000).

The advantage of studying the EL problem under the localized representation is significant. For instance, the likelihood of EL includes a vector of implied probabilities $(p(X_1,\theta),\ldots,p(X_n,\theta))$ where $\theta \in \Theta \subset \mathbb{R}^d$. Localization considers the probability vector $(p(X_1,\theta),\ldots,p(X_n,\theta))$ on a neighborhood of some θ^* and returns numbers instead of functions. In addition, a well-behaved local representation ensures the existence of the derivative of this representation. By definition, when the derivative exists, small changes will not blow up the approximation of the original likelihood ratio function and this representation is therefore robust to these changes. Apart from theoretical advantages, the method is also computationally attractive. Unlike global approaches where complexity grows exponentially with the dimension of the parameter, the local method can handle the growing number of $(p(X_1,\theta),\ldots,p(X_n,\theta))$ by cutting a large problem into many small, but easily computable local problems. The local approach can avoid some peculiar points that break down the computational routines.

2.2 EMPIRICAL LIKELIHOOD

EL considers a finite dimensional parameter θ and an increasing number of $(p(X_1,\theta),\ldots,p(X_n,\theta))$. EL simultaneously finds the optimal θ and the optimal $(p(X_1,\theta),\ldots,p(X_n,\theta))$ that satisfy the required moment constraints $\sum_{i=1}^{n} m(X_i,\theta)p(X_i,\theta) = 0$. Its criterion is:

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

where p_i is shorthand for $p(X_i,\theta)$ given the value θ. An explicit expression for the optimal p_i's can be derived using the Lagrangian method and gives the solution:

$$\tilde{p}_i(\theta) := \frac{1}{n}\frac{1}{1+\lambda_n^T m_i(\theta)},$$

where $\tilde{p}_i(\theta)$ is called the *implied probability*. The candidate solutions belong to the family

$$\mathcal{E}_\theta := \{\tilde{P}_\theta : \theta \in \Theta, \int m(X,\theta)d\tilde{P}_\theta = 0\},$$

where $d\tilde{P}_\theta(x_i) = \tilde{p}_i(\theta)d\mu$ for a counting measure μ[1]. The λ_n is the solution of:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{m_i(\theta)}{1+\lambda_n^T m_i(\theta)}\right] = 0. \tag{2.1}$$

Let the log-likelihood ratio of the implied probability between any two parameter values θ_1 and θ_2 be:

$$\Lambda_n(\theta_1,\theta_2) := \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{\tilde{p}_i(\theta_1)}{\tilde{p}_i(\theta_2)}\right]$$

and define the average log-likelihood ratio of the implied probability given θ and counting numbers $1/n$ as

$$\Lambda_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\log n\tilde{p}_i(\theta).$$

The constraint $0 \leq \tilde{p}_i \leq 1$ requires that the inequality $1+\lambda_n^T m_i(\theta) \geq 1/n$ always holds. The *population* $\lambda(\theta) := \lim_{n\to\infty}\lambda_n$ must lie in

1 The family $\mathcal{E}_\theta$ obtains both continuous measures and discrete measures. The definition will become clear once we introduce the infinite divisiblility concept.

a convex and closed set $\Gamma_\theta = \lim_{n\to\infty} \cup_{i=1}^{n} \Gamma_{\theta,i}$. For fixed n, the set (σ-algebra) $\Gamma_{\theta,n}$ is defined as a collection of subsets of

$$\{\lambda_n : 1 + \lambda_n^T m(X_i,\theta) \geq 1/n,\ i = 1,\ldots,n,\ \theta \in \Theta\}.$$

In the following, we will consider the case where the derivatives of $m(X,\theta)$ do not exist for some X. A weaker consistency result for the EL estimator is required. Here are the conditions:

Condition 2.1. (i) $M(\theta) := \mathbb{E}[m(X,\theta)]$ 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

$$\inf_{\theta: d(\theta,\theta_0)\geq\epsilon} |M(\theta)| > |M(\theta_0)| = 0,$$

where ϵ is an arbitrary value larger than zero and $d(\cdot,\cdot)$ is any distance function on Θ.

(iii) $m(X,\theta)$ is continuous in θ,

$$\lim_{\theta'\to\theta} \left\| m(X,\theta) - m(X,\theta') \right\| = 0.$$

(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) $\liminf_{\theta\to\infty} \|M(\theta)\| / b(\theta)$ is larger than 1, and
(3) $\limsup_{\theta\to\infty} \|m(X,\theta) - M(\theta)\| / b(\theta) < 1$.

(v) $\sum_{i=1}^{n} \left[m(x_i,\theta_0) m(x_i,\theta_0)^T\right] / n$ exists and has full rank.

Condition 2.1 (i) ensures the model is identified for a small neighborhood of θ_0. (ii) is a local separability condition. (iii) is used to obtain the continuity of the Lagrangian multiplier. (iv) is an envelope assumption; it is used to obtain some dominated convergence results. The one-point (Alexandroff) compactification allows us to let θ approach any boundary place 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 the continuous derivative of $m(X,\theta)$ and that the derivative is of full rank. Condition 2.1 is less restrictive because it allows for irregular cases where the usual "delta method" does not work, e.g. when $m(x,\theta)$ is non-differentiable. Condition 2.1 (i)-(iv) are the standard M-estimator conditions in Huber (1981) and are very weak in the context of parametric models.

Theorem 2.2. *If Condition 2.1 holds, then every sequence T_n satisfying*

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

will converge to θ_0 almost surely.

Remark 2.3. Kitamura and Stutzer (1997) relax the assumptions in Qin and Lawless (1994) and Kitamura et al. (2004) 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 the moment restriction, however, is assumed in order to obtain a valid approximation for the Lagrangian $\lambda(\theta)$. In this chapter, the assumptions are similar to the standard M-estimator conditions in Huber (1981), thus the differentiability assumption is not required.

2.3 GAUSSIAN PROPERTIES, METRIZATION AND LOCALIZATION OF EL

To get a standard result for the estimator, the criterion function has to satisfy some regularity conditions. The attempt in this chapter is to consider the situations where regularity conditions may be violated, so a weaker counterpart of the conditions is called for. Here the problem appears. The Lagrangian multiplier λ_n gives a dual of the constraint function. But the functional form of λ_n has no closed-form representation, since it is the solution of Equation (2.1) depending on sample size and parameter values. Because of this feature, the general techniques such as empirical processes of studying irregular behavior of the functions are not directly applicable. While the usual asymptotic theorems that rely on differentiability and smoothness of the criterion functions in moment-based estimations are too restrictive in the situations on which we focus, we need alternative conditions and specifications.

We come back to the probability family $\mathcal{E}_\theta$. If $\mathcal{E}_\theta$ belongs to an ideal space, e.g. a complete separable metric (Polish) space, the ordinary topology for assessing weak convergence of $\tilde{P} \in \mathcal{E}_\theta$

is *relative compactness*[2]. The relative compactness will induce a well-defined likelihood ratio function.

Condition 2.4. For any sequence

$$\Lambda_n((X_1,\ldots,X_n),\theta) = n^{-1}\sum_{i=1}^{n}\log n\tilde{p}(\theta,X_i)$$

evaluated at any fixed θ, there are constants a_n such that

$$\Lambda_n((X_1,\ldots,X_n),\theta) - a_n$$

forms a relatively compact sequence.

Lemma 2.5. *Given Condition 2.4, for every fixed n and θ, the random variable* $\log n\tilde{p}(\theta,X_i)$ *has bounded variance.*

Note that Condition 2.4 still allows a single (or a certain proportional) value of $\log n\tilde{p}_i(\theta)$ to go to infinity at some x_i if only the speed is slower than exponential rate of n. Condition 2.4 and Lemma 2.5 imply that $\Lambda_n(\theta)$ can be thought of as a well-defined random variable or a realization of a likelihood ratio process $\Lambda((X_1,\ldots,X_n),\theta)$ with indices n and θ. In the following subsection, we will see how to connect this process with a local representation.

2.3.1 *Approximation for an Infinitely Divisible Family*

A non-closed form λ_n induces a non-closed form vector

$$(\tilde{p}_1(\theta),\ldots,\tilde{p}_n(\theta)).$$

It is better to transfer the attention of the implied probability vectors to a family of probability measures

$$\mathcal{E}_\theta := \{\tilde{P}_\theta : \theta\in\Theta, \int m(X,\theta)d\tilde{P}_\theta = 0\},$$

where the discrete vector $(\tilde{p}_1(\theta),\ldots,\tilde{p}_n(\theta))$ satisfies

$$\sum_{i}^{n} m(x_i,\theta)\tilde{p}_i(\theta) = 0.$$

2 A sequence of statistics S_n associated with $\tilde{P}$ is *relatively compact* if for every $\epsilon>0$ there is a number $b(\epsilon)$ and an integer $N(\epsilon)$ such that $n\geq N(\epsilon)$ implies $\tilde{P}\{|S_n|>b(\epsilon)\}<\epsilon$.

If a random variable ξ, for every natural number n, can be represented as the sum

$$\xi = \xi_{1,n} + \xi_{2,n} + \cdots + \xi_{n,n}$$

of n i.i.d random variables $\xi_{1,n},\ldots,\xi_{n,n}$, then ξ is called *infinitely divisible* (Gnedenko and Kolmogorov, 1968, p. 78). A probability distribution is said to be infinitely divisible if and only if it can be represented as the distribution of the sum of an arbitrary number of i.i.d random variables. A family of such distributions is often referred to as an *infinitely divisible family*. In our case, for arbitrary sample size n and fixed θ, the log-likelihood ratio process is

$$\Lambda((X_1,\ldots,X_n),\theta) = \log n\tilde{p}(X_1,\theta) + \cdots + \log n\tilde{p}(X_n,\theta).$$

Every additional term $\log n\tilde{p}(X_i,\theta)$ is an i.i.d increment of this log-likelihood ratio process.

Condition 2.6. Let $\varpi_{n,\theta}(\cdot) = \sum_{i=1}^{n}\left[\frac{m(X_i,\theta)}{1+(\cdot)m(X_i,\theta)}\right]$. Given θ,

$$\lim_{n\to\infty}\varpi_{n,\theta}(\cdot)$$

is independent of $m(X_i,\theta)$ for $0 < i \leq n$.

In a localization approach, when θ is given, λ_n as a solution of the nonlinear equation (2.1) depends on the aggregated element $\varpi_{n,\theta}(\cdot)$. For sufficient large n, an aggregated element may be independent of its individual element. For example, the average of a summation of i.i.d variables will converge to a Gaussian random variable, but the i.i.d variable itself does not necessary to be Gaussian. When this is the case for λ_n, λ_n, which is independent of $m(X_i,\theta)$, is simply a stochastic factor in all $\log n\tilde{p}(X_i,\theta)$, $0 < i \leq n$.

For given a sufficient large n and θ, then

$$\log n\tilde{p}(X,\theta) := -\log(1+\lambda_n^T m(X,\theta))$$

is random and $(\log n\tilde{p}(X_1,\theta),\ldots,\log n\tilde{p}(X_n,\theta))$ is a random vector where λ_n is a stochastic factor for n-sample size problems. For sufficient large n, Condition 2.6 implies that $\log n\tilde{p}(X_i,\theta)$ are identically distributed and independent. Then one can think that the integral of the log-likelihood ratio process $\int \log\frac{dP_\theta}{dP_0}(x)dP_0(x)$ for given sample size n, $\sum_i \log n\tilde{p}(X_i,\theta)$, as representing an infinite divisible process ξ in n additive terms $\xi_{1,n} + \xi_{2,n} + \cdots +$

$\xi_{n,n}$[3]. Thus $\mathcal{E}_\theta$ does not merely include the family of distributions that satisfy the constraint $\int m(x,\theta)d\tilde{P}_\theta(x)$, it also requires the sample average of the log-likelihood ratio process of $\tilde{P}_\theta$ to be infinitely divisible. It is clear now that EL *inherits* the moment constraint from moment-based methods and *inherits* the infinitely divisibility from likelihood ratio based methods.

An infinitely divisible family $\mathcal{E}$ admits a representation $\mathcal{E} = \mathcal{E}_1 \times \cdots \times \mathcal{E}_n = \otimes_{i=1,\ldots,n}\mathcal{E}_i$ based on n copies of the so called divisor $\mathcal{E}_i$, where n could be arbitrarily large and $\times$ denotes the direct product. The family $\mathcal{E}$ is called *divisible* with divisor $\mathcal{E}_i$. There are several well known infinitely divisible families, e.g. Poisson and Gaussian families.

It has been proved by Gnedenko and Kolmogorov (1968, Theorem 17.5) that any infinitely divisible family can be approximated by a finite number of Poisson type measures. This is an extremely useful result. It basically means that the infinitely divisible family constructed by $\{\log n\tilde{p}(X,\theta)\}$ can be approximated by a finite number of Poisson measures[4]. We know that the Poisson family can be related to the Gaussian family, for example via Hellinger's affinity. We will use this property to deduce a representation of the likelihood ratio process.

Theorem 2.7. *If $\tilde{P}_\theta$ is infinitely divisible then when $n \to \infty$, the log-likelihood $\log d\tilde{P}_{\theta+\delta_n\tau_n}/d\tilde{P}_\theta$ can be approximated by a linear quadratic expression such that the difference*

$$\sum_{i=1}^{n} \log \frac{d\tilde{P}_{\theta+\delta_n\tau_n}}{d\tilde{P}_\theta}(x_i) - \left[\tau_n^T S_{\theta,n} - \frac{1}{2}\tau_n^T K_{\theta,n}\tau_n\right] \tag{2.2}$$

tends to zero in probability for any bounded sequence $\{\tau_n\}$ with a random vector $S_{\theta,n}$ and a deterministic matrix $K_{\theta,n}$.

The infinite divisible feature gives us a useful representation for the likelihood ratio process, a linear quadratic expression with a local parameter τ_n. With this expression, we can construct our estimator without bothering with non-linear optimization, since the parameter in (2.2) is re-parametrized by τ_n which appears linearly and quadratically in the equation. Furthermore, neither the computational algorithm nor the weakly convergent statistics involve any differentiation requirements.

3 More details about such a construction are discussed in Le Cam and Yang (2000, Chapter 5), although in most cases, they use $\log(1 + (p_\theta/p_\vartheta)^{-1/2} - 1)$ instead of $\log(p_\theta/p_\vartheta)$ directly.

4 We give a short description about Poissonization in the appendix.

Remark 2.8. In the proof, we will show a relation for univariate Gaussian families. For any pair of Gaussian measures G_θ and G_ϑ, there will be a linear-quadratic expression to relate them. Therefore, the integral of $(dG_\theta/dG_\vartheta)^{1/2}$ w.r.t. G_ϑ will have a linear quadratic representation. Then we show that if $\tilde{P}_\theta$ is infinitely divisible, $(d\tilde{P}_\theta/d\tilde{P}_\vartheta)^{1/2}$ will be approximately equal to $(dG_\theta/dG_\vartheta)^{1/2}$, so $(d\tilde{P}_\theta/d\tilde{P}_\vartheta)^{1/2}$ will also have a linear quadratic representation.

Remark 2.9. The linear-quadratic approximations to the log-likelihood ratios can possibly be used 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.

Remark 2.10. From a computational aspect, when confronted with the nonlinear optimization, the Hessian matrix of the problem in some cases is difficult to evaluate especially in regions that are either extremely flat or very erratic. It is then computationally more efficient to consider the local optimization and avoid a singular or non-invertible Hessian matrix rather than calculate the global second order derivative of the objective function.

Remark 2.11. Theorem 2.7 shows that with a proper choice of δ_n, the log-likelihood ratio can be approximated by a linear-quadratic representation. One of the main focus of this representation is the quadratic term. As we known, for a pair of Gaussian measures (G_θ, G_ϑ) with dominating measure μ we will have

$$\begin{aligned}
&\int \left(\frac{dG_\theta}{dG_\vartheta}\right)^{\frac{1}{2}} dG_\vartheta = \int dG_\theta^{\frac{1}{2}} dG_\vartheta^{\frac{1}{2}} d\mu \\
=&\mathbb{E}\exp\left\{\sum_{i=\theta,\vartheta}\frac{1}{2}\left[L(i) + \mathbb{E}\log(\frac{dG_i}{d\mu})\right]\right\} \\
=&\left[\exp -\frac{1}{4}\left(K(\theta,\theta) + K(\vartheta,\vartheta)\right)\right] \cdot \mathbb{E}\exp\left(\sum_{i=\theta,\vartheta}\frac{1}{2}L(i)\right) \qquad (2.3) \\
=&\exp\left\{\frac{1}{4}\left[2K(\theta,\vartheta) - K(\theta,\theta) - K(\vartheta,\vartheta)\right]\right\}, \qquad (2.4)
\end{aligned}$$

[5]where $L(i) := \{\log(dG_i/d\mu) - \mathbb{E}\log(dG_i/d\mu)\}$. The property of $L(i)$ includes that it is Gaussian with expectation $\mathbb{E}L(i) = 0$ and covariance kernel $K(\theta,\vartheta) = \mathbb{E}L(\theta)L(\vartheta)$ and we have

$$\mathbb{E}L(i)^2 = K(i,i).$$

Let $q(\theta,\vartheta) = -8\log\int dG_\theta^{\frac{1}{2}} dG_\vartheta^{\frac{1}{2}} d\mu$. Since the quadratic term is deterministic in the neighborhood of θ_0, we can use interpolation to find $K(\cdot,\cdot)$. With an arbitrary mid-point u, three-point interpolation gives us:

$$K(\theta,\vartheta) = -\left(q(\theta,\vartheta) - q(\theta,u) - q(u,\vartheta)\right).$$

For small $|\theta - \vartheta|$, to speed up the computation, one could use an approximated value $\Lambda_n(\theta,\vartheta)$ instead of $q(\theta,\vartheta)$[6].

2.3.2 *Comparison with Other Conditions*

The standard EL ratio can be put into the form of the linear quadratic representation in (2.2) but this requires some additional assumptions, e.g. differentiability of $m(X,\theta)$. The following proposition establishes this relation.

Proposition 2.12. *Suppose that in addition to Condition 2.1, the following holds*

(i) the model is just-identified, $\partial m(X,\theta)/\partial\theta < \infty$ for any X, the rank of $\mathbb{E}[\partial m(X,\theta)/\partial\theta]|_{\theta_0}$ equals $\dim(\theta)$,

(ii) $\frac{1}{n}\sum_{i=1}^n [m_i(\theta)m_i(\theta)^T]$ and $\frac{1}{n}\sum_{i=1}^n [\lambda_n^T m_i(\theta)]^2$ are both finite for any positive n, even as $n \to \infty$,

5 The derivation of (2.3) and (2.4) is as follows. Since $\mathbb{E}[\exp(\log(dG_i/d\mu))] = 1$, then we have

$$\mathbb{E}\exp\left[L(i) + \mathbb{E}\log(\frac{dG_i}{d\mu})\right] = \left[\mathbb{E}e^{L(i)}\right]\cdot e^{\mathbb{E}\log(\frac{dG_i}{d\mu})} = 1$$

By the log-normal property, $\mathbb{E}\exp L(i) = e^{\frac{1}{2}K(i,i)}$, we have

$$e^{\frac{1}{2}K(i,i)}\cdot e^{\mathbb{E}\log(\frac{dG_i}{d\mu})} = 1 \Longleftrightarrow \mathbb{E}\left[\log(\frac{dG_i}{d\mu})\right] = -\frac{1}{2}K(i,i)$$

thus we have (2.3). For $\mathbb{E}\exp[L(\theta) + L(\vartheta)]$, we have $2K(\theta,\vartheta)$. Combining $2K(\theta,\vartheta)$ and $K(i,i)$ gives us (2.4).

6 The concern is that the square root density computing may induce rounding error. In fact $\frac{1}{2}\log\int(dG_\theta/dG_\vartheta)^{1/2}dG_\vartheta$ approximately equal to $\frac{1}{2}\sum_i \log(dG_\theta/dG_\vartheta)(x_i)$ when x_i is generated by G_ϑ.

then the log-likelihood ratio between $\tilde{p}_{\theta_0}$ and $\tilde{p}_{\theta_0+\delta_n\tau}$ can be approximated by:

$$2\sum_{i=1}^{n}\log\frac{\tilde{p}_{\theta_0+\delta_n\tau_n}}{\tilde{p}_{\theta_0}}(x_i)=\delta_n\tau_n^T A_1+\frac{1}{2}\delta_n^2\tau_n^T A_2\tau_n^T+o_p(1) \qquad (2.5)$$

where A_1 *is* $\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\sum_{i=1}^{n}m_i(\theta_0)$ *and* A_2 is $\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta^T}$.

The expansion (2.5) is obtained simply by Taylor expansion and the result therefore does not apply to the nonstandard applications we are interested in. However, the result is intuitive as it mimics the standard Local Asymptotic Normal (LAN) property for parametric models, see e.g. van der Vaart (1998, pp 104). The relation between (2.5) and (2.2) is also quite clear: the first term is τ_n times a random vector, and the second term is its variance.

Remark 2.13. With the additional normality assumption on the average of $m_i(\theta_0)$ and assuming $\delta_n=n^{-1/2}$ we will of course have:

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

Asymptotic normality of the EL estimator is established by equation (2.5) with additional conditions on the continuity or the boundedness of second derivative of the moment restriction functions, e.g. Qin and Lawless (1994), Newey and Smith (2004) or Kitamura et al. (2004).

Remark 2.14. An alternative way of deducing this asymptotic normality is via Differentiability in Quadratic Mean (DQM). This entails the existence of a vector of measurable functions $S_{\theta_0,n}$ such that

$$\int\left[\tilde{p}_{\theta_0+\delta_n\tau}^{1/2}-\tilde{p}_{\theta_0}^{1/2}-\frac{1}{2}\delta_n\tau^T S_{\theta_0,n}\tilde{p}_{\theta_0}^{1/2}\right]^2 d\mu=o(\|\delta_n\|^2), \qquad (2.6)$$

where $\delta_n\to 0$. 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 along a path, the square root of the implied probability $\theta \mapsto \sqrt{\tilde{p}_\theta}$ is differentiable, then DQM basically means that a expansion of the square root of $\tilde{p}_\theta$ is valid and the remainder term is negligible in $L^2(\mu)$ norm. The term $S_{\theta,n}$ can be considered as the score function of the implied probability $\tilde{p}_\theta$ at θ_0. DQM implies that the condition does not require the pointwise definition of the derivative of $m(\theta, X)$ therefore it is less restrictive.

The implied probability includes the term $m(\theta, X)$ which is not always differentiable in nonstandard cases that we want to consider. It therefore deserves more effort to relax the restrictive condition on differentiability. In fact, Theorem 2.7 implies that the log-likelihood ratio belongs to the LAN family. The result is already good enough for constructing an efficient (or asymptotic sufficient) estimator. The expression in (2.2) is much weaker than the regular conditions and DQM. It only states that log-likelihood ratios of implied probabilities can be approximated by a linear-quadratic expression.

2.4 ESTIMATION

2.4.1 *Local Estimation*

By the result (2.2) in Theorem 2.7, we can study the behavior of a pair $(\tilde{P}_{\theta+\delta_n\tau_n}, \tilde{P}_\theta)$ by looking at the log-likelihood ratio process $\Lambda_n(\theta + \delta_n\tau_n, \theta)(X)$ with index τ_n. The log-likelihood ratio process admits linear quadratic approximations as $n \to \infty$, with the term $\tau_n S_n$ linear in τ_n and the term $\tau_n^T K_n \tau_n$ quadratic in τ_n. The numerical values of the approximation depend on the concentrated point θ and its local neighborhoods. With these ideas in mind, we will show the following steps of constructing a local type estimator. The explanation of each step is given after the definition.

Definition. Given Condition 2.1, 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 and restricted such that it lies in Θ_n (a δ_n-sparse discretization of Θ).

Step 2. Construct a matrix K_n with $K_{n,i,j} = u_i^T K_n u_j$, $i, j = 1, 2, \ldots, d$, given by

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

and $\{u_1, \ldots, u_d\}$ is a linearly independent set of directional vectors in $\mathbb{R}^d$ selected in advance.

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 statistic.

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 if it is larger than $\sum_i \log n\tilde{p}_{\theta_n^*}(x_i)$ then the estimator is T_n, otherwise it is θ_n^*. When $T_n \neq \theta_n^*$, if the difference between $\lambda_n(\theta_n^*)$ and $\lambda_n(T_n)$ is larger than a certain criterion value then go back to Step 2 and replace θ_n^* with T_n.

STEP 1 The *δ_n-sparse* (discretization of the) parameter space in Step 1 is suggested by Le Cam (see Le Cam and Yang (2000, p 125)). It requires a sequence of subsets $\Theta_n \subset \Theta$ satisfying the following conditions (i) that for any $\theta \in \Theta$ and any constant $b \in \mathbb{R}^+$, 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. In other words, asymptotically θ_n^* should be close enough to θ_0. Another interpretation of δ_n-sparsity is from a Bayesian perspective. That is for arbitrary priors, the corresponding posteriors essentially concentrate on the small vicinities shrinking at the rate δ_n.

STEP 2 As in the remark in previous section, the covariance matrix in Step 2 is an analog to the covariance kernel in Gaussian processes. For a stationary Gaussian process, the covariance kernel is smooth and differentiable in quadratic mean, the covariance kernel can be written as

$$\begin{aligned}
&\mathrm{Cov}\left(\frac{1}{\delta_n}(G_{\theta+u\delta_n} - G_\theta), \frac{1}{\delta_n}(G_{\vartheta+u\delta_n} - G_\vartheta)\right) \\
&= \frac{1}{\delta_n^2}\left(2C(\theta-\vartheta) - C(\theta-\vartheta+u\delta_n) - C(\theta-\vartheta-u\delta_n)\right) \\
&\to -\left.\frac{\partial^2 C(h)}{\partial h^2}\right|_{h=\theta-\vartheta},
\end{aligned}$$

where $C(\theta, \vartheta) := \mathrm{Cov}\{G_\theta, G_\vartheta\}$. Since K_n is an analog to the covariance kernel, the construction of K_n is nothing else but a finite difference of $\Lambda_n(\cdot,\cdot)$ which is analogous to the second derivative of the covariance kernel.

STEP 3 AND 4 With a control term K_n which is asymptotically determined, all the randomness of the log-likelihood ratio is contained in the first term, S_n. Step 3 is to extract the randomness from $\Lambda_n(\cdot,\cdot)$ and construct the linear term. Step 4 is to construct the estimator. To verify these two steps, we need to ensure that the covariance kernel in (2.2) is invertible.

Proposition 2.15. *The matrices $K_{\theta,n}$ in* (2.2) *are almost surely positive definite. Any cluster point K_θ of $K_{\theta,n}$ in $P_{\theta,n}$-law is invertible.*

If $K_n - K_{\theta,n}$ converges to zero, then K_n is also invertible. This result will be given in the following Theorem. If K_n is positive definite, by substituting $S_n = K_n\delta_n^{-1}(T_n - \theta_n^*)$ into the linear quadratic expression:

$$\tau_n^T K_n \delta_n^{-1}(T_n - \theta_n^*) - \frac{1}{2}\tau_n^T K_n \tau_n = -\frac{1}{2}\delta_n^{-2}\left[T_n - (\theta_n^* + \delta_n\tau_n)\right]^T K_n \times \left[T_n - (\theta_n^* + \delta_n\tau_n)\right] + \frac{1}{2}\delta_n^{-2}\left[T_n - \theta_n^*\right]^T K_n \left[T_n - \theta_n^*\right],$$

we have a quadratic expression of T_n and $(\theta_n^* + \delta_n\tau_n)$. The maximal value of this approximating representation of the log-likelihood ratio is achieved when $\theta_n^* + \delta_n\tau_n = T_n$. In other words, $\delta_n^{-1}(T_n - \theta_n^*)$ is the estimator for the local parameter τ_n.

Remark 2.16. The construction was originally proposed by Le Cam (1974). He supposed that there is a special interest in the likelihood function at particular points where Taylor's expansion fails, e.g. for the Laplace distribution. The advantage of the construction is that the quadratic term does not depend very much on the particular auxiliary estimation method that is used to obtain the value of θ_n^* and the construction is only determined in a local neighborhood of the particular point.

Remark 2.17. One may be concerned with the δ_n-consistency requirement for the auxiliary estimator. For a simple i.i.d. case, the δ_n is set to $n^{-1/2}$, the requirement is the same as asking for an $\sqrt{n}$-consistent auxiliary estimator. Any $\sqrt{n}$-consistent estimator should be, in principle, good enough from the estimation perspective, because the auxiliary estimator θ_n^* is at least in a neighborhood of θ_0. However, in practice, it may be hard to find a well behaved moment restriction function around θ_0. The use

of local EL estimator is to overcomes the problem and improve the auxiliary estimator. We suppose that θ_n^* is located within a range $n^{-1/2}$ of the true value, then a local method would give a refinement. When consistency and asymptotic normality are treated separately, one could take good care of consistency first and then use localization method to improve the final result or one could take care of the concentration of distribution first and then correct the bias by localization.

Theorem 2.18. *Given conditions 2.1 and 2.4, T_n, S_n and K_n have following properties:*

(i) $K_n^{-1}S_n - K_{\theta,n}^{-1}S_{\theta,n}$ and $K_n - K_{\theta,n}$ converge to zero in $\tilde{P}_{\theta,n}$-law where $(K_{\theta,n}, S_{\theta,n})$ is in (2.2)*.*

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

(iii) if Equation (2.6) *holds and the moment restrictions are just-identifying, the sequence of models $\{\tilde{P}_{\theta,n} : \theta \in \Theta\}$ is LAN and*

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

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

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 $\tau_n = \delta_n^{-1}(\theta_0 - \theta_n^*)$ can be used to construct a log likelihood ratio based on $P_{\theta_0+\tau_n\delta_n}$ that is asymptotically $\mathcal{N}(\tau_n, I_{\theta_0}^{-1})$. 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.

Remark 2.19. Some other articles also utilize local information based on an EL framework. Donald et al. (2003) propose re-sampling data from a local EL estimated distribution. Kitamura et al. (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. These methods construct smooth objective functions, implicitly or explicitly. Our solution is to discretize the parameter space and then construct local log-likelihood ratios as local objective functions.

Theorem 2.18 gives an asymptotic result on the weak convergence of the estimator. In the theorem, the limit distribution is based on a kind of Cramér-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 a technical concept of δ_n-*regularity*. This concept expresses the desirable requirement that a small change in the parameter should not change the distribution of estimator too much. For the estimator sequence T_n, if the difference between the distributions of $\delta_n^{-1}(T_n - \theta_0 - \delta_n\tau)$ and $\delta_n^{-1}(T_n - \theta_0)$ tends to zero under $P_{\theta_0+\delta_n\tau,n}$-law and $P_{\theta_0,n}$-law respectively, then T_n is called δ_n-regular at the point θ_0.

Theorem 2.20. *Given Condition 2.1 and letting W be a non-negative bowl shaped loss function, if T_n is δ_n-regular on all Θ, then for any estimator sequence Z_n of τ, one has*

$$\lim_{b\to\infty}\lim_{c\to\infty}\lim_{n\to\infty}\inf_{n}\sup_{|\tau|\le c}\mathbb{E}_{\theta_0+\delta_n\tau}\left[\min(b, W(Z_n-\tau))\right] \ge \mathbb{E}[W(\xi)]$$

where ξ has a Gaussian distribution $\mathcal{N}(0, K^{-1})$. The lower bound is achieved by $Z_n = \delta_n^{-1}(T_n - \theta_0)$.

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 $\min(b, W(Z_n - \tau_n))$. We let c go to infinity in order to cover a general case. The expectation $\mathbb{E}_{\theta_0+\delta_n\tau}[\cdot]$ is taken w.r.t. a measure $\mathcal{M}$ of the set $\{\theta : |\theta - \theta_0| \le \delta_n\tau_n\}$ while $\mathbb{E}[\cdot]$ is taken w.r.t. a distribution of $K^{-1/2} \times \mathcal{N}(0, I)$ on ξ.

The theorem can be interpreted as follows. When using the auxiliary estimator θ_n^* in the likelihood ratio, this induces randomness to the local parameter τ_n. By using the LAN result in Theorem 2.18, we can attach the local parameter τ_n with a Gaussian measure. By the Gaussian prior assumption of τ_n, one can express the convergent procedure as a procedure of updating a Gaussian prior, while for a centered Gaussian prior, this procedure is to update the prior covariance matrix Γ^{-1}. The δ_n-regularity condition implies that K_n will converge uniformly in a neighborhood of θ_0 for arbitrary measure $\mathcal{M}$. Thus the covariance will converge to a the posterior covariance matrix $(K+\Gamma)^{-1}$. The Gaussian randomness introduces a new random variable ξ that has the posterior covariance matrix $(K+\Gamma)^{-1}$. The lower bound of the Bayes risk of this Gaussian variable is obtained by letting Γ go to zero, corresponding to initial values of τ widely spread. This is the 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 an asymptotically optimal estimator.

2.4.2 *Global Estimation*

From our previous discussion, we can see that the construction of T_n relies on the assumption that, locally, the logarithms of likelihood ratios admit approximations of a quadratic nature:

$$\tau_n^T K_n \delta_n^{-1}(T_n - \theta_n^*) - \frac{1}{2}\tau_n^T K_n \tau_n = -\frac{1}{2}\delta_n^{-2}\left[T_n - (\theta_n^* + \delta_n\tau_n)\right]^T K_n \times \left[T_n - (\theta_n^* + \delta_n\tau_n)\right] + \frac{1}{2}\delta_n^{-2}\left[T_n - \theta_n^*\right]^T K_n \left[T_n - \theta_n^*\right].$$

The method is to use the auxiliary estimate from local expansions to construct the estimate T_n and covariance kernel K_n. As we can see, the kernel K_n controls the quality of representation. In the local case, because of δ-sparsity and the auxiliary θ_n^*, the kernel matrix $K = \lim K_n$ is deterministic. If the quadratic nature can be extended to a global region, namely linear quadratic representation valid on the whole or at least a large part of the parameter space as the linear-quadratic representation (2.2), then K is not necessarily fixed; in fact, it should vary with θ.

If we are worried about irregular situations, e.g. flatness, non-smoothness, non-differentiability, on a large part of Θ for Λ_θ, then we should turn to a global method that can mimic the good properties of our local estimator. The difficulty of extending the linear quadratic representation to all of the parameter space Θ comes from the randomness of K_θ. In this section, we will encounter this problem, where the form of the population K_θ could be random. In order to emphasize this property, we will use the notation:

$$\mathbb{K}_\theta(T_n - \theta) := (T_n - \theta)^T K_{\theta,n}(T_n - \theta) \tag{2.7}$$

for the global case, namely $\mathbb{K}_\theta(\cdot)$ is a quadratic random function on $\theta \in \Theta$. Conditional on θ, the linear quadratic approximation still holds, like (2.2). When K is random, the approximation belongs to the Locally Asymptotically Mixed Normal (LAMN) class. The reason for using quadratic form in (2.7) instead of the linear-quadratic one is that both "linear" and "quadratic" terms will be influenced by the choice of parametrization, it is better to use one term rather than two to express this variational effect of θ. Because the constructed estimator will depend on the value of θ, from now on, we will denote T_n as $T_{\theta,n}$.

For practical application, it is more convenient to work in a well-defined space rather than an arbitrary functional space on Θ. We want to attach to the parameter space Θ a vector space $\mathcal{F}(\Theta)$ induced by K_θ which is the "kernel" of this space $\mathcal{F}(\Theta)$. To see the necessity of doing this, one should compare to the situation in LAN. The LAN result can be stated as a pointwise result in θ. Thus the convergence and optimality properties are also valid pointwise like Theorem 2.18, or just a small region around θ like Theorem 2.20. If one wants to extend the result to the whole Θ space, namely, a uniform result on Θ, then one needs an appropriate uniformity condition. If the subspace $\mathcal{F}(\Theta)$ is constructed by K_θ, then an estimate using K_θ will be uniformly valid. The following condition is to define such a subspace.

Condition 2.21. (i) The vector space $\mathcal{F}(\Theta)$ is a linear subspace with finite signed measures and with finite support on Θ. Let

$$\mathcal{F}(\Theta) := \{\nu : \nu(\Theta) = 0\}.$$

We require that for infinite number of observations, any measure $\nu \in \mathcal{F}(\Theta)$ for centering $T_\theta - \theta$ is constructed by K_θ.

(ii) The inner product for $\mathcal{F}(\Theta)$ is defined by bilinear function

$$\langle \theta, \vartheta \rangle_K = \frac{1}{2}\{K_{\theta+\vartheta} - K_\theta - K_\vartheta\} = -\frac{1}{2}\{K_{\theta-\vartheta} - K_\theta - K_\vartheta\}$$

and $\theta \mapsto \sqrt{K_\theta}$ is the semi-norm on $\mathcal{F}(\Theta)$.

(iii) Let $\bar{\Lambda}_\theta(\vartheta, \theta^*) = -\{\mathbb{K}_\theta(T_\theta - \vartheta) - \mathbb{K}_\theta(T_\theta - \theta^*)\}/2$. The function $\bar{\Lambda}_\theta(\cdot,\cdot)$ is additive on its domain such that

$$\bar{\Lambda}_\theta(\vartheta, u) = \bar{\Lambda}_\theta(\vartheta, \theta^*) + \bar{\Lambda}_\theta(\theta^*, u).$$

Condition (i) is to give a tractable structure for the problem. The *null space* $\mathcal{F}(\Theta)$ is of particular interest. For any subset in Θ, K_θ is a quadratic form. The restriction of K_θ is to symmetrize $K(\cdot,\cdot)$. The restriction of $T_\theta - \theta$ is to make T as a centering estimator. (ii) is also called *the polarization condition* which ensures that the parallelogram law holds. This condition attaches a kind of Hilbert space characteristics to the linear space $\mathcal{F}(\Theta)$. One can compare (ii) with Theorem 2.7 where the canonical Gaussian family is attached to the Hilbert space. The condition basically says that we want the constructed quadratic K_θ to inherit such a property automatically. (iii) imposes a feature of presumed midpoints as Step 2 in local estimator's construction. This relation is to connect the quadratics with logarithms.

The motivation of these conditions is to ensure that the entire family $\mathcal{E}_\theta$ is globally approximable by a heteroskedastic Gaussian family where the log-likelihood contains constructed K_θ only and K_θ varies slowly enough as θ varies. Then conditional on any local value θ, one could also approximate $\mathcal{E}_\theta$ by a Gaussian family. The heteroskedastic approximation is uniformly feasible over θ. These conditions are extracted from the properties of *Quadric*, the solution rings of algebraic linear quadratic equations. The simplification allows us to work on a more concrete topological structure rather than on rings.

The aim of a global construction is to find centering values of K_θ and an estimator $T_{\theta,n}$ based on K_θ. Consider the EL log-likelihood ratio

$$\Lambda_n(s) = \frac{1}{n}\sum_{i=1}^{n} \log n\tilde{p}_i(s) = \Lambda_n(s,\theta_0),$$

where θ_0 induces the counting measure. For any initial parameter value t, minimizing $\Lambda_n(s)$ is equivalent to finding some ideal centering values which make the difference

$$\Lambda_n(s) + \frac{1}{2}[\hat{\mathbb{K}}(T_{\theta,n} - s) - \hat{\mathbb{K}}(T_{\theta,n} - t)]$$

tends to zero in probability by updating t, where $\hat{\mathbb{K}} = \mathbb{K}_{T_{\theta,n}}$ and $T_{\theta,n}$ are our estimators.

Definition. Given conditions 2.1 and 2.21, the global estimators are constructed as follow:

Step 1. Give two initial values t and s, compute the EL value $\Lambda_n(t)$.

Step 2. Find a $T_{\theta,n}$ to solve the linear equation system:

$$\langle T_{\theta,n} - s, t - s\rangle_{K_t} = \bar{\Lambda}_t(s,t) + \frac{1}{2}\mathbb{K}_t(s-t) \tag{2.8}$$

where the inner product and $\bar{\Lambda}_t(\cdot,\cdot)$ are defined in condition 2.21 and $\mathbb{K}_t(\cdot)$ is constructed by Step-2 in the local method .

Step 3. Adjust the parameter t by solving the quadratic problem:

$$\frac{1}{2}[\hat{\mathbb{K}}(T_{\theta,n} - s) - \hat{\mathbb{K}}(T_{\theta,n} - t)] = \Lambda_n(s).$$

If $|t^* - t|$ is larger than a certain critical value, e.g. 10^{-5}, then go back to step 1 and use t^* as the starting value.

It is a recursive algorithm type construction. Step 2 and 3 involve heavier computational tasks compared to the local method. But the whole setting avoids computing derivative and is feasible for arbitrary values s and t.

2.5 ROBUSTNESS

The robustness we consider here is essentially Huber's idea that an estimator is insensitive to perturbations of the model. This includes insensitivity to outliers if we think of outliers as being generated by a different model with a small probability. When the assumed structure of the model is incorrect or the DGP is wrongly specified, one can detect, in principle, the misspecification 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 analyze an inference procedure that explicitly takes into account that the model can, to a certain extent, be misspecified.

The sensitivity of EL estimation results from the unboundedness of the moment constraints. We borrow Huber's setting to illustrate this problem. Consider our estimator T_n as a statistical functional of an empirical measure P_n such that

$$\frac{1}{n}\sum_{i=1}^{n} m(T_n, X_i) = \frac{1}{n}\sum_{i=1}^{n} m(T(P_n), X_i) = 0. \tag{2.9}$$

The functional $T(P_n)$ it is defined as:

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

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 x's 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 asks for stronger condition. Because $IF(x, T, P_n)$ is defined by the functional derivative in 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. The influence function is unbounded if $m(\theta)$ is unbounded or if the derivative is not defined. An unbounded influence function implies an unbounded asymptotic bias of a statistic at a single-point contamination of the model.

White (1982) shows that Maximum Likelihood (ML) defined in terms of the Kullback-Leibler divergence is robust. EL inherits a lot of properties from ML, 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 associated with these outliers' moment restriction functions will give strong penalties so that the values of λ will stay close to zero independently of the value of θ. The implied density $\tilde{p}_i$ of each outlier's moment restriction function equals $1/n$. When the sample size is very small, this means that the effects of relative weights on the outliers are strong, the criterion function $\sum_i^n \log n\tilde{p}_i(\theta)$ will be quite flat on θ. In large samples, the intrinsically misspecified EL estimator will have a slower convergence rate, although it may be consistent.

As we see, the statistical functional of EL estimation is a solution of (2.9). The moment restriction function $m(T(P_n), x)$ as in the previous discussion is discontinuous on those peculiar points causing unbounded "influence function" of $T(P_n)$ and (2.9) is non-differentiable on these points as well, thus EL is not a robust estimation procedure 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. Localization can prevent such misbehavior.

If a peculiar point x_k drives $T(P_n)$ unbounded in a direction $u_i \in \Theta$, then the moment restrictions function $m(T(P_n), x_i)$ becomes unbounded and λ_n may not be zero because

$$\sum_i m(T(P_n), x_i) \neq 0.$$

However, we have a strong belief that the model is correctly specified. Thus we force λ_n to zero for the moment restriction and exclude the effect from the peculiar point. We achieve this goal by selecting a direction of the local estimation which guides λ_n to zero.

Note that the auxiliary estimator in local EL, θ_n^* in the range of order δ_n, does admit a good quadratic approximation of the log-likelihood ratios. This regularizes the matrix in the quadratic term to be positive semi-definite. In step 2 of the local EL's construction, we let

$$K_{n,i,j} = -\left\{\Lambda_n[\theta_n^* + \delta_n(u_i + u_j), \theta_n^*] - \Lambda_n[\theta_n^* + \delta_n u_i, \theta_n^*] - \Lambda_n[\theta_n^* + \delta_n u_j, \theta_n^*]\right\}.$$

The direction u in $K_{n,i,j}$ is constructed by a bisection type method. The correctly specified model must satisfy an auxiliary condition $f(\theta_0) = \lambda_n^T(\sum_i m(\theta_0, x_i)) = 0$. $\lambda_n(\theta_0)$ is dual parameter of $\sum_i m(\theta_0, x_i)$, when $\sum_i m(\theta_0, x_i) = 0$ is zero then $\lambda_n(\theta_0)$ is zero. Suppose $f(\theta_n^*)$ is positive. We select a direction u such that $f(\theta_n^* + u) = -f(\theta_n^*)$. By the mean value theorem, $f(\theta)$ must have at least one root $f(\theta) = 0$ between $\tilde{\theta}$ and θ_n^*. The constructed $K_{n,i,j}$ and S_n using this direction will make the local estimator T_n leave from θ_n^* to θ_0.

Remark 2.22. 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, the moment constraint functions are correctly specified, although certain sample points may lead to discontinuity and the unboundedness problems previously described. Applying mis-specification test may reject this essentially 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 2.23. Schennach (2007) shows that an Exponential Tilting Empirical Likelihood (ETEL) estimation is robust and almost as efficient as EL. However, it is still less efficient in the higher order than EL. Moreover, the procedure of ETEL changes the divergence criterion in the intermediate step[7]. 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.

2.6 CONCLUSION

In this chapter, we propose a new local EL method. We discuss its construction and have derived theoretical properties. The construction is based on the infinite divisibility property which is one of the crucial features in stochastic processes; to the best of our knowledge, this feature has not yet been applied to EL. When the implied probability of EL is embedded in the infinitely divisible class, the log-likelihood ratio admits a local representation. Our local estimator is built on the basis of this representation. The consistency, local asymptotic normality, and asymptotic optimality of this estimator have been established. These results depend on conditions that are weaker than usual and allow for applications when the standard regularity conditions are violated.

7 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.

APPENDIX TO CHAPTER 2

PROOF OF THEOREMS

Proof of Theorem 2.2

The Lagrangian of EL is

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

where λ and γ are Lagrange multipliers. Setting the partial derivative of L w.r.t. p_i equal to zero will give $\gamma = n$ and the implied probability $\tilde{p}_i = 1/(\gamma + n\lambda_n^T m_i(\theta))$. By the implicit function theorem, the partial derivative of $\sum_{i=1}^{n} \log \tilde{p}_i$ w.r.t. λ gives a function $Im(\cdot,\cdot)$ of λ_n and θ such that

$$\begin{aligned} \frac{\partial \sum \log \tilde{p}_i}{\partial \lambda} &:= Im(\lambda_n, \theta) = 0, \qquad (2.10) \\ \Longrightarrow \frac{1}{n} \sum_{i=1}^{n} \frac{m_i(\theta)}{1 + \lambda_n^T m_i(\theta)} &= \sum_{i=1}^{n} \tilde{p}_i(\theta) m_i(\theta) \end{aligned}$$

where λ_n is unique for fixed n and θ. Note that $Im(\lambda_n, \theta) = 0$ for $\forall \theta \in \Theta$ and θ is continuous hence $Im(\cdot)$ is continuous in θ. By the continuity of $m(X, \theta)$ and the representation of $Im(\cdot)$, we know that λ_n is also continuous on θ. The proof of the uniqueness of $\lambda(\theta)$ is as follows: because the set $\Gamma(\theta) = \lim_{n\to\infty} \cap_{i=1,\ldots,n} \{\lambda | 1 + \lambda^T m(X_i, \theta) > 1/n\}$ is convex if it does not vanish, the function of $\log p$ is strictly concave on λ, so $\lambda(\theta)$ exists and is unique.

With these, the properties of likelihood ratio are shown in as follows. Equation (2.10) can be re-written as

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

Condition 2.1 (v) states that $n^{-1}\sum_i^n m_i(\theta)m_i(\theta)^T$ is positive definite, let $\mathbf{c}$ be larger than any eigenvalue of $n^{-1}\sum_i^n m_i(\theta)m_i(\theta)^T$ and let v be the corresponding eigenvector. The convex combination of $m_i(\theta)m_i(\theta)^T$ over $\{\tilde{p}_i(\theta)\}$ in $(*)$ is bounded by $v^T\mathbf{c}v$. Let $E_v = v^T\mathbf{c}v$. According to condition 2.1 (iv), $m_i(\theta)$ has an envelop function $b(\theta)$ such that $\liminf_\theta |m(\theta,X)|/b(\theta) \geq 1$, then

$$\lim_{n\to\infty} |\lambda_n|/b^{'}(\theta) \geq 1$$

for any θ where $b^{'}(\theta) = b(\theta)/E_v$.

Let's first prove the existence of $\Lambda(\theta)$:

$$\lim_{n\to\infty}\int \log\frac{1}{n}\frac{n}{1+\lambda(n,\theta)^T m(x,\theta)}dP(x) = \tag{2.11}$$

$$\mathbb{E}\lim_{n\to\infty}\log\frac{1}{1+\lambda(n,\theta)^T m(X,\theta)} = \mathbb{E}\log\frac{1}{1+\lambda(\theta)^T m(X,\theta)} = \Lambda(\theta).$$

The first convergence is by the LLN and the second equation is obtained by the dominated convergence Theorem, since $[1+\lambda(\theta)^T m(x,\theta)]^{-1}$ is bounded and $\lambda(\theta)$ exists.

Next we prove the continuity of $\Lambda(\theta)$. The envelop functions $b^{'}(\theta)$ and $b(\theta)$ are integrable and continuous (Condition 2.1), $\lambda(\theta)^T m(X,\theta)$ is bounded by a continuous function. Thus $\Lambda(\theta)$ is continuous and is bounded by an envelop function $b^{''}(\theta) = \max(b^{'}(\theta), b(\theta))$ such that

$$\sup_\theta \|\Lambda_n(\theta) - \Lambda(\theta)\|/b^{''}(\theta) < 1 \tag{2.12}$$

Now prove the identifiability of EL estimation. Choose a compact set $\Theta_c \subset \Theta$ such that for given ϵ

$$\sup_{\theta\in\Theta_C} |\Lambda(\theta)|/b^{''}(\theta) \geq 1-\epsilon.$$

By (2.12), LLN applied to $\Lambda_n(\theta)$ implies

$$\begin{aligned}\sup_\theta \frac{\|\Lambda_n(\theta)-\Lambda(\theta)\|}{b''(\theta)} &< \frac{\|\Lambda_n(\theta)\| - \|\Lambda(\theta)\| + 2\|\Lambda(\theta)\|}{b''(\theta)} - \epsilon \\ &< \frac{\|\Lambda_n(\theta)\| - \|\Lambda(\theta)\| + 2\sup_{\theta\in\Theta_C}|\Lambda(\theta)|}{b''(\theta)} - \epsilon \\ &< \frac{\|\Lambda_n(\theta)-\Lambda(\theta)\| + 2\sup_{\theta\in\Theta_C}|\Lambda(\theta)|}{b''(\theta)} < 1-3\epsilon\end{aligned}$$

The first inequality uses triangle inequality, the second one uses supremum property, and the third one uses triangle inequality again. Therefore

$$|\Lambda_n(\theta) - \Lambda(\theta)| \leq (1 - 3\epsilon)b''(\theta)$$
$$\leq \frac{1-3\epsilon}{1-\epsilon} \sup_{\theta \in \Theta_C} |\Lambda(\theta)| \leq (1-\delta) \sup_{\theta \in \Theta_C} |\Lambda(\theta)|$$

for $\forall \theta \in \Theta_c$. This inequality implies

$$\sup_{\theta \in \Theta_c} |\Lambda_n(\theta)| \leq \sup_{\theta \in \Theta_c} |\Lambda(\theta)| + \epsilon$$

asymptotically for any $\theta \in \Theta_c$. Thus if $\theta_0 \in \Theta_c$, then

$$\{T_n \subset \Theta_c\} \subset \left\{ \sup_{\theta \in \Theta_c} \Lambda_n(\theta) \leq \Lambda(\theta_0) + o_p(1) \right\},$$

where the probability of the event on the right side converges to one as $n \to \infty$. Because the compact set Θ could be shrinking to an arbitrary neighborhood of θ_0, the EL estimator T_n is consistent.

Proof of Theorem 2.7

Before proving the Theorem, we need to introduce a relation for univariate Gaussian families. For any pair of Gaussian measures in $\mathcal{G}_\Theta = \{G_\theta, \theta \in \Theta\}$, $G_\theta \subset \mathcal{E}_\theta$, there will be an expression to relate both of them as follows:

$$dG_\theta = \exp\left[\langle Y_\vartheta, \theta\rangle - \frac{1}{2}\|\theta\|^2\right] dG_\vartheta, \tag{2.13}$$

$$\exp\left[\frac{1}{2}\langle Y_\vartheta, \vartheta + \theta\rangle\right]$$

where $\vartheta, \theta \in \Theta$. The bilinear product in this expression is $\langle Y_\vartheta, \theta\rangle = \int_0^1 Y_\vartheta(t) G_\theta(dt)$ where Y_ϑ is a univariate Gaussian process. This is a random variable (functional integral or Wiener integral) with mean zero and variance $\|\theta\|^2 \leq \infty$ [8]. If dG_θ and dG_ϑ are defined as (2.13), the integral of $(dG_\theta / dG_\vartheta)^{1/2}$ w.r.t. G_ϑ will has a linear quadratic representation.

8 This expression is called weak form expression and is often used for generalizing Gaussian processes.

Proof. Le Cam and Yang (2000, Proposition 4.1) show that the affinity between two Poissonized $d\tilde{P}_\theta, d\tilde{P}_\vartheta$ is

$$\int \sqrt{d\tilde{P}_\theta d\tilde{P}_\vartheta} = \exp\left\{-\frac{1}{2}\|\theta - \vartheta\|^2\right\}.$$

Since Gnedenko and Kolmogorov (1968, Theorem 17.5) show that finite many number of Poisson type measures can approximate any infinitely divisible family and EL is embedded in an infinitely divisible family, we know the above expression is applicable over here. The Hellinger affinity for Gaussian family is

$$\int \sqrt{dG_\theta dG_\vartheta} = \int \exp\left[\frac{1}{2}\langle Y_\vartheta, \vartheta + \theta\rangle - \frac{1}{4}(\|\theta\|^2 + \|\vartheta\|^2)\right] dG_\vartheta.$$

The Gaussian property of $\langle Y_\vartheta, \vartheta + \theta\rangle$implies that is log-normal distributed, then by log-normal property there is:

$$\int \exp\left[\frac{1}{2}\langle Y_\vartheta, \vartheta + \theta\rangle\right] dG_t = \exp\left(\frac{1}{8}\|\theta + \vartheta\|^2\right).$$

Because only metric distance is going to be studied in $\int \sqrt{dG_\theta dG_\vartheta}$, we attach a Hilbert space to $\mathcal{G}$. The parallelogram identity for Hilbert space induces

$$\|\theta + \vartheta\|^2 + \|\theta - \vartheta\|^2 = 2\left(\|\theta\|^2 + \|\vartheta\|^2\right),$$

so

$$2\left(\|\theta\|^2 + \|\vartheta\|^2\right) + \|\theta + \vartheta\|^2 = -\|\vartheta - \theta\|^2$$

Therefore, $\sqrt{dG_\theta dG_\vartheta} = \exp(-\|\theta - \vartheta\|^2/8)$ which is isometric to $\int \sqrt{d\tilde{P}_\vartheta d\tilde{P}_\theta} = \exp(-\|\theta - \vartheta\|^2/2)$. If Fubini's theorem holds, the expression

$$2\log \int \left(\frac{d\tilde{P}_\theta}{d\tilde{P}_\vartheta}\right)^{\frac{1}{2}} d\tilde{P}_\vartheta \approx 8\log \int \left(\frac{dG_\theta}{dG_\vartheta}\right)^{\frac{1}{2}} dG_\vartheta$$

implies

$$\int \left[\log \frac{d\tilde{P}_\theta}{d\tilde{P}_\vartheta}\right] d\tilde{P}_\vartheta = 4 \int \left[\log \frac{dG_\theta}{dG_\vartheta}\right] dG_\vartheta$$

so that we can use the Gaussian expression (2.13) for the log-likelihood ratio process.

By Karhunen–Loève Theorem (Kallenberg, 2002), the Gaussian process Y_θ can be expressed as

$$Y_\theta = \sum_{j=1}^{\infty} \xi_j \mathbf{u}_j(\theta)$$

where $\{\mathbf{u}_j\}$ constitutes an orthonormal basis for the Hilbert space $\mathcal{G}$ and ξ_j are Gaussian random variables and stochastically independent. Now let $\mathbf{u}_j(\cdot) = \sum_i^m \tau_i \mathbf{e}_i(\cdot)$ where $\mathbf{e}$ is a unit basis for the local parameter space and τ_i are linear coefficients for $\mathbf{e}_i(\cdot)$. Let j indicate the index of a basis on the Hilbert space and i indicate the index of a basis on the local parameter space. Then the inner product in the Hilbert space can be expressed using local parameter coordinates such that $\langle Y_\vartheta, \theta \rangle = \sum_i^m \tau_i \theta_i \langle \mathbf{e}(\vartheta), \xi \rangle = \tau^T(\theta\tilde{\xi})$ where $\tilde{\xi}$ is also Gaussian because of the linear property. Let $\theta\tilde{\xi} = S'_\theta$ and $\mathbb{E}(\theta\tilde{\xi})^2 = K'_\theta$, then

$$\|\theta\|^2 = \mathbb{E}[\tau^T(\theta\xi)]^2 = \tau^T K'_\theta \tau.$$

From (2.13), we have

$$\int \left[\log \frac{d\tilde{P}_\vartheta}{d\tilde{P}_\theta}\right] d\tilde{P}_\theta = \tau^T S_\theta - \frac{1}{2}\tau^T K_\theta \tau.$$

where $S_\theta = \int S'_\theta d\tilde{P}_\theta$ and $K_\theta = \int K'_\theta d\tilde{P}_\theta$. For a finite dimensional Gaussian vector based on n realizations Gaussian process, we have the sample counterparts τ_n, $S_{\theta,n}$ and $K_{\theta,n}$. We conclude that the EL ratio is approximately equal to the log-likelihood ratio of $\mathcal{G}$, which for the sample of size n is $\tau_n^T S_{\theta,n} - \tau_n^T K_{\theta,n}\tau_n/2$. □

Proof of Theorem 2.18

Proof. (i) When θ is given, by equation (2.2)

$$\begin{aligned}\Lambda_n(\theta + \delta_n\tau_n, \theta) =& \tau_n^T S_{\theta,n} - \frac{1}{2}\tau_n^T K_{\theta,n}\tau_n + o_{\tilde{p}_\theta}(1) \qquad (2.14)\\ =& -\frac{1}{2}\Big[(K_{\theta,n}^{-1}S_{\theta,n} - \tau_n^T)^T K_{\theta,n}(K_{\theta,n}^{-1}S_{\theta,n} - \tau_n^T))\\ & -(S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n})\Big] + o_{\tilde{p}_\theta}(1).\end{aligned}$$

Similarly,

$$\begin{aligned}\Lambda_n(\theta + \delta_n\tau_n, \theta) &= \tau_n^T K_n \delta_n^{-1}(T_n - \theta_n^*) - \frac{1}{2}\tau_n^T K_n \tau_n \qquad (2.15)\\ &= -\frac{1}{2}\Big[(\delta_n(T_n - \theta) - \tau_n^T)^T K_n(\delta_n(T_n - \theta) - \tau_n^T)) \qquad (2.16)\\ &\quad -(\delta_n(T_n - \theta))^T K_n(\delta_n(T_n - \theta))\Big].\end{aligned}$$

The difference between (2.14) and (2.15) tends to zero as $n \to \infty$. Non-negativity of K_n and $K_{\theta,n}$ shows that each of the four

quadratic terms in (2.16) and (2.14) must be non-negative. If $S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$ converges to $(\delta_n(T_n-\theta))^T K_n(\delta_n(T_n-\theta))$, then

$$(\delta_n(T_n-\theta)-\tau_n^T)^T K_n(\delta_n(T_n-\theta)-\tau_n^T)) \to (K_{\theta,n}^{-1}S_{\theta,n}-\tau_n^T)^T K_{\theta,n}(K_{\theta,n}^{-1}S_{\theta,n}-\tau_n^T)).$$

So one can conclude that $K_n - K_{\theta,n} \to 0$ and $S_n - S_{\theta,n} \to 0$.

Now consider the opposite case $(\delta_n(T_n-\theta))^T K_n(\delta_n(T_n-\theta)) \nrightarrow S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$. By a standard property of quadratic functions, we can have for some positive-definite matrix C

$$(\delta_n(T_n-\theta))^T K_n(\delta_n(T_n-\theta)) + C \to S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$$

Then for some vector Δ such that $\delta_n \Delta^T K_n \Delta \delta_n = C$, there is

$$(\delta_n(T_n-\theta+\Delta))^T K_n(\delta_n(T_n-\theta+\Delta)) \to S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$$

So $T_n + \Delta$ is optimal estimator for τ_n, because

$$(\delta_n(T_n-\theta+\Delta)-\tau_n^T)^T K_n(\delta_n(T_n-\theta+\Delta)-\tau_n^T)) \to (K_{\theta,n}^{-1}S_{\theta,n}-\tau_n^T)^T K_{\theta,n}(K_{\theta,n}^{-1}S_{\theta,n}-\tau_n^T)).$$

But this contradicts with our definition of T_n.

Thus $(\delta_n(T_n-\theta))^T K_n(\delta_n(T_n-\theta))$ converges to $S_{\theta,n}^T K_{\theta,n}^{-1} S_{\theta,n}$. It implies K_n converges to $K_{\theta,n}$ in probability and $\delta_n(T_n-\theta)$ converges to $K_{\theta,n}^{-1}S_{\theta,n}$.

(ii) By Proposition 2.15, we know that clustering points K_θ of $K_{\theta,n}$ are 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 $K_\theta^{-1}S_{\theta,n}$. The Gaussian variable $S_{\theta,n}$ is second moment bounded. So the term $\delta_n(T_n-\theta)$ is bounded in probability.

(iii) We know the DQM condition implies (2.2), thus the linear-quadratic equation (2.2) may coincide with S_n and K_n by (i). The log-likelihood process can be rewritten as a centered log-likelihood process $\Xi_n(\cdot)$ plus a shift item $b_n(\cdot)$:

$$\delta_n \Lambda_n(\theta,\vartheta)(x) = \overbrace{\frac{1}{n}\delta_n \sum_{i=1}^n \log\frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x_i) - \int \log\frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x)dP_0}^{\Xi_n(\theta)} + \underbrace{\int \log\frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x)dP_0}_{b_n(\theta)} + o_p(1).$$

Let $\delta_n = n^{-1/2}$. Given fixed $\lambda(\cdot)$ values in the constraint of equation (2.1), Theorem 2.7 says that $\log\frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x_i)$ in $\Xi_n(\eta)$ can

be replaced by a linear quadratic formulae w.r.t. τ_n, namely $\log \frac{\tilde{p}_\theta}{\tilde{p}_\vartheta}(x_i)$ belongs to a smooth functional class $\mathcal{C}^2$. Therefore the process $\theta \mapsto \Xi_n(\theta)$ is an empirical process and $\Xi_n(\theta) \rightsquigarrow \Xi(\theta)$ by Donsker's Theorem, see van der Vaart (1998, Example 19.9) where $\Xi(\theta)$ is a Gaussian process. Note that $\Xi(\theta)$ has mean $\int \Xi(\theta) dP_0 = 0$ and covariance kernel $\mathbb{E}\Xi^2(\theta)$ under P_0. The log-normal property implies that $\mathbb{E}\exp[\Xi(\theta) + b(\theta)] = 1$ with the expectation taken under P_0 and $b(\theta) = \lim_{n\to\infty} b_n(\theta)$. Log normal property of $\exp\Xi(\cdot)$ gives $b(\theta) = -(1/2)\mathbb{E}\Xi^2(\theta)$. By Proposition 2.12 and equation (2.2), we can show that

$$\begin{aligned}\Xi_n(\theta) =& S_{\theta,n}\\ b(\theta) =& -\frac{1}{2}K_\theta,\end{aligned}$$

and when $\theta = \theta_0$

$$\begin{aligned}\Xi_n(\theta_0) =& \mathbb{E}\left[\frac{\partial m(X,\theta_0)}{\partial\theta}^T \left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\right]\delta_n\sum_i^n m_i(\theta_0)\\ b(\theta_0) =& -\frac{1}{2}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T \left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta^T}.\end{aligned}$$

□

Proof of Theorem 2.20

Discussion: 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 expectation conditional on the local parameter τ. Note that the measure $\mathcal{M}$ has not yet been specified. If one can in Bayesian fashion give a prior distribution on $\mathcal{M}$, then what we need to study is the posterior distributions given this "local prior measures". In fact, the δ_n-sparse condition already implies that for arbitrary priors, the corresponding posteriors concentrate on the small shrinking neighborhood of θ_0.

Proof. First look at the population log-likelihood ratio

$$\begin{aligned}\Lambda(\theta+\tau,\theta) = -\frac{1}{2}\Big[&(K_\theta^{-1}S_\theta - \tau)^T K_\theta (K_\theta^{-1}S_\theta - \tau)\\ &-(S_\theta^T K_\theta^{-1} S_\theta)\Big] + o_{\tilde{p}_\theta}(1)\end{aligned}$$

which implies that the term $(K_\theta^{-1}S_\theta - \tau)^T K_\theta (K_\theta^{-1}S_\theta - \tau))$ is χ^2 distributed. The quadratic form of a Gaussian variable ξ, $\xi^T\xi$,

can generate exactly the same distribution. As Theorem 2.7 shows that the approximation of Gaussian family is feasible. For any value of θ, there will be such a ξ_θ whose distribution is equivalent to $K_\theta^{-1}S_\theta - \tau$ and has the variance $K_\theta^{-1/2}$. Then we have the expression

$$\tau = K_\theta^{-1}S_\theta - \xi_\theta,$$

which shows that τ consists of two Gaussian variables $K_\theta^{-1}S_\theta$ and ξ_θ. Thus we are able to impose a Gaussian structure on the measure $\mathcal{M}$.

Now we can look at the expectation $\min(b, \mathbb{E}[W(Z_n - \tau)|\theta_0 + \delta_n\tau])$ which is bounded by b. Since both "prior" and "posterior" concentrate around θ_0 and are Gaussian, the updating information only occurs for covariance matrix. Let τ be a Gaussian random variable centered at 0 with inverse covariance Γ. The conjugate property indicates the posterior of τ can be written as:

$$Z_n = \delta_n^{-1}(\tilde{T}_n - \theta_0) = (K_n + \Gamma)^{-1/2}K_n\delta_n^{-1}(T_n - \theta_0),$$

especially when $\Gamma = 0$, $Z_n = \delta_n^{-1}(T_n - \theta_0)$. By Anderson's Lemma[9], for bounded W, there is

$$\mathbb{E}[W(Z_n - \tau)|\theta_0 + \delta_n\tau] \geq \mathbb{E}[W(Z_n)|\theta_0 + \delta_n\tau].$$

Since $K_n\delta_n^{-1}(T_n - \theta_0) \sim \mathcal{N}(0, I)$, the lower bound of $\mathbb{E}[W(Z_n - \tau)|\theta_0 + \delta_n\tau]$ is

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

The measure of $\theta_0 + \delta_n\tau$ is replaced by $K_n + \Gamma$ because of the Gaussian property, namely the update of covariance matrix. Note that K_n and Γ are 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

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

9 For a symmetric distribution, shifting an integral function of it to a new position will product higher expected value, see van der Vaart (1998, Lemma 8.5).

for Δ with a small enough Euclidean norm $|\Delta|$. Especially, when Γ go to zero or say the measure $\mathcal{M}$ degenerates to a point eventually, $Z_n = \delta_n^{-1}(T_n - \theta_0)$ obtains the lower bound $\mathbb{E}[W(K_\theta^{-1/2}) \times \mathcal{N}(0, I)]$. If $W = 1$ and $K_\theta = K$, by Theorem 2.18(iii) we achieve the efficient bound of semi-parametric estimators. □

OTHER TECHNICAL DETAILS

Proof of Lemma 2.5

Proof. Note that for every fixed n and θ, $\log n\tilde{p}(\theta, X_i)$ is an independent random variable. Condition 2.4 is often used to deduce the upper bound of likelihood ratio function. The purpose of imposing this assumption is to get a bounded variance of $\log n\tilde{p}(\theta, X_i)$. Compactness implies, for a given ϵ, the existence of a number $a_n < \infty$ such that $\Pr\{|\Lambda_n(\theta)| > a_n/2\} < \epsilon/2$. Let $\Lambda_{n,k}(\theta) = n^{-1}\sum_i^k \log n\tilde{p}(\theta, X_i)$. Levy's inequality says that for any $k \leq n$ and $a_n \geq 0$, it holds that

$$\Pr\left[\sup_k |\Lambda_{n,k}(\theta)| \geq \frac{a_n}{2}\right] \leq 2\Pr\left[|\Lambda_n(\theta)| \geq \frac{a_n}{2}\right] < \epsilon.$$

Then by taking differences, $n^{-1}\log n\tilde{p}_k(\theta) = \Lambda_{n,k}(\theta) - \Lambda_{n,k-1}(\theta)$, we have

$$\begin{aligned}\Pr\left[\sup_k \frac{|\log n\tilde{p}_k|}{n} \geq a_n\right] &\leq \Pr\left[\sup_k \frac{|\Lambda_{n,k}(\theta) - \Lambda_{n,k-1}(\theta)|}{n} \geq a_n\right] \\ \leq \Pr\left[\sup_k \frac{2|\Lambda_{n,k}(\theta)|}{n} \geq a_n\right] &< 2\Pr\left[|\Lambda_n(\theta)| \geq \frac{a_n}{2}\right] < \epsilon.\end{aligned}$$

It indicates that for every $a_n > 0$ the quantity

$$\sup_k \Pr\left[n^{-1}\log n\tilde{p}(\theta, X_k)\right] \geq a_n$$

tends to zero as n goes to infinity. Thus the random variable $\log n\tilde{p}(\theta, X_i)$ has bounded variance. □

Poisson Approximation for Arbitrary Infinitely Divisible Families

Let $\phi(t)$ and $\phi_n(t)$ be the characteristic functions of distributions in $\mathcal{E}$ and $\mathcal{E}_n$. By the infinitely divisible property, $\phi(t) = [\phi_n(t)]^n$

or $\phi_n(t) = [\phi(t)]^{1/n}$. Two characteristic functions have the following relation:

$$n(\phi_n(t) - 1) = n(\sqrt[n]{\phi(t)} - 1) = n\left(e^{\frac{1}{n}\log\phi(t)} - 1\right)$$
$$= n\left(1 + \frac{1}{n}\log\phi(t) + o(\frac{1}{n}) - 1\right) \to \log\phi(t),$$

or say $\exp(n(\phi_n(t) - 1)) \to \phi(t)$. The concrete construction of characteristic function in $\mathcal{E}_{\theta,n}$ depends on the discrete Fourier transform of $\Lambda(X,\theta)$ on j segments e.g. $\inf\Lambda(X) < c_1 < c_2 < \cdots < c_j < \sup\Lambda(X)$ which implies that

$$\lim_{j\to\infty}\sum_{k=1}^{j} a_k(i)e^{itc_k} = \int e^{it\Lambda(X)}dF_n = \phi_n(t),$$

where $a_n(k) = n(F_n(c_k) - F_n(c_{k-1}))$ is the Fourier coefficient[10] and F_n is the measure for $\Lambda_n(\theta)$. Combined with the expression above, one can see that a characteristic function of finite many number of Poisson measures (compound Poisson measures) approximates $\phi(t)$:

$$\exp\sum_{i=1}^{j}(na_i)\left(e^{it\Lambda(x_i,\theta)} - 1\right) \to \phi(t) \tag{2.17}$$

where $j \to \infty$ and $\{na_i\}_{i=1,\ldots,j}$ converges to a measure. To see the argument of (2.17), let $V(\cdot)$ be a Poisson process (a random measure) with Poisson parameter γ such that $\mathbb{E}V(\mathcal{A}) = \gamma(\mathcal{A})$ for a set $\mathcal{A}$. For any function v in infinite divisible family, the characteristic function of v is $\phi(t) = \exp\{\int(e^{itv} - 1)d\gamma\}$.

The approximation can be viewed as constructing a new family which approximately equals the infinite divisible $\mathcal{E}_\theta$. Firstly select a Poisson variable ν (again a random measure) such that $\mathbb{E}\nu(\Lambda(X)) = 1$ for any log-likelihood ratio $\Lambda(X)$ and then carry out n-draws from the direct product $\otimes_{i=1,\ldots,\nu}\mathcal{E}_{\theta,i}$, ν copies $\mathcal{E}_{\theta,i}$. The result is called a poissonized family.

Proof of Proposition 2.12

The proof is based on Taylor expansions. Note that

$$m(x,\theta_0 + \delta_n\tau) = m(x,\theta_0) + \delta_n\frac{\partial m(x,\theta_0)}{\partial\theta^T}\tau + o_p(\delta_n^2). \tag{2.18}$$

10 The Stieltjes sum, a discrete version of stochastic integral.

Let $\theta \in \{\theta || \theta - \theta_0| \leq |\tau|\delta_n\}$, $|\tau|$ is a vector with elements equal to their absolute values. 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:

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

It implies:

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

It follows from the approximation of λ above. Using equation (2.18) we can further simplify the terms involving $\theta+\delta_n\tau$. We obtain for the middle term:

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

With the big bracket becoming

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

where $O(n^{-1/2}(\log\log n)^{1/2})$ is used to bound the difference of the sample average and the expectation of a random vector. Thus the local EL is

$$2\sum_{i=1}^{n}\log\frac{\tilde{p}_{\theta_0+\delta_n\tau_n}}{\tilde{p}_{\theta_0}}(x_i)=\delta_n\tau_n^T A_1++\frac{1}{2}\delta_n^2\tau_n^T A_2\tau_n^T+o_p(1)$$

where A_1 is $\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\sum_{i=1}^{n}m_i(\theta_0)$ and A_2 is $\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta}^T\left(\mathbb{E}m(X,\theta_0)m(X,\theta_0)^T\right)^{-1}\mathbb{E}\frac{\partial m(X,\theta_0)}{\partial\theta^T}$.

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

$$\lim_{n\to\infty}A_n\cdot\sum_{i=1}^{n}[m_i(\theta_0+\delta_n\tau)-m_i(\theta_0)]/n=o_p(1)$$

with $A_n=\sum_{i=1}^{n}m_i(\theta_0)\left(\mathbb{E}m(x,\theta_0)m(x,\theta_0)^T\right)^{-1}$ by the continuity of $m_i(\theta)$.

Proof of Proposition 2.15

To prove K_θ is invertible, we will prove K_θ is almost surely positive definite. Le Cam's first lemma implies that

$$\mathbb{E}\exp\left[\tau^T S_\theta-\frac{1}{2}\tau^T K_\theta\tau\right]=1. \tag{2.19}$$

Because (2.19) holds for all τ, we can use a symmetrized method to simplify (2.19). For a given 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

$$\mathbb{E}[(\cosh\tau^T S_\theta)\exp(-\tau^T K_\theta\tau/2)]=1. \tag{2.20}$$

Assume there is some τ such that $\tau^T K_\theta\tau$ is negative, then

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_{\{\tau^T K_\theta\tau>0\}}(\cosh\tau^T S_\theta)\exp(-\tau^T K_\theta\tau/2)\right] \\ \leq\mathbb{E}\left[(\cosh\tau^T S_\theta)\exp(-\tau^T K_\theta\tau/2)\right]=1\end{aligned} \tag{2.21}$$

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 $(\cosh \tau^T S_\theta) > 1$,

$$\underbrace{\mathbb{E}\left[\mathbb{I}_{\{\tau^T K_\theta \tau > 0\}} (\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau/2)\right]}_{>0} +$$
$$\underbrace{\mathbb{E}\left[\mathbb{I}_{\{\tau^T K_\theta \tau \leq 0\}} (\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau/2)\right]}_{\geq 1}$$
$$= \mathbb{E}\left[(\cosh \tau^T S_\theta) \exp(-\tau^T K_\theta \tau/2)\right] > 1$$

we have a contradiction with equation (2.20) unless the set $\{\tau^T K_\theta \tau \leq 0\}$ is empty. Therefore, K_θ is positive definite and hence invertible.