A Primer on the Large Sample Analysis of Sieve Estimation

Ariel Boyarsky*

January 10, 2021

1 Preface to Notes

An immediate disclaimer, nothing in this note is original. Rather, the purpose of this document is to collect several useful results in the theory of sieve estimators. Hopefully, this note will help future researchers quickly familiarize themselves with the status of the current literature on sieves and learn the techniques needed to make original contributions to the line of study. One can think of this note as a slightly less intensive version of Chen (2007) including some results omitted from that paper due to my own particular interests. Several researchers are to thank for the results exposited in this note. An incomplete list of statisticians and econometricians whose work is represented here is: Chunrong Ai, Timothy Christensen, Ulf Grenander, Zhipeng Liao, Demian Pouzo, Xiaotong Shen, Halbert White, Jeffrey Wooldridge, and especially Xiaohong Chen.

Assumptions abound classical data analysis. Often these assumptions are not driven by subject-matter expertise or theory but are rather made to simplify analysis. To this end semi/nononparametric methods allow researchers to step away from possibly unpalatable assumptions and allow the data to "speak". This way the researcher can be confident in the knowledge that their results are driven by the empirical evidence and not a possibly mistaken assumption. Sieve estimation techniques have been developed to subsume a large class of semi/nonparametric methods, such as neural networks, series estimators, nonparametric MLE, etc. A particular benefit in the economic literature is the ability to estimate structural models with nonparametric parameters. Thus, results are driven by theory and data.

2 General Preliminaries of Sieve Estimation

We begin by defining the sieve extremum estimator. We will restrict ourselves to the simplest case with i.i.d. data given by $Z = (Z_1, \dots, Z_n)$, $Z_i \in \mathbb{R}^{d_z}$, and a regular ($\sqrt{n}$-estimable) functional f. In the most general case, what we call **semi-nonparametric**, we are interested in estimating the parameter $\theta_0 = (\beta_0, h_0) \in B \times \mathcal{H} = \Theta$. Here $B \subset \mathbb{R}^k$ and compact while $\mathcal{H}$ is some infinite-dimensional space. First we need to define a **population criterion**,

$$Q : \Theta \to \mathbb{R}$$

Then we can also define the **empirical criterion,**

$$\hat{Q}_n : \Theta \to \mathbb{R}$$

Since Θ is an infinite dimensional parameter is very difficult to maximize a function over this space. In fact, it is not a $\sqrt{n}$-estimable problem (Shen 1997). Thus, the innovation of sieve estimators is to maximize the criterion over a finite-dimensional approximation of the full parameter space. In this way, the analyst makes a "promise" that she will let the size of the parameter space slowly grow to infinite. The finite

*ariel.boyarsky@yale.edu. This note is purely for educational purposes, all errors are my own. For more sophisticated and thorough coverage see Chen (2007). The main purpose of this note is to present the main ideas of the sieve estimator, present the standard assumptions required, and elucidate how asymptotic and inferential results are proven for newcomers to the literature. This note is a work in progress.

dimensional spaces we maximize over are called sieves, $\Theta_n \subset \Theta_{n+1} \subset \cdots \subset \Theta$, are typically (for nice results) compact, non-decreasing spaces. We also need to assume there exists a projection $\pi_n : \Theta \to \Theta_n$ that lets us project our infinite-dimensional parameters onto a finite-dimensional space. In particular, we want that,

$$d(\theta, \pi_n \theta) \to 0 \text{ as } n \to \infty$$

It is also necessary to define $P_n(\theta)$ as an extension of π_n to the full space. So that $P_n\theta = \pi_n\theta$ when $\theta \in \Theta$.

Finally we can define the **approximate sieve extremum estimate**, $\hat{\theta}_n$, as an approximate maximizer of the empirical criterion,

$$\hat{Q}_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta_n} \hat{Q}_n(\theta) - O_p(\eta_n) \text{ for } \eta_n \to 0 \text{ as } n \to \infty \tag{1}$$

When $\eta_n = 0$ this is the **sieve extremum estimator**.

2.1 Identification

Identification of $\hat{\theta}_n$ requires only mild assumptions. Theorem 2.2 of White and Wooldridge shows that $\hat{\theta}_n$ is well defined and measurable under the following conditions. First let $d(\cdot, \cdot)$ be a distance metric on Θ.

Assumption 1. *(Sieve Identification)*

1. *$\hat{Q}_n(\theta)$ is a measurable function of the data $\{Z_{ij}\}_{ij \in \mathbb{N}} \forall\, \theta \in \Theta_k$.*
2. *The spaces Θ_k are compact under $d(\cdot, \cdot)$.*
3. *$\hat{Q}_n(\theta)$ is upper semi-continuous on the sieve spaces Θ_k under $d(\cdot, \cdot)$.*

The next identification-type result is due to White and Wooldridge (1991)

Theorem 1. *Let $(\Omega, \mathcal{F}, \Pr)$ be a complete probability space and let (Θ, d) be a metric space. Let $\hat{Q}_n : \otimes \times \times_n \to \mathbb{R}$ be $\mathcal{F} \times \mathcal{B}(\Theta_n)$-measurable. Then under Assumption 1 there exists a $\mathcal{F} \times \mathcal{B}(\Theta_n)$-measurable sequence $\hat{\theta}_n : \Omega \to \Theta_n$ such that for all $\omega \in \Omega$ we have,*

$$\hat{Q}_n(\omega, \hat{\theta}_n(\omega)) = \inf_{\theta \in \Theta_n} \hat{Q}_n(\omega, \theta)$$

Proof. The proof is due to White and Wooldridge (1991) but relies on an earlier result of Debreu (1967). To see this result denote $v_n(\omega) = \inf_{\theta \in \Theta_n} \hat{Q}_n(\omega, \theta)$ and consider the set $\left\{(w,t) \in \Omega \times \Theta_n | \hat{Q}_n(w,t) = v_n(w)\right\}$ is $\mathcal{F} \times \mathcal{B}(\Theta_n)$-measurable. Furthermore, $\hat{\theta}_n$ exists due to the Kuratowski and Ryll-Nardzewski measurable selection theorem (introduced below) that there exists some $\hat{\theta}_n \in v_n(\omega)$ that returns a single value and is $\mathcal{B}(\Theta_n)$-measurable. □

Definition 1. Suppose $F : X \to 2^Y$ be some mapping. A **selection** is some mapping $f : X \to Y$ such that $f(x) \in F(x)$ for all $x \in X$. In particular, for a given input x $f(x)$ returns only a single value of $F(x)$. This is a special case of a *choice function*.

Theorem 2 (Kuratowski and Ryll-Nardzewski Measurable Selection Theorem (Theorem 6.9.3, Bogachev, Measure Theory))**.** *Suppose X is a Banach space and that $\mathcal{B}(X)$ is the Borel σ-algebra on X. Let $(\Omega, \mathcal{B}, P)$ be a measure space and suppose there exists $F : \Omega \to X$. Furthermore suppose that for any open set $U \subset X$ we have,*

$$\{\omega : F(\omega) \cap U \neq \varnothing\} \in \mathcal{B}$$

Then F has a selection f that is measurable with respect to $\mathcal{B} \times \mathcal{B}(X)$.

3 Sieve M-Estimation

We say that an estimator is **Sieve M-Estimation** if the empirical criterion is a sample average. That is,,

$$\hat{Q}_n(\theta, Z) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i)$$

l is a sieve likelihood or a single observation criterion. Notice that this may differ from the typical likelihood function in that it need not define a density.

Example 1 (Partially Linear Model)**.** Consider the partially linear model,

$$y = h_0(x_1) + x_2^T\beta + \epsilon$$

where $\mathbb{E}[y - \left(h_0(x_1) + x_2^T\beta\right)|x, y] = \mathbb{E}\left[\epsilon|x, y\right] = 0$. Then the empirical criterion is given by,

$$\sup_{\theta\in\Theta_n} \hat{Q}_n = \sup_{\beta\in B,\, h\in\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^{n}\left[y_i - h(x_1) - x_2^T\beta\right]$$

3.1 The Sieve M-Estimator

The estimator is given by,

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} \log l(\theta, Z_i)$$

Recall that $\Theta_n = B \times \mathcal{H}_{k(n)}$. We define $\mathcal{H}_{k(n)}$ as a linear sieve space spanned by a $k(n)$-dimensional basis,

$$\mathcal{H}_{k(n)} = \left\{ h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{k(n)} \alpha_k p_k(\cdot) = \alpha' p^{k(n)}(\cdot) \right\}$$

where $\{p_k\}_{k=1}^{k(n)}$ are a sequence of known basis functions. These could be wavelets, splines, Fourier series, polynomials, etc. Furthermore, we let $k(n) \to \infty$ as $n \to \infty$.

3.2 Large Sample Properties

To begin the discussion of asymptotic properties we need to define several more concepts. First we write the Kullback-Leibler information as,

$$K(\theta_0, \theta) = \mathbb{E}_0\left[l(\theta_0, Z) - l(\theta, Z)\right]$$

Also we define the empirical process indexed by g as,

$$\mathbb{G}_n(g) = \frac{1}{n}\sum_{i=1}^{n}\left(g\left(Z_i\right) - \mathbb{E}_0\left[g(Z_i)\right]\right)$$

Next we need to characterize the score-process of the estimator. To this end we need to define the pathwise derivative of the likelihood. Parameterize the path from θ_0 to θ by $\theta(\theta_0, t) \in \Theta$ for $t \in [0, 1]$ such that $\theta(\theta_0, 0) = \theta_0$ and $\theta(\theta_0, 1) = \theta_1$.

Assumption 2 (Pathwise Differentiability)**.** *Suppose that the pathwise derivative of $l(\theta(\theta_0, t), Z)$ exists at $t = 0$. Define this by $l'_{\theta_0}[\theta - \theta_0, Z]$.*

3.2.1 Likelihood and Score Function Constructions

An equivalent way to state this assumption is,

$$\frac{\partial l(\theta_0, Z)}{\partial(\theta)}[\theta - \theta_0] \equiv \lim_{t \to 0} \frac{l(\theta_0 + t[\theta - \theta_0], Z) - l(\theta_0, Z)}{t}$$

Furthermore, we assume that $l'_{\theta_0}[\theta - \theta_0, Z] - \mathbb{E}_0\left[l'_{\theta_0}[\theta - \theta_0, Z]\right]$ is linear in $\theta - \theta_0$. Next suppose that we approximate $l(\theta, Z) - l(\theta_0, Z)$ by it's pathwise derivative $l'_{\theta_0}[\theta - \theta_0, Z]$. The remainder of this approximation is given by,

$$r[\theta - \theta_0, Z] = l(\theta, Z) - l(\theta_0, Z) - l'_{\theta_0}[\theta - \theta_0, Z]$$

Next define,

$$V_{\theta_0} = \text{clspan}\left(\{\theta - \theta_0 : \theta \in \Theta\}\right)$$

Here we denotes *clspan* as the completion of the linear span. We choose the inner product on this space $\langle \cdot, \cdot \rangle$ such that $Var_0\left(l'_{\theta_0}[\cdot, Z]\right) = ||\cdot||^2$. An example of this is the Fisher inner product. The norm for this inner product is,

$$||\theta - \theta_0||^2 = \mathbb{E}\left[\frac{\partial l(\theta_0, Z)}{\partial \theta}[\theta - \theta_0]\right]^2$$

Note that this implies $||\theta||$ is the Fisher information at θ

3.2.2 Construction of Functional of Interest

Recall that we are interested in a functional of the parameter, $f(\theta)$. Since we intend to conduct inference on this functional we need to apply certain regularity conditions to this functional as well.

Similar to the score function we assume that, $f'_i[\theta - \theta_0]$ is linear in $\theta - \theta_0$ and that

$$||f'_i|| = \sup_{\{\theta \in \Theta : ||\theta - \theta_0|| > 0\}} \frac{|f'_i[\theta - \theta_0]|}{||\theta - \theta_0||} < \infty$$

Then, it is necessary to assume that the functional of interest converges.

Assumption 3 (Convergent Functional)**.**

$$|f_i(\theta) - f_i(\theta_0) - f'_i[\theta - \theta_0]| \leq a_n ||\theta - \theta_0||^{\omega}$$

as $||\theta - \theta_0|| \to 0$.

3.2.3 Sieve Riesz Representation

First let us recall the Riesz Representation Theorem for Hilbert Spaces.

Theorem 3. *Let T be a bounded linear functional on a Hilbert space, H, then there exists some $g \in H$ such that for any $f \in H$ we have,*

1. *Representation:*
$$T(f) = \langle f, g \rangle$$
2. *Operator Norm:*
$$||T||_{op} = \sup_{v \in H, v \neq 0} \frac{||Tv||}{||v||} = \sup_{v \in H, v \neq 0} \frac{||gv||}{||v||} = ||g||_{op}$$

With this theorem established we can apply the Riesz representation theorem for Hilbert spaces to find a sieve representor $v^* \in V_{\theta_0}$such that $v^* = (v_1^*, \ldots, v_n^*)^T$,

$$\langle \theta - \theta_0, v_i^* \rangle = f'_i[\theta - \theta_0] \tag{2}$$

3.2.4 Conditions for Asymptotic Consistency

In this section we present the conditions necessary for consistency of the Sieve extremum estimator allowing for a noncompact infinite-dimensional parameter space for both well-posed and ill-posed problems. The result is due to Chen (2007). Begin by defining a psuedo metric on Θ denoted by $d(\cdot,\cdot)$.

Example 2 (Semi-nonparametric Problem)**.** Let $\Theta = B \times \mathcal{H}$ and define,

$$d(\theta,\theta') = ||\beta - \beta'||_{l^2} + ||h - h'||_{L^2}$$

Here the norm on the finite-dimensional parameter is simply the euclidean norm and we use it's generalization to the Lebsegue integral for the functional space, $||h||_{L^2} = \left(\int |h|^2\right)^{1/2}$.

We then need the following conditions. For clarity I repeat some of the assumptions already made above. In particular, Assumption 1 will be restated so that it is clearer how each condition is required.

Condition 1 (Identification)**.** 1. $Q(\theta_0) > -\infty$ and if $Q(\theta_0) = \infty$ then $Q(\theta) < \infty$ for all other $\theta \in \Theta_n \backslash \{\theta_0\}$ for $n \geq 1$.

2. There is a nonincreasing positive function δ and a positive function g such that for all $\epsilon > 0$ and all $n \geq 1$ we have,
$$Q(\theta_0) - \sup_{\{\theta \in \Theta_n : d(\theta,\theta_0) \geq \epsilon\}} Q(\theta) \geq \delta(k) g(\epsilon) > 0$$

Condition 2 (Sieve spaces)**.** 1. $\Theta_n \subset \Theta_{n+1} \subset \cdots \subset \Theta_{k(n)} \subset \Theta$ for all $n \geq 1$.

2. Recall the projection, $\pi_n \theta_0 \in \Theta_k$. Suppose that $d(\theta_0, \pi_n \theta_0) \to 0$ as $n \to \infty$.

Condition 3 (Continuity)**.** 1. Suppose that for $n \geq 1$,$Q(\theta)$ is upper semi-continuous (e.s.c.) on Θ_k under the metric d. Recall that this means that for all $\varepsilon > 0$ and some $\theta \in \Theta_n$ there exists $\delta > 0$ and $\theta' \in B_\delta(\theta)$ such that $Q(\theta') < Q(\theta) + \varepsilon$.

2. $\left|Q(\theta_0) - Q(\pi_{k(n)}\theta_0)\right| = o(\delta(k(n))$

Condition 4 (Compactness)**.** The sieve spaces, Θ_n, are compact under d.

Condition 5 (Uniform Convergence)**.** 1. For all $n \geq 1$ we have $\text{plim}_{n\to\infty} \sup_{\theta \in \Theta_n} \left|\hat{Q}_n(\theta) - Q(\theta)\right| = 0$

2. $\sup_{\theta \in \Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right| = o_p(\delta(k(n))$

3. $\eta_{k(n)} = o(\delta(k(n))$

Remark 1. Condition 1 is straightforward it is needed so that the true parameter maximizes the criterion and is identifiable. Condition 2 defines the sieve spaces and ensure that as we increase the size of the spaces the projection of the true parameter becomes arbitrarily close to the true parameter. Condition 3 is also needed so that we may identify the true parameter through maximizing the criterion. It further ensures that the criterion is continuous about the project of the true parameter as it grows closer to the actual true parameter. Condition 4 is a relaxation of the typical M-estimation assumption that the parameter space be compact. Instead, we maximize over compact sieve spaces that may be inside a non-compact parameter space. Condition 5 is a uniform law of large numbers, ensuring that the empirical criterion will converge uniformly to the true criterion in large samples. Furthermore, it controls the rate at which the two coincide.

Theorem 4. *Suppose Conditions 1-5 hold. Let $\hat{\theta}_n$ be the approximate sieve extremum estimator. Then $d(\hat{\theta}_n, \theta) \xrightarrow{p} 0$.*

Proof. This proof is due to Chen (2007). We have previously seen that $\hat{\theta}_n$ is well-defined and measurable. For all $\epsilon > 0$ we know that $\sup_{\left\{\theta \in \Theta_{k(n)} : d(\theta,\theta_n) \geq \epsilon\right\}} Q(\theta)$ exists from the compactness of the sieve spaces and

the upper semi-continuity of the criterion. Furthermore by the definition of the sieve extremum estimator we know,

$$\begin{aligned}\Pr[d(\hat{\theta}_n, \theta_0) > \epsilon] &\leq \Pr[\sup_{\{\theta \in \Theta_{k(n)} : d(\theta, \theta_n) \geq \epsilon\}} \hat{Q}_n(\theta) \geq \hat{Q}_n(\pi_{k(n)}\theta_0) - O(\eta_{k(n)})] \\ &\leq P_1 + P_2\end{aligned} \tag{3}$$

where

$$\begin{aligned}P_1 &\equiv \Pr\left[\sup_{\theta \in \Theta_{k(n)} : d(\theta, \theta_0) \geq \epsilon} \left|\hat{Q}_n(\theta) - Q(\theta)\right| > \sup_{\theta \in \Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right|\right] \\ &\leq \Pr\left[\sup_{\theta \in \Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right| > \sup_{\theta \in \Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right|\right] \\ &= 0\end{aligned}$$

and

$$P_2 \equiv \Pr\left[\sup_{\theta \in \Theta_{k(n)} : d(\theta, \theta_0) \geq \epsilon} Q(\theta) \geq Q(\pi_{k(n)}\theta_0) - 2 \sup_{\theta \in \Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right| - O(\eta_{k(n)})\right]$$

To see how P_1 and P_2 are defined notice that P_1 is the probability that $\left|\hat{Q}_n(\theta) - Q(\theta)\right|$ are close within an ϵ-ball of θ_0. Since we control the deviation of the empirical criterion from the true criterion by $\sup_{\theta \in \Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right|$ in P_2 we simply take Equation 3 and switch out the empirical criterion for the true criterion. This is done twice and so we adjust for this by $2 \sup_{\theta \in \Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right|$. Then surely the probability that Equation 3 holds is less than or equal to $P_1 + P_2$. Furthermore, notice that the lower bound in P_1 states that the supremum in a smaller space is greater than the supremum in the full space. Furthermore, by Condition 5 we have that $\text{plim}_{n \to \infty} \sup_{\theta \in \Theta_n} \left|\hat{Q}_n(\theta) - Q(\theta)\right| = 0$ and so $P_1 = 0$ as shown. Now let us consider P_2 after rearranging,

$$\begin{aligned}P_2 &= \Pr\left[2 \sup_{\theta \in \Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right| + Q(\theta_0) - Q(\pi_{k(n)}\theta_0) + O(\eta_{k(n)}) \geq Q(\theta_0) - \sup_{\theta \in \Theta_{k(n)} : d(\theta, \theta_0) \geq \epsilon} Q(\theta)\right] \\ &\geq \delta(k(n)) g(\epsilon) \to 0\end{aligned}$$

The inequality follows from Condition 1that $Q(\theta_0) - \sup_{\theta \in \Theta_{k(n)} : d(\theta, \theta_0) \geq \epsilon} Q(\theta) \geq \delta(k) g(\epsilon)$. Which must go to 0 since the left hand side is bounded from above by Condition 3 so that $Q(\theta_0) - Q(\pi_{k(n)}\theta_0) = o_p(\delta(k(n)))$ and Condition 5 so that $\sup_{\theta \in \Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right| = o_p(\delta(k(n)))$. □

3.2.5 Conditions for Asymptotic Normality

With the main mechanisms established we will now present the conditions necessary to establish asymptotic normality of the Sieve M-Estimator. For clarity, I will restate some of the assumptions already used above. In this section, I call these assumptions conditions to differentiate them from previously discussed ideas. These conditions are due to Shen (1997), Chen and Shen (1998), Shen et al. (2005), and Chen (2007).

Condition 6 (Convergent Functional)**.**

$$|f_i(\theta) - f_i(\theta_0) - f_i'[\theta - \theta_0]| \leq a_n ||\theta - \theta_0||^{\omega}$$

as $||\theta - \theta_0|| \to 0$.

Condition 7 (Bounded Functional)**.**

$$||f'|| < \infty$$

Condition 8 (Projection to Sieve Space)**.** Suppose there exists $\pi_n v^* \in \Theta_n$ such that,

$$||\pi_n v^* - v^*|| \times \left|\left|\hat{\theta}_n - \theta_0\right|\right| = o_p(n^{-1/2})$$

Condition 9 (Stochastic Equicontinuity)**.**

$$\sup_{\{\theta \in \Theta_n : ||\theta - \theta_0|| \leq \delta_n\}} \mathbb{G}_n \left(r(\theta - \theta_0, Z) - r[\pi_n \theta(\theta, \epsilon_n) - \theta_0, Z] \right) = O_p(\epsilon_n^2)$$

Condition 10 (KL Convergence)**.**

$$\sup_{\{\theta \in \Theta_n : ||\theta - \theta_0|| \leq \delta_n\}} [K(\theta_0, \pi_n \theta(\theta, \epsilon_n)) - K(\theta_0, \theta)] \\ -\frac{1}{2} \left[||\theta(\theta, \epsilon_n) - \theta_0||^2 - ||\theta_0 - \theta||^2 \right] = O(\epsilon_n^2)$$

Condition 11 (Projection Approximation Error)**.**

$$\sup_{\{\theta \in \Theta_n : ||\theta - \theta_0|| \leq \delta_n\}} ||\theta(\theta, \epsilon_n) - \pi_n \theta(\theta, \epsilon_n)|| = O(\epsilon_n \delta_n^{-1})$$

and

$$\sup_{\{\theta \in \Theta_n : ||\theta - \theta_0|| \leq \delta_n\}} \mathbb{G}_n(l'_{\theta_0}[\theta(\theta, \epsilon_n) - \pi_n(\theta(\theta, \epsilon_n), Z] = O_p(\epsilon_n^2)$$

Condition 12 (Vanishing Gradient)**.**

$$\sup_{\{\theta \in \Theta_n : ||\theta - \theta_0|| \leq \delta_n\}} \mathbb{G}_n(l'_{\theta_0}[\theta - \theta_0, Z]) = O_p(\epsilon_n)$$

Condition 13 (Gaussian Score Process)**.**

$$\sqrt{n} \mathbb{G}_n(l'_{\theta_0}[v^*, Z]) \to \mathcal{N}(0, \sigma_{v^*}^2)$$

with $\sigma_{v^*}^2 = \lim_{n \to \infty} \frac{1}{n} Var_0 \left(\sum_{i=1}^n l'_{\theta_0}[v^*, Z_i] \right) > 0$.

Remark 2. Condition 9 implies that remainder from the linear approximation of the criterion by it's derivative satisfies stochastic equicontinuity. Condition 10 requires that the Kullback-Leibler divergence is equivalent to $||\cdot||^2$. This controls the quadratic behavior of the criterion. Condition 11 ensures that θ_0 is in the interior of Θ. Condition 13 is automatically implied when Z is i.i.d. However, when data is not i.i.d. it allows us to establish asymptotic normality of the Sieve M-estimator.

Theorem 5. *Under Conditions 6-13 we have that when* $\left|\left|\hat{\theta}_n - \theta_0\right|\right|^\omega = o_p(\frac{1}{\sqrt{n}})$. *Then,*

$$\sqrt{n} \left(f(\hat{\theta}_n) - f(\theta_0) \right) \to \mathcal{N}(0, \sigma_{v^*}^2)$$

with $\sigma_{v^*}^2 = \lim_{n \to \infty} \frac{1}{n} Var_0 \left(\sum_{i=1}^n l'_{\theta_0}[v^*, Z_i] \right) > 0$.

The idea of this proof follows Shen (1997), an extension to the weakly dependent case can be found in Chen and Shen (1998). We will strive to relate $Q(\hat{\theta}_n) - Q(\theta_0)$ to $f(\hat{\theta}_n) - f(\theta_0)$. Using the sieve Riesz representor we can approximate $f(\hat{\theta}_n) - f(\theta_0)$ by $\left\langle \hat{\theta}_n - \theta_0, v^* \right\rangle$ which itself can be approximated by $\frac{1}{n} \sum_{i=1}^n l'_{\theta_0}[v^*, Z_i]$. Then the result will follow by a typical central limit theorem.

Proof. This proof is originally due to Shen (1997) Theorem 1. We assume the data is i.i.d. and so condition 13 is unneeded.

Step 1: Control the Criterion Loss

Take $\pi_n \theta_n \in \{\pi_n \theta_n \in \Theta_n : ||\pi_n \theta_n - \theta_0|| \leq \delta_n\}$. Then begin with,

$$r[\pi_n \theta_n - \theta_0, Z] = l(\pi_n \theta_n, Z) - l(\theta_0, Z) - l'_{\theta_0}[\pi_n \theta_n - \theta_0, Z]$$

Summing yields,

$$n^{-1}\sum_{i=1}^{n} r[\pi_n\theta_n - \theta_0, Z] = \hat{Q}_n(\pi_n\theta_n) - \hat{Q}_n(\theta_0) - n^{-1}\sum_{i=1}^{n} l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z_i]$$

Rearranging,

$$\hat{Q}_n(\pi_n\theta_n) = \hat{Q}_n(\theta_0) + n^{-1}\sum_{i=1}^{n} l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z_i] + n^{-1}\sum_{i=1}^{n} r[\pi_n\theta_n - \theta_0, Z]$$

Next we add and subtract the expectation of the remainder and pathwise derivative,

$$\begin{aligned}\hat{Q}_n(\pi_n\theta_n) =& \hat{Q}_n(\theta_0) \\ &+ n^{-1}\sum_{i=1}^{n} l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z_i] - \mathbb{E}_0\left[l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z]\right] \\ &+ n^{-1}\sum_{i=1}^{n} r[\pi_n\theta_n - \theta_0, Z_i] - \mathbb{E}_0\left[r[\pi_n\theta_n - \theta_0, Z]\right] \\ &+ \mathbb{E}_0\left[l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z_i]\right] + \mathbb{E}_0\left[r[\pi_n\theta_n - \theta_0, Z]\right]\end{aligned}$$

Using the notation for empirical processes we can rewrite this as,

$$\begin{aligned}\hat{Q}_n(\pi_n\theta_n) =& \hat{Q}_n(\theta_0) \\ &+ \mathbb{G}_n(l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z]) \\ &+ \mathbb{G}_n(r[\pi_n\theta_n - \theta_0, Z]) \\ &+ \mathbb{E}_0\left[l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z_i]\right] + \mathbb{E}_0\left[r[\pi_n\theta_n - \theta_0, Z]\right]\end{aligned}$$

Expanding the remainder yields,

$$\begin{aligned}\hat{Q}_n(\pi_n\theta_n) =& \hat{Q}_n(\theta_0) + \mathbb{G}_n(l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z]) + \mathbb{G}_n(r[\pi_n\theta_n - \theta_0, Z]) \\ &+ \mathbb{E}_0\left[l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z_i]\right] - \mathbb{E}_0\left[l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z_i]\right] \\ &+ \mathbb{E}_0\left[l(\pi_n\theta_n, Z) - l(\theta_0, Z)\right]\end{aligned}$$

Noting that the last term is just the Kullback-Leibler information we see this yields Equation (9.2) in Shen (1997),

$$\begin{aligned}\hat{Q}_n(\pi_n\theta_n) =& \hat{Q}_n(\theta_0) - K(\theta_0, \pi_n\theta_n) + \mathbb{G}_n(l'_{\theta_0}[\pi_n\theta_n - \theta_0, Z]) \\ &+ \mathbb{G}_n(r[\pi_n\theta_n - \theta_0, Z])\end{aligned} \tag{4}$$

Then we substitute $\hat{\theta}_n$ for $\pi_n\theta_n$ to get,

$$\begin{aligned}\hat{Q}_n(\hat{\theta}_n) =& \hat{Q}_n(\theta_0) - K(\theta_0, \hat{\theta}_n) + \mathbb{G}_n(l'_{\theta_0}\left[\hat{\theta}_n - \theta_0, Z\right]) \\ &+ \mathbb{G}_n(r[\hat{\theta}_n - \theta_0, Z])\end{aligned} \tag{5}$$

Step 2: Plug in Path to θ_0

We want to consider the difference between the criterion of the estimator and the projection of some θ in the δ-ball around θ_0, $\hat{Q}_n(\hat{\theta}_n) - Q(\pi_n\theta_n)$. So we need t plug in some θ_n such that $||\theta_n - \theta_0|| \leq \delta$ as was our requirement to construct Equation 4 in Step 1. Consider $\theta(\hat{\theta}_n, \epsilon) = (1-\epsilon_n)\hat{\theta}_n + \epsilon_n(v^* + \theta_0)$. Then,

$$\begin{aligned}\left|\left|(1-\epsilon_n)\hat{\theta}_n + \epsilon_n(v^* + \theta_0) - \theta_0\right|\right| &= \left|\left|(1-\epsilon_n)\hat{\theta}_n + \epsilon_n v^* + (1-\epsilon_n)\theta_0\right|\right| \\ &= \left|\left|(1-\epsilon_n)\left(\hat{\theta}_n - \theta_0\right) + \epsilon_n v^*\right|\right| \leq \delta_n\end{aligned}$$

Since as we have seen $d(\hat{\theta}_n, \theta_0) \xrightarrow{p} 0$. Now we subtract the two formulations (note we use linearity in the empirical processes) and substitute in $\theta(\theta, \epsilon_n)$ for θ_n,

$$\begin{aligned}
\hat{Q}_n(\hat{\theta}_n) =& \hat{Q}_n(\pi_n\theta(\theta,\epsilon_n)) - \left[K(\theta_0,\hat{\theta}_n) - K(\theta_0,\pi_n\theta(\theta,\epsilon_n))\right] \\
&+ \mathbb{G}_n(l'_{\theta_0}\left[\hat{\theta}_n - \pi_n\theta(\theta,\epsilon_n), Z\right]) \\
&+ \mathbb{G}_n(r[\hat{\theta}_n - \theta_0, Z] - -r[\pi_n\theta(\theta,\epsilon_n) - \theta_0, Z]) \\
=& \hat{Q}_n(\pi_n\theta(\theta,\epsilon_n)) - \frac{1}{2}\left[\left|\left|\theta_0 - \hat{\theta}_n\right|\right|^2 - ||\pi_n\theta(\theta,\epsilon_n) - \theta_0||^2\right] \\
&+ \mathbb{G}_n(l'_{\theta_0}\left[\hat{\theta}_n - \pi_n\theta(\theta,\epsilon_n), Z\right]) \\
&+ O_p(\epsilon_n^2)
\end{aligned}$$

Where we use Condition 9 to kill the last term and use Condition 10 to substitute out the Kullback-Leibler information criterion loss. Now we notice that from the definition of the estimator,

$$\hat{Q}_n(\hat{\theta}_n) - \hat{Q}_n\left(\pi_n\theta(\theta,\epsilon_n)\right) \geq Q_n(\hat{\theta}_n) - \sup_{\theta\in\Theta_n}\hat{Q}_n(\theta) \geq -O(\epsilon_n^2)$$

Since by Condition 11 we have at most $||\theta(\theta,\epsilon_n) - \pi_n\theta(\theta,\epsilon_n)|| = O(\epsilon_n\delta_n^{-1})$. So we have,

$$\begin{aligned}
-O(\epsilon_n^2) \leq& -\frac{1}{2}\left[\left|\left|\theta_0 - \hat{\theta}_n\right|\right|^2 - ||\pi_n\theta(\theta,\epsilon_n) - \theta_0||^2\right] \\
&+ \mathbb{G}_n(l'_{\theta_0}\left[\hat{\theta}_n - \theta(\hat{\theta}_n,\epsilon_n), Z\right]) + O_p(\epsilon_n^2)
\end{aligned}$$

Plugging in our linear path from earlier,

$$\begin{aligned}
-O(\epsilon_n^2) \leq& -\frac{1}{2}\left[\left|\left|\theta_0 - \hat{\theta}_n\right|\right|^2 - \left|\left|\pi_n\theta(\hat{\theta}_n,\epsilon_n) - \theta_0\right|\right|^2\right] \\
&+ \mathbb{G}_n(l'_{\theta_0}\left[\epsilon_n(v^* - \left(\hat{\theta}_n - \theta_0\right), Z\right]) + O_p(\epsilon_n^2)
\end{aligned}$$

Next we want to expand the projection in the norm and apply the polarization identity. Notice that by the polarization identity,

$$\begin{aligned}
\left|\left|\pi_n\theta(\hat{\theta}_n,\epsilon_n) - \theta_0\right|\right|^2 &= \left|\left|(1-\epsilon)\left(\hat{\theta}_n - \theta_0\right) + \epsilon_n v^*\right|\right|^2 \\
&= \left|\left|(1-\epsilon)\left(\hat{\theta}_n - \theta_0\right)\right|\right|^2 + ||\epsilon_n v^*||^2 + 2\left\langle(1-\epsilon)\left(\hat{\theta}_n - \theta_0\right), \epsilon_n v^*\right\rangle
\end{aligned}$$

So,

$$\begin{aligned}
-O(\epsilon_n^2) \leq& -\frac{1}{2}\left[1 - (1-\epsilon_n)^2\right]\left|\left|\theta_0 - \hat{\theta}_n\right|\right| + \frac{1}{2}||\epsilon_n v^*||^2 \\
&+ (1-\epsilon)\left\langle\left(\hat{\theta}_n - \theta_0\right), \epsilon_n v^*\right\rangle \\
&+ \mathbb{G}_n(l'_{\theta_0}\left[\epsilon_n(v^* - \left(\hat{\theta}_n - \theta_0\right), Z\right]) + O_p(\epsilon_n^2) \\
\leq& -\epsilon_n\left|\left|\theta_0 - \hat{\theta}_n\right|\right| + \frac{1}{2}\epsilon_n^2\left|\left|\theta_0 - \hat{\theta}_n\right|\right| \\
&+ (1-\epsilon)\left\langle\hat{\theta}_n - \theta_0, \epsilon_n v^*\right\rangle \\
&+ \mathbb{G}_n(l'_{\theta_0}\left[\epsilon_n v^*, Z\right]) + O_p(\epsilon_n^2) \\
\leq& (1-\epsilon)\left\langle\hat{\theta}_n - \theta_0, \epsilon_n v^*\right\rangle + \mathbb{G}_n(l'_{\theta_0}\left[\epsilon_n v^*, Z\right]) + O_p(\epsilon_n^2)
\end{aligned}$$

Recall that we define $\epsilon_n = o(n^{-1/2})$. This implies,

$$-(1-\epsilon)\left\langle \hat{\theta}_n - \theta_0, \epsilon_n v^* \right\rangle - \mathbb{G}_n(l'_{\theta_0}[\epsilon_n v^*, Z]) \leq O(\epsilon_n^2) + O_p(\epsilon_n^2) = o_p(n^{-1/2})$$

Thus we can justify the approximation,

$$\begin{aligned}
-(1-\epsilon)\left\langle \hat{\theta}_n - \theta_0, \epsilon_n v^* \right\rangle &= \mathbb{G}_n(l'_{\theta_0}[\epsilon_n v^*, Z]) + o_p(n^{-1/2}) \\
-\left\langle \hat{\theta}_n - \theta_0, \epsilon_n v^* \right\rangle - \mathbb{G}_n(l'_{\theta_0}[\epsilon_n v^*, Z]) &= o_p(n^{-1/2}) \\
\left|\left\langle \hat{\theta}_n - \theta_0, \epsilon_n v^* \right\rangle - \mathbb{G}_n(l'_{\theta_0}[\epsilon_n v^*, Z])\right| &= o_p(n^{-1/2})
\end{aligned}$$

The last line follows because we can always swap v^* for $-v^*$.

Step 3: Asymptotic Distribution of Functional Loss

By Condition 6 we have,

$$\begin{aligned}
\left|f(\hat{\theta}_n) - f(\theta_0) - f'_{\theta_0}\left[\hat{\theta}_n - \theta_0\right]\right| \leq a_n \left|\left|\hat{\theta}_n - \theta_0\right|\right|^{\omega} \\
\iff f(\hat{\theta}_n) - f(\theta_0) = f'_{\theta_0}\left[\hat{\theta}_n - \theta_0\right] + o_p\left(a_n \left|\left|\hat{\theta}_n - \theta_0\right|\right|^{\omega}\right) \\
= \left\langle \hat{\theta}_n - \theta_0, v^* \right\rangle + o_p(n^{-1/2}) \\
= \mathbb{G}_n(l'_{\theta_0}[v^*, Z]) + + o_p(n^{-1/2})
\end{aligned}$$

Where we use the definition of the sieve Riesz representer from equation 2 to get the second equality and we use the approximation from Step 2 for the final equality. The result then follows from classical CLT under i.i.d. data (alternatively in the case of dependent data we require that the data satisfy Condition 13 as is the case with β-mixing data as shown in Chen and Shen (1998)). □

3.3 Sieve Inference via Sieve QLR

In this section, we will describe the sieve quasi likelihood ratio (QLR) statistic as a generalization of the typical likelihood ratio statistic. An alternative test is the sieve Wald (or t) statistic. We test,

$$H_0 : f(\theta_0) = f_0 \longleftrightarrow H_1 : f(\theta_0) \neq f_0$$

Define the **Sieve quasi likelihood ratio** statistic as,

$$QLR_n(f_0) \equiv n\left(\inf_{\theta \in \Theta_{k(n)} : f(\theta) = f_0} \hat{Q}_n(\theta) - \hat{Q}_n(\hat{\theta}_n)\right)$$

3.4 Irregular Functionals

4 Sieve Minimum Distance Estimation