Lec 21: Nonparametric Sieve Estimation

Weiping Zhang

2019.12.16

Introduction

- In this lecture we focus on the sieve estimation of nonparametric regression models. Unlike the kernel estimator which is a local estimator, the sieve estimator is a global estimator in the sense that it estimates the function of interest over its domain in a single step.
- Sieve estimators include the series estimators and spline estimators as special cases.
- We only consider sieve estimators that are linear in their base functions. For an in-depth treatment of general linear or nonlinear sieve estimation, see Chen (2007, 2011).

- In comparison with the kernel method, the sieve method has several advantages.
  - One is its computational easiness. Sieve methods are easily identified by their use of approximating functions such as splines, power or trigonometric series. Since the approximation is of global nature, one can estimate the whole function of interest in a single pass and thus saves on computational time.
  - it is much easier to impose certain structure or restrictions (e.g., additivity) in sieve estimation than in kernel estimation. Sieve methods are particularly well-suited to certain classes of problems.
  - Third, for some econometric problems (e.g., nonparametric regression with endogeneity) sieve methods offer a much easier solution than kernel methods.

Motivation

- To motivate the method of sieves, consider the nonparametric regression

$$Y_i = m(X_i) + u_i, i = 1, \ldots, n. \tag{1}$$

where X_i is a $q \times 1$ random vector, m is a smooth function and u_i is the disturbance term such that $E(u_i|X_i) = 0$ and $E(u_i^2|X_i) = \sigma^2(X_i)$.

- A naive approach to estimating m is to solve the following minimization problem

$$\hat{m}(\cdot) = \arg \min_{m(\cdot) \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - m(X_i)]^2, \tag{2}$$

where $\mathcal{M}$ is afunctional space for $m(\cdot)$ to lie in.

- If $\mathcal{M}$ is very restrictive, e.g., $\mathcal{M} = \{m : m(x) = \beta_0 + \beta' x, (\beta_0, \beta')' \in \mathbb{R}^{q+1}\}$, the esitmator $\hat{m}(x)$ may reduce to parametric estimator of $m(x)$.
- If $\mathcal{M}$ is rich enough, e.g., we do not impose any restrictions on $\mathcal{M}$, then we may obtain an estimator $\hat{m}(\cdot)$ that simply interpolates the data. In the latter case, there is no hope for the estimator to converge to the true regression function. This motivates us to impose some restrictions on the function space $\mathcal{M}$.
- Grenander (1981) suggests that we attempt the above minimization problem within a subspace of the parameter space and then allow the subspace to grow with the sample size n. This sequence of subspaces from which the estimator is obtained is called a "**sieve**", and the resulting procedure is called the method of sieves. The approximating subspaces are called the *sieve spaces*.

- To be specific, we consider the following sieve space

$$\mathcal{M}_n = \{g : g(x) = \sum_{j=1}^{J_n} \beta_j \phi_j(x)\}, \tag{3}$$

  where $\{\phi_j\}$ is a sequence of basis functions and J_n denotes the number of approximating terms in the above sieve space when the sample size is fixed at n. When we allow J_n to grow to infinity together with n, the sieve space $\mathcal{M}$ is expanding so that functions inside it can be used to approximate various smooth functions at certain rate.

- The sieve estimator of $m(\cdot)$ is now given by

$$\hat{m}(\cdot) = \arg \min_{m(\cdot) \in \mathcal{M}_n} \frac{1}{n} \sum_{i=1}^{n} [Y_i - m(X_i)]^2, \tag{4}$$

- Assuming that belongs to $\mathcal{M}$, it can be shown that $\hat{m}(x)$ is consistent with $m(x)$ under two conditions:
  - $\mathcal{M}_n$ grows with n so that it becomes dense in $\mathcal{M}$ as $n \to \infty$,
  - $\mathcal{M}_n$ does not grow too fast.
- In the above example, J_n controls the growth of $\mathcal{M}_n$ so that it has to grow with n but cannot grow too fast.
  Mathematically, a minimum requirement of J_n is that

$$1/J_n + J_n/n \to 0 \quad as \quad n \to \infty.$$

Hölder Class and Linear Sieves

- The most popular class of functions considered in the nonparametric literature is the Hölder smoothness class. Following Chen (2007), suppose that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_q$ is the Cartesian product of the compact intervals $\mathcal{X}_1, \ldots, \mathcal{X}_q$. Let $\gamma \in (0, 1]$. A real-valued function g is said to satisfy a Hölder condition with expoenent γ if there is a positive constant c such that

$$|g(x) - g(\tilde{x})| \leq c\|x - \tilde{x}\|^{\gamma} \ for \ all \ x, \tilde{x} \in \mathcal{X}$$

where $\|x\| = (\sum_{i=1}^{q} x_j^2)^{1/2}$ denotes the Euclidean norm. Given a q-tuple nonnegative integers, $\alpha = (\alpha_1, \ldots, \alpha_q)$, set $|\alpha| = \alpha_1 + \cdots + \alpha_q$ and let D^{α} denote the differential operator defined by

$$D^{\alpha} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_1^{\alpha_q}}.$$

- Let $[p]$ be the largest integer that is strictly less than p so that $p = [p] + \gamma$. A real-valued function g on $\mathcal{X}$ is said to p-smooth if $g(\cdot)$ is $[p]$ times continuously differentiable on $\mathcal{X}$ and $D^{\alpha}g$ satisfies a Hölder condition with exponent γ for all α with $|\alpha| = [p]$.
- Denote the class of all p-smooth real-valued functions on $\mathcal{X}$ by $\Lambda^{p}(\mathcal{X})$ which is called a Hölder Class. Denote the space of all $[p]$-times continuously differentiable functions on $\mathcal{X}$ by $C^{[p]}(\mathcal{X})$. A Hölder ball with smoothness $p = [p] + \gamma$ is defined as

$$C_c^{[p]}(\mathcal{X}) = \Big\{ g \in C^{[p]}(\mathcal{X}) : \sup_{|\alpha| \le [p]} \sup_{x \in \mathcal{X}} |D^{\alpha} g(x)| \le c, \sup_{|\alpha| \le [p]} \sup_{x \in \mathcal{X}} \frac{|D^{\alpha} g(x) - D^{\alpha} g(\tilde{x})|}{\|x - \tilde{x}\|^{\gamma}} \le c \Big\}.$$

- The Hölder or p-smooth class of functions are popular because a p-smooth function can be approximated well by various linear sieves. Essentially, the Hölder class is a class of functions that admit a Taylor expansion with a well behaved remainder term.
- A sieve is called a finite dimensional linear sieve if it is a linear span of finitely many known basis functions. Linear sieves include power series, Fourier series, splines, and wavelets, and they form a large class of sieves useful for sieve extreme estimation.
- Chen (2007, pp. 5570-5572) provides some examples of commonly used linear sieves for univariate functions with support $\mathcal{X} = [0, 1]$. Below are three examples

- **Polynomials**. Let $Pol(J_n)$ denote the space of polynomials on [0, 1] of degree J_n or less; that is,

$$Pol(J_n) = \Big\{ \sum_{j=0}^{J_n} a_j x^j : x \in [0,1], a_j \in \mathbb{R} \Big\}.$$

  $Pol(J_n)$ serves as the simplest linear sieves and the justification for the polynomial basis is the famous Stone-Weierstrass theorem.

- **Trigonometric polynomials**. Let $TriPol(J_n)$ denote the space of trigonometric polynomials on [0, 1] of degree J_n or less; that is,

$$TriPol(J_n) = \Big\{ a_0 + \sum_{j=0}^{J_n} [a_j cos(2j\pi x) + b_j sin(2j\pi x)] : x \in [0,1], a_j, b_j \in \mathbb{R} \Big\}.$$

- Let $CosPol(J_n)$ denote the space of cosine polynomials on [0, 1] of degree J_n or less; that is,

$$CosPol(J_n) = \Big\{ a_0 + \sum_{j=0}^{J_n} a_j cos(2j\pi x) : x \in [0,1], a_j \in \mathbb{R} \Big\}.$$

- Let $SinPol(J_n)$ denote the space of sine polynomials on [0, 1] of degree J_n or less; that is,

$$SinPol(J_n) = \Big\{ a_0 + \sum_{j=0}^{J_n} b_j sin(2j\pi x) : x \in [0,1], b_j \in \mathbb{R} \Big\}.$$

- It is well known that $TriPol(J_n)$ is well suited for approximating periodic functions on [0,1], $CosPol(J_n)$ is well suited for approximating aperiodic functions on [0,1], and $SinPol(J_n)$ can approximate functions vanishing at the boundary points, i.e., $g(0) = g(1) = 0$.

- **Univariate Splines.** Let $t_0, t_1, \ldots, t_{J_n}, t_{J_n+1}$ be real number with $0 = t_0 < t_1 < \cdot < t_{J_n} < t_{J_n+1} = 1$. we can partition [0,1] into $J_n + 1$ subintervals $I_j = [t_j, t_{j+1}), j = 0, 1, \ldots, J_n - 1$ and $I_{J_n} = [t_{J_n}, t_{J_n+1}]$. It is typically assumed that the knots $t_1, \ldots, t_{J_n}$ have bounded mesh ratio:

$$\frac{\max_{0 \le j \le J_n}(t_{j+1} - t_j)}{\min_{0 \le j \le J_n}(t_{j+1} - t_j)} \le c \tag{5}$$

  for some constant $c > 0$. Let $r \ge 1$ be an integer. A function on [0,1] is **spline of order** r or equivalently, **of degree** $m = r - 1$, with knots $t_1, \ldots, t_{J_n}$ if

  - it is a polynomial of degree m or less on each subinterval $I_j, j = 0, 1, \ldots, J_n$; and
  - (for $m \ge 1$) it is $(m-1)$-times continuously differentiable on [0,1].

- Such spline functions constitute a linear space of dimension $J_n + r$. For a fixed integer $r \geq 1$, let $Spl(r, J_n)$ denote the spaces of splines of order r (or degree $r-1$) with J_n knots satisfying (5); that is

$$Spl(r, J_n) = \Big\{ \sum_{k=0}^{r-1} a_k x^k + \sum_{j=0}^{J_n} b_j (x - t_j)_+^{r-1} : x \in [0,1], a_j \in \mathbb{R} \Big\}.$$

- It turns out this linear sieve has very good properties in empirical applications and is usually preferred to the polynomials and trigonometric polynomials.

- The extension from the [0,1] support to an arbitrary compact support on the real line is straightforward.
- In the general case, one uses the tensor-product to generate linear sieves of multivariate functions from linear sieves of univariate functions. To see this, let $\mathbb{G}_l$ be a linear space of functions on $\mathcal{X}_l$ for $1 \leq l \leq q$. The tensor product, $\mathbb{G}$, of $\mathbb{G}_1, \ldots, \mathbb{G}_q$ is defined as the space of functions on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_q$ spanned by the functions $\prod_{l=1}^{q} g_l(x_l)$, where $g_l \in \mathbb{G}_l$ for $1 \leq l \leq q$.

Weighted Smoothness Classes and Linear Sieves

- In many applications, the parameters of interest are functions with unbounded support. In this case, the Hölder Class is not well suited. Chen (2007, pp. 5573-5574) presents two finite-dimensional linear sieves that can approximate functions with unbounded supports well.
- Let $L_p(\mathcal{X}, w), 1 \leq p < \infty$, denote the class of real-valued functions g such that $\int_{\mathcal{X}} |g(x)|^p w(x) dx < \infty$ for a smooth weight function $w : \mathcal{X} \mapsto (0, \infty)$. Below we borrow Chen's examples.
- **Hermite Polynomials** Hermite polynomial series $\{H_k : k = 1, 2, \ldots\}$ is an orthonormal basis of $L_2(\mathcal{X}, w)$ with $w(x) = exp(-x^2/2)$, where $H_1(x) = \pi^{-1/4}$ and for all $k \geq 2$,

$$H_k(x) = (-1)^k exp(-x^2/2) \frac{d^k}{dx^k} [exp(-x^2/2)].$$

Let $HPol(J_n)$ denote the space of Hermite polynomials on $\mathbb{R}$ of degree J_n or less; that is,

$$HPol(J_n) = \Big\{ \sum_{j=1}^{J_n+1} a_j H_j(x) exp(-x^2/2) : x \in \mathbb{R}, a_j \in \mathbb{R} \Big\}.$$

It turns out any function in $L_2(x) = \{g(x) : \int_{\mathcal{X}} g^2(x)dx < \infty\}$ can be approximated well by the linear sieve $HPol(J_n)$ as $J_n \to \infty$.

- Nonlinear sieves can also be used for sieve extreme estimation. A popular class of nonlinear sieves is the single hidden layer feed forward Artificial Neural Networks (ANN). The sieve spaces can also be infinite-dimensional. For interested readers, we refer them directly to Chen (2007) for details.

Sieve estimator

- Consider a nonparametric regression (1), we now discuss how to estimate $m(\cdot)$ using the method of sieves. Letting $p^{K_n}(\cdot)$ denote the $K_n \times 1$ vector of the first K_n approximating functions

$$p^K(x) = (p_1(x), \ldots, p_K(x))'$$

where we suppress the dependence of K_n on n. By the approximation property of the basis functions, there exists a $K \times 1$ vector β such that $m(x) - \beta' p^K(x) \to 0$ for all x. This motivates us to rewrite (1)

$$Y_i = \beta' p^K(X_i) + [m(X_i) - \beta' p^K(X_i) + u_i] = \beta' p^K(X_i) + e_i, \tag{6}$$

where $e_i = m(X_i) - \beta' p^K(X_i) + u_i$ denotes the new error term.

- Then we can estimate the unknown finite dimensional object β by the least squares regression of Y_i on $p^K(X_i)$, i.e.,

$$\hat{\beta} = (P'P)^- P'Y \tag{7}$$

where P denotes the $n \times K$ matrix whose ith row is given by $p^K(X_i)'$, $Y = (Y_1, \ldots, Y_n)'$ and $(\cdot)^-$ denotes the generalized inverse of $(\cdot)$. The estimator of $m(x)$ is given by

$$\hat{m}(x) = \hat{\beta}' p^K(x). \tag{8}$$

- If one is also interested in estimating the derivatives of $m(x)$, one can also obtain their estimates based on $\hat{\beta}$. For example

$$\frac{d\hat{m}(x)}{dx} = \hat{\beta}' \frac{dp^K(x)}{dx}$$

is estimating the first derivative of $m(x)$.

Convergence Rates

- Stone (1985, 1986), Andrews (1991), Newey (1995, 1997), Huang (2001), and de Jong (2002) study the asymptotic results for sieve estimators. Here we follow Newey (1997) and recognize its limitation.
- To establish the consistency and convergence rates for $\hat{m}(x)$, we make the following assumptions.
  (A1) $\{X_i, Y_i\}$ is IID and X_i has compact support $\mathbf{X}$.
  (A2) $Var(u_i|X_i)$ is bounded a.s.
  (A3) (i) Let $\Phi_K = E[p^K(X_i)p^K(X_i)']$.
  $0 < c_1 \leq \lambda_{min}(\Phi_K) \leq \lambda_{max}(\Phi_K) \leq c_2 < \infty$ uniformly in K.
  (ii) There exists $\alpha > 0$ such that
  $\sup_{x \in \mathcal{X}} \|m(x) - \beta' p^K(x)\| = O(K^{-\alpha})$ for some $\beta \in \mathbb{R}^K$. (iii) There exists a sequence of constants $\zeta_0(K)$ such that
  $\sup_{x \in \mathcal{X}} \|p^K(x)\| \leq \zeta_0(K)$.
  (A4) As $n \to \infty, K \to \infty$, and $\zeta_0(K)^2 K/n \to 0$.

Theorem

Under assumptions A1-A4, we have

(1) $\int [\hat{m}(x) - m(x)]^2 dF(x) = O_p(K/n + K^{-2\alpha})$;

(2) $n^{-1} \sum_{i=1}^{n} [\hat{m}(X_i) - m(X_i)]^2 = O_p(K/n + K^{-2\alpha})$;

(3) $\sup_{x \in \mathcal{X}} |\hat{m}(x) - m(x)| = O_p(\zeta_0(K)(\sqrt{K/n} + +K^{-\alpha})$;

where F is the CDF of X_i.

Proof

- Here we sketch the proof of the above theorem. First, we claim that

$$\|\hat{\beta} - \beta\| = O_p(\sqrt{K/n} + K^{-\alpha}) \tag{9}$$

and apply this result to prove (i)-(iii) in the theorem. Then we will prove (9).
To show (i), note that by the Cauchy-Schwarz inequality, (9) and Assumptions A3(i)-(ii)

$$\begin{aligned}
&\int [\hat{m}(x) - m(x)]^2 dF(x) \\
&= \int \{(\hat{\beta} - \beta)' p^K(x) + [\beta' p^K(x) - m(x)]\}^2 dF(x) \\
&\leq 2(\hat{\beta} - \beta)' \int p^K(x) p^K(x)' dF(x) (\hat{\beta} - \beta) \\
&\quad + 2 \int [\beta' p^K(x) - m(x)]^2 dF(x)
\end{aligned}$$

$$\leq 2\|\hat{\beta} - \beta\|^2 \lambda_{max}(E[p^K(X_1)p^K(X_1)'])$$
$$+ 2 \sup_{x \in \mathcal{X}} \|m(x) - \beta' p^K(x)\|^2$$
$$= 2\|\hat{\beta} - \beta\|^2 O(1) + O(K^{-2\alpha}) = O_p(K/n + K^{-2\alpha}).$$

For (ii), letting $\Phi_{Kn} = n^{-1} \sum_{i=1}^n p^K(X_i)p^K(X_i)'$, we have by the Cauchy-Schwarz inequality, (9) and Assumptions A3(i)-(ii),

$$\frac{1}{n} \sum_{i=1}^n [\hat{m}(X_i) - m(X_i)]^2$$
$$= \frac{1}{n} \sum_{i=1}^n \{(\hat{\beta} - \beta)' p^K(X_i) + [\beta' p^K(X_i) - m(X_i)]\}^2$$
$$\leq 2(\hat{\beta} - \beta)' \Phi_{Kn} (\hat{\beta} - \beta) + 2 \frac{1}{n} \sum_{i=1}^n [\beta' p^K(X_i) - m(X_i)]^2$$
$$\leq 2\|\hat{\beta} - \beta\|^2 \lambda_{max}(\Phi_{Kn}) + O(K^{-2\alpha}) = O_p(K/n + K^{-2\alpha}).$$

provided we can show $\lambda_{max}(\Phi_{Kn}) = O_p(1)$. Under Assumptions A1 and A4, we can readily show that $E\|\Phi_{Kn} - \Phi_K\|^2 = O(\zeta_0(K)^2 K/n) = o(1)$ implying that

$$\|\Phi_{Kn} - \Phi_K\| = O_p(\zeta_0(K)\sqrt{K/n}) = o_p(1) \tag{10}$$

by Chebyshev inequality. It follows that from the definition of maximum eigenvalue that

$$\begin{aligned}
\lambda_{max}(\Phi_{Kn}) &= \max_{\|z\|=1} \{z'\Phi_K z + z'(\Phi_{Kn} - \Phi_K)z\} \\
&\leq \lambda_{max}(\Phi_K) + \|\Phi_{Kn} - \Phi_K\| \\
&= \lambda_{max}(\Phi_K) + o_p(1) = O_p(1).
\end{aligned} \tag{11}$$

This proves (ii). For (iii), we have by the triangle inequality, (9), and Assumption A3(iii),

$$\begin{aligned}\sup_{x\in\mathcal{X}}|\hat{m}(x)-m(x)| &= \sup_{x\in\mathcal{X}}|(\hat{\beta}-\beta)'p^K(x)+[\beta'p^K(x)-m(x)]| \\ &\leq \sup_{x\in\mathcal{X}}\|p^K(x)\|\|\hat{\beta}-\beta\|+\sup_{x\in\mathcal{X}}|\beta'p^K(x)-m(x)| \\ &= \zeta_0(K)O_p(\sqrt{K/n}+K^{-1\alpha})+O(K^{-\alpha}) \\ &= O_p[\zeta_0(K)(\sqrt{K/n}+K^{-\alpha})].\end{aligned}$$

We now prove (9). Let $\mathbf{m}=(m(X_1),\ldots,m(X_n))'$ and $U=(u_1,\ldots,u_n)'$. Then

$$\begin{aligned}\hat{\beta}-\beta &= (P'P)^-P'Y-\beta \\ &= (P'P)^-P'[P\beta+U+(\mathbf{m}-P\beta)]-\beta\end{aligned}$$

$$
\begin{aligned}
&= (P'P)^{-}P'U + (P'P)^{-}P'(\mathbf{m} - P\beta) \\
&= \Phi_{Kn}^{-}n^{-1}P'U + \Phi_{Kn}^{-}n^{-1}P'(\mathbf{m} - P\beta) \\
&=: A_{1n} + A_{2n},\ say. \qquad (12)
\end{aligned}
$$

Using the arguments like those used to obtain (11), we can show that

$$\lambda_{min}(\Phi_{Kn}) \geq \lambda_{min}(\Phi_K) - o_p(1) \geq \underline{c}_{\Phi}/2 \qquad (13)$$

with probability approaching 1 (w.p.a.1) as $n \to \infty$ under Assumption A3(i). That is, w.p.a.1, Φ_{Kn} is invertible and we can write Φ_{Kn}^{-} as Φ_{Kn}^{-1}, the usual inverse of Φ_{Kn}. It follows that w.p.a.1,

$$
\begin{aligned}
\|A_{2n}\|^2 &= \|\Phi_{Kn}^{-} n^{-1} P'(\mathbf{m} - P\beta)\|^2 \\
&= n^{-2} tr\{P'(\mathbf{m} - P\beta)(\mathbf{m} - P\beta)' P \Phi_{Kn}^{-1} \Phi_{Kn}^{-1}\} \\
&= n^{-2} tr\{\Phi_{Kn}^{-1/2} P'(\mathbf{m} - P\beta)(\mathbf{m} - P\beta)' P \Phi_{Kn}^{-1/2} \Phi_{Kn}^{-1}\} \\
&\leq \lambda_{max}(\Phi_{Kn}^{-1}) n^{-2} tr\{\Phi_{Kn}^{-1/2} P'(\mathbf{m} - P\beta)(\mathbf{m} - P\beta)' P \Phi_{Kn}^{-1/2}\} \\
&= \lambda_{max}(\Phi_{Kn}^{-1}) n^{-2} tr\{(\mathbf{m} - P\beta)(\mathbf{m} - P\beta)' P (P'P)^{-1} P'\} \\
&\leq [\lambda_{max}(\Phi_{Kn})]^{-1} n^{-2} \|\mathbf{m} - P\beta\|^2 = O_p(K^{-2\alpha})
\end{aligned}
$$

where the first inequality follows from the fact that $tr(AB) \leq \lambda_{max}(B) tr(A)$ for any symmetric matrix A and positive semidefinite matrix B, the last inequality follows from the same fact and the fact that $P(P'P)^{-1}P$ is a projection matrix with maximum eigenvalue 1, and the last equality follows from (13) and the fact that $n^{-1}\|\mathbf{m} - P\beta\|^2 = O_p(K^{-2\alpha})$ by Assumption A3(ii).

Similarly, we can readily show that

$$\begin{aligned}\|A_{1n}\|^2 = \|\Phi_{Kn}^{-1} n^{-1} P'U\|^2 &= n^{-2} tr(P'UU'P\Phi_{Kn}^{-1}\Phi_{Kn}^{-1}) \\ &\leq [\lambda_{min}(\Phi_{Kn})]^{-2} n^{-2} \|P'U\|^2 = O_p(K/n)\end{aligned}$$

where the last equality follows (13) and the fact that $n^{-2}\|P'U\|^2 = O_p(K/n)$ by Markov inequality. Consequently, we have by the Minkowski inequality that

$$\|\hat{\beta} - \beta\| \leq \|A_{1n}\| + \|A_{2n}\| = O_p(\sqrt{K/n} + K^{-\alpha}).$$

This completes the proof of the theorem.

Asymptotic Normality

- To study the asymptotic normality of $\hat{m}(x)$, we add the following assumption
  Assumption A5. (i) $E(u_i^4|X_i)$ is bounded a.s., and $\sigma^2(X_i) = E(u_i^2|X_i) \geq \underline{c}_{\sigma^2} > 0$ a.s. (ii) $nK^{-2\alpha} = o(1)$.
  Note that Assumption A5(i) specifies the moment conditions on the error term and A5(ii) is imposed to ensure that the asymptotic bias term is asymptotically smaller than the asymptotic variance term as there is no way to estimate the asymptotic bias for sieve estimates.

- Using (12), we have

$$\begin{aligned}\hat{m}(x) - m(x) &= (\hat{\beta} - \beta)' p^K(x) + [\beta' p^K(x) - m(x)] \\ &= p^K(x)' \Phi_{Kn}^{-} n^{-1} P' U + p^K(x)' \Phi_{Kn}^{-} n^{-1} P' (\mathbf{m} - P\beta) \\ &\quad + [\beta' p^K(x) - m(x)] \\ &= b_{1n} + b_{2n} + b_{3n}, \ say \end{aligned} \tag{14}$$

We shall show that b_{1n} contributes to the asymptotic variance of ˆ $\hat{m}(x)$ and $b_{2n} + b_{3n}$ contributes to its asymptotic bias, which is asymptotically negligible under Assumption A5(ii). Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\sigma^2(X_i) = Var(u_i|X_i)$. To derive the asymptotic variance, note that

$$\begin{aligned} Var(\sqrt{n} b_{1n}|\mathbf{X}) &= p^K(x)' \Phi_{Kn}^{-} Var(n^{-1/2} P' U|\mathbf{X}) \Phi_{Kn}^{-} p^K(x) \\ &= p^K(x)' \Phi_{Kn}^{-} Var(n^{-1/2} \sum_{i=1}^{n} P^K(X_i) u_i|\mathbf{X}) \Phi_{Kn}^{-} p^K(x) \end{aligned}$$

$$= p^K(x)'\Phi_{Kn}^- n^{-1} \sum_{i=1}^n Var(P^K(X_i)u_i|\mathbf{X})\Phi_{Kn}^- p^K(x)$$
$$= p^K(x)'\Phi_{Kn}^- \Sigma_{Kn}\Phi_{Kn}^- p^K(x)$$

where $\Sigma_{Kn} = n^{-1}\sum_{i=1}^n P^K(X_i)p^K(X_i)\sigma^2(X_i)$. Noting that $E(\sqrt{n}b_{1n}|\mathbf{X}) = 0$, by the variance decomposition formula,

$$Var(\sqrt{n}b_{1n}) = E[Var(\sqrt{n}b_{1n}|\mathbf{X})] + Var(E(\sqrt{n}b_{1n}|\mathbf{X}))$$
$$= p^K(x)'E[\Phi_{Kn}^-\Sigma_{Kn}\Phi_{Kn}^-]p^K(x).$$

The White's heteroskedasticity-robust estimate of $\hat{\Sigma}_{Kn}$ is given by $\hat{\Sigma}_K = n^{-1}\sum_{i=1}^n p^K(X_i)p^K(X_i)'\hat{u}_i^2$ where $\hat{u}_i = Y_i - \hat{\beta}'p^K(X_i)$.

- It follows that we can estimate the asymptotic variance of $\sqrt{n}b_{1n}$ by

$$\hat{V}_K = p^K(x)'\Phi_{Kn}^{-1}\hat{\Sigma}_K\Phi_{Kn}^{-1}p^K(x).$$

  The following theorem establishes the asymptotic normality of $\hat{m}(x)$.

- **Theorem** Suppose Assumptions A1-A5 hold. Assume that $\|p^K(x)\| \geq \underline{c}_p > 0$, then

$$\sqrt{n}\hat{V}_k^{-1/2}[\hat{m}(x) - m(x)] \rightsquigarrow N(0,1).$$

Proof

- Let $V_K = p^K(x)'\Phi_{Kn}^{-1}\Sigma_K\Phi_{Kn}^{-1}p^K(x)$ where $\Sigma_K = E[p^K(X_i)p^K(X_i)'\sigma^2(X_i)]$. By (14) and Slutsky lemma, we can prove the theorem by showing that (i) $V_K \geq c\|p^K(x)\|^2$ for some $c > 0$, (ii) $\hat{V}_K - V_K = o(V_k)$, (iii) $\sqrt{n}V_K^{-1/2}b_{1n} \rightsquigarrow N(0,1)$, (iv) $\sqrt{n}V_K^{-1/2}b_{2n} = o_p(1)$, and (v) $\sqrt{n}V_K^{-1/2}b_{3n} = o_p(1)$. To see this, note that

$$
\begin{aligned}
&\sqrt{n}\hat{V}_K^{-1/2}[\hat{m}(x) - m(x)] \\
&= \sqrt{n}V_K^{-1/2}[\hat{m}(x) - m(x)] + \sqrt{n}(\hat{V}_K^{-1/2} - V_K^{-1/2})[\hat{m}(x) - m(x)] \\
&= \sqrt{n}V_K^{-1/2}[\hat{m}(x) - m(x)] \\
&\quad + \sqrt{n}V_K^{-1/2}(\hat{V}_K^{1/2} - V_K^{1/2})V_K^{-1/2}[\hat{m}(x) - m(x)]
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{n} V_K^{-1/2} [\hat{m}(x) - m(x)] \\
&\quad + \sqrt{n} V_K^{-1/2} (\hat{V}_K^{1/2} - V_K^{1/2}) V_K^{-1/2} [\hat{m}(x) - m(x)] \\
&= \sqrt{n} V_K^{-1/2} [\hat{m}(x) - m(x)] \\
&\quad + \frac{\hat{V}_K - V_K}{\hat{V}_K^{1/2} (\hat{V}_K^{1/2} + V_K^{1/2})} \sqrt{n} V_K^{-1/2} [\hat{m}(x) - m(x)]
\end{aligned}
$$

The first term is asymptotically $N(0,1)$ given (iii)-(v) and the second term is $o_p(1)$ given (ii)-(v). The proof of (ii) is given in Lemma 1 below. We now prove the other parts in turn.

For (i), noting that $\lambda_{min}(\Sigma_K) \geq \underline{c}_{\sigma^2}\lambda_{min}(\Phi_K) \geq \underline{c}_{\sigma^2}\underline{c}_{\Phi}$ we have

$$\begin{aligned} V_K &= p^K(x)'\Phi_K^{-1}\Sigma_K\Phi_K^{-1}p^K(x) \\ &\geq \lambda_{min}(\Sigma_K)p^K(x)'\Phi_K^{-1}\Phi_K^{-1}p^K(x) \\ &\geq \lambda_{min}(\Sigma_K)[\lambda_{max}(\Phi_K)]^{-2}p^K(x)'p^K(x) \\ &\geq \underline{c}_{\sigma^2}\underline{c}_{\Phi}\bar{c}_{\Phi}^2 > 0. \end{aligned}$$

So (i) follows by choosing $c = \underline{c}_{\sigma^2}\underline{c}_{\Phi}\bar{c}_{\Phi}^2 > 0$.
We now prove (iii),

$$\begin{aligned} \sqrt{n}V_K^{-1/2}b_{1n} &= \sqrt{n}V_K^{-1/2}p^K(x)'\Phi_{Kn}^{-}n^{-1}P'U \\ &= \sqrt{n}V_K^{-1/2}p^K(x)'\Phi_K^{-}n^{-1}P'U \\ &\quad + \sqrt{n}V_K^{-1/2}p^K(x)'[\Phi_{Kn}^{-} - \Phi_K^{-}]n^{-1}P'U \\ &= B_{1n,1} + B_{1n,2}, \quad say. \end{aligned}$$

For $B_{1n,1}$ we can readily apply the Liapounov CLT and demonstrate that $B_{1n,1} \rightsquigarrow N(0,1)$. For $B_{1n,2}$,

$$
\begin{aligned}
E(B_{1n,2}^2|\mathbf{X}) &= n^{-1}V_K^{-1}p^K(x)'[\Phi_{Kn}^- - \Phi_K^-] \\
&\quad \cdot E[P'UU'P|\mathbf{X}][\Phi_{Kn}^- - \Phi_K^-]p^K(x) \\
&\leq \lambda_{max}(n^{-1}E[P'UU'P|\mathbf{X}])V_K^{-1}p^K(x)'[\Phi_{Kn}^- - \Phi_K^-][\Phi_{Kn}^- - \Phi_K^-]p^K(x) \\
&\leq \lambda_{max}(n^{-1}E[P'UU'P|\mathbf{X}])\{V_K^{-1}\|p^K(x)\|^2\}\|\Phi_{Kn}^- - \Phi_K^-\|^2 \\
&= O_p(1)O_p(1)o_p(1) = o_p(1)
\end{aligned}
$$

where we use the fact that $n^{-1}E[P'UU'P|\mathbf{X}] = n^{-1}\sum_{i=1}^n p^K(X_i)p^K(X_i)\sigma^2(X_i)$ has bounded maximum eigenvalue under Assumptions A3(i) and A5. Hence $B_{1n,2} = o_p(1)$ by the conditional Chebyshev inequality. Thus (iii) follows.

For (iv), we have by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\sqrt{n}V_K^{-1/2}|b_{2n}| = n^{-1/2}V_K^{-1/2}|p^K(x)'\Phi_{Kn}^{-}P'(\mathbf{m}-P\beta)| \\
&\leq V_K^{-1/2}\{p^K(x)'\Phi_{Kn}^{-}(n^{-1}P'P)\Phi_{Kn}^{-}p^K(x)\}^{1/2}\|\mathbf{m}-P\beta\|^2 \\
&= V_K^{-1/2}\{p^K(x)'\Phi_{Kn}^{-}p^K(x)\}^{1/2}\|\mathbf{m}-P\beta\|^2 \\
&\leq \{V_K^{-1/2}\|p^K(x)\|\}[\lambda_{max}(\Phi_{Kn}^{-})]^{1/2}O_p(n^{1/2}K^{-\alpha}) \\
&= O(1)O_p(1)o_p(1) = o_p(1).
\end{aligned}$$

By the same token we can show $\sqrt{n}V_K^{-1/2}|b_{3n}| = o_p(1)$. This completes the proof of the theorem.

Lemma

Suppose the conditions in the above theorem hold. Then $\hat{V}_k - V_k = o_p(V_K)$.

Proof. For (ii), we have

$$\begin{aligned}
\hat{V}_K - V_K &= p^K(x)'[\Phi_{Kn}^{-1}\hat{\Sigma}_K\Phi_{Kn}^{-1} - \Phi_K^{-1}\Sigma_K\Phi_K^{-1}]p^K(x) \\
&= p^K(x)'(\Phi_{Kn}^{-1} - \Phi_K^{-1})\hat{\Sigma}_K\Phi_{Kn}^{-1}p^K(x) \\
&+ p^K(x)'\Phi_K^{-1}(\hat{\Sigma}_K - \Sigma_K)\Phi_{Kn}^{-1}p^K(x) \\
&+ p^K(x)'\Phi_K^{-1}\Sigma_K(\Phi_{Kn}^{-1} - \Phi_K^{-1})p^K(x) \\
&= C_{1n} + C_{2n} + C_{3n},\ say
\end{aligned}$$

Noting that $\|\Phi_{Kn} - \Phi_K\| = O_p(\zeta_0(K)\sqrt{K/n})$, we have

$$\|\Phi_{Kn}^{-1} - \Phi_K^{-1}\|^2 = \|\Phi_{Kn}^{-1}(\Phi_{Kn} - \Phi_K)\Phi_K^{-1}\|^2$$

$$
\begin{aligned}
&= tr(\Phi_{Kn}^{-1}(\Phi_{Kn} - \Phi_K)\Phi_K^{-1}\Phi_K^{-1}(\Phi_{Kn} - \Phi_K)\Phi_{Kn}^{-1}) \\
&\leq \lambda_{max}(\Phi_K^{-1}\Phi_K^{-1})tr(\Phi_{Kn}^{-1}(\Phi_{Kn} - \Phi_K)\Phi_{Kn}^{-1}) \\
&\leq \lambda_{max}(\Phi_K^{-1}\Phi_K^{-1})\lambda_{max}(\Phi_{Kn}^{-1}\Phi_{Kn}^{-1})\|\Phi_{Kn} - \Phi_K\|^2 \\
&= O(1)O_p(1)O_p(\zeta_0(K)^2 K/n).
\end{aligned}
$$

That is, $\|\Phi_{Kn} - \Phi_K\| = O_p(\zeta_0(K)\sqrt{K/n}) = o_p(1)$. By the Cauchy-Schwarz inequality and the fact that $\|\Phi_{Kn} - \Phi_K\| = o_p(1)$,

$$
\begin{aligned}
|C_{1n}| &= |p^K(x)'(\Phi_{Kn}^{-1} - \Phi_K^{-1})\hat{\Sigma}_K\Phi_{Kn}^{-1}p^K(x)| \\
&\leq \{p^K(x)'(\Phi_{Kn}^{-1} - \Phi_K^{-1})(\Phi_{Kn}^{-1} - \Phi_K^{-1})p^K(x)\}^{1/2} \\
&\quad \cdot \{p^K(x)'\Phi_{Kn}^{-1}\hat{\Sigma}_K\hat{\Sigma}_K\Phi_{Kn}^{-1}p^K(x)\}^{1/2} \\
&\leq \|p^K(x)\|^2\|\Phi_{Kn}^{-1} - \Phi_K^{-1}\|\lambda_{max}(\hat{\Sigma}_K)\lambda_{max}(\Phi_{Kn}^{-1}) \\
&= \|p^K(x)\|^2 o_p(1).
\end{aligned}
$$

Similarly, $|C_{3n}| = \|p^K(x)\|^2 o_p(1)$ by (10).
For C_{2n}, we first make the following decomposition

$$\begin{aligned}
\hat{\Sigma}_K - \Sigma_K &= \frac{1}{n}\sum_{i=1}^{n} p^K(X_i)p^K(X_i)'[Y_i - \hat{\beta}' p^K(X_i)]^2 \\
&\quad - E[p^K(X_i)p^K(X_i)'\sigma^2(X_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n} p^K(X_i)p^K(X_i)'u_i^2 - E[p^K(X_i)p^K(X_i)'u_i^2]] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} p^K(X_i)p^K(X_i)'[m(X_i) - \hat{m}(X_i)]^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n} p^K(X_i)p^K(X_i)'u_i[m(X_i) - \hat{m}(X_i)] \\
&= D_{1n} + D_{2n} + D_{3n}, \quad say.
\end{aligned}$$

Then

$$
\begin{aligned}
C_{2n} &= p^K(x)'\Phi_K^{-1}D_{1n}\Phi_{Kn}^{-1}p^K(x) + p^K(x)'\Phi_K^{-1}D_{2n}\Phi_{Kn}^{-1}p^K(x) \\
&\quad + p^K(x)'\Phi_K^{-1}D_{3n}\Phi_{Kn}^{-1}p^K(x) \\
&= C_{2n,1} + C_{2n,2} + C_{2n,3},\ say.
\end{aligned}
$$

Noting that $\|D_{1n}\| = O_p(K/\sqrt{n}) = o_p(1)$ by Chebyshev inequality, we have by Cauchy-Schwarz inequality

$$
\begin{aligned}
|C_{2n,1}| &\leq |p^K(x)'\Phi_K^{-1}D_{1n}\Phi_{Kn}^{-1}p^K(x)| \\
&\leq \{p^K(x)'\Phi_K^{-1}D_{1n}D_{1n}\Phi_K^{-1}p^K(x)\}^{1/2}\{p^K(x)'\Phi_{Kn}^{-1}\Phi_{Kn}^{-1}p^K(x)\}^{1/2} \\
&\leq \|D_{1n}\|\{\lambda_{max}(\Phi_K^{-1})\lambda_{max}(\Phi_{Kn}^{-1})\}\|p^K(x)\|^2 \\
&= o_p(1)O_p(1)O_p(1)\|p^K(x)\|^2 = o_p(1)\|p^K(x)\|^2.
\end{aligned}
$$

Cross-validation

- For each observation i, we can create a leave-one-out prediction error by estimating the coefficients using the observations excluding i. That is, define the leave-one-out es

$$\hat{\beta}_{-i} = (P'_{-i}P_{-i})^{-}P'Y_{-i} \tag{15}$$

  and the prediction error

$$\tilde{e}_i = Y_i - \hat{\beta}'_{-i}P^K(X_i)$$

- Therefore, the cross-validation criterion is

$$CV(K) = \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2.$$

- To implement cross-validation selection, a user first has to select the set of models $K = 1, \ldots, M_n$ over which to search. For example, if using a power series approximation, a user has to first determine the highest power, or if using a spline a user has to determine the order of the spline and the maximum number of knots.
- Unfortunately, there is no precise guidance on how to determine the number of models M_n.
- We introduce some alternative criterion commonly used to model selection.

- **Mallows criterion** (Mallows, 1973):

$$Mallows(K) = \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2 + 2\hat{\sigma}^2 K$$

where $\hat{\sigma}^2$ is preliminary estimator of $E(e_i^2)$. Li (1987) provided conditions under which Mallows selection is asymptotically optimal, but Andrews (1991b) shows that its optimality fails under heteroskedasticity

- **AIC** (Akaike, 1973):

$$AIC(K) = nlog(\sum_{i=1}^{n}\tilde{e}_i^2) + 2K$$

AIC selection has similar asymptotic properties as Mallows selection, in that it is asymptotically optimal under conditional homoskedasticity but not under heteroskedasticity.

- **BIC**(Schwarz, 1978):

$$BIC(K) = nlog(\sum_{i=1}^{n} \tilde{e}_i^2) + log(n)K$$

BIC has the property of consistent model selection: When the true model is a finite dimensional series, BIC will select that model with probability approaching one as the sample size increases. However, when there is no finite dimensional true model, then BIC tends to select overly parsimonious mode.

- **PLS** A different approach to selection is the class of penalized least squares estimators.

$$\hat{\beta}_\lambda = \arg\min_\beta \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \beta' p^K(X_i))^2 + \sum_{j=1}^{K} p_\lambda(\beta_j)$$

where $p_\lambda(u)$ is a non-negative symmetric penalty function and λ is a tuning parameter.

Estimation of Density Functions

- Let $X_1, \ldots, X_n$ be drawn from a density function $f(\cdot)$. To estimate the density function by the method of sieves, we assume that the logarithm of the true density function admits the following series expansion

$$logf(x) = \sum_{j=0}^{\infty} \beta_j \phi_j(x),$$

  for some basis system $\{\phi_j(x)\}$.
- Let $K_n \to \infty$ as $n \to \infty$. We can approximate $logf(x)$ by

$$logf_{n0}(x;\beta) = \sum_{j=0}^{K_n} \beta_j \phi_j(x).$$

- Let $\beta = (\beta_0, \beta_1, \ldots, \beta_{J_n})'$. To ensure that $exp(\sum_{j=0}^{K_n} \beta_j \phi_j(x))$ is a density function, we need to normalize $f_n(x)$ by

$$C_n(\beta) = \int exp(\sum_{j=0}^{K_n} \beta_j \phi_j(x))dx,$$

which ensures that $f_n(x, \beta) = f_{n0}(x; \beta)/C_n(\beta)$ to be proper density function. The sieve method chooses β to minimize the following criterion function

$$\begin{aligned} Q_n(\beta) &= -n^{-1} \sum_{i=1}^{n} log f_n(X_i; \beta) \\ &= -n^{-1} \sum_{i=1}^{n} \Big\{ \sum_{j=0}^{K_n} \beta_j \phi_j(X_i) - log C_n(\beta) \Big\} \end{aligned}$$

The resulting estimator $\hat{\beta}$ is called the sieve maximum likelihood estimator (MLE), which is a special case of the general sieve extreme estimators.