TA Session 9: Preview of Semi/Nonparametric and Review of ARMA

Shuowen Chen[1]

EC708: PhD Econometrics I (Spring 2020)

[1] Materials are based on Hamilton (1994), Hayashi (2000), Ichimura & Todd (2007), Li & Racine (2007) and teaching slides by Bruce Hansen, Jean–Jacques Forneron, Hiroaki Kaido and Pierre Perron. Students interested in a formal exposure to the topics are strongly recommended to take EC711 and EC712. Empirically–oriented students are recommended to take EC732.

Outline

- Preview of Semi/Nonparametric Econometrics
  - Local Nonpar: Kernel Density Estimation
  - Two Examples
    1. Nonparametric IV (Series/Sieve Estimation)
    2. Semiparametric: Partially Linear Model (Kernel Regression)
- Review of *ARMA* Process
  - Stationarity and Ergodicity
  - Wold's Decomposition
  - Stationarity and Identification of *ARMA*

Benefits and Challenges of Flexible Models

Why Nonparametric?

- All models are wrong, some are useful (George Box)
- Identification can come from parametric assumptions
- Nonparametric theory admits that all models are misspecified
- Objects of interest beyond conditional mean (e.g., quantile)

Challenges

- Identification
- Bias and variance tradeoff
- Rate of convergence (slower than $\sqrt{n}$)
- Tunning parameters (incomplete data–driven rules)
- Computation (curse of dimensionality)

Semiparametrics

- Middle child of parametric and nonparametric
- Recent work includes neural nets as first step (Gao & Li, 2019)

Deep Learning

- GAN in structural modeling: Kaji, Manresa & Pouliot (2020)
- GAN in Monte Carlo simulations: Athey et al. (2019)

Distribution and Density Estimations

Denote X as a r.v. with continuous distribution $F(x)$ and density $f(x)$.
Goal: estimate $F(x)$ and $f(x)$ from a random sample $\{X_1, \ldots, X_n\}$

- Empirical CDF:
$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \le x)$$
- Define $\mathbb{G}_n = \sqrt{n}(\mathbb{F}_n(x) - F(x))$, by SLLN and CLT[2]
$$\mathbb{F}_n \xrightarrow{a.s.} F(x), \quad \mathbb{G}_n \xrightarrow{d} \mathcal{N}(0, F(x)(1 - F(x)))$$
- Estimating f relies on numerical derivatives
$$\begin{aligned}\widehat{f}(x) &= \frac{\mathbb{F}_n(x+h) - \mathbb{F}_n(x-h)}{2h} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2h} \mathbb{1}(x - h \le X_i \le x + h)\end{aligned}$$

[2]This is a special example of empirical process, which is vital for theories of semi/nonparametric, bootstrap and statistical learning. See Kosorok (2008) for a textbook introduction.

Roles of Kernel and Bandwidth

- Define $k(x) = \frac{1}{2}\mathbb{1}(|x| \leq 1)$, then

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)$$

- The estimator counts percentage of obs close to x
- Kernel $k(\cdot)$: how much weight to put on each obs around x
  - The $k(x)$ here is **uniform** kernel: gives weight $\frac{1}{2h}$ to each obs between $x - h$ and $x + h$ and zeros to all others
  - There are many other kernel functions
- Bandwidth h: how many obs around x to include
  - Important for bias and variance
  - Requires some tunning

▸ Kernel Functions ▸ Kernel Examples

Bias and Variance of Kernal Density Estimator

Denote $m_2(k)$ as the second moment of the uniform kernel and $f^{(2)}$ as the second order derivative of f

$$Bias(\widehat{f}(x)) = \frac{1}{2} f^{(2)}(x) h^2 m_2(k) + o(h^4)$$

Denote $R(k) = \int k(x)^2 dx$

$$Var(\widehat{f}(x)) = \frac{f(x)R(k)}{nh} + O(\frac{1}{n})$$

Remarks:

- An increase in h increases bias while decreases variance
- Why? Large h means we use obs far away, which leads to more inaccuracy. Small h means we use few obs and variability of estiamtes is high.
- The tradeoff carries over to other nonparametric methods

Measure of Estimation Precision

Recall definition of mean squared errors:

$$MSE(\widehat{f}(x)) = \mathbb{E}(\widehat{f}(x) - f(x))^2 = Bias^2(\widehat{f}(x)) + Var(\widehat{f}(x))$$

Use an asymptotic approximation:

$$AMSE(\widehat{f}(x)) = \left[\frac{1}{2}f^{(2)}(x)h^2 m_2(k)\right]^2 + \frac{f(x)R(k)}{nh}$$

A global measure: Asymptotic Mean Integrated Squared Error

$$AMISE = \int AMSE(\widehat{f}(x))dx = \frac{m_2^2(k)}{4}R(f^{(2)})h^4 + \frac{R(k)}{nh},$$

where $R(f^{(2)}) = \int (f^{(2)}(x))^2 dx$

- The value of h that minimizes $AMISE$ is the asymptotically optimal bandwidth
- Stata uses Silverman Rule–of–Thumb bandwidth
- Another common stratety is cross validation (cross-fitting)

Quick Notes on Inference

- Hypothesis testing requires estimation of variance
- Bias estimation is important in nonparametric estimation
- Bypass bias using undersmoothing: set h smaller than optimal
- In practice, compute the optimal bandwidth and set h smaller
- Difference between pointwise and uniform convergence, hence pointwise and uniform confidence bands[3]

[3] Also applies to quantile and distribution regressions.

Nonparametric IV

Consider the following model:

$$Y = g(X) + u, \quad \mathbb{E}(u|X) \neq 0, \quad \mathbb{E}(u|Z) = 0$$

What does g mean?

$$\mathbb{E}(Y|Z=z) = \int g(x)f(x|z)dx$$

- $\mathbb{E}(Y_i|Z_i = z)$ and $f(x|z)$ can be consistently estimated
- But solution of g is not necessarily unique[4]: ill–posedness
- **Trouble it causes**: even if consistent estimators of $\mathbb{E}(Y|Z=z)$ and $f(x|z)$ were plugged in, no consistent estimator of g
- Newey and Powell (2003) put restrictions on space of g: compact set under the Sobolev norm

[4] We would need g to be a continuous functional of $\mathbb{E}(Y|Z)$ and $f(x|z)$, or completeness of condiditonal expectation of functions of x conditional on z, which is both esoteric and hard to test in nonparametric models (Canay, Santos and Shaikh, 2012). In linear models completeness is like rank conditions.

Nonparametric IV: Series Estimation

- Suppose g can be approximated using series approximation

$$g(x) \approx \sum_{j=1}^{\mathcal{J}} \gamma_j p_j(x)$$

- Now we have

$$\mathbb{E}(Y|Z=z) \approx \sum_{j=1}^{\mathcal{J}} \gamma_j \mathbb{E}(p_j(x)|Z=z)$$

Estimation Procedures:

1. Series estimation of Z on basis functions $\{p_j(x)\}$ and get $\widehat{\mathbb{E}(p_j|Z)}$
2. OLS on Y against $\widehat{\mathbb{E}(p_j|Z)}$ to get $\{\widehat{\gamma}_j\}$

Remarks:

- $\widehat{g}(x) = \sum_{j=1}^{\mathcal{J}} \widehat{\gamma}_j p_j(x)$, uniformly consistent
- Rate of convergence and asymptotic normality are tricky, need to adjust for standard errors

▸ Series Estimation

Semiparametric: Partially Linear Model

Consider the model

$$Y = X\beta + g(Z) + e, \quad \mathbb{E}(e|X,Z) = 0$$

- Parametric part: linear specification for X
- Nonparametric part: how Z enters the regression
- Allow for heteroskedasticity $Var(e|X,Z) = \sigma^2(X,Z)$

We follow the approach by Robinson (1988)

$$Y - \mathbb{E}(Y|Z) = (X - \mathbb{E}(X|Z))\beta + e$$

- If we could observe $\mathbb{E}(Y|Z)$ and $\mathbb{E}(X|Z)$, then OLS applies
- Need to consider identification though:

$$(X - \mathbb{E}(X|Z))'(X - \mathbb{E}(X|Z))$$

X cannot contain a constant, none of X can be a deterministic function of Z.

Feasible Estimation: Two–Step Procedure

- In practice, need to estimate $\mathbb{E}(Y|Z)$ and $\mathbb{E}(X|Z)$
- Denote $\widehat{Y} = Y - \widehat{\mathbb{E}}(Y|Z)$ and $\widehat{X} = X - \widehat{E}(X|Z)$, OLS estimator

$$\widehat{\beta} = [\widehat{X}'\widehat{X}]^{-1}\widehat{X}'\widehat{Y},$$

- Essentially FWL theorem, decomposition details

$$\widehat{\beta} - \beta = [\widehat{X}'\widehat{X}]^{-1}\widehat{X}'\Big(e - [\widehat{\mathbb{E}}(Y|Z) - \mathbb{E}(Y|Z)] + [\widehat{\mathbb{E}}(X|Z) - \mathbb{E}(X|Z)]\beta\Big)$$

- Use undersmoothing to get rid of bias (blue terms)
- Work with variance as in OLS (red)
- **Double Machine Learning** method (Chernozhukov et al., 2017) replaces with first–step machine learning to cope with more covariates in Z

▸ Nadaraya–Watson kernel estimator

Stationarity

Strict Stationarity

- For any admissible $t_1, \cdots, t_n$ and any k, the joint probability distribution of $\{x(t_1), \cdots, x(t_n)\}$ is identical to that of $\{x(t_1 + k), \cdots, x(t_n + k)\}$
- Invariant probability structure under a shift of the time origin

Stationarity up to Order m

- For any admissible $t_1, \cdots, t_n$ and any k, all joint moments up to order m of $\{x(t_1), \cdots, x(t_n)\}$ exist and are identical to those of $\{x(t_1 + k), \cdots, x(t_n + k)\}$

(Weak) Stationarity: $m = 2$

- First two moments independent of time

$$\mathbb{E}[x(t)] = \mu, \quad \mathbb{E}[x(t)^2] = \mu_2$$

- $\mathbb{E}[x(t)x(s)]$ is a function of $t - s$ only
- Variance is time–invariant and covariance only depends on $t - s$

Autocovariance and Autocorrelation

For a stationary process $\{x_t\}$ with mean μ and variance σ^2,

- Autocavariance Function $R(\tau)$:

$$R(\tau) = \mathbb{E}[(x_t - \mu)(x_{t-\tau} - \mu)]$$

- Autocorrelation Function $\rho(\tau)$:

$$\rho(\tau) = \frac{R(\tau)}{R(0)},$$

  where $R(0) = Var(x_t) = \sigma^2$

Remarks:

- $|R(\tau)| \leq R(0)$ and $|\rho(\tau)| \leq \rho(0) = 1, \forall \tau$
- If $\{x_t\}$ is real–valued, then $R(-\tau) = R(\tau)$ and $\rho(-\tau) = \rho(\tau)$

Ergodicity

- If you have time series with sample size T in S parallel universes:

$$\{x_t^s\}_{t=1}^T, \quad s = 1, \ldots, S$$

- Ensemble mean of t–th observation: $E(x_t) = plim_{S\to\infty} \frac{1}{S}\sum_{s=1}^S x_t^s$
- Meaning? Average cross all different states at time t
- In practice, only have one time series $\{x_t\}_{t=1}^T$ and thus time average $\overline{x} = \frac{1}{T}\sum_{t=1}^T x_t$

Question: When is time average consistent for ensemble mean?

- Stationarity: $\mathbb{E}(x_t) = \mu$, $\mathbb{E}[(x_t - \mu)(x_{t-\tau} - \mu)] = R(\tau)$
- Absolute summability: $\sum_{\tau=0}^{\infty} |R(\tau)| < \infty$
- Result[5]: $\sqrt{T}(\overline{x} - \mu) \to \mathcal{N}(0, \sum_{\tau=-\infty}^{\infty} R(\tau))$

Remark:

- Important for simulation–based methods on dynamic models (Duffie & Singleton, 1993)

[5] Chapter 7 of Hamilton (1994); Chapter 6 of Hayashi (2000) has further discussions on Gordin's conditions.

Wold's Decomposition Theorem

Any stationary process can be represented in terms of a linear combination of i.i.d. random variable with mean 0

$$x_t = d_t + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$$

- $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ and $\varepsilon_t = x_t - \mathbb{E}[x_t | x_{t-s}, s \geq 1]$
- ε_t : white noise; forecast error from the optimal estimator of x_t using past information
- d_t uncorrelated with ε_{t-j}
- $d_t = \mathbb{E}(d_t | x_{t-s}, s \geq 1)$ (linearly deterministic component)

Remarks:

- Stationary process can be represented as an ARMA process
- For estimation in practice need some further assumptions on ψ_j
- Limiting distribution of $\{x_t\}$ depends on rate of decay of weights of moving average representation Long and Short Memory

Autoregressive Moving Average Process

An $ARMA(p,q)$ process is defined as a random process $\{x_t\}$ such that

$$A(L)x_t = B(L)e_t$$

- e_t is i.i.d. $(0, \sigma_e^2)$
- $A(L) = 1 - a_1 L - a_2 L^2 - \cdots - a_p L^p$ (Autoregressive)
- $B(L) = 1 + b_1 L + b_2 L^2 + \cdots + b_q L^q$ (Moving Average)
- Lag operator: $L^p x_t \equiv x_{t-p}$

Alternative representation of the process:

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=0}^{q} e_{t-j}$$

Remarks:

- $\{x_t\}$ is an $ARMA(p,q)$ process with mean μ if $\{x_t - \mu\}$ is an $ARMA(p,q)$ process
- When $A(L) = B(L) = 1$, $\{x_t\}$ is called white noise[6]

[6] The name is due to spectral theory, which won't be covered in this session.

Stationarity of $AR(p)$ Process

Consider the following process

$$x_t = a_1 x_{t-1} + \cdots + a_p x_{t-p} + e_t$$

- For stationarity, needs the roots of

  $$A(z) := 1 - a_1 z - \cdots - a_p z^p = 0$$

  all lie outside the unit circle
- Can express $A(L) = \prod_p^{i=1} \left(1 - \frac{1}{\mu_i^*} L\right)$, where μ_i^*'s are roots of $A(z) = 0$
- Can express $\{x_t\}$ as an infinite MA process

  $$x_t = A^{-1}(L) e_t = \sum_{i=0}^{\infty} k_i e_{t-i},$$

  where k_i's are functions of μ_i^*'s

Stationarity of $MA(q)$ Process

Consider the following process

$$x_t = e_t + b_1 e_{t-1} + \cdots + b_q e_{t-q}$$

- $\mathbb{E}(x_t) = 0$
- $Var(x_t) = (1 + b_1^2 + \cdots + b_q^2)\sigma_e^2$
- Other orders of autocovariance:

$$\mathbb{E}[x_t x_{t-\tau}] = \begin{cases} \sigma_e^2[b_\tau + b_{\tau+1}b_1 + \cdots + b_q b_{q-\tau}] & \tau = 1, 2, \ldots, q \\ 0 & \tau > q \end{cases}$$

Remarks:

- $MA(q)$ is always stationary
- Autocorrelation $\rho(\tau)$ cut off after q lags

Invertibility of $MA(q)$ Process

Consider two $MA(1)$ processes

$$x_t = (1 - b_1 L)e_t, \quad e_t \sim i.i.d.(0, \sigma_e^2) \tag{1}$$

$$x_t = \left(1 - \frac{1}{b_1}L\right)\tilde{e}_t, \quad \tilde{e}_t \sim i.i.d.(0, c^2\sigma_e^2) \tag{2}$$

- Same first, second moments and autocorrelation function
- Indistinguishable

Recursive substitution leads to $e_t = x_t - b_1 e_{t-1} = \sum_{i=0}^{\infty}(-b_1)^i x_{t-i}$

- Mean–squared convergent only if $|b_1| < 1$: root of $B(L) = 0$ lies outside the unit circle
- For $MA(q)$ process, invertible if all roots of $B(z) = 0$ lie outside the unit circle, which leads to

$$e_t = B^{-1}(L)x_t = \sum_{i=0}^{\infty} h_i x_{t-i}$$

- Invertibility: express $MA(q)$ as an $AR(\infty)$ process with coefficients $\{h_i\}$ such that $\sum_{i=0}^{\infty}|h_i| < \infty$, hence ensuring that $\sum_{i=0}^{\infty} h_i x_{t-i}$ is mean–squared convergent (Key for VAR analysis)

Stationarity and Identification of $ARMA(p, q)$

Stationarity

- Roots of $A(z) = 0$ lie outside the unit circle

Identification

- Roots of $B(z) = 0$ lie outside the unit circle
- $A(L)$ and $B(L)$ have no common factors
- Why? Suppose $A(L) = C(L)D(L)$ and $B(L) = C(L)E(L)$, where $C(L)$ is a polynomial with order r, then $A(L)x_t = B(L)e_t$ is observationally equivalent to the $ARMA(p - r, q - r)$

$$D(L)x_t = E(L)e_t$$

Short and Long Memory Stationary Processes

Short Memory

- $\sum_{j=0}^{\infty} j|\psi_j| < \infty$ or $\sum_{\tau=0}^{\infty} |\rho(\tau)| < \infty$
- Applies to stationary and invertible $ARMA(p, q)$ processes

Long Memory

- If $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ is satisfied while $\sum_{j=0}^{\infty} j|\psi_j| < \infty$ is not
- One type: fractionally integrated process. For example: $(1-L)^d x_t = e_t$, where $d \in (0, 0.5)$. Then
$$\psi_j = \frac{d(d+1)\cdots(d+j-1)}{j!}$$
For large j, $\rho(\tau) \approx c\tau^{2d-1}$ for some constant, hence
$$\sum_{i=1}^{\infty} |\rho(i)| \approx c \sum_{i=0}^{\infty} i^{2d-1} = \infty$$
When $2d = 1 < 0$, autocorrelation hyperbolic rate of decay, much slower than geometric decay
- $ARIMA(p, d, q)$: $A(L)(1-L)^d x_t = B(L)e_t$

Kernel Functions

- $k(x) : \mathbb{R} \to \mathbb{R}$ such that $\int k(x)dx = 1$
- Kernel is nonnegative if $k(x) \geq 0 \; \forall x$. In this case it is a PDF.
- j–th moment: $m_j(k) = \int x^j k(x)dx$
- Symmetric kernel: $k(x) = k(-x) \; \forall x$
- Order ν: the order of the first non–zero moment. Kernels with order larger than 2 have negative parts and hence not probability densities. Dubbed bias–reducing kernels

◂ Back

Kernel Examples

1. Triangular Kernel

$$k(x) = (1 - |x|)\mathbb{1}(|x| \leq 1)$$

2. Epanechnikov Kernel

$$k(x) = \frac{3}{4}(1 - x^2)\mathbb{1}(|x| \leq 1)$$

3. Gaussian Kernel

$$k(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(- \frac{x^2}{2}\Big)$$

Remarks:

- Gaussian kernel has infinite support
- Triangular kernel indifferentiable at 0. Consistency and asymptotic distribution proofs often use mean value theorem which requires differentiability.

Back

Series Estimation

$$Y = g(X) + u, \quad g(X) = \mathbb{E}(Y|X)$$

- Kernel based methods[7] estimate values of $g(x)$ for each x individually: only consider observations close to x
- Series estimation is global: one regression estimates $g(x)$ **for all** x
- How? Allow for flexible function form

$$g \approx \sum_{j=1}^{\mathcal{J}} \gamma_j g_j$$

  where $\{g_j\}_{j=1}^{\mathcal{J}}$ is a set of basis functions and $\{\gamma_j\}_{j=1}^{\mathcal{J}}$ are coefficients to be estimated
- Replaces infinite–dimensional estimation with estimation over a large finite–dimensional space (**Sieve Space**, Chen, 2007)
- Neural network is also sieve (Chen and White, 1999)

[7] Other methods include k–means (Bonhomme, Lamadon & Manresa, 2019), nearest neighbor matching and local polynomial.

Series Estimation: Intuition and Basis Functions

- $\sup_{x\in\mathcal{S}} |\sum_{j=1}^{\mathcal{J}} \gamma_j g_j(x) - g(x)| = O(\mathcal{J}^{-\alpha})$ for some $\alpha > 0$ and compact set $\mathcal{S}$. Requires g to be differentiale up to some order.
- Two common choices of basis functions:
  1. Power Series: $g_j = x^j$
  2. Splines: $f : \mathbb{R} \to \mathbb{R}$ is a k–th order spline with knot points at $t_1 < \cdots < t_m$ if f is a polynomial of degree j on each of the intervals $(-\infty, t_1], [t_1, t_2], \cdots, [t_m, \infty)$ and $f^{(l)}$ is continuous at $t_1, \cdots, t_m$ for each $l = 0, 1, \cdots, k-1$. Total dimension is $j = k + m + 1$
- Other basis: wavelets, B–splines, Bernstein polynomials, etc.
- Adding number of knots or terms of power series, more flexible fit and decreases bias, analogous to shrinking bandwidth for kernels
- Choosing $\mathcal{J}$? Mallow's C_L, generalized cross validation, leave–one–out cross validation

◂ Back

Kernel Regression

Consider the model[8]

$$Y = g(X) + e, \quad g(X) = \mathbb{E}(Y|X)$$

Goal is to estimate $g(X)$. First note that

$$\mathbb{E}(Y|X = x) = \int y f_{Y|X}(y|X = x)dy = \frac{\int y f_{Y,X}(y, x)dy}{f_X(x)}$$

- We can estimate $f_X(x)$ using kernel density estimation $\widehat{f}_X(x)$
- What about $f_{Y,X}(y, x)$? We can extend density estimation to multivariate case with product kernel: $K(y, x) = k(y)k(x)$

$$\widehat{f}_{Y,X}(y, x) = \frac{1}{nh_X h_Y} \sum_{i=1}^{n} k\Big(\frac{Y_i - y}{h_Y}\Big) k\Big(\frac{X_i - x}{h_X}\Big)$$

- Some algebra gives

$$\int y \widehat{f}_{Y,X}(y, x)dy = \frac{1}{nh_X} \sum_{i=1}^{n} k\Big(\frac{X_i - x}{h_X}\Big) Y_i$$

[8] For simplicity I assume X is univariate, but can be multivariate.

Nadaraya–Watson Kernel Estimator

This gives us the following $\widehat{g}(x) := \widehat{\mathbb{E}}(Y|X = x)$:

$$\widehat{g}(x) = \frac{\sum_{i=1}^{n} Y_i k\left(\frac{X_i - x}{h_X}\right)}{\sum_{i=1}^{n} k\left(\frac{X_i - x}{h_X}\right)}$$

Denote $k_{h_X}(X_i - x) := \frac{k\left(\frac{X_i - x}{h_X}\right)}{\sum_{i=1}^{n} k\left(\frac{X_i - x}{h_X}\right)}$, then we have

$$\widehat{g}(x) = \sum_{i=1}^{n} k_{h_X}(X_i - x) Y_i$$

Remarks:

- Point estimator: different for each value of x
- Weighted average of Y_i, more weight to observations such that X_i is close to x
- Poor performance at end points (one side observations), use local polynomial as an alternative