Introduction to Nonparametric/Semiparametric Econometric Analysis: Implementation

Yoichi Arai

National Graduate Institute for Policy Studies

2014 JEA Spring Meeting (14 June)

Motivation: RD Estimates of the Effect of Head Start Assistance by Ludwig and Miller (2007, QJE)

Variable	Nonparametric		
Bandwidth	9	18	36
Number of obs. with nonzero weight	[217, 310]	[287, 674]	[300, 1877]
1968 HS spending per child			
RD estimate	137.251	114.711	134.491^{**}
	(128.968)	(91.267)	(62.593)
1972 HS spending per child			
RD estimate	182.119^{*}	88.959	130.153^{*}
	(148.321)	(101.697)	(67.613)
Age 5–9, Mortality, 1973–83			
RD estimate	-1.895^{**}	-1.198^{*}	-1.114^{**}
	(0.980)	(0.796)	(0.544)
Blacks age 5–9, Mortality, 1973–83			
RD estimate	−2.275	-2.719^{**}	−1.589
	(3.758)	(2.163)	(1.706)

Observations

- Estimates can change dramatically by the choice of bandwidths.
- Statistical significance can also change depending on the choice of bandwidths.

Lessons

- It would be nice to have objective criterion to choose bandwidths!

MSE (MISE): Measures of Discrepancy

Suppose your objective is to estimate

- some function $f(x)$ (f evaluated at x), or
- some function f over entire support.

Let $\hat{f}_h$ be some estimator based on a bandwidth h.

Most Popular Measures of Discrepancy of $\hat{f}$ from the true objective f

- $MSE(x) = E[\{\hat{f}_h(x) - f(x)\}^2]$ (Local Measure).
- $MISE = \int E[\{\hat{f}_h(x) - f(x)\}^2]dx$ (Global Measure).
- MSE and $MISE$ changes depending on the function f as well as estimation methods.

Kernel Functions

Table : Popular Choices for Kernel Functions

Name	Function
Normal	$(2\pi)^{-1/2} e^{-x^2/2}$
Uniform	$(1/2)1\{\lvert x\rvert < 1\}$
Epanechnikov	$(3/4)(1 - x^2)1\{\lvert x\rvert < 1\}$
Triangular	$(1 - \lvert x\rvert)1\{\lvert x\rvert < 1\}$

Practical Choices

- It is well-known that nonparametric estimates are not very sensitive to the choice of kernel functions.
- For estimating a function at interior points or globally, a common choice is the Epanechnikov kernel (Hodges & Lehmann, 1956).
- For estimating a function at boundary points (by LLR), a popular choice is that the triangular kernel (Cheng, Fan & Marron,1997).

Bandwidth Selection

> Contrary to the selection of kernel functions, it is well-known that estimates are sensitive to the choice of bandwidths.

In the following, we briefly explain 3 popular approaches for bandwidth selection

1. Rule of Thumb Bandwidth
2. Plug-In Method
3. Cross-Validation

See Silverman (1986) for more about basic treatment on density estimation.

AMSE for Kernel Density Estimators

Given a random sample $\{X_i, i = 1, 2, \ldots, n\}$, we are interested in estimating its density f.

For the kernel density estimator

$$\hat{f}_h = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

the asymptotic approximation of the MSE (AMSE) is given by

$$AMSE(x) = \left\{\frac{\mu_2}{2} f^{(2)}(x) h^2\right\}^2 + \frac{\kappa_2 f(x)}{nh}$$

where $f^{(r)}$ is the r-th derivative of f.

Similarly, the asymptotic approximation of the MISE (AMISE) is given by

$$AMISE = \frac{\mu_2^2}{4}\left\{\int f^{(2)}(x)^2 dx\right\} h^4 + \frac{\kappa_2}{nh}$$

Optimal Bandwidth

Bandwidths that minimize the AMSE and AMISE are given, respectively, by

$$h_{AMSE} = C(K)\left\{\frac{f(x)}{f^{(2)}(x)^2}\right\}^{1/5} n^{-1/5}$$

and

$$h_{AMISE} = \underbrace{C(K)}_{\text{depends on } K} \underbrace{\left\{\frac{1}{\int f^{(2)}(x)^2 dx}\right\}^{1/5}}_{\text{depends on } f} n^{-1/5}$$

where $C(K) = \{\kappa_2/\mu_2^2\}^{1/5}$. Both h_{AMSE} and h_{AMISE} depend on 3 things

1. K (Kernel function),
2. f (true density including the 2nd derivative $f^{(2)}$),
3. n (sample size).

In addition, h_{AMSE} depends on the evaluation point x.

Bandwidth Selection I: Rule of Thumb Bandwidth

Rule of Thumb (ROT) Bandwidth can be obtained by specifying, for h_{AMISE},

- Gaussian kernel for K, and
- Gaussian density with variance σ^2 for f,

implying

$$h_{ROT} = 1.06\sigma n^{-1/5}.$$

Remark

- In practice, we use an estimated $\hat{\sigma}$ for σ.
- This is the default bandwidth used by Stata command `kdensity`.
- Obviously, h_{ROT} works well if the true density is Gaussian.
- Not necessarily works well if the true density is not Gaussian.

Bandwidth Selection II: Plug-In Method

Rather than assuming Gaussian density, the plug-in method estimates

- f and $f^{(2)}$ for h_{AMSE},
- $\psi = \int f^2(x)^2 dx$ for h_{AMISE}.

A standard kernel density and density derivative estimator is given by

$$\hat{f}_{a_1}(x) = \frac{1}{na_1}\sum_{i=1}^{n} K\left(\frac{x - X_i}{a_1}\right), \quad \hat{f}_{a_2}^{(d)}(x) = \frac{1}{na_2^{d+1}}\sum_{i=1}^{n} K^{(d)}\left(\frac{x - X_i}{a_2}\right)$$

ψ can be estimated by

$$\hat{\psi} = n^{-1}\sum_{i=1}^{n} \hat{f}_{a_3}^{(4)}(X_i).$$

Remark

- These require to choose the bandwidths a_1, a_2 and a_3.
- Those are usually chosen by a simple rule such as the ROT rule.
- The plug-in method introduced here is often called direct plug-in (DPI).

Bandwidth Selection II: Plug-In Method

There exists a more sophisticated method proposed by Sheather and Jones (1991, JRSS B).

- The pilot bandwidths such as a_1, a_2, a_3 can be written as a function of h.
- Determine the bandwidths h and the pilot bandwidths simultaneously.

The bandwidth chosen in this manner is called the solve-the-equation (STE) rule.

Remark

- Simulation studies show the STE bandwidths perform very well.
- The DPI and STE bandwidths can be obtained by the Stata command `kdens`.
- See also Wand and Jones (1994) for more about these bandwidths.

Bandwidth Selection III: Cross Validation

Least Squares Cross Validation (LSCV) bandwidth minimizes

$$LSCV(h) = \int \hat{f}_h(x)^2 dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{-i,h}(X_i)$$

where the leave-one-out kernel density estimator is given by

$$\hat{f}_{-i,h}(x) = \frac{1}{n-1} \sum_{j \neq i}^{n} \left(\frac{x - X_j}{h} \right).$$

Rationale for the LSCV

- Observe that
$$\int (\hat{f}_h(x) - f(x))^2 dx = R(\hat{f}_h) + \int f(x)^2 dx.$$
where
$$R(\hat{f}_h) = \int \hat{f}_h(x)^2 dx - 2 \int \hat{f}_h(x) f(x) dx.$$
- Then we can show that
$$E[LSCV(h)] = E[R(\hat{f})].$$

Bandwidth Selection III: Cross Validation

Some Remarks on the LSCV

- The LSCV is based on the global measure by construction.
- The LSCV requires numerical optimization.
- Then the LSCV can be computationally very intensive.
- Some simulation studies show that the LSCV bandwidth tends to be very dispersed.

AMSE for the Local Linear Regression

Given a random sample $\{(Y_i, X_i), i = 1, 2, \ldots, n\}$, we are interested in estimating the regression function

$$m(x) = E[Y_i | X_i = x].$$

The local linear regression can be obtained by minimizing

$$\sum_{i=1}^{n} \{y_i - \alpha - \beta(X_i - x)\}^2 K\left(\frac{X_i - x}{h}\right)$$

and the resulting $\hat{\alpha}$ estimates $m(x)$.

The AMSE for the LLR is given by

$$AMSE(x) = \frac{\mu_2^2}{4} m^{(2)}(x) h^4 + \frac{\kappa_2 \sigma^2(x)}{nhf(x)}$$

LLR is popular because of design adaptation property especially at boundary points. (See Fan and Gijbels, 1996.)

Optimal Bandwidth for the Local Linear Regression

The optimal bandwidth is given by

$$h_{AMSE} = C(K)\left\{\frac{\sigma^2(x)}{m^{(2)}(x)^2 f(x)}\right\}^{1/5} n^{-1/5}.$$

For global estimation, the commonly used bandwidth minimizes

$$\int AMSE(x)w(x)dx$$

where $w(x)$ is a weighting function and it is given by

$$h_{AMISE} = \underbrace{C(K)}_{\text{depends on } K} \underbrace{\left\{\frac{\int \sigma^2(x)w(x)/f(x)dx}{\int m^{(2)}(x)^2 w(x)dx}\right\}^{1/5}}_{\text{depends on } m^{(2)},\sigma^2,f, \text{ and } w} n^{-1/5}.$$

Bandwidth Selection I: Plug-In Method

The plug-in bandwidth is given by

$$h_{ROT} = C(K)\left\{\frac{\hat{\sigma}^2 \int w(x)dx}{\sum_{i=1}^{n} \hat{m}^{(2)}(X_i)^2 w(X_i)}\right\}^{1/5}.$$

where $\hat{\sigma}^2$ and $\hat{m}^{(2)}$ are obtained by the global polynomial regression of order 4.
Remark:

- A possible choice for $w(x)$ is the uniform kernel constructed to cover 90% of the sample.
- This is the default bandwidth used by the Stata command `lpoly`.
- This bandwidth is also called the ROT bandwidth.

Bandwidth Selection II: Cross-Validation

The bandwidth based on the cross-validation minimizes

$$CV(h) \equiv \sum_{i=1}^{n} \{y_i - \hat{f}_{-i,h}(X_i)\}^2$$

where $\hat{f}_{-i,}$ is the leave-one-out LLR estimates.

That is

$$h_{CV} = \arg\min_h CV(h)$$

Remark

- This bandwidth can be obtained by the Stata command `locreg`

Bandwidth Selection III: More Sophisticated Method

Remember that the AMSE for the LLR is given by

$$AMSE(x) = \frac{\mu_2^2}{4} m^{(2)}(x) h^4 + \frac{\kappa_2 \sigma^2(x)}{nhf(x)}.$$

There exists a method to obtain the finite sample approximation of the whole bias and variance component proposed by Fan, Gijbels, Hu and Huang (1996).

Let $\widehat{MSE}(x, h)$ be a finite sample approximation of the AMSE. Then the refined bandwidth is given by

$$h_R = \arg\min_h \int \widehat{MSE}(x, h) dx$$

Remark

- This bandwidth works better than the plug-in bandwidth but not universally.
- There exist several modified bandwidths.

Sharp RD Design

Let

- Y_1, Y_0: potential outcomes for treated and untreated,
- Y: observed outcome, $Y = DY_1 + (1 - D)Y_0$,
- D be a binary indicator for treatment status, 1 for treated and 0 for untreated.

In the sharp RD design, the treatment D is determined by the assignment variable Z

$$D = \begin{cases} 1 & \text{if } Z \geq c \\ 0 & \text{if } Z < c \end{cases}$$

where c is the cut-off point.

- We can show that the ATE at the cut-off point is defined and represented by

$$E[Y_1 - Y_0 | Z = c] = \lim_{z \to c+} E[Y | Z = z] - \lim_{z \to c-} E[Y | Z = z].$$

Illustration of Sharp RDD

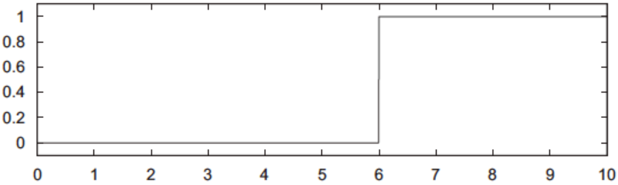

Fig. 1. Assignment probabilities (SRD).

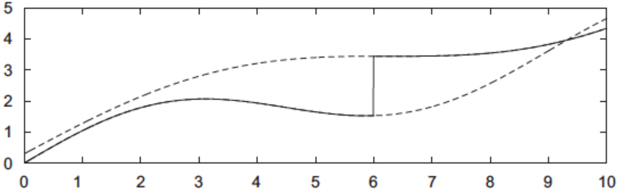

Fig. 2. Potential and observed outcome regression functions.

Figures are taken from Imbens & Lemiux (2008).

Local Approach versus Global Approach

Local Approach

- It suffices to assume local continuity.
- Robust to outliers and discontinuities.

Global Approach

- Assumes global smoothness.
- Obviously vulnerable to outliers and discontinuities.
- Can use more observations.

Currently, it is popular to employ the LLR (local approach) to estimate the RD estimator.

Bandwidth Selection

- It is important to note that our objective is to estimate not $\lim_{z \to c+} E[Y|Z=z]$ (or $\lim_{z \to c-} E[Y|Z=z]$) but the ATE at the cut-off point.

Existing Approaches for Bandwidth Selection

1. Ad-hoc Approach: Choose optimal bandwidths to estimate $\lim_{z \to c+} E[Y|Z=z]$ (or $\lim_{z \to c-} E[Y|Z=z]$).
2. Local CV: Local Version of Cross-Validation (quasi-local criterion)
3. Optimal Bandwidth with Regularization proposed by Imbens and Kalyanaraman (2012)
4. Simultaneous Selection of Optimal Bandwidths proposed by Arai and Ichimura (2014)

Bandwidth proposed by Imbens and Kalyanaraman (2012)

Basic Idea

- Use a single bandwidth to estimate the ATE at the cut-off point.
- Propose the bandwidth that minimizes the AMSE and modify it with regularization term.

Let f be the density of Z,

$$m_1(c) = \lim_{z \to c+} E[Y|Z = z], \quad m_0(c) = \lim_{z \to c-} E[Y|Z = z],$$
$$\sigma_1^2(c) = \lim_{z \to c+} Var[Y|Z = z], \quad \sigma_0^2(c) = \lim_{z \to c-} Var[Y|Z = z].$$

Then the AMSE for the RD estimator is given by

$$AMSE_n(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(c) - m_0^{(2)}(c) \right] h^2 \right\}^2 + \frac{v}{nhf(c)} \left\{ \sigma_1^2(c) + \sigma_0^2(c) \right\}.$$

where b_1 and v are the constants that depend on the kernel function.

Bandwidth proposed by Imbens and Kalyanaraman (2012)

Then the optimal bandwidth is given by

$$h_{opt} = C_K \left\{ \frac{\sigma_1^2(c) + \sigma_0^2(c)}{f(c)\left(m_1^{(2)}(c) - m_0^{(2)}(c)\right)^2} \right\} n^{-1/5}$$

The bandwidth proposed by IK is

$$h_{IK} = C_K \left\{ \frac{\hat{\sigma}_1^2(c) + \hat{\sigma}_0^2(c)}{\hat{f}(c)\left[\left(\hat{m}_1^{(2)}(c) - \hat{m}_0^{(2)}(c)\right)^2 + \hat{r}\right]} \right\} n^{-1/5}$$

where $\hat{r}$ is, what they term, a regularization term.

Remark

- h_{opt} can be very large when $m_1^{(2)}(c) - m_0^{(2)}(c)$ is small.
- The regularization term is basically to avoid the small denominator.

Bandwidth proposed by Arai and Ichimura (2014)

Basic Idea

- Choose two bandwidths simultaneously.
- Propose the bandwidth that minimizes the AMSE with the second-order bias term.

With two bandwidths, the AMSE is given by

$$AMSE_n(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(c) h_1^2 - m_0^{(2)}(c) h_0^2 \right] \right\}^2 + \frac{v}{nf(c)} \left\{ \frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0} \right\}.$$

Arai and Ichimura (2014) show

> The minimization problem of the AMSE is not well-defined because the bias-variance trade-off breaks down.

Bandwidth proposed by Arai and Ichimura (2014)

Instead, Arai and Ichimura (2014) propose the bandwidth h_{MMSE} that minimizes

$$MMSE_n(h) = \frac{b_1^2}{4}\left[\hat{m}_1^{(2)}(c)h_1^2 - \hat{m}_0^{(2)}(c)h_0^2\right]^2 + \left[\hat{b}_{2,1}(c)h_1^3 - \hat{b}_{2,0}(c)h_0^3\right]^2 + \frac{v}{n\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{h_1} + \frac{\hat{\sigma}_0^2(c)}{h_0}\right\},$$

where the second term is the squared second-order-bias term.

Observations

- The bias of the RD estimator based on h_{IK} can be large for some designs.
- The RD estimator based on h_{MMSE} is robust to designs.
- The Stata ado file to implement the bandwidth is available at http://www3.grips.ac.jp/~yarai/.

Ludwig and Miller (2007) Data Revisited

Variable	MMSE	IK
1968 Head Start spending per child		
Bandwidth	[26.237, 45.925]	19.012
RD estimate	110.590	108.128
	(76.102)	(80.179)
1972 Head Start spending per child		
Bandwidth	[22.669, 42.943]	20.924
RD estimate	105.832	89.102
	(79.733)	(84.027)
Age 5–9, Mortality, 1973–1983		
Bandwidth	[8.038, 14.113]	7.074
RD estimate	-2.094^{***}	-2.359^{***}
	(0.606)	(0.822)
Blacks age 5–9, Mortality, 1973–1983		
Bandwidth	[22.290, 25.924]	9.832
RD estimate	-2.676^{***}	-1.394
	(1.164)	(2.191)

Reference I

- Arai, Y. and H. Ichimura (2014) Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator. *GRIPS Working Paper* 14-03.
- Cheng, M.Y., J. Fan and J.S. Marron (1997) On Automatic Boundary Corrections. *AoS*, 25, 1691-1708.
- Fan, J. and I. Gijbels (1996) *Local polynomial modeling and its applications*. Chapman and Hall.
- Fan, J, I. Gijbels., T.C. Hu and L.S. Huang (1996) A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, 6, 113-127.
- Hodges, J.L. and E.L. Lehmann (1956) The efficiency of some nonparametric competitors of the *t*-test. *AMS*, 27, 324-335.
- Imbens, G.W. and K. Kalyanaraman (2012) Optimal bandwidth choice for the regression discontinuity estimator. *REStud*, 79, 933-959.

Reference II

- Imbens, G.W. and T. Lemieux (2008) Regression discontinuity designs: A guide to practice. *JoE*, 142, 615-635.
- Ludwig, J. and D.L. Miller (2007) Does head start improve children's life change?Evidence from a regression discontinuity design. *QJE*, 122, 159-208.
- Sheather, S.J. and M.C. Jones (1991) A reliable data-based bandwidth selection method for kernel density estimation. *JRSS B*, 53, 683–690.
- Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.
- Wand, M.P. and M.C. Jones (1994) *Kernel Smoothing*. Chapman and Hall.