Non- and Semiparametric Estimation Methods

Juan Carlos Escanciano

Department of Economics
Indiana University
http://mypage.iu.edu/~jescanci/
Outline

Objective of this talk

Nonparametric Estimation

Semiparametric Estimation

Implementation

Conclusions

Some Theory
Motivating Example

- Labor participation: $Y_i \in \{0, 1\}, \; X_i \in \mathbb{R}^d$.

- **Parametric** Probit model: $\mathbb{E}(Y_i|X_i) = \Phi(X_i'\theta_0), \; \theta_0$ unknown

- **Semiparametric** Single-Index model: $\mathbb{E}(Y_i|X_i) = \eta(X_i'\theta_0), \; \theta_0$ and $\eta$ unknown

- **Nonparametric** model: $\mathbb{E}(Y_i|X_i) = m(X_i), \; m$ unknown.
Objective of this talk

Present an introduction to Semi- and Non-parametric methods.

- **Advantage**: Robust to misspecification.
- **Disadvantage**: Curse of dimensionality.
- Semiparametric offers a compromise.

- **Warning**: This will not be a mathematically rigorous talk!
Inception: The empirical distribution

- $X_i$ denotes a r.v, e.g. income of household $i$.
- The probability distribution of $X_i$ can be characterized by its cdf: $F(x) := \mathbb{P}(X_i \leq x) = \mathbb{E}\left(1_{(-\infty,x]}(X_i)\right)$.
- Natural estimator: the empirical distribution function (edf)

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x]}(X_i).$$

- $F_n$ is a (random) cdf. For example $F_n(25.000) = 0.128$. 
Empirical cdf in R

```r
y <- rnorm(10)
Fn <- ecdf(y)
plot(Fn, verticals=TRUE, do.p=TRUE, lwd=3 )
```
Nonparametric Estimation

Non- and Semiparametric Methods
Figure 1: Nonparametric ecdf labor household log income 1989-1990
Asymptotic Theory (under iid)

- $F_n(x)$ is unbiased for $F(x)$.
- The strong LLN yields, for fixed $x \in \mathbb{R}$,
  \[ F_n(x) \rightarrow_{a.s.} F(x) := \mathbb{E}[1_{(-\infty, x]}(X_i)]. \]
- Moreover, also for fixed $x \in \mathbb{R}$, the standard CLT yields
  \[ \sqrt{n}(F_n(x) - F(x)) \rightarrow_d N(0, F(x)(1 - F(x))). \]
- Parametric rates of convergence for nonparametric estimates.
- Functional versions: e.g. Glivenko-Cantelli (1933).
Density Estimation

- **Parametric**: $f(x, \theta_0)$, estimate $\theta_0$ by MLE.
- **Nonparametric**: based on

$$f(x) := \lim_{h \to 0} \frac{F(x + h) - F(x - h)}{2h}$$

we can estimate $f$ by the naive kernel estimator

$$\hat{f}_h(x) := \frac{F_n(x + h) - F_n^-(x - h)}{2h}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{1\{x-h \leq X_i \leq x+h\}}{2h}$$

$$= \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \{1 \left| \frac{x-X_i}{h} \right| \leq 1\}.$$
Density Estimation

- $h$ is the so-called **bandwidth** parameter.
- We can write the naive estimator as

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

where $K(u) = 0.51 \{ |u| \leq 1 \}$. 
- Different from the **histogram**.
Density Estimation: cont.

- More generally, we can replace \( K(u) = 0.51\{|u|\leq 1\} \) with a continuous function that integrates up to one.
- This leads to Rosenblatt-Parzen’s kernel density estimator.
- Popular kernels are Gaussian and Epanechnikov

\[
K(u) = \frac{1}{\sqrt{2\pi}} e^{-0.5u^2}
\]

\[
K(u) = \frac{3}{4} (1 - u^2) 1\{|u|\leq 1\}.
\]

- Often, we also require \( K \) to be symmetric.
Density estimation in R

```r
his <- hist(y)
bw <- npudensbw(dat=y, ckerttype="gaussian", bwmethod="cv.ls",
               bandwidth.compute=TRUE)
# y Bandwidth(s): 0.0103764
densy <- npudens(bws=bw, tdat=y)
```
Figure 2: Histogram of Log income 1989-1990
Figure 3: Gaussian KD of Log income 1989-1990. CV band.
Figure 4: Gaussian KD of Log income 1989-1990. 0.1*CV band.
Figure 5: Gaussian KD of Log income 1989-1990. 5*CV band.
Density Estimation: cont.

Properties of $\hat{f}_h$: under iid and $f^{(k)}(x) = \partial^k f(x)/\partial x^k$.

- By change of variables $u = -(x - y)/h$, $y = x + uh$,

\[
\mathbb{E} \left( \hat{f}_h(x) \right) = \mathbb{E} \left( \frac{1}{h} K \left( \frac{x - X_i}{h} \right) \right) \\
= \int \frac{1}{h} K \left( \frac{x - y}{h} \right) f(y) dy \\
= \int K(u) f(x + uh) du \\
= f(x) + hf^{(1)}(x) \int uK(u) du + h^2 f^{(2)}(x) \kappa_2 + o(h^2) \\
= f(x) + O(h^2 b(x)).
\]

- $\hat{f}_h$ is biased, with a bias proportional to $h^2 f^{(2)}(x)$. 

Non- and Semiparametric Methods
Density Estimation: cont.

Similarly,

\[
\text{Var}\left(\hat{f}_h(x)\right) = \frac{1}{n} \text{Var}\left(\frac{1}{h} K \left(\frac{x - X_i}{h}\right)\right)
\]

\[
\approx \frac{1}{n} \int \frac{1}{h^2} K^2 \left(\frac{x - y}{h}\right) f(y) dy
\]

\[
\approx \frac{1}{nh} f(x) \int K^2 (u) du
\]

\[
\equiv \frac{1}{nh} \nu(x)
\]

Small \( h \) \( \Longrightarrow \) small bias but big variance. Trade off.

\[
\text{MSE}(x) \approx h^2 b^2(x) + (nh)^{-1} \nu(x).
\]

Optimal choice \( h^* \approx (\nu(x)/4b(x)) n^{-1/5}. \)
How to choose the kernel and bandwidth?

- Choice of $K$ is not as important as the choice of $h$.

- Rule-of-thumb: $h = 1.06\sigma n^{-1/5}$.

- Cross-Validation: choose the bandwidth that minimizes an estimator of the IMSE.

- Uniform-in-bandwidth theory is needed.
Density Estimation: Asymptotics

- Consistency: \( MSE(x) \to 0 \) as \( n \to \infty, h \to 0, nh \to \infty \).
- Asymptotic normality: \( nVar \left( \hat{f}_h(x) \right) \) blows up! parametric convergence is not possible. Rate \( \sqrt{nh} \).
- \( \sqrt{nh} \left( \hat{f}_h(x) - f(x) - h^2 b^2(x) \right) \to_d N(0, \nu(x)) \)
- When \( h \approx n^{-1/5} \implies \sqrt{nh} \approx n^{2/5} \) is the fastest rate of convergence; cf. Stone.
Density Estimation: Bias reduction

- If $f$ is smoother, we can use higher-order kernels, i.e.

$$\int u^p \cdot K(u) du = 0,$$

for $p = 1, \ldots, r - 1$, and $\int u^r \cdot K(u) du < \infty$. Then,

$$\mathbb{E} \left( \hat{f}_h(x) \right) = f(x) + O(h^r b_r(x)).$$

Variance is still the same.

- $K$ is negative. Higher-order kernel estimates behaved poorly numerically unless $n$ is very large.
Density Estimation: Multivariate density estimation

- Joint density estimation of \((Y, X)\):

\[
\hat{f}_h(y, x) = \frac{1}{nh^2} \sum_{i=1}^{n} K \left( \frac{y - Y_i}{h} \right) K \left( \frac{x - X_i}{h} \right).
\]

- Bias is still proportional to \(h^2\), but variance is proportional to \(n^{-1}h^{-2}\). Optimal MSE when \(h^* \approx n^{-1/6}\), so \(MSE^* \approx n^{2/6}\).

- More generally, for \(J\)-dimensional, bias\(\approx h^2\), \(Var \approx n^{-1}h^{-J}\), \(h^* \approx n^{-1/(J+4)}\) and \(MSE^* \approx n^{2/(J+4)}\).

- **Curse of dimensionality**: The larger \(J\) the slower the rate of convergence.
Nonparametric regression

- \((Y_i, X_i)\) such that \(Y_i = m(X_i) + \varepsilon_i\), with \(\mathbb{E}(\varepsilon_i \mid X_i) = 0\).
- Define the integrated regression function

\[
R(x) = \mathbb{E} \left( Y_i 1_{(-\infty, x]}(X_i) \right) \\
= \mathbb{E} \left( m(X_i) 1_{(-\infty, x]}(X_i) \right) \\
= \int_{-\infty}^{x} m(z)f(z)dz.
\]

- Note that \(R^{(1)}(x) = m(x)f(x)\) or \(m(x) = R^{(1)}(x)/f(x)\).
- \(R\) can be estimated by the sample analog.
Nonparametric regression estimation: NW

The last expression suggests the estimator

\[ \hat{m}_h(x) : = \frac{R_n(x + h) - R_n(x - h)}{2h} \frac{1}{\hat{f}_h(x)} \]

\[ = \frac{1}{n\hat{f}_h(x)} \sum_{i=1}^{n} \frac{Y_i 1_{x-h \leq X_i \leq x+h}}{2h} \]

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

\[ = \sum_{i=1}^{n} Y_i w_{i,n}(x) . \]

This is the so-called Nadaraya-Watson (NW) estimator.
Properties of NW estimator

- **Bias:** $\mathbb{E}(\hat{m}_h(x)) - m(x) \approx h^2 B(x)$, for some $B(x)$.

- **Variance:** $\text{Var}(\hat{m}_h(x)) \approx (nh)^{-1} V(x)$, for

  $$V(x) = \left[ \int K^2(u) \, du \right] \mathbb{E} \left( \varepsilon_i^2 \mid X_i = x \right) / f(x).$$

- Asymptotic normality holds, and pointwise confidence intervals.

- As before, higher order kernels can be used to reduce bias (numerically bad unless $n$ is large).

- Extension to multivariate regression as before.
How to chose the bandwidth in regression?

- Rule-of-thumb and Plug-in methods: $h_i = a_i n^{-1/(4+d)}$.


- Modified AIC procedure: for small samples performs better than LS, see Li and Racine.
Other estimators: Local Linear

- Another interpretation of NW:
  \[
  \hat{m}_h(x) = \arg \min_m \sum_{i=1}^{n} (Y_i - m)^2 K \left( \frac{x - X_i}{h} \right).
  \]

- More generally,
  \[
  (\hat{M}_h(x), \hat{D}_h(x)) = \arg \min_{M,D} \sum_{i=1}^{n} (Y_i - M - (x - X_i)D)^2 K \left( \frac{x - X_i}{h} \right).
  \]

- One can show \( \hat{M}_h(x) \) consistently estimates \( m(x) \) (and \( \hat{D}_h(x) \) consistently estimates \( \partial m(x)/\partial x \))

- We could also fit a local quadratic, cubic, etc.
Other estimators: Local Linear

Advantages of local polynomial:

- Speeds convergence rates, similar to higher order kernels
- Automatically provides derivative estimates.
- It can behave better near boundaries of the data

Disadvantages:

- It can be more numerically unstable, more sensitive to outliers.
Other estimators: Series

- Assume the support of $X$ is contained in $[0,1]$
- Let $\{\psi_j\}_{j=0}^\infty$ be a complete basis of $L_2([0,1])$, e.g. Legendre polynomials.
- Then, if $m \in L_2([0,1])$, then $m(x) = \sum_{j=1}^{\infty} m_j \psi_j(x)$.
- Estimator $\hat{m}(x) = \sum_{j=1}^{J_n} \hat{m}_j \psi_j(x)$, where $\hat{m}_j$ are estimated by OLS of $Y_i$ on $\psi_1(X_i), ..., \psi_{J_n}(X_i)$.
- Bias $\approx J^{-\alpha}$ and Variance $\approx J/n$, where $\alpha$ depends on the basis and smoothness of $m$.
- For splines or power series $\alpha = r/d$, $r$ is the number of continuous derivatives, and $d$ dimension of $X_i$. 
Other targets and estimators

- **Other sieves**: we aim to estimate the probability of working given $X$, say $p(X)$. With $\Lambda(z) := \exp(z)/(1 + \exp(z))$, define $\hat{p}(z) = \Lambda(\sum_{l=0}^{J_n} \hat{p}_l \psi_l(z))$, where $\hat{p} = (\hat{p}_1, ..., \hat{p}_{J_n})'$ solves

$$
\max_{p \in \mathbb{R}^{J_n}} \sum_{i=1}^{n} Y_i \ln \Lambda \left( \sum_{l=1}^{J_n} p_l \psi_l(X_i) \right) + (1 - Y_i) \ln \left[ 1 - \Lambda \left( \sum_{l=1}^{J_n} p_l \psi_l(X_i) \right) \right]
$$

and where $J_n \to \infty$ as $n \to \infty$.

- Nearest neighbor.

- **Other targets**: conditional distributions, quantiles, etc...
A Semiparametric approach to the labor participation example: \( Y_i = 1(X'_i \theta_0 - \varepsilon_i > 0), \varepsilon_i \perp X_i \), with cdf \( \eta \).

Model: \( \mathbb{E}(Y_i|X_i) = \eta(X'_i \theta_0), \theta_0 \) and \( \eta \) unknown.

Ichimura’s (1993) SLS:

\[
\hat{\theta} = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \hat{\eta}_{i\theta} \right\}^2 t_{ni},
\]

where \( \hat{\eta}_{i\theta} \) is a nonparametric estimator of \( \mathbb{E}(Y_i|X'_i \theta) \), \( \hat{t}_{ni} \) is a trimming sequence, e.g. \( \hat{t}_{ni} = \mathbb{I}(X_i \in A) \).

For example:

\[
\hat{\eta}_{i\theta} = \frac{1}{n h \hat{f}_{h\theta}(X'_i \theta)} \sum_{j=1}^{n} Y_j K \left( \frac{X'_i \theta - X'_j \theta}{h} \right)
\]
Asymptotics of Ichimura’s estimator

- **FOC:**

\[
0 = \sqrt{n} \partial_\theta S_n(\hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ Y_i - \hat{\eta}_{i\theta} \} \partial_\theta \hat{\eta}_{i\theta} \hat{t}_{ni},
\]

where \( \partial_\theta \hat{\eta}_{i\theta} := \partial \hat{\eta}_{i\theta} / \partial \theta |_{\theta = \hat{\theta}}. \)

- **Mean Value Argument:**

\[
\sqrt{n}(\hat{\theta} - \theta_0) = G_n^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ Y_i - \hat{\eta}_{i\theta_0} \} \partial_\theta \hat{\eta}_{i\theta_0} \hat{t}_{ni},
\]

where \( G_n = n^{-1} \sum_{i=1}^{n} \hat{t}_{ni} \partial_\theta \hat{\eta}_{i\hat{\theta}} \partial_\theta' \hat{\eta}_{i\hat{\theta}} \) and \( \bar{\theta} \) is such that \( |\bar{\theta} - \theta_0| \leq |\hat{\theta} - \theta_0| \) a.s.
By the uniform consistency of $\partial_\theta \hat{\eta}_{i\hat{\theta}}$

$$G_n \rightarrow_p \Lambda_0 =: E[\partial_\theta \eta_{i\theta_0} \partial'_\theta \eta_{i\theta_0}],$$

where $\partial_\theta \eta_{i\theta_0} := \partial \eta_{i\theta} / \partial \theta |_{\theta = \theta_0}$.

A Stochastic Equicontinuity argument yields

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \hat{\eta}_{i\hat{\theta}}\} \partial_\theta \hat{\eta}_{i\theta_0} \hat{t}_{ni} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \eta_{i\theta_0}\} \partial_\theta \eta_{i\theta_0} + o_P(1),$$

An application of Linderberg-Lévy CLT then yields

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \Lambda_0^{-1} \Sigma \Lambda_0^{-1}),$$

where $\Sigma = E[var(Y_i|X_i)\partial_\theta \eta_{i\theta_0} \partial'_\theta \eta_{i\theta_0}]$. 
Klein and Spady’s (1993) Estimator

- Binary choice \( Y_i = 1(X_i'\theta_0 - \varepsilon_i > 0) \), \( \varepsilon_i \) independent of \( X_i \), with cdf \( \eta \).

- Semiparametric likelihood estimator:

\[
\hat{\theta} = \arg \max_{\theta \in \Theta} \mathcal{L}_n (\theta) := \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i \log[\hat{\eta}_{i\theta}] + (1 - Y_i) \log[1 - \hat{\eta}_{i\theta}] \right\} \tilde{t}_{in}
\]

where \( \hat{\eta}_{i\theta} \) as above and \( \tilde{t}_{in} \) trimming to avoid zero in \( \log \).
Semiparametric Estimation

FOC:

\[ 0 = \sqrt{n} \partial_{\hat{\theta}} \mathcal{L}_n(\hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ Y_i - \hat{\eta}_{i\hat{\theta}} \right\} \hat{\psi}_{i\hat{\theta}} \tilde{t}_{in} \]

where \( \hat{\psi}_{i\hat{\theta}} := \partial_{\theta} \hat{\eta}_{i\hat{\theta}} [\hat{\eta}_{i\hat{\theta}} (1 - \hat{\eta}_{i\hat{\theta}})]^{-1} \).

Mean Value Argument:

\[ \sqrt{n}(\hat{\theta} - \theta_0) = H_n^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ Y_i - \hat{\eta}_{i\theta_0} \right\} \hat{\psi}_{i\hat{\theta}} \tilde{t}_{in} + o_P(1), \]

for a matrix \( H_n \) satisfying

\[ H_n \to_p \Delta_0 \equiv E[\partial_{\theta} \eta_{i\theta_0} \partial_{\theta}^t \eta_{i\theta_0} [\eta_{i\theta_0} (1 - \eta_{i\theta_0})]^{-1}]. \]

By a SE argument and the CLT

\[ \sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, \Delta_0^{-1}). \]
Linear Regression with Unknown Heteroscedasticity

Model:

\[ Y_i = X_i' \theta_0 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i|X_i] = 0 \text{ a.s.} \]

Robinson’s (1987) optimal GLS estimator:

\[
\hat{\theta}_n = \left( \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}^{-2}(X_i)X_iX'_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}^{-2}(X_i)X_iY_i \right) = \\
\theta_0 + \left( \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}^{-2}(X_i)X_iX'_i \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}^{-2}(X_i)X_i\varepsilon_i \right)
\]

where \( \hat{\sigma}^2(X_i) \) estimates consistently \( \sigma^2(X_i) := \mathbb{E}[\varepsilon_i^2|X_i] \).
Under some mild consistency conditions

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{\sigma}^{-2}(X_i)X_iX_i' = \mathbb{E} \left[ \sigma^{-2}(X_i)X_iX_i' \right] + o_P(1).
\]

Thus we focus on the analysis of

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\sigma}^{-2}(X_i)X_i\varepsilon_i.
\]

By SE argument

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\sigma}^{-2}(X_i)X_i\varepsilon_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma^{-2}(X_i)X_i\varepsilon_i + o_P(1).
\]
Impact of Nuisance Parameters

- In the examples considered so far there is no contribution to the asymptotic distribution of $\hat{\theta}_n$ from estimating nuisance parameters.
- However, in many other examples there is a contribution. This will have an impact on how standard errors and confidence intervals are computed.
- Unfortunately, accounting for this impact is technically difficult.
A Generic Example

- Suppose we aim to estimate $\theta_0 = \mathbb{E}(m(Z, \eta))$, where $\eta$ is an unknown function.

- $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \hat{\eta})$, where $\hat{\eta}$ is a nonpar. estimator of $\eta$.

- Impact from estimating $\eta$?
Main tool: Stochastic Equicontinuity

- Under general conditions, a SE argument yields

\[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(Z_i, \hat{\eta}) - \mathbb{E}(m(Z_i, \hat{\eta})) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(Z_i, \eta) - \mathbb{E}(m(Z_i, \eta)). \]

- or

\[ \sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{m(Z_i, \eta) - \mathbb{E}(m(Z_i, \eta))\} + \Delta, \]

where \( \Delta = \sqrt{n} \left[ \mathbb{E}(m(Z_i, \hat{\eta})) - \mathbb{E}(m(Z_i, \eta)) \right]. \)
Impact of Nuisance Parameters

- **Parametric case:** A Mean Value Argument leads to

  \[ \Delta \approx \sqrt{n} (\hat{\eta} - \eta) \left[ \mathbb{E} (\hat{m}(Z_i, \eta)) \right]. \]

- **Nonparametric case:** Functional derivatives, Riesz's Representation and Bias calculations.
Towards a General Theory

Informally: $\eta \rightarrow \mathbb{E}(m(Z_i, \eta))$ is functional with derivative $V$ (continuous linear operator)

\[
\sqrt{n} \left[ \mathbb{E}(m(Z_i, \hat{\eta})) - \mathbb{E}(m(Z_i, \eta)) \right] \approx \sqrt{n} V(\hat{\eta} - \eta) \\
\approx \sqrt{n} \mathbb{E}(v(Z)(\hat{\eta} - \eta)(Z)) \\
\approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(Z_i, \theta_0, \eta)
\]
Average derivatives: Powell, Stock and Stoker (1989)

- Estimate $\theta_0 := \mathbb{E} \left[ \partial \mathbb{E} \left( Y_i \mid X_i \right) / \partial X_i \right]$.
- $\mathbb{E} \left[ \partial \mathbb{E} \left( Y_i \mid X_i \right) / \partial X_i \right] = \mathbb{E} \left[ \dot{\eta}(X_i' \beta_0)X_i' \right] \beta_0 = k \beta_0$.
- Estimate $\theta_0 = \mathbb{E} [\eta(X_i)]$ by

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\eta}_h(X_i).$$

- AD are generally interpretable, but curse of dimensionality.
Overcoming the curse of dimensionality in nonparametric regression.

Robinson (1988): \( Y_i = X_i' \theta_0 + \eta(Z_i) + \varepsilon_i. \)

\[
\hat{\theta} = \left( \mathbb{E}_n \left[ \hat{X}_Z' \hat{X}_Z \right] \right)^{-1} \mathbb{E}_n \left[ \hat{X}_Z \hat{Y}_Z \right],
\]

where \( \hat{W}_Z = W - \hat{E} [W|Z]. \)

Other alternatives: additive regression, functional coefficients, transformation models...
Policy or Treatment Evaluation: CATE

- $D$ is the treatment indicator.
- $Y(1)$ is the outcome under treatment.
- $Y(0)$ is the outcome without treatment.
- **Missing** data problem: We only observe $Y = Y(1) \cdot D + Y(0) \cdot (1 - D)$.
- Treatment is **unconfounded**: $(Y(1), Y(0))$ independent of $D$, conditional on $X$. Trade off identifi-flex.
Policy or Treatment Evaluation: CATE

Parameter of interest: \( \tau(X) := \mathbb{E}(Y(1) - Y(0)|X) \).

Propensity score: \( p(X) = \mathbb{E}(D|X) \).

Note that by the unconfoundedness,

\[
\tau(X) = \mathbb{E}[Y(1)|X, D = 1] - \mathbb{E}[Y(0)|X, D = 0] \\
= \mathbb{E}[Y|X, D = 1] - \mathbb{E}[Y|X, D = 0].
\]
Moreover, the ATE (cf. Rosenbaum and Rubin, 1983)

\[
\tau_{ATE} = \mathbb{E}[\tau(X)]
= \mathbb{E} \left[ \frac{YD}{p(X)} - \frac{Y(1-D)}{1-p(X)} \right].
\]

This suggests the estimator

\[
\hat{\tau}_{ATE} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_iD_i}{\hat{p}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{p}(X_i)},
\]

where \( \hat{p} \) is the SLE (cf. Hirano, Imbens and Ridder, 2003).
Suppose $Y = Y^* D \equiv H(X^\top \theta_0, e) D$, where we do not observe $Y^*$ when $D = 0$.

Nonparametric selection: $D = \mathbb{1}[g_0(X) - u \geq 0]$.

The unobserved errors $e$ and $u$, though independent of $X$, are correlated with each other.

Define $g_0(X) \equiv E(D|X)$. Then,

$$E(Y|X) = F[X^\top \theta_0, g_0(X)].$$
Suppose \( Y = H(Z^\top \alpha_0 + W \gamma_0, e) \) where \( W \) is an endogenous regressor with \( W = s_0(Z) + u \), and \( e \) and \( u \) are unobserved correlated error terms.

Assume a ‘control function’ approach: \( e|Z, u \sim e|u \).

Define \( g_0(X) := W - s_0(Z) \). Then,

\[
\mathbb{E}(Y|X) = F[X^\top \theta_0, g_0(X)].
\]
Endogenous Variables: Nonparametric IV

- Model:
  \[ Y = g(X) + \varepsilon, \quad \mathbb{E}[\varepsilon|Z] = 0. \]

- The variable of interest \( X \) is endogenous, in the sense that \( \mathbb{E}[\varepsilon|X] \neq 0 \). \( Z \) is an instrument.

- Nonparametric extension of classical linear IV, where \( g(X) = \theta'_0 X \).

- How to estimate \( g \)? Difficult problem.
Endogenous Variables: Nonparametric IV

- The exclusion restriction yields
  \[ \mathbb{E}[Y|Z] = \mathbb{E}[g(X)|Z]. \]

- Multiply both sides by \( f_Z(z) \), the density of \( Z \), to obtain
  \[ m(z) = Ag(z), \]
  where \( m(z) := \mathbb{E}[Y|Z = z]f_Z(z) \), and \( A \) is the integral operator
  \[ Ag(z) = \int_0^1 g(x)f(x,Z)(x,z)dx, \]

- Estimator of \( g \), from \( A^*m(z) = A^*Ag(z) \),
  \[ \hat{g} = \left( \hat{A}^*\hat{A} \right)^{-1} \hat{A}\hat{m}. \]
Endogenous Variables: Nonparametric IV

Horowitz (2011, 2012):

\[ \hat{g}_n(x) = \sum_{j=1}^{J_n} \hat{g}_j \psi_j(x), \]

where \( G = (\hat{g}_1, \ldots, \hat{g}_{J_n})' = (Z_n'X_n)^{-1}(Z_n'Y_n)^{-1} \), where \( Z_n \) is the \( n \times J_n \) matrix with elements \( \psi_j(Z_i) \), \( X_n \) is the \( n \times J_n \) matrix with elements \( \psi_j(X_i) \) and \( Y_n = (Y_1, \ldots, Y_n)' \).

Implementation: standard IV estimator.
Two references on implementation

J.S. Racine


H. Ichimura and P. Todd

*Implementing nonparametric and semiparametric estimators*  
Application to Job Training Evaluation

- Data from Dehejia and Wahba (1999), originally from LaLonde (1986).
- Variables: age, educ, black, hisp, married, nodegr, re74, re75, re78, u74, u75, treat.
- Denote $m_j(X) := \mathbb{E}[Y|X, D = j]$. 

Non- and Semiparametric Methods
Application to Job Training Evaluation

- Estimate the ATE based on:
  - Parametric specification:
    \[
    m_j(x, \delta_j) = \delta_{j;1} + \delta_{j;2} \text{age} + \delta_{j;3} \text{age}^2 + \delta_{j;4} \text{educ} + \delta_{j;5} \text{educ}^2 \\
    + \delta_{j;6} \text{black} + \delta_{j;7} \text{hisp} + \delta_{j;8} \text{married} + \delta_{j;9} \text{nodegr} \\
    + \delta_{j;10} \text{re74} + \delta_{j;11} \text{re74}^2 + \delta_{j;12} \text{re75} + \delta_{j;13} \text{re75}^2 \\
    + \delta_{j;14} \text{u74} + \delta_{j;15} \text{u75}. \ j = 0, 1.
    \]
  - Using local linear estimators for \( m_1(\cdot) \) and \( m_0(\cdot) \):
    \[
    m_j(x) := m_j(\text{age,educ,black,hisp,married,nodegr,re74,re75}),
    \]
    for \( j = 0, 1. \)
    Also compute 95% bootstrap confidence interval.
# Application to Job Training Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{\tau}_{ATE}$</th>
<th>$\hat{\tau}_{ATT}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric:</td>
<td>1617.969</td>
<td>1843.092</td>
</tr>
<tr>
<td></td>
<td>(466.8807,2826.3136)</td>
<td>(506.8876,3117.5643)</td>
</tr>
<tr>
<td>Nonparametric:</td>
<td>2076.923</td>
<td>2440.282</td>
</tr>
<tr>
<td></td>
<td>(302.3463,2720.8618)</td>
<td>(306.0942,3235.5631)</td>
</tr>
</tbody>
</table>
Application to Job Training Evaluation

Compute the propensity score-based estimator of the ATE based on:

\[ p(x, \delta) = \Lambda(\delta_1 + \delta_2 \text{age} + \delta_3 \text{age}^2 + \delta_4 \text{educ} + \delta_5 \text{educ}^2 + \delta_6 \text{black} + \delta_7 \text{hisp} + \delta_8 \text{married} + \delta_9 \text{nodegr} + \delta_{10} \text{re74} + \delta_{11} \text{re74}^2 + \delta_{12} \text{re75} + \delta_{13} \text{re75}^2 + \delta_{14} \text{u74} + \delta_{15} \text{u75}), \]

\[ \hat{\tau}_{ATE} = 1466.449 \text{ with a 95\% CI (200.976,3108.783).} \]
Application to Job Training Evaluation

Histogram of 'Propensity Score' Parametric

Histogram of 'Propensity Score' Nonparametric

Non- and Semiparametric Methods
An example: Binary Choice with Endogeneity

- Binary choice model: $Y = \mathbb{I}[Z^\top \alpha_0 + W\gamma_0 - e > 0]$ and

$$
\mathbb{E}(Y|X) = F[Z^\top \alpha_0 + W\gamma_0, g_0(X)],
$$

where $F$ is the conditional cdf of $e$ given $u$.

- $g_0(X) := W - \mathbb{E}(W|Z)$.

- See Blundell and Powell (2004) and Rothe (2009).

Application to Migration

- EJL(2011): model migration decisions ($Y$ equals 1 if migrates or 0 otherwise) with an endogenous regressor ($W=$income) and other exogenous covariates ($Z=$[state,education,family size,age]).
- SLS Estimator:

$$\hat{\beta} := \arg \min_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i - \hat{F}_i(W_i(\beta, \hat{g}) | \beta, \hat{g}) \right]^2 \hat{a}_i.$$ 

- Data from 1990 wave of PSID: $n=4582$. 

Non- and Semiparametric Methods
Figure 6: Nonparametric Age Effects on Labor Income and Migration Probabilities
Binary Choice with Endogeneity: Application to Migration

\[ Y = \mathbb{I}(\text{State} + \beta_{0;1}\text{Edu} + \beta_{0;2}\text{Size} + \beta_{0;3}\text{Age} + \beta_{0;4}\log(\text{Income}) - e \geq 0) \]

\[ \log(\text{Income}) = g_0(\text{State, Edu, Size, Age}) + u \]

\[ \hat{g}(\text{State, Edu, Size, Age}) \] uses generalized kernels as in Racine & Li (2004). Bandwidths chosen by Least-Squares cross-validation.

(I) Probit; (II) Ichimura’s (1993) SLS; (III) 2-Steps SLS.
## Estimated Marginal Effects

<table>
<thead>
<tr>
<th>Variable</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edu*</td>
<td>-0.0052 (0.0119)</td>
<td>-0.0162 (0.0095)</td>
<td>0.0329 (0.0106)</td>
</tr>
<tr>
<td>0 - 1</td>
<td>-0.0052 (0.0119)</td>
<td>-0.0166 (0.0296)</td>
<td>0.0355 (0.0114)</td>
</tr>
<tr>
<td>Size*</td>
<td>0.0028 (0.0035)</td>
<td>-0.0018 (0.0034)</td>
<td>0.0115 (0.0044)</td>
</tr>
<tr>
<td>1 - 2</td>
<td>0.0028 (0.0033)</td>
<td>-0.0017 (0.0305)</td>
<td>0.0036 (0.0110)</td>
</tr>
<tr>
<td>2 - 3</td>
<td>0.0028 (0.0034)</td>
<td>-0.0018 (0.0301)</td>
<td>0.0081 (0.0109)</td>
</tr>
<tr>
<td>3 - 4</td>
<td>0.0028 (0.0035)</td>
<td>-0.0018 (0.0296)</td>
<td>0.0124 (0.0112)</td>
</tr>
<tr>
<td>Age*</td>
<td>-0.0039 (0.0005)</td>
<td>-0.0037 (0.0017)</td>
<td>-0.0067 (0.0018)</td>
</tr>
<tr>
<td>log(Income)*</td>
<td>-0.0322 (0.0060)</td>
<td>-0.0617 (0.0270)</td>
<td>-0.0570 (0.0149)</td>
</tr>
</tbody>
</table>

Non- and Semiparametric Methods
Conclusions

- Trade off between Flexibility and Efficiency.
- Semiparametric models are a compromise.
- Implementation is easier than in the past.
- Even if parametric is preferred, check semiparametric.
More

- Semi-nonparametric estimation (NPMLE, Sieves).
- Semiparametric Efficiency.
- Choice of smoothing parameters in Semiparametrics.
- Hypothesis testing.
- ...
Some Math: A Formal Theory


Model: \( \theta_0 \in \Theta, \Theta \subset \mathbb{R}^{d_\theta}, h_0 \in \mathcal{H}, \) s.t.

\[
\mathbb{E}[m(Z, \theta_0, h_0(Z))] = 0,
\]  

(1)

where \( m(\cdot, \theta, h(\cdot)) \) from \( \mathbb{R}^{d_Z} \) to \( \mathbb{R}^{d_m} \), for each \( \gamma := (\theta, h(\cdot)) \in \Gamma := \Theta \times \mathcal{H}. \)

\( h_0(\cdot) \) may contain \( \theta \) as additional arguments. We assume there is a first-step nonparametric estimator \( \hat{h}(\cdot) \) for \( h_0(\cdot) \) available with certain convergence properties as specified in Assumption A1 and A2 below.
A Formal Theory

- Define the norms $\|g\|_\infty = \sup_{Z \in \mathcal{Z}} |g(Z)|$ and $\|g\|_r := (\mathbb{E} [|g(Z)|^r])^{1/r}$, where $\text{calZ}$ is the support of $Z$.

- The function space $\mathcal{H}$ is endowed with a pseudo-metric $\| \cdot \|_\mathcal{H}$ which is a sup-norm metric with respect to the $\theta$-arguments and a pseudo-metric with respect to $z$. For example, $\|h\|_\mathcal{H} := \sup_\theta \|h(\cdot, \theta)\|_\infty$ or $\|h\|_\mathcal{H} := \sup_\theta \|h(\cdot, \theta)\|_r$. 
Let $M_n(\theta, h) := \mathbb{E}_n[m(Z, \theta, h(Z))]$, which is the sample analog of $M(\theta, h) := \mathbb{E}[m(Z, \theta, h(Z))]$.

$\|A\| := (\text{tr}(A'\gamma A))^{1/2}$ for some fixed symmetric pd $\gamma$.

Estimator

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \|M_n(\theta, \hat{h})\|.$$
**Assumption A1:** Suppose that \( \theta_0 \in \Theta \) satisfies \( M(\theta_0, h_0) = 0 \). In addition, assume

(i) The estimator \( \hat{\theta} \) satisfies

\[
\left\| M_n \left( \hat{\theta}, \hat{h} \right) \right\| \leq \inf_{\theta \in \Theta} \left\| M_n \left( \theta, \hat{h} \right) \right\| + o_P(1). \tag{2}
\]

(ii) Identification: for all \( \varepsilon > 0 \), there exists \( \eta(\varepsilon) > 0 \) such that

\[
\inf_{\theta: |\theta - \theta_0| \geq \varepsilon} \left\| M(\theta, h_0) \right\| \geq \left\| M(\theta_0, h_0) \right\| + \eta(\varepsilon). \tag{3}
\]

(iii) Uniform Continuity: uniformly for all \( \theta \in \Theta \), \( M(\theta, h) \) is continuous in \( h \) at \( h = h_0 \) with respect to the metric \( \| \cdot \|_H \).

(iv) \( \left\| \hat{h} - h_0 \right\|_H = o_P(1) \).

(v) Uniform convergence: for all sequences of positive numbers \( \{\delta_n\} \to 0 \),

\[
\sup_{\theta \in \Theta, \|h - h_0\|_H \leq \delta_n} \left\| M_n(\theta, h) - M(\theta, h) \right\| = o_P(1). \tag{4}
\]
Consistency

Theorem

Under Assumption A1, it holds that $|\hat{\theta} - \theta_0| = o_P(1)$. 
Asymptotic Normality: Functional Derivatives

Define \( \Theta_\delta := \{ \theta \in \Theta : |\theta - \theta_0| \leq \delta \} \) and
\( \mathcal{H}_\delta := \{ h \in \mathcal{H} : \| h - h_0 \|_{\mathcal{H}} \leq \delta \} \)

For each \( \theta \in \Theta_\delta \), we say that \( M(\theta, h) \) is pathwise differentiable at \( h \in \mathcal{H}_\delta \) in the direction \( [\bar{h} - h] \) if\( \{ h + \lambda (\bar{h} - h) : \lambda \in [0, 1] \} \subset \mathcal{H} \) and

\[
\lim_{\lambda \to 0} \frac{M(\theta, h + \lambda (\bar{h} - h)) - M(\theta, h)}{\lambda} \text{ exists;}
\]

the derivative is denoted as \( V_h(\theta, h) [\bar{h} - h] \). For the weak convergence we need the following assumptions.
**Assumption A2:** Suppose that \( \theta_0 \) is in interior of \( \Theta \), and that \(|\hat{\theta} - \theta_0| = o_P(1)\). In addition, assume

(i) The estimator \( \hat{\theta} \) also satisfies

\[
\| M_n \left( \hat{\theta}(\tau), \hat{h} \right) \| \leq \inf_{\theta \in \Theta_\delta} \| M_n \left( \theta, \hat{h} \right) \| + o_P \left( n^{-1/2} \right). \tag{5}
\]

(ii) Smoothness in \( \theta \): the map \( \theta \to M(\theta, h_0) \) is continuously differentiable at \( \theta_0 \), with derivative \( V_{\theta_0} \) that is of full rank.

(iii) Smoothness in \( h \): for each \( \theta \in \Theta_\delta \), the pathwise derivative \( V_h(\theta, h_0) [h - h_0] \) of \( M(\theta, h) \) at \( h = h_0 \) exists in all directions \([h - h_0] \in \mathcal{H} \); and for all \((\theta, h) \in \Theta_\delta \times \mathcal{H}_\delta \) with a positive sequence \( \delta_n \to 0 \), it holds that

\[
\| M(\theta, h) - M(\theta, h_0) - V_h(\theta, h_0) [h - h_0] \| \leq c \| h - h_0 \|_{\mathcal{H}}^2 \tag{6}
\]

for a constant \( c > 0 \), and

\[
\| V_h(\theta, h_0) [h - h_0] - V_h(\theta_0, h_0) [h - h_0] \| \leq o(1) \delta_n. \tag{7}
\]
(iv) \( \Pr \left( \hat{h} \in \mathcal{H} \right) \to 1 \), and \( \left\| \hat{h} - h_0 \right\|_{\mathcal{H}} = o_P \left( n^{-1/4} \right) \).

(v) Stochastic Equicontinuity: for all sequences of positive numbers \( \delta_n \to 0 \),

\[
\sup_{|\theta - \theta_0| \leq \delta_n, \|h - h_0\|_{\mathcal{H}} \leq \delta_n} \left\| M_n (\theta, h) - M (\theta, h) - M_n (\theta_0, h_0) \right\| = o_P \left( n^{-1/2} \right). \tag{8}
\]

(vi) \( \sqrt{n} V_h (\theta_0, h_0) \left[ \hat{h} - h_0 \right] \) admits an asymptotic expansion:

\[
\sqrt{n} V_h (\theta_0, h_0) \left[ \hat{h} - h_0 \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi (Z_i, \theta_0, h_0) + o_P (1),
\]

and the function \( s (Z, \theta_0, h_0, \tau) := m (Z, \theta_0, h_0) + \phi (Z, \theta_0, h_0) \) satisfies a CLT.
Theorem

Under Assumption A1 and A2, \( \hat{\theta} \) is CAN with

\[
\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \Sigma),
\]

where \( \Sigma = A KA', \ A := [V'_{\theta_0} \gamma V_{\theta_0}]^{-1} V'_{\theta_0} \gamma \) and

\( K := E[s(Z_i, \theta_0, h_0) s(Z_i, \theta_0, h_0)'] \).
Some References: Surveys

Some References: Books

Some References: Books