Nonlinear Models of Econometric Analysis

September 2011

Introduction

- Linear econometric models are widely popular in economics.
- Most people run OLS and 2SLS.
- However, there are questions not addressed by OLS or 2SLS.
- Linear Models might be misspecified.
- Alternatives: nonlinear models, nonparametric models, semiparametric models.
- Some nonlinear models can be implemented in Stata: for example quantile regression, discrete choice models
- Some can not, for example auction models, dynamic discrete choice models, nonlinear models of demand and oligopolistic competition.

Introduction

- Nonlinear Models are challenging, in terms of both numerical implementation and econometric (statistical analysis).
- Econometric analysis focuses more on the statistical properties of nonlinear models.
- But numerical implementation is equally, if not more, difficult!
- Some nonlinear models, such quantile regression and discrete choice models, can be computed as efficiently as linear models. However, other models are far more difficult.
- Ken Judd's "Numerical Methods in Economics" is a good starting point.

Examples of the difficulty of numerical implementation:

- (Knittel)

  http://www.nber.org/papers/w14080
- (Judd)

  http://economics.uchicago.edu/Skrainka-HighPerformanceQuad.pdf

Course Materials

- Current course materials are posted at:

  `http://www.stanford.edu/~doubleh/eco273`

  `http://www.stanford.edu/~doubleh/condensedcourse/`

Outline of Materials

- Review of General Nonlinear Estimator Theory
- Nonparametric Regression, Application to Auctions
- Quantile Regression
- Simulation, Computation, Markov Chain Monte Carlo (MCMC) and Bayesian Methods
- Bootstrap and Subsampling
- Time permitting: Treatment Effect Models

Lecture 2: Consistency of M-estimators

Instructor: Han Hong

Department of Economics
Stanford University

Prepared by Wenbo Zhou, Renmin University

References

- Takeshi Amemiya, 1985, Advanced Econometrics, Harvard University Press
- Newey and McFadden, 1994, Chapter 36, Volume 4, The Handbook of Econometrics.

- Distinction between global and local consistency.
- Global condition: If Θ is compact,
  - $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$,
  - $Q(\theta) < Q(\theta_0)$ for $\theta \neq \theta_0$,

  then $\hat{\theta} \xrightarrow{p} \theta_0$, where $\hat{\theta} = argmax_{\theta \in \Theta} Q_n(\theta)$
- Local condition: If N is a neighborhood around θ_0,
  - $\sup_{\theta \in N} \left| \frac{\partial Q_n(\theta)}{\partial \theta} - \frac{\partial Q(\theta)}{\partial \theta} \right| \xrightarrow{p} 0$,
  - $Q(\theta) < Q(\theta_0)$ for $\theta \neq \theta_0$ and $\theta \in N$,

  then $\inf_{\theta \in \hat{\Theta}} ||\theta - \theta_0|| \xrightarrow{p} 0$, where $\hat{\Theta}$ denotes the set of θ for which $\frac{\partial Q_n(\theta)}{\partial \theta} = 0$.
- For the local consistency condition, check

$$(1) \ \frac{\partial Q(\theta_0)}{\partial \theta} = 0 \text{ and } (2) \ \frac{\partial^2 Q(\theta_0)}{\partial \theta \partial \theta'} \text{ negative definite.}$$

- Let $L(y_1, \ldots, y_n, \theta)$ be the JOINT density for i.i.d data $y_1, \ldots, y_n$, then
$$Q_n(\theta) \equiv \frac{1}{n} \log L(y_1, \ldots, y_n, \theta) = \frac{1}{n} \sum_{t=1}^{n} \log f(y_t, \theta).$$
- Change assumptions to
  - θ_0 is identified, i.e. $\theta \neq \theta_0 \Rightarrow f(y_t, \theta) \neq f(y_t, \theta_0)$,
  - $E \sup_{\theta \in \Theta} |\log f(y; \theta)| < \infty$.
- Identification implies $Q(\theta) < Q(\theta_0)$ since
$$E \frac{\log f(y; \theta)}{\log f(y; \theta_0)} < \log E \frac{f(y; \theta)}{f(y; \theta_0)} = \log \int f(y; \theta) \, dy = \log 1 = 0.$$
- Condition 2 is a dominance condition for stochastic equicontinuity.
- MLE consistency holds even if you have a parameter dependent support of the data.

- In general case when y_t is not i.i.d,

$$E \log L(y_1, \ldots, y_n; \theta) \leq \log EL(y_1, \ldots, y_n; \theta_0)$$

still holds but to justify the strict $<$ is harder.

- When global condition fails or Θ is not compact, local condition may hold.

- Example: Mixture of normal distributions.

$$y_t \sim \lambda N\left(\mu_1, \sigma_1^2\right) + (1-\lambda) N\left(\mu_2, \sigma_2^2\right),$$

$$L = \prod_{t=1}^{n} \left[\frac{\lambda}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(y_t - u_1)^2}{2\sigma_1^2} \right) + \frac{1-\lambda}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(y_t - u_2)^2}{2\sigma_2^2} \right) \right].$$

Set $u_1 = y_1$ and let $\sigma_1 \to 0$, then L increases to ∞. Hence global MLE cannot be consistent, but local MLE is.

- $Q_n(\theta) = g_n(\theta)' W g_n(\theta)$, for $g_n(\theta) = \frac{1}{n}\sum_{t=1}^n g(z_t, \theta)$, and W is the positive definite weighting matrix. If
  - $\sup_{\theta\in\Theta} |g_n(\theta) - Eg(z_t, \theta)| \xrightarrow{p} 0$,
  - $Eg(z_t, \theta) = 0$ iff $\theta = \theta_0$,

  then $\hat{\theta} \equiv argmax_\theta Q_n(\theta) \xrightarrow{p} 0$.
- Global identification in nonlinear GMM model is usually difficult and "assumed".
- But identification in linear models usually reduces to condition that the sample var-cov matrix for regressors is full rank, i.e
  - $Ex_t x_t'$ for iid models,
  - $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n x_t x_t'$ for fixed regressors.
- For least square, $\frac{1}{n}\sum_{t=1}^n (y_t - x_t'\beta)^2 \xrightarrow{p} E(y - x'\beta)^2$. Iff $Ex_t x_t'$ full rank,

$$E(y - x'\beta)^2 - E(y - x'\beta_0)^2 = E[x'(\beta - \beta_0)]^2$$
$$= (\beta - \beta_0)' Ex_t x_t' (\beta - \beta_0) > 0 \quad \text{if} \quad \beta \neq \beta_0.$$

- Conditional τth quantile of y_t given x_t is a linear regression function $x_t'\beta_0$, i.e. $Pr\left(y_t \leq x_t'\beta_0 | x_t\right) \equiv F_y\left(x_t'\beta_0 | x_t\right) = \tau$.
- The $\tau = \frac{1}{2}$th quantile is the median.
- Population moment condition:

$$E\left(\tau - 1\left(y_t \leq x_t'\beta_0\right)\right) x_t = E\left(\tau - Pr\left(y_t \leq x_t'\beta_0 | x_t\right)\right) x_t = 0.$$

- Sample moment condition:

$$\begin{aligned} 0 &\approx \frac{1}{n}\sum_{t=1}^{n} x_t \left(\tau - 1\left(y_t \leq x_t'\hat{\beta}\right)\right) \\ &= \frac{1}{n}\sum_{t=1}^{n} x_t \left[\tau 1\left(y > x_t'\hat{\beta}\right) - (1-\tau) 1\left(y_t \leq x_t'\hat{\beta}\right)\right]. \end{aligned}$$

- Integrate the condition back to obtain the convex objective function $Q_n(\beta)$.

- Objective function for QR:

$$\begin{aligned} Q_n(\beta) =& \frac{1}{n}\sum_{t=1}^{n}\left[\tau - 1\left(y_t \leq x_t'\beta\right)\right]\left(y_t - x_t'\beta\right) \\ =& \frac{1}{n}\sum_{t=1}^{n}\left[\tau\left(y_t - x_t'\beta\right)^{+} + (1-\tau)\left(y_t - x_t'\beta\right)^{-}\right] \end{aligned}$$

- When $\tau = \frac{1}{2}$, $Q_n(\beta) = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - x_t'\beta\right|$ becomes the Least Absolute Deviation (LAD) regression, which looks for the conditional median.

- Also, that $Ex_t x_t'$ is full rank implies global consistency for the linear quantile regression model.

$Q_n(\beta)$ for QR has two features:

- $Q_n(\beta)$ is convex so that pointwise convergence is sufficient for uniform convergence over compact Θ and the parameter space does not have to be compact.
- No moment conditions are needed for y_t to obtain pointwise convergence, this is done by subtracting $Q_n(\beta_0)$, and $Q_n(\beta) - Q_n(\beta_0) \xrightarrow{p} Q(\beta) - Q(\beta_0)$, by applying triangular inequality.

Concavity and noncompact parameter set: when $Q_n(\theta)$ is concave for maximization (or convex for minimization), then

- pointwise convergence $\Rightarrow$ uniform convergence.
- $Q(\theta)$'s local maximization $\Rightarrow$ global consistency.

Uniform Convergence (in probability)

- Definition: $\hat{Q}(\theta)$ converges in probability to $Q(\theta)$ uniformly over the compact set $\theta \in \Theta$ if

$$\forall \epsilon > 0, \lim_{T \to \infty} P\left(\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| > \epsilon\right) = 0.$$

- Consistency of M-Estimators: If
  - $Q_T(\theta)$ converges in probability to $Q(\theta)$ uniformly,
  - $Q(\theta)$ continuous and uniquely maximized at θ_0,
  - $\hat{\theta} = argmax Q_T(\theta)$ over compact parameter set Θ,

  plus continuity and measurability for $Q_T(\theta)$, then $\hat{\theta} \xrightarrow{p} \theta_0$.

- Consistency of estimated var-cov matrix: Note that it is sufficient for uniform convergence to hold over a shrinking neighborhood of θ_0.

First think about sequence of deterministic functions $f_n(\theta)$.

- Uniform Equicontinuity for $f_n(\theta)$:

$$\lim_{\delta \to 0} \sup_n \sup_{|\theta' - \theta| < \delta} |f_n(\theta') - f_n(\theta)| = 0.$$

- What if $f_n(\theta)$ may be discontinuous but the size of the jump goes to 0?
- Asymptotic uniform equicontinuity for $f_n(\theta)$:

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \sup_{|\theta' - \theta| < \delta} |f_n(\theta') - f_n(\theta)| = 0.$$

- Uniform convergence of $f_n(\theta)$:
  Θ compact, $\sup_{\theta \in \Theta} |f_n(\theta)| \longrightarrow 0$ if and only if $f_n(\theta) \longrightarrow 0$ for each θ and f_n is asymptotically uniformly equicontinuous.

Then the stochastic case $Q_n(\theta)$.

- Definition:

  A sequence of random functions $Q_n(\theta)$ is stochastic uniform equicontinuity if $\forall \epsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\left(\sup_{|\theta - \theta'| < \delta} |Q_n(\theta) - Q_n(\theta')| > \epsilon \right) = 0.$$

- Uniform convergence in probability:

  If $Q_n(\theta) \xrightarrow{p} 0$ for each θ, and $Q_n(\theta)$ is stochastic equicontinuous on $\theta \in \Theta$ compact, then

$$\sup_{\theta \in \Theta} |Q_n(\theta)| \xrightarrow{p} 0.$$

- Simple sufficient condition for stochastic equicontinuity.
  - where the objective function is smooth, differentiable, etc.

- Lipschitz condition: For $\forall \theta, \theta' \in \Theta$, if
$$|Q_n(\theta) - Q_n(\theta')| \leq B_n d(\theta, \theta'),$$
where $\lim_{\delta \to 0} \sup_{|\theta - \theta'| < \delta} d(\theta, \theta') = 0$ and $B_n = O_p(1)$, then $Q_n(\theta)$ is stochastic equicontinuous.

- Example: Suppose $Q_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} f(z_t, \theta)$, z_t iid, $f(z_t, \theta)$ differentiable with $f_\theta(z_t, \theta)$, then by Taylor, for $\bar{\theta} \in (\theta, \theta')$,
$$|Q_n(\theta) - Q_n(\theta')| \leq \frac{1}{n} \sum_{t=1}^{n} |f_\theta(z_t, \bar{\theta})| |\theta - \theta'|.$$
If $b(z_t) = \sup_{\theta \in \Theta} f_\theta(z_t, \theta)$ is such that $Eb(z_t) < \infty$, then the Lipschitz condition holds with $B_n = \frac{1}{n} \sum_{t=1}^{n} b(z_t)$.

- But what to do when the Lipschitz condition is not applicable?
- Uniform WLLN

  Θ compact, y_t iid, $g(y_t, \theta)$ continuous in θ for each y_t a.s., $Eg(y_t, \theta) = 0$, $E \sup_{\theta \in \Theta} |g(y_t, \theta)| < \infty$, then $\forall \epsilon > 0$,

$$\lim_{n \to \infty} P\left(\sup_{\theta \in \Theta} |\frac{1}{n} \sum_{t=1}^{n} g(y_t, \theta)| > \epsilon \right) = 0.$$

Proof: Use pointwise convergence + stochastic equicontinuity.

1. $E \sup_{\theta \in \Theta} |g(y_t, \theta)| < \infty \Longrightarrow E|g(y_t, \theta)| > \infty$ for each θ, so use SLLN 2 to conclude $\frac{1}{n}\sum_{t=1}^{n} g(y_t, \theta) \overset{a.s.(p)}{\longrightarrow} 0$ for each θ.
2. Verify stochastic equicontinuity for $\frac{1}{n}\sum_{t=1}^{n} g(y_t, \theta)$:

$$\sup_{|\theta - \theta'| < \delta} |\frac{1}{n}\sum_{t=1}^{n} g(y_t, \theta) - g(y_t, \theta')|$$

$$\leq \sup_{|\theta - \theta'| < \delta} \frac{1}{n}\sum_{t=1}^{n} |g(y_t, \theta) - g(y_t, \theta')|$$

$$\leq \frac{1}{n}\sum_{t=1}^{n} \sup_{|\theta - \theta'| < \delta} |g(y_t, \theta) - g(y_t, \theta')|.$$

Therefore

$$\begin{aligned}
&\lim_{\delta\to 0}\limsup_{n\to\infty} P\left(\sup_{|\theta-\theta'|<\delta}|\frac{1}{n}\sum_{t=1}^{n} g\left(y_t,\theta\right)-g\left(y_t,\theta'\right)|>\epsilon\right)\\
\leq &\lim_{\delta\to 0}\limsup_{n\to\infty} P\left(\frac{1}{n}\sum_{t=1}^{n}\sup_{|\theta-\theta'|<\delta}|g\left(y_t,\theta\right)-g\left(y_t,\theta'\right)|>\epsilon\right)\\
\leq &\lim_{\delta\to 0}\limsup_{n\to\infty}\frac{E\sum_{t=1}^{n}\sup_{|\theta-\theta'|<\delta}|g\left(y_t,\theta\right)-g\left(y_t,\theta'\right)|}{n\epsilon}\\
= &\lim_{\delta\to 0} E\sup_{|\theta-\theta'|<\delta}|g\left(y_t,\theta\right)-g\left(y_t,\theta'\right)|
\end{aligned}$$

Finally use (uniform b/o compact Θ) continuity of $g\left(y_t,\theta\right)$ and DOM. Since $\lim_{\delta\to 0}\sup_{|\theta-\theta'|<\delta}|g\left(y_t,\theta\right)-g\left(y_t,\theta'\right)|$ almost surely, and
$E\sup_{\delta}\sup_{|\theta-\theta'|<\delta}|g\left(y_t,\theta\right)-g\left(y_t,\theta'\right)|<E2\sup_{\theta}|g\left(y_t,\theta\right)|<\infty$.

Lecture 3: Asymptotic Normality of M-estimators

Instructor: Han Hong

Department of Economics
Stanford University

Prepared by Wenbo Zhou, Renmin University

References

- Takeshi Amemiya, 1985, Advanced Econometrics, Harvard University Press
- Newey and McFadden, 1994, Chapter 36, Volume 4, The Handbook of Econometrics.

The General Framework

- Everything is just some form of first order Taylor Expansion:

$$\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = 0 \Longleftrightarrow \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + \sqrt{n}\left(\hat{\theta} - \theta_0\right)\frac{\partial^2 Q_n(\theta^*)}{\partial \theta \partial \theta'} = 0.$$

$$\begin{aligned}\sqrt{n}\left(\hat{\theta} - \theta_0\right) &= -\left(\frac{\partial^2 Q_n(\theta^*)}{\partial \theta \partial \theta'}\right)^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \\ &\overset{LD}{=} -\left(\frac{\partial^2 Q(\theta_0)}{\partial \theta \partial \theta'}\right)^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \xrightarrow{d} N\left(0, A^{-1}BA^{-1}\right)\end{aligned}$$

where

$$A = E\left(\frac{\partial^2 Q(\theta_0)}{\partial \theta \partial \theta'}\right),\ B = Var\left(\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta}\right)$$

- In MLE, $\frac{\partial Q_n(\theta)}{\partial \theta} = \frac{1}{n}\frac{\partial \log L(\theta)}{\partial \theta}$. $\frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta'} = \frac{1}{n}\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'}$.
- Information matrix:

$$E\frac{\partial^2 \log L(\theta_0)}{\partial \theta \partial \theta'} = -E\frac{\partial \log L(\theta_0)}{\partial \theta}\frac{\partial \log L(\theta_0)}{\partial \theta'}.$$

  by using interchange of integration and differentiation.

- So $A = -B$, and

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, -A^{-1}\right) = N\left(0, \left(-\lim \frac{1}{n} E\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'}\right)^{-1}\right).$$

- What if interchanging integration and differentiation is not possible?
- Example: If $y \in (\theta, \infty)$, then $E\frac{\partial \log f(y;\theta)}{\partial \theta} = f(\theta)$.

- $Q_n(\theta) = g_n(\theta)' W g_n(\theta)$, $g_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} g(z_t, \theta)$.
- Asymptotic normality holds when the moment functions only have first derivatives.
- Denote $G_n(\theta) = \frac{\partial g_n(\theta)}{\partial \theta}$, $\theta^* \in [\theta_0, \hat{\theta}]$, $\hat{G}_n \equiv G_n(\hat{\theta})$, $G_n^* \equiv G_n(\theta^*)$, $G = EG_n(\theta_0)$, $\Omega = E\left(g(z, \theta_0) g(z, \theta_0)'\right)$.

$$
\begin{aligned}
0 = \hat{G}_n' W g_n(\hat{\theta}) &= \hat{G}_n' W \left(g_n(\theta_0) + G_n^*(\hat{\theta} - \theta_0)\right) \\
\Longrightarrow \sqrt{n}(\hat{\theta} - \theta_0) &= (\hat{G}_n' W G_n^*)^{-1} \hat{G}_n' W \sqrt{n} g_n(\theta_0) \\
&\overset{LD}{=} (G' W G)^{-1} G' W \sqrt{n} g_n(\theta_0) \overset{LD}{=} (G' W G)^{-1} G' W \times N(0, \Omega) \\
&= N\left(0, (G' W G)^{-1} G' W \Omega W G (G' W G)^{-1}\right)
\end{aligned}
$$

- Efficient choice of $W = \Omega^{-1}$(or $W \propto \Omega^{-1}$),

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \left(G'\Omega^{-1}G\right)^{-1}\right).$$

- When G is invertible, W is irrelevant,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, G^{-1}\Omega G'^{-1}\right) = N\left(0, \left(G'\Omega^{-1}G\right)^{-1}\right).$$

- When $\Omega = \alpha G$(or $G \propto \Omega$),

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \alpha G^{-1}\right).$$

- Least square (LS): $g(z,\beta) = x(y - x\beta)$.
  - $G = Exx'$, $\Omega = E\varepsilon^2 xx'$, then

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, (Exx')^{-1}\left(E\varepsilon^2 xx'\right)(Exx')^{-1}\right),$$

    the so-called White's heteroscedasticity consistency standard error.
  - If $E\left[\varepsilon^2 | x\right] = \sigma^2$, then $\Omega = \sigma^2 G$ and

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \sigma^2 (Exx')^{-1}\right).$$

- Weighted LS: $g(z,\beta) = \frac{1}{E(\varepsilon^2|x)}(y - x'\beta)$.

$$G = E\frac{1}{E(\varepsilon^2|x)}xx' = \Omega \Longrightarrow \sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, G).$$

- Linear 2SLS: $g(z,\beta) = z(y - x\beta)$.
  - $G = Ezx'$, $\Omega = E\varepsilon^2 zz'$, $W = (Ezz')^{-1}$, then
    $$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, V).$$
  - If $E\varepsilon^2 zz' = \sigma^2 Ezz'$, $V = \sigma^2 \left[Exz' (Ezz')^{-1} Ezx'\right]^{-1}$.
- Linear 3SLS: $g(z,\beta) = z(y - x\beta)$.

  $G = Ezx'$, $\Omega = E\varepsilon^2 zz'$, $W = \left(E\varepsilon^2 zz'\right)^{-1}$, then

  $$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, V) \text{ for } V = \left[Exz' \left(E\varepsilon^2 zz'\right)^{-1} Ezx'\right]^{-1}.$$
- MLE as GMM: $g(z,\theta) = \frac{\partial \log f(z,\theta)}{\partial \theta}$.

  $G = -E\frac{\partial^2 \log f(z,\theta)}{\partial \theta \partial \theta'} = \Omega = E\frac{\partial \log f(z,\theta)}{\partial \theta}\frac{\partial \log f(z,\theta)}{\partial \theta'}$, then
  $$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, G^{-1}\right) = N(0, \Omega).$$

- GMM again:
  - Take linear combinations of the moment conditions to make

$$\text{Number of } g(z,\theta) = \text{Number of } \theta.$$

  - In particular, take $h(z,\theta) = G'Wg(z,\theta)$ and use $h(z,\theta)$ as the new moment conditions, then

$$\hat{\theta} = argmax_\theta \left[\frac{1}{n}\sum_{t=1}^{n} h(z_t,\theta)\right]' \left[\frac{1}{n}\sum_{t=1}^{n} h(z_t,\theta)\right]$$

is asymptotically equivalent to $\hat{\theta} = argmax_\theta g_n' W g_n$, where $G = E\frac{\partial h(z,\theta)}{\partial \theta} = G'WG$, $\Omega = Eh(z,\theta)h(z,\theta)' = G'W\Omega WG$.

- Quantile Regression as GMM:
  - $g(z,\beta) = (\tau - 1(y \leq x'\beta))x$, and W is irrelevant.
  - $G = E\frac{g(z,\beta)}{\partial\beta} = -E\frac{\partial 1(y \leq x'\beta)x}{\partial\beta}$. Proceeding with a "quick and dirty" way – take expectation before taking differentiation:

$$\begin{aligned} G =& \frac{\partial E1(y \leq x'\beta)x}{\partial\beta} = \frac{\partial ExF(y \leq x'\beta|x)}{\partial\beta} \\ =& Ex\frac{\partial F(y \leq x'\beta|x)}{\partial\beta} = Ef_y(x'\beta|x)xx' = Ef_u(0|x)xx'. \end{aligned}$$

  - Conditional on x, $\tau - 1(y \leq x'\beta_0) = \tau - 1(u \leq 0)$ is a Bernoulli r.v.$\Rightarrow E\left[(\tau - 1(y \leq x'\beta_0))^2 |x\right] = \tau(1-\tau)$, then

$$\Omega = EE\left[(\tau - 1(y \leq x'\beta_0))^2 |x\right] xx' = \tau(1-\tau)Exx'.$$

- Quantile Regression as GMM:
  - $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d}$
    $N\left(0, \tau(1-\tau)\left[Ef_u(0|x)xx'\right]^{-1} Exx' \left[Ef_u(0|x)xx'\right]^{-1}\right)$.
  - $f(0|x) = f(0)$ if homoscedastic, then $V = \frac{\tau(1-\tau)}{f(0)} Exx'$.
- Consistent estimation of G and Ω:
  - Estimated by $G \doteq \frac{1}{n}\sum_{t=1}^{n} \frac{\partial g\left(z_t, \hat{\theta}\right)}{\partial \theta}$.
  - For nonsmooth problems as quantile regression, use
    $\frac{Q_n\left(\hat{\theta}+2h_n\right)+Q_n\left(\hat{\theta}-2h_n\right)-2Q\left(\hat{\theta}\right)}{4h_n^2}$ to approximate.
    Require $h_n = o(1)$ and $1/h_n = o\left(1/\sqrt{n}\right)$.
  - For stationary data, heteroscedasticity and dependence will only affect estimation of Ω. For independent data, use White's heteroscedasticity-consistent estimate; for dependent data, use Newey-West's autocorrelation-consistent estimate.

- The initial guess $\tilde{\theta} \Rightarrow$ the next round guess $\bar{\theta}$.
- Newton-Raphson, use quadratic approximation for $Q_n(\theta)$.
- Gauss-Newton, use linear approximation for the first-order condition, e.g. GMM.
- If the initial guess is a $\sqrt{n}$ consistent estimate, more iteration will not increase (first-order) asymptotic efficiency.
- e.g. $\left(\tilde{\theta} - \theta_0\right) = O_p\left(\frac{1}{\sqrt{n}}\right)$, then $\sqrt{n}\left(\bar{\theta} - \theta_0\right) \stackrel{LD}{=} \sqrt{n}\left(\hat{\theta} - \theta_0\right)$, for $\hat{\theta} = argmax_\theta Q_n(\theta)$.

- $\phi(z_t)$ is called influence function if
  - $\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \phi(z_t) + o_p(1)$,
  - $E\phi(z_t) = 0$, $E\phi(z_t)\phi(z_t)' < \infty$.
- Think of $\sqrt{n}(\hat{\theta} - \theta_0)$ distributed as

$$\phi(z_t) \sim N\left(0, E\phi\phi'\right).$$

- Used for discussion of asymptotic efficiency, two step or multistep estimation, etc.

- For MLE,

$$\phi(z_t) = \left[-E\frac{\partial^2 \ln f(y_t,\theta_0)}{\partial\theta\partial\theta'}\right]^{-1}\frac{\partial \ln f(y_t,\theta_0)}{\partial\theta}$$
$$= \left[E\frac{\partial \ln f(y_t,\theta_0)}{\partial\theta}\frac{\partial \ln f(y_t,\theta_0)}{\partial\theta'}\right]^{-1}\frac{\partial \ln f(y_t,\theta_0)}{\partial\theta}.$$

- For GMM,

$$\phi = -\left(G'WG\right)^{-1}G'Wg(z_t,\theta_0),$$
$$\text{or} \quad \phi = -\left(E\frac{\partial h}{\partial\theta}\right)^{-1}h(z_t,\theta_0) \quad \text{for} \quad h(z_t,\theta_0) = G'Wg(z_t,\theta_0).$$

- Quantile Regression:

$$\phi(z_t) = \left[Ef(0|x)xx'\right]^{-1}(\tau - 1(u \le 0))x_t.$$

- Is MLE efficient among all asymptotically normal estimators?
- Superefficient estimator:
  Suppose $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$ for all θ. Now define

$$\theta^* = \begin{cases} \hat{\theta} & \text{if} \quad |\hat{\theta}| \geq n^{-1/4} \\ 0 & \text{if} \quad |\hat{\theta}| < n^{-1/4} \end{cases}$$

  then $\sqrt{n}(\theta^* - \theta_0) \xrightarrow{d} N(0, 0)$ if $\theta_0 = 0$, and
  $\sqrt{n}(\theta^* - \theta_0) \overset{LD}{=} \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$ if $\theta_0 \neq 0$.
- $\hat{\theta}$ is regular if for any data generated by $\theta_n = \theta_0 + \delta/\sqrt{n}$, for $\delta \geq 0$, $\sqrt{n}(\hat{\theta} - \theta_0)$ has a limit distribution that does not depend on δ.

- For regular estimators, influence function representation indexed by τ,

$$\sqrt{n}(\hat{\theta}(\tau) - \theta_0) \overset{LD}{=} \phi(z, \tau) \sim N\left(0, E\phi(\tau)\phi(\tau)'\right),$$

- $\hat{\theta}(\bar{\tau})$ is efficient than $\hat{\theta}(\tau)$ if it has a smaller var-cov matrix.
- A necessary condition is that $Cov\left(\phi(z, \tau) - \phi(z, \bar{\tau}), \phi(z, \bar{\tau})\right) = 0$ for all τ including $\bar{\tau}$.
- The following are equivalent:

$$\begin{aligned} & Cov\left(\phi(z, \tau) - \phi(z, \bar{\tau}), \phi(z, \bar{\tau})\right) = 0 \\ \Longleftrightarrow & Cov\left(\phi(z, \tau), \phi(z, \bar{\tau})\right) = Var\left(\phi(z, \bar{\tau})\right) \\ \Longleftrightarrow & E\phi(z, \tau)\phi(z, \bar{\tau})' = E\phi(z, \bar{\tau})\phi(z, \bar{\tau})' \end{aligned}$$

Newey's efficiency framework:

- Classify estimators into the GMM framework with
$$\phi(z,\tau) = D(\tau)^{-1} m(z,\tau).$$
- For the class indexed by $\tau = W$, given a vector $g(z,\theta_0)$,
$$D(\tau) \equiv D(W) = G'WG \text{ and}$$
$$m(z,\tau) \equiv m(z,W) = G'Wg(z,\theta_0).$$
- Consider MLE among the class of GMM estimators, so that τ indexes any vector of moment function having the same dimension as θ. In this case,
$$D(\tau) \equiv D(h) = -E\frac{\partial h}{\partial \theta} \text{ and } m(z,\tau) = h(z_t,\theta_0).$$

- For this particular case where $\phi(z,\tau) = D(\tau)^{-1} m(z,\tau)$,

  $$E\phi(z,\tau)\phi(z,\bar{\tau})' = E\phi(z,\bar{\tau})\phi(z,\bar{\tau})' \Longrightarrow$$

  $$D(\tau)^{-1} Em(z,\tau) m(z,\bar{\tau}) D(\bar{\tau})^{-1} = D(\bar{\tau})^{-1} Em(z,\bar{\tau}) m(z,\bar{\tau}) D(\bar{\tau})^{-1}.$$

- If $\bar{\tau}$ satisfies $D(\tau) = Em(z,\tau) m(z,\bar{\tau})$ for all τ, then both sides above are the same $D(\bar{\tau})^{-1}$ and so efficient.

- Examples. Check $D(\tau) = Em(z,\tau) m(z,\bar{\tau})$.

- GMM with optimal weighting matrix:

  $$D(\tau) = G'WG, \quad m(z,\tau) = m(z,W) = G'Wg(z,\theta_0).$$

  To check $D(\tau) = Em(z,\tau) m(z,\bar{\tau}) = G'W\Omega\bar{W}G$,

  $$G'WG = G'W\Omega\bar{W}G \Longrightarrow \Omega\bar{W} = I \Longrightarrow \bar{W} = \Omega^{-1}.$$

- MLE better than any GMM:

$$D(\tau) = -E\frac{\partial h(z,\theta_0)}{\partial \theta}, \quad m(z,\tau) = h(z,\theta_0).$$

To check $D(\tau) = Eh(z,\theta_0)\bar{h}(z,\theta_0)$, use the generalized information matrix equality:

$$\begin{aligned} 0 &= \frac{\partial Eh(z,\theta_0)}{\partial \theta} = \frac{\partial}{\partial \theta}\int h(z,\theta) f(z,\theta)\, dz \\ &= \int \frac{\partial h(z,\theta)}{\partial \theta} f(z,\theta)\, dz + \int h(z,\theta)\frac{\partial \ln f(z,\theta)}{\partial \theta} f(z,\theta)\, dz \\ &= E\frac{\partial h(z,\theta_0)}{\partial \theta} + Eh(z,\theta_0)\frac{\partial \ln f(z,\theta_0)}{\partial \theta} \end{aligned}$$

$\Longrightarrow \bar{h}(z,\theta_0) = \frac{\partial \ln f(y,\theta_0)}{\partial \theta}$, the score function for MLE.

General Framework:

- First step estimator $\sqrt{n}\left(\hat{\gamma}-\gamma_0\right)=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\phi\left(z_t\right)+o_p\left(1\right)$.
- Estimate $\hat{\theta}$ by

$$\frac{\partial Q_n(\hat{\theta},\hat{\gamma})}{\partial\theta}=\frac{1}{n}\sum_{t=1}^{n}\frac{q(z_t,\hat{\theta},\hat{\gamma})}{\partial\theta}=0\overset{\text{Let}}{=}\frac{1}{n}\sum_{t=1}^{n}h(z_t,\hat{\theta},\hat{\gamma}).$$

- Let

$$H\left(z,\theta,\gamma\right)=\frac{\partial h\left(z,\theta,\gamma\right)}{\partial\theta},\quad\Gamma\left(z,\theta,\gamma\right)=\frac{\partial h\left(z,\theta,\gamma\right)}{\partial\gamma};$$

$$H=EH\left(z_t,\theta_0,\gamma_0\right),\quad\Gamma=E\Gamma\left(z,\theta_0,\gamma_0\right);$$

$$h=h\left(\theta_0,\gamma_0\right).$$

- Then just taylor expand: $\frac{1}{\sqrt{n}}\sum h\left(z_t,\hat{\theta},\hat{\gamma}\right)=0$

$$\iff \frac{1}{\sqrt{n}}\sum h\left(\theta_0,\hat{\gamma}\right)+\frac{1}{n}\sum H\left(\theta^*,\hat{\gamma}\right)\sqrt{n}\left(\hat{\theta}-\theta_0\right)=0\Longrightarrow$$

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)=-\left[\frac{1}{n}\sum H\left(\theta^*,\hat{\gamma}\right)\right]^{-1}\frac{1}{\sqrt{n}}\sum h\left(\theta_0,\hat{\gamma}\right)$$

$$\overset{LD}{=}-H^{-1}\left[\frac{1}{\sqrt{n}}\sum h\left(\theta_0,\gamma_0\right)+\frac{1}{n}\sum \Gamma\left(\theta_0,\gamma^*\right)\sqrt{n}\left(\hat{\gamma}-\gamma_0\right)\right]$$

$$\overset{LD}{=}-H^{-1}\left[\frac{1}{\sqrt{n}}\sum h+\Gamma\left(\frac{1}{\sqrt{n}}\sum \phi\left(z_t\right)+o_p\left(1\right)\right)\right]$$

$$\overset{LD}{=}-H^{-1}\left[\frac{1}{\sqrt{n}}\sum h+\Gamma\frac{1}{\sqrt{n}}\sum \phi\left(z_t\right)\right].$$

So that $\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,V\right)$ for

$$V=H^{-1}E\left(h+\Gamma\phi\right)\left(h'+\phi'\Gamma'\right)H^{-1'}.$$

- GMM both first stage $\hat{\gamma}$ and second stage $\hat{\theta}$:
  - $\phi = -M^{-1} m(z)$, for some moment condition $m(z, \gamma)$.
  - $h(\theta, \hat{\gamma}) = G' W g(z, \theta, \hat{\gamma})$ so that $H = G' W G$, $\Gamma = G' W \frac{\partial g}{\partial \gamma} \equiv G' W G_\gamma$ for $G_\gamma \equiv \frac{\partial g}{\partial \gamma}$.
  - Plug these into the above general case.
- If $W = I$, and G is invertible, then this simplies to

$$V = G^{-1} \left[\Omega + (Eg\phi') G_\gamma' + G_\gamma (E\phi g') + G_\gamma (E\phi\phi') G_\gamma' \right] G^{-1'}.$$

- Again if you have trouble differentiating $\frac{\partial g(\theta,\gamma)}{\partial \theta}$ or $\frac{\partial g(\theta,\gamma)}{\partial \gamma}$, then simply take expectation before differentiation, just replace H and Γ by $\frac{\partial Eg(\theta,\gamma)}{\partial \theta}$ and $\frac{\partial Eg(\theta,\gamma)}{\partial \gamma}$.

Lecture 4: Basic Nonparametric Estimation

Instructor: Han Hong

Department of Economics
Stanford University

2011

- There can be many meanings to "nonparametrics".
- One meaning is optimization over a set of function.
- For example, given the sample of observations $x_1, \ldots, x_n$, find a distribution function under which the joint probability of $x_1, \ldots, x_n$ is maximized.
  - This is also called "nonparametric maximum likelihood".
- The meaning of "nonparametric" for now is density estimate and estimation of conditional expectations.

- One motivation is to first use the histogram to estimate the density:

$$\frac{1}{2h}\frac{\#\ of\ x_i\ in\ (x-h,x+h)}{n} = \frac{1}{2h}\frac{1}{n}\sum_{t=1}^{n} 1\left(x-h \leq x_i \leq x+h\right)$$
$$= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}1\left(\frac{|x-x_i|}{h} \leq 1\right)$$

- $\frac{1}{2}1\left(|x| \leq 1\right)$ is the uniform density over $(-1,1)$, called the uniform kernel.

- Generally, use other density function $K\left(\cdot\right)$ to get

$$\hat{f}\left(x\right) = \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{x-x_i}{h}\right).$$

- Another motivation is to estimate the distribution function $F(x)$ by

$$\hat{F}(x) = \frac{1}{n}\sum_{t=1}^{n} 1(x_i \leq x),$$

but you can't differentiate it to get the density.

- Replace $1(x_i \leq x)$ by $G\left(\frac{x_i - x}{h}\right)$ where $G(\cdot)$ is any smooth distribution function ($G(\infty) = 1, G(-\infty) = 0$), and $h \to 0$.

- In practice, take h as some small but fixed number, like 0.1.

- So let $K = G'(\cdot)$, differentiate $\hat{F}(x)$ to get

$$\hat{f}(x) = \frac{1}{nh}\sum_{t=1}^{n} K\left(\frac{x_i - x}{h}\right) \text{ or } \frac{1}{nh^d}\sum_{t=1}^{n} K\left(\frac{x_i - x}{h}\right) \text{ if } x \in R^d.$$

- Estimate $E(y|x)$ or more generally $E(g(y)|x)$ for some function $g(\cdot)$, or things like conditional quantiles.
- Local weighting: use observations x_i close to x.
  - Take a neighborhood $\mathcal{N}$ around x and the size of $\mathcal{N}$ should shrink to 0 but not too fast.
  - Average over those y_i for which $x_i \in \mathcal{N}$.
  - More generally give more weights to those y_i if x_i is close to x, and less weights to those y_i if x_i is far away from x.
- For weights $W_n(x, x_i)$ such that

  $(1) \sum_{i=1}^n W_n(x, x_i) = 1, \quad (2)\ W_n(x, x_i) \to 0$ if $x_i \neq x$,

  $(3)\ \max_{1 \leq i \leq n} |W_n(x, x_i)| \to 0$ as $n \to \infty$,

  estimate $E(y|x)$ by $\sum_{i=1}^n W_n(x, x_i) Y_i$.

- Anything you do parametrically, if you do that only for x_i close to x, then you become "nonparametric".
- Local nonparametric estimates:
  - kernel smoothing
  - k-nearest neighborhood (k-NN)
  - local polynomials
- Global nonparametric estimates:
  - series (sieve)
  - splines
- The focus today is kernel.

- Use density weighting for the weights $W_n(x, x_i)$, then get the kernel estimator of $E(y|x)$.
- If x_i is one-dimensional, let

$$W_n(x, x_i) = \frac{\frac{1}{nh} K\left(\frac{x-x_i}{h}\right)}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}, \quad \text{satisfying} \quad \sum_{i=1}^{n} W_n(x, x_i) = 1.$$

- The kernel estimator of $E(y|x)$ is

$$\sum_{i=1}^{n} W_n(x, x_i) Y_i = \sum_{i=1}^{n} \frac{\frac{1}{nh} K\left(\frac{x-x_i}{h}\right)}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} Y_i = \frac{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) Y_i}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}.$$

- If $x_i \in R^d$, use the multidimension density function and replace h with h^d.

- Estimate $\gamma(x)$ and $f(x)$ separately for

$$E(y|x) = \frac{E(y|x) f(x)}{f(x)} = \frac{\int yf(y,x)\,dy}{f(x)} = \frac{\gamma(x)}{f(x)}$$

- $\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$.
- For $\hat{\gamma}(x)$, plug

$$\hat{f}(x,y) = \frac{1}{nh^{d+1}} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \bar{K}\left(\frac{y_i-y}{h}\right)$$

into $\int yf(y,x)\,dy$, and let $u = (y_i - y)/h$:

$$\int y\hat{f}(y,x)\,dy = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \int y \frac{1}{h} \bar{K}\left(\frac{y_i-y}{h}\right) dy$$

$$= \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \int (y_i + uh)\bar{K}(u)\,du = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i.$$

- Another view for $\hat{\gamma}(x)$: think of $\int y\hat{f}(y,x)\,dy$ as $\int ydP$, where P is the measure over y defined by

$$P(y_i \leq y, x_i = x) = \frac{d}{dx}P(y_i \leq y, x_i \leq x)$$

$$\overset{estimate}{=} \frac{d}{dx}\frac{1}{n}\sum_{i=1}^{n} 1(y_i \leq y)\, G\left(\frac{x_i - x}{h}\right) = \frac{1}{nh^d}\sum_{i=1}^{n} 1(y_i \leq y)\, K\left(\frac{x_i - x}{h}\right)$$

- Plug in this estimate of P into $\int ydP$:

$$\int yd\hat{P} = \int yd\frac{1}{nh^d}\sum_{i=1}^{n} 1(y_i \leq y)\, K\left(\frac{x_i - x}{h}\right)$$

$$= \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)\int yd1(y_i \leq y) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i$$

- Only need to care $\hat{\gamma}(x)$ since $\hat{f}(x)$ is just a special case of $\hat{\gamma}(x)$ if $y_i \equiv 1$ identically.
- Convenient forms of kernel (density) function:
  - Uniform kernel $\frac{1}{2}1(|u| \leq 1)$;
  - Triangular kernel: $(1 - |u|)1(|u| \leq 1)$;
  - Quartic, epanechniknov, gaussian, etc.
- Estimating derivatives: as long as kernel is smooth differentiable, just simply differentiate $\hat{\gamma}(x)$:

$$\hat{\gamma}^{(k)}(x) = \frac{1}{nh^{k+d}} \sum_{i=1}^{n} K^{(k)}\left(\frac{x_i - x}{h}\right) y_i$$

- Other two major weighting schemes for $W_{ni}(x)$.
- k-nearest neighborhood (k-NN)
  - Use k closest neighbors of point x instead of fixed one.
  - Weight these k neighbors equally or according to distances.
  - Example: use any kernel density weight $K(\cdot)$.
- Local polynomial
  - Run a kth polynomial regression using observations over $|x_i - x| \leq h$.
  - The degree k corresponds to the order of the kernel.

- Series (Sieve)
  - The only difference between series and local polynomials is that you run the polynomials using all observations, instead of only a shrinking neighborhood $(x-h, x+h)$.
  - Instead of fixing k, let $k \to \infty$.
  - Instead of using polynomials, use family of orthogonal series of functions, like trigonometric function, etc.
- Splines
  - Find a twice differentiable function $g(x)$ that minimizes $\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(x)^2 dx$, for some $\lambda > 0$.
  - $\lambda \int g''(x)^2 dx$ is to penalize the roughness of the estimate g.
  - This will give a cubic polynomial with continuous second derivatives.

- Curse of dimensionality:
  For a given bandwidth (window size), the higher dimension x, the less data in a neighborhood with bandwidth h.
- If both $h \to 0$ and $nh^d \to \infty$, then the estimate is consistent.
- How about the speed at which estimator converges?
- Conclusion:
  Suppose the true function $\gamma(x)$ is pth degree differentiable, all pth derivative bounded uniformly over x. Then the optimal bandwidth h_{opt} is $n^{-\frac{1}{2p+d}}$, and the best rate at which $\hat{\gamma}(x)$ can approach $\gamma(x)$ is $O_p\left(n^{-\frac{p}{2p+d}}\right)$.

- The problem here is the bias and variance trade-off.
  - The smaller the h, the smaller the bias, but the less observations you have, thus the large the variance.
- Criterion: $total\ error = bias + estimation\ error$, or MSE.
- The bias is $O_p(h^p)$.
  - Use p bounded derivatives condition and taylor expansion.
- The variation is $O_p\left(\frac{1}{\sqrt{nh^d}}\right)$.
  - Think of $\bar{x} - \mu = O_p\left(\frac{1}{\sqrt{n}}\right)$, by analogy with nh^d.
- Total error is $O_p\left(h^p + \frac{1}{\sqrt{nh^d}}\right)$.

- Find a h to minimize total error,

$$h_{opt} = O\left(n^{-\frac{1}{2p+d}}\right).$$

- Then the (pointwise) optimal rate of convergence is

$$O\left(h_{opt}^{p}\right) = O\left(\frac{1}{\sqrt{nh^{d}}}\right) = O\left(n^{-\frac{p}{2p+d}}\right).$$

- It is not possible to have $\sqrt{n}$ convergence for nonparametric estimates since $\frac{p}{2p+d} < \frac{1}{2}$.

- Sometimes $n^{1/4}$ rate of convergence is needed for getting rid of the second order terms for semiparametric estimators, which means $p > d/2$.

- The optimal bandwidth of $\gamma^{(k)}(x)$ is of the same order as that of estimating $\gamma(x)$ itself.
- The bias is $O_p(h^{p-k})$, and the variation is $O_p\left(\frac{1}{h^k\sqrt{nh^d}}\right)$.
- The total error is $O_p\left(h^{p-k}+\frac{1}{h^k\sqrt{nh^d}}\right)$.
- Find a h to minimize this again,

$$h_{opt}=n^{-\frac{1}{2p+d}}.$$

- Then the best convergence rate is

$$O_p\left(n^{p-k}\right)=O_p\left(\frac{1}{h^k\sqrt{nh^d}}\right)=O_p\left(n^{-\frac{p-k}{2p+d}}\right).$$

- A kernel of order r is defined as those $K(\cdot)$ for which:

$$\int K(u)\,du = 1, \quad \int K(u)\,u^q du = 0, \forall q = 1, \ldots, r-1,$$
$$\int |u^r K(u)|\,du < \infty.$$

- Bias of kernel estimates $= E\hat{\gamma}(x) - \gamma(x)$

$$\begin{aligned} E\hat{\gamma}(x) =& E\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) Y_i = \int \frac{1}{h^d} K\left(\frac{x-x_i}{h}\right) E(y_i|x_i)\, f(x_i)\, dx_i \\ =& \int \frac{1}{h^d} K\left(\frac{x-x_i}{h}\right) \gamma(x_i)\, dx_i = \int K(u)\,\gamma(x+uh)\,du \\ =& \gamma(x) + \sum_{j=1}^{r-1} h^j \frac{\gamma^{(j)}}{j!} \int u^j k(u)\,du + h^r \frac{1}{r!} \int \gamma^{(r)}(x^*)\,u^r K(u)\,du \end{aligned}$$

- If $\gamma(x)$ has pth bounded derivatives and the kernel is of order r, then the bias $= h^{\min(p,r)}$.

- Variance of kernel estimates:

$$Var\left(\hat{\gamma}(x)\right) = \frac{1}{n^2 h^{2d}} \sum_{i=1}^{n} Var\left(K\left(\frac{x - x_i}{h}\right) Y_i\right)$$

$$= \frac{1}{nh^{2d}} E\left(K^2\left(\frac{x - x_i}{h}\right) Y_i^2\right) - \frac{1}{nh^{2d}} \left(EK\left(\frac{x - x_i}{h}\right) Y_i\right)^2$$

$$= \frac{1}{nh^d} \int \frac{1}{h^d} K^2\left(\frac{x - x_i}{h}\right) E\left(y_i^2 | x_i\right) f\left(x_i\right) dx_i - \frac{1}{n} \left(E\frac{1}{h^d} K\left(\frac{x - x_i}{h}\right) Y_i\right)^2$$

$$= \frac{1}{nh^d} \int \frac{1}{h^d} K^2\left(\frac{x - x_i}{h}\right) g\left(x_i\right) dx_i + O\left(\frac{1}{n}\right)$$

$$= \frac{1}{nh^d} \int K^2\left(u\right) g\left(x + uh\right) du + O\left(\frac{1}{n}\right)$$

$$= \frac{1}{nh^d} \int K^2\left(u\right) g\left(x\right) du + \frac{1}{nh^d} h \int K^2\left(u\right) g'\left(x^*\right) udu + O\left(\frac{1}{n}\right)$$

$$= \frac{1}{nh^d} \int K^2\left(u\right) g\left(x\right) du + O\left(\frac{1}{nh^d} h\right) + O\left(\frac{1}{n}\right) = O\left(\frac{1}{nh^d}\right)$$

- If use $h \sim h_{opt}$, the asymptotic distribution will depend on both the bias and the variance.
- If use $h << h_{opt}$, i.e., $\frac{h}{h_{opt}} \to 0$, the asymptotic distribution has no bias in but the convergence rate is not the fastest.
- Example: consider $d = 1$, $r = 2$, then $h_{opt} = n^{-\frac{1}{2p+d}} = n^{-\frac{1}{5}}$.
- Find the asymptotic distribution of

$$\sqrt{nh_{opt}}\left(\hat{m}(x) - m(x)\right) = h_{opt}^{-2}\left(\hat{m}(x) - m(x)\right),$$

for $\hat{m}(x) = \frac{\hat{\gamma}(x)}{\hat{f}(x)}$.

- Linearization

$$\hat{m}(x) - m(x) \approx \frac{1}{f(x)} \left(\hat{\gamma}(x) - \gamma(x)\right) - \frac{\gamma(x)}{f(x)^2} \left(\hat{f}(x) - f(x)\right)$$

- As seen above, $E\hat{\gamma}(x) - \gamma(x) = \frac{1}{2} h^2 \gamma''(x) \int u^2 K(u)\, du$.
- $E\hat{f}(x) - f(x) = \frac{1}{2} h^2 f''(x) \int u^2 K(u)\, du$,
  since $\gamma(x) = m(x) f(x)$ and $m(x) \equiv 1$.
- Therefore,

$$\begin{aligned} Eh_{opt}^{-2} \left(\hat{m}(x) - m(x)\right) &= \frac{1}{2} \left(\frac{\gamma''}{f} - \frac{m}{f} f''\right) \int u^2 K(u)\, du \\ &= \frac{1}{2} \left(\frac{m'' f + 2m' f' + m f''}{f} - \frac{m}{f} f''\right) \int u^2 K(u)\, du \\ &= \frac{1}{2} \frac{2m'(x) f'(x) + m''(x) f(x)}{f(x)} \int u^2 K(u)\, du. \end{aligned}$$

- As seen above, for $g(x) = E(y^2|x) f(x)$,
  $Var\left(\sqrt{nh}(\hat{\gamma}(x) - \gamma(x))\right) \to g(x) \int K^2(u)\, du$.
- $Var\left(\sqrt{nh}(\hat{f}(x) - f(x))\right) \to f(x) \int K^2(u)\, du$ since for density estimate where $y \equiv 1$, $g(x) = f(x)$.
- The covariance between $\hat{\gamma}(x)$ and $\hat{f}(x)$:

$$Cov\left(\sqrt{nh}(\hat{\gamma}(x) - \gamma(x)), \sqrt{nh}(\hat{f}(x) - f(x))\right) \to \gamma(x) \int K^2(u)\, du.$$

- Therefore, use the delta method

$$Var\left(\sqrt{nh}(\hat{m}(x) - m(x))\right) = Var\left(\sqrt{nh}\left(\frac{1}{f}\hat{\gamma} - \frac{m}{f}\hat{f}\right)\right)$$

$$= \left(\frac{1}{f^2} E(y^2|x) f - \frac{2}{f^2} m\gamma + \frac{m^2}{f^2} f\right) \int K^2(u)\, du$$

$$= \frac{1}{f(x)} \left(E(y^2|x) - m(x)^2\right) \int K^2(u)\, du = \frac{1}{f(x)} \sigma^2(x) \int K^2(u)\, du$$

- To summarize: $\sqrt{nh}\left(\hat{m}\left(x\right)-m\left(x\right)\right)\xrightarrow{d}$

$$N\left(\frac{m''\left(x\right)f\left(x\right)+2m'\left(x\right)f'\left(x\right)}{2f\left(x\right)}\int u^{2}K\left(u\right)du,\frac{1}{f\left(x\right)}\sigma^{2}\left(x\right)\int K^{2}\left(u\right)du\right)$$

- If use a undersmooth bandwidth $h<<n^{-1/5}$, say $h=n^{-1/4}$,

$$\sqrt{nh}\left(\hat{m}\left(x\right)-m\left(x\right)\right)\xrightarrow{d}N\left(0,\frac{1}{f\left(x\right)}\sigma^{2}\left(x\right)\int K^{2}\left(u\right)du\right)$$

- If use h_{opt} to draw the confidence interval around $\hat{m}\left(x\right)$, consistent bias term is needed.

- However, $\gamma''\left(x\right)$ can NOT be estimated consistently using h_{opt}. Instead, use a oversmoothed bandwidth, say $g=n^{-1/6}$.

- Good fit of estimate:
  - Minimize $\sum_{i=1}^{n}\left(\hat{m}\left(x_i\right)-m\left(x_i\right)\right)^2$.
  - If replace $m\left(x_i\right)$ with y_i, we will get perfect fit 0 since as $h \to 0$, $\hat{m}\left(x_i\right)=y_i$.
- Another way to think about this,

$$\sum_{i=1}^{n}\left(\hat{m}\left(x_i\right)-y_i\right)^2=\sum_{i=1}^{n}\left(\hat{m}\left(x_i\right)-m\left(x_i\right)-\epsilon_i\right)^2$$

$$=\underbrace{\sum_{i=1}^{n}\left(\hat{m}\left(x_i\right)-m\left(x_i\right)\right)^2}_{\text{what we want}}+\underbrace{\sum_{i=1}^{n}\epsilon_i^2}_{\text{unrelated}}\underbrace{-2\sum_{i=1}^{n}\left(\hat{m}\left(x_i\right)-m\left(x_i\right)\right)\epsilon_i}_{\text{the trouble}}.$$

- Expectation of trouble term:

$$E\sum_{i=1}^{n}\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{x_i-x_j}{h}\right)\epsilon_j\epsilon_i=\frac{1}{nh}\sum_{i=1}^{n}K\left(0\right)\sigma^2=\frac{1}{h}\sigma^2K\left(0\right)$$

- Cross validation
  - Leave-one-out estimate $\hat{m}_{-i}\left(x_i\right)=\frac{1}{(n-1)h}\sum_{j\neq i}^{n}K\left(\frac{x_j-x_i}{h}\right)y_i$
  - Minimize cross-validation function

$$CV\left(h\right)=\sum_{i=1}^{n}\left(m_{-i}\left(x_i\right)-y_i\right)^2$$

- Penalizing function
  - Consistent trouble term estimate $K\left(0\right)\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{m}\left(x_i\right)\right)^2$
  - Minimize penalizing function

$$G(h)=\sum_{i=1}^{n}\left(\hat{m}\left(x_i\right)-y_i\right)^2+2K\left(0\right)\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{m}\left(x_i\right)\right)^2$$

- It is essentially equivalent to high order kernel.
- It doesn't make any difference if you are just running a simple kernel regression.
- If the objective function is only convex with positive $K(\cdot)$, say, run a nonparametric quantile regression, then operationally the Jacknife method is very useful in preserving the convexity of the objective function.

- It is useful to obtain optimal bandwidth and optimal uniform convergence rate, i.e., for $\sup_{x\in X}|\hat{\gamma}(x)-\gamma(x)|$.
- Again, consider the bias-variance tradeoff.
- The bias $\sup_{x\in X}|\hat{\gamma}(x)-\gamma(x)|$ for rth order kernel is $O_P(h^p)$.
- The error $\sup_{x\in X}|\hat{\gamma}(x)-E\hat{\gamma}(x)|$ is $O_p\left(\left(\frac{nh^d}{\log n}\right)^{-1/2}\right)$.
  - Use Berstein inequality in the proof.
- Minimize total error $O_P(h^p)+O_p\left(\left(\frac{nh^d}{\log n}\right)^{-1/2}\right)$.

Applications of Nonparametric methods to Structural Estimation of Auction Models

- Parametric: Paarsch (1992, Journal of Econometrics)
- Nonparametric: Guerre, Perrigne, Vuong (Econometrica, 2000)

First Price Auctions
(Independent and Private Value Models)

J auctions

I bidders per auction with

(i.i.d) evaluations $v_{ij} \sim f(v)$

↓

across both auctions and bidders

(1) The Theory model:

~~MAX~~ consider bidder 1

without loss of generality

$$\max_b (v_1 - b)\, P(b > \max_{j=2,\dots,n} \beta(v_j))$$

$$= (v_1 - b)\, P(\max_{j=2,\dots,n} v_j \le \beta^{-1}(b))$$

$$= (v_1 - b)\, F_n(\beta^{-1}(b))^{I-1}$$

$$= (v_1 - b)\, G(\beta^{-1}(b))$$

FOC gives us

$$\frac{(v_1 - b)\, g(\beta^{-1}(b))}{\beta'(\beta^{-1}(b))} - G(\beta^{-1}(b)) = 0.$$

In Equilibrium $\beta(v_1) = b$,

So that

$$(v_1 - \beta(v_1))\, g(v_1) - G(v_1)\, \beta'(v_1) = 0$$

$$\Rightarrow v_1\, g(v_1) = \beta(v_1)\, g(v_1) + G(v_1)\, \beta'(v_1)$$

$$= \frac{\partial}{\partial v_1}\left[\beta(v_1)\, G(v_1)\right]$$

Note that $G(0) = F(0)^{I-1} = 0$

$\beta(0) = 0$.

Thus,

$$\beta(v_i)\, G(v_i) = \int_0^{v_i} x\, g(x)\, dx$$

$$\Rightarrow \beta(v_i) = \frac{1}{G(v_i)} \int_0^{v_i} x\, g(x)\, dx$$

Now

$$G(v_i) = F(v_i)^{I-1}, \quad g(x) = (I-1) F(x)^{I-2} f(x)$$

Thus

$$\beta(v) = \frac{1}{F(v)^{I-1}} \int_0^{v} x\, (I-1)\, F(x)^{I-2} f(x)\, dx$$

and

$$\beta'(v) = \frac{1}{F(v)^{I-1}}\, v\,(I-1)\, F(v)^{I-2} f(v)$$

$$- \frac{(I-1) f(v)}{F(v)^{I-2}} \int_0^{v} x\,(I-1)\, F(x)^{I-2} f(x)\, dx$$

$$= \frac{v(I-1)}{F(v)} f(v) - f(v)\, F(v)(I-1)\, \beta(v)$$

(2) If we make parametric assumption about $F(v;\theta)$

e.g. Normal $\underbrace{(\mu, \sigma^2)}_{\theta}$ or log-normal

(a) $F(v;\theta)$

(b) let $b = \beta(v;\theta)$

Then (Paarsch 1992) use maximum likelihood method,

$$\hat{\theta} = \arg\max_{\theta} \log \prod_{j=1}^{J} \prod_{i=1}^{I} \underbrace{g(b_{ij}|\theta)}_{\downarrow}$$

density of Bids

$$= \log \prod_{j=1}^{J} \prod_{i=1}^{n} f(\beta^{-1}(b_{ij};\theta)) \frac{1}{\beta'_{\theta}(\beta^{-1}(b_{ij};\theta))}$$

Numerically (1) Compute $\beta(v;\theta)$

(2) Invent $\beta^{-1}(v;\theta)$

(3) maximize wrt θ.

3 Density Estimation

- Let h denote the length of the cells in the histogram

- Let f denote the density and F the cdf, then:

$$f(x_0) = \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h}$$

- A first (naive) estimator of a density would be to use the height of cells in a histogram.

$$\begin{aligned} \widehat{f}_{HIST}(x_0) &= \frac{1}{N} \sum_{i=1}^{N} \frac{1(x_0 - h < x_i < x_0 + h)}{2h} \\ &= \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} 1\left(\left| \frac{x_i - x_0}{h} \right| < 1 \right) \end{aligned}$$

- This corresponds to the probability of falling into a bin of length 2h.

- In practice, note that this estimate of the density will be discontinuous.

- A more desirable (and efficient!) way to estimate the density would be to smooth out the discontinuities.

- A kernel density estimator generalizes our histogram estimator to:

$$\widehat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} K\left(\frac{x_i - x_0}{h}\right)$$

- where K takes the place of the indicator function above.

- K is called a kernel function and h is smoothing parameter called a bandwidth.

- We will make the following assumptions about the kernel function:

(i) K(z) is symmetric around 0

(ii) $\int K(z)dz = 1$, $\int zK(z)dz = 0$, $\int |K(z)|\, dz < \infty$

(iii) (a) either $K(z) = 0$ for $|z| > z_0$ or (b) $|z|\, K(z) \to 0$ as $|z| \to \infty$

(iv) $\int z^2 K(z)dz = \kappa < \infty$

- We will commonly assume that $z \in [-1, 1]$ as a normalization on the domain in the case of (iii) a.

- Some commonly used kernels are:

$$\begin{aligned}
&\text{uniform } 1\left(|z| < 1\right) \\
&\text{Epanechnikov } \frac{3}{4}(1 - z^2) \times 1\left(|z| < 1\right) \\
&\text{normal } (2\pi)^{-1/2} \exp(-z^2/2)
\end{aligned}$$

- Note that as h is larger, larger weights are given to observations further away from x_0.

- That is, larger values of h smooth the observations more heavily.

- In an application, we will want $h \to 0$ as $N \to \infty$ so that in the limit (at an appropriate rate).

- Thus, we only include observations in an arbitrarily small neighborhood in our density estimate $\widehat{f}(x_0)$.

- In choosing the bandwidth, we will face a tradeoff between the bias of $\widehat{f}(x_0)$, denoted $b(x_0)$, and the variance of $\widehat{f}(x_0)$, denoted V$[\widehat{f}(x_0)]$.

$$\begin{aligned} b(x_0) &= E[\widehat{f}(x_0)] - f(x_0) = \frac{1}{2}h^2 f''(x_0) \int z^2 K(z) dz \\ V[\widehat{f}(x_0)] &= \frac{1}{Nh} f(x_0) \int K(z)^2 dz + o(\frac{1}{Nh}) \end{aligned}$$

- Note that a small h decreases the bias but increases the variance.

- In the limit, we it is desirable to let $h \to 0$ and $Nh \to \infty$ so that both the bias and the variance eventually become zero.

- It can be shown that $\widehat{f}(x_0)$ is pointwise consistent if $h \longrightarrow 0$ and $Nh \longrightarrow \infty$

- Uniform consistency if $Nh/lnN \longrightarrow \infty$ (this requires more smoothing).

- It can be shown that the kernel is (pointwise) asymptotically normal,

$$(Nh)^{1/2}\left(\widehat{f}(x_0) - f(x_0) - b(x_0)\right) \rightarrow^d N[0, f(x_0)\int K(z)^2 dz]$$

- This is potentially complicated object to compute.

- A practical alternative is to use a resampling procedure such as the bootstrap.

- Another important choice is the bandwidth.

- This can be found by minimizing the expected mean square error.

- There are also plug in estimates (such as Silvverman's plug in estimate).

4 Example-Part 1.

- Next, we consider the problem of the identification and estimation of auction models.

- In an auction, the economist sees the distribution of bids.

- The economist wishes to infer bidder's private information and utility functions.

- Key papers in the literature are Paarsch (1992), Elyakime, Laffont, Loisel and Vuong (1994) and Guerre, Perrigne and Vuong (2000).

5 First Price Auction Examples.

- Consider the first price auction with independent private values.

- In the model, there are $i = 1, ..., N$ symmetric bidders with valuation v_i for a single and indivisible object.

- Valuations are iid with cdf $F(v)$ and pdf $f(v)$.

- In the auction, bidders simultaneously submit sealed bids b_i.

- Bidder i's vNM utility is

$$u_i(b_1, ..., b_n, v_i) \equiv \begin{cases} v_i - b_i & \text{if } b_i > b_j \text{ for all } i \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

- Let $\pi_i(b_i; v_i)$ denote the expected profit of bidder i where ϕ is the inverse of the bid function:

$$\pi_i(b_i; v_i) \equiv (v_i - b_i)F(\phi(b))^{N-1}. \qquad (2)$$

- The first order condition for maximizing expected profits (2) implies that

$$v = b + \frac{F(\phi(b))}{f(\phi(b))\phi'(b)(N-1)} \ . \qquad (3)$$

- This looks hard to deal with.

- Guerre, Perrigne and Vuong (2000) propose an alternative approach.

- The econometrician observes $t = 1, ..., T$ independent replications of the auction described above.

- For each auction t, the econometrician observes all of the bids $b_{i,t}$.

- The object that GPV wish to estimate is $F(v)$.

- Let $G(b) = F(\phi(b_i))$ denote the equilibrium distribution of the bids.

- If we substitute $G(b)$ into (**??**) allows us to write expected utility as:

$$(v_i - b_i)G(b_i)^{N-1}.$$

The first order conditions can now be written as:

$$(v_i - b_i)\,(N-1)\,g(b_i) - G(b_i) = 0 \tag{4}$$

$$v_i = b_i + \frac{G(b_i)}{(N-1)g(b_i)} \tag{5}$$

- Let $\widehat{G}$ and $\widehat{g}$ denote estimates of G and g

- we can form an estimate $\widehat{v}_{i,t}$ of bidder i's private information $v_{i,t}$ in auction t by substituting these terms into (5):

$$\widehat{v}_{i,t} = b_{i,t} + \frac{\widehat{G}(b_{i,t})}{(N-1)\widehat{g}(b_{i,t})} \tag{6}$$

To summarize, the estimator proposed by GPV:

1. Given bids $b_{i,t}$ for $i = 1, ..., N$ and $t = 1, ..., T$, estimate the distribution and density of bids $\widehat{G}(b)$ and $\widehat{g}(b)$.

2. Compute $\widehat{v}_{i,t}$ for $i = 1, ..., N$ and $t = 1, ..., T$ using equation (6). Use the empirical cdf of the $\widehat{v}_{i,t}$ to estimate F.

- This idea turns out to be quite general.

- The distribution of bids can be used to recover private information even in multiple unit auctions or auctions with dynamics.

- These estimators have been applied to offshore oil drilling, procurement, electronic commerce and treasury bill markets.

- There are still some interesting research questions left, however, particularly in the common values case.

Quantile Regression

In OLS

$$y = x\beta + \varepsilon$$

we assume

$$E(\varepsilon | x) = 0,$$

Same as

$$E(y|x) = x\beta.$$

But we can also estimate $E(y|x)$ nonparametrically

e.g.

$$\hat{E}(y|x) = \frac{\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right) y_i}$$

Both $f(y|x)$ and $F(y|x)$, or the entire distribution of y given y, can be estimated nonparametrically

Instead of ~~conditional~~ mean, what about quantiles and conditional quantiles?

$$Q_\tau(Y) = \inf\{q : P(Y \le q) \ge \tau\}$$

$$Q_\tau(Y|x) = \inf\{q : P(Y \le q|x) \ge \tau\}$$

~~Usually, $\tau = 0.5$~~

Note.

$$Q_\tau(y) = q \iff P(y \le q) = \tau.$$

$Q_\tau(Y)$ can ~~be~~ be estimated by the $(100\times\tau)$th order statistic.

How about $Q_\tau(Y|x)$

Note that

$$\hat{F}_Y(q|x) = \hat{P}(Y \le q|x) = \hat{E}(1(Y \le q)|x)$$

$$= \frac{\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) 1(Y_i \le q)}{\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)}$$

Define $Q_\tau(Y|x) = q(\tau)$

$\Leftrightarrow \quad \hat{P}(Y \le q(\tau)|x) = \tau.$

$$\hat{F}_Y(q(\tau)|x) = \tau = \underbrace{\frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \mathbb{1}(Y_i \le q(\tau))}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}}_{\text{call this } g(q)}$$

Use the bisection method (or any other root finder to solve for q such that

$$\hat{F}_Y(q|x) = \tau \Leftrightarrow g(q) = \tau.$$

$$\Rightarrow \quad q(\tau)$$

But most of the time we don't have enough data to run nonparametric quantiles

In OLS, we assume

$$E(y|x) = x'\beta$$

In Quantile Regress, assume

$$Q_\tau(y|\tau) = x'\beta_\tau$$

Reference:

Koenker & Bassett, 1978, Econometrica

Chernozhukov & Hansen 2005, Econometrica

In Stata

qreg y x, τ

by Default $\tau = \frac{1}{2}$.

① $\tau = \frac{1}{2}$, $\text{med}(Y|X) = x\beta$

$$\hat{\beta} = \arg\min \frac{1}{n} \sum_{i=1}^{n} |Y_i - X_i\beta|$$

More Generally, $\forall \tau$.

$$\min_{\beta} Q_\tau(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\tau - 1(y_i \le x_i\beta)\right](y_i - x_i\beta)$$

"Check function"

$$\rho_\tau(u) = \left(\tau - 1(u \le 0)\right) u$$

then

~~$Q_\tau(\beta) = \frac{1}{n}\sum_{i=1}^{n}$~~

$$Q_\tau(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - x_i\beta)$$

$$= \begin{cases} \tau(y_i - x_i\beta) & \text{if } y_i > x_i\beta \\ (\tau - 1)(y_i - x_i\beta) & \text{if } y_i < x_i\beta \end{cases}$$

So for $\tau = \frac{1}{2}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\right)|y_i - x_i\beta|$$

$\frac{1}{2}$ not important

For one Quantile

$P(u_t < 0) = \tau$ or $P(u_t < 0 \mid x_t) = \tau$

0 is the τth quantile of the error term $u_t \Leftrightarrow Q_\tau(u_t \mid x_t) = 0$

$$Q_\tau(y \mid x_t) = Q_\tau(x_t\beta + u_t \mid x_t) = x_t\beta + Q_\tau(u_t \mid x_t) = x_t\beta.$$

Often times people look at multiple Quantiles.

$$Q_\tau(y_i \mid x_t) = x_t\beta(\tau)$$

problem: Now Quantiles might cross.

The τth percentile

$$q_\tau = \arg\min_q E\,\rho_\tau(y-q)$$

$$E\,\rho_\tau(y-q) = (\tau-1)\int_{-\infty}^{q}(y-q)\,dF(y) + \tau\int_{q}^{\infty}(y-q)\,dF(y)$$

FOC:

$$(\tau-1)\left[-\int_{-\infty}^{q} dF(y)\right] + \tau\cdot\left[-\int_{q}^{\infty} dF(y)\right] = 0$$

$$\Rightarrow \quad (\tau-1)\,F_Y(q) + \tau\cdot\underbrace{\left(1-F_Y(q)\right)}_{\downarrow\ 1-\tau} = 0$$

Solution $\quad F_Y(q) = \tau$

Consistency of Quantile Regression

A Special case of M estimators

① $\beta \subset B$ compact (can be removed)

② $Q_\tau(y|x) = x'\beta \iff P(y \le x\beta | x) = \tau$

or

$$y = x\beta + \varepsilon \qquad P(\varepsilon \le 0 | x) = \tau$$

③ $E f_\varepsilon(0|x) xx'$ is nonsingular and finite.

Step 1: $\sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - x_i\beta) - E\rho_\tau(y_i - x_i\beta) \right| \xrightarrow{p} 0.$

a) Pointwise LLN of sample $Q_n(\beta)$

b) Stochastic Equicontinuity

- VC – class
- Convexity lemma

Step 2:

$E\,\rho_\tau(y - x\beta)$ is uniquely maximized at β_τ

$$E\,\rho_\tau(y_i - x_i\beta) = E_x\, E_{y|x}\,\rho_\tau(y - x\beta)$$

$$= E_x\left[\int_{-\infty}^{x'\beta} (\tau - 1)(y - x\beta) f(y|x)\, dy + \int_{x\beta}^{\infty} \tau(y - x\beta) f(y|x)\, dy\right]$$

$$\Rightarrow \frac{\partial E[\rho_\tau(y - x\beta)]}{\partial \beta} = E_x\Big[x(1-\tau)\, P(y < x\beta \mid x) - \tau\, P(y > x\beta \mid x)\Big]$$

$$= 0 \quad \text{at} \quad \beta_\tau$$

$$\frac{\partial^2 E\, \rho_\tau(y - x\beta)}{\partial \beta \partial \beta'} = E\, xx'\, f_y(x\beta \mid x)$$

$$= E\, xx'\, f_\varepsilon(0 \mid x) \quad \text{at } \beta_\tau$$

$$f_{y \mid x}(x\beta \mid x) = f_\varepsilon(0 \mid x)$$

since $y = x\beta + \varepsilon$

Asymptotic Distribution

Non standard.

$$\frac{\partial Q_n(\beta)}{\partial \beta} \approx \frac{1}{n} \sum_{i=1}^{n} \left(\tau - 1(y < x\beta)\right) x$$

$$\sqrt{n}\, \frac{\partial Q_n(\beta)}{\partial \beta} \xrightarrow{d} N\left(0,\ \tau(1-\tau)\, E\, xx'\right)$$

$$\frac{\partial^2 Q_n(\beta)}{\partial\beta\,\partial\beta'} \simeq \frac{\partial^2 \mathbb{E} Q_n(\beta)}{\partial\beta\,\partial\beta}$$

$$\simeq \frac{\partial^2}{\partial\beta\,\partial\beta'} \mathbb{E}\, \rho_\tau(y - x\beta)$$

$$= \frac{\partial}{\partial\beta} \mathbb{E}\left(\tau - 1(y < x\beta)\right) x$$

$$= \frac{\partial}{\partial\beta} \mathbb{E}\left(\tau - P(y < x\beta \mid x)\right) x$$

$$= -\mathbb{E}\, f_y(x\beta \mid x)\, xx'$$

$$\Downarrow$$

$$H$$

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0,\ H^{-1} \underset{\Downarrow}{\Omega} H^{-1}\right)$$

$$\tau(1-\tau)\, \mathbb{E} xx'$$

Formally, Pollard (1990, Econometric Theory)

Shows that

$$\sum_{i=1}^{n} \rho_\tau(y_i - x_i\beta) = \sum_{i=1}^{n} \rho_\tau(\varepsilon_i)$$

$$+ \sqrt{n}(\beta - \beta_\tau)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tau - 1(y_i < x_i\beta_\tau)\right)x_i$$

$$+ \sqrt{n}(\beta - \beta_\tau)' E f_{\varepsilon}(0|x)\, xx' \sqrt{n}(\beta - \beta_\tau)$$

$$+ o_p(1)$$

$$\sqrt{n}(\hat{\beta} - \tilde{\beta}) = o_p(1)$$

$\hat{\beta} \rightarrow$ min the LHS

$\tilde{\beta} \rightarrow$ min the RHS

s.

$$\sqrt{n}(\tilde{\beta} - \beta_\tau) = \left(E_x f_{\varepsilon}(0|x)xx'\right)^{-1} \frac{1}{\sqrt{n}}\sum\left(\tau - 1(y_i \le x_i\beta)\right)x_i$$

$$\xrightarrow{d} N(0, H^{-1}\Omega H^{-1})$$

$Q = \tau(1-\tau) E xx'$ is

easy to estimate.

$$\tau(1-\tau)\ \frac{1}{n}\sum x_i x_i'$$

How about

$$H = E\, f_\varepsilon(0|x)\, xx'.$$

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_\varepsilon(0|x_i)\, x_i x_i'$$

where $\hat{f}_\varepsilon(0|x_i) = \dfrac{\frac{1}{nh}\sum_{j=1}^{n} k\left(\frac{\hat{\varepsilon}_{ij}}{h}\right) k\left(\frac{x_j - x_i}{h}\right)}{\sum_{j=1}^{n} k\left(\frac{x_j - x_i}{h}\right)}$

homoscedastic case

$$f_\varepsilon(0|x) = f_\varepsilon(0)$$

$$\hat{f}_\varepsilon(0) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{\hat{\varepsilon}_i}{h}\right) \qquad \hat{\varepsilon}_i = y_i - x_i\hat{\beta}$$

It turns out

$\hat{f}$ can be

$$\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{\hat{\varepsilon}_i}{h}\right) X_i X_i'$$

need $h \to 0$

$$\frac{nh}{\log n} \to \infty$$

↓

Some previous papers require $\sqrt{n}\, h \to \infty$:
too strong

$$E\ \frac{1}{h} k\left(\frac{\varepsilon}{h}\right) XX' \to E\, f_\varepsilon(0|x)\, XX'$$

Consider $\hat{H} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k\left(\frac{\hat{\varepsilon}_i}{h}\right) x_i x_i'$

Intuitively $\hat{\varepsilon}_i \to \varepsilon_i$

Consider $\tilde{H} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k\left(\frac{\varepsilon_i}{h}\right) x_i x_i'$

want to show

(1) $E\tilde{H} \to E f_\varepsilon(\cdot | x_i) x_i x_i'$

(2) $\mathrm{var}(\tilde{H}) \to 0.$

Show (1).

$$E\left(\frac{1}{h} k\left(\frac{\varepsilon}{h}\right) x_i x_i'\right)$$

$$= \int\int \frac{1}{h} k\left(\frac{\varepsilon}{h}\right) x x' f(\varepsilon | x) d\varepsilon f(x) dx$$

let $u = \frac{\varepsilon}{h} \Rightarrow \varepsilon = uh$

$\partial u = (\partial \varepsilon) \cdot \frac{1}{h}$ $\quad d\varepsilon = (du) h$

$$= \int\int \frac{1}{h} k(u) x x' f_\varepsilon(uh | x) f(x) h \, du \, dx$$

$$= \int\int K(u)\, x x' f_\varepsilon(uh \mid x)\, f(x)\, du\, dx$$

$$= \int \left[\int K(u) f_\varepsilon(\underbrace{uh}_{\to 0} \mid x)\, du \right] x x' f(x)\, dx$$

$$\to f_\varepsilon(0 \mid x)$$

$$\to \int \left[\underbrace{\int K(u)\, du}_{1} \right] f_\varepsilon(0 \mid x)\, x x' f(x)\, dx$$

$$= \int f_\varepsilon(0 \mid x)\, x x' f(x)\, dx$$

$$= E\left[f_\varepsilon(0 \mid x)\, x x' \right]$$

(2) Variance. Suppose X_i is a scalar

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\left(\frac{\varepsilon_i}{h}\right)X_i X_i'\right) = \frac{1}{nh^2}\mathrm{Var}\left(k\left(\frac{\varepsilon_i}{h}\right)X_i X_i'\right)$$

$$= \frac{1}{nh^2}\left(E\left[k^2\left(\frac{\varepsilon_i}{h}\right)X_i^4\right] - \left(E\,k\left(\frac{\varepsilon_i}{h}\right)X_i^2\right)^2\right)$$

$$= \frac{1}{nh}\int\int \frac{1}{h}k^2\left(\frac{\varepsilon}{h}\right)x^4 f_{\varepsilon}(\varepsilon|x)\,d\varepsilon\, f(x)\,dx$$

$$- \underbrace{\frac{1}{nh}E\left(\frac{1}{h}k\left(\frac{\varepsilon_i}{h}\right)X_i^2\right)^2}_{O\left(\frac{1}{n}\right)}$$

If $u_i = \frac{\varepsilon_i}{h} \Rightarrow u_i h = \varepsilon_i \Rightarrow (du_i)\,h = d\varepsilon_i$

$$= \frac{1}{nh}\int\int \frac{1}{h}k^2(u_i)X_i^4 f_{\varepsilon}(u_i h|X_i) f(x_i)\,h\,du_i\,dx_i - O\left(\frac{1}{n}\right)$$

($u_i h \to 0$)

$$= \frac{1}{nh}\left[\underbrace{\int k^2(u)\,du}_{\text{constant}}\right]\left[\underbrace{\int f_{\varepsilon}(0|x)\,x_i^4 f(x)\,dx}_{\text{constant}}\right] - O\left(\frac{1}{n}\right)$$

$$= \frac{1}{nh}\cdot \text{const} - O\left(\frac{1}{n}\right)$$

$\to 0$ if $nh \to \infty$

Summary

$$\sqrt{n}(\hat{\beta}_\tau - \beta) \xrightarrow{A} N(0, \hat{H}^{-1} \hat{S} \hat{H}^{-1})$$

where

$$\hat{S} = \tau(1-\tau) \frac{1}{n} \sum_{i=1}^{n} X_i X_i'$$

$$\hat{H} = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{\hat{\xi}_i}{h}\right) X_i X_i'$$

Quantile Regression as GMM

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta)$$

$$\rho_\tau(u) = (\tau - 1(u<0))\, u$$

Approximate first order condition

$$\sum_{i=1}^{n} X_i \left[\tau - 1(y_i \le x_i'\hat{\beta}_\tau < 0)\right] \approx o_p(1)$$

GMM Style Quantile Estimator

$$\min_\beta \frac{1}{n} \sum_{i=1}^{n} X_i \left(\tau - 1(y_i \le x_i'\beta)\right)' W \frac{1}{n} \sum_{i=1}^{n} X_i \left(\tau - 1(y_i \le x_i'\beta)\right)$$

Is X_i the optimal instrument ?!

· You can also use

$$\min_{\beta} \left\| \frac{1}{n} \sum_{i=1}^{n} g(x_i)\left(\tau - 1(y_i < x_i\beta)\right) \right\|_{W}$$

where $\|x\|_W = x' W x$

optimal W

$$\mathrm{var}\left(g(x_i)\left(\tau - 1(y_i < x_i\beta)\right)\right)^{-1}$$

$$= \left[\tau(1-\tau)\, E\, g(x_i) g(x_i)' \right]^{-1}$$

$$\simeq \frac{1}{\tau(1-\tau)} \left(\frac{1}{n} \sum_{i=1}^{n} g(x_i) g(x_i)' \right)^{-1}$$

e.g. $\quad g(x_i) = 1(x_i > 0)$

Endogeneity (Chernozhukov & Hansen 2005) Econometrica

$$y_i = \alpha D_i + X_i'\beta + \varepsilon_i$$

D_i - endo

X_i' - Exogenous

Let Z be the instrument, assume.

τ-the Quantile of $\varepsilon | Z = 0$

$\Leftrightarrow$

$$P(\varepsilon < 0 | Z) = \tau.$$

This is different from $E(\varepsilon | Z) = 0$, weaker than $\varepsilon \perp Z$.

Quantile IV.

$$\operatorname{argmin} \left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \left(\tau - 1(y_i < \alpha D_i + X_i'\beta) \right) \right\|_W$$

for $\|x\|_W = x'Wx$.

Optimal weighting matrix is

$$W^{-1} = \operatorname{Var}\left(Z \left(\tau - 1(y_i < \alpha D_i + X_i'\beta) \right) \right)$$

$$= \left(\tau(1-\tau) E Z_i Z_i' \right)^{-1}$$

$$\sim \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i'$$

· Computation.

Often D is a scalar

$$y_i = \alpha D_i + x_i\beta + \varepsilon_i$$

$$y_i - \alpha D_i = x_i\beta + \varepsilon_i$$

If α is known.

qreg $y_i - \alpha D_i$ on x_i z_i

should estimate $(\beta, 0)$.

choose α so that the coeff on z_i in

qreg $y_i - \alpha D_i$ on x_i z_i

is as close as possible to zero.

Endogeneity (Chernozhukov & Hansen 2005)
Econometrica

$$y_i = \alpha D_i + X_i'\beta + \varepsilon_i$$

D_i - endo

X_i' - Exogenous

Let Z be the instrument, assume.

τ-the Quantile of $\varepsilon | Z = 0$

$\Leftrightarrow$

$$P(\varepsilon < 0 | Z) = \tau.$$

This is different from $E(\varepsilon | Z) = 0$,

weaker than $\varepsilon \perp Z$.

Quantile IV.

$$\operatorname{argmin} \left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \left(\tau - 1(y_i < \alpha D_i + X_i'\beta) \right) \right\|_W$$

for $\|x\|_W = x'Wx$.

Optimal weighting matrix is

$$W^{-1} = Var\left(Z \left(\tau - 1(y_i < \alpha D_i + X_i'\beta) \right) \right)$$

$$= \left(\tau(1-\tau) E Z_i Z_i' \right)^{-1}$$

$$\sim \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i'$$

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0,\ (GWG)^{-1}GW\Sigma WG(GWG)^{-1})$$

$$\Omega = \tau(1-\tau)\,E z_i z_i' \approx \tau(1-\tau)\frac{1}{n}\sum_{i=1}^{n} z_i z_i'$$

$$G = E\left[\frac{\partial}{\partial\beta} z_i\,(\tau - 1(y_i < x_i\beta))\right]$$

$$\approx E\,\frac{\partial}{\partial\beta} z_i\,(\tau - P(y_i < x_i\beta \mid z_i))$$

$$= E\,\frac{\partial}{\partial\beta} z_i\,(\tau - F_{y_i}(x_i\beta \mid z_i))$$

$$= E\,z_i x_i'\, f_{y_i}(x_i\beta \mid z_i, x_i)$$

$$= E\,z_i x_i'\, f_{\varepsilon_i}(y_i - x_i\beta \mid z_i, x_i)$$

$$= E\,z_i x_i'\, f_{\varepsilon}(0 \mid z_i, x_i)$$

$$\hat{G} \approx \frac{1}{n}\sum_{i=1}^{n} z_i x_i'\,\frac{1}{h}K\left(\frac{y_i - x_i\hat{\beta}}{h}\right)$$

Can show that if $h \to 0$, $\sqrt{n}h/\log n \to \infty$

then $\hat{G} \xrightarrow{p} G$.

Efficiency of Quantile IV estimator.

Special case of conditional moment model / optimal choice of instrument.

$$E[\rho(y_i, x_i, \beta) | z_i] = 0.$$

$$E[\tau - 1(y_i < x_i \beta) | z_i] = 0$$

Recall optimal Instrument

$$\phi(z_i) = \sigma^2(z_i)^{-1} E\left[\frac{\partial}{\partial \beta} \rho(y_i, x_i, \beta) | z_i\right]$$

In the nonsmooth case

$$\sigma^2(z_i) = Var\left(\tau - 1(y_i - x_i\beta_\tau) \mid z_i\right)$$

$$= \tau(1-\tau)$$

$$\frac{\partial}{\partial\beta} E\left[\varphi(y_i, x_i, \beta) \mid z_i\right]$$

$$= \frac{\partial}{\partial\beta} E\left[\tau - 1(y_i \le x_i\beta) \mid z_i\right]$$

$$= -\frac{\partial}{\partial\beta} E\left[P(y_i < x_i\beta \mid x_i, z_i) \mid z_i\right]$$

$$= E\left[x_i f_{y_i}(x_i\beta \mid x_i, z_i) \mid z_i\right]$$

$$= -E\left[f_\varepsilon(0 \mid x_i, z_i) x_i \mid z_i\right]$$

$$\hat{g}(z_i) = \frac{1}{\tau(1-\tau)} \hat{E}\left[\hat{f}_\varepsilon(0 \mid x_i, z_i) x_i \mid z_i\right]$$

$$= \frac{\frac{1}{n}\sum_{j=1}^{n} \frac{1}{h} k\left(\frac{y_j - x_j\hat{\beta}}{h}\right) x_j \frac{1}{h} k\left(\frac{z_j - z_i}{h}\right)}{\frac{1}{n}\sum_{j=1}^{n} \frac{1}{h} k\left(\frac{z_j - z_i}{h}\right)}$$

Note that for the special case

$X_i = Z_i$

Optimal IV $= \frac{1}{\tau(1-\tau)} E[X_i f(0|X_i)|X_i]$

$= \frac{1}{\tau(1-\tau)} X_i f(0|X_i)$

GMM:

$$\frac{1}{h}\sum_{i=3}^{n} \frac{1}{\tau(1-\tau)} X_i \hat{f}(0|X_i)\left[\tau - 1(y_i < x_i\beta)\right] \approx 0$$

This corresponds to "Weighted" Quantile Regression

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=3}^{n} \hat{f}(0|X_i) \cdot \rho_\tau(y_i - x_i\beta)$$

Feasible Weighted QR

Asymptotically $\approx$ infeasible weighted QR.

Random Sample Generation and Simulation of Probit Choice Probabilities

Based on sections 9.1-9.2 and 5.6 of Kenneth Train's
Discrete Choice Methods with Simulation

Presented by Jason Blevins
Applied Microeconometrics Reading Group
Duke University

21 June 2006

"Anyone attempting to generate random numbers by deterministic means is, of course, living in a state of sin."
—John Von Neumann, 1951

Outline

- Density simulation and sampling
  - Univariate
  - Truncated univariate
  - Multivariate Normal
  - Accept-Reject Method for truncated densities
  - Importance sampling
  - Gibbs sampling
  - The Metropolis-Hastings Algorithm
- Simulation of Probit Choice Probabilities
  - Accept-Reject Simulator
  - Smoothed AR Simulators
  - GHK Simulator

Simulation in Econometrics

- Goal: approximate a conditional expectation which lacks a closed form.
- Statistic of interest: $t(\epsilon)$, where $\epsilon \sim F$.
- Want to approximate $\mathbb{E}[t(\epsilon)] = \int t(\epsilon) f(\epsilon) d\epsilon$.
- Basic idea: calculate $t(\epsilon)$ for R draws of ϵ and take the average.
  - Unbiased: $\mathbb{E}\left[\frac{1}{R}\sum_{r=1}^{R} t(\epsilon^r)\right] = \mathbb{E}[t(\epsilon)]$
  - Consistent: $\frac{1}{R}\sum_{r=1}^{R} t(\epsilon^r) \xrightarrow{p} \mathbb{E}[t(\epsilon)]$
- This is straightforward *if* we can generate draws from F.
- In discrete choice models we want to simulate the probability that agent n chooses alternative i.
  - Utility: $U_{n,j} = V_{n,j} + \epsilon_{n,j}$ with $\epsilon_n \sim F(\epsilon_n)$.
  - $B_{n,i} = \{\epsilon_n \,|\, V_{n,i} + \epsilon_{n,i} > V_{n,j} + \epsilon_{n,j} \; \forall j \neq i\}$.
  - $P_{n,i} = \int \mathbb{1}_{B_{n,i}}(\epsilon_n) \, f(\epsilon_n) d\epsilon_n$.

Random Number Generators

- True Random Number Generators:
  - Collect entropy from system (keyboard, mouse, hard disk, *etc.*)
  - Unix: `/dev/random`, `/dev/urandom`
- Pseudo-Random Number Generators:
  - Linear Congruential Generators ($x_{n+1} = ax_n + b \mod c$): fast but predictable, good for Monte Carlo
  - Nonlinear: more difficult to determine parameters, used in cryptography
- Desirable properties for Monte Carlo work:
  - Portability
  - Long period
  - Computational simplicity
- DIEHARD Battery of Tests of Randomness, Marsaglia (1996)

Uniform and Standard Normal Generators

- Canned:
  - Matlab: `rand()`, `randn()`
  - Stata: `uniform()`, `invnormal(uniform())`
- Known algorithms:
  - Box-Muller algorithm
  - Marsaglia and Zaman (1994): `mzran`
  - Numerical Recipes, Press et al. (2002): `ran1`, `ran2`, `ran3`, `gasdev`

Simulating Univariate Distributions

- Direct vs. indirect methods.
- Transformation
  - Let $u \sim \mathrm{N}(0,1)$. Then $v = \mu + \sigma u \sim \mathrm{N}\left(\mu, \sigma^2\right)$ and
  - $w = \mathrm{e}^{\mu+\sigma u} \sim \mathrm{Lognormal}\left(\mu, \sigma^2\right)$.
- Inverse CDF transformation:
  - Let $u \sim \mathrm{N}(0,1)$. If $F(\epsilon)$ is invertible, then $\epsilon = F^{-1}(u) \sim F(\epsilon)$.
  - Only works for univariate distributions

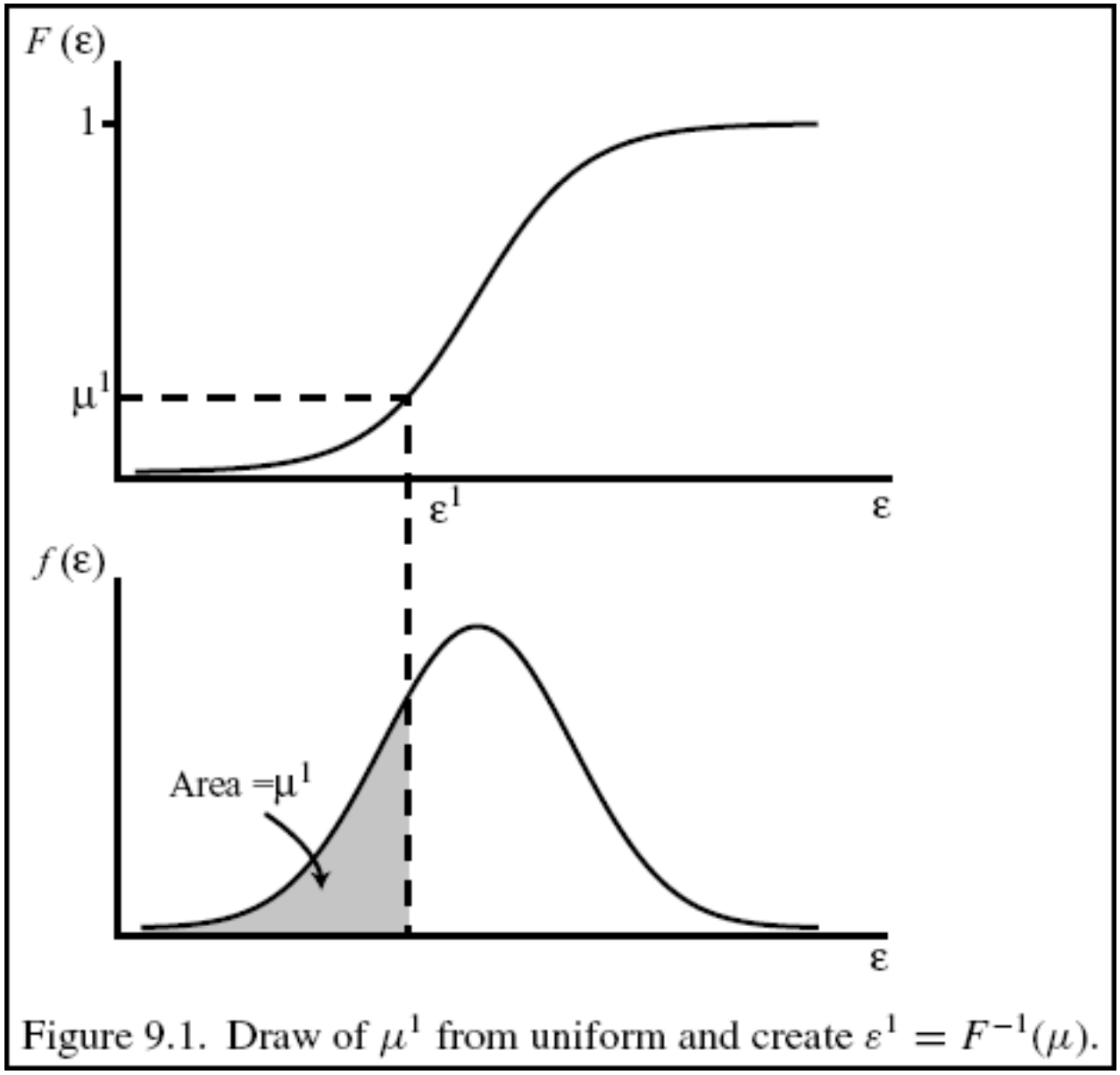


Figure 9.1. Draw of μ^1 from uniform and create $\varepsilon^1 = F^{-1}(\mu)$.

Truncated Univariate Distributions

- Want to draw from $g(\epsilon \mid a \leq \epsilon \leq b)$.
- Conditional density in terms of unconditional distribution $f(\epsilon)$:

$$g(\epsilon \mid a \leq \epsilon \leq b) = \begin{cases} \frac{f(\epsilon)}{F(b)-F(a)}, & \text{if} \quad a \leq \epsilon \leq b \\ 0, & \text{otherwise} \end{cases}$$

- Drawing is analogous to using the inverse CDF transformation.
- Let $\mu \sim \mathrm{U}(0,1)$ and define $\bar{\mu} = (1-\mu)F(a) + \mu F(b)$. $\epsilon = F^{-1}(\bar{\mu})$ is necessarily between a and b.

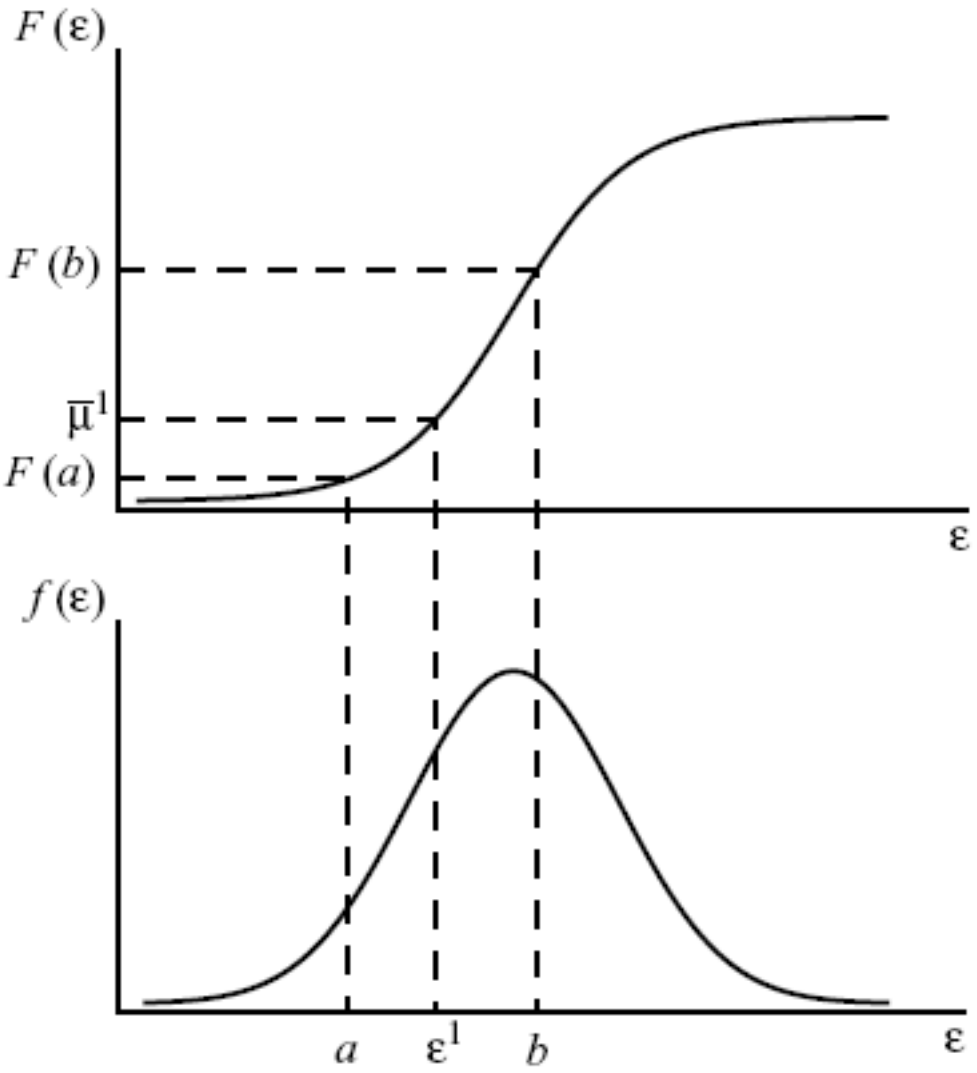


Figure 9.2. Draw of $\bar{\mu}^1$ between $F(a)$ and $F(b)$ gives draw ε^1 from $f(\varepsilon)$ between a and b.

The Multivariate Normal Distribution

- Assuming we can draw from $\mathrm{N}(0, 1)$, we can generate draws from any multivariate normal distribution $\mathrm{N}(\mu, \Omega)$.
- Let $LL^\top$ be the Cholesky decomposition of Ω and let $\eta \sim \mathrm{N}(0, I)$.
- Then, since a linear transformation of a Normal r.v. is also Normal:

$$\epsilon = \mu + L\eta \sim \mathrm{N}(\mu, \Omega)$$

$$\mathbb{E}[\epsilon] = \mu + L\mathbb{E}[\eta] = \mu$$

$$\begin{aligned}
\mathrm{Var}(\epsilon) &= \mathbb{E}\left[(L\eta)(L\eta)^\top\right] \\
&= \mathbb{E}\left[L\eta\eta^\top L^\top\right] \\
&= L\,\mathbb{E}\left[\eta\eta^\top\right] L^\top \\
&= L\,\mathrm{Var}(\eta)\, L^\top = \Omega
\end{aligned}$$

The Accept-Reject Method for Truncated Densities

- Want to draw from a multivariate density $g(\epsilon)$, but truncated so that $a \leq \epsilon \leq b$ with $a, b, \epsilon \in \mathbb{R}^l$.
- The truncated density is $f(\epsilon) = \frac{1}{k} g(\epsilon)$ for some normalizing constant k.
- Accept-Reject method:
  - Draw ϵ^r from $f(\epsilon)$.
  - Accept if $a \leq \epsilon^r \leq b$, reject otherwise.
  - Repeat for $r = 1, \ldots, R$.
- Accept on average kR draws.
- If we can draw from f, then we can draw from g without knowing k.
- Disadvantages:
  - Size of resulting sample is random if R is fixed.
  - Hard to determine required R.
  - Positive probability that no draws will be accepted.
- Alternatively, fix the number of draws to accept and repeat until satisfied.

Importance Sampling

- Want to draw from f but drawing from g is easier.
- Transform the target expectation into an integral over g:

$$\int t(\epsilon) f(\epsilon) d\epsilon = \int t(\epsilon) \frac{f(\epsilon)}{g(\epsilon)} g(\epsilon) d\epsilon.$$

- Importance Sampling: Draw ϵ^r from g and weight by $\frac{f(\epsilon^r)}{g(\epsilon^r)}$.
- The weighted draws constitute a sample from f.
- The support of g must cover that of f and $\sup \frac{f}{g}$ must be finite.
- To show equivalence, consider the CDF of the weighted draws:

$$\begin{aligned} \int \frac{f(\epsilon)}{g(\epsilon)} \mathbb{1}(\epsilon < m)\, g(\epsilon) d\epsilon &= \int_{-\infty}^{m} \frac{f(\epsilon)}{g(\epsilon)} g(\epsilon) d\epsilon \\ &= \int_{-\infty}^{m} f(\epsilon) d\epsilon = F(m) \end{aligned}$$

The Gibbs Sampler

- Used when it is difficult to draw from a joint distribution but easy to draw from the conditional distribution.
- Consider a bivariate case: $f(\epsilon_1, \epsilon_2)$.
- Drawing iteratively from conditional densities converges to draws from the joint distribution.
- The Gibbs Sampler: Choose an initial value ϵ_1^0.
  - Draw $\epsilon_2^0 \sim f_2(\epsilon_2 \mid \epsilon_1^0)$, $\epsilon_1^1 \sim f_1(\epsilon_1 \mid \epsilon_2^0), \ldots,$ $\epsilon_1^t \sim f_1(\epsilon_1 \mid \epsilon_2^{t-1})$, $\epsilon_2^t \sim f_2(\epsilon_2 \mid \epsilon_1^t)$.
  - The sequence of draws $\{(\epsilon_1^0, \epsilon_2^0), \ldots, (\epsilon_1^t, \epsilon_2^t)\}$ converges to draws from $f(\epsilon_1, \epsilon_2)$.
- See Casella and George (1992) or Judd (1998).

The Gibbs Sampler: Example

- $\epsilon_1, \epsilon_2 \sim \mathrm{N}(0, 1)$.
- Truncation: $\epsilon_1 + \epsilon_2 \leq m$.
- Ignoring truncation, $\epsilon_1 \,|\, \epsilon_2 \sim \mathrm{N}(0, 1)$.
- Truncated univariate sampling:

$$\mu \sim \mathrm{U}(0, 1)$$
$$\bar{\mu} = (1 - \mu)\Phi(0) + \mu\Phi(m - \epsilon_2)$$
$$\epsilon_1 = \Phi^{-1}(\mu\Phi(m - \epsilon_2))$$

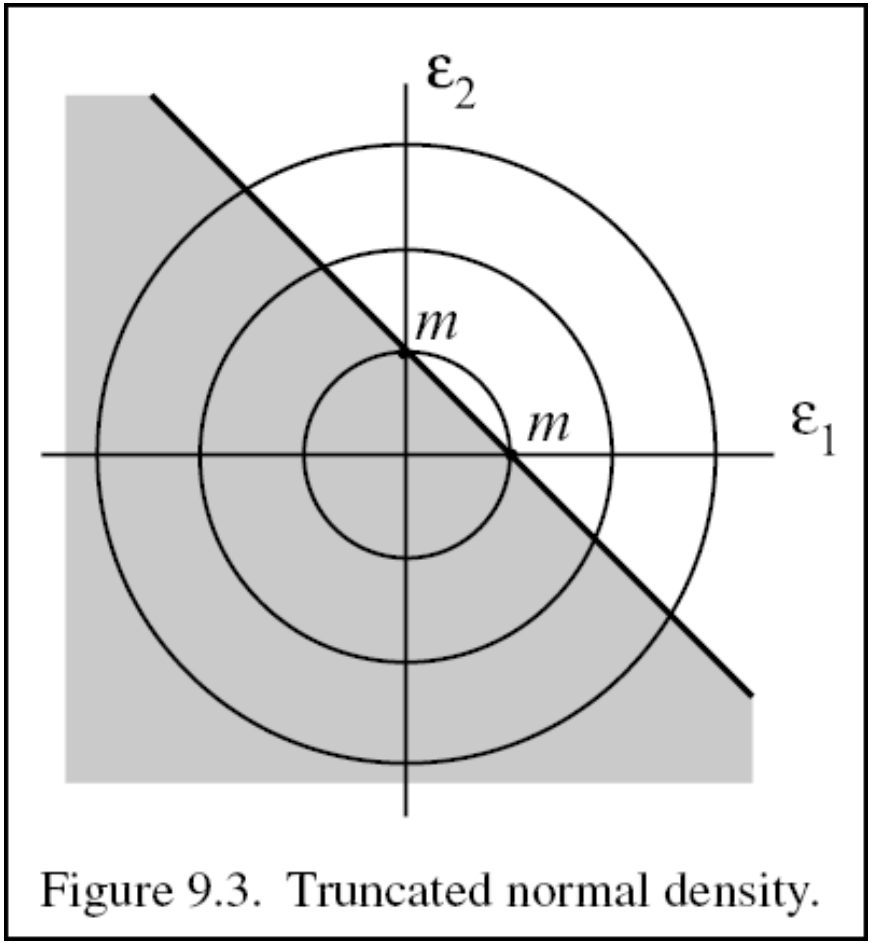


Figure 9.3. Truncated normal density.

The Metropolis-Hastings Algorithm

- Only requires being able to evaluate f and draw from g.
- Metropolis-Hastings Algorithm:
  1. Let ϵ^0 be some initial value.
  2. Choose a trial value $\tilde{\epsilon}^1 = \epsilon^0 + \eta$, $\eta \sim g(\eta)$, where g has zero mean.
  3. If $f(\tilde{\epsilon}^1) > f(\epsilon^0)$, accept $\tilde{\epsilon}^1$.
  4. Otherwise, accept $\tilde{\epsilon}^1$ with probability $f(\tilde{\epsilon}^1)/f(\epsilon^0)$.
  5. Repeat for many iterations.
- The sequence $\{\epsilon^t\}$ converges to draws from f.
- Useful for sampling truncated densities when the normalizing factor is unknown.
- Description of algorithm: Chib and Greenberg (1995)

Calculating Probit Choice Probabilities

- Probit Model:
  - Utility: $U_{n,j} = V_{n,j} + \epsilon_{n,j}$ with $\epsilon_n \sim \mathrm{N}(0, \Omega)$.
  - $B_{n,i} = \{\epsilon_n \,|\, V_{n,i} + \epsilon_{n,i} > V_{n,j} + \epsilon_{n,j} \; \forall j \neq i\}$.
  - $P_{n,i} = \int_{B_{n,i}} \phi(\epsilon_n) d\epsilon_n$.
- Non-simulation methods:
  - Quadrature: approximate the integral using a specifically chosen set of evaluation points and weights (Geweke, 1996, Judd, 1998).
  - Clark algorithm: maximum of several normal r.v. is itself approximately normal (Clark, 1961, Daganzo et al., 1977).
- Simulation methods:
  - Accept-reject method
  - Smoothed accept-reject
  - GHK (Geweke-Hajivassiliou-Keane)

The Accept-Reject Simulator

- Straightforward:
  1. Draw from distribution of unobservables.
  2. Determine the agent's preferred alternative.
  3. Repeat R times.
  4. The simulated choice probability for alternative i is the proportion of times the agent chooses alternative i.
- General:
  - Applicable to any discrete choice model.
  - Works with any distribution that can be drawn from.

The Accept-Reject Simulator for Probit

- Let $B_{n,i} = \{\epsilon_n \,|\, V_{n,i} + \epsilon_{n,i} > V_{n,j} + \epsilon_{n,j}, \ \forall j \neq i\}$. The Probit choice probabilities are:

$$P_{n,i} = \int \mathbb{1}_{B_{n,i}}(\epsilon_n)\, \phi(\epsilon_n) d\epsilon_n.$$

- Accept-Reject Method:
  1. Take R draws $\{\epsilon_n^1, \ldots, \epsilon_n^R\}$ from $\mathrm{N}(0, \Omega)$ using the Cholesky decomposition $LL^\top = \Omega$ to transform *iid* draws from $\mathrm{N}(0, 1)$.
  2. Calculate the utility for each alternative: $U_{n,j}^r = V_{n,j} + \epsilon_{n,j}^r$.
  3. Let $d_{n,j}^r = 1$ if alternative j is chosen and zero otherwise.
  4. The simulated choice probability for alternative i is:

$$\hat{P}_{n,i} = \frac{1}{R} \sum_{r=1}^{R} d_{n,i}^r$$

The Accept-Reject Simulator: Evaluation

- Main advantages: simplicity and generality.
- Can also be applied to the error differences in discrete choice models.
  - Slightly faster
  - Conceptually more difficult
- Disadvantages:
  - $\hat{P}_{n,i}$ will be zero with positive probability.
  - $\hat{P}_{n,i}$ is a step function and the simulated log-likelihood is not differentiable.
  - Gradient methods are likely to fail (gradient is either 0 or undefined).

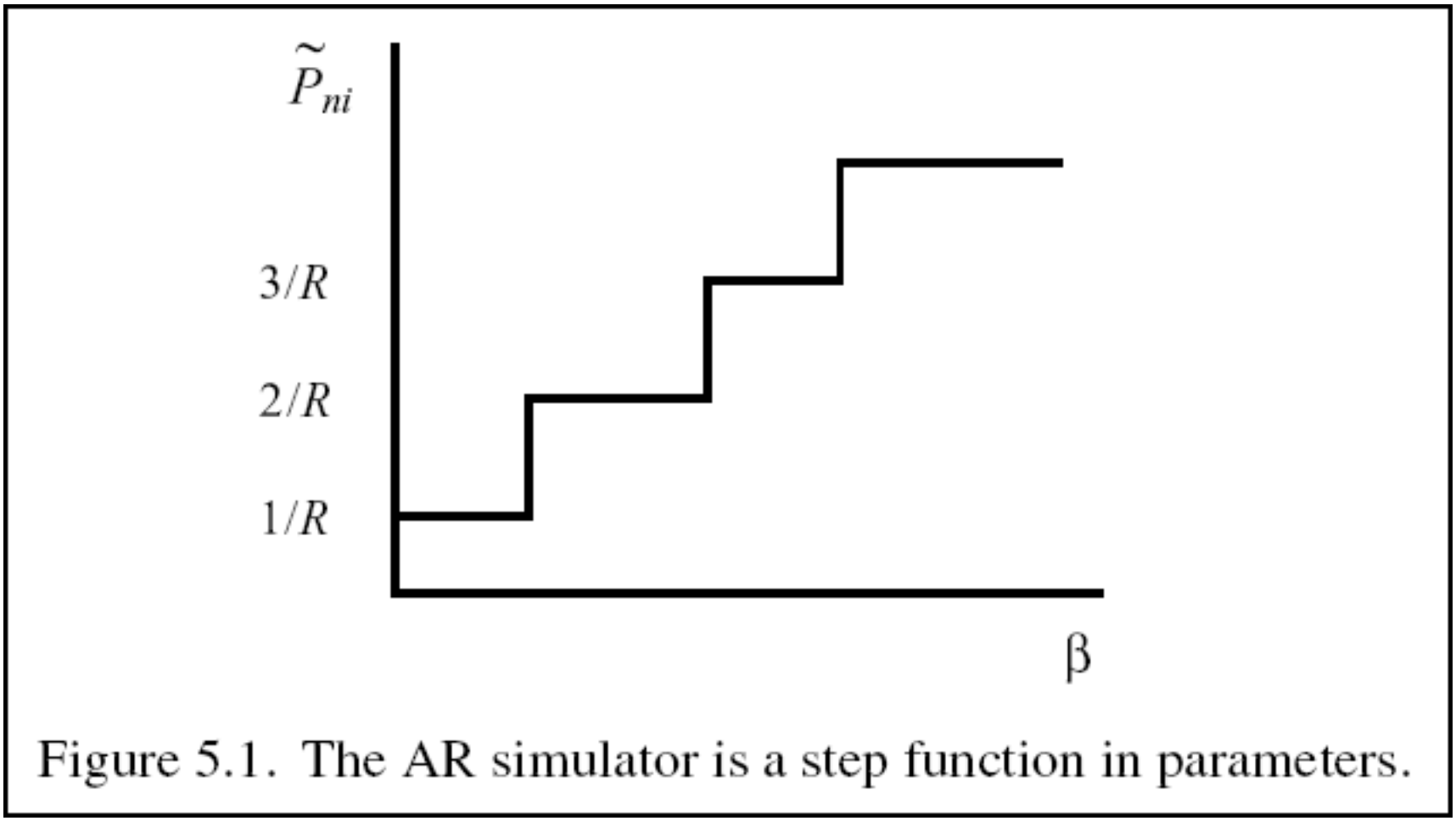


Figure 5.1. The AR simulator is a step function in parameters.

The Smoothed Accept-Reject Simulator

- Replace the indicator function with a general function of $U_{n,j}$ for $j = 1, \ldots, J$ that is:
  - increasing in $U_{n,i}$ and decreasing in $U_{n,j}$ for $j \neq i$,
  - strictly positive, and
  - twice differentiable.
- McFadden (1989) suggested the Logit-smoothed AR simulator:
  1. Draw $\epsilon_n^r \sim \mathrm{N}(0, \Omega)$, for $r = 1, \ldots, R$.
  2. Calculate $U_{n,j}^r = V_{n,j} + \epsilon_{n,j}^r \quad \forall j, r$.
  3. Calculate the smoothed choice function for each simulation to find $\hat{P}_{n,i}$:

$$S_i^r = \frac{\exp(U_{n,i}^r/\lambda)}{\sum_{j=1}^{J} \exp(U_{n,j}^r/\lambda)},$$

$$\hat{P}_{n,i} = \frac{1}{R} \sum_{r=1}^{R} S_i^r$$

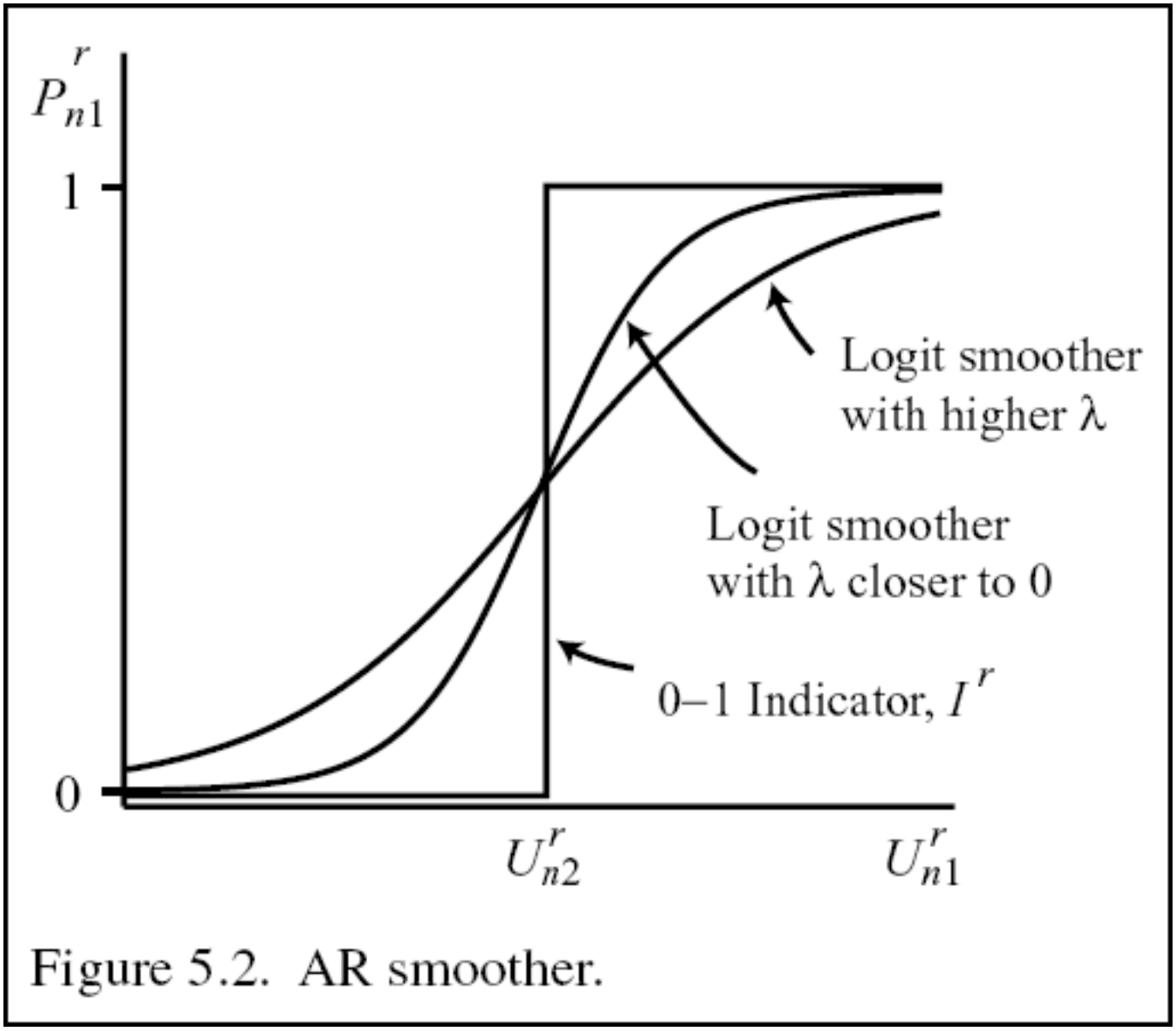


Figure 5.2. AR smoother.

The Smoothed Accept-Reject Simulator: Evaluation

- Simulated log-likelihood using smoothed choice probabilities is... smooth.
- Slightly more difficult to implement than AR simulator.
- Can provide a behavioral interpretation.
- Choice of smoothing parameter λ is arbitrary.
- Objective function is modified.
- Use alternative optimization methods instead (simulated annealing)?

The GHK Simulator

- GHK: Geweke, Hajivassiliou, Keane.
- Simulates the Probit model in differenced form.
- For each i, simulation of $P_{n,i}$ uses utility differences relative to $U_{n,i}$.
- Basic idea: write the choice probability as a product of conditional probabilities.
- We are much better at simulating univariate integrals over $N(0,1)$ than those over multivariate normal distributions.

GHK with Three Alternatives

- An example with three alternatives:

$$U_{n,j} = V_{n,j} + \epsilon_{n,j},\ j = 1, 2, 3 \quad \text{with} \quad \epsilon_n \sim \mathrm{N}(0, \Omega)$$

- Assume Ω has been normalized for identification.
- Consider $P_{n,1}$. Difference with respect to $U_{n,1}$:

$$\tilde{U}_{n,j,1} = \tilde{V}_{n,j,1} + \tilde{\epsilon}_{n,j,1},\ j = 2, 3 \quad \text{with} \quad \tilde{\epsilon}_{n,1} \sim \mathrm{N}\left(0, \tilde{\Omega}_1\right)$$

$$P_{n,1} = \mathbb{P}\left(\tilde{U}_{n,2,1} < 0, \tilde{U}_{n,3,1} < 0\right) = \mathbb{P}\left(\tilde{V}_{n,2,1} + \tilde{\epsilon}_{n,2,1} < 0, \tilde{V}_{n,3,1} + \tilde{\epsilon}_{n,3,1} < 0\right)$$

- $P_{n,1}$ is still hard to evaluate because $\tilde{\epsilon}_{n,j,1}$'s are correlated.

GHK with Three Alternatives

- One more transformation. Let $L_1 L_1^\top$ be the Cholesky decomposition of $\tilde{\Omega}_1$:

$$L_1 = \begin{pmatrix} c_{aa} & 0 \\ c_{ab} & c_{bb} \end{pmatrix}$$

- Then we can express the errors as:

$$\tilde{\epsilon}_{n,2,1} = c_{aa}\eta_1$$
$$\tilde{\epsilon}_{n,3,1} = c_{ab}\eta_1 + c_{bb}\eta_2$$

  where η_1, η_2 are *iid* $\mathrm{N}(0,1)$.
- The differenced utilities are then

$$\tilde{U}_{n,2,1} = \tilde{V}_{n,2,1} + c_{aa}\eta_1$$
$$\tilde{U}_{n,3,1} = \tilde{V}_{n,3,1} + c_{ab}\eta_1 + c_{bb}\eta_2$$

GHK with Three Alternatives

- $P_{n,1}$ is easier to simulate now:

$$\begin{aligned} P_{n,1} &= \mathbb{P}\left(\tilde{V}_{n,2,1} + c_{aa}\eta_1 < 0, \tilde{V}_{n,3,1} + c_{ab}\eta_1 + c_{bb}\eta_2 < 0\right) \\ &= \mathbb{P}\left(\eta_1 < -\frac{\tilde{V}_{n,2,1}}{c_{aa}}\right) \mathbb{P}\left(\eta_2 < -\frac{\tilde{V}_{n,3,1} + c_{ab}\eta_1}{c_{bb}} \,\middle|\, \eta_1 < -\frac{\tilde{V}_{n,2,1}}{c_{aa}}\right) \\ &= \Phi\left(-\frac{\tilde{V}_{n,2,1}}{c_{aa}}\right) \int_{-\infty}^{-\tilde{V}_{n,2,1}/c_{aa}} \Phi\left(-\frac{\tilde{V}_{n,3,1} + c_{ab}\eta_1}{c_{bb}}\right) \phi(\eta_1) d\eta_1 \end{aligned}$$

- First term only requires evaluating the standard Normal CDF.
- Integral is over a truncated univariate standard Normal distribution.
- The 'statistic' in this case is the standard Normal CDF.

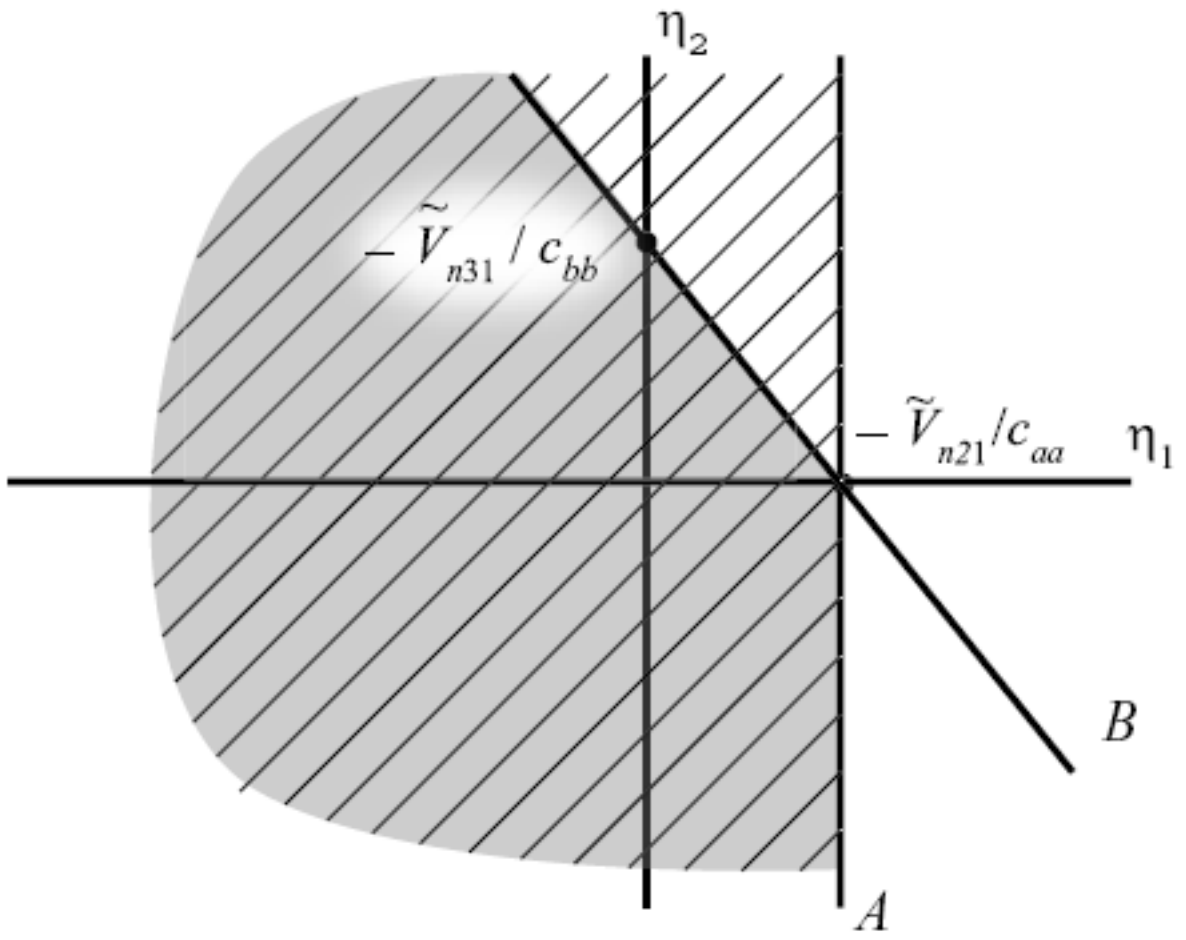


Figure 5.3. Probability of alternative 1.

GHK with Three Alternatives: Simulation

$$\Phi\left(-\frac{\tilde{V}_{n,2,1}}{c_{aa}}\right)\int_{-\infty}^{-\frac{\tilde{V}_{n,2,1}}{c_{aa}}}\Phi\left(-\frac{\tilde{V}_{n,3,1}+c_{ab}\eta_1}{c_{bb}}\right)\phi(\eta_1)d\eta_1 = k\int_{-\infty}^{\bar{\eta}_1} t(\eta_1)\phi(\eta_1)d\eta_1$$

1. Calculate $k = \Phi\left(-\frac{\tilde{V}_{n,2,1}}{c_{aa}}\right)$.

2. Draw η_1^r from $\mathrm{N}(0,1)$ truncated at $-\tilde{V}_{n,2,1}/c_{aa}$ for $r = 1,\ldots,R$: Draw $\mu^r \sim \mathrm{U}(0,1)$ and calculate $\eta_1^r = \Phi^{-1}\left(\mu^r\Phi\left(-\frac{\tilde{V}_{n,2,1}}{c_{aa}}\right)\right)$.

3. Calculate $t^r = \Phi\left(-\frac{\tilde{V}_{n,3,1}+c_{ab}\eta_1^r}{c_{bb}}\right)$ for $r = 1,\ldots,R$.

4. The simulated choice probability is $\hat{P}_{n,1} = k\frac{1}{R}\sum_{r=1}^{R} t^r$

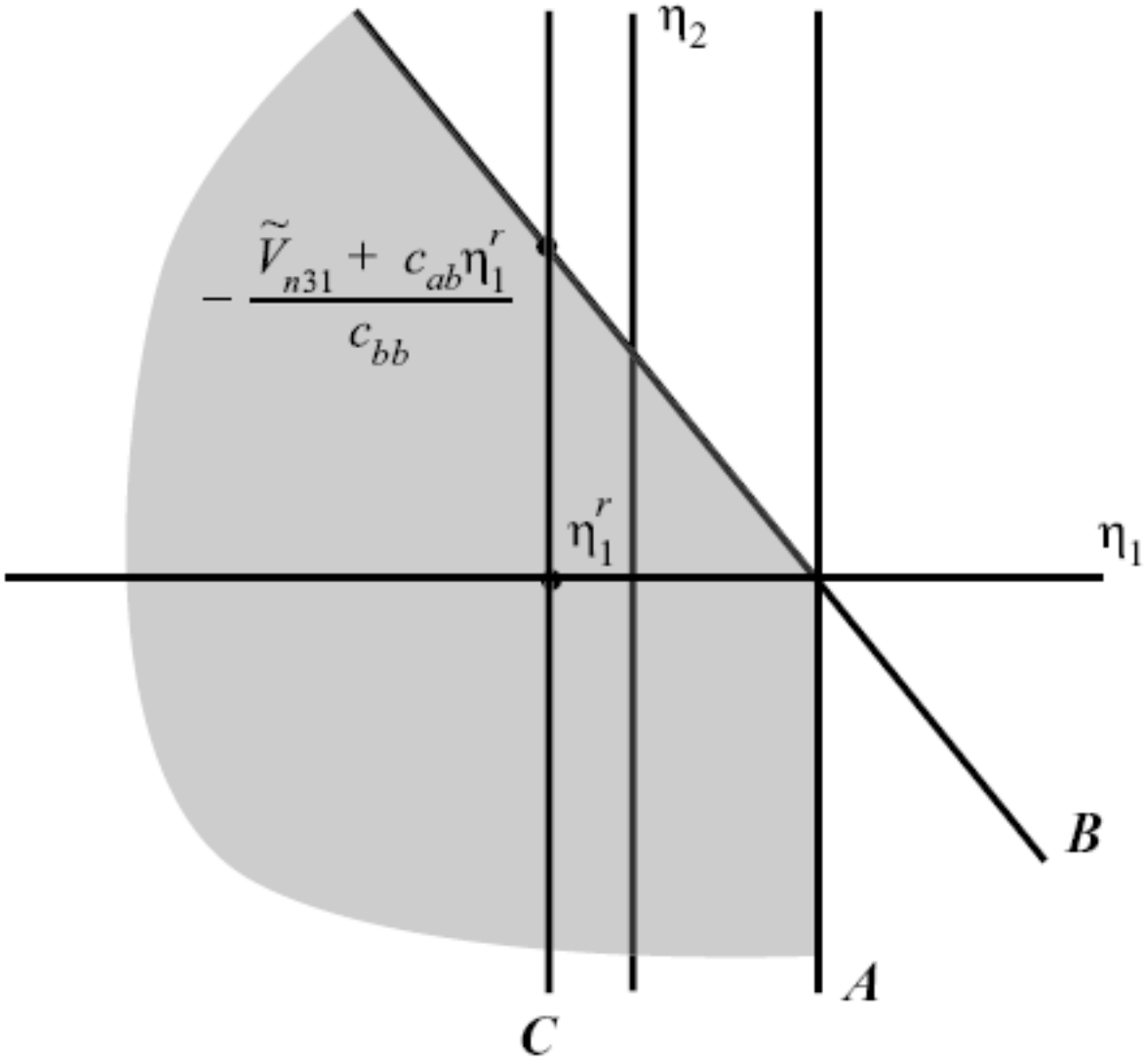


Figure 5.4. Probability that η_2 is in the correct range, given η_1^r.

GHK as Importance Sampling

$$P_{n,1} = \int \mathbb{1}_B(\eta)\, g(\eta) d\eta$$

where $B = \{\eta \,|\, \tilde{U}_{n,j,i} < 0 \; \forall j \neq i\}$ and $g(\eta)$ is the standard Normal PDF.

- Direct (AR) simulation involves drawing from g and calculating $\mathbb{1}_B(\eta)$.
- GHK draws from a different density $f(\eta)$ (the truncated normal):

$$f(\eta) = \begin{cases} \frac{\phi(\eta_1)}{\Phi(-\tilde{V}_{n,1,i}/c_{11})} \frac{\phi(\eta_2)}{\Phi(-(\tilde{V}_{n,2,i}+c_{21}\eta_1)/c_{22})} \cdots, & \text{if} \quad \eta \in B \\ 0, & \text{otherwise} \end{cases}$$

- Define $\hat{P}_{i,n}(\eta) = \Phi(-\tilde{V}_{n,1,i}/c_{11})\Phi(-(\tilde{V}_{n,2,i} + c_{21}\eta_1)/c_{22}) \cdots$.
- $f(\eta) = g(\eta)/\hat{P}_{n,i}(\eta)$ on B.
- $P_{n,i} = \int \mathbb{1}_B(\eta)\, g(\eta) d\eta = \int \mathbb{1}_B(\eta) \frac{g(\eta)}{g(\eta)/\hat{P}_{i,n}(\eta)} f(\eta) d\eta = \int \hat{P}_{i,n}(\eta) f(\eta) d\eta$

References

George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46:167–174, 1992.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–335, 1995.

Charles E. Clark. The greatest of a finite set of random variables. *Operations Research*, 9:145–162, 1961.

Carlos F. Daganzo, Fernando Bouthelier, and Yosef Sheffi. Multinomial probit and qualitative choice: A computationally efficient algorithm. *Transportation Science*, 11:338–358, 1977.

John Geweke. Monte Carlo simulation and numerical integration. In Hans M. Amman, David A. Kendrick, and John Rust, editors, *Handbook of Computational Economics*, volume 1, Amsterdam, 1996. North Holland.

Kenneth L. Judd. *Numerical Methods in Economics*. MIT Press, Cambridge, MA, 1998.

George Marsaglia. DIEHARD: A battery of tests of randomness. `http://www.csis.hku.hk/~diehard`, 1996.

George Marsaglia and Arif Zaman. Some portable very-long-period random number generators. *Computers in Physics*, 8:117–121, 1994.

Daniel McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57:995–1026, 1989.

William H. Press, William T. Vetterling, Saul A. Teukolsky, and Brian P. Flannery. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2002.

Markov Chain Monte Carlo Methods

John Geweke

Department of Economics
University of Iowa

Presentation at ICE 06, Chicago, July, 2006

Part I

- $\theta^{(m)} \sim p\left(\theta|\theta^{(m-1)}, C\right), \quad (m = 1, 2, 3, \ldots)$
- If C is specified correctly, then

$$\theta^{(m-1)} \sim p\left(\theta|I\right), \theta^{(m)} \sim p\left(\theta|\theta^{(m-1)}, C\right) \Longrightarrow \theta^{(m)} \sim p\left(\theta|I\right).$$

- Better yet, if

$$\theta^{(m-1)} \sim p\left(theta|J\right), \theta^{(m)} \sim p\left(\theta|\theta^{(m-1)}, C\right) \Longrightarrow \theta^{(m)} \sim p\left(\theta|J\right)$$

then $J = I$. And even better,

$$p\left(\theta^{(m)}|\theta^{(0)}, C\right) \xrightarrow{d} p\left(\theta|I\right), \quad \forall \theta^{(0)} \in \Theta.$$

- If $p\left(\theta^{(m)}|\theta^{(0)}, C\right) \xrightarrow{d} p\left(\theta|I\right), \quad \forall\theta^{(0)} \in \Theta$, then we can approxiate $E\left[h\left(\omega\right)|I\right]$ by
  - iterating the chain B ("burn-in") times:
  - drawing $\omega^{(m)} \sim p\left(\omega|\theta^{(m)}\right), (m = 1, \ldots, M)$;
  - Computing

$$\bar{h}^M = M^{-1} \sum_{m=1}^{M} h\left(\omega^{(m)}\right).$$

- Blocking: $\theta' = \left(\theta'_{(1)}, \ldots, \theta'_{(B)}\right)$.
- Some notation: corresponding to any subvector $\theta_{(b)}$,

$$\theta'_{<(b)} = \left(\theta'_{(1)}, \ldots, \theta'_{(b-1)}\right), (b = 2, \ldots, B), \theta_{<(1)} = \emptyset$$
$$\theta'_{>(b)} = \left(\theta'_{(b+1)}, \ldots, \theta'_{(B)}\right), (b = 1, \ldots, B-1), \theta_{>(B)} = \emptyset$$
$$\theta'_{-(b)} = \left(\theta'_{<(b)}, \theta'_{>(b)}\right)$$

- Very important: choose the blocking so that

$$\theta_{(b)} \sim p\left(\theta_{(b)} | \theta_{-(b)}, I\right)$$

is possible.

Imagine $\theta^{(0)} \sim p\left(\theta|I\right)$, and then in succession

$$
\begin{aligned}
\theta^{(1)}_{(1)} &\sim p\left(\theta_{(1)}|\theta^{(0)}_{-(1)}, I\right), \\
\theta^{(1)}_{(2)} &\sim p\left(\theta_{(2)}|\theta^{(1)}_{<(2)}, \theta^{(0)}_{>(2)}, I\right), \\
\theta^{(1)}_{(3)} &\sim p\left(\theta_{(3)}|\theta^{(1)}_{<(3)}, \theta^{(0)}_{>(3)}, I\right), \\
&\vdots, \\
\theta^{(1)}_{(b)} &\sim p\left(\theta_{(b)}|\theta^{(1)}_{<(b)}, \theta^{(0)}_{>(b)}, I\right), \\
&\vdots, \\
\theta^{(1)}_{(B)} &\sim p\left(\theta_{(B)}|\theta^{(1)}_{<(B)}, \theta^{(0)}_{>(B)}, I\right)
\end{aligned}
$$

We have $\theta^{(1)} \sim p\left(\theta|I\right)$.

Now repeat

$$
\begin{aligned}
\theta_{(1)}^{(2)} &\sim p\left(\theta_{(1)}|\theta_{-(1)}^{(1)}, I\right), \\
\theta_{(2)}^{(2)} &\sim p\left(\theta_{(2)}|\theta_{<(2)}^{(2)}, \theta_{>(2)}^{(1)}, I\right), \\
\theta_{(3)}^{(2)} &\sim p\left(\theta_{(3)}|\theta_{<(3)}^{(2)}, \theta_{>(3)}^{(1)}, I\right), \\
&\vdots \\
\theta_{(b)}^{(2)} &\sim p\left(\theta_{(b)}|\theta_{<(b)}^{(2)}, \theta_{>(b)}^{(1)}, I\right), \\
&\vdots \\
\theta_{(B)}^{(2)} &\sim p\left(\theta_{(B)}|\theta_{<(B)}^{(2)}, \theta_{>(B)}^{(1)}, I\right).
\end{aligned}
$$

We have $\theta^{(2)} \sim p\left(\theta|I\right)$.

- The general step in the Gibbs sampler is

$$\theta_{(b)}^{(m)} \sim p\left(\theta_{(b)} | \theta_{<(b)}^{(m)}, \theta_{>(b)}^{(m-1)}, I\right)$$

  for $b = 1, \ldots, B$ and $m = 1, 2, \ldots$

- This defines the Markov chain

$$p\left(\theta^{(m)} | \theta^{(m-1)}, G\right) = \prod_{b=1}^{B} p\left[\theta_{(b)}^{(m)} | \theta_{<(b)}^{(m)}, \theta_{>(b)}^{(m-1)}, I\right].$$

- Key property:

$$\theta^{(0)} \sim p\left(\theta | I\right) \Rightarrow \theta^{(m)} \sim p\left(\theta | I\right).$$

- Potential problems: disjoint support.

- What it does: $\theta^* \sim q\left(\theta^* | \theta^{(m-1)}, H\right)$
- Then

$$P\left(\theta^{(m)} = \theta^*\right) = \alpha\left(\theta^* | \theta^{(m-1)}, H\right)$$
$$P\left(\theta^{(m)} = \theta^{(m-1)}\right) = 1 - \alpha\left(\theta^* | \theta^{(m-1)}, H\right)$$

where

$$\alpha\left(\theta^* | \theta^{(m-1)}, H\right) = \min\left\{\frac{p\left(\theta^* | I\right) / q\left(\theta^* | \theta^{(m-1)}, H\right)}{p\left(\theta^{(m-1)} | I\right) / q\left(\theta^{(m-1)} | \theta^*, H\right)}, 1\right\}.$$

- If we define

$$u\left(\theta^{*}|\theta,H\right)=q\left(\theta^{*}|\theta,H\right)\alpha\left(\theta^{*}|\theta,H\right)$$

- then

$$\begin{aligned}P\left(\theta^{(m)}=\theta^{(m-1)}|\theta^{(m-1)}=\theta,H\right)=&r\left(\theta|H\right)\\=&1-\int_{\Theta}u\left(\theta^{*}|\theta,H\right)d\nu\left(\theta^{*}\right).\end{aligned}$$

- Notice that

$$P\left(\theta^{(m)}\in A|\theta^{(m-1)}=\theta,H\right)=\int_{A}u\left(\theta^{*}|\theta,H\right)d\nu\left(\theta^{*}\right)+r\left(\theta|H\right)I_{A}\left(\theta\right).$$

$$u\left(\theta^{*}|\theta, H\right) = q\left(\theta^{*}|\theta, H\right)\alpha\left(\theta^{*}|\theta, H\right)$$

We can write the transition density in one line making use of the Dirac delta function, an operator with the property

$$\int_{A} \delta_{\theta}\left(\theta^{*}\right) f\left(\theta^{*}\right) d\nu\left(\theta^{*}\right) = f\left(\theta\right) I_{A}\left(\theta\right)$$

Then

$$\begin{aligned} p\left(\theta^{(m)}|\theta^{(m-1)}, H\right) =& u\left(\theta^{(m)}|\theta^{(m-1)}, H\right) \\ &+ r\left(\theta^{(m-1)}|H\right)\delta_{\theta^{(m-1)}}\left(\theta^{(m)}\right). \end{aligned}$$

$$\alpha\left(\theta^{*}|\theta^{(m-1)},H\right)=\min\left\{\frac{p\left(\theta^{*}|I\right)/q\left(\theta^{*}|\theta^{(m-1)},H\right)}{p\left(\theta^{(m-1)}|I\right)/q\left(\theta^{(m-1)}|\theta^{*},H\right)},1\right\}$$

- Special case 1, original Metropolis (1953):

$$q\left(\theta^{*}|\theta,H\right)$$
$$\Longrightarrow\alpha\left(\theta^{*}|\theta^{(m-1)},H\right)=\min\left[p\left(\theta^{*}|I\right),1\right]$$

- Important example: random walk Metropolos chain

$$q\left(\theta^{*}|\theta,H\right)=q\left(\theta^{*}-\theta|H\right),$$

where $q\left(\cdot|H\right)$ is symmetric about zero.

$$\alpha\left(\theta^{*}|\theta^{(m-1)},H\right)=\min\left\{\frac{p\left(\theta^{*}|I\right)/q\left(\theta^{*}|\theta^{(m-1)},H\right)}{p\left(\theta^{(m-1)}|I\right)/q\left(\theta^{(m-1)}|\theta^{*},H\right)},1\right\}$$

- Special case 2, *Metropolis independence chain*:

$$q\left(\theta^{*}|\theta,H\right)=q\left(\theta^{*}|H\right)$$

$$\Longrightarrow\alpha\left(\theta^{*}|\theta^{(m-1)},H\right)=\min\left\{\frac{p\left(\theta^{*}|I\right)/q\left(\theta^{*}|H\right)}{p\left(\theta^{(m-1)}|I\right)/q\left(\theta^{(m-1)}|H\right)},1\right\}$$

$$=\min\left\{\frac{w\left(\theta^{*}\right)}{w\left(\theta^{(m-1)}\right)},1\right\}$$

where $w\left(\theta\right)=p\left(\theta|I\right)/q\left(\theta|H\right)$.

- A two part argument - Part 1:
- Suppose any transition probability density function $p\left(\theta^{(m)}|\theta^{(m-1)}, T\right)$ satisfies the *reversibility condition*

$$p\left(\theta^{(m-1)}|I\right) p\left(\theta^{(m)}|\theta^{(m-1)}, T\right) = p\left(\theta^{(m)}|I\right) p\left(\theta^{(m-1)}|\theta^{(m)}, T\right)$$

with respect to $p(\theta|I)$. Then

$$\begin{aligned}
&\int_{\Theta} p\left(\theta^{(m-1)}|I\right) p\left(\theta^{(m)}|\theta^{(m-1)}, T\right) \, d\nu\left(\theta^{(m-1)}\right) \\
&= \int_{\Theta} p\left(\theta^{(m)}|I\right) p\left(\theta^{(m-1)}|\theta^{(m)}, T\right) \, d\nu\left(\theta^{(m-1)}\right) \\
= p\left(\theta^{(m)}|I\right) &\int_{\Theta} p\left(\theta^{(m-1)}|\theta^{(m)}, T\right) \, d\nu\left(\theta^{(m-1)}\right) = p\left(\theta^{(m)}|I\right).
\end{aligned}$$

and so $p(\theta|I)$ is an *invariant density* of the Markov chain.

- Part 2 of the argument (How Hastings did it):
- Suppose we don't know the probability $\alpha\left(\theta^{*}|\theta^{(m-1)}, H\right)$, but we want $p\left(\theta^{(m)}|\theta^{(m-1)}, H\right)$to be reversible with respect to $p\left(\theta|I\right)$:
-

$$p\left(\theta^{(m-1)}|I\right) p\left(\theta^{(m)}|\theta^{(m-1)}, H\right) = p\left(\theta^{(m)}|I\right) p\left(\theta^{(m-1)}|\theta^{(m)}, H\right).$$

- Trivial if $\theta^{(m-1)} = \theta^{(m)}$. For $\theta^{(m-1)} \neq \theta^{(m)}$ we need

$$\begin{aligned} & p\left(\theta^{(m-1)}|I\right) q\left(\theta^{*}|\theta^{(m-1)}, H\right) \alpha\left(\theta^{*}|\theta^{(m-1)}, H\right) \\ & = p\left(\theta^{*}|I\right) q\left(\theta^{(m-1)}|\theta^{*}, H\right) \alpha\left(\theta^{(m-1)}|\theta^{*}, H\right). \end{aligned}$$

- $$p\left(\theta^{(m-1)}|I\right) q\left(\theta^{*}|\theta^{(m-1)}, H\right) \alpha\left(\theta^{*}|\theta^{(m-1)}, H\right)$$
$$= p\left(\theta^{*}|I\right) q\left(\theta^{(m-1)}|\theta^{*}, H\right) \alpha\left(\theta^{(m-1)}|\theta^{*}, H\right).$$

- Suppose without loss of generality that

$$p\left(\theta^{(m-1)}|I\right) q\left(\theta^{*}|\theta^{(m-1)}, H\right) > p\left(\theta^{*}|I\right) q\left(\theta^{(m-1)}|\theta^{*}, H\right).$$

- Set $\alpha\left(\theta^{(m-1)}|\theta^{*}, H\right) = 1$ and

$$\alpha\left(\theta^{*}|\theta^{(m-1)}, H\right) = \frac{p\left(\theta^{*}|I\right) q\left(\theta^{(m-1)}|\theta^{*}, H\right)}{p\left(\theta^{(m-1)}|I\right) q\left(\theta^{*}|\theta^{(m-1)}, H\right)}.$$

Why does the Metropolis-Hastings algorithm work?

- The goal is to verify the reversibility condition:

$$f(x) f\left(x'|x\right) = f\left(x'\right) f\left(x|x'\right)$$

- Note that according to the Gibbs sampler:

$$f\left(x'|x\right) = \int f\left(x'|y\right) f\left(y|x\right) dy = \int \frac{f\left(x', y\right) f\left(x, y\right)}{f\left(y\right) f\left(x\right)} dy.$$

- Therefore

$$\begin{aligned} f(x) f\left(x'|x\right) =& f(x) \int f\left(x'|y\right) f\left(y|x\right) dy \\ =& \int \frac{f\left(x', y\right) f\left(x, y\right)}{f\left(y\right)} dy. \end{aligned}$$

- This is obviously exchangable in x and x'. Hence the reversibility condition holds.

Lecture 6, Bayes Estimators

Department of Economics
Stanford University

September, 2008

- Prior $\pi(\theta)$. likelihood $f(\mathbf{x}|\theta)$.
- Posterior density

$$p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\,\pi(\theta)}{\int f(\mathbf{x}|\theta)\,\pi(\theta)\,d\theta}.$$

- In general, computing $p(\theta|\mathbf{x})$ is difficult.
- Exception: conjugate family. Let $\mathcal{F}$ denote the class of likelihoods $f(\mathbf{x}|\theta)$. A class Π of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$.
- The conjugate family for the normal mean when variances are known is normal.

- $\{X_t\}, t = 1, \ldots, n$ i.i.d. $X_t \sim N\left(\mu, \sigma^2\right)$. σ^2 known.
- Prior $\pi\left(\mu\right) \sim N\left(\mu_0, \lambda_0\right)$, μ_0, λ_0 known.
- Posterior distribution

$$p\left(\mu|\mathbf{X}\right) \sim N\left(\frac{\lambda^2\bar{x} + \frac{\sigma^2}{n}\mu_0}{\lambda^2 + \frac{\sigma^2}{n}}, \frac{\frac{\sigma^2}{n}\lambda^2}{\lambda^2 + \frac{\sigma^2}{n}}\right).$$

- Write $t_0 = 1/\lambda^2$, $\bar{t} = n/\sigma^2$: precision parameters.

$$p\left(\mu|\mathbf{X}\right) \sim N\left(\frac{t_0\mu_0 + \bar{t}\bar{x}}{t_0 + \bar{t}}, \frac{1}{t_0 + \bar{t}}\right)$$

- prior mean and sample mean are weighted by their precisions.
- Posterior precision sum of prior and data precisions.

- Bayesian point estimator.
- minimizes posterior expected loss functions:

$$\hat{\theta} = \min_{\theta \in \Theta} \int \rho\left(\theta - \tilde{\theta}\right) p\left(\tilde{\theta}|\mathbf{x}\right) d\tilde{\theta}.$$

- If $\rho(x) = x^2$, square loss:

$$\hat{\theta} = \int \tilde{\theta} p\left(\tilde{\theta}|\mathbf{x}\right) d\tilde{\theta} \quad \text{posterior mean.}$$

- In the normal example:

$$\hat{\mu} = \frac{t_0 \mu_0 + \bar{t}\bar{x}}{t_0 + \bar{t}}.$$

- Other posterior locations, or loss functions, can be used.
- Posteriot interval: region under $p(\theta|\mathbf{x})$ with a given area.

No frequentists shall be denied the pleasure of Bayesian techniques

Department of Economics
Stanford University

November, 2011

A MCMC Approach to Classical Estimation

Victor Chernozhukov Han Hong

August 14, 2002

An Example

Consider MM estimator for Instrumental Median Model:

$$E\big(\tau - 1(Y \leq D'\theta)\big)Z = 0.$$

Maximize criterion

$$L_n(\theta) = -n g_n(\theta) W(\theta) g_n(\theta)$$

with

$$g_n(\theta) = \sum \big(\tau - 1(Y_i \leq D_i'\theta)\big) Z_i$$

and

$$W(\theta) = [\tau(1-\tau)\frac{1}{n}\sum Z_i Z_i']$$

Computing this is a disaster.

Smoothing will not solve the problem either.

Some other examples:

auction models in Donald and Paarsch(1993)

continuous-updated estimator in Hansen and Heaton

possibly nonlinear censored median regression

nonlinear IV wit with many local optima (nonlinear IV)

maximum score

What we do:

1. formally interpret

$$C \cdot \exp(L_n(\theta))$$

as *beliefs* or *q-posterior* about the parameter

this is clearly a non-bayesian way of forming beliefs and learning a parameter

2. easily draw via MCMC a sample

$$S = (\theta^{(1)}, ..., \theta^{(k)})$$

whose marginal distribution is

$$C \cdot \exp(L_n(\theta))$$

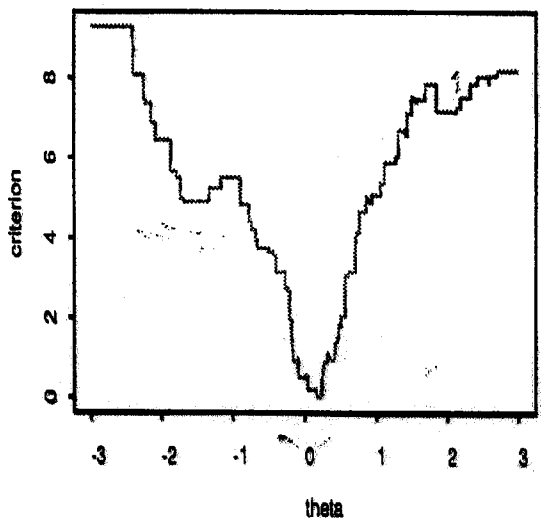
3. compute the q-posterior mean or median of S

take that as an estimator

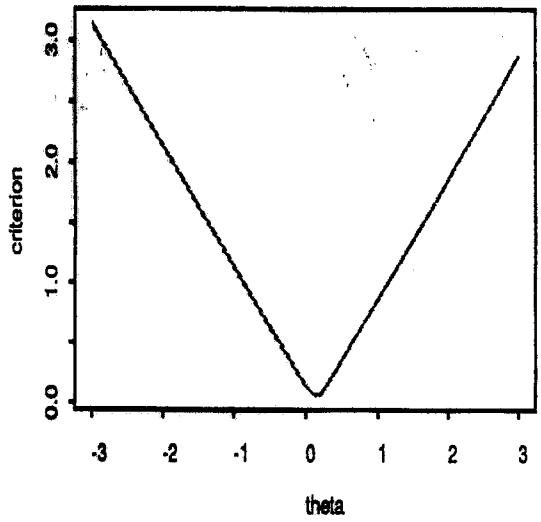
4. compute q-posterior quantiles of S to state confidence intervals

In step 3 and 4 we implicitly solve well-defined convex problems

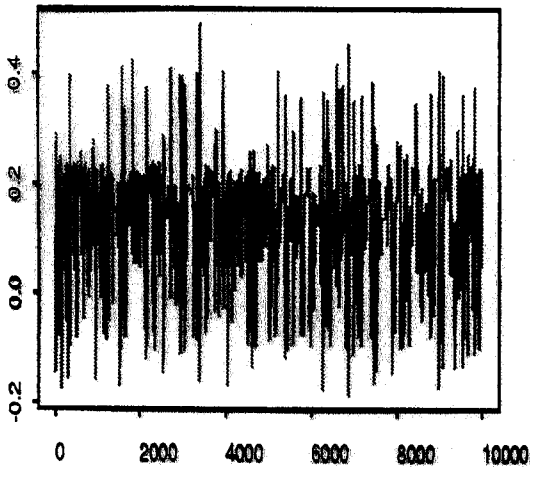
Criterion for IV-QR



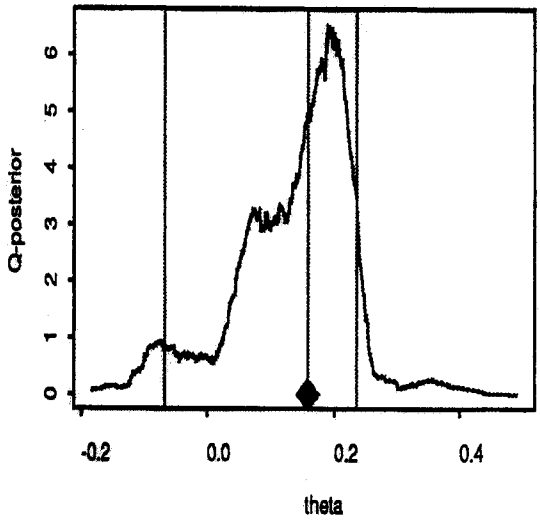
Criterion for QB-Estimation



Markov Chain Sequence



Q-Posterior for Theta



[I. Computation

Definition 1 (*Metropolis with Quasi-Posteriors*)
Given quasi-posterior density $p_n(\theta)$, known up to a constant, and a prespecified conditional density $q(\theta'|\theta)$, generate $(\theta^{(0)}, ..., \theta^{(T)})$ by,

1. *Choose a starting value $\theta^{(0)}$.*
2. *Generate ξ from $q(\xi|\theta^{(j)})$*
3. *Update $\theta^{(j+1)}$ from $\theta^{(j)}$ for $j = 1, 2, ...,$ using*

$$\theta^{(j+1)} = \begin{cases} \xi & \text{with probability} \quad \rho(\theta^{(j)}, \xi) \\ \theta^{(j)} & \text{with probability} \quad 1 - \rho(\theta^{(j)}, \xi) \end{cases},$$

where

$$\rho(x, y) = \min\left(\frac{p_n(y)q(x|y)}{p_n(x)q(y|x)}, 1\right).$$

- Implication of the algorithm:

$$\frac{1}{B}\sum_{t=1}^{B} f(\theta^{(t)}) \xrightarrow{p} \int_{\Theta} f(\theta)p_n(\theta)d\theta.$$

- Application to Q-Bayes Estimation:

Theorem 1 *For any convex and p_n-integrable ρ_n*

$$\arg\min_{\theta\in\Theta}\left[\frac{1}{T}\sum_{j=1}^{T}\rho_n(\theta^{(j)}-\theta)\right] \xrightarrow{p} \hat{\theta} = \arg\min_{\tilde{\theta}\in\Theta}\left[\int_{\Theta}\rho_n(\tilde{\theta}-\theta)p_n(\theta)d\theta,\right]$$

provided that $\hat{\theta}$ is uniquely defined.

- Q-Bayes Estimator and Simulated Annealing:

$$\lim_{\lambda\to\infty}\frac{\int_{\Theta}\theta e^{\lambda L_n(\theta)}\pi(\theta)\,d\theta}{\int_{\Theta}e^{\lambda L_n(\theta)}\pi(\theta)\,d\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}} L_n(\theta)$$

As $\lambda\to\infty$, the sequence of quasi-likelioods converge to a dirac measure.

Questions:

Will this work ?

consistency

asymptotic normality

correct coverage

Conditions? Can we handle

the Median IV example (discontinuous)

GMM with Huber-Pakes-Pollard conditions

regressions with Huber-LeCam conditions

Formal Definitions

finite-sample criterion function

$$L_n(\theta)$$

motivation of extremum estimators: learning by analogy

$n^{-1}L_n \to M$,

so extremum estimator $\to \theta_0$, the extremum of M

$L_n(\theta)$ is not a likelihood function generally

But

$$p_n(\theta) = \frac{e^{L_n(\theta)}\pi(\theta)}{\int_\Theta e^{L_n(\theta)}\pi(\theta)d\theta} \tag{1}$$

is a proper belief or a *quasi-posterior* about θ.

$\pi(\theta)$ is a weight or prior density that is strictly positive and continuous over Θ.

p_n is not a true posterior, since

p_n is generally created through non-Bayesian learning

$\rho_n(u)$ is a *penalty or bernoullian loss function*:

i. $\rho_n(u) = n||u||^p$, a p-th power loss, including the $p = 2$ squared loss,

ii. $\rho_n(u) = \sqrt{n}\sum_{j=1}^{k} |u_k|$, an absolute deviation loss,

iii. $\rho_n(u) = \sqrt{n}\sum_{j=1}^{k}(\tau_j - 1(u_j \leq 0))u_j$, a check (Koenker-Bassett) loss function.

Definition 1 *The class of QBE minimize the expected loss under the belief p_n:*

$$\begin{aligned}\widehat{\theta} &= \arg\min_{z\in\Theta}\left[E_{p_n}(\rho_n(z-\tilde{\theta}))\right] \\ &= \arg\min_{z\in\Theta}\left[\int_{\Theta}\rho_n(z-\theta)\left(\frac{e^{L_n(\theta)}\pi(\theta)}{\int_{\Theta}e^{L_n(\theta)}\pi(\theta)d\theta}\right)d\theta\right].\end{aligned} \tag{2}$$

Conventional convex loss functions lead to:

q-posterior means $\leftarrow \|\cdot\|_2$

q-posterior quantiles $\leftarrow |\cdot|$

q-posterior mode $\leftarrow \|\cdot\|_p$, p =large

Motivation revisited:

1. Generic Computability using Markov Chain Monte-Carlo

- Easy to compute $\hat{\theta}$ by drawing a sample whose marginal distribution is p_n

- MCMC revolutionized Bayesian implementation

- This paper: studies the formal properties of non-bayesian – but similarly defined – estimators that are obtainable by MCMC

- Hope: estimators will have good formal properties and will become a good approach to "hard" non-bayesian problems

2. Econometric preferences satisfy Savage Axioms of choice under uncertainty,

$\hat{\theta}$ maximizes a subjective expected utility, given data

I. "Regular" Cases

Assumption 1 (Parameter) *θ_0 belongs to the interior of a compact convex subset Θ of $\mathbb{R}^d$.*

Assumption 2 (Loss Function) *Loss function has two properties:*

i. *$\rho_n(u)$ is of the form $\rho(a_n u)$, where $\rho(u) > 0$ and $= 0$ iff $u = 0$, and $a_n = \sqrt{n}$*

ii. *ρ is convex and $\rho(h) \leq 1 + \|h\|^p$*

iii. *$\int \rho(u - z)e^{-u'Au}du$ attains unique minimum at finite ξ for any positive definite A.*

Assumption 3 (Identification) *For any $\delta > 0$, there is $\epsilon > 0$:*

$$P\left\{ \sup_{|\theta - \theta_0| \geq \delta} \frac{1}{n}\Big(L_n(\theta) - L_n(\theta_0) \Big) \leq -\epsilon \right\} \longrightarrow 1.$$

Assumption 4 (Local Asymptotic Normality, LAN) *For θ in a ball at θ_0,*

i. $L_n(\theta) - L_n(\theta_0) =$

$$(\theta - \theta_0)'\Delta_n(\theta_0) - \tfrac{1}{2}(\theta - \theta_0)'[nJ(\theta_0)](\theta - \theta_0) + R_n(\theta),$$

ii. $\Delta_n(\theta_0)/\sqrt{n} \xrightarrow{d} N(0, \Omega)$,

iii. *Ω and $J(\theta_0)$ are positive definite constant matrices,*

iv. *for each $\epsilon > 0$ there is sufficiently small $\delta > 0$ and large $M > 0$ such that*

$$(a) \quad \limsup_n P\left\{\sup_{M/\sqrt{n} \le |\theta - \theta_0| \le \delta} \frac{|R_n(\theta)|}{n|\theta - \theta|^2} > \epsilon\right\} < \epsilon,$$

$$(b) \quad \limsup_n P\left\{\sup_{|\theta - \theta_0| \le M/\sqrt{n}} \frac{|R_n(\theta)|}{1} > \epsilon\right\} = 0.$$

Comments:

1. Assumptions are generally patterned but differ from after Bickel and Yahav, LeCam.

2. Differences due to non-likelihood formulation, no quadratic mean differentiability style restrictions (yet).

3. Conditions merge together Huber-style conditions with Bertein-von-Misses style conditions to handle Median IV or Pakes-Pollard GMM criterions for example.

Illustrate the sensibility of Assumption 4.iv, consider a usual Cramer-Amemiya type restriction.

Lemma 1 *Assumption 4.iv holds with*

$$\Delta_n(\theta_0) = \nabla_\theta L_n(\theta_0) \text{ and } J(\theta_0) = \nabla_{\theta\theta'} M(\theta_0),$$

if for $\delta > 0$, $L_n(\theta)$ *is twice differentiable in* θ *when* $|\theta - \theta_0| < \delta$

$$\nabla_\theta L_n(\theta_0)/\sqrt{n} \xrightarrow{d} N(0, \Omega)$$

and for each $\epsilon > 0$,

$$P\Big(\sup_{|\theta-\theta_0|<\delta} \Big| \nabla_{\theta\theta'} L_n(\theta)/n - \nabla_{\theta\theta'} M(\theta) \Big| > \epsilon \Big) \longrightarrow 0$$

where $M(\theta)$ *is twice continuously differentiable at* θ_0.

The Asymptotic Results for "Regular" Cases

Using the obtained earlier beliefs

$$p_n(\theta) = \frac{e^{L_n(\theta)}\pi(\theta)}{\int_\Theta e^{L_n(\theta)}\pi(\theta)d\theta}$$

re-define them to form the quasi-posterior

$$p_n^*(h) = \sqrt{n}p_n(h/\sqrt{n} + J(\theta_0)^{-1}\Delta_n(\theta_0)/\sqrt{n}) -$$

for the rescaled parameter centered at the "score":

$$h = \sqrt{n}(\theta - \theta_0) - J(\theta_0)^{-1}\Delta_n(\theta_0)/\sqrt{n}.$$

Theorem 1 (Beliefs in Large Sample) *Under assumptions 1 - 4 for any* $\alpha \geq 0$,

$$\int_{H_n} \left(1 + |h|^\alpha\right)\left|p_n^*(h) - p_\infty^*(h)\right|dh \xrightarrow{p} 0,$$

and

$$p_\infty^*(h) = \frac{\sqrt{|J(\theta_0)|}}{\sqrt{(2\pi)^d}} \cdot e^{-\frac{1}{2}h'J(\theta_0)h}.$$

For large n, the belief $p_n(\theta)$ is approximately a **random normal density**

$$\text{random mean} = \theta_0 + J(\theta_0)^{-1}\Delta_n(\theta_0)/n$$

and constant *variance* parameter

$$\text{variance} = J(\theta_0)^{-1}/n.$$

Theorem 1 includes the classical likelihood settings and the Berstein-Von Mises Theorems for the posterior likelihood:

$$2\|p_n^* - p_\infty^*\|_{TV} = \int_{H_n} \left| p_n^*(h) - p_\infty^*(h) \right| dh \xrightarrow{p} 0.$$

Convergence in the total variation norm is not sufficient for convergence of such location functionals as the posterior means.

Theorem 1 provides a stronger result: convergence of α-th moments:

$$\left(\int_{H_n} |h|^\alpha p_n^*(h)dh\right) - \left(\int_{H_n} |h|^\alpha p_\infty^*(h)dh\right) \xrightarrow{p} 0.$$

This plus convexity arguments give Theorem 2.

Theorem 2 (QBE in Large Sample) *Under assumption 1-4*

$$\sqrt{n}(\hat{\theta} - \theta_0) - Z_n \xrightarrow{p} 0$$

where

$$Z_n = \arg\min_z \left[\int_{\mathbb{R}^d} \rho(z-u) p_\infty^*(u - J(\theta_0)^{-1}\Delta_n(\theta_0))\ du \right],$$

and

$$Z_n \xrightarrow{d} Z_\infty = \arg\min_z \left[\int_{\mathbb{R}^d} \rho(z-u) p_\infty^*(u - W)\ du \right]$$

with

$$W = N(0, J(\theta_0)^{-1}\Omega(\theta_0)J(\theta_0)^{-1}).$$

Furthermore, if $\rho(h) = \rho(-h)$ *for all* h,

$$Z_n = J(\theta_0)^{-1}\Delta_n(\theta_0) \quad \textit{hence} \quad \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} W.$$

If

$$\int_{\mathbb{R}^d} h p^*_\infty(h)dh = 0 = EW,$$

$$\int_{\mathbb{R}^d} hh' p^*_\infty(h)dh \equiv J(\theta_0)^{-1} = \text{Var}(W) \equiv J(\theta_0)^{-1}\Omega(\theta_0)J(\theta_0)^{-1},$$

or

$$\Omega(\theta_0) = J(\theta_0) \qquad (*)$$

the quasi-posteriors are valid for classical inference

(*): GMM with Hansens' weights, minimum distance, empirical likelihood, and weighted extremum estimators.

Inference about

$g(\theta_0)$,

and g is differentiable at θ_0. Define

$$F_{g,n}(x) = \int_{\theta\in\Theta: g(\theta)\le x} p_n(\theta)d\theta.$$

and

$$Q_{g,n}(\alpha) = \inf\{x : F_{g,n}(x) \ge \alpha\}.$$

Theorem 3 (Validity of Beliefs in Large Sample) *Suppose that the generalized information equality holds:*

$$J(\theta_0) = \Omega(\theta_0).$$

Then

$$P(g(\theta_0) \leq Q_{g,n}(\alpha)) \to \alpha.$$

"Speaking Poinwtise"

$$
\begin{aligned}
& e^{\,L_n(\theta_0+u/\sqrt{n})-L_n(\theta_0)} \\
= \;& e^{u'\Delta_n(\theta_0)/\sqrt{n}-\frac{1}{2}u'J(\theta_0)u} + o_p(1) \\
= \;& e^{-\frac{1}{2}(u-J(\theta_0)^{-1}\Delta_n(\theta_0)/\sqrt{n})'J(\theta_0)(u-J(\theta_0)^{-1}\Delta_n(\theta_0))} \cdot C_n + o_p(1) \\
= \;& e^{-\frac{1}{2}\,h'J(\theta_0)h} \cdot C_n + o_p(1).
\end{aligned}
$$

Proof Outline.

Ignoring random constants:

$$\int (1+|h|^\alpha)|p_n^*(h) - p_\infty^*(h)|dh$$

is bounded over three areas:

(i) $|h| \le M$,

(ii) $M \le |h| \le \delta\sqrt{n}$,

(iii) $|h| \ge \delta\sqrt{n}$.

respectively

1. by

$$
I = \sup_{|h|\le M} |h|^\alpha \Big| \exp(-\frac{1}{2}h'J(\theta_0)h + R_n(\theta)) - \exp(-\frac{1}{2}h'J(\theta_0)h) \Big| \xrightarrow{p} 0 \tag{3}
$$

2. by

$$II = \int_{M \le |h| \le \delta\sqrt{n}} (1 + |h|^\alpha) \exp\left(-\frac{1}{2} h' J(\theta_0) h + \left| R_n(T_n + \frac{h}{\sqrt{n}}) \right| \right)$$

which by setting M large and δ small can be made bounded by (except an ϵ probability event)

$$\int_{M \le |h| \le \infty} (1 + |h|^\alpha) \exp\left(-\frac{1}{2} h' J(\theta_0) h + |h|^2 \lambda \right)$$

where $\lambda < \frac{1}{4}\text{mineig}(J(\theta_0))$,

set M large to make the entire term as small as we like

3. by with probability going to 1:

$$III = \int_{|\theta - \theta_0| \ge \delta/2} (1 + |\theta|^\alpha) \exp(L_n(\theta) - L_n(\theta_0)) d\theta$$
$$\le \int_{|\theta - \theta_0| \ge \delta/2} (1 + |\theta|^\alpha) \exp(-n\epsilon) d\theta \to 0$$

■

Assumptions for "Nonregular" Cases

Assumption 5 (Parameter) *θ_0 belongs to a compact convex subset Θ of Euclidian space $\mathbb{R}^d$.*

Assumption 6 (Loss Function) **i.** *$\rho_n(u)$ is of the form $\rho(a_n u)$, where $\rho(u) > 0$ and $= 0$ iff $u = 0$, for some $a_n \to \infty$*

ii. *ρ is convex and $\rho(h) \leq 1 + \|h\|^p$.*

The focus here is on

$$\ell_n(u) \equiv L_n(\theta_0 + u/a_n) - L_n(\theta_0).$$

Assumption 7 *Suppose as $a_n \to \infty$ there is a random process $\ell_\infty(\cdot)$ such that*

i **(Marginal Convergence)**

$$(\ell_n(u_j), j \leq J) \xrightarrow{d} (\ell_\infty(u_j), j \leq J).$$

ii **(L^1-continuity)** *for any compact set K and bounded Lipschitz f*

$$\int_K f(u) e^{\ell_n(u/a_n)} du \xrightarrow{d} \int_K f(u) e^{\ell_\infty(u)} du.$$

iii **(Small Tails)** *for* $\alpha \geq \alpha^*$ *and each* $\epsilon > 0$, *there exists* $M > 0$ *such that*

$$\limsup_{n\to\infty} P\Big(\int_{|u|>M} |u|^{\alpha} e^{\ell_n(u)} du > \epsilon\Big) < \epsilon.$$

Lemma 2 *Suppose that assumption 7.i hols, and that for any compacts set* K *and for some small* $\delta > 0$ *and some* $M > 0$.

(i) $\sup_{n,z\in K} E e^{\ell_n(z)} < \infty$,

(ii) $\sup_{n,|t-s|<\delta} E|e^{\ell_n(t)/2} - e^{\ell_n(s)/2}|^2 < C|t-s|^{\alpha}$,

(iii) $\sup_n E|z|^p e^{\ell_n(z)} < C(M)/(|z| \ln^2 |z|)$ *for* $|z| > M$,

then assumptions 7.ii - 7.iii hold.

Theorem 4 (General Cases) *Under assumption 5-7,*

$$a_n(\hat{\theta} - \theta_0) \xrightarrow{d} Z_\infty,$$
$$Z_\infty = \arg\min_z \Big[\int_{\mathbb{R}^d} \rho(z-u) e^{L_\infty(u)} du.\Big]$$

provided that Z *is unique a.s.*

Some Non-Standard Examples (briefly):

Andrews, paramater-on-boundary papers: $a_n = n^{-1/3}$

$$\ell_\infty(z) = W'z - z'\frac{1}{2}J(\theta)z \text{ if } z \in V,$$
$$\ell_\infty(z) = -\infty \text{ if } z \notin V,$$
$$V = \{v : \theta_0 + v\delta \in \Theta, \text{ for some scalar } \delta > 0\}.$$

Kim and Pollard, Maximum Score:

$$L_n(\beta) = -n^{-1/3}\sum_{i=1}^{n} |\delta_i - 1(X'\beta > 0)|,$$

Marginal limit:

$$\ell_n(u) = (L_n(\beta_0 + n^{-1/3}u) - L_n(\beta_0))$$

is given by

$$\ell_\infty(u) = -u'Vu + W(u),$$

where $W(u)$ is a mean-zero Gaussian process with a complicated covariance kernel.

Non-Standard Auction and Equilibrium Search Models, cf. Chernozhukov and Hong(2001):

$$\ell_\infty(z) \equiv \exp(z'EX[p(X) - q(X)]) \times \exp[\int_E l_z(j,x)d\mathbf{N}(j,x)], \text{ where}$$

$\mathbf{N}(\cdot) \equiv \sum_{i=1}^\infty 1[(J_i, \mathcal{X}_i) \in \cdot] + \sum_{i=1}^\infty 1[(J_i', \mathcal{X}_i') \in \cdot]$,

$(\mathcal{X}_i), (\mathcal{X}_i')$ are i.i.d. with d.f. F_X, and $(\mathcal{X}_i')$ is independent of $(\mathcal{X}_i)$,

$$\begin{array}{ll} J_i \equiv \Gamma_i/p(\mathcal{X}_i), & \Gamma_i \equiv \quad \mathcal{E}_1 + ... + \mathcal{E}_i, \\ J_i' \equiv \Gamma_i'/q(\mathcal{X}_i'), & \Gamma_i' \equiv -(\mathcal{E}_1' + ... + \mathcal{E}_i'), \end{array} \tag{4}$$

$(\mathcal{E}_i)$ and $(\mathcal{E}_i')$ are two i.i.d., mutually independent sequences of standard exponential random variables that are also independent of $(\mathcal{X}_i)$ and $(\mathcal{X}_i')$.

IV. Some Applications

- Generalized Method of Moments

$$Em_i(\theta_0) = 0.$$

$$L_n(\theta) = -\frac{1}{2} n \left(g_n(\theta)\right)' W_n(\theta) \left(g_n(\theta)\right),$$

$$g_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_i(\theta),$$

$$W_n(\theta) = W(\theta) + o_p(1), \text{ where } W(\theta) = \left[\lim_n \operatorname{Var}[\sqrt{n} g_n(\theta)]\right]^{-1},$$

Proposition 1 (*GMM*) *Suppose that (5.1)-(5.3) hold, and assumption 1-3 hold, and that over an open ball at* θ_0

(i) $m_i(\theta)$ *is stationary and strongly mixing,*

(ii) $G(\theta) = \nabla_\theta E m_i(\theta)$ *is continuous and bounded,*

(iii) $J(\theta) = G(\theta)'W(\theta)G(\theta)$ *is continuous and positive definite,*

(iv) $\Delta_n(\theta_0) = \sqrt{n} g_n(\theta_0)W(\theta_0)G(\theta_0) \xrightarrow{d} N(0, \Omega(\theta_0))$, *and* $\Omega(\theta_0) = G(\theta_0)'W(\theta_0)G(\theta_0)$

(v) for any $\delta_n \to 0$, *the Huber's condition hold*

$$\sup_{|\theta-\theta_0|\leq\delta_n} \left[\frac{\sqrt{n}|\,(g_n(\theta) - g_n(\theta_0)) - (Eg_n(\theta) - Eg_n(\theta_0))\,|}{1 + \sqrt{n}|\theta - \theta_0|} \right] = o_p(1)$$

then assumption 4 holds and therefore all of the conclusions of Theorems 1-3 hold, with Δ_n, $\Omega(\theta_0)$ *and* $J(\theta_0)$ *defined above.*

- Regressions or M-estimators.

$$L_n(\theta) = \sum_{i=1}^{n} m_i(\theta).$$

Proposition 2 (M-Regression) *Suppose assumptions 1-3 hold for the criterion function specified above with the following additional properties: over an open ball at* θ_0

(i) $m_i(\theta)$ *is stationary, strongly mixing sequence,*

(ii) $|m_i(\theta_1) - m_i(\theta_2)| \leq c_i|\theta_1 - \theta_2|$, $Ec_i^2 < \infty$, $J(\tau) = \nabla_{\theta\theta'} E[m_i(\theta)]$ *is continuous and p.d.*

(iii) $\psi_i(\theta) = \nabla_\theta m_i(\theta)$, *exists a.s.*, $n^{-1/2}\sum_{i=1}^{n}\psi_i(\theta) \xrightarrow{d} N(0, \Omega(\theta_0))$

(iv) $E[m_i(\theta) - m_i(\theta_0) - \psi_i(\theta_0)'(\theta - \theta_0)]^2 = o(|\theta - \theta_0|^2)$

The conclusions of Theorems 1 and 2 hold, and if $J(\theta_0)^{-1} = \Omega(\theta_0)$ *the conclusions of Theorem 3 holds.*

Generalized Empirical Likelihood

For a set of moment equations that defines an economic parameter of interest $Em_i(\theta_0) = 0$. Define

$$L_n(\theta, \gamma) \equiv \sum_{i=1}^{n} s[m_i(\theta)'\gamma(\theta)],$$

Then $L_n(\theta) = L_n(\theta, \hat{\gamma}(\theta))$, where $\hat{\gamma}(\theta)$ solves

$$\hat{\gamma}(\theta) \equiv \arg \inf_{\gamma \in \Gamma} L_n(\theta, \gamma). \tag{5}$$

The scalar function $s(\cdot)$ is strictly convex, finite, and four times differentiable function on an open *interval* of $\mathbb{R}$ containing 0, denoted $\mathcal{V}$, and is equal to $+\infty$ outside such an interval. $s(\cdot)$ is normalized so that both $s'(0) = 1$ and $s''(0) = 1$. The choices of the function $s(v) = -\ln(1-v), \exp(v), (1+v)^2/2$ lead to the well-known empirical likelihood, exponential tilting, and continuously updating GMM estimator.

- **Example 1**: Instrumental Median Regression
- The moment condition:

$$m_i(\theta) = (\tau - 1(Y_i \leq q(D_i, X_i, \theta))\, Z_i,$$

- Conditions of proposition 1 are satisfied under mild conditions:
  - (i) (Y_i, D_i, X_i, Z_i) is iid data sequence, and $E[m_i(\theta_0)Z_i] = 0$ and θ_0 is identifiable.
  - (ii) $\{m_i(\theta) = (\tau - 1(Y \leq q(D, X, \theta)))\, Z, \theta \in \Theta\}$ is Donsker, $E \sup_\theta |m_i(\theta)|^2 < \infty$,
  - (iii) $G(\theta) = \nabla_\theta E m_i(\theta) = E f_{Y|D,X,Z}(q(D, X, \theta))Z\nabla_\theta q(D, X, \theta$ is continuous,
  - (iv) $J(\theta) = G(\theta)'[\text{Var } m_i(\theta)]^{-1}\, G(\theta)$ is continuous and p.d. in an open ball at θ_0.
- Weighting matrix

$$\hat{W}(\theta) = \left[\frac{1}{n}\sum_{i=1}^{n} m_i(\theta)\, m_i(\theta)'\right].$$

- **Example 2:** Censored QR and Nonlinear QR

$$L_n(\beta) = -\sum_{i=1}^{n} \omega_i \rho_\tau (Y_i - \max(0, g(X_i;\beta)) \mid$$

QBE is asymptotically equivalent to Powell's estimator. With

$$\omega_i = \frac{1}{\tau(1-\tau)} E1(g(X_i;\beta_0) > 0) f_{Y_i|X_i}(g(X_i;\beta_0))$$

the quasi-posterior quantiles are valid for inference purposes.

- Simulation Examples: censored regression model

$$Y^* = \beta_0 + X'\beta + u$$
$$X \sim N(0, I_3), \quad u = X_2^2 N(0,1),$$
$$Y = \max(0, Y^*)$$

- quasi-Bayes estimator to the Powell CQR objective function

$$L_n(\beta) = -\sum_{i=1}^{n} |Y_i - \max(0, X_i'\beta)|$$

$$\sum_{i=1}^{n} |Y_i^* - X_i'\beta|$$

- **Simulation Example: Instrumental Median Regression**

$$Y = D'\beta + u, \quad u = \sigma(D)\epsilon,$$
$$D = \exp N(0, I_3), \quad \epsilon = N(0,1)$$
$$\sigma(D) = (1 + \textstyle\sum_{i=1}^{3} D_{(i)})/5$$

- **Instrument moment condition**

$$g_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{2} - 1(Y_i \leq \alpha + D'\beta))Z, \text{ where } Z = (1, D, D^2).$$

- **Weight matrix**

$$W = \left[\frac{1}{n}\sum_{i=1}^{n}(\frac{1}{2} - 1(Y_i \leq \alpha + D'\beta))^2 Z_i Z_i'\right]^{-1}.$$

$\tau(1-\tau)\frac{1}{n}\sum^{n} z_i z_i'$

Table 1. Comparison of quasi-bayesian estimators with censored quantile regressi estimates obtained using iterated linear programming (100 simulation runs)

Estimator	rmse	mad	mean bias	med. bias	medad
n=400					
Q-posterior-mean	0.473	0.378	0.138	0.134	0.34
Q-posterior-median	0.465	0.372	0.131	0.137	0.344
Iterated LP(10)	0.518	0.284	0.04	0.016	0.17
	3.798	0.827	-0.568	-0.035	0.24
n=1600					
Q-posterior-mean	0.155	0.121	-0.018	0.0097	0.0897
Q-posterior-median	0.155	0.121	-0.02	0.0023	0.092
Iterated LP(7)	0.134	0.106	0.04	0.067	0.085
	3.547	0.511	0.023	-0.384	0.087

Table 2. Comparison of quasi-bayesian estimators with quantile regression

Estimator	rmse	mad	mean bias	med. bias	medad
n=200					
Q-mean	.0747	.0587	.0174	.0204	.0478
Q-median	.0779	.0608	.0192	.136	.0519
QR	.0787	.0628	.0067	.0092	0.051
n=800					
Q-mean	.0425	.0323	-.0018	-.0003	0.028
Q-median	.0445	.0339	-.0023	.0001	.0295
QR	.0498	.0398	.0007	.0025	.0356

Table 3. Comparison of quasi-bayesian inference with standard inference

Inference	coverage	length
n=200		
Q-equal tail	.943	.377
Q-symmetric(around mean)	.941	.375
QR: Rank-Inversion	.884	.285
QR: HS	.659	.177

Inference	coverage	length
n=800		
Q-equal tail	.92	.159
Q-symmetric(around mean)	.917	.158
QR: Rank-Inversion	.887	.148
QR: HS	.602	.082

Lecture 11: Bootstrap

Instructor: Han Hong

Department of Economics
Stanford University

2011

- Replace the real world by the bootstrap world:
  - Real World: Population(F_0) $\longrightarrow$ Sample(F_1): $X_1, \ldots, X_n$.
  - The bootstrap world: Sample(F_1): $X_1, \ldots, X_n \longrightarrow$ Bootstrap Sample $F_2 = X_1^*, \ldots, X_n^*$.
- We care about functional of $F_0 : \theta(F_0)$, the bootstrap principle says that we estimate $\theta(F_0)$ by $\theta(F_1)$.
- The only problem is how to define $\theta(F_0)$, and the bootstrap resample is only useful for defining this function for $\theta(F_1)$.
- A bootstrap resample is a sample of size n, drawn independently with replacement from the empirical distribution F_1, i.e., $P(X_i^* = X_j | F_1) = n^{-1}$, $1 \leq i, j \leq n$.

- The simplist example: the mean.

$$\theta(F_0) = \mu = \int x dF(x).$$

The bootstrap estimate is

$$\theta(F_1) = \int x dF_1(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = E(X_i^*|F_1)$$

- Similarly, for the variance.

$$\theta(F_0) = \sigma^2 = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$$

$$\theta(F_1) = \hat{\sigma}^2 = \int x^2 d\hat{F}(x) - \left(\int x d\hat{F}(x)\right)^2$$

$$= E(X_i^{*2}|F_1) - (E(X_i^*|F_1))^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X})^2$$

- Both of these drawing X_i^* from F_1 is called nonparametric bootstrap.

- In regression models, $y_i = x_i'\beta + \epsilon_i$, the nonparametric bootstrap (for estimating the distribution of $\hat{\beta}$, say) draws (y_i^*, x_i^*) from the JOINT empirical distribution of (y_i, x_i). It is also possible to draw from $\hat{\epsilon}_i = y_i - x_i'\hat{\beta}$ fixing the x_i's.

- With d dimension data you can find many different ways of resampling, depending on your assumptions about the relation among y_i, x_i, for example.

- You can also modify your bootstrap resample scheme by taking into account a priori information you have about X_i, say if you know X_i is symmetric around 0, then you might want to resample from the $2n$ vector $X_i, -X_i, i = 1, \ldots, n$.

- If you know F_0 is from a parametric family, say $\mathcal{E}\left(\lambda = \mu^{-1}\right)$, then you may want to resample from $F\left(\lambda\right) = \mathcal{E}(\hat{\lambda})$ instead of the empirical distribution F_1.
  - If you choose MLE, then it is $\hat{\lambda} = \frac{1}{\hat{\mu}} = \frac{1}{\bar{X}}$. So you resample from an exponential distribution with mean $\bar{X}$.
  - But we will only discuss nonparametric bootstrap today.
- The bootstrap principle again: The whole business is to find the definition of the functional $\theta\left(F_0\right)$.
  - It is often the solution $t = \theta\left(F_0\right)$ to $E\left[f\left(F_1, F_0; t\right)|F_0\right] = 0$.
  - Since we don't know F_0, the bootstrap version is to estimate t by $\hat{t}$ s.t. $E\left[f\left(F_2, F_1; \hat{t}\right)|F_1\right] = 0$.
- Examples are bias reduction and confidence interval.

- Need $t = E\left(\theta\left(F_1\right) - \theta\left(F_0\right)|F_0\right)$. The bootstrap principle suggests estimating by $\hat{t} = E\left(\theta\left(F_2\right) - \theta\left(F_1\right)|F_1\right)$.

- For example,
  $\theta\left(F_0\right) = \mu^2 = \left(\int x dF_0\left(x\right)\right)^2$, then $\theta\left(F_1\right) = \bar{X}^2 = \left(\int x dF_1\left(x\right)\right)^2$.

$$E\left(\theta\left(F_1\right)|F_0\right) = E_{F_0}\left(\mu + n^{-1}\sum_{i=1}^{n}\left[X_i - \mu\right]\right)^2 = \mu^2 + n^{-1}\sigma^2$$

$$\Longrightarrow t = n^{-1}\sigma^2 = O\left(n^{-1}\right)$$

$$E\left(\theta\left(F_2\right)|F_1\right) = E_{F_1}\left(\bar{X} + n^{-1}\sum_{i=1}^{n}\left[X_i^* - \bar{X}\right]\right)^2 = \bar{X}^2 + n^{-1}\hat{\sigma}^2$$

$$\Longrightarrow \hat{t} = n^{-1}\hat{\sigma}^2 \quad \text{where} \quad \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

- So the bootstrap bias-corrected estimate of μ^2 is:

$$\theta(F_1) - \hat{t} = 2\theta(F_1) - E(\theta(F_2)|F_1) = \bar{X}^2 - n^{-1}\hat{\sigma}^2$$

  Its bias is:

$$E\left[\bar{X}^2 - n^{-1}\hat{\sigma}^2 - \mu^2|F_0\right] = n^{-1}\sigma^2 - n^{-1}\left(1 - n^{-1}\right)\sigma^2 = n^{-2}\sigma^2$$

  So the bias is reduced by an order of $O\left(n^{-1}\right)$, compared to the uncorrected estimate.

- For this problem, the one step bootstrap bias correction does not completely eliminate the bias.(It turns out bootstrap iteration will do)

- But another resample scheme, the jacknife, can eliminate bias completely for this example.

- In general, let $\hat{\theta}$ be an estimator using all data and $\hat{\theta}_{-i}$ be the estimator obtained by omitting observation i.
- The ith jacknife pseudovalue is given as $\theta_i^* = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$.
- The Jacknife estimator is the average of these n of θ_i^*:
$$\hat{\theta}_J \equiv \frac{1}{n}\sum_{i=1}^{n}\theta_i^*.$$
- In this example, $\hat{\theta} = \bar{X}^2$. $\hat{\theta}_{-i} = \left(\frac{1}{n-1}\sum_{j\neq i} X_j\right)^2$. So
$$\hat{\theta}_J = n\bar{X}^2 - (n-1)\left(\frac{1}{n-1}\sum_{j\neq i} X_j\right)^2$$
which is unbiased.

- Look for a one-sided confidence interval of the form $(-\infty, \hat{\theta} + t)$ with coverage probability of α:

$$P\left(\theta(F_0) \leq \hat{\theta} + t\right) = \alpha \Longrightarrow P\left(\theta(F_0) - t \leq \hat{\theta}\right) = \alpha.$$

- The bootstrap version becomes $P\left(\theta(F_1) - \hat{t} \leq \theta(F_2)\right) = \alpha$. So $-\hat{t}$ is $(1-\alpha)$th quantile of $\theta(F_2) - \theta(F_1)$ conditional on $\theta(F_1)$.

- Usually the distribution function of $\theta(F_2) - \theta(F_1)$ conditional on F_1 is difficult to calculate, as difficult as $\theta(F_1) - \theta(F_0)$ conditional on $\theta(F_0)$.

- But as least the former can be simulated (since you know F_1), while the later can't (since you don't know F_0).

- To simulate the distribution of $\theta(F_2) - \theta(F_1)$ conditional on F_1

  (1) Independently draw B (a very big number, say 100,000) bootstrap resamples $X_b^*, b = 1, \ldots, B$ from F_1, where each $X_b^* = (X_{b1}^*, \ldots, X_{bn})^*$, each X_{bi}^* is independent draw from the empirical distribution.

  (2) For each X_b^*, calculate $\theta_b^* = \theta(X_b^*)$. Then simply use the empirical distribution of X_b^*, or any smoothed version of it, to approximate the distribution of $\theta(F_2) - \theta(F_1)$ conditional on F_1.

  This approximation can be arbitrary close as $B \to \infty$.

- Almost just the same as the confidence interval problem.
- Consider a statistics(like OLS coefficient $\hat{\beta}$, t-statistics) $T_n = T_n(X_1, \ldots, X_n)$, want to know its distribution function:

$$P_n(x, F_0) = P(T_n \leq x | X_1, \ldots, X_n \sim iid\ F_0)$$

- But don't know F_0, so use the bootstrap principle,

$$P_n(x, F_1) = P(T_n^* \leq x | X_1^*, \ldots, X_n^* \sim iid\ F_1)$$

- Again when $P_n(x, F_1)$ can't be analytically computed, it can be approximated arbitrary well by

$$P_n(x, F_1) \approx \frac{1}{B} \sum_{b=1}^{B} 1(T_{nb}^* \leq x)$$

for $T_{nb}^* = T_n(X_{b1}^*, \ldots, X_{bn}^*)$.

- Note again the schema in the bootstrap approximation.

$$P_n(x, F_0) \overset{1}{\approx} P_n(x, F_1) \overset{2}{\approx} \frac{1}{B} \sum_{b=1}^{B} 1(T_{nb}^* \leq x)$$

1. The statistical error: introduced by replacing F_0 with F_1, the size of error as $n \to \infty$ can be analyzed through asymptotic theory, e.g. Edgeworth expansion.
2. The numerical error: introduced by approximating F_1 using simulation. Should disappear as $B \to \infty$. It has nothing to do with n-asymptotics and statistical error.

- Similarly, standard error of T_n

$$\sigma^2(T_n) \approx \sigma^2(T_n^*) \approx \frac{1}{B} \sum_{b=1}^{B} \left(T_{nb}^* - \frac{1}{B} \sum_{b=1}^{B} T_{nb}^* \right)^2$$

- Whether the bootstrap works or not (in the consistency sense of whether $P\left(T_n^* \leq x|F_1\right) - P\left(T_n \leq x|F_0\right) \longrightarrow 0$) need to be analyzed case by case.
- $\sqrt{n}$ consistent, asymptotically normal test statistics can be bootstrapped, but it is not known whether other things may work.
- Example of inconsistency, nonparametric bootstrap fails.

  Take $F \sim U\left(0, \theta\right)$, and $X_{(1)}, \ldots, X_{(n)}$ is the order statistics of the sample, so $X_{(n)}$ is the maximum. It is naturally to estimate θ using $X_{(n)}$.

- $\frac{\theta - X_{(n)}}{\theta}$ converges at rate n to $\mathcal{E}(1)$, since for $x > 0$:

$$P\left(n\frac{\theta - X_{(n)}}{\theta} > x\right) = P\left(X_{(n)} < \theta - \frac{\theta x}{n}\right) = P\left(X_i < \theta - \frac{\theta x}{n}\right)^n$$
$$= \left(\frac{1}{\theta}\left(\theta - \frac{\theta x}{n}\right)\right)^n = \left(1 - \frac{x}{n}\right)^n \overset{n\to\infty}{\longrightarrow} e^{-x}$$

  In particular, the limiting distribution is continuous.

- But this is not the case for bootstrapped distribution, $X^*_{(n)}$. The bootstrapped version is naturally $n(X_{(n)} - X^*_{(n)})/X_{(n)}$. But

$$P\left(n\frac{X_{(n)} - X^*_{(n)}}{X_{(n)}} = 0\right) = \left(1 - \left(1 - \frac{1}{n}\right)^n\right) \overset{n\to\infty}{\longrightarrow} \left(1 - e^{-1}\right) \approx 0.63$$

  So there is a big probability mass at 0 in the limiting distribution of the bootstrap sample.

- It turns out that in this example parametric bootstrap would work although nonparametric bootstrap fails. But there are many examples where even parametric bootstrap will fail.
- An alternative to bootstrap, called subsample, proposed by Romano(1998), which include the jacknife as a special case, is almost always consistent, as long as the subsample size m satisfies $m \to \infty$ and $m/n \to 0$. The jacknife case $m = n - 1$ does not satisfy the general consistency condition. Serial correlation in time series also creates problem for naive nonparametric bootstrap. Subsample is one way out.
- The other alternative is to resample blocks instead of individual observations(Fitzenberg(1998)).
- However, both of these will only give consistency but not the 2nd order benefit of edgeworth expansion.

- So if in most cases bootstrap only works when asymptotic theory works, why use bootstrap?
- Some conceivable benefits are:
  - Don't want to waste time deriving asymptotic variance, although $\sqrt{n}$ consistency and asym normality is known. Let the computer do the job.
  - Avoid bandwidth selection in estimating var-cov of quantile regression type estimators. Bandwidth is needed for either kernel estimate of the conditional density $f(0|x_t)$ or for numerical derivatives.
  - For asymptotic pivotal statistics, bootstrapping is equivalent to automatically doing edgeworth expansion.

- An exact (or asymptotic) pivotal statistics T_n is one whose (or asymptotic) distribution does not depend on unknown parameters $\forall n$.
- Denote pivotal statistics by T_n and nonpivotal ones by S_n.
- If know that $F \sim N\left(\mu, \sigma^2\right)$, then
  - $S_n = \sqrt{n}\left(\bar{X} - \mu\right) \sim N\left(0, \sigma^2\right)$ is nonpivotal since unknown σ^2. The bootstrap estimate is $N\left(0, \hat{\sigma}^2\right)$, so there is error in approximating the distribution of S_n.
  - $T_n = \sqrt{n-1}\frac{\left(\bar{X}-\mu\right)}{\hat{\sigma}^2} \sim t_{n-1}$ for $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$. The bootstrap estimate is also t_{n-1}. No error here.
- If T_n is exact pivotal, need not bootstrap at all. Either look up a table or simulate. But most statistics are asymptotic pivotal.

- No matter what F is, for t-statistics the CLT says $P\left(T_n \leq x\right) \overset{n\to\infty}{\longrightarrow} \Phi\left(x\right)$, so it is asymptotically pivotal.
- But the CLT doesn't say how fast $P\left(T_n \leq x\right)$ tends to $\Phi\left(x\right)$.
- The Edgeworth expansion describes it:

$$P_n\left(x, F_0\right) \equiv P\left(T_n \leq x | F_0\right) = \Phi\left(x\right) + G\left(x, F_0\right)\frac{1}{\sqrt{n}} + O\left(n^{-1}\right)$$

The bootstrap version is:

$$P_n\left(x, F_1\right) \equiv P\left(T_n^* \leq x | F_1\right) = \Phi\left(x\right) + G\left(x, F_1\right)\frac{1}{\sqrt{n}} + O_p\left(n^{-1}\right)$$

- The Edgeworth expansion can be carried out up to many terms in power of $n^{-1/2}$. Expansion up to the 2nd term:

$$P_n\left(x, F_0\right) \equiv P\left(T_n \leq x | F_0\right) = \Phi\left(x\right) + G\left(x, F_0\right)\frac{1}{\sqrt{n}} + H\left(x, F_0\right)\frac{1}{n} + O\left(n^{-\frac{3}{2}}\right)$$

Consider error in approximating $P_n(x, F_0)$:

- Error of CLT:

$$P_n(x, F_0) - \Phi(x) = G(x, F_0)\frac{1}{\sqrt{n}} + O\left(n^{-1}\right) = O\left(\frac{1}{\sqrt{n}}\right)$$

- Error of Bootstrap:

$$P_n(x, F_0) - P_n(x, F_1) = G(x, F_0)\frac{1}{\sqrt{n}} - G(x, F_1)\frac{1}{\sqrt{n}} + O_p\left(n^{-1}\right) = \\ \left(G(x, F_0) - G(x, F_1)\right)\frac{1}{\sqrt{n}} + O_p\left(n^{-1}\right) = O_p\left(n^{-1}\right)$$

since $\sqrt{n}(F_1 - F_0) = O_p(1)$, and assuming $G(x, F)$ is smooth and differentiable in the 2nd argument:

$$G(x, F_1) - G(x, F_0) = O_p(F_1 - F_0) = O_p\left(\frac{1}{\sqrt{n}}\right).$$

- So if your sample size is 100, By CLT you commit an error of (roughly) 0.1, but by bootstrap 0.01, big improvement??

- However, this improvement doesn't work for nonpitoval statistics, say S_n: by CLT

$$P\left(S_n \leq x\right) \overset{n\to\infty}{\longrightarrow} \Phi\left(\frac{x}{\sigma}\right).$$

- The corresponding Edgeworth expansion is:

$$P_n\left(x, F_0\right) \equiv P\left(S_n \leq x | F_0\right) = \Phi\left(\frac{x}{\sigma}\right) + G\left(x/\sigma, F_0\right)\frac{1}{\sqrt{n}} + O\left(n^{-1}\right)$$

The bootstrap version is:

$$P_n\left(x, F_1\right) \equiv P\left(S_n^* \leq x | F_1\right) = \Phi\left(\frac{x}{\hat{\sigma}}\right) + G\left(x/\hat{\sigma}, F_1\right)\frac{1}{\sqrt{n}} + O\left(n^{-1}\right)$$

Consider error in approximating $P_n(x, F_0)$:

- Error of CLT: need to replace σ by $\hat{\sigma}$.

  $P_n(x, F_0) - \Phi(x/\hat{\sigma}) =$
  $\Phi(x/\sigma) - \Phi(x/\hat{\sigma}) + G(x/\sigma, F_0)\frac{1}{\sqrt{n}} + O(n^{-1}) = O\left(\frac{1}{\sqrt{n}}\right)$

- Error of Bootstrap:

  $P_n(x, F_0) - P_n(x, F_1) = \Phi(x/\sigma) - \Phi(x/\hat{\sigma}) + G(x/\sigma, F_0)\frac{1}{\sqrt{n}} -$
  $G(x/\hat{\sigma}, F_1)\frac{1}{\sqrt{n}} + O_p(n^{-1}) = O_p(n^{-1/2})$

  This is because both $F_1 - F_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\hat{\sigma} - \sigma = O_p\left(\frac{1}{\sqrt{n}}\right)$.

- No improvement compared to CLT. This is because now the 1st term $\Phi(x/\sigma)$ does not cancelled with $\Phi(x/\hat{\sigma})$.

- The implication of this is that bootstrapping provides better approximation to two sided symmetric test(or symmetric confidence interval) compared to one sided test(or confidence interval).
- Assume $G(x, F_0)$ is an even function in x.
- One-sided test: reject if $T_n \leq x$ (or $T_n > x$), the approximaton error being:

$$\begin{aligned} P_n(x, F_0) - P_n(x, F_1) &= G(x, F_0) \frac{1}{\sqrt{n}} - G(x, F_1) \frac{1}{\sqrt{n}} + O_p\left(n^{-1}\right) \\ &= O_p\left(n^{-1}\right) \end{aligned}$$

- Two sided test: reject if $|T_n| \geq x \Leftrightarrow (T_n > x \cup T_n < -x)$, then

$$\begin{aligned}
&P(|T_n| > x) = P(T_n > x) + P(T_n < -x) \\
=&\left[1 - \Phi(x) - G(x, F_0)\frac{1}{\sqrt{n}} - H(x, F_0)\frac{1}{n} - O\left(n^{-3/2}\right)\right] \\
+&\left[\Phi(-x) + G(-x, F_0)\frac{1}{\sqrt{n}} + H(-x, F_0)\frac{1}{n} + O\left(n^{-3/2}\right)\right] \\
=&2\Phi(-x) - 2H(x, F_0)\frac{1}{n} + O\left(n^{-3/2}\right)
\end{aligned}$$

- So the approximation error is:

$$\begin{aligned}
P(|T_n^*| > x|F_1) - P(|T_n| > x) &= 2[H(x, F_0) - H(x, F_1)]\frac{1}{n} + O\left(n^{-3/2}\right) \\
&= O_p\left(n^{-3/2}\right)
\end{aligned}$$

Smaller by an order of $O_p\left(n^{-1/2}\right)$.

- Only look at $G(x, F_0)$ but not higher order terms like $H(x, F_0)$
- Simply take $X_1, \ldots, X_n$ iid $EX_i = 0, Var(X_i) = 1$. So $T_n = \sqrt{n}\bar{X}$
- Recall the characteristic function for T_n: by X_i iid assumption

$$\phi_{T_n}(t) = Ee^{itT_n} = Ee^{it\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i} = \left(Ee^{i\frac{t}{\sqrt{n}}X_i}\right)^n$$
$$= \left[\phi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n = e^{n\log\phi_X\left(\frac{t}{\sqrt{n}}\right)}$$

- Taylor expand this around $\frac{t}{\sqrt{n}} = 0$:

$$n \log \phi_X \left(\frac{t}{\sqrt{n}} \right)$$

$$= n \log \phi_X (0) + n \frac{\phi'_X (0)}{\phi_X (0)} \frac{t}{\sqrt{n}} + n \frac{1}{2} \left[\frac{\phi''_X (0)}{\phi_X (0)} - \frac{(\phi'_X (0))^2}{\phi_X (0)^2} \right] \left(\frac{t}{\sqrt{n}} \right)^2$$

$$+ n \frac{1}{3!} \left[\frac{\phi'''_X (0)}{\phi_X (0)} - 3 \frac{\phi'_X (0) \phi''_X (0)}{\phi_X (0)^2} + 2 \frac{\phi'_X (0)^3}{\phi_X (0)^3} \right] \left(\frac{t}{\sqrt{n}} \right)^3 + O \left(\frac{t}{\sqrt{n}} \right)^4$$

- Recall that $\phi_X (0) = 1$, $\phi'_X (0) = iEX = 0$, $\phi''_X (0) = i^2 EX^2 = -1$, $\phi'''_X (X) = i^3 EX^3 \equiv -i\mu^3$:

$$n \log \phi_X \left(\frac{t}{\sqrt{n}} \right) = -\frac{1}{2} t^2 - \frac{i}{6} \mu^3 \frac{t^3}{\sqrt{n}} + O \left(\frac{t^4}{n} \right)$$

$$\Phi_{T_n} (t) = e^{n \log \phi_X \left(\frac{t}{\sqrt{n}} \right)} = e^{-t^2/2} \exp \left(-\frac{i}{6} \mu^3 \frac{t^3}{\sqrt{n}} + O \left(\frac{t^4}{n} \right) \right)$$

$$= e^{-t^2/2} \left[1 - \frac{i}{6} \mu^3 \frac{t^3}{\sqrt{n}} + O \left(n^{-1} \right) \right]$$

- Use the Inversion Formula: for $\phi_X(t) = Ee^{itX} = \int e^{itx} f(x)\,dx$, there is $f(x) = \frac{1}{2\pi}\int e^{-ixt}\phi_X(t)\,dt$
- For example, the characteristic function of $N(0,1)$ is $e^{-t^2/2}$, so $e^{-t^2/2} = \int e^{itx}\phi(x)\,dx$, so $\phi(x) = \frac{1}{2\pi}\int e^{-ixt}e^{-t^2/2}dt$.
- Now applying this to $X = T_n$:

$$
\begin{aligned}
f_{T_n}(x) &= \frac{1}{2\pi}\int e^{-ixt}\phi_{T_n}(t)\,dt = \frac{1}{2\pi}\int e^{-ixt}e^{n\log\phi_X\left(\frac{t}{\sqrt{n}}\right)}dt \\
&= \frac{1}{2\pi}\int e^{-ixt}e^{-\frac{t^2}{2}}\left[1 - \frac{i}{6}\mu^3\frac{t^3}{\sqrt{n}} + O\left(n^{-1}\right)\right]dt \\
&= \frac{1}{2\pi}\int e^{-ixt}e^{-\frac{t^2}{2}}dt - \frac{i}{6}\frac{\mu^3}{\sqrt{n}}\left(\frac{1}{2\pi}\int e^{-ixt}e^{-\frac{t^2}{2}}t^3dt\right) \\
&= \frac{1}{2\pi}\int e^{-ixt}e^{-\frac{t^2}{2}}dt - \frac{i}{6}\frac{1}{(-i)^3}\frac{\mu^3}{\sqrt{n}}\left[\frac{d}{dx^3}\left(\frac{1}{2\pi}\int e^{-ixt}e^{-\frac{t^2}{2}}t^3dt\right)\right] \\
&+ O\left(n^{-1}\right) = \phi(x) - \frac{1}{6}\frac{\mu^3}{\sqrt{n}}\phi'''(x) + O\left(n^{-1}\right)
\end{aligned}
$$

- So

$$P\left(T_n \leq x\right) = \int^{x} f_{T_n}\left(u\right) du = \Phi\left(x\right) - \frac{1}{6}\frac{\mu^3}{\sqrt{n}}\phi''\left(x\right) + O\left(n^{-1}\right)$$

.

- So

$$G\left(x, F_0\right) = -\frac{\mu^3}{6}\phi''\left(x\right) = \frac{\mu^3}{6}\left(1 - x^2\right)\phi\left(x\right),$$

by noting that $\phi'\left(x\right) = -x\phi\left(x\right)$, and $\phi''\left(x\right) = -\phi\left(x\right) + x^2\phi\left(x\right)$. Note that $G\left(x, F_0\right)$ is an even function.

Lecture 13: Subsampling vs Bootstrap

Dimitris N. Politis, Joseph P. Romano, Michael Wolf

2011

- $R_n\left(x_n, \theta\left(P\right)\right) = \tau_n\left(\hat{\theta}_n - \theta\left(P\right)\right)$
- Example:
  - $\hat{\theta}_n = \bar{X}_n, \tau_n = \sqrt{n}, \theta = EX = \mu\left(P\right)$
  - $\hat{\theta} = \min X_n, \tau_n = n, \theta\left(P\right) = sup\{x : F\left(x\right) \leq 0\}$
- Define: $J_n\left(P\right)$, the distribution of $\tau_n\left(\hat{\theta}_n - \theta\left(P\right)\right)$ under P. For real $\hat{\theta}_n$,

$$J_n\left(x, P\right) \equiv Prob_P\left(\tau_n\left(\hat{\theta}_n - \theta\left(P\right)\right) \leq x\right)$$

- Since P is unknown, $\theta\left(P\right)$ is unknown, and $J_n\left(x, P\right)$ is also unknown.

- The bootstrap estimate $J_n(x, P)$ by $J_n(x, \hat{P}_n)$, where $\hat{P}_n$ is a consistent estimate of P in some sense.
  - For example, take $\hat{P}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$

$$sup_x \left| \hat{P}_n(x) - P(x) \right| \xrightarrow{a.s.} 0$$

- Similarly estimate $(1-\alpha)$th quantile of $J_n(x, P)$ by $J_n(x, \hat{P}_n)$: i.e. Estimate $J_n^{-1}(x, P)$ by $J_n^{-1}(x, \hat{P}_n)$.
- Usually $J_n(x, \hat{P}_n)$ can't be explicitly calculated, use MC:

$$J_n\left(x, \hat{P}_n\right) \approx \frac{1}{B}\sum_{i=1}^{B} 1\left(\tau_n\left(\hat{\theta}_{n,i} - \hat{\theta}_n\right) \leq x\right)$$

for $\hat{\theta}_{n,i} = \hat{\theta}\left(X_{1,i}^*, \ldots, X_{n,i}^*\right)$.

- When bootstrap works, for each x,

$$J_n(x, \hat{P}_n) - J_n(x, P) \xrightarrow{p} 0 \Longrightarrow J_n^{-1}(1-\alpha, \hat{P}_n) - J_n^{-1}(1-\alpha, P) \xrightarrow{p} 0$$

- When should Bootstrap "work"? Need local uniformity in weak convergence:
  - Usually $J_n(x, P) \longrightarrow J(x, P)$.
  - Usually $\hat{P}_n \xrightarrow{a.s.} P$ in some sense, say $\sup_x |\hat{P}_n(x) - P(x)| \xrightarrow{a.s.} 0$
  - Suppose for each sequence P_n s.t. $P_n \to P$, say $\sup_x |P_n - P| \to 0$, it is also true that $J_n(x, P_n) \longrightarrow J(x, P)$, then it must be true that a.s. $J_n(x, \hat{P}_n) \longrightarrow J(x, P)$
  - So it ends up having to show for $P_n \to P$, $J_n(x, P_n) \to J(x, P)$, use triangular array formulation.

- Sample mean with finite variance.
  - $\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0$.
  - $\theta(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{a.s.} \theta(F) = E(X)$.
  - $\sigma^2(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \xrightarrow{a.s.} \sigma^2(F) = Var(X)$.
  - Use Linderberg-Feller for the triangular array, applied to the deterministic sequence of P_n such that:

    1) $\sup_x |P_n(x) - P_n(x)| \to 0$; 2) $\theta(P_n) \to \theta(P)$;

    3) $\sigma^2(P_n) \to \sigma^2(P)$,

    it can be shown that $\sqrt{n}(\bar{X}_n - \theta(P_n)) \xrightarrow{d} N(0, \sigma^2)$ under P_n.
  - Since $\hat{P}_n$ satisfies 1,2,3 a.s., therefore $J_n(x, \hat{P}_n) \xrightarrow{a.s.} J(x, P)$
- So "local uniformity" of weak convergence is satisfied here.

- Order Statistics:

  $F \sim U(0, \theta)$, and $X_{(1)}, \ldots, X_{(n)}$ is the order statistics of the sample, so $X_{(n)}$ is the maximum:

$$P\left(n\frac{\theta - X_{(n)}}{\theta} > x\right) = P\left(X_{(n)} < \theta - \frac{\theta x}{n}\right) = P\left(X_i < \theta - \frac{\theta x}{n}\right)^n$$
$$= \left(\frac{1}{\theta}\left(\theta - \frac{\theta x}{n}\right)\right)^n = \left(1 - \frac{x}{n}\right)^n \overset{n\to\infty}{\longrightarrow} e^{-x}$$

- The bootstrap version:

$$P\left(n\frac{X_{(n)} - X^*_{(n)}}{X_{(n)}} = 0\right) = \left(1 - \left(1 - \frac{1}{n}\right)^n\right)$$
$$\overset{n\to\infty}{\longrightarrow} \left(1 - e^{-1}\right) \approx 0.63$$

- Degenerate U-statistics:

  Take $w(x,y) = xy$, $\theta(F) = \int\int w(x,y)\,dF(x)\,dF(y) = \mu(F)^2$.

$$\hat{\theta}_n = \theta\left(\hat{F}_n\right) = \frac{1}{n(n-1)}\sum\sum_{i\neq j} X_i X_j$$

$$S(x) = \int xy dF(y) = x\mu(F)$$

- If $\mu(F) \neq 0$ it is known that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, 4Var\left(S\left(X\right)\right)\right) = N\left(0, 4\left(\mu^2 EX^2 - \mu^4\right)\right)$$

  The bootstrap works.

- But if $\mu(F) = 0 \Longrightarrow \theta(F) = 0$:

$$\theta(\hat{F}_n) = \frac{1}{n(n-1)} \sum\sum_{i \neq j} X_i X_j = \bar{X}_n^2 - \frac{1}{n} \frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2 = \bar{X}_n^2 - \frac{S_n^2}{n}$$

$$n\left(\theta\left(\hat{F}_n\right) - \theta(F)\right) = n\bar{X}_n^2 - S_n^2 \xrightarrow{d} N(0, \sigma^2) - \sigma^2$$

- However the bootstrap version of $n\left[\theta\left(\hat{F}_n^*\right) - \theta\left(\hat{F}_n\right)\right]$:

$$n\left(\left[\bar{X}_n^{*2} - \frac{1}{n} S_n^{*2}\right] - \left[\bar{X}_n^2 - \frac{1}{n} S_n^2\right]\right) = n\bar{X}_n^{*2} - S_n^{*2} - n\bar{X}_n^2 + S_n^2$$
$$\approx n\left(\bar{X}_n^{*2} - \bar{X}_n^2\right)$$
$$= \left[\sqrt{n}\left(\bar{X}_n^* - \bar{X}_n\right)\right]^2 + 2\sqrt{n}\left(\bar{X}_n^* - \bar{X}_n\right)\sqrt{n}\bar{X}_n$$
$$\xrightarrow{d} N(0, \sigma^2)^2 + 2N(0, \sigma^2)\sqrt{n}\bar{X}_n$$

- iid case: Y_i block of size b from $(X_1, \ldots, X_n)$, $i = 1, \ldots, q$, for $q = \binom{n}{b}$. Let $\hat{\theta}_{n,b,i} = \hat{\theta}(Y_i)$ calculated with ith block of data.

- Use the empirical distribution of $\tau_b(\hat{\theta}_{n,b,i} - \hat{\theta})$ over the q pseudo-estimates to approximate the distribution of $\tau_n(\hat{\theta} - \theta)$:

$$\text{Approximate} \qquad J_n(x, P) = P\left(\tau_n\left(\hat{\theta}_n - \theta\right) \leq x\right)$$

$$\text{by} \qquad L_{n,b}(x) = q^{-1} \sum_{i=1}^{q} 1\left(\tau_b\left(\hat{\theta}_{n,b,i} - \hat{\theta}_n\right) \leq x\right)$$

- Claim: If $b \to \infty$, $b/n \to 0$, $\tau_b/\tau_n \to 0$, as long as $\tau_n(\hat{\theta} - \theta) \xrightarrow{d}$ something,

$$J_n(x, P) - L_{n,b}(x) \xrightarrow{p} 0$$

Subsampling:

- Each subset of size b comes from the TRUE model. Since $\tau_n(\hat{\theta}_n - \theta) \xrightarrow{d} J(x, P)$, so as long as $b \to \infty$:

$$\tau_b(\hat{\theta}_b - \theta) \xrightarrow{d} J(x, P)$$

  The distributions of $\tau_n(\hat{\theta}_n - \theta)$ and $\tau_b(\hat{\theta}_b - \theta)$ should be close.

- But $\tau_b(\hat{\theta}_b - \theta) = \tau_b(\hat{\theta}_b - \hat{\theta}_n) + \tau_b(\hat{\theta}_n - \theta)$. Since

$$\tau_b\left(\hat{\theta}_n - \theta\right) = O_p\left(\frac{\tau_b}{\tau_n}\right) = o_p(1)$$

  The distributions of $\tau_b(\hat{\theta}_b - \theta)$ and $\tau_b(\hat{\theta}_b - \hat{\theta}_n)$ should be close.

- The distribution of $\tau_b(\hat{\theta}_b - \hat{\theta}_n)$ is estimated by the empirical distribution over $q = \binom{n}{b}$ pseudo-estimates.

Bootstrap:

- Recalculate the statistics from the ESTIMATED model $\hat{P}_n$.
- Given that $\hat{P}_n$ is close to P, hopefully $J_n(x, \hat{P}_n)$ is close to $J_n(x, P)$ (Or to $J(x, P)$, the limit distribution).
- But when bootstrap fails

$$\hat{P}_n \longrightarrow P \nRightarrow J_n\left(x, \hat{P}_n\right) \longrightarrow J(x, P)$$

- Assumptions: $\tau_n(\hat{\theta}_n - \theta) \xrightarrow{d} J(x, P)$, $b \to \infty$, $\frac{b}{n} \to 0$, $\frac{\tau_b}{\tau_n} \to 0$.
  Need to show: $L_{n,b}(x) - J(x, P) \xrightarrow{p} 0$.

- Since $\tau(\theta_n - \theta) \xrightarrow{p} 0$, it is enough to show

$$U_{n,b}(x) = q^{-1} \sum_{i=1}^{q} 1\left(\tau_b\left(\hat{\theta}_{n,b,i} - \theta\right) \leq x\right) \xrightarrow{p} J(x, P).$$

  $U_{n,b}(x)$ is a bth order U-statistics with kernel function bounded by $(-1, 1)$.

- $U_{n,b}(x) - J(x, P) = U_{n,b}(x) - EU_{n,b}(x) + EU_{n,b}(x) - J(x, P)$, it is enough to show

$$U_{n,b}(x) - EU_{n,b}(x) \xrightarrow{p} 0 \quad \text{and} \quad EU_{n,b}(x) - J(x, P) \to 0$$

- But

$$EU_{n,b}(x) - J(x,P) = J_b(x,P) \to 0$$

- Use Hoeffding exponential-type inequality (Serfling(1980), Thm A. p201):

$$\begin{aligned} P\left(U_{n,b}(x) - J_b(x,P) \geq \epsilon\right) &\leq \exp\left(-2\frac{n}{b}\epsilon^2 / \left[1-(-1)\right]\right) \\ &= \exp\left(-\frac{n}{b}t^2\right) \longrightarrow 0 \qquad \text{as } \frac{n}{b} \longrightarrow \infty. \end{aligned}$$

- So

$$\begin{aligned} L_{n,b}(x) - J(x,P) = {} & L_{n,b}(x) - U_{n,b}(x) + U_{n,b}(x) \\ & - J_b(x,P) + J_b(x,P) - J(x,P) \xrightarrow{p} 0. \end{aligned}$$

Q.E.D.

- Respect the ordering of the data to preserve correlation.

$$\hat{\theta}_{n,b,t} = \hat{\theta}_b\left(X_t, \ldots, X_{t+b-1}\right), \quad q = T - b + 1.$$

$$L_{n,b}(x) = \frac{1}{q}\sum_{i=1}^{q} 1\left(\tau_b\left(\hat{\theta}_{n,b,t} - \hat{\theta}_n\right) \leq x\right)$$

- Assumption: $\tau_n(\hat{\theta}_n - \theta) \xrightarrow{d} J(x, P)$, $b \to \infty$, $\frac{b}{n} \to 0$, $\frac{\tau_b}{\tau_n} \to 0$, $\alpha(m) \to 0$.
- Result: $L_{n,b}(x) - J(x, P) \xrightarrow{p} 0$.
- Most difficult part: To show $\tau_n(\hat{\theta}_n - \theta) \xrightarrow{d} J(x, P)$.

- Can treat iid data as time series, or even using non-overlapping blocks $k = \left[\frac{n}{b}\right]$, but using $\binom{n}{b}$ more efficient.
- For example, if

$$\bar{U}_n(x) = k^{-1} \sum_{j=1}^{k} 1\left(\tau_b\left[R_{n,b,j} - \theta(P)\right] \leq x\right)$$

then

$$U_{n,b}(x) = E\left[\bar{U}_n(x) \,|\mathcal{X}_n\right] = E\left[1\left(\tau_b\left[R_{n,b,j} - \theta(P)\right] \leq x\right) |\mathcal{X}_n\right]$$

for $\mathcal{X}_n = \left(X_{(1)}, \ldots, X_{(n)}\right)$.

- $U_{n,b}(x)$ is better than $\bar{U}_n(x)$ since $\mathcal{X}_n$ is sufficient statistics for iid data.

- Hypothesis Testing: $T_n = \tau_n t_n (X_1, \ldots, X_n)$,

$$G_n (x, P) = Prob_p (\tau_n \leq x) \overset{P \in P_0}{\longrightarrow} J (x, P)$$

$$\hat{G}_{n,b} (x) = q^{-1} \sum_{i=1}^{q} 1 (T_{n,b,i} \leq x) = q^{-1} \sum_{i=1}^{q} 1 (\tau_b t_{n,b,i} \leq x)$$

As long as $b \to \infty$, $\frac{b}{n} \to 0$, then under $P \in P_0$:

$$\hat{G}_{n,b} (x) \longrightarrow G (x, P)$$

If under $P \in P_1$, $T_n \to \infty$, then $\forall x$, $\hat{G}_{n,b} (x) \to 0$.

- Key difference with confidence interval: don't need $\frac{\tau_b}{\tau_n} \to 0$, because don't need to estimate θ_0 but assumed known under the null hypothesis.

- Assume that $\tau_n = n^\beta$, for some unknown $\beta > 0$. Estimate β using different size of subsampling distribution.
- Key idea: Compare the shape of the empirical distributions of $\hat{\theta}_b - \hat{\theta}_n$ for different values of b to infer the value of β.
- Let $q = \binom{n}{b}$ for iid data, or $q = T - b + 1$ for time series data:

$$L_{n,b}(x|\tau_b) \equiv q^{-1} \sum_{a=1}^{q} 1\left(\tau_b\left(\hat{\theta}_{n,b,a} - \hat{\theta}_n\right) \leq x\right)$$

$$L_{n,b}(x|1) \equiv q^{-1} \sum_{a=1}^{q} 1\left(\hat{\theta}_{n,b,a} - \hat{\theta}_n \leq x\right)$$

- This implies

$$L_{n,b}(x|\tau_b) = L_{n,b}\left(\tau_b^{-1}x|1\right) \equiv t$$

- $x = L_{n,b}^{-1}(t|\tau_b) = \tau_b\left(\tau_b^{-1}x\right) = \tau_b L_{n,b}^{-1}(t|1)$
- Since $L_{n,b}(x|\tau_b) \xrightarrow{p} J(x,P)$, if $J(x,P)$ is continuous and increasing, it can be infered that

$$L_{n,b}^{-1}(t|\tau_b) = J^{-1}(t,P) + o_p(1)$$

- Same as

$$\tau_b L_{n,b}^{-1}(t|1) = J^{-1}(t,P) + o_p(1)$$

- So

$$b^{\beta} L_{n,b}^{-1}(t|1) = J^{-1}(t,P) + o_p(1)$$

- Assuming $J^{-1}(t,P) > 0$, or $t > J(0,P)$, take log.

- For different b_1 and b_2, then this becomes

$$\beta \log b_1 + \log\left(L_{n,b_1}^{-1}(t|1)\right) = \log J^{-1}(t,P) + o_p(1)$$
$$\beta \log b_2 + \log\left(L_{n,b_2}^{-1}(t|1)\right) = \log J^{-1}(t,P) + o_p(1)$$

- Different out the "fixed effect"

$$\beta(\log b_1 - \log b_2) = \log\left(L_{n,b_2}^{-1}(t|1)\right) - \log\left(L_{n,b_1}^{-1}(t|1)\right) + o_p(1)$$

- So estimate β by

$$\begin{aligned}\hat{\beta} &= (\log b_1 - \log b_2)^{-1}\left(\log\left(L_{n,b_2}^{-1}(t|1)\right) - \log\left(L_{n,b_1}^{-1}(t|1)\right)\right)\\ &= \beta + (\log b_1 - \log b_2)^{-1} \times o_p(1)\end{aligned}$$

- Take $b_1 = n^{\gamma_1}$, $b_2 = n^{\gamma_2}$, $(1 \geq \gamma_1 > \gamma_2 > 0)$

$$\hat{\beta} - \beta = ((\gamma_1 - \gamma_2)\log n)^{-1} o_p(1) = o_p\left((\log n)^{-1}\right)$$

- How to know $t > J(0, P)$,

$$L_{n,b}(0|\tau_b) = L_{n,b}(0|1) = J(0, P) + o_p(1)$$

So estimating $J(0, P)$ not a problem.

- Alternatively, take $t_2 \in (0.5, 1)$, take $t_1 \in (0, 0.5)$

$$b^{\beta}\left(L_{n,b}^{-1}(t_2|1) - L_{n,b}^{-1}(t_1|1)\right) = J^{-1}(t_2|P) - J^{-1}(t_1|P) + o_p(1)$$

$$\beta \log b + \log\left(L_{n,b}^{-1}(t_2|1) - L_{n,b}^{-1}(t_1|1)\right) = \log\left(J^{-1}(t_2|P) - J^{-1}(t_1|P)\right) + o_p(1)$$

- $\hat{\beta} = (\log b_1 - \log b_2)^{-1}$
$\left[\log\left(L_{n,b_2}^{-1}(t_2|1) - L_{n,b_2}^{-1}(t_1|1)\right) - \log\left(L_{n,b_1}^{-1}(t_2|1) - L_{n,b_1}^{-1}(t_1|1)\right)\right]$

- Take $b_1 = n^{\gamma_1}$, $b_2 = n^{\gamma_2}$ $(1 > \gamma_1 > \gamma_2 > 0)$, $\hat{\beta} - \beta = o_p((\log n)^{-1})$.

- $\hat{\tau}_n = n^{\hat{\beta}}$

$$L_{n,b}\left(x|\hat{\tau}_b\right) = q^{-1}\sum_{a=1}^{q} 1\left(\hat{\tau}_b\left(\hat{\theta}_{n,b,a} - \hat{\theta}_n\right) \leq x\right)$$

Can show that

$$\sup_x \left| L_{n,b}\left(x|\hat{\tau}_b\right) - J\left(x, P\right)\right| \xrightarrow{p} 0.$$

- Problem: imprecise in small samples.
  - In variation estimation, best choice of b gives $O(n^{-1/3})$ error rate.
  - Parameter estimates, if model is true, gives $O(n^{-1/2})$ error rate.
  - Bootstrap pivotal statistics, when applicable, gives even better than $O(n^{-1/2})$ error rate.