Sample Selection Models with Monotone Control Functions

Ruixuan Liu [1] Zhengfei Yu [2]

[1]Emory University

[2]University of Tsukuba

Indiana University Econometrics Seminar
April 2, 2019

The sample selection model

The sample selection problem arises when observations are not taken from a random sample of the population.

- Wage equations [Roy(1950); Heckman and Honore(1990); Schafgans (1998, 2000)]
- Female labor supply [Heckman(1974), Gronau(1974), Arellano and Bonhomme (2017)]
- Schooling choice [Cameron and Heckman(1998)]
- Unionism status [Lee(1978), Lemieux(1998)]
- Immigration [Borjas(1987); Chiquiar and Hanson(2005)]

The sample selection model

- A prototypical sample selection model consists of the following outcome and selection equations:

$$
\begin{aligned}
Y_i^* &= X_i'\beta_0 + \varepsilon_i, \quad \text{(Outcome)} \\
D_i &= \mathbb{I}\{W_i'\gamma_0 + \nu_i > 0\}, \quad \text{(Selection)} \\
Y_i &= Y_i^* D_i, \quad \text{for } i = 1, \cdots, n,
\end{aligned}
$$

  where (Y_i, D_i, X_i', W_i') are observed variables and (ε_i, ν_i) are latent error terms.
- The unknown parameters: regression coefficients β_0, γ_0, and the joint distribution $F_{\varepsilon\nu}$ of the latent errors.
- The conditional mean of $Y_i$is

$$\mathbb{E}[Y_i|X_i, W_i, D_i = 1] = X_i'\beta_0 + \lambda_0(W_i'\gamma_0),$$

  $\lambda_0(W_i'\gamma_0) = \mathbb{E}[\varepsilon_i|\nu_i > -W_i'\gamma_0]$ corrects for the sample selection bias; known as the control function.

Parametric sample selection models

- Heckman's selection model:
  Bivariate normal (ε_i, ν_i) with correlation ρ
  $\Longrightarrow \lambda_0(W_i'\gamma_0) = \rho\sigma_\varepsilon \phi(W_i'\gamma_0)/\Phi(W_i'\gamma_0)$
  The control function λ_0 is parametric and monotone.

- Lee's generalized selection model:
  Maintain the Gaussian copula for (ε_i, ν_i), yet allow for arbitrary (but known) marginal distributions:

$$\lambda_0(t) = \rho\sigma_\varepsilon \left\{ \frac{\phi(\Phi^{-1} \circ F_\nu(t))}{F_\nu(t)} \right\}.$$

  A selection with t-marginal distributions [Marchenko and Genton (2012, JASA)]:
  λ_0 is parametric and monotone.

Non/semi-parametric selection models

- The conditional mean specification

$$\mathbb{E}[Y_i|X_i, W_i, D_i = 1] = X_i'\beta_0 + \lambda_0(W_i'\gamma_0)$$

  leads to a partial linear and single index model (Li and Racine, 2007).
- Without any distributional assumption on (ε_i, ν_i)
  $\Longrightarrow$ λ_0 is nonparametric and subject to some smoothness assumptions.
  $\Longrightarrow$ kernel or sieve type estimators are used to estimate the nonparametric components.
- Powell (1987), Gallant and Nychka (1987), Robinson(1988), Newey (2009), Ahn and Powell (1993), Andrews and Schafgans (1998), Li and Wooldridge (2002), Das et al. (2003)...

"While there have been substantial theoretical advances in weakening the parametric structure used to secure identification of the models used in the early work, progress in implementing these procedures in practical empirical problems has been slow and empirical applications of semi-parametric methods have been ***plagued by issues of sensitivity of estimates to choices of smoothing parameters, trimming parameters, and the like.****"*

Heckman and Vytlacil (2007, Handbook 6B)

Elie: *So, in the 1980s and 1990s, we see ingenious papers that provided insightful results on approaches to combining stochastic restrictions, support conditions and some functional form assumptions to get point identification in a wide variety (of mostly nonlinear) models. Also, the estimation approaches were non-trivial. But, on the other hand, it is also safe to say that* ***this literature has not had as much impact on empirical work directly. Do you think it is a problem?***
Chuck: I very much view it as a problem.

THE ET INTERVIEW: Charles Manski, Interviewed by Elie Tamer)

What we do...

- A semiparametric sample selection model:
  - No parametric distributional assumption on error terms (ε_i, ν_i)
  - Maintains the monotonicity of the control function $\lambda_0(\cdot)$.
- A simple sufficient condition for a monotone λ_0: ε_i and ν_i satisfy certain dependence restriction; i.e., right tail increasing (decreasing) for positively (negatively) dependent pairs.
- A new semiparametric estimation method:
  - Isotonic estimates of nonparametric components;
  - no tuning parameter to be chosen by users (bandwidth, number of basis functions, trimming sequences...)
- A new selectivity test:
  - Test whether $\lambda_0(\cdot)$ is a constant vs. decreasing/increasing function.

Main Contributions

- We state a simple sufficient condition that leads to the monotonicity of the control function.
  - The condition only depends on the copula function.
  - The monotone shape is shared by a much larger family beyond the original Heckman model.
- We develop new semiparametric estimation and testing procedures that do not require the selection of any tuning parameter.
  - Easy to implement with existing R-packages.
- Technical contributions to semiparametric two-step estimation involving shape-restricted components.
  - Results in Chen et al. (2003), Lee and Ichimura (2010), or Chen et al. (2014) are not directly applicable.
  - Cosslett (1991) has proposed a tuning-parameter-free estimation procedure, but only established consistency using a sample-splitting trick.

Related literature

Non/semi-parametric selection models:

- Semiparametric estimation: Powell (1987), Robinson(1988), Newey (2009), Ahn and Powell (1993), Li and Wooldridge (2002), Das et al. (2003)...
- Copula models: Abbring and Heckman (2007), Fan and Wu (2010), Fan et al. (2017), Arellano and Bohnomme (2017), Maasoumi and Wang (2018+).
- Generalized Roy model and Program Evaluation: Heckman et al. (2003), Heckman and Vytlacil (2005), Brinch et al. (2017), Mogstad et al. (2018), Kline and Walters (2019).

Shape restricted estimation applied to single index models:

- Cosslett (1983, 1991): consistency result only.
- Groeneboom and Hendrickx (2018).

Shape restriction in econometric models:

- Matzkin (1991, 1993), Horowitz and Lee (2017), Chetverikov et al. (2018).

Monotone control function

A sufficient condition on (ε, ν) for the monotonicity of the control function λ_0: right tail increasing ($RTI(\varepsilon|\nu)$)

Def. [Esary and Proschan, 1972, Annals] A random variable ε is said to be right tail increasing in ν, if $P\{\varepsilon > s|\nu > t\}$ is an increasing function of t for all s.

- A (positive) dependence condition that is stronger than positive quadrant dependence (PQD), yet weaker than stochastic increasing (SI).
- It relies on the copula function, regardless of the marginal distribution.
- Right tail decreasing ($RTD(\varepsilon|\nu)$) also yields a monotone λ_0.

A Simple Roy Model

- Y_1 and Y_0 are wages attached to different sectors (or different education levels):

$$Y_1 = X'\beta_1 + u_1, \tag{1}$$
$$Y_0 = X'\beta_0 + u_0,$$

  with an observable switching cost (or price) $C = \tilde{W}'\beta_C$.
- The individual self selects into the sector with a higher wage modulo the switching cost:

$$D = \mathbb{I}\{X'(\beta_1 - \beta_0) - \tilde{W}'\beta_C + (u_1 - u_0) > 0\}. \tag{2}$$

  One only observes the wage corresponding to sector 1; i.e., $Y = D \times Y_1$.
- $W = (X', \tilde{W}')'$, $\gamma_0 = ((\beta_1 - \beta_0)', \beta_C')'$, and $\nu = u_1 - u_0$ in the selection equation.
- $RTI(\varepsilon|\nu)$ means that when $\nu = u_1 - u_0$ is larger, it is more likely that u_1 is large as well.

Monotone control function

$RTI(\varepsilon|\nu)$ is a dependence property that only depends on the copula function.

- A copula is a bivariate distribution function C from $[0,1]^2$ to $[0,1]$ with uniform margins.
- A copula links marginal distributions to the joint. $F_{\varepsilon,\nu}(s,t) = C(F_\varepsilon(s), F_\nu(t))$ for continuous r.v. ε and ν. (Sklar's Theorem)
- Let C be the copula of the joint dist (ε, ν), then for all $u \in [0,1]$,
  1. $RTI(\varepsilon|\nu) \iff (1-u-v+C(u,v))/(1-v)$ is increasing in v,
  2. $RTI(\varepsilon|\nu) \iff (u-C(u,v))/(1-v)$ is decreasing in v.

Monotone control function

Theorem (monotone CF) $RTI(\varepsilon|\nu) \Longrightarrow$ control function $\lambda_0(t) = \mathbb{E}(\varepsilon|\nu > -t)$ is deceasing.

$$
\begin{aligned}
\mathbb{E}[\varepsilon|\nu > t] &= \int_0^{+\infty} s dF_{\varepsilon|\nu>t}(s) + \int_{-\infty}^{+0} s dF_{\varepsilon|\nu>t}(s) \\
&= \int_0^{+\infty} \left(1 - F_{\varepsilon|\nu>t}(s)\right) ds - \int_{-\infty}^{0} F_{\varepsilon|\nu>t}(s) ds \\
&= \int_0^{+\infty} \frac{1 - F_\varepsilon(s) - F_\nu(t) + C(F_\varepsilon(s), F_\nu(t))}{1 - F_\nu(t)} ds \\
&\quad - \int_{-\infty}^{0} \frac{F_\varepsilon(s) - C(F_\varepsilon(s), F_\nu(t))}{1 - F_\nu(t)} ds.
\end{aligned}
$$

The first integrand is increasing in $F_\nu(t)$ and the second integrand is decreasing in $F_\nu(t)$;
$\Longrightarrow \mathbb{E}(\varepsilon|\nu > t)$ increasing in t.

Gaussian copula models

- The control function in Heckman (1979) has the well-known form depending on the inverse Mill's ratio:

$$\lambda_0(t) = \rho\sigma_\varepsilon \left\{ \frac{\phi(t)}{\Phi(t)} \right\}. \tag{3}$$

- $\lambda_0(t)$ is monotone, whereas the direction depends on whether $\rho > 0$ or < 0.
- In fact, the monotonicity property here only depends on the Gaussian copula $C(u, v; \rho) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$ and the sign of its correlation coefficient denoted by ρ .
- A straightforward calculation shows that the partial derivative of any Gaussian copula is

$$\frac{\partial}{\partial v} C(u, v; \rho) = \Phi \left(\frac{\Phi^{-1}(u) - \rho\Phi^{-1}(v)}{\sqrt{1 - \rho^2}} \right). \tag{4}$$

Archimedean copula models

- When the copula function is Archimedean, i.e., $C(u,v) = \psi^{[-1]}(\psi(u) + \psi(v))$ with ψ as the generator function.
- $RTI(\varepsilon|\nu)$ is equivalent to Oakes' (1989) cross-ratio function being greater or equal to 1; i.e., $CR(u) \geq 1$ for any u, where

$$CR(u) = -u\frac{\psi^{(2)}(u)}{\psi^{(1)}(u)}. \tag{5}$$

- For Clayton copula

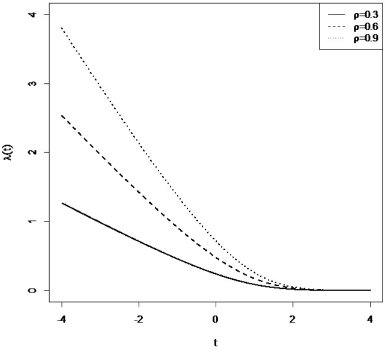
$$C(u,v;\alpha) = (u^{-\alpha} + v^{-\alpha} - 1)^{-1/\alpha}, \quad 0 \leq u,v \leq 1, \tag{6}$$

where the parameter $\alpha \geq 0$. The cross-ratio function $CR(u) = \alpha + 1$ for any $u \in [0,1]$, which is always greater or equal to 1.

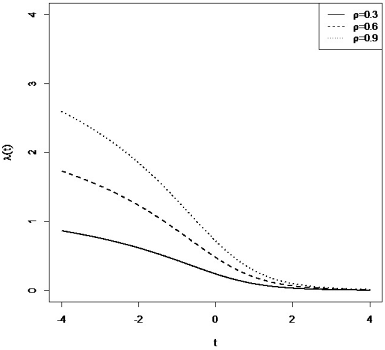
Control Functions for Different Margins

Figure: Plots of $\lambda(t) = \mathbb{E}[\varepsilon|\nu \geq -t]$; A Gaussian copula with correlation ρ.

(a)Marginal: $N(0,1)$.



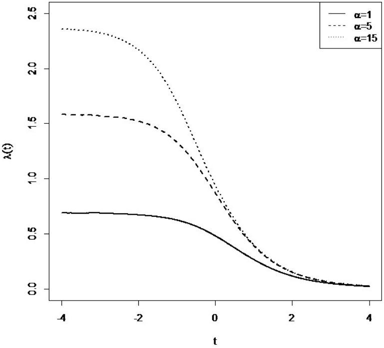
(b) Marginal : $t(5)$.



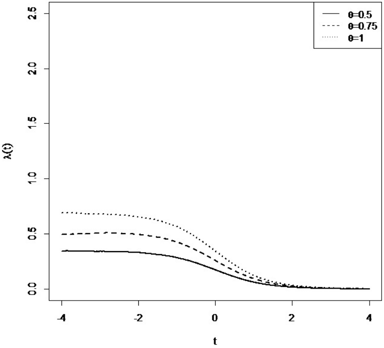
Control Functions for Different Copulas

Figure: Plots of $\lambda(t) = \mathbb{E}[\varepsilon|\nu \geq -t]$; A Marginal $t(5)$.

(a)Clayton Copula.



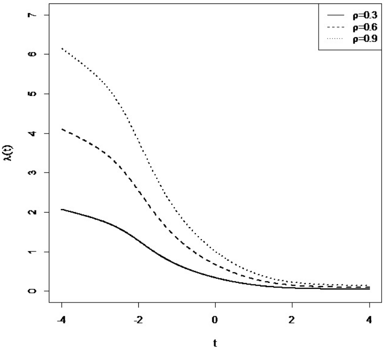
(b) FGM Copula.



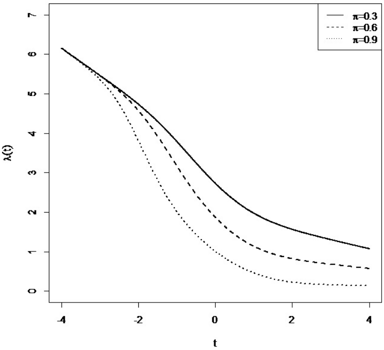
Control Functions for Normal Mixtures

Figure: Plots of $\lambda(t) = \mathbb{E}[\varepsilon|\nu \geq -t]$; A Normal Mixture.

(e) Different correlation ρ.



(f) Different mixing coefficient π.



Monotone Control Functions with Different Shapes

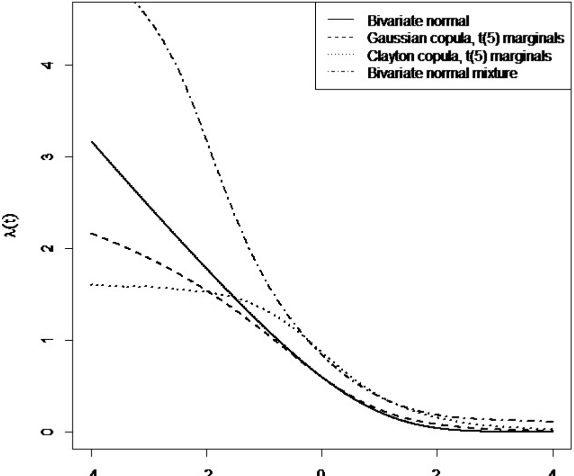


Figure: Control function λ_0 for different joint distribution of error terms

Estimation

A two-stage semiparametric estimation approach:

- two nonparametric components: F_ν and λ;
- both are monotone $\Longrightarrow$ nonparametric MLE (NPMLE)/isotonic for F_ν and λ;
- no tuning parameter determined by users;
- resulting estimators F_ν and λ are piecewise constant functions.

Estimation: Stage 1

Estimate coefficients γ_0 in the selection equation in the presence of an unknown distribution F_ν

- We adopt Groeneboom and Hendrikx (2018 Annals of Statistics).
- for any fixed γ, compute the NPMLE of $\widehat{F}_{nv}(\cdot;\gamma)$ by

$$\max_F \sum_{i=1}^{n} \left[\bar{D}_i \log F(-W_i'\gamma) + (1-\bar{D}_i)\log(1-F(-W_i'\gamma))\right],$$

  where $\bar{D}_i \equiv 1 - D_i$ (Cosslett, 1983, EMCA).
- Estimate γ_0 by solving estimating equations (moment conditions):

$$\frac{1}{n}\sum_{i=1}^{n} W_i \left[\bar{D}_i - \hat{F}_{n\nu}(-W_i'\hat{\gamma}_n;\hat{\gamma}_n)\right] = 0.$$

Stage 1: NPMLE of $F_{n\nu}(\cdot;\gamma)$

- For a fixed γ, let $V_i^{(\gamma)} = -W_i'\gamma$
- Sort $V_{(1)}^{(\gamma)} \leq \cdots \leq V_{(n)}^{(\gamma)}$, and let $\bar{D}_{(i)}^{(\gamma)}$ be the indicator associated with the i-th order statistic $V_{(i)}^{(\gamma)}$.
- The NPMLE $\hat{F}_{n\nu}(\cdot;\gamma)$ is the left derivative of the (greatest) convex minorant of the cumulative sum diagram:

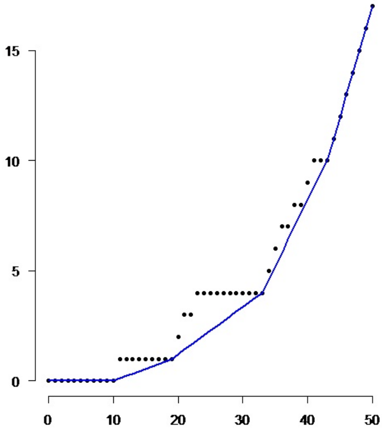
$$P_0 = (0,0), \quad P_i = \left(i, \sum_{j=1}^{i} \bar{D}_{(j)}^{(\gamma)}\right) \quad \text{for } i = 1, \cdots, n.$$

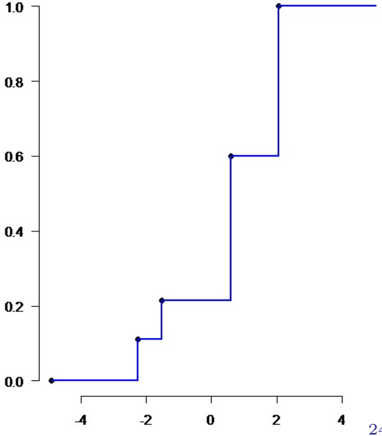
- (greatest) convex minorant = the maximal convex function lying entirely below the diagram of points
- The left derivative at P_i determines the value of $F_{n\nu}(\cdot;\gamma)$ at $V_{(i)}^{(\gamma)}$ and hence on $[V_{(i)}^{(\gamma)}, V_{(i+1)}^{(\gamma)})$.
- NPMLE was first proposed by Ayer et al. (1955).

Greatest Convex Minorant and Its Derivative

The left panel: cumulative sum diagram (black dots) and the (greatest) convex minorant (blue line).

The right panel: left derivative of the (greatest) convex minorant (=NPMLE).





Pool Adjacent Violators Algorithm (PAVA)

[Brunk (1955), Barlow et al. (1972), Robertson et al. (1988)]
A simple example: Suppose that $\bar{D}_i$ sorted according to $V_{(i)}$ is $\{0, 1, 0, 0, 1, 0, 1, 1\}$. Then PAVA proceeds as

$$0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1$$

$$0 \quad \underbrace{1 \quad 0} \quad 0 \quad 1 \quad 0 \quad 1 \quad 1$$

$$0 \quad \underbrace{\frac{1}{2}(2) \qquad 0} \quad 1 \quad 0 \quad 1 \quad 1$$

$$0 \quad \tfrac{1}{3}(3) \quad \underbrace{1 \quad 0} \quad 1 \quad 1$$

$$0 \quad \tfrac{1}{3}(3) \quad \tfrac{1}{2}(2) \quad 1 \quad 1$$

The NPMLE is a step function with jumps: $0 \to \frac{1}{3}$ at $V_{(2)}$, $\frac{1}{3} \to \frac{1}{2}$ at $V_{(5)}$, and $\frac{1}{2} \to 1$ at $V_{(7)}$.

Estimation: Stage 2

- Given $\hat{\gamma}_n$, estimate the slope coefficient β_0 in the outcome equation:

$$\mathbb{E}[Y_i | X_i, W_i, D_i = 1] = X_i' \beta_0 + \lambda(W_i' \gamma_0),$$

- a partial linear model with a monotone nonparametric component λ; see Huang (2002).
- let $i = 1, ..., n_1$ denotes the subsample with $D_i = 1$.
- estimate β_0 and λ by the least squares estimator under the monotonicity restriction of λ:

$$(\hat{\beta}_n, \hat{\lambda}_n) = \arg \min_{\beta \in \mathbf{B}, \boldsymbol{\lambda}_n \in \mathcal{D}} \sum_{i=1}^{n_1} \left[Y_i - X_i' \beta - \lambda(W_i' \hat{\gamma}_n) \right]^2 .$$

Estimation: Stage 2

Compute $(\hat{\beta}_n, \hat{\lambda}_n)$:

1. For any value β, use NPMLE (isotonic regression) to compute $\hat{\lambda}_n(\cdot;\beta)$;
   similar to Stage 1: $\hat{\lambda}_n(\cdot;\beta)$ is the left derivative of the (greatest) convex minorant of the cumulative sum diagram:
   $$P_0 = (0,0), \quad P_i = \left(i, \sum_{j=1}^{i} (Y_{(j)} - X'_{(j)}\beta) \right) \quad \text{for } i = 1, ..., n_1,$$
   where $(Y_{(j)} - X'_{(j)}\beta)$ corresponds to the j-th order statistic $V_i^{(\hat{\gamma}_n)} = -W_i'\hat{\gamma}_n, i = 1, ..., n_1$.
2. Minimize the objective function w.r.t β.

R package "isotone" for isotonic regression.

Estimation: Stage 2

An alternative algorithm to compute $(\hat{\beta}_n, \hat{\lambda}_n)$:

- Estimate β and λ simultaneously by a cone projection.
- $\mathcal{D} = \{t \in \mathbb{R}^{n_1} : At \leq 0\}$ where A is an $(n-1) \times n$ matrix with $A_{i,i} = 1, A_{i,i+1} = -1$ and others 0.
- The projection of an n_1-vector onto $\mathcal{D}$ has the form: $b_0 \mathbf{1}_{n_1} + \sum_{j=1}^{M_n} b_j \mathbf{e}_j, b_j > 0$ for $j = 1, ..., M_n$, where $\mathbf{e}_j$'s are some columns of the matrix $A'(AA')^{-1}$[Meyer (2013)].
- R package "coneproj" for cone projection.

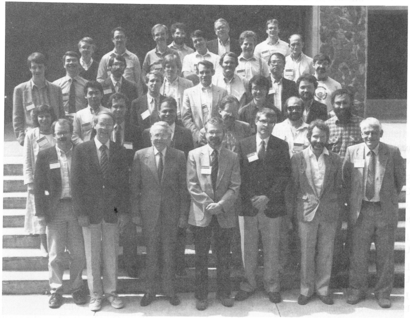
Cosslett's two-stage approach

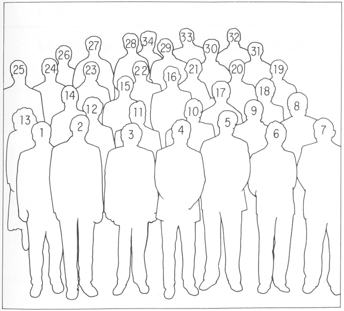
- **Stage 1**: Profile maximum likelihood estimation (Cosslett, 1983) for the selection equation:

$$\max_{\gamma,F} \sum_{i=1}^{n} \left[\bar{D}_i \log F(-W_i'\gamma) + (1-\bar{D}_i) \log(1 - F(-W_i'\gamma)) \right],$$

- The estimated marginal d.f. $\tilde{F}_{n\nu}(\cdot)$ is a step-wise function that is constant on a finite number K_n of intervals $I_j = [c_{i-1}, c_j)$, for $j = 1, ..., K_n$ and $c_0 = -\infty, c_{K_n} = +\infty$.
- **Stage 2**: Cosslett (1991) estimates the outcome equation while approximating the control function $\lambda(\cdot)$ by K_n indicator variables $\{\mathbb{I}(W'\tilde{\gamma}_n \in I_j)\}_{j=1}^{K_n}$.
- Only consistency results are available for both β and γ.

Figure: A conference 30 years ago...





Participants of the Fifth International Symposium in Economic Theory and Econometrics, Duke University, May 1988.

1 T. S. Thompson	2 K. Land	3 G. Kozmetsky	4 W. A. Barnett
5 G. E. Tauchen	6 P. C. B. Phillips	7 P. Huber	8 P. M. Robinson
9 H. White	10 H. Bierens	11 H. Vinod	12 K. J. Singleton
13 R. L. Matzkin	14 A. Han	15 J. H. Stock	16 C. Sims
17 H. Ichimura	18 J. L. Powell	19 D. Pollard	20 L. F. Lee
21 A. R. Pagan	22 L. P. Hansen	23 A. Lewbell	24 D. A. Hsieh
25 W. K. Newey	26 S. R. Cosslett	27 C. F. Manski	28 J. Geweke
29 A. R. Gallant	30 T. M. Stoker	31 G. Chamberlain	32 D. Andrews
33 J. J. Heckman	34 M. Lavine		

Comparisons with existing methods

Compare with the sieve-type estimator[Newey(2009)].

- Sieve: $\lambda_n(\cdot) = \sum_{j=1}^{K_n} b_j P_j(\cdot)$, where $P_1(\cdot), \cdots, P_{K_n}(\cdot)$ are basis functions in the sieve space.
  K_n is specified by users; $P_j(\cdot)$ is a smooth function.
- Our estimator: $\boldsymbol{\lambda}_n = b_0 \mathbf{1}_{n_1} + \sum_{j=1}^{M_n} b_j \mathbf{e}_j$, $b_j > 0$ for $j = 1, ..., M_n$.
  M_n is determined by data; each $\mathbf{e}_j$ is an n-vector.

Compare with Cosslett (1991).

- Stage 2 approximates the control function λ_0 by indicators $\mathbb{I}(W'\tilde{\gamma}_n \in I_j), j = 1, ..., K_n$, where the end points of intervals $I_j, j = 1, ...K_n$ are jumps of NPMLE of F_ν in Stage 1.
- The estimated control function is not necessarily monotone and its jump locations are determined by first stage estimates.
- only consistency results are established.

Regularity conditions

Condition

There exists a local neighborhood $\mathcal{N}_0$ around γ_0 such that for any $\gamma \in \mathcal{N}_0$, $W'\gamma$ is a non-degenerate random variable conditional on X.

- We strengthen the identification condition (A-2) in Heckman and Vytlacil (2007b).

Condition

The true monotone control function λ_0 is continuously differentiable with its derivative denoted by $\dot{\lambda}(\cdot)$. Moreover, its inverse denoted by $\lambda_0^{-1}(\cdot)$ is globally Lipschitz continuous.

- The true monotone control function and its inverse function are also smooth.

Asymptotic Properties

- The consistency and asymptotic distribution of Stage 1 coefficient $\hat{\gamma}_n$ follows from Groeneboom and Hendrikx (2018 Annals).
- Main task: determine the asymptotic contribution of the estimated $\hat{\gamma}_n$ to $\hat{\beta}_n$.
- Nonstandard problem: the estimated control function is piecewise-constant $\Longrightarrow$ cannot be differentiated to determine the asymptotic contribution of $\hat{\gamma}_n$.
- Combine the characterization of isotonic regression [Huang (2002), Mammen and Yu (2007), Cheng (2009)] and empirical process theory [Groeneboom and Hendrikx (2018), Balabdaoui et.al (2016, 2017)]

Asymptotic Properties

Theorem (Consistency and Rate of Convergence)

$$|\hat{\beta}_n - \beta_0| + \| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \| = O_p(n^{-1/3}\log n).$$

For the nonparametric component, we use the following L_2 norm to metrize its convergence:

$$\| \hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0) \|^2 \equiv \int \left(\hat{\lambda}_n(w'\hat{\gamma}_n) - \lambda_0(w'\gamma_0)\right)^2 f_{W|D=1}(w)dw,$$

where $f_{W|D=1}(\cdot)$ is the conditional density of W given $D = 1$.

More notations

We adopt the following notations from Newey (2009).

- Let $U_i \equiv D_i(X_i - \mathbb{E}[X_i|D_i = 1, W_i'\gamma_0])$;
  $\epsilon_i \equiv D_i(Y_i - X_i'\beta_0 - \lambda_0(V_i))$
  $H_\beta \equiv \mathbb{E}[U_i U_i']$; $\Sigma \equiv \mathbb{E}[\epsilon_i^2 U_i U_i']$; $H_\gamma \equiv \mathbb{E}[U_i \lambda_0'(V_i) W_i]$.
- V_γ is AsyVar of $\hat{\gamma}_n$ in Stage 1.
- $V_\gamma = A^{-1} B A^{-1}$, where
  $A \equiv \mathbb{E}\left[f_{\nu 0}(-W'\gamma_0)\left\{W - E[W|W'\gamma_0]\right\}^{\otimes 2}\right]$,
  $B \equiv \mathbb{E}\left[\left\{(F_{\nu 0}(-W'\gamma_0) - \bar{D})(W - E[W|W'\gamma_0])\right\}^{\otimes 2}\right]$.

Asymptotic properties

Theorem (Asymptotic Normality)

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \Rightarrow \mathbb{N}(0, V_\beta),$$

where

$$V_\beta \equiv H_\beta^{-1}\left(\Sigma + H_\gamma V_\gamma H_\gamma'\right) H_\beta^{-1}.$$

$H_\beta^{-1} \Sigma H_\beta^{-1}$: AsyVar of $\hat{\beta}_n$ when γ_0 is known.
$H_\beta^{-1}\left(H_\gamma V_\gamma H_\gamma'\right) H_\beta^{-1}$: the effect of estimating γ_0 in Stage 1.

Testing for selectivity

- H_0 : no sample selection bias.
- Heckman model: H_0 : the coefficient on the inverse Mills equals to zero.
- Non/semi-parametric model: $H_0 : \lambda$ is constant vs. $H_0 : \lambda$ is a non-constant smooth function [Christofides et al. 2003].
- In our setup with the monotone control function (assume decreasing):
  $H_0 : \lambda$ is constant; $H_1 : \lambda$ is a decreasing function
- A likelihood ratio type test:

$$T_n = \frac{\sum_i (R^2_{ols,i} - R^2_{de,i})}{\sum_i R^2_{ols,i}},$$

$R_{ols,i}$ = OLS residuals;
$R_{de,i}$ = residuals under shape-restricted estimation.

Testing for selectivity

- The critical value can be calculated through bootstrap.
- Resample the centered OLS residuals.
- Distribution of T_n from the bootstrapped sample $\approx$ the distribution of original $T_n \Longrightarrow$ validity of the test.
- Under H_0, $T_n = o_p(1)$ [Zhang, 2002, Annals]; under H_1, $T_n \rightarrow_p$ a positive constant when λ deviates from a constant in a nontrivial way $\Longrightarrow$ power goes to 1 as n increases.

Monte Carlo simulations

- Selection and outcome equations

$$
\begin{aligned}
Y_i &= D_i\left(\beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i\right), \\
D_i &= \mathbb{I}\{-1 + \gamma_1 X_{1i} + \gamma_2 X_{2i} + \gamma_3 \tilde{W}_i + \nu_i > 0\},
\end{aligned}
$$

- (ε, ν): a mixture of two bivariate normal distributions.
- Compare monotone CF estimator with Heckman's two-step estimator and a kernel-type estimator (Klein-Spady + Robinson).
- The kernel estimator: two bandwidths $c_1 \times h_{cv,1}$, $c_2 \times h_{cv,1}$, where $(h_{cv,1}, h_{cv,2})$ are the bandwidths selected by cross-validation.
  - examine the sensitivity w.r.t. different (c_1, c_2).
- Simulation: monotone CF estimator is
  - robust to non-normal error terms;
  - comparable to the kernel estimator using "good" bandwidths and outperforms the latter using "bad" bandwidths.

Table: Median bias and mean absolute error for estimators of β_1

Method	Bwd	$n = 1000$		$n = 2000$	
	(c_1, c_2)	Med.bias	MAE	Med.bias	MAE
Monotone CF		.0952	.1172	.0690	.0842
Heckit		.1767	.1881	.1862	.1870
Kernel	$(1, 1)$	-.0224	.0945	-.0072	.0676
	$(1, 2)$	-.0693	.1156	-.0483	.0817
	$(1, 3)$	-.1208	.1506	-.0969	.1163
	$(1, 4)$	-.1720	.1881	-.1465	.1552
	$(1/2, 1)$	.0393	.1635	.0418	.1180
	$(1/2, 2)$	-.0074	.1607	.0109	.1144
	$(1/2, 3)$	-.0685	.1666	-.0413	.1216
	$(1/2, 4)$	-.1264	.1819	-.0982	.1411
	$(2, 1)$	-.0442	.0985	-.0224	.0705
	$(2, 2)$	-.0919	.1247	-.0605	.0878
	$(2, 3)$	-.1472	.1634	-.1073	.1237
	$(2, 4)$	-.1906	.1997	-.1562	.1617

Tuning Parameters: Less is More

- For kernel type estimators, the cross-validated bandwidths (computationally more intensive) work better than the 'plug-in' version.
- High-order expansions may not be easy, since the two-stage estimation is involved.
- Our method is much easier to implement.
- Our estimated distribution of the error term and the control function are automatically monotone.

A real-data example

- Data from Schafgans (1998,JAE): Second Malaysian Family Life Survey
- Re-examine the wage equations of the Malaysian Chinese.
- Y : log wage; D : labor market participation; $X = (exper, exper^2, primary, secondary, fail, urban)$;
- $W = (X, unearn, house, land)$. The exclusion restriction builds on non-wage income/wealth.
- Consider male and female workers separately.
- Heckman's two-step estimator; our monotone CF estimator; Schafgans (1998)'s kernel estimator (Ichimura + Robinson).

A real-data example

Decreasing or increasing control function λ ?

- Heckman's model: coefficient on IMR is 0.3891 for men and -0.2787 for women. (IMR is decreasing.)
- The kernel estimates of control function λ (see next page).
- Our selectivity test.

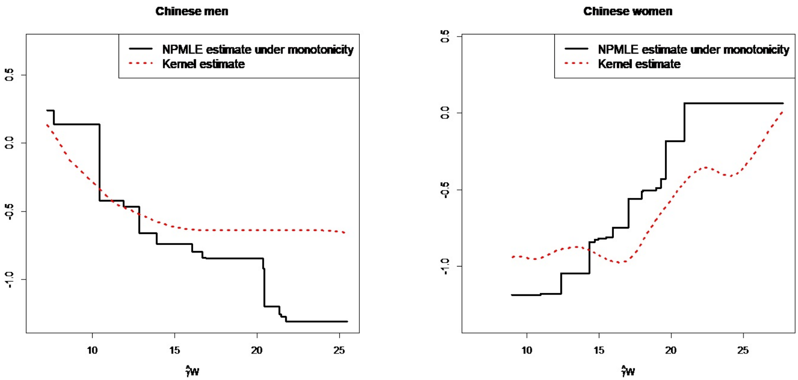
$\Longrightarrow$ Decreasing CF for men and Increasing CF for women.

Table: The p-values for testing the presence of sample selection bias

	Heckit t-test	Monotone CF		Kernel-based test
		Dec	Inc	
Men	.174	**.080**	.528	.866
Women	.357	.592	**.134**	.060

Note: Dec CF for men and Inc CF for women are in bold font.

Figure: The estimated control function $\hat{\lambda}(\hat{\gamma}'W)$



A real-data example

Table: Wage equation for Chinese males; number of total obs =1,190; number of working obs =559.

	Heckit	Monotone CF	Schafgans (1998)
Exper	.1109	.1237	.1051
	[.0887, .1331]	[-.0937, .1396]	[.0837, .1265]
Exper.sq	-.1840	-.2130	-.1750
	[-.2316, -.1364]	[-.2452, -.1411]	[-.2213, -.1287]
Prim.sch	.0235	.0260	.0232
	[-.0205, .0674]	[-.0345, .0760]	[-.0184 .0648]
Secon.sch	.1638	.1693	.1565
	[.1404, .1872]	[.1381, .1924]	[.1341, .1789]
Fail	-.1142	-.1148	-.1298
	[-.2416, .0132]	[-.2496, .0095]	[-.2455 -.0141]
Urban	.0751	.0543	.1047
	[-.0376, .1878]	[-.0311, .1837]	[-.0025, .2119]

Note: The confidence intervals for the Monotone CF estimate are calculated from 500 bootstrap samples.

A real-data example

Table: Wage equation for Chinese females; number of total obs =1,298; number of working obs =371.

	Heckit	Monotone CF	Schafgans (1998)
Exper	.0551	.0394	.0564
	[.0298, .0804]	[.0171, .0870]	[.0295 .0833]
Exper.sq	-.0511	-.0142	-.0635
	[-.1138, .0116]	[-.1274, .0411]	[-.1289, .0019]
Prim.sch	.1094	.1299	.0965
	[.0460, .1728]	[.0002, .1917]	[.0383, .1547]
Secon.sch	.1451	.0859	.0821
	[.0892, .2010]	[.0426, .1885]	[-.0016, .1658]
Fail	-.2145	-.2999	-.4214
	[-.3754, -.0536]	[-.4360, -.0914]	[-.6872, -.1556]
Urban	.0275	.0091	.0163
	[-.0974, .1520]	[-.1082, .1390]	[-.1038, .1364]

Empirical findings

- For male workers, the estimates are similar across three approaches.
- For female workers, the estimated coefficients on *Secon.sch* and *Fail* are quite different between Heckit and kernel approaches.
  - For these two coefficients, estimates under monotone CF are closer to the kernel estimate.
- The negative sorting for female workers might be due to assortative matching of marriage or the discrimination in the labor market.
- Oaxaca decomposition: gender-wage differential explained by observed characteristics
  OLS: 17.09%; Heckman's model: 21.16%; Kernel: 25.12%; Monotone CF: 28.78%.

Extension: A panel selection model

We consider a panel selection model by Kyriazidou (1997)

$$
\begin{aligned}
Y_{it} &= D_{it}(X_{it}'\beta_0 + \alpha_i + \varepsilon_{it}), \\
D_{it} &= \mathbb{I}\{W_{it}'\gamma_0 + \eta_i + \nu_{it} > 0\}.
\end{aligned}
$$

where we only observe the dependent variable for the selected sample with $D_{it} = 1$,; i.e., $Y_{it} = Y_{it}^* D_{it}$ for $i = 1, \cdots, n$ and $t = 1, 2$.

Condition

(i). η_i in the selection equation is independent of W_i and ν_{it}.
(ii). ε_{it} is independent of $\nu_{it'}$ given ν_{it} for $t \neq t'$.

$\Longrightarrow$ α_i is a fixed effect and η_i is a random effect.
$\Longrightarrow$ This is stronger than Kyriazidou (1997) but weaker than Wooldridge (1995).

Panel selection: estimation

We have the following identity:

$$\begin{aligned}&\mathbb{E}[Y_{i1} - Y_{i2}|D_{i1} = 1, D_{i2} = 1, W_i]\\=&(X_{i1} - X_{i2})'\beta_0 + \lambda_{01}(W_{i1}'\gamma_0) - \lambda_{02}(W_{i2}'\gamma_0),\end{aligned}$$

where

$$\lambda_{0t}(W_{it}'\gamma_0) = \int \underbrace{\mathbb{E}[\varepsilon_{it}|\nu_{it} > -W_{it}'\gamma_0 - \eta_i]}_{\text{Monotone}} dF_\eta(\eta_i) \quad \text{for} \quad t = 1, 2.$$

In our **Stage 2**, given $\hat{\gamma}_{nt}$, we estimate the (differenced) outcome equation under the shape restriction for λ_t:

$$\min_{\beta \in \mathbf{B}, \lambda_1, \lambda_2 \in \mathcal{D}} \sum_{D_{i1} = D_{i2} = 1} \left[\Delta Y_i - \Delta X_i'\beta - \lambda_2(W_{i2}'\hat{\gamma}_{n2}) + \lambda_1(W_{i1}'\hat{\gamma}_{n1})\right]^2,$$

where $\Delta Y_i \equiv (Y_{i2} - Y_{i1})$ and $\Delta X_i \equiv (X_{i2} - X_{i1})$.

Conclusion

- We propose a semiparametric sample selection model with a monotonicity constraint on the selection correction function.
- The model lies between the original Heckman selection model and the non/semi-parametric selection model with no restriction on the control function.
- The monotonicity is justified by converting an intuitive dependence concept into conditions such as RTI or RTD.
- The model imposes no parametric distributional assumptions and delivers automatic semiparametric estimation and testing.
- Both the simulation and empirical application illustrate the utility of our proposal.