Identification and Estimation of Average Causal Response Function in a high-dimensional Sample Selection Model

Jun Cai[1], Jian Zhang[2], Yahong Zhou[3]

[1]HUST, [2]NKU, [3]SUFE

August,19,2024

Outline

Introduction

Model Setting and Identification

Estimation

Asymptotics

Monte Carlo Simulations

Application

Extension

Table of Contents

Introduction

We derive the identification and estimation of a semiparametric ACRF with sample selection in a high-dimensional covariate environment.

- An average causal effect (ACR) is usually defined as the expected difference between the outcomes of the treated, and what these outcomes would have been in the absence of treatment, especially for multi-valued treatment (Angrist and Imbens,1995)
- ACR has been widely applied in treatment effect literature with many interesting applications such as drug dosage, hours of exam preparation, cigarette smoking, and years of schooling in the treatment effect literature (Abadie, 2003)
- high-dimensional covariates $=>$ model the endogenous treatment in a more flexible way and justify the validity of IV

What we do?

We considers identification and estimation of a semiparametric ACRF in a high-dimension framework with an application to US Job Corps data:

- Propose the identification moment for ACRF with endogenous treatment and derive Neyman orthogonal moments to estimate two semi-parametric estimators based on it: HDSS and HDSS-series;
- Derived asymptotics for the proposed estimators and both of them are proved to be consistent and asymptotically normal, and Monte Carlo simulations demonstrate that ACRF performs better than the existing IV estimators in many empirically relevant scenarios;

What we do?

We considers identification and estimation of a semiparametric ACRF in a high-dimension framework with an application to US Job Corps data:

- Derive bounds on the proposed ACRF with one single IV with more complex selection mechanism (i.e., the treatment status affects the selection process)
- Apply the proposed methods to NJCS data to evaluate the causal response of residential component and yields new insights with consideration of heterogeneous causal effects with high-dimensional covariates.

Possible contributions

Our model owns four distinct features: high-dimensional setup, nonparametric response function, sample selection, and nontrival empirical findings. Our work may

- contribute to the high-dimensional treatment effect literature (Chernozhukov et al.,2018;Fan et al.,2022) by deriving a set of Neyman orthogonal moments with three nuisance parameters and utilizing the double machine learning techniques to estimate the proposed functional estimators;
- extend ACR to be ACRF which can be varying on covariates and estimate both of them in a unified framework (Angrist and Imbens,1995;Abadie,2003; Callaway et al., 2024);

Possible contributions

Our model owns four distinct features: high-dimensional setup, nonparametric response function, sample selection, and nontrivial empirical findings. Our work may

- consider the identification and estimation of heterogeneous average causal effect function with sample selection and derive bounds on the ACRF with one single IV, which extends the treatment effect bounds in Lee (2009), Chen and Flores (2015) and more recently Bartalotti et al(2023);
- contribute to the broad literature on evaluation of the effectiveness of US Job Corps program (JC) and recent debate on its reform (Chen et al, 2018; Huber et al, 2020; Strittmatter 2019; Thrush,2018).

Table of Contents

Model Setting

Consider a sample selection model with heterogeneous treatment function $m(X_1, D)$ and high-dimensional covariates X_2

$$Y = S \cdot Y^* \tag{1}$$

$$Y^* = m(X_1, D) + g(X_2) + U \tag{2}$$

$$D = D(X_1, X_2, X_3, V) \tag{3}$$

$$S = S(X_1, X_2, \varepsilon) \tag{4}$$

- $X_1 \in R^{d_1}$: low dimensional covariates, $X_2 \in R^p$: high-dimensional covariates
- The $m(X_1, D)$ and $g(X_2)$ are unknown functions and separate additive
- $D(X_1, X_2, X_3, V)$: the treatment equation and $S(X_1, X_2, \varepsilon)$: the selection equation
- (U, V, ε) is joint errors which may be correlated with each others, and X_3 is an instrument variable for binary treatment D

Parameter of Interest

The parameter of interest is

$$\theta(X_1) = m(X_1, 1) - m(X_1, 0), \tag{5}$$

which may vary by X_1, and

$$\mathbb{ACR} = E[\theta(X_1)|S = 1] = E[m(X_1, 1) - m(X_1, 0)|S = 1]. \tag{6}$$

- Parameter in Eq.(5) is an average causal response function (ACRF) and Parameter in Eq.(6) is the average causal response (ACR) for binary treatment (Angrist and Imbens,1995;Abadie,2003;Callaway et al., 2024)
- ACRF could be regarded as a conditional average treatment effect (CATE) under strong assumps ($Y(1) - Y(0)$ is identical for all individuals)

Identification of Parameter of Interest

Assumption 1

Given X_1 and X_2, X_3 is independent of (U, V, ε).

Since $S = S(X_1, X_2, \varepsilon)$, this assumption implies that X_3 is independent of selection S and unobserved heterogeneity U (i.e. the source of endogeneity) for given values of X_1 and X_2. This is an analog of exclusive restriction.

Assumption 2

$P(S = 1|X_1) > 0$ *with probability one.*

For almost all possible values of X_1, outcome Y is observed $(S = 1)$ with positive probability. This allows us to identify the casual effect $\theta(x_1)$ for any given value of x_1.

Assumption 3

Let $\mu(X_1, X_2, X_3) = E[D|X_1, X_2, X_3, S = 1]$. *The propensity score function* $\mu(\cdot)$ *satisfies that*

$$P(\mu(X_1, X_2, X_3) \neq E[\mu(X_1, X_2, X_3)|X_1, X_2, S = 1]|X_1 = x_1, S = 1) > 0.$$

This assumption implies that X_3 truly affects D. This is an analog of relevant condition.

Summary of Assumptions 1 to 3: X_3 can exogenously affect treatment assignment D without altering the sample selection mechanism S, this tells us that X_3 is a valid instruments in our context.

Identification of Parameter of Interest

$$
\begin{aligned}
& E[Y|X_1, X_2, X_3, S=1] \\
=& E[m(X_1,0) + (m(X_1,1) - m(X_1,0))D + g(X_2) + U|X_1, X_2, X_3, S=1] \\
=& m(X_1,0) + (m(X_1,1) - m(X_1,0))E[D|X_1,X_2,X_3,S=1] + g(X_2) \\
& + E[U|X_1,X_2,X_3,S=1] \\
=& m(X_1,0) + g(X_2) + E[U|X_1,X_2,X_3,S=1] \\
& + (m(X_1,1) - m(X_1,0))E[D|X_1,X_2,X_3,S=1] \\
=& m(X_1,0) + g(X_2) + f(X_1,X_2) \\
& + (m(X_1,1) - m(X_1,0))E[D|X_1,X_2,X_3,S=1] \\
=& \tilde{m}(X_1,X_2) + \theta(X_1)\mu(X_1,X_2,X_3)
\end{aligned}
$$

The slope coefficient $\theta(X_1)$ is identified by exploring the ratio of the variation in $E[Y|X_1, X_2, X_3, S=1]$ to the variation in $\mu(X_1, X_2, X_3)$ caused exogenously by the change of X_3.

Table of Contents

Neyman Orthogonal Moments

Recall:

$$E[Y|X_1, X_2, X_3, S=1] = \widetilde{m}(X_1, X_2) + \theta(X_1)\mu(X_1, X_2, X_3) \quad (7)$$

Conditioning on $(X_1, X_2, S=1)$, by LIE:
$E[Y|X_1, X_2, S=1] = \widetilde{m}(X_1, X_2) + \theta(X_1)\widetilde{\mu}(X_1, X_2)$,
where $\widetilde{\mu}(X_1, X_2) = E[D|X_1, X_2, S=1]$. Therefore,

$$\widetilde{m}(X_1, X_2) = h(X_1, X_2) - \theta(X_1)\widetilde{\mu}(X_1, X_2), \quad (8)$$

where $h(X_1, X_2) = E[Y|X_1, X_2, S=1]$.
Also, Eq.(7) can be written as a moment condition

$$E[Y - \widetilde{m}(X_1, X_2) - D\theta(X_1)|X_1, X_2, X_3, S=1] = 0. \quad (9)$$

Plug (8) into (9),

$$E\Big[Y - h(X_1, X_2) - \theta(X_1)(D - \widetilde{\mu}(X_1, X_2))\Big|X_1, X_2, X_3, S=1\Big] = 0. \quad (10)$$

Neyman Orthogonal Moments

Since X_2 is of high dimension, Neyman orthogonal moments can be derived based on the identification strategy as follows

$$E\Big[(\mu(X_1, X_2, X_3) - \widetilde{\mu}(X_1, X_2)) \cdot \Big[Y - h(X_1, X_2) - \theta(X_1)(D - \widetilde{\mu}(X_1, X_2))\Big] \Big| X_1 = x_1, S = 1\Big] = 0 \tag{11}$$

- It follows a similar idea as Example 1.1 in Chernozhukov et al. (2018)
- There are three nuisance parameters $\eta_0 = (\mu(\cdot), h(\cdot), \widetilde{\mu}(\cdot))$.
- We can verify the Neyman orthogonality condition holds with respect to the nuisance parameters.

Estimation

Based on the Neyman orthogonal moment in Eq.(11), we can solve

$$\theta(x_1) = \frac{E\left[(\mu(X_1, X_2, X_3) - \widetilde{\mu}(X_1, X_2))(Y - h(X_1, X_2))\middle| X_1 = x_1, S = 1\right]}{E\left[(\mu(X_1, X_2, X_3) - \widetilde{\mu}(X_1, X_2))(D - \widetilde{\mu}(X_1, X_2))\middle| X_1 = x_1, S = 1\right]}, \quad (12)$$

- Note that the denominator is non-zero by Assumption 3.
- $\theta(x_1)$ is a ratio of two conditional expectation and a multiple step procedure is proposed.
- Depends on different techniques used in the last step, we proposed two estimators: HDSS and HDSS-series.

Estimation Procedure: HDSS

Step 1 For each $k = 1, 2, \ldots, K$, we estimates within sample I_k^c

(i) Let $\widetilde{X}_3$ be (X_1, X_2, X_3) or a series of functions of (X_1, X_2, X_3). Consider

$$P(D = 1 | X_1, X_2, X_3, S = 1) \approx \Lambda(\widetilde{X}_3' \alpha).$$

α can be estimated on the subsample I_k^c by logistic regression with ℓ_1 penalty, denoted by $\widehat{\alpha}_{-k}$.

$$\widehat{\mu}(X_1, X_2, X_3; I_k^c) = \widehat{E}[D | X_1, X_2, X_3, S = 1]|_{I_k^c} = \Lambda(\widetilde{X}_3' \widehat{\alpha}_{-k});$$

(ii) Let $\widetilde{X}_2$ be (X_1, X_2) or a series of functions of (X_1, X_2). Regress Y on $\widetilde{X}_2$ with ℓ_1 penalty (i.e. LASSO), and obtain

$$\widehat{h}(X_1, X_2; I_k^c) = \widehat{E}[Y | X_1, X_2, S = 1]|_{I_k^c} = \widetilde{X}_2' \widehat{\gamma}_{-k};$$

(iii) Similar to **Step 1(i)**, we estimate

$$P(D = 1 | X_1, X_2, S = 1) \approx \Lambda(\widetilde{X}_2' \nu)$$

on the subsample I_k^c by logistic regression with ℓ_1 penalty, denoted by $\widehat{\nu}_{-k}$. It yields

$$\widehat{\widehat{\mu}}(X_1, X_2; I_k^c) = \widehat{E}[D | X_1, X_2, S = 1]|_{I_k^c} = \Lambda(\widetilde{X}_2' \widehat{\nu}_{-k}).$$

Estimation Procedure: HDSS

Step 2 Denote $K_h(x_1; X_{1i}) = \frac{1}{h^{d_1}} K\left(\frac{X_{1i}-x_1}{h}\right)$.

$$\widehat{\theta}_{Ker}(x_1) = \frac{\sum_{k=1}^{K} \sum_{i \in I_k} S_i \Delta\hat{\mu}_i \cdot \Delta Y_i \cdot K_h(x_1; X_{1i})}{\sum_{k=1}^{K} \sum_{i \in I_k} S_i \Delta\hat{\mu}_i \cdot \Delta D_i \cdot K_h(x_1; X_{1i})} \tag{13}$$

where

$$\begin{aligned} \Delta\hat{\mu}_i &:= \widehat{\mu}(X_{1i}, X_{2i}, X_{3i}; I_k^c) - \widehat{\tilde{\mu}}(X_{1i}, X_{2i}; I_k^c) \\ \Delta Y_i &:= Y_i - \widehat{h}(X_{1i}, X_{2i}; I_k^c) \\ \Delta D_i &:= D_i - \widehat{\tilde{\mu}}(X_{1i}, X_{2i}; I_k^c) \end{aligned}$$

Which is denoted as High-dimensional sample selection estimator, i.e., HDSS estimator.

Estimation Procedure: HDSS-series

To avoid the boundary bias introduced by nonparametric kernel estimation, we also propose a series estimation procedure which is more precise and robust to boundary points within the range of X_1.

$$E\Big[p_k(X_1)\big(\mu(X_1,X_2,X_3)-\widetilde{\mu}(X_1,X_2)\big)\Big[Y-h(X_1,X_2)-$$
$$\theta(X_1)\big(D-\widetilde{\mu}(X_1,X_2)\big)\Big]\Big|S=1\Big]=0,\quad k=1,2,\ldots$$

Thus by series approximation $\theta(X_1)\approx\sum_{k=1}^{K_n}\beta_k p_k(x_1)$,

$$E\Big[p_k(X_1)\big(\mu(X_1,X_2,X_3)-\widetilde{\mu}(X_1,X_2)\big)\Big[Y-h(X_1,X_2)-$$
$$\sum_{k=1}^{K_n}\beta_k p_k(x_1)\big(D-\widetilde{\mu}(X_1,X_2)\big)\Big]\Big|S=1\Big]\approx 0,\quad k=1,2,\ldots,K_n \tag{14}$$

Estimation Procedure: HDSS-series

Step 1 The same as HDSS estimator;

Step 2 Denote z_i, w_i, q_i and their estimates $\widehat{z}_i$, $\widehat{w}_i$, $\widehat{q}_i$ as

$$z_i = \begin{pmatrix} p_1(X_{1i})\big(\mu - \widetilde{\mu}(X_{1i}, X_{2i})\big) \\ p_2(X_{1i})\big(\mu - \widetilde{\mu}(X_{1i}, X_{2i})\big) \\ \vdots \\ p_{K_n}(X_{1i})\big(\mu - \widetilde{\mu}(X_{1i}, X_{2i})\big) \end{pmatrix}; \widehat{z}_i = \begin{pmatrix} p_1(X_{1i})\big(\widehat{\mu} - \widehat{\widetilde{\mu}}(X_{1i}, X_{2i})\big) \\ p_2(X_{1i})\big(\widehat{\mu} - \widehat{\widetilde{\mu}}(X_{1i}, X_{2i})\big) \\ \vdots \\ p_{K_n}(X_{1i})\big(\widehat{\mu} - \widehat{\widetilde{\mu}}(X_{1i}, X_{2i})\big) \end{pmatrix};$$

$$w_i = \begin{pmatrix} p_1(X_{1i})\big(D_i - \widetilde{\mu}(X_{1i}, X_{2i})\big) \\ p_2(X_{1i})\big(D_i - \widetilde{\mu}(X_{1i}, X_{2i})\big) \\ \vdots \\ p_{K_n}(X_{1i})\big(D_i - \widetilde{\mu}(X_{1i}, X_{2i})\big) \end{pmatrix}; \quad \widehat{w}_i = \begin{pmatrix} p_1(X_{1i})\big(D_i - \widehat{\widetilde{\mu}}(X_{1i}, X_{2i})\big) \\ p_2(X_{1i})\big(D_i - \widehat{\widetilde{\mu}}(X_{1i}, X_{2i})\big) \\ \vdots \\ p_{K_n}(X_{1i})\big(D_i - \widehat{\widetilde{\mu}}(X_{1i}, X_{2i})\big) \end{pmatrix};$$

$$q_i = Y_i - h(X_{1i}, X_{2i}); \qquad \widehat{q}_i = Y_i - \widehat{h}(X_{1i}, X_{2i}).$$

Estimation Procedure: HDSS-series

Step 2 Continued

Then Eq.(14) is equivalent to $E[z_i(q_i - w_i'\beta^{K_n})] \approx 0$. A series estimator is

$$\widehat{\beta^{K_n}} = \Big(\sum_{i=1}^{N} S_i \widehat{z}_i \widehat{w}_i'\Big)^{-1}\Big(\sum_{i=1}^{N} S_i \widehat{z}_i \widehat{q}_i\Big) \tag{15}$$

$$\widehat{\theta}_{Series}(x_1) = p^{K_n}(x_1)'\widehat{\beta^{K_n}}. \tag{16}$$

In practise, the polynomial order K_n is chose by leave-one-out Cross-Validation. We denote this estimator as HDSS-series estimator.

Table of Contents

Asymptotics for HDSS Estimator (I)

Under assumptions 1-3, sample splitting, kernel estimation and first stage converge rate assumptions, we obtain following asymptotic linear representation for ACRF based on kernel estimation:

Theorem 1

If Assumptions 1 to 6 hold and $x_1 \in \mathcal{T}$ is an interior point, then

$$\sup_{x \in \mathcal{T}} \left| \widehat{\theta}_{Ker}(x_1) - \theta(x_1) \right| = O_p\left(h^s + \left(\log(n)/(nh^{d_1}) \right)^{1/2} \right),$$

$$\widehat{\theta}_{Ker}(x_1) - \theta(x_1) - h^s B(x_1) = \mathbb{V}(x_1)^{-1} \frac{1}{h^{d_1}} (\mathbb{P}_n - \mathbb{P}) \frac{S_i \eta_i^D \nu_i^Y}{P(S=1)} K\left(\frac{X_{1i} - x_1}{h} \right) + R_\theta(x_1)$$

where $\eta_i^D = \mu(X_{1i}, X_{2i}, X_{3i}) - \widetilde{\mu}(X_{1i}, X_{2i})$,
$\nu_i^Y = Y_i - h(X_{1i}, X_{2i}) - \theta(X_{1i}) \left(D_i - \widetilde{\mu}(X_{1i}, X_{2i}) \right)$, *and*

$$\sup_{x \in \mathcal{T}} |R_\theta(x_1)| = o_p\left(h^s + \left(\log(n)/(nh^{d_1}) \right)^{1/2} \right).$$

$B(x_1)$ *and* $\mathbb{V}(x_1)$ *are defined in our paper.*

Note that $E[\eta_i^D \nu_i^Y | X_{1i}, S_i = 1] = 0$ by the Neyman orthogonal moment. The asymptotic linear representation above implies an asymptotic normal distribution.

Asymptotics for HDSS Estimator (II)

To reduce boundary bias, we assume a subset $\mathcal{T}$ of $supp(X_1)$ that excludes the boundary area and propose a trimmed average casual response estimator as follows

$$\widehat{\mathbb{ACR}}_{\mathcal{T},Ker} = \frac{1}{n_{TS}} \sum_{i=1}^{n} S_i 1\left(X_{1i} \in \mathcal{T}\right) \widehat{\theta}_{Ker}(X_{1i}),$$

where $n_{TS} = \sum_{i=1}^{n} S_i 1\left(X_{1i} \in \mathcal{T}\right)$.

Theorem 2

If Assumptions 1 to 7 hold and $nh^{2s} \to 0$, then

$$\sqrt{n}\left(\widehat{\mathbb{ACR}}_{\mathcal{T},Ker} - \mathbb{ACR}_{\mathcal{T}}\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{acr}),$$

where

$$\sigma^2_{acr} = \frac{1}{P(S=1)} E\left[\left\{\eta_i^D \nu_i^Y\right\}^2 \frac{1\left(X_{1i} \in \mathcal{T}\right) f_{X_1|S=1}(X_{1i})^2}{P(X_1 \in \mathcal{T}|S=1)^2 \mathbb{V}(X_{1i})^2} \middle| S_i = 1\right] + \frac{Var\left(\theta(X_{1i}) \middle| X_{1i} \in \mathcal{T}, S_i = 1\right)}{P(X_1 \in \mathcal{T}, S=1)}.$$

The variance consists of : (i) estimation of $\theta(X_{1i})$; (ii) taking average of $\theta(X_{1i})$.

Asymptotics for HDSS-series Estimator (I)

Theorem 3

If Assumptions 1 to 4, Assumptions 8 to 9 hold, then

(i) $\sup_{x_1 \in \mathcal{X}_1} |\widehat{\theta}_{Series}(x_1) - \theta(x_1)| = O_p\left(K_n^{-\alpha}\zeta_0(K_n)^2 + \zeta_0(K_n)/\sqrt{n}\right) = o_p(1)$.

(ii) *Denote* $\widehat{\Sigma}_{\theta,K_n}(x_1)$ *as*

$$\widehat{\Sigma}_{\theta,K_n}(x_1) = p^{K_n}(x_1)'\left(n^{-1}\sum_{i=1}^{n} S_i \widehat{z}_i \widehat{w}_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} S_i \widehat{z}_i \widehat{z}_i' \{\widehat{q}_i - \widehat{w}_i'\beta^{K_n}\}^2\right) \left(n^{-1}\sum_{i=1}^{n} S_i \widehat{w}_i \widehat{z}_i'\right)^{-1} p^{K_n}(x_1).$$

If x_1 *satisfies that* $\liminf_{n\to\infty} \frac{\|p^{K_n}(x_1)\|}{\zeta_0(K_n)} = c(x_1) > 0$, *then*

$$\sqrt{n}\widehat{\Sigma}_{\theta,K_n}(x_1)^{-1/2}\left(\widehat{\theta}_{Series}(x_1) - \theta(x_1) - \mathfrak{B}_n(x_1)\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\mathfrak{B}_n(x_1) = O(K_n^{-\alpha}\zeta_0(K_n)^2)$ *is a bias term defined in our paper.*

This thm presents both consistency and asymptotic normality of HDSS-series est. A standard t-test is applicable for inference here.

Asymptotics for HDSS-series Estimator (II)

Our suggested series estimator of average causal response is

$$\widehat{\mathbb{ACR}}_{Series} = \frac{\sum_{i=1}^{n} S_i \widehat{\theta}_{Series}(X_{1i})}{\sum_{i=1}^{n} S_i},$$

Theorem 4

Denote $\Sigma_{acr} = E\left[S_i\{\eta_i^D \nu_i^Y\}^2 q(X_{1i})^2\right] + \frac{Var\left(\theta(X_{1i}) \middle| S_i=1\right)}{P(S_i=1)}$ *and*

$$\widehat{\Sigma}_{acr} = \left({n_S}^{-1}\sum_{i=1}^{n} S_i p^{K_n}(X_{1i})'\right)\left(n^{-1}\sum_{i=1}^{n} S_i \widehat{z}_i \widehat{w}_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} S_i \widehat{z}_i \widehat{z}_i' \{\widehat{q}_i - \widehat{w}_i' \beta^{K_n}\}^2\right)\cdot$$

$$\left(n^{-1}\sum_{i=1}^{n} S_i \widehat{w}_i \widehat{z}_i'\right)^{-1}\left({n_S}^{-1}\sum_{i=1}^{n} S_i p^{K_n}(X_{1i})\right) + \frac{1}{n_S}\sum_{i=1}^{n} S_i\left\{\widehat{\theta}_{Series}(X_{1i}) - \overline{\theta}_S\right\}^2$$

where $n_S = \sum_{i=1}^{n} S_i$, $\overline{S} = n^{-1}\sum_{i=1}^{n} S_i$, *and* $\overline{\theta}_S = n_S^{-1}\sum_{i=1}^{n} S_i \widehat{\theta}_{Series}(X_{1i})$. *If Assumption 1 to Assumption 4, Assumption 8 to Assumption 10 hold, then* $\sqrt{n}\left(\widehat{\mathbb{ACR}}_{Series} - \mathbb{ACR}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{acr})$ *and* $\widehat{\Sigma}_{acr} \xrightarrow{p} \Sigma_{acr}$ *as* $n \to \infty$.

Comments: HDSS v.s. HDSS-series

- As for the estimation of ACRF $\theta(x_1)$, both estimators are asymptotic normally distributed. However, kernel-based HDSS estimator has to exclude the boundary area of $supp(X_1)$ while HDSS-series estimator presents more robustness near the boundary.

- The asymptotic variance of both estimators for $\mathbb{ACR}$ have two components: (i) The first part is because of the estimation of $\theta(\cdot)$; (ii) The second part is the variance from averaging $\theta(X_{1i})$. Moreover, both estimators have $\sqrt{n}-$convergent rates.

A Short Note for Efficiency

- According to Theorem 2 and Theorem 4, if we ignore trimming and assume all regular assumptions hold, both estimators for $\mathbb{ACR}$ have the same asymptotic variance, i.e. $\Sigma_{acr} = \sigma^2_{acr}$. This can be verified algebraically.
- If we further assume that there is no sample selection, this asymptotic variance is also the same as semiparametric efficient variance for average linear regression function, see Graham and de Xavier Pinto(2022). But our approach attains this efficiency in a high-dimensional setting.

Table of Contents

Monte Carlo Settings

We consider the following DGP with a benchmark case

$$Y = S * Y^*$$

$$Y^* = X_1 + 2X_1 D + X_1^2 D + 3X_{21} + 4X_{22}^2 + U \quad (DGP1)$$

$$\text{with} \quad m(X_1, D) = X_1 + 2X_1 D + X_1^2 D, \quad g(X_2) = 3X_{21} + 4X_{22}^2$$

$$D = 1\{V < X_1 + 2X_{22} + 3X_3\}$$

$$S = 1\{\varepsilon < 2 + 2X_1 + X_{22}\}$$

$$\text{where} \quad X_1 = (0.5X_{21} + W)/\sqrt{1.25} \sim N(0,1) \text{ and is correlated with } X_{21};$$

$$\text{with} \quad X_2 \sim N(0, \Sigma_p), (i,j)\text{-th element: } \Sigma_{ij} = \rho^{|i-j|} \text{ (Default: } \rho = 0.6)$$

$$\text{and} \quad W \sim N(0,1), \quad X_3 \sim N(0,1)$$

$$\begin{pmatrix} U \\ V \\ \varepsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0.5 & 0.5 \\ 0.5 & \sigma_v^2 & 0 \\ 0.5 & 0 & \sigma_\varepsilon^2 \end{pmatrix}\right).$$

With $\sigma_u^2 = \sigma_v^2 = \sigma_\varepsilon^2 = 1$.

...

Monte Carlo Settings

The true function of θ and the the average causal response estimator (ACR) is

$$\theta(X_1) = 2X_1 + X_1^2$$

$$E[\theta(X_1)|S = 1] = 2E[X_1|S = 1] + E[X_1^2|S = 1].$$

Note that $Corr(X_1, X_{21}) \neq 0$, thus we compare results of the proposed method with nonpara 2SLS estimators.

Monte Carlo Settings

We also investigate several scenarios with exponential decaying coefficients on high-dimensional X_2, binary X_3, and discrete X_1:

- DGP2: $Y^* = X_1 + 2X_1 D + X_1^2 D + \sum_{j=1}^{p}(0.8)^{j-1}X_{2j} + U$ with $m(X_1, D) = X_1 + 2X_1 D + X_1^2 D, \quad g(X_2) = \sum_{j=1}^{p}(0.8)^{j-1}X_{2j}$
- DGP3: Binary IV, $X_3 \sim Binomial(0.5)$
- DGP4: Discrete X_1 with equal prob of $1/3$ $Pr(X_1 = 0) \approx Pr(X_1 = 1) \approx Pr(X_1 = 2) \approx 1/3$

$$X_1 = \begin{cases} 0, & 0.5X_{21} + W < -\frac{\sqrt{5}}{4}, \\ 1, & -\frac{\sqrt{5}}{4} \leq 0.5X_{21} + W < \frac{\sqrt{5}}{4}, \\ 2, & 0.5X_{21} + W \geq \frac{\sqrt{5}}{4}. \end{cases} \text{ and is correlated with } X_{21}.$$

Simulation tables: The first row of each panel reports the Bias and the second row reports the RMSE. Replicate 100 times.

Simulation Results for DGP 1

Table: ACRF with High Dimension Sample Selection Model by DGP 1

$\theta(X_1)$		NPIV-oracle	NPIV-lasso	HDSS-nonorth	HDSS	HDSS-series
	ACR	-0.028	-0.216	-0.310	-0.056	-0.015
		0.226	0.313	0.542	0.351	0.243
	$X_1 = -1$	-0.079	-0.552	-1.083	0.195	0.201
		0.353	0.732	2.635	0.470	0.486
	$X_1 = 0$	0.063	0.110	-0.118	0.035	0.006
$N = 500, p = 100$		0.188	0.190	0.474	0.216	0.234
	$X_1 = 0.5$	0.054	0.071	-0.318	-0.038	-0.032
		[illegible]	0.215	0.526	0.235	0.211
	$X_1 = 1$	-0.010	-0.200	-0.622	-0.172	-0.065
		0.205	0.392	0.826	0.368	0.278
	ACR	0.043	-0.172	0.574	0.068	0.072
		0.186	0.264	7.626	0.489	0.230
	$X_1 = -1$	-0.070	-0.665	-1.459	0.231	0.280
		0.371	0.846	2.319	0.513	0.554
	$X_1 = 0$	0.071	0.074	0.007	0.087	0.044
$N = 500, p = 200$		0.181	0.150	0.391	0.212	0.199
	$X_1 = 0.5$	0.073	0.086	-0.260	-0.005	0.008
		0.191	0.196	0.496	0.215	0.169
	$X_1 = 1$	0.030	-0.129	-0.701	-0.136	-0.029
		0.222	0.352	0.890	0.335	0.273
	ACR	-0.019	-0.214	0.061	-0.022	-0.001
		0.109	0.241	[illegible]	0.356	0.105
	$X_1 = -1$	-0.062	-0.220	0.319	0.038	0.077
		0.187	0.336	0.475	0.227	0.206
	$X_1 = 0$	0.042	0.111	0.118	0.021	-0.016
$N = 2000, p = 100$		0.093	0.151	0.234	0.107	0.084
	$X_1 = 0.5$	0.023	-0.025	-0.273	-0.038	-0.022
		0.101	0.126	0.353	0.135	0.102
	$X_1 = 1$	-0.042	-0.358	-0.735	-0.132	-0.043
		0.129	0.395	0.794	0.220	0.140
	ACR	-0.010	-0.229	0.056	0.178	0.008
		0.105	0.251	1.253	2.364	0.095
	$X_1 = -1$	-0.068	-0.311	-0.450	0.027	0.080
		0.180	0.376	0.592	0.221	0.216
	$X_1 = 0$	0.048	0.115	0.115	0.026	-0.001
$N = 2000, p = 200$		0.094	0.160	0.242	0.105	0.088
	$X_1 = 0.5$	0.037	-0.002	-0.244	-0.033	-0.015
		0.107	0.115	0.329	0.129	0.112
	$X_1 = 1$	-0.019	-0.337	-0.687	-0.118	-0.042
		0.124	0.368	0.751	0.203	0.137

Simulation Results for DGP 2

Table: ACRF with High Dimension Sample Selection Model by DGP 2

		NPIV-oracle	NPIV-lasso	HDSS-nonorth	HDSS	HDSS-series
$N = 500, p = 100$	ACR	0.039	-0.218	-0.366	-0.074	-0.034
		0.474	0.321	0.597	0.349	0.250
	$X_1 = -1$	-0.098	-0.552	-1.014	0.166	0.152
		0.776	0.729	2.488	0.455	0.482
	$X_1 = 0$	0.038	0.133	-0.101	0.028	0.018
		0.442	0.223	0.485	0.241	0.227
	$X_1 = 0.5$	[illegible]	0.090	-0.327	-0.042	-0.034
		[illegible]	0.234	0.529	0.239	0.233
	$X_1 = 1$	[illegible]	-0.193	-0.681	-0.176	-0.083
		0.641	0.394	0.871	0.386	0.316
$N = 500, p = 200$	ACR	0.098	-0.175	0.349	0.046	0.066
		0.459	0.270	5.903	0.409	0.251
	$X_1 = -1$	-0.287	-0.674	-1.532	0.178	0.216
		0.957	0.852	2.596	0.512	0.529
	$X_1 = 0$	-0.013	-0.100	0.038	0.094	0.055
		0.419	0.182	0.378	0.218	0.211
	$X_1 = 0.5$	0.121	0.107	-0.245	0.010	0.016
		0.508	0.207	0.510	0.223	0.181
	$X_1 = 1$	0.253	-0.127	-0.741	-0.134	-0.035
		0.680	0.351	0.925	0.352	0.295
$N = 2000, p = 100$	ACR	-0.017	-0.220	0.048	-0.044	-0.020
		0.226	0.247	2.052	0.349	0.108
	$X_1 = -1$	-0.055	-0.219	0.283	0.037	0.075
		0.406	0.335	0.453	0.227	0.205
	$X_1 = 0$	-0.042	0.137	0.142	0.017	-0.019
		0.202	0.174	0.247	0.106	0.085
	$X_1 = 0.5$	-0.018	-0.004	-0.273	-0.037	-0.026
		0.244	0.129	0.354	0.134	0.104
	$X_1 = 1$	0.017	-0.354	-0.795	-0.138	-0.046
		0.312	0.392	0.852	0.225	0.152
$N = 2000, p = 100$	ACR	0.017	-0.208	0.033	0.200	0.011
		0.207	0.233	1.389	2.490	0.096
	$X_1 = -1$	-0.130	-0.311	-0.408	0.021	0.072
		0.473	0.377	0.567	0.226	0.218
	$X_1 = 0$	-0.047	0.145	0.137	0.027	-0.000
		0.196	0.186	0.255	0.113	0.092
	$X_1 = 0.5$	0.009	0.025	-0.248	-0.034	-0.015
		0.229	0.125	0.333	0.132	0.113
	$X_1 = 1$	0.075	-0.325	-0.731	-0.117	-0.040
		0.307	0.360	0.790	0.205	0.149

Simulation Results for DGP 3

Table: ACRF with High Dimension Sample Selection Model by DGP 3

$\theta(X_1)$		NPIV-oracle	NPIV-lasso	HDSS-nonorth	HDSS	HDSS-series
$N = 500, p = 100$	ACR	-0.035	-0.378	-1.827	0.174	-0.019
		0.292	0.456	11.126	4.911	0.457
	$X_1 = -1$	-0.023	-0.365	-5.223	0.200	0.010
		0.361	0.565	21.627	0.676	0.534
	$X_1 = 0$	0.050	0.036	-0.881	0.008	0.111
		0.238	0.091	1.457	0.278	0.332
	$X_1 = 0.5$	[illegible]	-0.129	-0.988	-0.113	0.028
		[illegible]	0.253	1.304	0.459	0.446
	$X_1 = 1$	[illegible]	-0.533	-0.666	-0.060	-0.126
		0.408	0.695	1.718	0.978	0.852
$N = 500, p = 200$	ACR	0.042	-0.321	0.140	-9.360	0.036
		0.262	0.406	4.558	102.737	0.458
	$X_1 = -1$	[illegible]	-0.530	-42.147	0.203	0.059
		0.421	0.776	361.237	0.611	0.638
	$X_1 = 0$	0.075	[illegible]	-0.086	0.074	0.177
		0.235	[illegible]	17.385	0.316	0.376
	$X_1 = 0.5$	0.065	-0.051	-1.026	-0.057	0.111
		0.287	0.217	1.468	0.405	0.362
	$X_1 = 1$	0.021	-0.421	-1.006	-0.046	-0.095
		0.385	0.600	1.645	0.865	0.617
$N = 2000, p = 100$	ACR	-0.044	-0.365	[illegible]	-0.315	-0.021
		0.169	0.401	[illegible]	2.138	0.234
	$X_1 = -1$	0.042	0.029	[illegible]	0.051	0.035
		0.191	0.204	0.918	0.282	0.234
	$X_1 = 0$	0.034	0.027	[illegible]	0.007	0.010
		0.126	0.068	0.536	0.164	0.150
	$X_1 = 0.5$	-0.021	-0.239	-0.806	-0.096	-0.020
		0.169	0.269	0.982	0.212	0.150
	$X_1 = 1$	-0.110	-0.679	-1.029	-0.172	-0.074
		0.251	0.732	1.471	0.389	0.364
$N = 2000, p = 200$	ACR	0.006	-0.305	0.373	-0.060	0.088
		0.151	0.347	3.443	2.765	0.243
	$X_1 = -1$	0.019	-0.038	-0.900	0.033	0.043
		0.173	0.199	1.179	0.243	0.249
	$X_1 = 0$	0.058	0.020	-0.228	0.025	0.006
		0.127	0.050	0.542	0.175	0.146
	$X_1 = 0.5$	0.024	-0.192	-0.797	-0.069	0.031
		0.154	0.229	0.976	0.250	0.184
	$X_1 = 1$	-0.045	-0.564	-1.032	-0.059	0.085
		0.205	0.632	1.364	0.445	0.334

Simulation: HDSS(Trimmed) v.s. HDSS(Not trimmed)

Table: Estimation ACR based on HDSS with Trimming

Design	Sample Size	X2 dim (p)	HDSS		HDSS-trimmed	
			Bias	RMSE	Bias	RMSE
DGP1	500	100	-0.049	0.350	-0.078	0.226
	500	200	0.078	0.491	0.001	0.232
	2000	100	-0.052	0.359	-0.062	0.132
	2000	200	0.192	2.365	-0.020	0.100
DGP2	500	100	-0.074	0.349	-0.104	0.244
	500	200	0.046	0.409	-0.017	0.237
	2000	100	-0.044	0.349	-0.051	0.128
	2000	200	0.200	2.490	-0.028	0.105
DGP3	500	100	0.179	4.911	-0.209	0.493
	500	200	-9.357	102.737	-0.136	0.447
	2000	100	-0.305	2.437	-0.138	0.302
	2000	200	-1.093	13.825	-0.073	0.265

Notes: The construction of trimmed estimator is as follow: if the Non-trimmed HDSS estimator = mean (A_i/B_i), then the trimming level is defined as $mean(B_i) * h^2 * n^{-1/2}$, and we dropped the observations with $|B_i| <$ trimming level. Replicate 100 times.

Simulation Results for DGP 4

Table: ACRF with High Dimension Sample Selection Model by DGP 4

$\theta(X_1)$		NPIV-oracle	NPIV-lasso	HDSS-nonorth	HDSS	HDSS-series
	ACR	-0.027	0.119	-0.979	-0.011	-0.011
		0.214	0.319	1.006	0.232	0.232
	$X_1 = 0$	0.019	0.059	-4.907	0.158	0.158
		0.218	0.146	5.093	0.315	0.315
$N = 500, p = 100$	$X_1 = 1$	[illegible]	0.348	0.951	-0.027	-0.027
		[illegible]	0.443	1.006	0.194	0.194
	$X_1 = 2$	[illegible]	-0.065	-1.061	-0.086	-0.086
		0.280	0.560	1.197	0.361	0.361
	ACR	0.025	0.223	-0.992	0.015	0.015
		0.196	0.352	1.018	0.212	0.212
	$X_1 = 0$	0.003	0.081	-5.983	0.181	0.181
		0.224	0.183	6.288	0.317	0.317
$N = 500, p = 200$	$X_1 = 1$	-0.008	0.414	1.152	-0.023	-0.023
		0.216	0.489	1.227	0.246	0.246
	$X_1 = 2$	0.007	0.037	-1.002	-0.160	-0.160
		0.287	0.539	1.162	0.397	0.397
	ACR	-0.008	0.092	-0.609	0.012	0.012
		0.112	0.162	0.622	0.106	0.106
	$X_1 = 0$	0.034	0.090	-1.712	0.082	0.082
		0.116	0.132	1.727	0.137	0.137
$N = 2000, p = 100$	$X_1 = 1$	-0.018	0.211	0.368	0.008	0.008
		0.101	0.241	0.386	0.095	0.095
	$X_1 = 2$	-0.012	-0.033	-0.503	-0.024	-0.024
		0.154	0.226	0.573	0.139	0.139
	ACR	-0.007	0.100	-0.673	0.013	0.013
		0.105	0.163	0.683	0.109	0.109
	$X_1 = 0$	0.015	0.068	-2.155	0.082	0.082
		0.116	0.121	2.178	0.132	0.132
$N = 2000, p = 200$	$X_1 = 1$	-0.018	0.234	0.448	-0.003	-0.003
		0.108	0.264	0.469	0.093	0.093
	$X_1 = 2$	-0.019	-0.045	-0.508	-0.033	-0.033
		0.139	0.224	0.579	0.164	0.164

Table of Contents

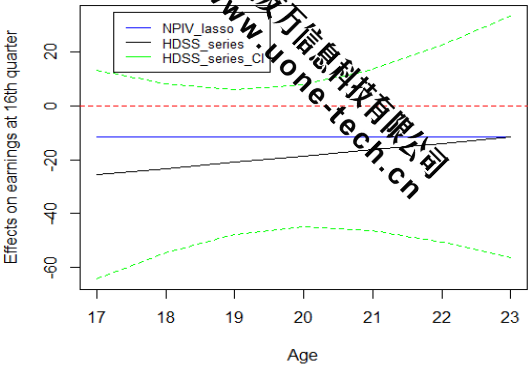
Application: Residential Component in Job Corps Program

- We apply the proposed methods to explore the average causal effect of residential component and its heterogeneity within the Job Corps program (JC) in US using the National Job Corps Study (NJCS) data.
- After enrolling in JC (i.e., S=1), participants are provided a residential choice based on their preferences. Enrollees can choose to reside in the training center or to live at home and commute to the training center every day (i.e., D=0 or 1).
- We use the prediction of residence choice as IV for the self-selected residential component, and include high-dimensional controls following Schochet and Burghardt (2007).
- About 13 percent of participants chose to be nonresidential and resided at home (Schochet et al., 2008).

Application: Heterogeneous Effects by Age

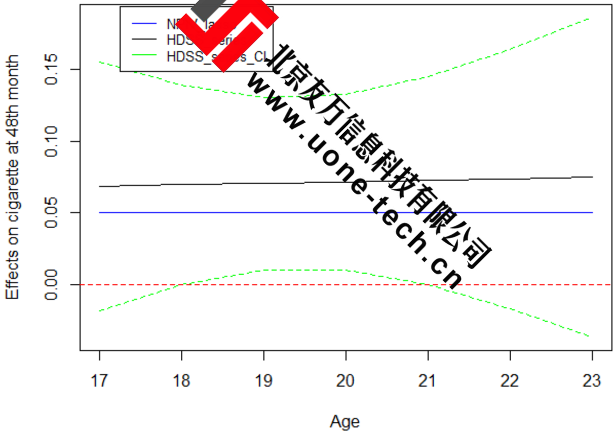
We investigate the ACRF with continuous covariate X_1, i.e., age, of earning at 16th quarter and cigarette occurrence at 48th month after randomization.

Figure: ACRF on Earnings at 16th Quarter by Age



Application: Heterogeneous Effects by Age

Figure: ACRF of Cigarette Occurrence at 48th Month by Age



Application: Heterogeneous Effects by Gender

We investigate the ACRF with binary covariate X_1, i.e., gender, of earning at 16th quarter and cigarette occurrence at 48th month after randomization.

Table: ACRF Estimates by Gender

Gender	Earnq16	Cigarm48
Female	-19.1	0.080**
	[-51.6,13.5]	[0.007,0.153]
Male	-18.4	0.059*
	[-61.6,24.7]	[-0.036,0.154]
ACR	-18.7 (14.5)	0.068** (0.032)

Notes: The point-wise 95% confidence intervals are in brackets and standard errors in parentheses for ACR. $*** = p < 0.01$; $** = p < 0.05$; $* = p < 0.10$.

Application: Heterogeneous Effects by Ethnicity

We also investigate the ACRF with discrete covariate X_1, i.e., ethnicity, of earning at 16th quarter and cigarette occurrence at 48th month after randomization.

Table: ACRF Estimates by Ethnicity

Ethnicity	Earnq16	Cigarm48
White	-21.7	0.041
	[21.9,-65.3]	[0.156,-0.074]
Black	-23.5^{*}	0.070^{**}
	[2.6,-49.6]	[0.129,0.106]
Hispanic	-15.3	0.099^{*}
	[29.5,-60.1]	[0.201,-0.003]

Notes: The point-wise 95% confidence intervals are reported in the brackets. Standard errors are reported in parentheses for ACR. $*** = p < 0.01$; $** = p < 0.05$; $* = p < 0.10$.

Application: Residential Component in Job Corps Program

- The residential component has a negative but insignificant effects on the earnings; however, a positive (detrimental) and significant effect on the risky behavior outcome.
- The significant detrimental effect on the the cigarette occurrence varies by age, gender and ethnicity. Younger female group and Black youth are more vulnerable to this detrimental effect.
- Overall, the ACR of residential component is negligible on earnings but significant and detrimental on risky behavior outcomes such as cigarette occurrence.

Table of Contents

Partial identification when D affects sample selection

D is also allowed to enter S, which yields a partial identi result:

$$S = S(X_1, X_2, D, \varepsilon) = S_e(X_1, X_2, X_3, V, \varepsilon).$$

Theorem 5

If the selection equation in Eq.(4) is replaced by $S = S(X_1, X_2, D, \varepsilon)$, Assumption 1 and 3 hold, $\mathcal{Y} = supp(Y^*)$ is bounded with $y_{min} = \min_{y \in \mathcal{Y}} y$ and $y_{max} = \max_{y \in \mathcal{Y}} y$, and $p(X_1, X_2, X_3) = P(S = 1|X_1, X_2, X_3) \geq C_e > 0$ with probability one, then $\theta(x_1)$ is partially identified by outer region $[\theta_{LB}(x_1), \theta_{UB}(x_1)] = [\theta_m(x_1) - v(x_1), \theta_m(x_1) + v(x_1)]$ with

$$\theta_m(x_1) = \frac{E\left[\left(\mu(X_1, X_2, X_3) - \widetilde{\mu}(X_1, X_2)\right)\left(Y - \widetilde{\mu}(X_1, X_2)\right)\Big| X_1 = x_1, S = 1\right]}{E\left[\left(\mu(X_1, X_2, X_3) - \widetilde{\mu}(X_1, X_2)\right)\left(D - \widetilde{\mu}(X_1, X_2)\right)\Big| X_1 = x_1, S = 1\right]},$$

$$v(x_1) = \frac{E\left[\left|\mu(X_1, X_2, X_3) - \widetilde{\mu}(X_1, X_2)\right| \cdot BD(X_1, X_2, X_3)\Big| X_1 = x_1, S = 1\right]}{E\left[\left(\mu(X_1, X_2, X_3) - \widetilde{\mu}(X_1, X_2)\right)\left(D - \widetilde{\mu}(X_1, X_2)\right)\Big| X_1 = x_1, S = 1\right]},$$

$$BD(X_1, X_2, X_3) = \min\left\{1, \frac{1 - \widetilde{p}(X_1, X_2)p(X_1, X_2, X_3)}{\widetilde{p}(X_1, X_2)p(X_1, X_2, X_3)}\right\} \cdot (y_{max} - y_{min}),$$

where $\widetilde{p}(X_1, X_2) = P(S = 1|X_1, X_2)$.

Conclusion

- This paper identifies and estimates a semiparametric ACR, first proposed by Angrist and Imbens (1995) and Abadie(2003), with sample selection in a high-dimensional covariate environment.
- The proposed ACRF is shown to be consistent and asymptotically normal. Monte Carlo simulations demonstrate that ACRF performs better than the existing IV estimators (such as NPIV-lasso).
- The empirical study evaluates the heterogeneous effect of the residential component in US Job Corps program with proposed ACRF and ACR, and yields new insights with a large set of controls.
- We also relax the selection-on-observables assump on selection process, and derive bounds on the proposed ACRF with one single IV with selection-on-unobservables (i.e., D affects the selection process).

Thank You!

yahong.zhou@mail.shufe.edu.cn