Model Assisted Inference for Conditional Average Treatment Effects

Adam Baybutt Manu Navjeevan

UCLA

May 11, 2022

Introduction

Motivation: Identification of the conditional average treatment effect (CATE) often relies on conditioning on a high dimensional set of control variables. Limited non-parametric estimation results in high dimensional settings, so nuisance parameter consistency in high dimensional settings typically relies on the functional form being correctly specified.

Research Goal: Develop nonparametric estimators and inference procedures for the CATE in a high dimensional setting that remain valid even if one of the nuisance parameters is misspecified.

Table of Contents

Setting

Assume a potential outcomes framework where the researcher observes i.i.d data consisting of:

- Realized outcome $Y \in \mathbb{R}$ where:

$$Y = DY_1 + (1 - D)Y_0.$$

- Treatment decision $D \in \{0, 1\}$
- Control variables $Z = (X, Z_1) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_z - d_x}$ such that $Y_1, Y_0 \perp D | Z$
  - Asymptotic analysis will allow $d_z \gg n$ but d_x is fixed.

The target parameter of interest is:

$$g(x) = \mathbb{E}[Y_1 | X = x].$$

Setting

To estimate $g(x)$ use the aIPW signal:

$$Y(\pi_0, m_0) = \frac{DY}{\pi_0(Z)} + \left(\frac{D}{\pi_0(Z)} - 1\right) m_0(Z).$$

where $\pi_0(Z)$ and $m_0(Z)$ are nuisance parameters that need to be estimated.

- Propensity Score: $\pi_0(Z) = \Pr(D = 1|Z)$
- Mean Regression: $m_0(Z) = \mathbb{E}[Y|D = 1, Z]$

Under Identifying Assumption:

$$g(x) = \mathbb{E}[Y_1|X = x] = \mathbb{E}[Y(\pi_0, m_0)|X = x].$$

Setting

The aIPW signal $Y(\pi, m)$ has two properties of interest:

1. Neyman Orthogonality:

$$\mathbb{E}[\partial_\pi Y(\pi_0, m_0)|Z] = \mathbb{E}[\partial_m Y(\pi_0, m_0)|Z] = 0.$$

   - Allows us estimate $g(x)$ at the nonparametric rate even when $\hat{\pi}$ and $\hat{m}$ converge slowly to π_0 and m_0 (Semenova and Chernozhukov, 2020).

Setting

The aIPW signal $Y(\pi, m)$ has two properties of interest:

1. Neyman Orthogonality:

$$\mathbb{E}[\partial_\pi Y(\pi_0, m_0)|Z] = \mathbb{E}[\partial_m Y(\pi_0, m_0)|Z] = 0.$$

2. Double Robustness: So long one of either $m = m_0$ or $\pi = \pi_0$,

$$\begin{aligned} g(x) &= \mathbb{E}[Y_1|X = x] \\ &= \mathbb{E}[Y(\pi, m)|X = x] \end{aligned}$$

   - Allows us to obtain a consistent estimator for $g(x)$ even when $\hat{\pi} \to_p \bar{\pi} \neq \pi_0$ or $\hat{m} \to_p \bar{m} \neq m_0$ (but not both).

Setting

However, Neyman Orthogonality does not tell us anything about the derivatives away from π_0 and m_0. In particular for $\bar{\pi} \neq \pi_0$ and $\bar{m} \neq m_0$ we may have:

$$\mathbb{E}[\partial_\pi Y(\bar{\pi}, \bar{m})|X] \neq 0 \quad \text{or} \quad \mathbb{E}[\partial_m Y(\bar{\pi}, \bar{m})|X] \neq 0.$$

In general, this prevents us from estimating $g(x)$ at the non-parametric rate when either $\hat{\pi} \to \bar{\pi} \neq \pi_0$ or $\hat{m} \to \bar{m} \neq m_0$.

Setting

However, Neyman Orthogonality does not tell us anything about the derivatives away from π_0 and m_0. In particular for $\bar{\pi} \neq \pi_0$ and $\bar{m} \neq m_0$ we may have:

$$\mathbb{E}[\partial_\pi Y(\bar{\pi}, \bar{m})|X] \neq 0 \quad \text{or} \quad \mathbb{E}[\partial_m Y(\bar{\pi}, \bar{m})|X] \neq 0.$$

In general, this prevents us from estimating $g(x)$ at the non-parametric rate when either $\hat{\pi} \to \bar{\pi} \neq \pi_0$ or $\hat{m} \to \bar{m} \neq m_0$.

- In a high dimensional setting consistency of $\hat{\pi}$ to π_0 or $\hat{m}$ to m_0 depends on correctly specifying their functional forms. In general, this prevents conducting valid inference on $g(x)$ under misspecification.

Setting

However, Neyman Orthogonality does not tell us anything about the derivatives away from π_0 and m_0. In particular for $\bar{\pi} \neq \pi_0$ and $\bar{m} \neq m_0$ we may have:

$$\mathbb{E}[\partial_\pi Y(\bar{\pi}, \bar{m})|X] \neq 0 \quad \text{or} \quad \mathbb{E}[\partial_m Y(\bar{\pi}, \bar{m})|X] \neq 0.$$

In general, this prevents us from estimating $g(x)$ at the non-parametric rate when either $\hat{\pi} \to \bar{\pi} \neq \pi_0$ or $\hat{m} \to \bar{m} \neq m_0$.

- In a high dimensional setting consistency of $\hat{\pi}$ to π_0 or $\hat{m}$ to m_0 depends on correctly specifying their functional forms. In general, this prevents conducting valid inference on $g(x)$ under misspecification.
- However, by carefully designing the estimating equations for $\hat{\pi}$ and $\hat{m}$, we can regain partial control of the derivatives of $Y(\cdot, \cdot)$ away from π_0 and m_0. This allows us to estimate $g(x)$ at the non-parametric rate and conduct inference.

Setting: Average Treatment Effects

When estimating the average treatment effect (as opposed to the CATE), Tan (2020) shows how to set up the estimating equations for the propensity score and the mean regression to gain control of the derivative of $Y(\cdot, \cdot)$ at $\bar{\pi}$ and $\bar{m}$, the probability limits of $\hat{\pi}$ and $\hat{m}$, respectively.

- Assumes a logistic regression form for the propensity score and a linear form for the mean regression, which we will follow.
  - Results are valid even if one of the functional forms is misspecified
- Logistic regression and mean regression paramaeters are estimated using an ℓ_1 (LASSO) penalty under sparsity.
- We will modify the estimating equations of Tan to allow for nonparametric estimation and inference.

Contribution

Project Contribution: Modify Tan (2020) estimating equations for $\hat{\pi}$ and $\hat{m}$ so that the CATE, $g(x) = \mathbb{E}[Y_1 \mid X = x]$, can be estimated at the non-parametric rate even when the functional form for either $\hat{\pi}$ or $\hat{m}$ is misspecified.

Contribution

Project Contribution: Modify Tan (2020) estimating equations for $\hat{\pi}$ and $\hat{m}$ so that the CATE, $g(x) = \mathbb{E}[Y_1 \mid X = x]$, can be estimated at the non-parametric rate even when the functional form for either $\hat{\pi}$ or $\hat{m}$ is misspecified.

- We will assume a logistic regression form for the propensity score and a linear regression form for the mean regression. These will be estimated using an ℓ_1-penalty to deal with $d_z \gg n$.
  - Control on the derivatives of $Y(\bar{\pi}, \bar{m})$ using our estimating equations depends on these specific functional forms, which is why this is called model assisted or model specific inference.
  - If one wanted to control using other functional forms, would have to set up estimating equations differently.
  - Improving on past work on model assisted inference, will use a data-dependent method to pick the ℓ_1 penalty parameter

Contribution

Project Contribution: Modify Tan (2020) estimating equations for $\hat{\pi}$ and $\hat{m}$ so that the CATE, $g(x) = \mathbb{E}[Y_1 \mid X = x]$, can be estimated at the non-parametric rate even when the functional form for either $\hat{\pi}$ or $\hat{m}$ is misspecified.

- After estimating the nuisance parameters $\hat{\pi}$ and $\hat{m}$, we will take a series regression approach to obtain a non-parametric estimator $\hat{g}(x)$ for $g(x)$.
- This estimator will converge to $g(x)$ at the non-parametric rate and we will be able to use it to conduct valid inference on $g(x)$.

Prior Literature

- Double Machine Learning/LASSO: Belloni et. al (2012), Belloni, Chernozhukov, Hansen (2014), CCDDHNR (2018)
- High Dimensional M-Estimation Tan (2017), Chetverikov & Sørenson (2021)
- Model Assisted Inference: Smucler et. al (2019), Tan (2020), Wu et. al (2021)
- Nonparametric Estimation: Newey (1997), Belloni et. al (2015), Semenova and Chernozhukov (2020)

Table of Contents

Estimation Procedure: Overview

The estimation procedure proceeds in the following steps:

1. Select a set of k nonnegative basis functions of X to use in the series approximation of $g(x)$, $p^k(x) = (p_1(x), \ldots, p_k(x))' \in \mathbb{R}^k$.

Estimation Procedure: Overview

The estimation procedure proceeds in the following steps:

1. Select a set of k nonnegative basis functions of X to use in the series approximation of $g(x)$, $p^k(x) = (p_1(x), \ldots, p_k(x))' \in \mathbb{R}^k$.

2. For each basis function, $p_j(x)$, estimate $\hat{\pi}_j$ and $\hat{m}_j$ using estimating equations described below.

Estimation Procedure: Overview

The estimation procedure proceeds in the following steps:

1. Select a set of k nonnegative basis functions of X to use in the series approximation of $g(x)$, $p^k(x) = (p_1(x), \ldots, p_k(x))' \in \mathbb{R}^k$.

2. For each basis function, $p_j(x)$, estimate $\hat{\pi}_j$ and $\hat{m}_j$ using estimating equations described below.

3. Calculate $\hat{\beta}^k$ via

$$\hat{\beta}^k = \mathbb{E}_n \left[p^k(x) p^k(x)' \right]^{-1} \mathbb{E}_n \begin{bmatrix} p_1(x) Y(\hat{\pi}_1, \hat{m}_1) \\ \vdots \\ p_k(x) Y(\hat{\pi}_k, \hat{m}_k) \end{bmatrix}$$

and let $\hat{g}(x) = p^k(x)' \hat{\beta}^k$. Let $k \to \infty$ with n to achieve a consistent estimator of $g(x)$.

Estimation Procedure: Estimating Equations

Assume a logistic regression form for the propensity score and a linear regression form for the mean regression:

$$\pi(z;\gamma) = \frac{1}{1+\exp(-\gamma' z)} \quad \text{and} \quad m(z;\alpha) = \alpha' z$$

For each basis function $p_j(x)$ estimate $\hat{\gamma}_j$ and $\hat{\alpha}_j$ via:

$$\hat{\gamma}_j = \arg\min_{\gamma} \mathbb{E}_n[p_j(X)\{De^{-\gamma' Z} + (1-D)\gamma' Z\}] + \lambda_{j,\gamma}\|\gamma\|_1$$

$$\hat{\alpha}_k = \arg\min_{\alpha} \frac{1}{2}\mathbb{E}_n[p_j(X)De^{-\hat{\gamma}_j' Z}\{Y - \alpha' Z\}^2] + \lambda_{j,\alpha}\|\alpha\|_1$$

and let $\hat{\pi}_j(z) = \pi(z;\hat{\gamma}_j)$, $\hat{m}_j(z) = \pi(z;\hat{\alpha}_j)$.

Estimation Procedure: Estimating Equations

For each basis function $p_j(x)$ estimate $\hat{\gamma}_j$ and $\hat{\alpha}_j$ via:

$$\hat{\gamma}_j = \arg\min_{\gamma} \mathbb{E}_n[p_j(X)\{De^{-\gamma' Z} + (1-D)\gamma' Z\}] + \lambda_{j,\gamma}\|\gamma\|_1 \tag{1}$$

$$\hat{\alpha}_k = \arg\min_{\alpha} \frac{1}{2}\mathbb{E}_n[p_j(X)De^{-\hat{\gamma}_j' Z}\{Y - \alpha' Z\}^2] + \lambda_{j,\alpha}\|\alpha\|_1 \tag{2}$$

and let $\hat{\pi}_j(z) = \pi(z;\hat{\gamma}_j)$, $\hat{m}_j(z) = \hat{\alpha}_j' z$.

- These are the Tan (2020) estimating equations, but weighted by the polynomial basis function $p_j(x)$.
- Improving on Tan (2020), will use a data driven method to select $\lambda_{j,\gamma}, \lambda_{j,\alpha}$

Selection Intuition | Selection Procedure | Procedure Validity

Estimation Procedure: Estimating Equations

Under standard assumptions, the estimators $\hat{\gamma}_j$ and $\hat{\alpha}$ will converge uniformly over $j \in \{1, \ldots, k\}$ to:

$$\bar{\gamma}_j := \arg\min_{\gamma} \mathbb{E}[p_j(X)\{De^{-\gamma' Z} + (1-D)\gamma' Z\}]$$

$$\bar{\alpha}_j := \arg\min_{\alpha} \frac{1}{2}\mathbb{E}[p_j(X)De^{-\bar{\gamma}' Z}\{Y - \alpha' Z\}^2]$$

For analysis we will let $\bar{\pi}_j(z) = \pi(z; \bar{\gamma})$ and $\bar{m}_j(z) = \bar{\alpha}_j' z$.

- So long as the functional forms of either the propensity score or the mean regression are correctly specified, $\bar{\pi}_j = \pi_0$ and $\bar{m}_j = m_0$

Estimation Intuition: Why k estimating procedures?

We now provide some intuition for the estimating procedure. Consider the following first order expansion:

$$\mathbb{E}_n\left[p_j(X)Y(\hat{\pi},\hat{m})\right] - \mathbb{E}_n\left[p_j(X)Y(\pi_0,m_0)\right]$$

$$= \mathbb{E}_n\left[p_j(X)\nabla_{\pi,m}Y(\pi_0,m_0)\right]\begin{bmatrix}\hat{\pi}-\pi_0\\ \hat{m}-m_0\end{bmatrix} + o_p(n^{-1/2}).$$

Estimation Intuition: Why k estimating procedures?

We now provide some intuition for the estimating procedure. Consider the following first order expansion:

$$\mathbb{E}_n\left[p_j(X)Y(\hat{\pi},\hat{m})\right] - \mathbb{E}_n\left[p_j(X)Y(\pi_0,m_0)\right]$$
$$= \mathbb{E}_n\left[p_j(X)\nabla_{\pi,m}Y(\pi_0,m_0)\right]\begin{bmatrix}\hat{\pi}-\pi_0\\ \hat{m}-m_0\end{bmatrix} + o_p(n^{-1/2}).$$

- Under conditional unconfoundedness $\mathbb{E}_n[p_j(X)Y(\pi_0,m_0)] \approx \mathbb{E}_n[p_j(X)Y_1]$.
- From Neyman Orthogonality of the signal $Y(\cdot,\cdot)$ right hand side will go to zero "fast enough" even if the nuisance parameters converge slowly to π_0 and m_0
  - The gradient being mean zero conditional on Z means we could use a single propensity score and outcome regression estimation

Estimation Intuition: Why k estimating procedures?

However, if the functional form of either π and m are misspecified, our estimators will not converge to π_0 and m_0. Instead, we should consider the Taylor expansion around their probability limits.

$$\mathbb{E}_n\left[p_j(X)Y(\hat{\pi}_j, \hat{m}_j)\right] - \mathbb{E}_n\left[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)\right]$$

$$= \mathbb{E}_n\left[p_j(X)\nabla_{\pi,m}Y(\bar{\pi}_j, \bar{m}_j)\right] \begin{bmatrix} \hat{\pi}_j - \bar{\pi}_j \\ \hat{m}_j - \bar{m}_j \end{bmatrix} + o_p(n^{-1/2}).$$

Estimation Intuition: Why k estimating procedures?

However, if the functional form of either π and m are misspecified, our estimators will not converge to π_0 and m_0. Instead, we should consider the Taylor expansion around their probability limits.

$$\mathbb{E}_n\left[p_j(X)Y(\hat{\pi}_j, \hat{m}_j)\right] - \mathbb{E}_n\left[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)\right]$$
$$= \mathbb{E}_n\left[p_j(X)\nabla_{\pi,m}Y(\bar{\pi}_j, \bar{m}_j)\right]\begin{bmatrix}\hat{\pi}_j - \bar{\pi}_j \\ \hat{m}_j - \bar{m}_j\end{bmatrix} + o_p(n^{-1/2}).$$

- So long as either $\bar{\pi} = \pi_0$ or $\bar{m} = m_0$, we have by Double Robustness:

$$\mathbb{E}\left[Y(\bar{\pi}, \bar{m}) \mid X\right] = \mathbb{E}\left[Y(\pi_0, m_0) \mid X\right] = \mathbb{E}\left[Y_1 \mid X\right].$$

So, $\mathbb{E}_n\left[p_j(X)Y(\bar{\pi}, \bar{m})\right] \approx \mathbb{E}_n\left[p_j(X)Y_1\right]$.

Estimation Intuition: Why k estimating procedures?

However, if the functional form of either π and m are misspecified, our estimators will not converge to π_0 and m_0. Instead, we should consider the Taylor expansion around their probability limits.

$$\mathbb{E}_n\left[p_j(X)Y(\hat{\pi}_j, \hat{m}_j)\right] - \mathbb{E}_n\left[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)\right]$$
$$= \mathbb{E}_n\left[p_j(X)\nabla_{\pi,m}Y(\bar{\pi}_j, \bar{m}_j)\right]\begin{bmatrix}\hat{\pi}_j - \bar{\pi}_j \\ \hat{m}_j - \bar{m}_j\end{bmatrix} + o_p(n^{-1/2}).$$

- In general Neyman Orthogonality does not give us:

$$\mathbb{E}\left[\nabla_{\pi,m}Y(\bar{\pi}, \bar{m}) \mid Z\right] = 0.$$

So we cannot use this to say $\mathbb{E}_n\left[p_j(X)\nabla_{\pi,m}Y(\bar{\pi}, \bar{m})\right] \approx 0$.

Estimation Intuition: Why k estimating procedures?

Instead, we will directly use the first order conditions of the estimating equations to control terms in the Taylor expansion. Since we have assumed specific parametric forms consider expansions around the parameters $\bar{\gamma}_j$ and $\bar{\alpha}_j$

$$\mathbb{E}_n\left[p_j(X)Y(\hat{\pi}_j, \hat{m}_j)\right] - \mathbb{E}_n\left[p_j(X)Y(\bar{\pi}_j, \bar{m}_j)\right]$$
$$= \mathbb{E}_n\left[p_j(X)\nabla_{\gamma_j,\alpha_j}Y(\bar{\pi}_j, \bar{m}_j)\right]\begin{bmatrix}\hat{\gamma}_j - \bar{\gamma}_j \\ \hat{\alpha}_j - \bar{\alpha}_j\end{bmatrix} + o_p(n^{-1/2}).$$

- Instead of controlling the derivative when we approach $\bar{\pi}_j$ and $\bar{m}_j$ from any direction, we will only control the derivative in certain directions. Hence model assisted or model specific inference

Estimation Intuition: Why k estimating procedures?

After substituting in the forms of $\bar{\pi}(z) = \pi(z; \bar{\gamma})$ and $\bar{m}(z) = \bar{\alpha}' z$ we obtain:

$$Y(\bar{\pi}_j, \bar{m}_j) = DY(1 + e^{-\bar{\gamma}_j' Z}) - D(1 + e^{-\bar{\gamma}_j' Z})\bar{\alpha}' Z + \bar{\alpha}' Z$$

So that

$$\mathbb{E}[p_j(X)\nabla_{\gamma_j, \alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] = \mathbb{E}\begin{bmatrix} -p_j(X) D e^{-\bar{\gamma}' Z}(Y - \bar{\alpha}' Z) Z \\ p_j(X)\{D(1 + e^{\bar{\gamma}' Z}) Z + Z\} \end{bmatrix}.$$

Estimation Intuition: Why k estimating procedures?

Looking at the first order conditions of our estimating equations

$$\bar{\gamma}_j := \arg\min_{\gamma} \mathbb{E}[p_j(X)\{De^{-\gamma' Z} + (1-D)\gamma' Z\}]$$

$$\bar{\alpha}_j := \arg\min_{\alpha} \frac{1}{2}\mathbb{E}[p_j(X)De^{-\bar{\gamma}' Z}\{Y - \alpha' Z\}^2]$$

We can see that:

$$\mathbb{E}\overbrace{\underbrace{\begin{bmatrix} p_j(X)\{D(1+e^{\bar{\gamma}' Z})Z + Z\} \\ p_j(X)De^{-\bar{\gamma}' Z}(DY - \bar{\alpha}' Z)Z \end{bmatrix}}_{\text{First order condition of } \bar{\alpha}_j}}^{\text{First order condition of } \bar{\gamma}_j} = 0.$$

Estimation Intuition: Why k estimating procedures?

The first order conditions of our estimating equations give exactly control over the gradient in the Taylor expansion:

$$\mathbb{E}[p_j(X)\nabla_{\gamma_j,\alpha_j}Y(\bar{\pi}_j,\bar{m}_j)] = \mathbb{E}\begin{bmatrix} -p_j(X)De^{-\bar{\gamma}'Z}(Y-\bar{\alpha}'Z)Z \\ p_j(X)\{D(1+e^{\bar{\gamma}'Z})Z+Z\} \end{bmatrix}$$

But,

$$\mathbb{E}\overbrace{\underbrace{\begin{bmatrix} p_j(X)\{D(1+e^{\bar{\gamma}'Z})Z+Z\} \\ p_j(X)De^{-\bar{\gamma}'Z}(Y-\bar{\alpha}'Z)Z \end{bmatrix}}_{\text{First order condition of } \bar{\alpha}_j}}^{\text{First order condition of } \bar{\gamma}_j} = 0.$$

This gives $\mathbb{E}_n\left[p_j(X)\nabla_{\gamma_j,\alpha_j}Y(\bar{\pi}_j,\bar{m}_j)\right] \approx 0.$

Estimation Intuition: Why k estimating procedures?

This control is not obtained by using other estimating equations. Consider taking a standard approach to estimate the logistic and mean regression parameters:

$$\bar{\gamma}^{\text{MLE}} = \arg\min_{\gamma} \mathbb{E}[\log(1+\exp(\gamma' Z)) - D\gamma' Z]$$

$$\bar{\alpha}^{\text{OLS}} = \arg\min_{\alpha} \frac{1}{2}\mathbb{E}[(Y - \alpha_j' Z)^2]$$

This gives first order conditions:

$$\mathbb{E}\overbrace{\left[\begin{array}{c} \frac{1}{1+\exp(\bar{\gamma}^{\text{MLE}\prime} Z)} Z - DZ \\ (Y - \bar{\alpha}' Z)Z \end{array}\right]}^{\substack{\text{First order condition of } \bar{\gamma}^{\text{MLE}} \\ \text{First order condition of } \bar{\alpha}^{\text{OLS}}}} = 0$$

Estimation Intuition: Why k estimating procedures?

The MLE logit and OLS first order conditions do not directly provide control over the gradient:

$$\mathbb{E}[p_j(X)\nabla_{\gamma,\alpha}Y(\bar{\pi}^{\text{MLE}},\bar{m}^{\text{OLS}})] = \mathbb{E}\begin{bmatrix} -p_j(X)De^{-\bar{\gamma}^{\text{MLE}\prime}Z}(Y-\bar{\alpha}^{\text{OLS}\prime}Z)Z \\ p_j(X)\{D(1+e^{\bar{\gamma}^{\text{OLS}\prime}Z})Z+Z\} \end{bmatrix}$$

$$\neq \mathbb{E}\overbrace{\underbrace{\begin{bmatrix} \frac{1}{1+\exp(\bar{\gamma}^{\text{MLE}\prime}Z)}Z - DZ \\ (Y-\bar{\alpha}'Z)Z \end{bmatrix}}_{\text{First order condition of } \bar{\alpha}^{\text{OLS}}}}^{\text{First order condition of } \bar{\gamma}^{\text{MLE}}} = 0$$

Estimation Intuition: Why k estimating procedures?

Moreover, control over the gradient $\mathbb{E}[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)]$ is not provided by the first order conditions for $\bar{\gamma}_s, \bar{\alpha}_s$, $s \neq j$.

$$\mathbb{E}[p_j(X)\nabla_{\gamma_j,\alpha_j} Y(\bar{\pi}_j, \bar{m}_j)] = \mathbb{E}\begin{bmatrix} -p_j(X)De^{-\bar{\gamma}'Z}(Y - \bar{\alpha}'Z)Z \\ p_j(X)\{D(1+e^{\bar{\gamma}'Z})Z + Z\} \end{bmatrix}$$

$$\neq \underbrace{\overbrace{\mathbb{E}\begin{bmatrix} p_s(X)\{D(1+e^{\bar{\gamma}'Z})Z + Z\} \\ p_s(X)De^{-\bar{\gamma}'Z}(Y - \bar{\alpha}'Z)Z \end{bmatrix}}^{\text{First order condition of } \bar{\gamma}_s}}_{\text{First order condition of } \bar{\alpha}_s} = 0.$$

Estimation Intuition: Why k estimating procedures?

Want to show a fast rate of convergence for any linear transformation of the vector

$$\mathbb{E}_n \begin{bmatrix} p_1(X)Y(\hat{\pi}_1, \hat{m}_1) \\ \vdots \\ p_k(X)Y(\hat{\pi}_k, \hat{m}_k) \end{bmatrix}.$$

so need to estimate $(\hat{\pi}_1, \hat{m}_1), \ldots, (\hat{\pi}_k, \hat{m}_k)$ via separate estimating equations.

- Each pair of estimating equations for $\hat{\gamma}_j$ and $\hat{\alpha}_j$ only allows us to control the gradient for one term in this vector
- Need k pairs of estimating equations to control the gradient of the whole vector.
- Each estimating equation varies only in the polynomial weight, $p_j(X)$.

Table of Contents

Results: Pointwise Normality

Now present the main result of the paper. We will first present an overall result that will rely on some first stage and second stage building blocks that we will cover later on.

Focus on pointwise results for now, but uniform inference results are also developed and are in a draft paper.

Results: Pointwise Normality

Proposition 1 (Pointwise Normality)

Suppose that the conditions of Theorems 1-4 hold. Then so long as either the logistic propensity score model or linear outcome regression model is correctly specified we have for any $\alpha \in S^{k-1}$ *there is a variance estimator* $\widehat{\Omega}$ *such that,*

$$\sqrt{n}\frac{\alpha'(\widehat{\beta}^k - \beta^k)}{\|\alpha'\widehat{\Omega}^{1/2}\|} \to_d N(0,1). \tag{3}$$

if additionally approximation error is negligible relative to the estimation error then for $\widehat{s}(x) := \widehat{\Omega}^{1/2}p^k(x)$*,*

$$\sqrt{n}\frac{\widehat{g}(x) - g(x)}{\|\widehat{s}(x)\|} \to_d N(0,1). \tag{4}$$

Results: First Stage Notation

First stage results will be given in terms of the sparsity indices as well as in terms of the growth rate of the basis functions. It will be helpful to make some notations here.

Denote the sparsity sets by:

$$\mathcal{S}_{\gamma,j} \equiv \{l : \gamma_{j,l} \neq 0\}$$

$$\mathcal{S}_{\alpha,j} \equiv \{l : \alpha_{j,l} \neq 0\}$$

and let:

$$s_k \equiv \max_{1 \leq j \leq k} \left\{ |\mathcal{S}_{\gamma,j}| + |\mathcal{S}_{\alpha,j}| \right\}$$

$$\xi_{k,\infty} \equiv 1 \vee \sup_{x \in \text{supp}(X)} \|p^k(X)\|_\infty$$

Results: First Stage Assumptions

Assumption 1 (First Stage Convergence)

1. $\max_{0\leq j\leq p}|Z_j|\leq C_0$ almost surely for some constant $C_0>0$.
2. $\bar{\pi}_j(Z)\geq B_0$ almost surely for a constant $B_0\in\mathbb{R}$ for all j.
3. The errors $Y_1-\bar{m}_j(X)$ are uniformly subgaussian conditional on Z and over all j.

   Uniformly Subgaussian
4. An empirical compatibility condition holds uniformly over all j for the matrices $\Sigma_{\gamma,j}=\mathbb{E}_n[p_j(X)De^{-\bar{\gamma}_j'Z}ZZ']$ and the sparsity sets $S_{\gamma,j}$ and $S_{\alpha,j}$.

   Empirical Compatibility Condition
5. There exists a constant c_u such that, for all j,

$$\min_{1\leq l\leq d_z}\mathbb{E}[\bar{U}_{\gamma,j}^2Z_l^2]\geq c_u \quad\text{and}\quad \min_{1\leq l\leq d_z}\mathbb{E}[\bar{U}_{\alpha,j}^2Z_l^2]\geq c_u.$$

6. The following sparsity bounds hold

$$\frac{\xi_{k,\infty}}{\ln(d_zn)}\to 0,\ \frac{\xi_{k,\infty}s_k^2\ln^5(d_zn)}{n}\to 0,\quad\text{and}\quad \frac{\xi_{k,\infty}^4\ln^7(d_zkn)}{n}\to 0.$$

Results: First Stage Convergence

Theorem 1 (No Effect of First Stage Bias)

Suppose that Assumption 1 holds and the penalty parameters are chosen as in (5)-(6). *If, in addition* $s_k \xi_{k,\infty}^2 k^{1/2} \ln(d_z)/n^{1/2} \to 0$ *we have that*

$$\sup_{1 \le j \le k} \left| \mathbb{E}_n[p_j(X) Y(\hat{\pi}_j, \hat{m}_j)] - \mathbb{E}_n[p_j(X) Y(\bar{\pi}_j, \bar{m}_j)] \right| = o_p(n^{-1/2} k^{-1/2})$$

Proof Idea.

Roughly follows the discussion above. First order conditions and ℓ_1 norm bounds give control of first order bias. Keep track of constants and union bound over $j \in \{1, \ldots, k\}$. □

Results: Second Stage Setup

For each $j = 1, \ldots, k$, so long as either the functional form of $\pi(\cdot)$ or $m(\cdot)$ is correctly specified:

$$\begin{aligned} g(x) &= \mathbb{E}[Y_1 \mid X = x] \\ &= \mathbb{E}[Y(\bar{\pi}_j, \bar{m}_j) \mid X = x] \end{aligned}$$

so we can write $\forall j = 1, \ldots, k$

$$\begin{aligned} Y(\bar{\pi}_j, \bar{m}_j) &= g(X) + \epsilon_j \\ &= g_k(X) + r_k + \epsilon_j. \end{aligned}$$

where $\mathbb{E}[\epsilon_j \mid X = x] = 0$ for all $j = 1, \ldots, k$, $g_k(x)$ denotes the L^2-projection of $g(x)$ onto the linear subspace spanned by $p^k(x) = \big(p_1(x), \ldots, p_k(x)\big)$, r_k is the approximation error, $r_k = g(x) - g_k(x)$.

Results: Second Stage Matrices

Depending on the rate of decay of approximation error, the asymptotic variance of our second stage estimator is governed by one of the following matrices

$$\tilde{\Omega} := Q^{-1}\mathbb{E}[\{p^k(X) \circ (\epsilon^k + r_k)\}\{p^k(X) \circ (\epsilon^k + r_k)\}']Q^{-1}$$
$$\Omega_0 := Q^{-1}\mathbb{E}[\{p^k(X) \circ \epsilon^k\}\{p^k(X) \circ \epsilon^k\}']Q^{-1}$$

where $Q = \mathbb{E}[p^k(X)p^k(x)']$ and $\circ$ represents the Hadamard element-wise product. These matrices will be estimated using the plug-in empirical analog

$$\widehat{\Omega} := \widehat{Q}^{-1}\mathbb{E}_n[\{p^k(X) \circ \widehat{\epsilon}^k\}\{p^k(X) \circ \widehat{\epsilon}^k\}']\widehat{Q}^{-1}$$

where $\widehat{Q}^{-1} = \mathbb{E}_n[p^k(X)p^k(X)']$.

Results: Second Stage Assumptions

Assumption 2 (Second-Stage Consistency Assumptions)

(i) Uniformly over all n, the eigenvalues of $Q = \mathbb{E}[p^k(x)p^k(x)']$ are bounded from above and away from zero.

(ii) The conditional variance of the error terms is uniformly bounded in the following sense. There exist constants $\underline{\sigma}^2$ and $\bar{\sigma}^2$ such that for any $j = 1, 2 \ldots$ we have that $\underline{\sigma}^2 \leq \mathrm{Var}(\epsilon_j \mid X) \leq \bar{\sigma}^2 < \infty$;

(iii) For each n and k there are finite constants c_k and ℓ_k such that for each $f \in \mathcal{G}$

$$\|r_k\|_{L,2} = \left(\mathbb{E}[r_k(x)^2]\right)^{1/2} \leq c_k \quad \text{and} \quad \|r_k\|_{L,\infty} = \sup_{x \in \mathcal{X}} |r_k(x)| \leq \ell_k c_k.$$

Results: Second Stage Assumptions

Assumption 3 (Uniform Integrability)

Let $\bar{\epsilon}_k := \max_{1 \leq j \leq k} |\epsilon_j|$. Assume that

(i) $\sup_{x \in \mathcal{X}} \mathbb{E}[\bar{\epsilon}_k^2 \, \mathbf{1}\{\bar{\epsilon}_k + \ell_k c_k > \delta \sqrt{n}/\xi_k\} \mid X = x] \to 0$ as $n \to \infty$ for any $\delta > 0$.

(ii) $\sup_{x \in \mathcal{X}} \mathbb{E}[\ell_k^2 c_k^2 \mathbf{1}\{\bar{\epsilon}_k + \ell_k c_k > \delta \sqrt{n}/\xi_k\} \mid X = x] \to 0$ as $n \to \infty$ for any $\delta > 0$.

Results: Second Stage Normality

Theorem 2 (Pointwise Normality)

Suppose that the conditions of Theorem 1 and Assumptions 2-3 hold. In addition suppose that $\xi_k^2 \log k/n \to 0$. *Then so long as either the logistic propensity score model or linear outcome regression model is correctly specified, for any* $\alpha \in S^{k-1}$:

$$\sqrt{n}\frac{\alpha'(\widehat{\beta}^k - \beta^k)}{\|\alpha'\Omega^{1/2}\|} \to_d N(0,1)$$

where generally $\Omega = \tilde{\Omega}$ *but if* $\ell_k c_k \to 0$ *then we can set* $\Omega = \Omega_0$. *Moreover, if the approximation error is negligible relative to the estimation error, namely if for* $s(x) := \Omega^{1/2} p^k(x)$, $\sqrt{n} r_k(x) = o(\|s(x)\|)$, *then* $\forall x \in \mathcal{X}$:

$$\sqrt{n}\frac{\widehat{g}(x) - g(x)}{\|s(x)\|} \to_d N(0,1)$$

Results: Variance Estimation

Last part is to show a consistent variance estimator. This will require a combination of second stage and first stage assumptions.

Results: Variance Estimation Assumption

Assumption 4 (Uniform Limit Theory)

Let $\bar{\epsilon}_k = \sup_{1\le j\le k} |\epsilon_j|$, $\alpha(x) := p^k(x)/\|p^k(x)\|$, and let

$$\xi_k^L := \sup_{\substack{x,x'\in\mathcal{X}\\ x\neq x'}} \frac{\|\alpha(x)-\alpha(x')\|}{\|x-x'\|}.$$

Further for any integer s let $\bar{\sigma}_k^s = \sup_{x\in\mathcal{X}} \mathbb{E}[|\bar{\epsilon}_k|^s | X=x]$. For some $m>2$ assume

(i) The regression errors satisfy $\sup_{x\in\mathcal{X}} \mathbb{E}[\max_{1\le i\le n} |\bar{\epsilon}_{k,i}|^m \mid X=x] \lesssim_P n^{1/m}$

(ii) The basis functions are such that (a) $\xi_k^{2m/(m-2)} \log k/n \lesssim 1$, (b) $(\bar{\sigma}_k^2 \vee \bar{\sigma}_k^m) \log \xi_k^L \lesssim \log k$, and (c) $\log \bar{\sigma}_k^m \xi_k \lesssim \log k$.

Results: First Stage Second Moment Convergence

Theorem 3 (Second Moment Convergence)

Suppose that the conditions of Theorem 1 hold and

$$\frac{\xi_{k,\infty}^5 s_k^2 k^2 \ln(d_z)}{n^{(m-1)/m}} \to 0$$

where $m > 2$ is as Assumption 4. Then the second moments of our estimator converge uniformly in empirical mean square at the following rate:

$$\xi_{k,\infty} \max_{1 \leq j \leq k} \mathbb{E}_n[p_j^2(X)(Y(\widehat{\pi}_j, \widehat{m}_j) - Y(\bar{\pi}_j, \bar{m}_j))^2] = o_p(k^{-2} n^{-1/m})$$

Results: Variance Estimation

Theorem 4 (Matrix Estimation)

Suppose that the conditions of Theorems 1-3 and Assumptions 1-4 hold. If in addition $\bar{R}_{1n} + \bar{R}_{2n} \lesssim (\log k)^{1/2}$ *for*

$$\bar{R}_{1n} := \sqrt{\frac{\xi_k^2 \log k}{n}}(n^{1/m}\sqrt{\log k} + \sqrt{k}\ell_k c_k)$$
$$\bar{R}_{2n} := \sqrt{\log k} \cdot \ell_k c_k$$

Then, so long as either the propensity score model or outcome regression model is correctly specified:

$$\|\widehat{\Omega} - \Omega\| \lesssim_P (v_n \vee \ell_k c_k)\sqrt{\frac{\xi_k^2 \log k}{n}} = o(1)$$

Table of Contents

Simulation Study: Setup

Set up simulations where

- $Z \in \mathbb{R}^{\text{dimZ}}$ is generated from a multivariate mean zero normal distribution with covariance matrix of the Toeplitz form

$$\text{Cov}(Z_i, Z_j) = \mathbb{E}[Z_i Z_j] = 2^{-|j-k|}.$$

- $X \sim \text{Unif}[1.5, 1]$ independently of Z
- $D \sim \text{Bernoulli}\left(\left\{1 + \exp(-X - \gamma' f(Z))\right\}^{-1}\right)$
- $\epsilon \sim N(0, 1)$ independently of D and Z
- $Y = 1 + D + X + \alpha' f(Z) + \epsilon$ so that $\mathbb{E}[Y_1|X] = 2 + X + \alpha' \mathbb{E}[f(Z)]$.

For correct specification $f(Z) = Z$, for mispecification set $f(Z) = \max\{Z, 0\}$ where the comparasion is made element wise.

Simulation Study: Preliminary Results

Following results are with 3 series terms, $n = 500$ and dimZ $= 100$

Regime	Estimating Equations	iMSE	95% CI	Pointwise SE	95% CI Uniform
Misspecified	MLE + OLS	6.152	0.59	0.985	0.89
Propensity	Model Assisted	12.987	0.65	0.989	0.98
Misspecified	MLE + OLS	7.463	0.28	1.067	0.89
Outcome Reg.	Model Assisted	4.634	0.63	0.9135	0.93
Neither	MLE + OLS	4.582	0.56	0.933	0.88
Misspecified	Model Assisted	11.593	0.69	0.952	0.97

Table of Contents

Conclusion

In this project we have

1. Shown convergence and asymptotic normality at the nonparametric rate of a novel CATE estimator under misspecification of one of the nuisance parameters.
2. Demonstrated a valid inference technique for the CATE based on our estimator that provides asymptotic point wise and uniform coverage even under nuisance parameter misspecification (model-assisted inference)

Remaining Work

The following are the main things remaining on the project

- Finish Simulations
  - Consider different basis functions, different sparsity and sample size settings
- Empirical Application (?)

Thank You!

Thank You!

Estimation Procedure: Penalty Parameter Selection

To obtain a finite sample bound on the ℓ_1 estimation error, we would like to use the distribution of the "true" residuals

$$\bar{U}_{\gamma,j} = -p_j(X)De^{-\bar{\gamma}'Z} + (1-D)$$
$$\bar{U}_{\alpha,j} = p_j(X)De^{-\bar{\gamma}'Z}(Y - \bar{\alpha}'Z)$$

and select the penalty parameters $\lambda_{\gamma,j}$ and $\lambda_{\alpha,j}$ dominate the scores below with high probability:

$$\|\mathbb{E}_n[\bar{U}_{\gamma,j}Z]\|_\infty \quad \text{and} \quad \|\mathbb{E}_n[\bar{U}_{\alpha,j}Z]\|_\infty,$$

However, we do not know the distribution of these residuals. Use of $\hat{\gamma}$ in the estimating equation for $\hat{\alpha}$ prevents use of standard data dependent procedures

◂ Back

Estimation Procedure: Penalty Parameter Selection

To get around this, Tan (2020) uses theoretical penalty parameters that dominate the scores with high probability.

- Theoretical penalty parameters typically depend on constants that are unknown to the researcher
- Even if known, this penalty parameter tends to be too large in practice

Back

Estimation Procedure: Penalty Parameter Selection

We incorporate a data-dependent method of selecting the penalty parameter, based on the multiplier bootstrap method of Chetverikov and Sørenson (2021).

1. Initialize a conservative guess of the penalty parameters, following guidance given in the paper.

Estimation Procedure: Penalty Parameter Selection

We incorporate a data-dependent method of selecting the penalty parameter, based on the multiplier bootstrap method of Chetverikov and Sørenson (2021).

1. Initialize a conservative guess of the penalty parameters, following guidance given in the paper.
2. Use this conservative guess to obtain pilot estimators, $\hat{\gamma}_j^{\text{pilot}}$ and $\hat{\alpha}_j^{\text{pilot}}$ and construct plug in estimates for the scores, $\widehat{U}_{\gamma,j}$ and $\widehat{U}_{\alpha,j}$.

Estimation Procedure: Penalty Parameter Selection

We incorporate a data-dependent method of selecting the penalty parameter, based on the multiplier bootstrap method of Chetverikov and Sørenson (2021).

1. Initialize a conservative guess of the penalty parameters, following guidance given in the paper.
2. Use this conservative guess to obtain pilot estimators, $\hat{\gamma}_j^{\text{pilot}}$ and $\hat{\alpha}_j^{\text{pilot}}$ and construct plug in estimates for the scores, $\widehat{U}_{\gamma,j}$ and $\widehat{U}_{\alpha,j}$.
3. Select the penalty parameters via a multiplier bootstrap procedure:

$$\lambda_{\gamma,j} = c_0 \times (1-\epsilon) \text{ quantile of } \max_{1\leq p\leq d_z} \left|\mathbb{E}_n[e\widehat{U}_{\gamma,j} Z_p]\right| \tag{5}$$

$$\lambda_{\alpha,j} = c_1 \times (1-\epsilon) \text{ quantile of } \max_{1\leq p\leq d_z} \left|\mathbb{E}_n[e\widehat{U}_{\alpha,j} Z_p]\right| \tag{6}$$

where $e \sim N(0,1)$ is generated independently of the data, $c_0, c_1 > 1$ and $\epsilon = o(k^{-1})$ is an error probability. Back

Results: Residual Estimation

Lemma 1 (Residual Estimation)

Suppose Assumption 1 holds and that $\hat{\gamma}_j^{pilot}$ *and* $\hat{\alpha}_j^{pilot}$ *are estimated using the estimating equations* (1)-(2) *with penalty parameters*

$$\lambda_{\gamma,j}^{pilot} = c_\gamma \sqrt{\frac{\ln^2(d_z n)}{n}} \quad \textit{and} \quad \lambda_{\alpha,j}^{pilot} = c_\alpha \sqrt{\frac{\ln^2(d_z n)}{n}}.$$

for some fixed constants $c_\gamma, c_\alpha > 0$. *If, in addition,* $k/n \to 0$*:*

$$\begin{aligned} \sup_{1 \le j \le k} \mathbb{E}_n \left[(\widehat{U}_{\gamma,j} - \bar{U}_{\gamma,j})^2 \right] &\lesssim_p \frac{\xi_{k,\infty} s_k \ln^2(d_z n)}{n} \\ \sup_{1 \le j \le k} \mathbb{E}_n \left[(\widehat{U}_{\alpha,j} - \bar{U}_{\alpha,j})^2 \right] &\lesssim_p \frac{\xi_{k,\infty}^2 s_k^2 \ln^2(d_z n)}{n} \end{aligned} \tag{7}$$

◂ Back

First Stage Assumptions: Uniformly Subgaussian

We assume the errors $\epsilon_j = Y_1 - \bar{m}_j(Z)$ are uniformly subgaussian conditional on Z and over all j in the following sense.

There exist finite constants $D_0, D_1 > 0$ such that, for all j:

$$D_0^2 \mathbb{E}\left[\exp\left(\epsilon_j^2 / D_0^2\right) - 1 \mid Z\right] \leq D_1^2.$$

First Stage Assumptions

First Stage Assumptions: Empirical Compatibility Condition

We say a matrix Σ obeys an empirical compatability constraint on a set $\mathcal{S} \subseteq \{1, \ldots, p\}$ with constants $\xi_0 > 1$ and $1 \geq \nu_0 > 0$ if for any $\delta \in \mathbb{R}^p$:

$$\sum_{j \notin \mathcal{S}} |\delta_j| \leq \xi_0 \sum_{j \in \mathcal{S}} |\delta_j| \implies \nu_0^2 \left(\sum_{j \in \mathcal{S}} |\delta_j| \right)^2 \leq |\mathcal{S}| \left(\delta' \Sigma \delta \right).$$

We assume that all the matrices $\Sigma_{\gamma,j} = \mathbb{E}_n[p_j(X) e^{-\bar{\gamma}' Z} Z Z']$ obey this empirical compatibility condition with the same constants ξ_0 and ν_0.

◂ First Stage Assumptions