High-Dimensional Causal Inference

Dominic DiSanto

Junwei Lu Reading Group - Summer 2024

July 11, 2024

Outline

Caveats

- We will focus exclusively on counterfactual/potential outcomes framework
- This excludes
  - Graphical Methods
  - Targeted MLE
    - https://www.khstats.com/blog/tmle/tutorial
    - https://tlverse.org/tlverse-handbook/tmle3.html

Papers of Focus

- *"Debiased Inverse Propensity Score Weighting for Estimation of Average Treatment Effects with High-Dimensional Confounders"*
  - Yuhao Wang & Rajen Shah [WS24]
- *"Causal Inference with High-dimensional Discrete Covariates"*
  - Zhenghao Zeng, Sivaraman Balakrishnan, Yanjun Han, Edward H. Kennedy [Zen+24]

Typical Causal Set-Up

- Estimand is ATE, $\tau = \mathbb{E}(Y(1) - Y(0))$
- Observe $T \in \{0,1\}^N$, $\mathbf{X} \in R^{N \times d}$ pre-treatment covariates
- Common assumptions:

$$\text{Unconfoundedness: } \{Y(1), Y(0)\} \perp\!\!\!\perp T \mid \mathbf{X}$$
$$\text{SUTVA: } Y_i = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$$
$$\text{Positivity: } \mathbb{P}(T_i = 1 \mid \mathbf{X}_i) =: \pi(\mathbf{x}) \in [\epsilon, 1-\epsilon] \text{ for } f(\mathbf{x}) > 0$$

AIPW

- IPW estimator $\hat{\tau}_{\text{IPW}} := \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}$
  - If $\hat{\pi} \xrightarrow{p} \pi$ consistent
- AIPW estimator $\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i(Y_i-\mu_i)}{\hat{\pi}(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)(Y_i-\mu_i)}{1-\hat{\pi}(X_i)}$
  - μ is some/any "augmentation" that ideally retains unbiasedness and reduces variance of our estimator
  - $\mu \perp T \mid X$ retains unbiasedness
  - $\hat{\mu} = (1-\hat{\pi}(X))\hat{r}_1(X) + \hat{\pi}(X)\hat{r}_0(x)$ is the common "AIPW"
    - For $r_j(X) = \mathbb{E}(Y(j) \mid X) = \mathbb{E}(Y \mid T = j, X)$

AIPW cont'd

- One can show $\sqrt{n}(\tau_{\text{AIPW}} - \tau) \xrightarrow{d} N(0, V)$
- Comparing $\hat{\tau}_{\text{AIPW}}$ to an oracle (in π, μ) τ_{AIPW}, we have

$$\left|\hat{\tau}_{AIPW} - \hat{\tau}^{*}_{AIPW}\right|$$
$$= O_P\left(\max_{w \in \{0,1\}} \mathbb{E}\left[\left(\hat{r}_w(X_i) - r_w(X_i)\right)^2\right]^{\frac{1}{2}} \mathbb{E}\left[\left(\hat{e}(X_i) - e(X_i)\right)^2\right]^{\frac{1}{2}}\right)$$

- If r_j, π are $n^{-1/4}$ estimable, $\sqrt{n}(\hat{\tau}_{\text{AIPW}} - \tau) \xrightarrow{d} N(0, V)$ and $\hat{\tau}_{\text{AIPW}}$ is semiparametrically efficient

Outline

"Double-Selection" Methods

When do we remain doubly-robust while performing model selection?

- "Double-selection" methods - Lasso variable selection followed by unpenalized but supplemented re-fitting
  - Under a (partially-) linear model $Y = \tau T + f(X) + \epsilon$, can recover asymptotic normality under double-Lasso selection [BCH12]
    - Requires $s_\pi \vee s_r = o\left(\sqrt{n}/\log p\right)$
    - See pg. 11 for details on implementation `https://arxiv.org/pdf/1201.0224`
  - Similar work by [Far15] with stricter $s_\pi s_r = o(\sqrt{n}/\log(p)^{1.5+\delta}),\ \delta > 0$ using group lasso, double-selection style
    - See pg. 21 for procedure, pg. 7 Corollary 1 for sparsity requirements `https://arxiv.org/pdf/1309.4686`

High-level, these methods require both r, π to be $\sqrt{n}/\log(p)$-sparse

Exploiting Sparsity Structure or “De-biasing” Methods

- [BWZ19] assume linear-logistic model and “ultra”-sparsity of *either* model, under weaker sparsity conditions on the latter
- “Double-robusty sparsity” when we have bounded $||\beta_\pi||_1, ||\beta_r||_1$ (see Theorem 1)
  - $s_\pi = o(\sqrt{n}/\log(p)), s_r = o(n/\log(p))$ and $||\beta_r||_1$ is large
  - $s_\pi = o(n^{3/4}/\log(p)), s_r = o(\sqrt{n}/\log(p))$
- Bias can decompose as $|\hat{\tau}_1 - \tau| \leq \left\| n^{-1} \sum_{i=1}^n \left[1 - W_i \left(1 + \exp\left(-X_i' \hat{\theta}_{(1)}\right)\right)\right] X_i \right\|_\infty \left\| \hat{\beta}_{(1)} - \beta_{(1)} \right\|_1$

Exploiting Sparsity Structure or "De-biasing" Methods

- [AIW18][1] require only $s_r = o(\sqrt{n}/\log(p))$ Athey paper https://arxiv.org/pdf/1604.07125
  - Estimate outcome coefficients $\hat{\beta}$ by lasso
  - Estimate balancing weights γ (see pg. 6/7)
  - $\hat{\mu}_{\mathrm{c}} = \bar{X}_{\mathrm{t}} \cdot \hat{\beta}_{\mathrm{c}} + \sum_{\{i:W_i=0\}} \gamma_i \left(Y_i^{\mathrm{obs}} - X_i \cdot \hat{\beta}_{\mathrm{c}} \right)$
  - $|\hat{\mu}_{\mathrm{c}} - \mu_{\mathrm{c}}| \leq \left\| \bar{X}_{\mathrm{t}} - \mathbf{X}_{\mathrm{c}}^{\top} \gamma \right\|_{\infty} \left\| \hat{\beta}_{\mathrm{c}} - \beta_{\mathrm{c}} \right\|_1 + \left| \sum_{\{i:W_i=0\}} \gamma_i \varepsilon_i \right|$
    - Here $\eta = Y(0) - X^T \beta_c$, i.e. outcome regression noise
  - Resulting estimator is $\sqrt{n}-$consistent and asymptotically normal under additional technical conditions

[1] Notation here for their primary endpoint, ATT, but extendable to ATE

Outline

Estimator

Estimand $\tau = \mathbb{E}(Y(1) - Y(0))$

$$\hat{\tau}_{\text{DIPW}} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i\,(Y_i - \hat{\mu}_i)}{\hat{\pi}_i} - \frac{(1-T_i)\,(Y_i - \hat{\mu}_i)}{1-\hat{\pi}_i}\right)$$

- $\hat{\pi}$ evaluated via lasso logistic regression
- $\hat{\mu}$ is evaluated via a quadratic program, an "orthogonalization" of the AIPW style augmentation
- Requires only $s_\pi = o(\sqrt{n}/\log p)$ for consistency, $o(1/\sqrt{\log n})$ estimation of regression models achieves semiparametric efficiency
    - Equivalent to $s_r = o(n/[\log n \log p])$ requirement
    - **Caveat:** Inference (i.e. CI's) require s_π assumptions

Estimation Procedure Outline

- Estimate $\hat{\pi} = \psi(x^T\hat{\gamma})$ with $\hat{\gamma}$ estimated via lasso on auxiliary data $\mathcal{D}_B$
  - Assumption: $s_\pi = o(\sqrt{n}\log(p))$
- Construct $\tilde{\mu}$, an estimate of μ_{ORA}, using $\mathcal{D}_B$
- Construct $\hat{\mu}$ using $\mathcal{D}_A$ by convex program above
- Plug-in and estimate $\hat{\tau}_{\mathrm{DIPW}}$, AIPW style estimator with $\hat{\mu}, \hat{\pi}$

Estimation Procedure

- Observe n, iid $(X, Y, T) \in \mathbb{R}^p \times \mathbb{R} \times \{0, 1\}$, say $\mathcal{D} = (\mathbf{X}, \mathbf{Y}, \mathbf{T})$
  - We will also use auxiliary datasets $\mathcal{D}_A = (\mathbf{X}_A, \mathbf{Y}_A, \mathbf{T}_A), \mathcal{D}_B$
  - Assume X, Y are σ_Y^2, σ_X^2 sub-Gaussian and $\max_{t\in\{0,1\}} |\mathbb{E}Y(t)| < m_Y$
- Assume a logistic model for the propensity $\pi(x) = \mathbb{P}(T = 1 \mid X = x) = \psi(x^T\gamma) := (1 + \exp(-x^T\gamma))^{-1}$
- Let $\mu_{\text{ORA}}(x) := (1 - \pi(x))r_1(x) + \pi(x)r_0(x)$
  - Recall $\mathbb{E}\tau_{\text{ORA}} = \tau$ if $\mu \perp T \mid X$

Estimation Procedure

- Let $\tilde{Y}_i := \frac{T_i Y_i (1-\hat{\pi}_i)}{\hat{\pi}_i} + \frac{(1-T_i) Y_i \hat{\pi}_i}{1-\hat{\pi}_i}$
- Estimate $\hat{\gamma}$ using $\mathcal{D}_B$
- Bias in $\hat{\tau}_{\text{IPW}}$ then becomes[2] determined by

$$\approx \left| \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{Y}_i - \mu_i \right) X_i^\top (\hat{\gamma} - \gamma) \right| \leq \frac{1}{n} \left\| \mathbf{X}^\top \tilde{\mathbf{Y}} - \mathbf{X}^\top \boldsymbol{\mu} \right\|_\infty \|\hat{\gamma} - \gamma\|_1$$

  - $\|\hat{\gamma} - \gamma\|_1 = o(s\sqrt{\log p/n})$ whp
  - So to control bias of $\hat{\tau}_{\text{DIPW}}$, we need only study $\frac{1}{n} \left\| \mathbf{X}^\top \tilde{\mathbf{Y}} - \mathbf{X}^\top \boldsymbol{\mu} \right\|_\infty$

[2] From comparing $\hat{\tau}_{\text{IPW}} - \tau_{\text{ORA}}$

Estimation Procedure

$$\frac{1}{n}\left\|\mathbf{X}^\top\tilde{\mathbf{Y}} - \mathbf{X}^\top\boldsymbol{\mu}\right\|_\infty \leq \left\|\frac{1}{n_A}\mathbf{X}_A^\top\left\{\tilde{\mathbf{Y}}_A - f\left(\mathbf{X}_A\right)\right\} - \frac{1}{n}\mathbf{X}^\top\{\boldsymbol{\mu} - f(\mathbf{X})\}\right\|_\infty + \left\|\frac{1}{n_A}\mathbf{X}_A^\top\left\{\tilde{\mathbf{Y}}_A - f\left(\mathbf{X}_A\right)\right\} - \frac{1}{n}\mathbf{X}^\top\{\tilde{\mathbf{Y}} - f(\mathbf{X})\}\right\|_\infty$$

- First term allows us to approximate μ by $\mathcal{D}_A$, respecting the requirement $\mu \perp T \mid X$
- Second term $\leq c\sqrt{\log p/\min\{n, n_A\}}$ under sub-Gaussian assumptions on X, Y, X_A, Y_A
- f is a fixed function, which we will choose

Estimation Procedure

- Considering $\left\|\frac{1}{n_A}\mathbf{X}_A^\top\left\{\tilde{\mathbf{Y}}_A - f\left(\mathbf{X}_A\right)\right\} - \frac{1}{n}\mathbf{X}^\top\{\boldsymbol{\mu} - f(\mathbf{X})\}\right\|_\infty \leq \eta$
- Select $\eta \asymp \sqrt{\log(p)/n}$, then

$$||\mathbf{X}^T\tilde{\mathbf{Y}} - \mathbf{X}^T\mu||_\infty||\hat{\gamma} - \gamma||_1 \leq \sqrt{\log(p)/n} \cdot s\sqrt{\log(p)/n} = s\log(p)/n$$

  - $o(n^{-1/2})$ under $x = o(\sqrt{n}/\log(p))$
- Remains to identify f, μ

Estimation Procedure

- **Lemma 1:** μ_{ORA} minimizes $V(\tau_{\text{ORA}})$.
  - So ideally $\mu \approx \mu_{\text{ORA}}$, but cannot regress $\tilde{\mathbf{Y}} \sim \mathbf{X}$ (as we require $\mu \perp \mathbf{T} \mid \mathbf{X}$)
  - Construct $\tilde{\mu}$ using $\mathcal{D}_B$, then estimate $\mu = \textit{argmin}||\mu - \tilde{\mu}(\mathbf{X})||_2^2$, using $\tilde{\mu} = f$

Thus estimate $\hat{\mu}$ by the convex program

$$\hat{\boldsymbol{\mu}} = \text{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^n} \frac{1}{n} \|\tilde{\mu}(\mathbf{X}) - \boldsymbol{\mu}\|_2^2$$

$$\text{subject to } \left\| \frac{1}{n_A} \mathbf{X}_A^\top \left\{ \tilde{\mathbf{Y}}_A - \tilde{\mu}\left(\mathbf{X}_A\right) \right\} - \frac{1}{n} \mathbf{X}^\top \{ \boldsymbol{\mu} - \tilde{\mu}(\mathbf{X}) \} \right\|_\infty \leq \eta$$

Estimation Procedure Outline

- Estimate $\hat{\pi} = \psi(x^T\hat{\gamma})$, $\hat{\gamma}$ estimated via lasso on auxiliary data $\mathcal{D}_B$
  - Assumption: $s_\pi = o(\sqrt{n}\log(p))$
- Construct $\tilde{\mu}$, an estimate of μ_{ORA}, using $\mathcal{D}_B$
- Construct $\hat{\mu}$ using $\mathcal{D}_A$ by convex program above
- Plug-in and estimate $\hat{\tau}_{\text{DIPW}}$, AIPW style estimator with $\hat{\mu}, \hat{\pi}$

Inference

- **See Theorem 3 (pg. 11) for asymptotic normality result and subsequent CI construction**
  - **Note new dependence on** $\|\mu - \mu_{\mathbf{ORA}}\|_\infty$ **(both) and** $s_\pi = o(\sqrt{n}/\log(p))$ **assumption (CI construction only)**

Additional Notes

- Asymptotic near-normality result $\sqrt{n}\left(\hat{\tau}_{DIPW} - \tau\right) = \delta + \sigma_\mu \zeta_1 + \sigma \zeta_2$ conditional on $\mathcal{D}$
  - $\delta < c(s + \sqrt{s \log(n)} \log(p)/\sqrt{n})$ whp
  - η is "near-Normal" in a Berry-Esseen sense, see Theorems 2 & 3
- Can use a sample-splitting procedure in place of hold-out/auxiliary data sets
- Can extend to link functions (say ϕ) beyond $\psi(x) = (1 + \exp(-x))^{-1}$, with conditions on ϕ', ϕ''

Outline

Set-Up

- Observe iid $(Y, X, A) \in \{0,1\} \times \{0, ..., d\}^K \times \{0,1\}$
  - $P(X = k) = p_k, k \in [d]$
  - $A \mid X = k \sim Ber(\pi_k)$
  - $Y \mid X = k, A = a \sim Ber(\mu_{ak})$
  - $q_{ak} = \mathbb{P}(X = k, A = a, Y = 1) = p_k \left[a\pi_k + (1-a)(1-\pi_k)\right] \mu_{ak}$
  - $w_k = \mathbb{P}(X = k, A = 1) = p_k \pi_k$
- Estimand is typical
  $\psi = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y|X, A = 1] - \mathbb{E}[Y|X, A = 0]]$
- Typical causal assumption, here positivity is on $\pi_k, k \in [d]$
  - Interesting/specific consideration of ϵ here as we might expect $\epsilon_n \to 0$ as $n \to \infty$ see Remark 1

Estimator Equivalence

Consider our suite of classical estimators

$$\widehat{\psi} = \sum_{k=1}^{d} \widehat{p}_k \left(\widehat{\mu}_{1k} - \widehat{\mu}_{0k}\right) = \sum_{k=1}^{d} \widehat{p}_k \left(\frac{\widehat{q}_{1k}}{\widehat{w}_k} - \frac{\widehat{q}_{0k}}{\widehat{p}_k - \widehat{w}_k}\right)$$

$$\widehat{\psi}_{\mathrm{reg}} = \mathbb{P}_n \left[\widehat{\mu}_{1X} - \widehat{\mu}_{0X}\right],$$

$$\widehat{\psi}_{\mathrm{ipw}} = \mathbb{P}_n \left[\frac{AY}{\widehat{\pi}_X} - \frac{(1-A)Y}{1-\widehat{\pi}_X}\right],$$

$$\widehat{\psi}_{\mathrm{dr}} = \mathbb{P}_n \left[\frac{A\left(Y - \widehat{\mu}_{1X}\right)}{\widehat{\pi}_X} + \widehat{\mu}_{1X} - \frac{(1-A)\left(Y - \widehat{\mu}_{0X}\right)}{1-\widehat{\pi}_X} - \widehat{\mu}_{0X}\right]$$

Claim: $\hat{\psi} = \hat{\psi}_{Reg} = \hat{\psi}_{\mathrm{IPW}} = \hat{\psi}_{\mathrm{AIPW}}$ for $\hat{\psi}$ constructed using sample-average plug-in estimators for μ, π, q, w

So we consider only

$$\widehat{\psi} = \sum_{k=1}^{d} \widehat{p}_k \left(\widehat{\mu}_{1k} - \widehat{\mu}_{0k}\right) = \sum_{k=1}^{d} \widehat{p}_k \left(\frac{\widehat{q}_{1k}}{\widehat{w}_k} - \frac{\widehat{q}_{0k}}{\widehat{p}_k - \widehat{w}_k} \right)$$
$$= \hat{\psi}_1 - \hat{\psi}_0$$
$$\psi = \sum_{k=1}^{d} p_k (\mu_{1k} - \mu 0k) = \sum_{k=1}^{d} p_k \left(\frac{q_{1k}}{w_k} - \frac{q_{0k}}{p_k - w_k} \right)$$
$$= \psi_1 - \psi_0$$

Estimation Rates

- See Proposition 3 for bias-derivation, requires $d = o(n)$ scaling
- See Proposition 4 for minimax lower bound contains $\frac{d^2}{n^2 \log^2 n}$ terms, that is $\hat{\psi}$ is minimax optimal up to log factors

References I

[AIW18] Susan Athey, Guido W. Imbens, and Stefan Wager. *Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions*. Jan. 31, 2018. arXiv: 1604.07125[econ,math,stat].

[BCH12] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. *Inference on Treatment Effects After Selection Amongst High-Dimensional Controls*. May 9, 2012. arXiv: 1201.0224[econ,stat].

[BWZ19] Jelena Bradic, Stefan Wager, and Yinchu Zhu. *Sparsity Double Robust Inference of Average Treatment Effects*. May 2, 2019. arXiv: 1905.00744[econ,math,stat].

References II

[Far15] Max H. Farrell. "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations". In: *Journal of Econometrics* 189.1 (Nov. 2015), pp. 1–23. arXiv: 1309.4686 [econ,math,stat].

[WS24] Yuhao Wang and Rajen D. Shah. *Debiased Inverse Propensity Score Weighting for Estimation of Average Treatment Effects with High-Dimensional Confounders*. Apr. 11, 2024. arXiv: 2011.08661 [math,stat].

[Zen+24] Zhenghao Zeng et al. *Causal Inference with High-dimensional Discrete Covariates*. May 5, 2024. arXiv: 2405.00118 [math,stat].

Misc. Undiscussed Papers I

- "Debiasing the Lasso: Optimal Sample Size for Gaussian Designs" - Adel Javanmard, Andrea Montanari
  https://arxiv.org/pdf/1508.02757
  - Discusses gap between lasso rate $s = o(n/\log p)$ and $s = o(\sqrt{n}/\log p)$ requirement in Slide 9 refs
- *"Deep Neural Networks for Estimation and Inference"* - Max H. Farrell, Tengyuan Liang, Sanjog Misra
  https://arxiv.org/pdf/1809.09953
  - Establishes conditions for semi-parametric efficiency of ATE when using deep NN's
- *"Program Evaluation and Causal Inference with High-Dimensional Data"* - Belloni, Chernozhukov, Fernandez-Val, Hansen
  https://arxiv.org/pdf/1311.2645
  - Contemporary with some of the "double-selection" methods discussed