Empirical Gateaux Derivatives for Causal Inference

Angela Zhou
Assistant Professor
USC Marshall Data Sciences and Operations
zhoua@usc.edu

Joint work with Yixin Wang and Michael Jordan

- Previously: robust decision-making to unobserved confounders
- This talk: computerized influence functions for causal inference
- Also: dynamic experience optimization; sequential causal inference with structure

Motivation

- State-of-the art ML + causal inference via *debiased* estimators (orthogonalized, doubly-robust)[1]
- Influence functions[2]
  "Taylor" expansions of functionals wrt. prob. distributions
  Derived by hand, guess-and-check, intuition ...
- Computerized influence functions[3] by finite differences[4]
  - Ex: Constrained MDP

[1] [Chernozhukov et al., 2018, Kennedy, 2022]
[2] Hines et al. [2022], Kennedy [2022]
[3] Chernozhukov et al. [2021]
[4] Carone et al. [2018], Frangakis et al. [2015]

Outline

- Background/setup
- Results
  - Characterization: mean under missingness (augmented IPW)
  - More complex examples
    - tabular infinite-horizon off policy optimization and evaluation for RL

Setup

- $\Psi(P)$ statistical functional of distribution P
  Observations $O \sim P$

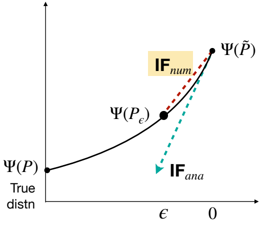
- **Plug-in evaluation** of Ψ
  Estimate distribution $\tilde{P}$, plug into Ψ
  If $O = (X, A, Y)$, for *density estimates* $\tilde{p}$:

$$\tilde{p}(y, A = 1, x) = \tilde{p}(x)\tilde{p}(A = 1 \mid x)\tilde{p}(y \mid A = 1, x)$$

- Example: Mean under missingness

$$\Psi(P) = \mathbb{E}[Y(1)] = \mathbb{E}[\mathbb{E}[Y \mid A = 1, X]] = \int \int y \frac{p(y, A=1, x)}{p(A=1, x)} p(x) dy dx$$

Influence functions of functionals



Cartoon (One-step estimator)

- Influence function[a]

$$\mathrm{IF}(o;P) = \frac{d\Psi\left(P + \epsilon(\delta_o - P)\right)}{d\epsilon}\bigg|_{\epsilon=0}$$

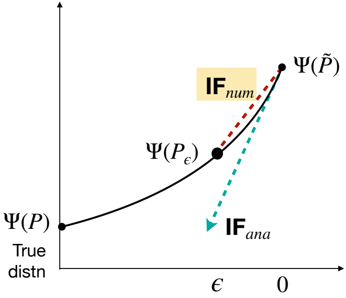
(If Gateaux diff'able)

- One-step estimator:

$$\Psi_{os}(P) = \Psi(P) + \mathbb{E}_n[\mathrm{IF}(O)]$$

[a] Hampel [1974], Huber [2004]

Influence functions of functionals



Cartoon (One-step estimator)

- **Finite difference approximation**

$$\begin{aligned}&\mathrm{IF}_{num}(o; P, \epsilon, \lambda)\\ &= \ \epsilon^{-1}(\Psi(P_{\epsilon,\lambda}) - \Psi(P))\end{aligned}$$

- Carone et al. [2018], Ichimura and Newey [2015]: smoothed Dirac $\tilde{\delta}^\lambda(o_i)$,

$$P_\epsilon^{o_i} = P_\epsilon^i = (1-\epsilon)P + \epsilon\tilde{\delta}^\lambda(o_i)$$

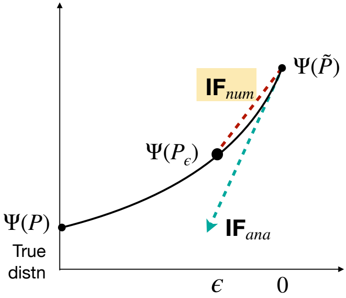
with $\tilde{\delta}^\lambda_{o_i}(o) = K_\lambda(o - o_i)$

a kernel, bandwidth λ

ϵ: finite-difference apx. error

λ: error from smoothing

Influence functions of functionals



Cartoon (One-step estimator)

- **Finite difference approximation**

$$\text{IF}_{num}(o; P, \epsilon, \lambda) = \epsilon^{-1}\left(\Psi(P_{\epsilon,\lambda}) - \Psi(P)\right)$$

- Carone et al. [2018], Ichimura and Newey [2015] show

$$\text{IF}(o; P) = \lim_{\lambda\to 0} \lim_{\epsilon \to 0} \text{IF}_{num}(o; P, \epsilon, \lambda)$$

Need to estimate $\tilde{P}$ for this to be useful for estimation!

$$\mathrm{IF}_{emp}(O_i) = \boxed{\frac{\Psi(\tilde{P}_\epsilon^i) - \Psi(\tilde{P})}{\epsilon}} \quad \text{empirical; smoothed \& estimated distns.}$$

$$\mathrm{IF}_{num}(O_i) = \frac{\Psi(P_\epsilon^i) - \Psi(P)}{\epsilon} \quad \text{numerical; smoothed \& true distns,}$$

$$\mathrm{IF}(O_i) = \frac{d}{d\epsilon}\Psi(P_\epsilon^i)\Big|_{\epsilon=0} \quad \text{analytical Gateaux derivative.}$$

- Q1: How does the empirical Gateaux derivative approximate the numerical derivative?
- Q2: What rates of numerical approximation preserve statistical properties? (i.e. rate double-robustness)

Computerized IF

Black-box evaluation of the functional

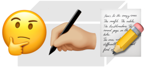
$$\tilde{p}(y, A = 1, x),\ \ \tilde{p}(A = 1, x),\ \ \tilde{p}(x)$$

$$\tilde{E}_P[Y \mid A = 1, X],\ \ \tilde{p}(A = 1 \mid x)$$

$$\Psi(P)$$

Analytical IF

$$\Psi(P)$$



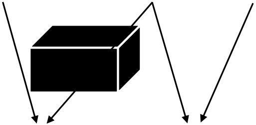
$$E_P[Y \mid A = 1, X],\ \ p(A = 1 \mid x)$$

Black-box estimation of nuisances

Computerized IF

Black-box evaluation of the functional

$\tilde{p}_{\epsilon}(y, A=1, x),\ \tilde{p}_{\epsilon}(A=1, x),\ \tilde{p}_{\epsilon}(x)$

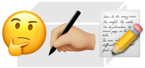


$\tilde{E}_{P_{\epsilon}}[Y \mid A=1, X],\ \tilde{p}_{\epsilon}(A=1 \mid x)$

$\Psi(P_{\epsilon})$

Analytical IF

$\Psi(P)$



$E_P[Y \mid A=1, X],\ p(A=1 \mid x)$

Black-box estimation of nuisances

$\lambda-$smoothed nuisance function,

$$\tilde{\mathbb{E}}_P[Y \mid A=1, X=x_0] = \int \mathbb{E}_P[Y \mid A=1, X=u]\, \tilde{\delta}_{x_0}^{\lambda}(u)\, du$$

Answering Q1+Q2

- Specialize to kernel density estimates, assume β-Holder smooth
- Error of perturbed nuisances in (ϵ, λ)
  Lemma: dim d,

$$\mathbb{E}[(\tilde{\mu}_\epsilon(X) - \mu(X))^2] = \mathbb{E}[(\tilde{\mu}(X) - \mu(X))^2] + O(\epsilon^2 \lambda^{-d})$$

- **Proposition/Corollary**: When the perturbation observation is the observation datapoint $O_i = (X_i, A_i, Y_i)$,

$$\tilde{\phi}(O_i) = \frac{\mathbb{I}[A_i = 1]}{\tilde{p}_\epsilon(A = 1 | X_i)} \left(Y_i - \mathbb{E}_{\tilde{P}}[Y \mid A = 1, X_i]\right) + \left(\mathbb{E}_{\tilde{P}_\epsilon}[Y \mid A = 1, X_i] - \Psi(\tilde{P})\right) + O(\lambda^\beta).$$

- **Theorem** (informal): achieve the parametric rate when $\epsilon \lambda^{-d/2} = o(n^{-\max(r_\mu, r_e)})$, and $\lambda^\beta = o(n^{-\frac{1}{2}})$
  (and plug-in nuisances are fast enough)

Case study

- Infinite-horizon off-policy optimization (analogous to [Tang et al., 2019])

$$(s, a, s', r, \dots)$$

$P(s' \mid s, a)$ (transition probability $\approx$ outcome model), $\frac{\mu_{\pi_e}(s,a)}{d(\tilde{s},\tilde{a})}$ (density ratio $\approx$ propensity)

- Primal/dual of linear-programming formulation:

$$\Psi_D(P) = \min_V \{(1-\gamma){\mu_0}^\top \mathrm{V} \colon \ (I - \gamma P_a)\,\mathrm{V} - \mathrm{r}_a \geq 0, \quad \forall a \in \mathcal{A}\}$$

$$\Psi_P(P) =$$

$$\max_\mu \left\{ \sum_{a \in \mathcal{A}} \mu_a^\top \mathrm{r}_a \colon \ \sum_{a \in \mathcal{A}} (I - \gamma P_a^\top)\mu_a = (1-\gamma)\mu_0, \ \mu_a \geq 0, \ \forall a \in \mathcal{A} \right\}$$

- V value function, μ^* the state-action occupancy under optimal distribution, r rewards

- **Proposition** (Analytical Gateaux derivative):
  Assume asymptotic linearity (nondegeneracy)
  Perturb to $o = (\tilde{s}, \tilde{a}, \tilde{s}')$:

$$\frac{d}{d\epsilon}\Psi_D(P_\epsilon)\Big|_{\epsilon=0} = (1-\gamma)V^*(\tilde{s}) - \frac{\mu^*(\tilde{s},\tilde{a})}{d(\tilde{s},\tilde{a})}\left(r(\tilde{s},\tilde{a}) + \gamma V^*(\tilde{s}') - V^*(\tilde{s})\right)$$

  perturbation analysis of optimization programs [Freund, 1985]
  (sensitivities = dual variables)

- **Proposition** (Empirical Gateaux derivative): for ϵ small enough to maintain the same active basis,

$$\epsilon^{-1}\left(\Psi\left(P_\epsilon\right) - \Psi(P)\right) = (1-\gamma){V_\epsilon}^*(\tilde{s}) - \frac{\mu^*(\tilde{s},\tilde{a})}{d_\epsilon(\tilde{s},\tilde{a})}\left(r(\tilde{s},\tilde{a}) + \gamma V^*\left(\tilde{s}'\right) - V_\epsilon^*(\tilde{s})\right) - \Psi_D(P) + O(\epsilon)$$

Illustration:
Suppose an analyst was doing model-based evaluation.
They estimated a transition probability model.

"Empirical gateaux derivatives" allow them to
approximate influence function adjustments
without estimating any additional nuisances,
and add any arbitrary constraints to the opt. problem.

Hence this can be a helpful first step.

Example epsilon-lambda plot

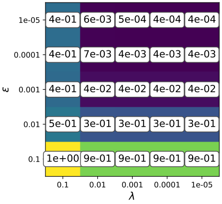


Figure: (ϵ, λ) plot for tuning.
Estimand varies if f.d. is unstable

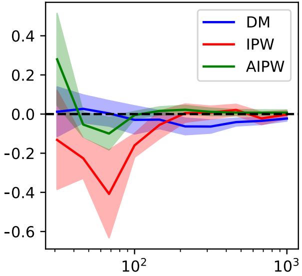


Figure: Convergence (AIPW from empirical Gateaux derivatives)

- Extremely simple example to illustrate (ϵ, λ) plot[5]
- Need more empirical work (incl. tuning f.d. scheme)

[5] Carone et al. [2018]

Thanks!

- Finite-difference computation of influence functions
- Optimization-based estimators
- New avenues for algorithm design?

zhoua@usc.edu

Related work

- Similar analytical derivations: Hines et al. [2022] derives influence functions by Gateaux derivatives; Kennedy [2022] suggests an "IF calculus"
- Automatic debiasing Chernozhukov et al. [2021]; variational characterization of Ichimura and Newey [2022]

Sensitivity analysis

- Let $P(Y \mid A=1, X) = P_{Y_1|X}$,

$$g(x) = \sup_{a(x) \leq W(x,y) \leq b(x), \forall y} \left\{ \mathbb{E}_{P_{Y_1|X}}[YW] \colon \mathbb{E}_{P_{Y_1|x}}[W] = 1, \forall x; \quad \right\} \quad (\mathrm{P}$$

-

$$\Psi^{opt}(P) = \mathbb{E}_{P_X}[g(X)] \qquad \text{optimization perspective}$$

$$\Psi^{f}(P) = \mathbb{E}_{P_X}[\mathbb{E}_{P_{Y_1|X}}[YW^*(X, Y)]] \qquad \text{closed-form, } W^*$$

M. Carone, A. R. Luedtke, and M. J. van der Laan. Toward computerized efficient estimation in infinite-dimensional models. *Journal of the American Statistical Association*, 2018.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

V. Chernozhukov, W. K. Newey, V. Quintas-Martinez, and V. Syrgkanis. Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv preprint arXiv:2104.14737*, 2021.

C. E. Frangakis, T. Qian, Z. Wu, and I. Diaz. Deductive derivation and turing-computerization of semiparametric efficient estimation. *Biometrics*, 71(4):867–874, 2015.

R. M. Freund. Postoptimal analysis of a linear program under simultaneous changes in matrix coefficients. In *Mathematical Programming Essays in Honor of George B. Dantzig Part I*, pages 1–13. Springer, 1985.

F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, pages 1–13, 2022.

P. J. Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *arXiv preprint arXiv:1508.01378*, 2015.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Z. Tang, Y. Feng, L. Li, D. Zhou, and Q. Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.