Doubly Robust Estimation with Split Training Data: Achieving Minimax Optimality with Undersmoothed Local Averaging Linear Smoothers

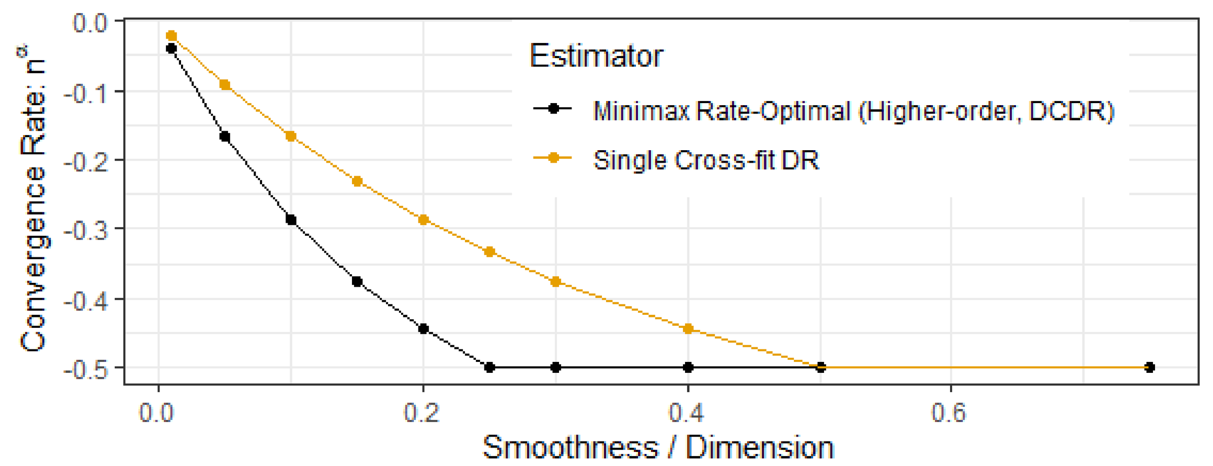
ALEC MCCLEAN, EDWARD H. KENNEDY, SIVARAMAN BALAKRISHNAN, AND LARRY WASSERMAN
DEPARTMENT OF STATISTICS AND DATA SCIENCE, CARNEGIE MELLON UNIVERSITY

MAIN IDEAS

- Doubly robust estimators with cross-fitting achieve favorable error guarantees, allowing for nonparametric nuisance function estimation while attaining parametric efficiency.
- However, with additional available structure, such as Hölder smoothness, we can construct even better estimators, including higher-order and ***double cross-fit doubly robust*** (DCDR) estimators [Robins et al. 2008, 2009, Newey and Robins, 2018, McGrath and Mukherjee, 2022].
- We construct DCDR estimators for the Expected Conditional Covariance and prove, with progressively stronger assumptions,
  1. a structure-agnostic linear expansion,
  2. adaptive minimal semiparametric efficiency,
  3. minimax rate-optimality, and
  4. slower-than-root-n inference.

SOME STRUCTURE –> FASTER RATES

With Hölder(s) smooth nuisance functions:



DATA, FUNCTIONAL, ASSUMPTIONS

Observe $3n$ iid observations $Z_i = \{X_i, A_i, Y_i\}_{i=1}^{3n}$, where

- $X \in \mathbb{R}^d$ are covariates,
- $A \in \{0, 1\}$ is a binary treatment, and
- $Y \in \mathbb{R}$ is an outcome.

Expected Conditional Covariance (ECC):

$$\psi_{ecc} = \mathbb{E}\{\text{cov}(A, Y \mid X)\},$$

ECC appears in the numerator of variance-weighted ATE [Li et al., 2011], measures causal influence [Díaz, 2023], and is used for conditional independence testing [Shah and Peters, 2020]. It is also the simplest mixed bias functional [Rotnitzky et al., 2019].

Un-centered efficient influence function:

$$\varphi_{ecc} = \{A - \pi(X)\}\{Y - \mu(X)\}$$

Nuisance Functions:

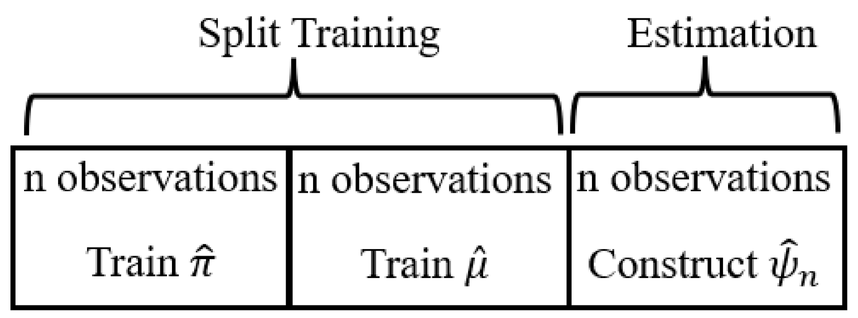
- $f(X)$ is the covariate density,
- $\pi(X) = \mathbb{P}(A = 1 \mid X)$ is the propensity score
- $\mu(X) = \mathbb{E}(Y \mid X)$ is the outcome regression

Smoothness: when relevant, we assume $\pi \in$ Hölder(α), $\mu \in$ Hölder(β), $f(X) \in$ Hölder($\alpha \vee \beta$). Hölder(s) smooth functions are approximately $s-1$ times differentiable with $s-1^{th}$ derivative Lipschitz.

THE DCDR ESTIMATOR

1. Split the data into three folds of n observations
2. Train *undersmoothed* $\widehat{\mu}$ and $\widehat{\pi}$ on *separate* folds
3. Construct the DCDR estimator on the third fold:

$$\widehat{\psi}_n = \frac{1}{n}\sum_{i=1}^{n}\widehat{\varphi}_{ecc}(Z_i) \equiv \frac{1}{n}\sum_{i=1}^{n}\{A - \widehat{\pi}(X_i)\}\{Y - \widehat{\mu}(X_i)\}$$



WHY DOUBLE SPLITTING + UNDERSMOOTHING

- Training the nuisance functions on the same data introduces a non-linearity bias, and splitting the training data removes this bias
- The single cross-fit estimator minimizes the nuisance function estimators' errors rather than the ECC estimator's error
- ***Undersmoothing*** the nuisance function estimators minimizes the bias of the DCDR estimator while averaging will prevent high variance, minimizing DCDR estimator's error

STRUCTURE-AGNOSTIC LINEAR EXPANSION

$$\widehat{\psi}_n - \psi_{ecc} = (\mathbb{P}_n - \mathbb{P})\varphi(Z) \quad \text{(CLT Term)}$$
$$+ O_{\mathbb{P}}\left(\underbrace{\|b_\pi\|\|b_\mu\|}_{\text{Bias}} + \underbrace{\sqrt{\frac{\mathbb{E}\|\widehat{\varphi} - \varphi\|^2 + \rho(\Sigma_n)}{n}}}_{\text{Standard Deviation}}\right),$$

where $\Sigma_n = \mathbb{E}\left(\text{cov}\left[\left\{\widehat{b}_\varphi(Z_1), ..., \widehat{b}_\varphi(Z_n)\right\}^T \mid X_1, ..., X_n\right]\right)$,

$\widehat{b}_\varphi(Z_i) = \mathbb{E}\{\widehat{\varphi}(Z_i) - \varphi(Z_i) \mid X_i, \textit{Training Data}\}$ is the conditional bias of $\widehat{\varphi}$, and $\rho(\cdot)$ is spectral radius.

- $\|b_\pi\|\|b_\mu\|$ is product of *biases* of $\widehat{\pi}$ and $\widehat{\mu}$.
- $\mathbb{E}\|\widehat{\varphi} - \varphi\|^2$ appears in usual analysis.
- $\rho(\Sigma_n)$ is new, but should be smaller than $\mathbb{E}\|\widehat{\varphi} - \varphi\|^2$.

Careful analysis of $\rho(\Sigma_n)$ shows

$$\frac{\rho(\Sigma_n)}{n} \lesssim \frac{\mathbb{E}\|\widehat{\varphi} - \varphi\|^2}{n}$$
$$+ \left(\overline{b}_\pi^2 + \overline{s}_\pi^2\right)\mathbb{E}\left[\text{cov}\{\widehat{\mu}(X_i), \widehat{\mu}(X_j) \mid X_i, X_j\}\right]$$
$$+ \underbrace{\left(\overline{b}_\mu^2 + \overline{s}_\mu^2\right)}_{\text{MSE of } \widehat{\mu}}\mathbb{E}\left[\underbrace{\text{cov}\{\widehat{\pi}(X_i), \widehat{\pi}(X_j) \mid X_i, X_j\}}_{\text{Covariance over training data of independent predictions}}\right]$$

General Applicability: these results apply with **any** nuisance function estimators under very weak assumptions. The initial linear expansion applies to all mixed bias functionals.

LINEAR SMOOTHERS

We consider three linear smoothers:

1. Local Polynomial Regression (LPR)
2. Orthonormal Series Regression (OSR)
3. "Covariate-density-adapted" Local Polynomial Regression (CDA-LPR)

CDA-LPR replaces estimated inverse sample design matrix with true inverse covariate density, *assuming this is known*.

For all estimators, $\mathbb{E}\left[\text{cov}\{\widehat{\eta}(X_i), \widehat{\eta}(X_j) \mid X_i, X_j\}\right] \lesssim \frac{1}{n}$, so

$$\widehat{\psi}_n - \psi_{ecc} = (\mathbb{P}_n - \mathbb{P})\varphi(Z) \quad \text{(CLT Term)}$$
$$+ O_{\mathbb{P}}\left(\|b_\pi\|\|b_\mu\| + \sqrt{\frac{\text{MSE}(\widehat{\pi}) + \text{MSE}(\widehat{\mu})}{n}}\right).$$

Minimizing remainder implies **undersmoothing is optimal!**

MINIMAL SEMIPARAMETRIC EFFICIENCY

With Hölder smooth π and μ, and DCDR estimator based on undersmoothed LPR with bandwidths scaling as $h_\mu, h_\pi \sim n^{-1/d}$ or OSR with regressor dimensions scaling as $k_\mu, k_\pi \sim n$, then

$$\begin{cases} \sqrt{\frac{n}{\mathbb{V}\{\varphi(Z)\}}}(\widehat{\psi}_n - \psi_{ecc}) \rightsquigarrow N(0,1) & \text{if } \frac{\alpha+\beta}{2} > d/4, \\ \sup_{\mathcal{P}_{\alpha,\beta}} \mathbb{E}\|\widehat{\psi}_n - \psi_{ecc}\| \lesssim n^{-\frac{\alpha+\beta}{d}} & \text{otherwise.} \end{cases}$$

Adaptivity: does not require knowledge of covariate density or true smoothness of π and μ.

MINIMAX RATE-OPTIMALITY

With Hölder smooth π and μ, known and smooth covariate density, and DCDR estimator based on CDA-LPR with one bandwidth scaling such that the estimator is consistent and the other scaling as $n^{\frac{-2}{2\alpha+2\beta+d}}$, then

$$\sup_{\mathcal{P}_{\alpha,\beta}} \mathbb{E}\|\widehat{\psi}_n - \psi_{ecc}\| \lesssim n^{-\left(\frac{2\alpha+2\beta}{2\alpha+2\beta+d} \wedge 1/2\right)}$$

The DCDR estimator can be **minimax rate-optimal**

SLOWER-THAN-ROOT-N INFERENCE

With Hölder smooth π and μ, known and smooth covariate density, and DCDR estimator based on CDA-LPR with one bandwidth scaling such that the estimator is consistent and the other scaling as $n^{-\frac{2+\varepsilon}{2\alpha+2\beta+d}}$ for $\varepsilon > 0$, then

$$\sqrt{\frac{n}{\mathbb{V}\{\widehat{\varphi}(Z) \mid \textit{Training Data}\}}}(\widehat{\psi}_n - \psi_{ecc}) \rightsquigarrow N(0,1)$$

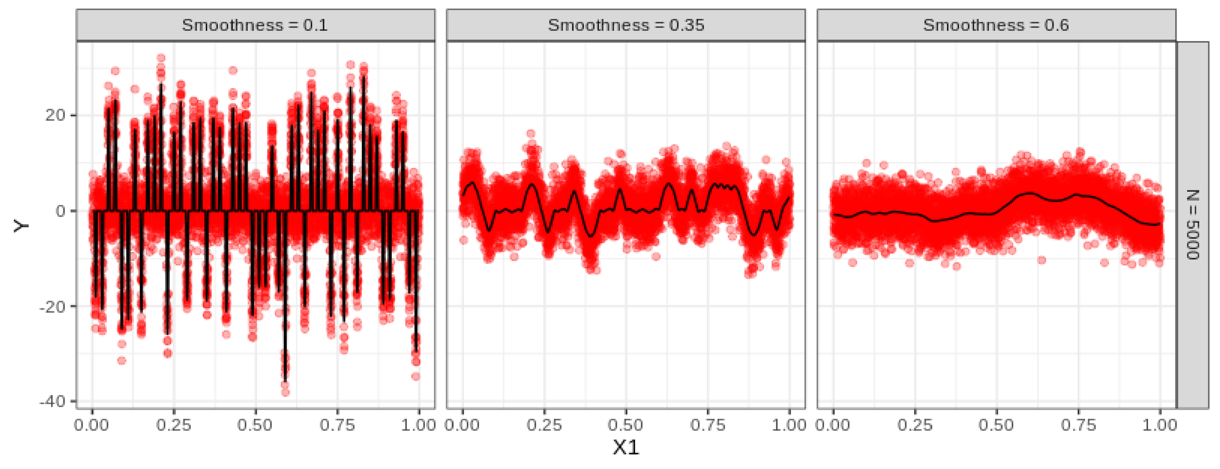
A CLT is feasible in the non-$\sqrt{n}$ regime.

(Proof inspired by Robins et al. 2015, Asymptotic Normality of Quadratic Estimators)

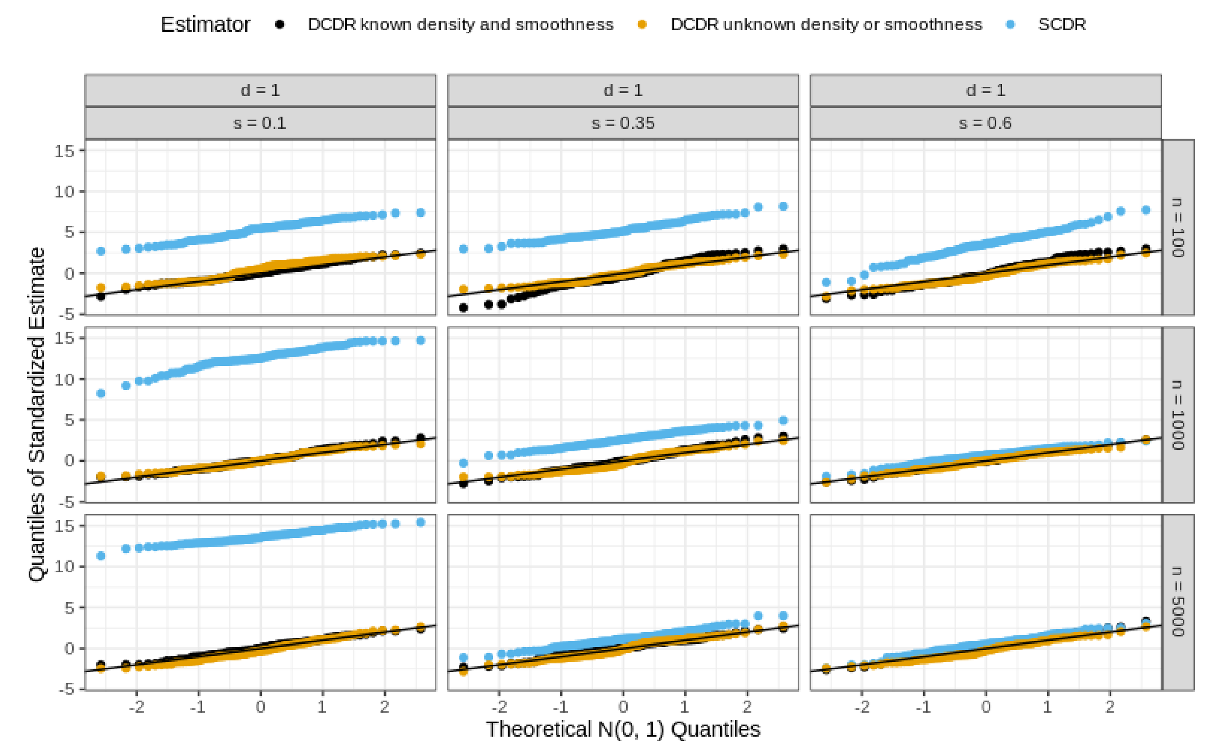
SIMULATIONS

To test our results, we constructed datasets with $X \sim Unif(0,1)$ and Hölder smooth nuisance functions using the lower-bound minimax construction (e.g., Tsybakov, 2009).

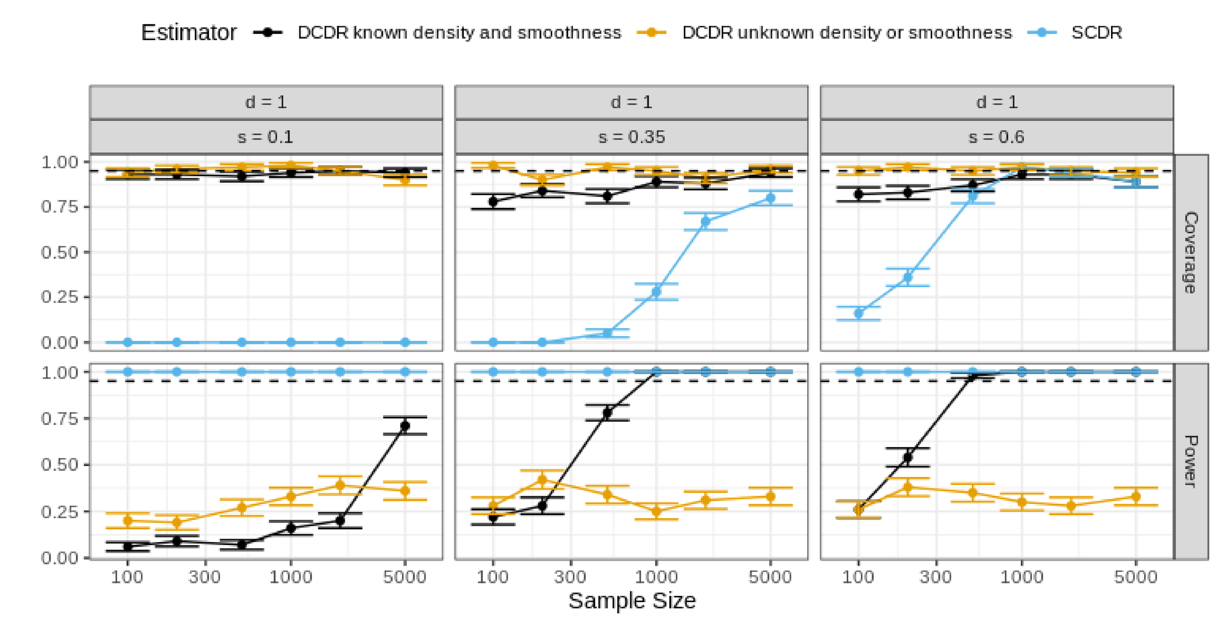
HÖLDER SMOOTH FUNCTIONS



QQ PLOTS



COVERAGE AND POWER



SIMULATION TAKEAWAYS AND CONTACT INFO

- Single split estimator no longer satisfies CLT for non-smooth nuisance functions ($s/d = 0.1$)
- DCDR estimators with known or unknown covariate density satisfy CLT for all smoothnesses
- Caveat: all estimators outperform theory because functions are only approximately Hölder smooth

alec@stat.cmu.edu, **alecmcclean.github.io**