On Doubly Robust Inference for Double Machine Learning in Semiparametric Regression (based on Dukes, Vansteelandt, Whitney, JMLR 2024)

Longtian Shi

Southern University of Science and Technology (SUSTech)

November 16, 2025

Outline

1. Problem statement & Motivation
2. Method: Drift Decomposition and Correction
3. Theoretical Results
4. Proof Sketch
5. Simulation (brief)

Problem: inference with ML-estimated nuisances

Partially linear model (PLM)

$$Y = \theta_0 A + m_0(L) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid A, L] = 0.$$

We want Valid inference (tests / CIs) for θ_0 while using flexible ML estimators for the nuisance functions $m_0(L) = \mathbb{E}[Y \mid A = 0, L]$ and $g_0(L) = \mathbb{E}[A \mid L]$.

- W denotes a random data following distribution P, then θ_0 is the unique solution to finite dimensional $\int \psi(w; \theta, \eta_0) dP(w) = 0$, where η_0 is the nuisance parameter that may be infinite dimensional estimated using semi-/non-parametric methods. Standard DML uses the orthogonal score $\psi(W; \theta, \eta) = \{A - g(L)\}\{Y - \theta A - m(L)\}$ with $0 = \frac{1}{n}\sum_{i=1}^{n} \psi(W_i; \theta, \hat{\eta})$ and cross-fitting to reduce overfitting bias. Let η^* be the limit of $\hat{\eta}$.
- $\mathbb{P}_n$ denotes the empirical measure; $Pf = \int f(x) dP(x)$ for a fixed f, so $P\psi(\theta_0, \eta_0) = 0$
- **Key issue:** How to conduct valid inference if we only have single consistency, e.g., $\hat{g} \to g_0$ but $\hat{m} \to m^* \neq m_0$?

Why standard DML inference can fail

Assume $V = -\partial P\psi(\theta, \eta)/\partial\theta|_{\theta=\theta_0}$ invertible. Supposing that $\hat{\eta}$ is obtained from an auxiliary sample, then one can show (see e.g. Theorem 5.31 of van der Vaart (2000)) that $\hat{\theta} - \theta_0 = V^{-1}(\mathbb{P}_n - P)\psi(\theta_0, \eta^*) + V^{-1}P\psi(\theta_0, \hat{\eta}) + o_P(n^{-1/2} + \|P\psi(\theta_0, \hat{\eta})\|)$, where $\|\cdot\|$ is the Euclidean norm. Therefore, the asymptotic behavior of $\hat{\theta}$ depends on $\hat{\eta}$ via the so-called 'drift' term $P\psi(\theta_0, \hat{\eta})$, which is the remainder from a linear expansion of $\hat{\theta}$, rather than the empirical term $(\mathbb{P}_n - P)\psi(\theta_0, \eta^*)$.

We are interested in estimators whose drift term can be written as $P\{d(\hat{g} - g_0)(\hat{m} - m_0)\}$ for some $d = d(W)$ that can be upper bounded, so the drift term is upper bounded by a term proportional to $||\hat{m} - m_0||_{P,2}||\hat{g} - g_0||_{P,2}$.

Consistency Holds but Inference Fails: The drift term's convergence is now dictated by the slower, consistent estimator. If $\|\hat{g} - g_0\|_{P,2} = o_P(n^{-\kappa})$ with $\kappa < 1/2$, the drift is also $o_P(n^{-\kappa})$.

- This is faster than $o_P(1)$ but slower than the required $o_P(n^{-1/2})$.
- No asymptotic linearity, invalidating standard variance estimators and normal approximations.

Dissecting the Drift Term

The first-order drift term, $P\psi(\theta_0, \hat{\eta})$, can be decomposed as:

$$P\psi(\theta_0, \hat{\eta}) = \underbrace{P\{d(\hat{g} - g^*)(\hat{m} - m^*)\}}_{R_1} + \underbrace{P\{d(\hat{g} - g^*)(m^* - m_0)\}}_{R_2} + \underbrace{P\{d(g^* - g_0)(\hat{m} - m^*)\}}_{R_3}$$

where $\hat{g} \to g^*$ and $\hat{m} \to m^*$.

- R_1: A second-order term, typically $o_P(n^{-1/2})$.
- R_2, R_3: Problematic first-order terms. If one model is correct, one of these vanishes.

Dissecting the Drift Term

The first-order drift term, $P\psi(\theta_0, \hat{\eta})$, can be decomposed as:

$$P\psi(\theta_0, \hat{\eta}) = \underbrace{P\{d(\hat{g} - g^*)(\hat{m} - m^*)\}}_{R_1} + \underbrace{P\{d(\hat{g} - g^*)(m^* - m_0)\}}_{R_2} + \underbrace{P\{d(g^* - g_0)(\hat{m} - m^*)\}}_{R_3}$$

where $\hat{g} \to g^*$ and $\hat{m} \to m^*$.

- R_1: A second-order term, typically $o_P(n^{-1/2})$.
- R_2, R_3: Problematic first-order terms. If one model is correct, one of these vanishes.

Assume $\hat{g} \to g_0$ (so $g^* = g_0$), but $\hat{m} \to m^* \neq m_0$ in PLM as an example,

$$\begin{aligned} R_2 =& \frac{1}{n}\sum_{n=1}^{n} \color{red}{\hat{G}(L_i)\{A_i - \hat{g}(L_i)\}} - (\mathbb{P}_n - P)\{G^*(A - g_0)\} - (\mathbb{P}_n - P)\{\hat{G}(A - \hat{g}) - G^*(A - g_0)\} \\ &+ P\{(\bar{G} - G^*)(g_0 - \hat{g}) + P\{(G^* - \hat{G})(g_0 - \hat{g})\}, \end{aligned}$$

where $G^*(L) = \mathbb{E}\{Y - \theta_0 A - m^*(L) | g^*(L)\}$ estimated by $\hat{G}(L)$ (later). If the empirical term is $o_p(n^{-1/2})$, then solving $0 = \frac{1}{n}\sum_{n=1}^{n}\{A_i - \hat{g}(L_i)\}\{Y_i - \theta_0 A_i - \hat{m}(L_i)\} - \hat{G}(L_i)\{A_i - \hat{g}(L_i)\}$ for estimating θ will yield a drift term upper bounded by $||\hat{G} - G^*||_{P,2}||\hat{g} - g_0||_{P,2}$.

Introducing Additional Nuisance Parameters

Key Insight

The problematic term R_2 or R_3 can be rewritten using iterated expectations.

Taking R_3 as an example, let's define a new nuisance parameter $M^*(L) := \mathbb{E}[A - g^*(L) | m_0(L)]$, which is a univariate regression of the residualized treatment $A - g^*(L)$ on the outcome model prediction $m_0(L)$. The term R_3 can be re-expressed and decomposed as:

$$R_3 \approx \underbrace{\mathbb{P}_n[\hat{M}(L)\{Y - \theta_0 A - \hat{m}(L)\}]}_{\text{Bias Correction}} - \underbrace{(\mathbb{P}_n - P)[M^*\{Y - \theta_0 A - m_0\}]}_{\text{Mean-Zero Linear Term}} + \text{rem}.$$

Introducing Additional Nuisance Parameters

Key Insight

The problematic term R_2 or R_3 can be rewritten using iterated expectations.

Taking R_3 as an example, let's define a new nuisance parameter $M^*(L) := \mathbb{E}[A - g^*(L) | m_0(L)]$, which is a univariate regression of the residualized treatment $A - g^*(L)$ on the outcome model prediction $m_0(L)$. The term R_3 can be re-expressed and decomposed as:

$$R_3 \approx \underbrace{\mathbb{P}_n[\hat{M}(L)\{Y - \theta_0 A - \hat{m}(L)\}]}_{\text{Bias Correction}} - \underbrace{(\mathbb{P}_n - P)[M^*\{Y - \theta_0 A - m_0\}]}_{\text{Mean-Zero Linear Term}} + \text{rem.}$$

If outcome model misspecified ($m^* \neq m_0$):

- Problematic term is R_2.
- We introduce:

$$G^*(L) := \mathbb{E}[Y - \theta_0 A - m^*(L) | g_0(L)]$$

If propensity score misspecified ($g^* \neq g_0$):

- Problematic term is R_3.
- We introduce:

$$M^*(L) := \mathbb{E}[A - g^*(L) | m_0(L)]$$

The Proposed Algorithm (1/2): Setup & Estimation

1. **Cross-Fitting Setup**: The method augments initial estimates $\hat{g}$ and $\hat{m}$ to annihilate the first-order bias. Split data into K folds, I_k. For each fold, train initial estimators $\hat{g}_k^c$ and $\hat{m}_k^c$ on the complement data I_k^c.

The Proposed Algorithm (1/2): Setup & Estimation

1. **Cross-Fitting Setup**: The method augments initial estimates $\hat{g}$ and $\hat{m}$ to annihilate the first-order bias. Split data into K folds, I_k. For each fold, train initial estimators $\hat{g}_k^c$ and $\hat{m}_k^c$ on the complement data I_k^c.
2. **Estimate Additional Nuisances**: On the main fold I_k, estimate $\tau^* = \{G^*, M^*\}$ using Nadaraya-Watson kernel regression.
   - For a point x in the support of $\hat{m}_k^c(L)$, define the kernel weights:

$$\varphi_j(L; x, h, \hat{\eta}) = K\left(\frac{x - \hat{m}_k^c(L)}{h}\right) \cdot \{A - \hat{g}_k^c(L)\}^{j-1}, \quad j = 1, 2$$

     where K is a kernel function (e.g., Gaussian) and $h > 0$ is a bandwidth.
   - Then, the NW estimators for $M^*(x)$ and similarly $G^*(x)$ is:

$$\hat{M}_k(x) = \frac{\frac{1}{n_k}\sum_{i \in I_k} \varphi_2(L_i; x, h, \hat{\eta})}{\frac{1}{n_k}\sum_{i \in I_k} \varphi_1(L_i; x, h, \hat{\eta})} = \frac{\sum_{i \in I_k} K\left(\frac{x - \hat{m}_k^c(L_i)}{h}\right) \cdot \{A_i - \hat{g}_k^c(L_i)\}}{\sum_{i \in I_k} K\left(\frac{x - \hat{m}_k^c(L_i)}{h}\right)}$$

$$\hat{G}_k(x) = \frac{\frac{1}{n_k}\sum_{i \in I_k} \rho_2(L_i; x, h, \hat{\eta})}{\frac{1}{n_k}\sum_{i \in I_k} \rho_1(L_i; x, h, \hat{\eta})} = \frac{\sum_{i \in I_k} K\left(\frac{x - \hat{g}_k^c(L_i)}{h}\right) \cdot \{Y_i - \theta_0 A_i - \hat{m}_k^c(L_i)\}}{\sum_{i \in I_k} K\left(\frac{x - \hat{g}_k^c(L_i)}{h}\right)}$$

The Proposed Algorithm (2/2): Correction & Final Score

3. **Update Nuisance Models**: Find scalar coefficients $\hat{\alpha}_k, \hat{\beta}_k$ via simple least squares on fold I_k to make the bias corrections orthogonal.

$$\sum_{i \in I_k} \hat{G}_k(L_i)\{A_i - \hat{g}_k^c(L_i) - \alpha \hat{G}_k(L_i)\} = 0 \implies \hat{\alpha}_k$$

$$\sum_{i \in I_k} \hat{M}_k(L_i)\{Y_i - \theta_0 A_i - \hat{m}_k^c(L_i) - \beta \hat{M}_k(L_i)\} = 0 \implies \hat{\beta}_k$$

The Proposed Algorithm (2/2): Correction & Final Score

3. **Update Nuisance Models**: Find scalar coefficients $\hat{\alpha}_k, \hat{\beta}_k$ via simple least squares on fold I_k to make the bias corrections orthogonal.

$$\sum_{i \in I_k} \hat{G}_k(L_i)\{A_i - \hat{g}_k^c(L_i) - \alpha \hat{G}_k(L_i)\} = 0 \implies \hat{\alpha}_k$$

$$\sum_{i \in I_k} \hat{M}_k(L_i)\{Y_i - \theta_0 A_i - \hat{m}_k^c(L_i) - \beta \hat{M}_k(L_i)\} = 0 \implies \hat{\beta}_k$$

4. **Final Score**: For each observation $i \in I_k$, the final, doubly robust score ψ^* is:

$$\psi^*(W_i; \theta_0, \hat{\eta}_k, \hat{\tau}_k) = \psi^*(W_i) := \underbrace{\{A_i - \tilde{g}_k(L_i)\}\{Y_i - \theta_0 A_i - \tilde{m}_k(L_i)\}}_{\text{Updated cross-moment}}$$
$$- \underbrace{\hat{G}_k(L_i)\{A_i - \tilde{g}_k(L_i)\}}_{\text{Correction for G}} - \underbrace{\hat{M}_k(L_i)\{Y_i - \theta_0 A_i - \tilde{m}_k(L_i)\}}_{\text{Correction for M}}$$

where $\tilde{g}_k = \hat{g}_k^c + \hat{\alpha}_k \hat{G}_k$ and $\tilde{m}_k = \hat{m}_k^c + \hat{\beta}_k \hat{M}_k$.

Main Result: Doubly Robust Asymptotic Linearity

Theorem (Theorem 1)

Under Assumptions 1-4, the proposed score statistic, based on ψ^*, is asymptotically linear. Specifically, the cross-fitted score average $\mathbb{P}_{n,k}\psi^*(\theta_0, \hat{\eta}_k^c, \hat{\tau}_k)$ can be expanded as:

$$\sqrt{n_k}\mathbb{P}_{n,k}\psi^*(\theta_0, \hat{\eta}_k^c, \hat{\tau}_k) = \sqrt{n_k}\mathbb{P}_{n,k}\psi_1(\theta_0, \eta^*, \tau^*) - I\{m^* = m_0\}\sqrt{n_k}\mathbb{P}_{n,k}\psi_2(\theta_0, \eta^*, \tau^*) \\ - I\{g^* = g_0\}\sqrt{n_k}\mathbb{P}_{n,k}\psi_3(\theta_0, \eta^*, \tau^*) + o_P(1)$$

where ψ_1, ψ_2, ψ_3 are fixed, mean-zero influence functions:

$$\psi_1(W; \theta_0, \eta, \tau) = \{A - g^*(L) - \alpha G^*(L)\}\{Y - \theta_0 A - m^*(L) - \beta M^*(L)\}$$
$$\psi_2(W; \theta_0, \eta, \tau) = M^*(L)\{Y - \theta_0 A - m^*(L) - \beta M^*(L)\}$$
$$\psi_3(W; \theta_0, \eta, \tau) = G^*(L)\{A - g^*(L) - \alpha G^*(L)\}$$

Main Result: Doubly Robust Asymptotic Linearity

Theorem (Theorem 1)

Under Assumptions 1-4, the proposed score statistic, based on ψ^*, is asymptotically linear. Specifically, the cross-fitted score average $\mathbb{P}_{n,k}\psi^*(\theta_0, \hat{\eta}_k^c, \hat{\tau}_k)$ can be expanded as:

$$\begin{aligned}\sqrt{n_k}\mathbb{P}_{n,k}\psi^*(\theta_0, \hat{\eta}_k^c, \hat{\tau}_k) = \sqrt{n_k}\mathbb{P}_{n,k}\psi_1(\theta_0, \eta^*, \tau^*) - I\{m^* = m_0\}\sqrt{n_k}\mathbb{P}_{n,k}\psi_2(\theta_0, \eta^*, \tau^*) \\ - I\{g^* = g_0\}\sqrt{n_k}\mathbb{P}_{n,k}\psi_3(\theta_0, \eta^*, \tau^*) + o_P(1)\end{aligned}$$

where ψ_1, ψ_2, ψ_3 are fixed, mean-zero influence functions:

$$\begin{aligned}\psi_1(W; \theta_0, \eta, \tau) &= \{A - g^*(L) - \alpha G^*(L)\}\{Y - \theta_0 A - m^*(L) - \beta M^*(L)\} \\ \psi_2(W; \theta_0, \eta, \tau) &= M^*(L)\{Y - \theta_0 A - m^*(L) - \beta M^*(L)\} \\ \psi_3(W; \theta_0, \eta, \tau) &= G^*(L)\{A - g^*(L) - \alpha G^*(L)\}\end{aligned}$$

This theorem ensures that the final estimator is asymptotically normal **if *either* the outcome model ($\hat{m}$) *or* the propensity score model ($\hat{g}$) is consistent,** achieving the double robustness.

Convergence Rates of Auxiliary Estimators

Theorem (Theorem 4: The convergence rate of $\hat{M}(x)$ (and $\hat{G}(x)$).)

$\hat{M}(x)$ is the NW estimator of $M^*(x) = \mathbb{E}[A - g^*(L) | m^*(L) = x]$, with $\hat{g}$ and $\hat{m}$ estimated on an auxiliary sample. Under standard kernel assumptions and conditions needed on empirical processes (e.g. $||g_0 - g^*||_{P,2} = O(1)$), kernel K is of VC-type with $\xi \geq e$, $\nu \geq 1$), then:

$$|\hat{M}(x) - M^*(x)| = O_p(h^{\vartheta} + \zeta_g + h^{-1}\zeta_m); \quad \mathbb{E}[\{\hat{M}(x) - M^*(x)\}^2] = O(h^{2\vartheta} + \zeta_g^2 + h^{-2}\zeta_m^2).$$

Here ϑ is the kernel order, h is the bandwidth ($h \to 0$, $nh^3 \to \infty$), and $\|g^* - \hat{g}^c\|_{P,2} = O_P(\zeta_g)$, and $\|m^* - \hat{m}^c\|_{P,2} = O_P(\zeta_m)$.

Convergence Rates of Auxiliary Estimators

Theorem (Theorem 4: The convergence rate of $\hat{M}(x)$ (and $\hat{G}(x)$).)

$\hat{M}(x)$ is the NW estimator of $M^*(x) = \mathbb{E}[A - g^*(L) | m^*(L) = x]$, with $\hat{g}$ and $\hat{m}$ estimated on an auxiliary sample. Under standard kernel assumptions and conditions needed on empirical processes (e.g. $||g_0 - g^*||_{P,2} = O(1)$), kernel K is of VC-type with $\xi \geq e$, $\nu \geq 1$), then:

$$|\hat{M}(x) - M^*(x)| = O_p(h^{\vartheta} + \zeta_g + h^{-1}\zeta_m); \quad \mathbb{E}[\{\hat{M}(x) - M^*(x)\}^2] = O(h^{2\vartheta} + \zeta_g^2 + h^{-2}\zeta_m^2).$$

Here ϑ is the kernel order, h is the bandwidth ($h \to 0$, $nh^3 \to \infty$), and $\|g^* - \hat{g}^c\|_{P,2} = O_P(\zeta_g)$, and $\|m^* - \hat{m}^c\|_{P,2} = O_P(\zeta_m)$.

- The convergence of $\hat{M}$ depends on the rates of **both** initial estimators, $\hat{g}$ and $\hat{m}$.
- This suggests that the optimal bandwidth for $\hat{M}(L)$ will now depend on the convergence rate of $\hat{m}(L)$, as well as using a larger bandwidth (undersmoothing) when $\hat{m}$ converges slowly.
- At least from the high quality estimation of M^* or G^*, one might prefer an inconsistent but quickly convergent estimator $\hat{g}$ rather than one that converges slowly to the truth.

Proof Sketch of Theorem 1 (Part 1/2)

1. **Decompose asymptotic linearity of the score statistic under single consistency:**

$$\sqrt{n_k}\left[\mathbb{P}_{n,k}\psi_1(\theta_0,\hat{\eta}_k^c,\hat{\tau}_k)-\mathbb{P}_{n,k}\psi_1(\theta_0,\eta^*,\tau^*)\right] = \underbrace{\mathbb{G}_{n,k}\left[\psi_1(\theta_0,\hat{\eta}_k^c,\hat{\tau}_k)-\psi_1(\theta_0,\eta^*,\tau^*)\right]}_{\mathcal{I}_1} + \underbrace{\sqrt{n_k}P\left[\psi_1(\theta_0,\hat{\eta}_k^c,\hat{\tau}_k)-\psi_1(\theta_0,\eta^*,\tau^*)\right]}_{\mathcal{I}_2},$$

where $\mathbb{G}_{n,k}=\sqrt{n_k}(\mathbb{P}_{n,k}-P)$.

2. **Controlling the Empirical Process Term $\mathcal{I}_1$:**
   - By sample splitting, $\hat{\eta}_k^c$ is fixed when conditioning on I_k^c
   - Consider the function class: $\mathcal{F}=\{\psi_1(\cdot;\hat{\eta}_k^c,\tau)-\psi_1(\cdot;\eta^*,\tau^*):\tau\in\mathcal{T}\}$.
   - Under Assumption 4, $\mathcal{F}$ is VC-type with bounded entropy:

$$\sup_Q \log N(\epsilon\|F\|_{Q,2},\mathcal{F},L_2(Q))\le \nu\log(\xi/\epsilon),\quad \xi\ge e,\nu\ge 1$$

   - Apply maximal inequality (Chernozhukov et al., 2014):

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_{n,k}(f)|\,\middle|\, I_k^c\right]\lesssim J(1,\mathcal{F},L_2)+n_k^{-1/2}J^2(1,\mathcal{F},L_2)$$

   - Entropy conditions ensure $J(1,\mathcal{F},L_2)<\infty$, hence $\mathcal{I}_1=o_P(1)$

Proof Sketch of Theorem 1 (Part 2/2)

3. **Analyzing the Bias Term $\mathcal{I}_2$:**
   - Expand using the product structure of the estimating function:

$$\mathcal{I}_2 = \sqrt{n_k} P \left[(\hat{g}_k^c - g^* + \alpha^* G^* - \hat{\alpha}_k \hat{G}_k)(\hat{m}_k^c - m^* + \beta^* M^* - \hat{\beta}_k \hat{M}_k) \right]$$

   - **Case 1: Both models consistent** ($g^* = g_0$, $m^* = m_0$)
     Standard DML case. By Cauchy-Schwarz and Assumption 3:

$$|\mathcal{I}_2| \leq \sqrt{n_k} \|\hat{g}_k^c - g_0\|_{P,2} \|\hat{m}_k^c - m_0\|_{P,2} = o_P(1)$$

   - **Case 2: Outcome or Propensity model misspecified** (e.g., if $g^* = g_0$, $m^* \neq m_0$)
     - ★ $G^* = \beta^* = 0$. Then $\mathcal{I}_2 = \sqrt{n_k} P \left[(\hat{g}_k^c - g_0)(\hat{m}_k^c - m_0 - \hat{\beta}_k \hat{M}_k) \right] + o_P(1)$
     - ★ Remainder is $O_P(\sqrt{n_k} \|\hat{g}_k^c - g_0\|_{P,2} \|\hat{G}_k - G^*\|_{P,2}) = o_P(1)$

Both $\mathcal{I}_1$ and $\mathcal{I}_2$ are $o_P(1)$,

$$\sqrt{n_k} \mathbb{P}_{n,k} \psi^*(\theta_0, \hat{\eta}_k^c, \hat{\tau}_k) = \sqrt{n_k} \mathbb{P}_{n,k} \psi^*(\theta_0, \eta^*, \tau^*) + o_P(1)$$

The right-hand side is a sum of i.i.d. mean-zero random variables, yielding asymptotic linearity.

Simulation design (summary)

- **Objective:** compare coverage, bias, RMSE, and test size of:
  - naive plug-in estimator,
  - standard DML (orthogonal score),
  - proposed corrected-score DR inference.
- **DGP:** PLM with controlled smoothness / sparsity to vary ML convergence rates; alternative scenarios where only m or g is correctly specified.
- **Estimation choices:** ML learners for $\hat{m}, \hat{g}$ (e.g. RF, GBM, LASSO depending on setting); G, M estimated with Nadaraya–Watson (univariate); cross-fitting with $K = 5$.
- **Metrics:** empirical bias, RMSE, 95% CI coverage, empirical size and power for tests.
- **Implementation note:** tune bandwidth h by cross-validation (on the univariate smoother), but avoid severe undersmoothing unless needed to control bias term.

Thank you!