Automatic doubly robust inference via calibration

Lars van der Laan, Alex Luedtke, Marco Carone

University of Washington, Department of Statistics

Summary

- **Doubly robust estimators** are widely used for estimating causal effects.
- **Consistency** requires only ***one*** nuisance function to be estimated well, but **asymptotic normality** typically requires ***both***. Inference is not doubly robust!
- To correct this mismatch, **we propose *calibrated DML***, providing doubly robust inference **by calibrating nuisance estimators.**

Calibrated DML Procedure

Estimate Nuisances → Calibrate → Debias

Calibrating the nuisances before debiasing ensures doubly robust asymptotic normality.

- ***Isotonic calibrated DML as special case***: standard DML augmented with a simple, tuning-free **post-hoc calibration** step using isotonic regression of labels on cross-fitted nuisance estimates.
- **Asymptotic normality** for linear functionals holds if ***either*** the outcome regression ***or*** Riesz representer (e.g., propensity score) is estimated well.
- A **bootstrap procedure** enables valid inference without extra nuisance estimation.

Properties of Calibrated DML

Estimator	PS only			OR only			Both		
	Cons.	Norm.	Eff.	Cons.	Norm.	Eff.	Cons.	Norm.	Eff.
G-comp	–	–	–	✓	×	×	–	–	–
IPW	✓	×	×	–	–	–	–	–	–
AIPW	✓	×	×	✓	×	×	✓	✓	✓
Cal. DML (ours)	✓	✓	✓*	✓	✓	✓*	✓	✓	✓

Properties of calibrated DML for ATE under correct specification of PS and/or OR.
PS: Propensity score. **OR**: Outcome regression. **Cons.**: Consistent. **Norm.**: Asym. normal. **Eff.**: Efficient.

Background: DML for linear functionals

- **Data structure:** $Z = (W, A, Y) \sim P_0$, where W is a vector of covariates, A is a treatment assignment, and Y is a real-valued outcome.
- **Target parameter:** a linear functional of the outcome regression.

$$\tau_0 := E_0[m(Z, \mu_0)] \quad \text{where} \quad \mu_0(a, w) := E_0[Y \mid A = a, W = w],$$

with $\mu \mapsto m(z, \mu)$ linear. For example, the ATE corresponds to $m(z, \mu) := \mu(1, w) - \mu(0, w)$.

- **Key fact:** there exists a **Riesz representer** α_0 such that

$$\tau_0 = E_0[\alpha_0(A, W)Y] \quad \text{(weighted average of the outcome).}$$

Debiased Machine Learning (DML)

Obtain estimators μ_n of μ_0 and α_n of α_0, and compute:

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \mu_n) + \frac{1}{n}\sum_{i=1}^{n} \alpha_n(A_i, W_i)\{Y_i - \mu_n(A_i, W_i)\}.$$

DML is *rate doubly robust*:

(i) **Consistent** if $\|\mu_n - \mu_0\| = o_p(1)$ or $\|\alpha_n - \alpha_0\| = o_p(1)$.

(ii) **Asymptotically normal** if $\|\mu_n - \mu_0\| \cdot \|\alpha_n - \alpha_0\| = o_p(n^{-1/2})$.

Objective: doubly robust inference

- **Our goal:** construct estimators that are **doubly robust asymptotically normal**.
  - Valid inference—e.g., confidence intervals and hypothesis testing—even if only one nuisance function is estimated well.
- **Formally:** $\sqrt{n}(\tau_n - \tau_0) \xrightarrow{d} N(0, \sigma_0^2)$ holds if *any* of the following:
  1. $\|\mu_n - \mu_0\| = o_p(n^{-1/4})$
  2. $\|\alpha_n - \alpha_0\| = o_p(n^{-1/4})$
  3. $\|\mu_n - \mu_0\| \cdot \|\alpha_n - \alpha_0\| = o_p(n^{-1/2})$

Nuisance calibration implies doubly robust inference

- We discover a link between doubly robust inference and model calibration—a technique typically used in prediction and classification.
- A predictor/model $f(\cdot)$ is empirically calibrated with respect to a loss $\ell(z, f)$ if its empirical risk cannot be improved by any transformation of its predictions:

$$\sum_{i=1}^{n} \ell(Z_i, f) = \min_{\theta} \sum_{i=1}^{n} \ell(Z_i, \theta \circ f).$$

Key Finding

Suppose nuisance estimators μ_n and α_n are empirically calibrated:

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \mu_n(A_i, W_i)\}^2 = \min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \theta(\mu_n(A_i, W_i))\}^2$$

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\alpha_n(A_i, W_i)^2 - 2\,m(Z_i, \alpha_n)\right\} = \min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\left\{\theta(\alpha_n(A_i, W_i))^2 - 2\,m(Z_i, \theta \circ \alpha_n)\right\}.$$

Then, DML is debiased and asymptotically normal even if one nuisance estimator is poorly estimated.

Calibration improves stability and quality of nuisances

- Calibration of the outcome regression implies unbiasedness:

$$\mu_n(a, w) = \frac{\sum_{i=1}^{n} \mathbf{1}\{\mu_n(A_i, W_i) = \mu_n(a, w)\}\, Y_i}{\sum_{i=1}^{n} \mathbf{1}\{\mu_n(A_i, W_i) = \mu_n(a, w)\}},$$

ensuring that the regression predictions do not systematically over- or under-estimate observed outcomes on average.

- Calibration of (inverse) propensity scores implies balance:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{\pi_n(W_i)} f(\pi_n(W_i)) = \frac{1}{n}\sum_{i=1}^{n} f(\pi_n(W_i)) \quad \text{for all } f,$$

ensuring that large inverse propensity weights meaningfully contribute to balance, rather than inflating variance without reducing bias.

How to Calibrate?

- **Post-hoc calibration** adjusts a model $f(\cdot)$ by minimizing empirical loss ℓ over a class of transformations applied to its outputs.
- **Histogram binning:** discretize the range of $f(\cdot)$, and within each bin, assign the prediction that minimizes empirical risk. This learns a piecewise constant transformation.
- **Isotonic regression:** a data-adaptive, tuning-free binning method that fits an optimal *monotone* transformation of the model's predictions.

Calibrated DML using isotonic calibration

Algorithm 1 Calibrated DML using isotonic calibration

Input: Dataset $\mathcal{D}_n = \{O_i : i = 1, \ldots, n\}$; number J of cross-fitting splits

1: Partition $\mathcal{D}_n$ into folds $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(J)}$
2: **for** $s = 1, \ldots, J$ **do**
3: Fit initial estimators $\mu_{n,s}, \alpha_{n,s}$ on $\mathcal{E}^{(s)} := \mathcal{D}_n \setminus \mathcal{C}^{(s)}$ **(cross-fitting)**
4: **end for**
5: Define $j(i) := s$ for $i \in \mathcal{C}^{(s)}$ (indicates fold membership)
6: **Fit calibrators using isotonic regression** with $\mathcal{F}_{iso}$ the class of monotone (↑) functions:

$$f_n \in \operatorname*{argmin}_{f \in \mathcal{F}_{\text{iso}}} \sum_{i=1}^{n} (Y_i - f(\mu_{n,j(i)}(A_i, W_i)))^2$$

$$g_n \in \operatorname*{argmin}_{g \in \mathcal{F}_{\text{iso}}} \sum_{i=1}^{n} \left[g(\alpha_{n,j(i)}(A_i, W_i))^2 - 2\, m(Z_i, g \circ \alpha_{n,j(i)}) \right]$$

7: Set calibrated estimators: $\mu^*_{n,s} := f_n \circ \mu_{n,s}$, $\alpha^*_{n,s} := g_n \circ \alpha_{n,s}$
8: **Compute calibrated estimator:**

$$\tau^*_n := \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \mu^*_{n,j(i)}) + \frac{1}{n}\sum_{i=1}^{n} \alpha^*_{n,j(i)}(A_i, W_i)(Y_i - \mu^*_{n,j(i)}(A_i, W_i))$$

9: **(Optional)** Bootstrap empirical means in computation of f_n, g_n, and τ^*_n to construct CIs.
10: **return** calibrated DML estimate τ^*_n and CI based on influence function or bootstrap.

Theory for Calibrated DML

- Define errors: $\Delta_{\mu,n,j} := \mu^*_{n,j} - \mu_0$, $\Delta_{\alpha,n,j} := \alpha^*_{n,j} - \alpha_0$.
- For a summary map $\varphi : \mathcal{W} \times \mathcal{A} \to \mathbb{R}$, define the projection

$$\Pi_\varphi f := \theta_f \circ \varphi, \quad \text{where} \quad \theta_f := \arg\min_{\theta} \|f - \theta \circ \varphi\|.$$

Note $\Pi_\varphi f(w, a) = E_0[f(W, A) \mid \varphi(W, A) = \varphi(w, a)]$.

Assumptions:

(i) *(Both converge to something)* $\|\mu_{n,j} - \overline{\mu}_0\| + \|\alpha_{n,j} - \overline{\alpha}_0\| = o_p(1)$ for some $\overline{\mu}_0, \overline{\alpha}_0$

(ii) *(At least one converges fast enough)* $\|\mu^*_{n,j} - \mu_0\| \wedge \|\alpha^*_{n,j} - \alpha_0\| = o_p(n^{-1/4})$

(iii) *(Error coupling for projections)* $\|(\Pi_{\mu^*_{n,j}} - \Pi_{\mu_0})\Delta_{\alpha,n,j}\| = O_p(\|\mu^*_{n,j} - \mu_0\|)$ and $\|(\Pi_{\alpha^*_{n,j}} - \Pi_{\alpha_0})\Delta_{\mu,n,j}\| = O_p(\|\alpha^*_{n,j} - \alpha_0\|)$

Doubly Robust Asymptotic Linearity

Under these conditions, we have $\tau^*_n - \tau_0 = P_n\chi_0 + o_p(n^{-1/2})$, where:

$$\chi_0(z) = \underbrace{m(z, \overline{\mu}_0) - P_0 m(z, \overline{\mu}_0) + \overline{\alpha}_0(a, w)\{y - \overline{\mu}_0(a, w)\}}_{\text{Usual IF at misspecified limits}}$$

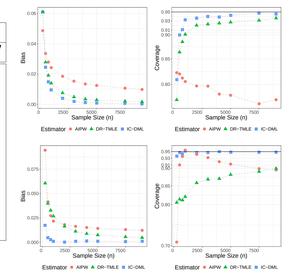
$$+ \underbrace{\mathbf{1}\{\overline{\alpha}_0 \neq \alpha_0\} \cdot s_0(a, w)\{y - \mu_0(a, w)\}}_{\text{From representer misspecification}} \qquad s_0(a, w) = \Pi_{\mu_0}(\alpha_0 - \overline{\alpha}_0)$$

$$+ \underbrace{\mathbf{1}\{\overline{\mu}_0 \neq \mu_0\} \cdot (m(z, r_0) - r_0(a, w)\alpha_0(a, w))}_{\text{From outcome misspecification}} \qquad r_0(a, w) = \Pi_{\alpha_0}(\mu_0 - \overline{\mu}_0)$$

Benchmarking on semi-synthetic data

(Left) Evaluation of AIPW vs. calibrated DML for ATE on semi-synthetic benchmarks, with both outcome regression and propensity scores estimated using gradient-boosted trees. (Right) Plots of bias and coverage in simulation studies.

Dataset	Bias		RMSE		Coverage	
	calDML	AIPW	calDML	AIPW	calDML	AIPW
ACIC-2017 (18)	0.28	**0.21**	**0.58**	0.70	**0.64**	0.27
ACIC-2017 (20)	**0.20**	1.6	**1.4**	2.0	**0.90**	0.32
ACIC-2017 (22)	0.035	**0.004**	**0.10**	0.11	**0.81**	0.56
ACIC-2017 (24)	**0.04**	0.30	**0.25**	0.35	**0.90**	0.32
ACIC-2018 (Aggr)	**7.1**	9.0	110	**97**	**0.69**	0.58
IHDP	0.13	**0.13**	0.46	**0.46**	**0.57**	**0.57**
Lalonde CPS	**0.084**	0.14	0.34	**0.22**	**0.75**	0.16
Lalonde PSID	0.039	**0.038**	0.44	**0.19**	**0.84**	0.46
Twins	**0.21**	0.22	**0.23**	0.24	**0.54**	0.51



doi: 10.48550/arXiv.2411.02771

Twitter: @LarsvanderLaan3

ACIC 2025

lvdlaan@uw.edu