Towards Trustworthy Statistical Inference with Black-Box AI Predictions

Jiwei Zhao, PhD
Associate Professor
Department of Statistics
Department of Biostatistics & Medical Informatics
University of Wisconsin-Madison

Professional Development Course/CE, JSM 2026
Thomas M. Menino Convention & Exhibition Center, CC-153C
8/2/2026

Outline

1. Introduction and Motivation [830-845am]
2. Prediction Powered Inference [845-930am]
3. with Multiple AI Predictions [930-1015am]
4. Break [1015-1030am]
5. under Covariate Shift [1030-1115am]
6. under Label Shift [1115-1215pm]
7. Take-Home Messages [1215-1230pm]

Outline

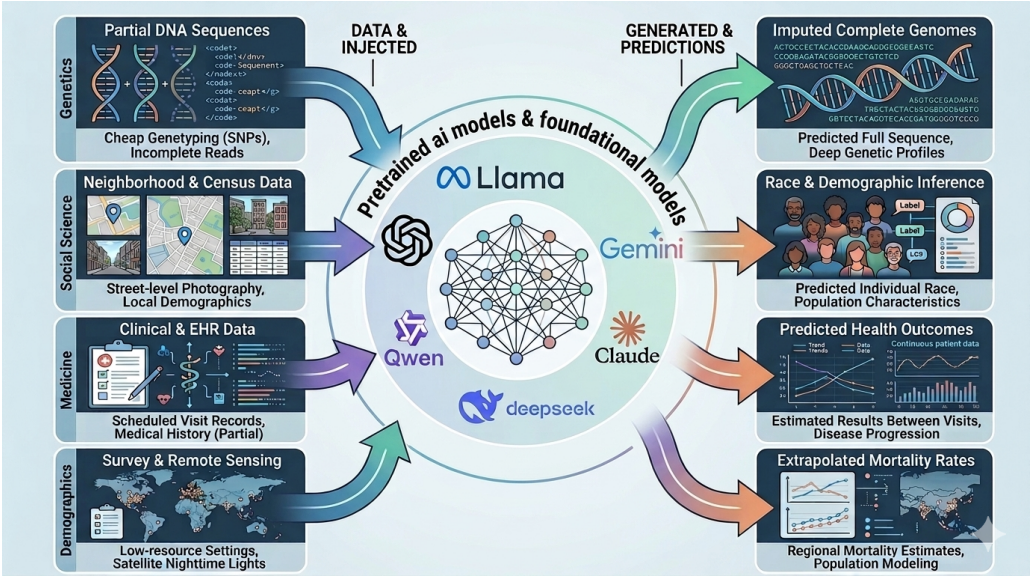
Big Picture

Researchers increasingly use predictions from pre-trained algorithms as stand-ins for hard-to-collect and expensive samples, as AI/ML tools become more accessible and scientists face new obstacles to data collection (e.g. rising costs, declining survey response rates)[1]

- In genetics, phenotypes based on cheap partial genome sequences stand in for more expensive complete sequences.
- Social scientists use census data and street level photography to predict an individual's race based on the composition of their neighborhood.
- Pragmatic trials vastly reduce costs by predicting outcomes that may have occurred between regularly scheduled clinic visits.
- Demographers in low-resource settings extrapolate difficult to collect data on mortality based on small surveys and easy-to-obtain measures like light intensity.

[1] Hoffman, K., Salerno, S., Afiaz, A., Leek, J., & McCormick, T. (2024). Do we really even need data?. arXiv preprint arXiv:2401.08702.

Big Picture (AI Version)



A Scientist's Dilemma

- Gold-standard experimental data are scarce: manual chart review, wet-lab assays, field visits, human annotation…
- ML predictions are cheap and abundant: AlphaFold structures, transformer gene-expression models, ResNet image classifiers, LLMs…
- How should a scientist use black-box AI predictions to draw statistically valid conclusions?

Can we get tighter inference than labeled-only,
with valid coverage guaranteed?

Two Tempting But Flawed Approaches

Classical approach

Ignore the predictions; use labeled data only.
Wasteful: throws away abundant unlabeled data; wide CIs, low power.

Imputation approach

Treat predictions $\widehat{Y}$ as if they were ground truth Y.
Invalid: ML errors $\Rightarrow$ biased point estimates, intervals that miss the truth.

Neither “trust the black box” nor “ignore it” is satisfactory.

Outline

Prediction-Powered Inference (PPI)

Angelopoulos, Bates, Fannjiang, Jordan, Zrnic (2023). "*Prediction-Powered Inference*," *Science* 382:669–674.

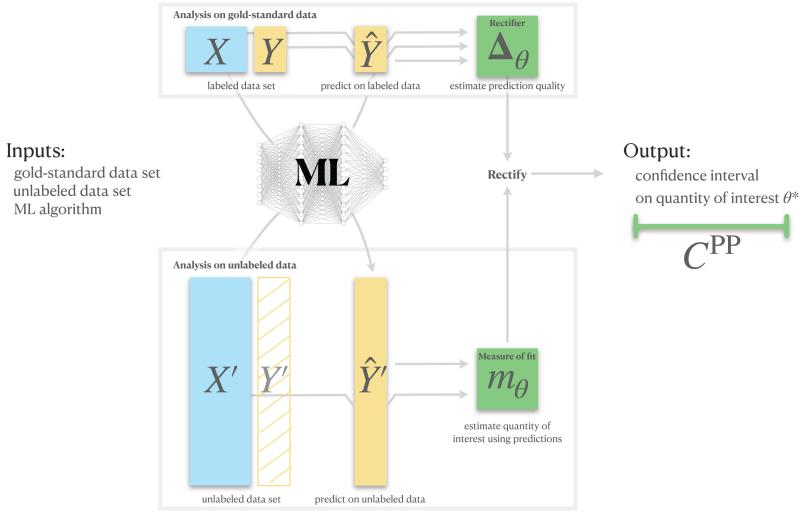
- Combines gold-standard data with ML predictions in a provably valid way.
- Produces CIs that are narrower than classical whenever predictions are reasonably accurate.
- Works for *any* ML algorithm, *any* underlying distribution.
- Applies to means, quantiles, regression coefficients, ... any convex M-estimator.

Data Structure and Target

- Labeled (gold-standard) sample: $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} \mathbb{P}$.
- Unlabeled sample: $(X'_1, \ldots, X'_N)$ with true but unseen $(Y'_1, \ldots, Y'_N)$, also from $\mathbb{P}$.
- An ML model produces predictions $\widehat{Y}_i$ on labeled and $\widehat{Y}'_i$ on unlabeled features.
- Regime of interest: $N \gg n$ with $n/N \to p \in (0, 1)$.
- Target estimand θ^*: the minimizer of a convex population risk,

$$\theta^* \;=\; \text{argmin}_\theta \; \mathrm{E}[L_\theta(X, Y)].$$

PPI: Protocol



[Image from Angelopoulos et al. 2023]

The Three-Ingredient Protocol

1. Estimand θ^*.

2. Measure of fit m_θ: on unlabeled data, using $\widehat{Y}'$, quantifies how consistent θ is with the *imputed* data.

3. Rectifier Δ_θ: on labeled data, captures how much ML prediction bias there is.

Prediction-powered confidence set:

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta : |m_\theta + \Delta_\theta| \leq w_\theta(\alpha)\}$$

If predictions are *perfect*, $\Delta_\theta \equiv 0$ and m_θ alone gives the CI on the combined dataset.

Width of the Confidence Set

For each coordinate j,

$$w_{\theta,j}(\alpha) \;=\; z_{1-\alpha/(2p)} \sqrt{\frac{\widehat{\sigma}^2_{\Delta_{\theta,j}}}{n} \;+\; \frac{\widehat{\sigma}^2_{m_{\theta,j}}}{N}}\,.$$

- Δ-variance $(1/n)$ $+$ m-variance $(1/N)$: both CLT terms.
- Large N crushes the m-term; remaining term dominated by rectifier noise.
- The better the ML model, the smaller var(Δ), hence the tighter the CI.

Example 1: Outcome Mean

Target: $\theta^* = \mathrm{E}[Y]$.

- Measure of fit $m_\theta = \theta - \frac{1}{N}\sum_{j=1}^{N} \widehat{Y}'_j$.
- Rectifier $\Delta_\theta = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)$.

$$\widehat{\theta}^{\mathrm{PP}} = \underbrace{\frac{1}{N}\sum_{j=1}^{N} \widehat{Y}'_j}_{\text{imputation on } \mathcal{U}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)}_{\text{bias correction on } \mathcal{L}}$$

Classical CI on Δ + classical CI on m, combined.

Example 2: q-Quantile

Target: $\theta^* = F_Y^{-1}(q)$.

- $m_\theta = -q + \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\{\widehat{Y}'_j \le \theta\}$
- $\Delta_\theta = \frac{1}{n}\sum_{i=1}^{n}\{\mathbf{1}(Y_i \le \theta) - \mathbf{1}(\widehat{Y}_i \le \theta)\}$

Median is a special case ($q = 1/2$); same recipe delivers valid CIs.

Example 3: Linear & Logistic Regression

Linear (OLS): $\theta^* = \text{argmin}_\theta \text{E}[(Y - X^{\text{T}}\theta)^2]$

$$m_\theta = \frac{1}{N}\sum_{j=1}^{N} X_j' {X_j'}^{\text{T}}\theta - \frac{1}{N}\sum_{j=1}^{N} X_j'\widehat{Y}_j' \qquad \Delta_\theta = \frac{1}{n}\sum_{i=1}^{n} X_i(\widehat{Y}_i - Y_i)$$

Logistic: $\theta^* = \text{argmin}_\theta \text{E}\big[-Y\theta^{\text{T}}X + \log(1 + e^{\theta^{\text{T}}X})\big]$

$$m_\theta = \frac{1}{N}\sum_{j=1}^{N} X_j'\{\mu_\theta(X_j') - \widehat{Y}_j'\} \qquad \Delta_\theta = \frac{1}{n}\sum_{i=1}^{N} X_i(\widehat{Y}_i - Y_i)$$

with $\mu_\theta(x) = 1/(1 + e^{-x^{\text{T}}\theta})$.

The Master Protocol: Convex M-Estimators

For any convex loss L_θ with (sub)gradient ∇L_θ:

$$m_\theta = \frac{1}{N}\sum_{j=1}^{N} \nabla L_\theta(X'_j, \widehat{Y}'_j)$$
$$\Delta_\theta = \frac{1}{n}\sum_{i=1}^{n} \{\nabla L_\theta(X_i, Y_i) - \nabla L_\theta(X_i, \widehat{Y}_i)\}$$

- Unifies mean, quantile, OLS, logistic, quantile regression, ...
- θ^* is a root of $\mathrm{E}[\nabla L_{\theta^*}(X, Y)] = 0$ (the non-degenerate case).
- Cover all rows of Table 1 of the paper.

PPI: Examples

Table 1. Prediction-powered inference for common statistical problems. Given a measure of fit m_θ and rectifier Δ_θ, prediction-powered inference computes a confidence interval as $C^{PP} = \{\theta \text{ such that } |m_\theta + \Delta_\theta| \leq w_\theta(\alpha)\}$, where $w_\theta(\alpha)$ is a constant that depends on the error level α (see Theorem S1 in the SM). Algorithms S1 to S6 are stated in the SM. The last row ("convex minimizer") refers to a method that generalizes the methods in previous rows.

Estimand	Measure of fit m_θ	Rectifier Δ_θ	Procedure
Mean outcome	$\theta - \frac{1}{N}\sum_{i=1}^{N} \hat{Y}'_i$	$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)$	Alg. S1
Median outcome	$\frac{1}{2N}\sum_{i=1}^{N} \text{sign}\left(\theta - \hat{Y}'_i\right)$	$\frac{1}{n}\sum_{i=1}^{n}\left(1\{Y_i \leq \theta\} - 1\left\{\hat{Y}_i \leq \theta\right\}\right)$	Alg. S2
q-quantile of outcome	$-q + \frac{1}{N}\sum_{i=1}^{N} 1\left\{\hat{Y}'_i \leq \theta\right\}$	$\frac{1}{n}\sum_{i=1}^{n}\left(1\{Y_i \leq \theta\} - 1\left\{\hat{Y}_i \leq \theta\right\}\right)$	Alg. S3
Linear regression	$\theta - (X')^+\hat{Y}'$	$X^+(\hat{Y} - Y)$	Alg. S4
Logistic regression	$\frac{1}{N}\sum_{i=1}^{N} X'_i\left(\frac{1}{1+e^{-\theta^T X'_i}} - \hat{Y}'_i\right)$	$\frac{1}{n}\sum_{i=1}^{n} X_i\left(\hat{Y}_i - Y_i\right)$	Alg. S5
Convex minimizer	$\frac{1}{N}\sum_{i=1}^{N} \nabla L_\theta\left(X'_i, \hat{Y}'_i\right)$	$\frac{1}{n}\sum_{i=1}^{n}\left(\nabla L_\theta\left(X_i, \hat{Y}_i\right) - \nabla L_\theta(X_i, Y_i)\right)$	Alg. S6

[Image from Angelopoulos et al. 2023]

Theorem (Coverage, SM Theorem S1)

Under iid sampling, convex L_θ, non-degeneracy $\mathrm{E}[\nabla L_{\theta^*}(X, Y)] = \mathbf{0}$, and $n/N \to p \in (0, 1)$:

$$\liminf_{n,N\to\infty} \Pr\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1 - \alpha.$$

- Valid for any ML algorithm (AlphaFold, GPT-4, XGBoost, ...).
- Valid under any data-generating distribution.
- Non-degeneracy holds whenever L is differentiable (or non-differentiable only on a measure-zero set).

Proof Sketch

1. CLT on the labeled side: $\sqrt{n}\,(\Delta_{\theta^*} - \mathrm{E}\Delta_{\theta^*}) \Rightarrow \mathcal{N}(0, \sigma_\Delta^2)$.
2. CLT on the unlabeled side: $\sqrt{N}\,(m_{\theta^*} - \mathrm{E}m_{\theta^*}) \Rightarrow \mathcal{N}(0, \sigma_m^2)$.
3. Independence across samples $\Rightarrow$ independent CLTs.
4. Non-degeneracy $\Rightarrow \mathrm{E}[m_{\theta^*} + \Delta_{\theta^*}] = 0$.
5. Union bound across coordinates for vector θ.

Bad predictions don't hurt validity — they only *inflate width*.

When Does PPI Beat Classical?

For the mean case (after some algebra):

$$\frac{1}{N}\text{var}(\widehat{Y}') \;+\; \frac{1}{n}\text{var}(\widehat{Y} - Y) \;<\; \frac{1}{n}\text{var}(Y).$$

When $N \gg n$, this reduces to

$$\boxed{\text{var}(\widehat{Y} - Y) \;<\; \text{var}(Y).}$$

Intuition: the ML model must explain away at least *some* variance of Y.

. . . And When It Doesn't

PPI *underperforms* classical when

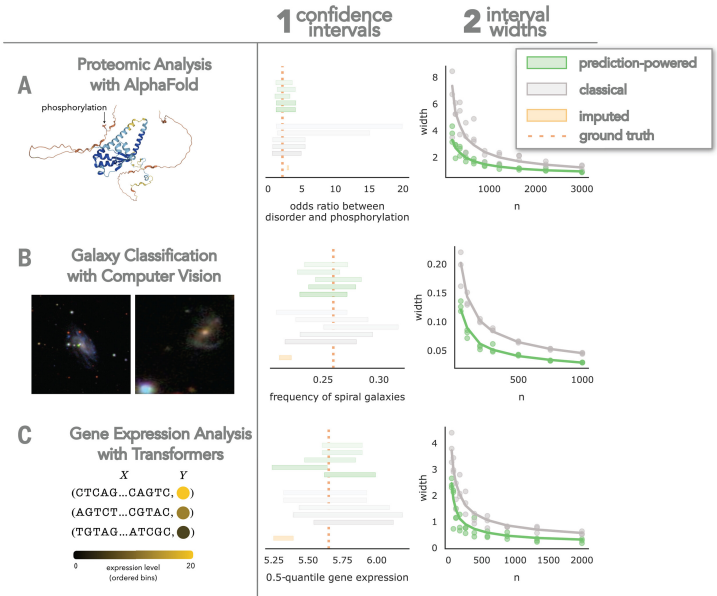
- Predictions are poor: $\text{var}(\widehat{Y} - Y)$ close to $\text{var}(Y)$.
- Unlabeled set is small: $N \not\gg n$; the $\text{var}(\widehat{Y}')/N$ term dominates.

Binary example ($Y \sim \text{Bern}(p)$, symmetric error rate η):

- $p = 0.5$: PPI helps only if $\eta < 25\%$.
- $p = 0.1$: PPI helps only if $\eta < 9.5\%$.

$\Rightarrow$ motivates Section 2: safety + adaptivity across multiple predictors.

Real-World Applications



(A) Proteomics with AlphaFold

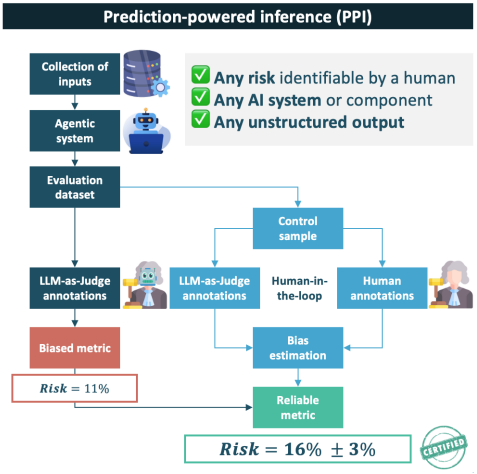
- **Question:** Are post-translational modifications (PTMs) more frequent in intrinsically disordered regions (IDRs)?
- Setup: $N \approx 10{,}803$ proteins, IDR labels cheap via *AlphaFold-predicted* relative solvent accessibility.
- Logistic regression coefficient (log-odds ratio).
- Imputation: wildly over-estimated odds ratio.
- PPI: valid, with much tighter CI than classical.

(C) Gene Expression Quantiles

- **Question:** 25%, 50%, 75% quantiles of gene expression induced by native yeast promoters.
- $N \approx 61{,}150$ promoter sequences; state-of-the-art transformer gives $\widehat{Y}$.
- Estimand: quantile(s).
- PPI gives much tighter CIs for all three quantiles, valid across trials.

GLIDE (Generated Label Inference & Debiasing Engine)

A Python library for rigorous evaluation of GenAI systems using hybrid human/proxy annotations, developed at Emerton Data.



PostPI: Post-Prediction Inference

- Wang et al. (2020)[2] modeled the relationship between the predicted and observed outcomes in the labeled subset of the data.

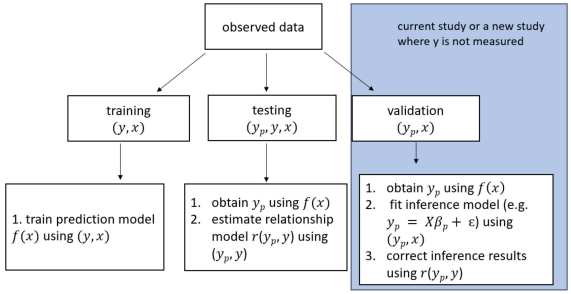
RESEARCH ARTICLE | PHYSICAL SCIENCES |

Methods for correcting inference based on outcomes predicted by machine learning

Siruo Wang, Tyler H. McCormick, and Jeffrey T. Leek | Authors Info & Affiliations

Edited by Robert Tibshirani, Stanford University, Stanford, CA, and approved October 6, 2020 (received for review January 24, 2020)

November 18, 2020 | 117 (48) 30266-30275 | https://doi.org/10.1073/pnas.2001238117



y: observed outcome x: observed covariate y_p: predicted outcome
$f(x)$: prediction model $r(y_p, y)$: relationship model

- Bootstrap procedure.

[2]Wang, S., McCormick, T., & Leek, J. (2020). Methods for correcting inference based on outcomes predicted by machine learning. Proceedings of the National Academy of Sciences, 117(48), 30266-30275.

PPI vs. Post-Prediction Inference (Wang et al. 2020)

- Wang et al. 2020: fit a correction model $r(\widehat{Y})$ on labeled data; apply to unlabeled.
- Requires strong parametric assumptions on the $Y \mid \widehat{Y}$ relationship.
- No general coverage guarantee; fails to cover in practice.
- PPI: valid under minimal distributional assumptions, broader class of estimands, extends to distribution shift.

PPI vs. Conformal Prediction

Conformal prediction	**Prediction-powered inference**
Prediction interval for a *single new* Y_{new}	CI for a *population parameter* θ^*
Exchangeability, non-conformity scores	CLT / empirical-Bernstein
Marginal coverage per instance	Population-level coverage
Not a substitute for PPI (ablation: inf-width intervals)	Targets scientific inference questions

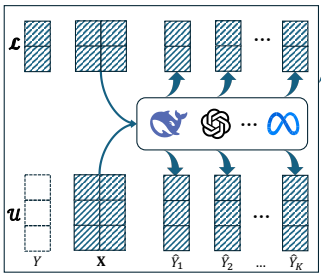
Both rely on a predictive model + labeled set, but answer *different* questions.

Key Takeaways

- PPI: a three-ingredient protocol $(\theta^*, m_\theta, \Delta_\theta)$ that fuses labeled + unlabeled + AI predictions.
- Validity for any ML, any distribution; power when $\text{var}(\widehat{Y} - Y) \ll \text{var}(Y)$.
- Distribution shift (covariate, label) accommodated under stronger structural assumptions.
- Limitations $\Rightarrow$ motivate the next three sections:
  - Section 2: multiple black-box predictors, safety & adaptivity.
  - Section 3: efficient inference under unknown covariate shift.
  - Section 4: efficient inference under unknown label shift.

Outline

SADA: Motivation



- We collect labeled data $\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ and unlabeled data $\mathcal{U} = \{\mathbf{x}_i, i = n+1, \ldots, N\}$.
- It is common to obtain predictions from multiple models or sources:
  - different models, such as GPT and Llama, often give different answers;
  - the quality of predictions from black-box models can be highly variable.
- Previous methods were designed for improving estimation efficiency rather than predictive accuracy.

SADA: Scope

1. Estimation and inference of a parameter $\boldsymbol{\theta}^* \in \Theta \subset \mathbb{R}^p$, defined by

$$\mathbb{E}\{\mathbf{s}(\mathbf{X}, Y; \boldsymbol{\theta}^*)\} = \mathbf{0}.$$

   - sample mean: $s(x, y; \theta) = y - \theta$;
   - least squares: $\mathbf{s}(\mathbf{x}, y; \boldsymbol{\theta}) = \mathbf{x}(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta})$.

- The evaluation metric is typically the MSE, $\mathbb{E}\{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^2\}$.

SADA: Scope (cont'd)

2. Prediction/Classification problem aiming to find $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ that minimizes the expected loss:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\{\ell(f_{\boldsymbol{\theta}}(\mathbf{X}), Y)\} =: \mathcal{R}(\boldsymbol{\theta}).$$

   - binary classification: classifier $f_{\boldsymbol{\theta}}(\mathbf{x}) \in (0, 1)$ and the cross-entropy loss

$$\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y) = -y\log\{f_{\boldsymbol{\theta}}(\mathbf{x})\} - (1-y)\log\{1 - f_{\boldsymbol{\theta}}(\mathbf{x})\}.$$

- The evaluation metric is typically the excess risk, $\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*)$.

SADA: Goal

Given multiple predicted labels with unknown quality, how can we aggregate them in a safe and data-adaptive manner to enhance statistical inference or improve predictive accuracy?

Inference: Illustration with Mean Estimation

	Unit	Features	Label	Multiple predicted labels
Labeled data $\mathcal{L}$	1	$\mathbf{x}_1$	y_1	$\widehat{\mathbf{y}}_1 = (\widehat{y}_{1,1}, \ldots, \widehat{y}_{K,1})^{\mathrm{T}}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n	$\mathbf{x}_n$	y_n	$\widehat{\mathbf{y}}_n = (\widehat{y}_{1,n}, \ldots, \widehat{y}_{K,n})^{\mathrm{T}}$
Unlabeled data $\mathcal{U}$	$n+1$	$\mathbf{x}_{n+1}$	?	$\widehat{\mathbf{y}}_{n+1} = (\widehat{y}_{1,n+1}, \ldots, \widehat{y}_{K,n+1})^{\mathrm{T}}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	N	$\mathbf{x}_N$	?	$\widehat{\mathbf{y}}_N = (\widehat{y}_{1,N}, \ldots, \widehat{y}_{K,N})^{\mathrm{T}}$

- Consider $\mathbf{s}(\mathbf{x}, y; \theta) = y - \theta$, which corresponds to the estimand $\theta^* = \mathbb{E}(Y)$.
- The naive estimator is $\widehat{\theta}^{\mathrm{nv}} = n^{-1} \sum_{i=1}^{n} y_i$, with $\mathrm{var}(\widehat{\theta}^{\mathrm{nv}}) = n^{-1}\mathrm{var}(Y)$.

Inference: Illustration with Mean Estimation

- We construct a family of unbiased estimators, indexed by weights $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)^{\mathrm{T}}$:

$$\widehat{\theta}(\boldsymbol{\omega}) := \frac{1}{n}\sum_{i=1}^{n} y_i + \sum_{k=1}^{K} \omega_k \left(\frac{1}{N-n} \sum_{i=n+1}^{N} \widehat{y}_{k,i} - \frac{1}{n}\sum_{i=1}^{n} \widehat{y}_{k,i} \right).$$

  - When $\boldsymbol{\omega} = 0$, it reduces to the naive/labeled-only estimator.
  - When $K = 1$, and $\omega = 1$, it turns out to be the PPI estimator.

- The variance, or equivalently, the mean squared error, of $\widehat{\theta}(\boldsymbol{\omega})$:

$$\mathbb{E}[\{\widehat{\theta}(\boldsymbol{\omega}) - \theta^*\}^2] = \underbrace{\frac{1}{n}\mathrm{var}(Y)}_{\mathrm{var}(\widehat{\boldsymbol{\theta}}^{\mathrm{nv}})} + \underbrace{\frac{N}{n(N-n)}\boldsymbol{\omega}^{\mathrm{T}}\mathrm{var}(\widehat{\mathbf{Y}})\boldsymbol{\omega} - \frac{2}{n}\boldsymbol{\omega}^{\mathrm{T}}\mathrm{cov}(\widehat{\mathbf{Y}}, Y)}_{\text{additional variance term}}. \quad (1)$$

Inference: Illustration with Mean Estimation

- (1) is a quadratic form of $\boldsymbol{\omega}$, which achieves the minimum at

$$\boldsymbol{\omega}^{\text{opt}} = \frac{N-n}{N}\text{var}(\widehat{\mathbf{Y}})^{-1}\text{cov}(\widehat{\mathbf{Y}}, Y).$$

- In practice, one can estimate $\boldsymbol{\omega}^{\text{opt}}$ as

$$\widehat{\boldsymbol{\omega}}^{\text{opt}} = \frac{N-n}{N}\left\{\frac{1}{N}\sum_{i=1}^{N}(\widehat{\mathbf{y}}_i - \overline{\widehat{\mathbf{y}}})(\widehat{\mathbf{y}}_i - \overline{\widehat{\mathbf{y}}})^{\text{T}}\right\}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{y}}_i - \overline{\widehat{\mathbf{y}}})(y_i - \overline{y})\right\},$$

  where $\overline{y} = n^{-1}\sum_{i=1}^{n} y_i$, and $\overline{\widehat{\mathbf{y}}} = N^{-1}\sum_{i=1}^{N}\widehat{\mathbf{y}}_i$.

- Our proposed SADA estimator is $\widehat{\theta}^{\text{sada}} = \widehat{\theta}(\widehat{\boldsymbol{\omega}}^{\text{opt}})$.

Safety & Guaranteed Efficiency Gain

- One can compute that

$$\mathbb{E}\{(\widehat{\theta}^{\text{sada}} - \theta^*)^2\} = \underbrace{\frac{1}{n}\text{var}(Y)}_{\text{naive estimator}} - \underbrace{\left(\frac{1}{n} - \frac{1}{N}\right)\text{cov}(\widehat{\mathbf{Y}}, Y)^{\mathrm{T}}\text{var}(\widehat{\mathbf{Y}})^{-1}\text{cov}(\widehat{\mathbf{Y}}, Y)}_{\text{efficiency gain}}.$$

- The efficiency gain is always non-negative regardless of the quality of the predictions.
- It vanishes to zero if and only if $\text{cov}(\widehat{\mathbf{Y}}, Y) = \mathbf{0}$, that means, none of the predictions are correlated with the ground truth Y.
- In this worst-case scenario, the optimal weight is automatically assigned to zero (i.e., $\boldsymbol{\omega} = \mathbf{0}$), reducing to the naive estimator.
- It's a SAFE aggregation.

Case 1: Oracle case

- If one of the black-box predictor $\widehat{Y}_k \equiv Y$ (no need to know which one!!)
- For example, $\widehat{Y}_1 \equiv Y$, then the optimal weight turns out to be

$$\boldsymbol{\omega}^{\text{opt}} = \frac{N-n}{N} \cdot (1, 0, \ldots, 0)^{\mathrm{T}}.$$

The SADA estimator turns out to be

$$\widehat{\theta}^{\text{sada}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{y}_{1i}, \quad \text{and} \quad \text{var}(\widehat{\theta}^{\text{sada}}) = \frac{1}{N} \text{var}(Y).$$

- It performs as if we knew the ground truth of the unlabeled data.

Case 2: Semiparametric efficiency

- If all the predictions are deterministic functions of $\mathbf{X}$, i.e., $\widehat{Y}_k \equiv \widehat{f}_k(\mathbf{X})$.
- The best prediction one can expect to fit the outcome is $\mathbb{E}(Y \mid \mathbf{X})$.
- If one of the black-box predictors $\widehat{Y}_k \equiv \mathbb{E}(Y \mid \mathbf{X})$ (no need to know which one!!)
- For example, $\widehat{Y}_1 \equiv \mathbb{E}(Y \mid \mathbf{X})$, then the optimal weight turns out to be

$$\boldsymbol{\omega}^{\text{opt}} = \frac{N-n}{N} \cdot (1, 0, \ldots, 0)^{\mathrm{T}}.$$

- Under some regularity conditions, $\sqrt{N}(\widehat{\theta}^{\text{sada}} - \theta^*) \xrightarrow{d} \mathcal{N}\{0, \mathbb{E}(\phi_{\text{eif}}^2)\}$, where $r = 1$ indicates labeled unit and 0 for unlabeled unit, $n/N \to \pi \in (0, 1)$ and

$$\phi_{\text{eif}}(r, \mathbf{x}, y) = \frac{r}{\pi}\{y - \mathbb{E}(Y \mid \mathbf{x})\} + \mathbb{E}(Y \mid \mathbf{x}) - \theta^*,$$

is the efficient influence function of θ^*.

Full Protocol

- We construct a family of unbiased estimators, $\widehat{\boldsymbol{\theta}}(\mathcal{W})$, indexed by the tuning parameters $\mathcal{W} = (\mathcal{W}_1^{\mathrm{T}}, \mathcal{W}_2^{\mathrm{T}}, \ldots, \mathcal{W}_K^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{(Kp)\times p}$, which solves

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{s}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) + \sum_{k=1}^{K}\mathcal{W}_k^{\mathrm{T}}\left\{\frac{1}{N-n}\sum_{i=n+1}^{N}\mathbf{s}(\mathbf{x}_i, \widehat{y}_{k,i}; \boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{s}(\mathbf{x}_i, \widehat{y}_{k,i}; \boldsymbol{\theta})\right\} = \mathbf{0}.$$

- If $\mathcal{W} = \mathbf{0}$, it turns out to be the naive estimator.
- When there is only one prediction $(K = 1)$:
  - if $\mathcal{W} = \mathbf{I}$, $\widehat{\boldsymbol{\theta}}(\mathcal{W})$ reduces to the PPI estimator;
  - if $\mathcal{W} = \omega\mathbf{I}$ with the tuning parameter ω selected optimally, it reduces to the PPI++ estimator.

Full Protocol

Proposition 1. Among the family of estimators $\widehat{\boldsymbol{\theta}}(\mathcal{W})$, the optimal tuning parameter, $\mathcal{W}^{\text{opt}}$, that minimizes the mean squared error loss such that $\mathbb{E}[\{\widehat{\boldsymbol{\theta}}(\mathcal{W}^{\text{opt}}) - \boldsymbol{\theta}^*\}^{\otimes 2}] \preceq \mathbb{E}[\{\widehat{\boldsymbol{\theta}}(\mathcal{W}) - \boldsymbol{\theta}^*\}^{\otimes 2}]$ for any $\mathcal{W} \in \mathbb{R}^{(Kp)\times p}$ is

$$\mathcal{W}^{\text{opt}} = \frac{N-n}{N}\text{var}\{\mathcal{S}(\mathbf{X}, \widehat{\mathbf{Y}}; \boldsymbol{\theta}^*)\}^{-1}\mathbb{E}\{\mathcal{S}(\mathbf{X}, \widehat{\mathbf{Y}}; \boldsymbol{\theta}^*)\mathbf{s}(\mathbf{X}, Y; \boldsymbol{\theta}^*)^{\mathrm{T}}\},$$

where $\mathcal{S}(\mathbf{x}, \widehat{\mathbf{y}}; \theta) := (\mathbf{s}(\mathbf{x}, \widehat{y}_1; \theta)^{\mathrm{T}}, \cdots, \mathbf{s}(\mathbf{x}, \widehat{y}_K; \theta)^{\mathrm{T}})^{\mathrm{T}}$.

- The proposed SADA estimator is $\widehat{\boldsymbol{\theta}}^{\text{sada}} = \widehat{\boldsymbol{\theta}}(\widehat{\mathcal{W}}^{\text{opt}})$, where $\widehat{\mathcal{W}}^{\text{opt}}$ is a consistent estimator of $\mathcal{W}^{\text{opt}}$.

Asymptotic Properties

Theorem (Safety)

Under some regularity conditions, the SADA estimator $\widehat{\boldsymbol{\theta}}^{\text{sada}}$ *has the asymptotic representation* $\sqrt{n}(\widehat{\boldsymbol{\theta}}^{\text{sada}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{H}^{-1}\boldsymbol{\Sigma}_{\text{opt}}\mathbf{H}^{-\mathbf{T}}\right)$. *More specifically, for its mean squared error, up to a negligible term, we have*

$$\mathbb{E}\{(\widehat{\boldsymbol{\theta}}^{\text{sada}} - \boldsymbol{\theta}^*)^2\} = \frac{1}{n}\mathbf{H}^{-1}\boldsymbol{\Sigma}_{\text{opt}}\mathbf{H}^{-\mathbf{T}},$$

where $\boldsymbol{\Sigma}_{\text{opt}} = \boldsymbol{\Sigma}_{\text{nv}} - (N-n)/N \cdot \boldsymbol{\Sigma}_g$, *and*

$$\boldsymbol{\Sigma}_g = \mathbb{E}\{\mathbf{s}(\mathbf{X}, Y; \boldsymbol{\theta}^*)\mathcal{S}(\mathbf{X}, \widehat{\mathbf{Y}}; \boldsymbol{\theta}^*)^{\mathrm{T}}\} var\{\mathcal{S}(\mathbf{X}, \widehat{\mathbf{Y}}; \boldsymbol{\theta}^*)\}^{-1}\mathbb{E}\{\mathcal{S}(\mathbf{X}, \widehat{\mathbf{Y}}; \boldsymbol{\theta}^*)\mathbf{s}(\mathbf{X}, Y; \boldsymbol{\theta}^*)^{\mathrm{T}}\}$$

is always positive semi-definite. It implies that $var(\widehat{\boldsymbol{\theta}}^{\text{sada}}) \preceq var(\widehat{\boldsymbol{\theta}}^{\text{nv}})$.

Asymptotic Properties

Theorem (Adaptivity)

(i) *Suppose* $\widehat{Y}_k \equiv Y$ *for some* k*, then*

$$\mathbb{E}\{(\widehat{\boldsymbol{\theta}}^{\text{sada}} - \boldsymbol{\theta}^*)^2\} = \frac{1}{N}\mathbf{H}^{-1}\boldsymbol{\Sigma}_{\text{nv}}\mathbf{H}^{-\mathbf{T}}.$$

Note that this is the same as the oracle estimator who knows the ground truth of the unlabeled data;

(ii) *Suppose* $n/N \to \pi \in (0,1)$ *as* $n, N \to \infty$*. Assume* $\widehat{Y}_k = \widehat{f}_k(\mathbf{X})$ *for* $k = 1, \ldots, K$*. Suppose* $\mathbf{s}(\mathbf{x}, \widehat{y}_{k'}; \boldsymbol{\theta}^*) = \mathbf{s}(\mathbf{x}, \widehat{f}_{k'}(\mathbf{x}); \boldsymbol{\theta}^*) \equiv \boldsymbol{\mu}(\mathbf{x})$ *for some* k'*, then we have* $\sqrt{N}(\widehat{\boldsymbol{\theta}}^{\text{sada}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}\{\mathbf{0}, \mathbb{E}(\boldsymbol{\Phi}_{\text{eif}}\boldsymbol{\Phi}_{\text{eif}}^{\mathrm{T}})\}$*, which attains the semiparametric efficiency bound.*

Simulation Design

$Y \sim \mathcal{N}(\theta^*, 1)$, $\theta^* = 0.5$, $n = 60$, $N = 200$, 1000 replications.

Two predictors, quality indexed by $\gamma \in [0, 1]$:

- $\widehat{Y}_1 = \gamma Y + (1 - \gamma)\varepsilon_1$ (better as $\gamma \uparrow$).
- $\widehat{Y}_2 = (1 - \gamma)Y + \gamma\varepsilon_2$ (worse as $\gamma \uparrow$).

Compare: naive (labeled only), PPI with $\widehat{Y}_1$, PPI with $\widehat{Y}_2$, PPI++ with each, SADA with both.

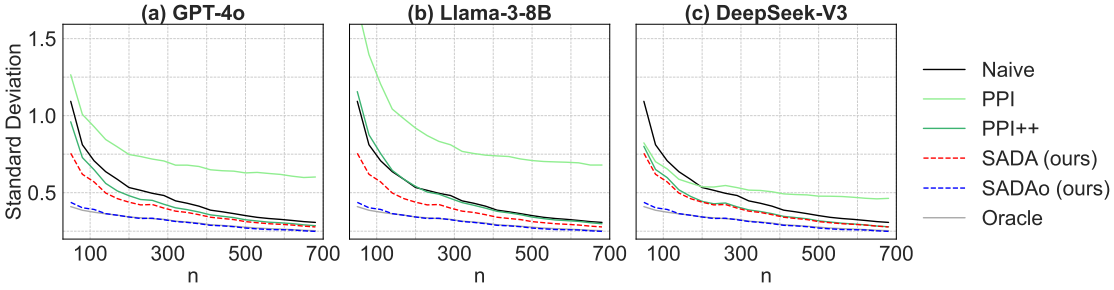
Relative Efficiency $\mathrm{sd}/\mathrm{sd}_{\mathrm{naive}}$

Method	$\gamma = 0$	0.3	0.5	0.7	1
PPI($\widehat{Y}_1$)	1.55	1.10	0.84	0.65	0.65
PPI($\widehat{Y}_2$)	0.65	0.66	0.85	1.10	1.55
PPI++($\widehat{Y}_1$)	0.99	0.93	0.80	0.64	0.55
PPI++($\widehat{Y}_2$)	0.55	0.64	0.80	0.93	0.99
SADA (both)	0.55	0.64	0.73	0.64	0.55

- Vanilla PPI: worse than naive when wrong model chosen.
- PPI++: safe but single-predictor.
- SADA: uniformly lowest sd; wins most strongly at $\gamma = 0.5$ by fusing both.

Application: Politeness of Online Requests

- **Motivation**: How certain linguistic features affect perceived politeness
- **Goal**: Regression coefficient for politeness scores on indicative modal features.
- **Black-box models:** GPT-4o, Llama-3-8B, and DeepSeek.



Prediction: Motivation

- In applications such as image classification, practitioners are primarily concerned with predictive accuracy rather than parameter estimation.
- Intuitively, $\text{var}(\widehat{\boldsymbol{\theta}}_1) \leq \text{var}(\widehat{\boldsymbol{\theta}}_2)$ does NOT imply $\mathcal{R}(\widehat{\boldsymbol{\theta}}_1) \leq \mathcal{R}(\widehat{\boldsymbol{\theta}}_2)$, where $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}\{\ell(f_{\boldsymbol{\theta}}(\mathbf{X}), Y)\}$.
- Suppose $\widehat{\mathcal{R}}(\boldsymbol{\theta})$ is an unbiased estimator of $\mathcal{R}(\boldsymbol{\theta})$, and $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{R}}(\boldsymbol{\theta})$.
- The excess risk

$$\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*) \leq \mathcal{R}(\widehat{\boldsymbol{\theta}}) - \widehat{\mathcal{R}}(\widehat{\boldsymbol{\theta}}) + \widehat{\mathcal{R}}(\boldsymbol{\theta}^*) - \mathcal{R}(\boldsymbol{\theta}^*) \leq 2 \sup_{\boldsymbol{\theta}\in\Theta} |\widehat{\mathcal{R}}(\boldsymbol{\theta}) - \mathcal{R}(\boldsymbol{\theta})|.$$

- Uniform concentration theory shows that the supremum depends on the complexity of Θ, the variance scale of $\widehat{\mathcal{R}}$, and the tail behavior of the empirical loss.

Prediction: Algorithm

- We consider a set of empirical loss functions:

$$\widehat{\mathcal{R}}(\boldsymbol{\theta}, \boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \sum_{k=1}^{K} \omega_i \left\{ \frac{1}{N-n} \sum_{i=n+1}^{N} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \widehat{y}_{k,i}) - \frac{1}{n}\sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \widehat{y}_{k,i}) \right\}.$$

- For a given $\boldsymbol{\theta}$,

$$\begin{aligned} \boldsymbol{\omega}^{\text{opt}}(\boldsymbol{\theta}) &= \arg\min_{\boldsymbol{\omega}} \mathbb{E}\{\widehat{\mathcal{R}}(\boldsymbol{\theta}, \boldsymbol{\omega}) - \mathcal{R}(\boldsymbol{\theta})\}^2 \\ &= \frac{N-n}{N} \text{var}\{\mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{X}), \widehat{\mathbf{Y}})\}^{-1} \text{cov}\{\mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{X}), \widehat{\mathbf{Y}}), \ell(f_{\boldsymbol{\theta}}(\mathbf{X}), Y)\}, \end{aligned}$$

  where $\mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}), \widehat{\mathbf{y}}) = (\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), \widehat{y}_1), \ell(f_{\boldsymbol{\theta}}(\mathbf{x}), \widehat{y}_2), \ldots, \ell(f_{\boldsymbol{\theta}}(\mathbf{x}), \widehat{y}_K))^{\text{T}}$.

- The SADA estimator is defined as

$$\widehat{\boldsymbol{\theta}}^{\text{sada}} = \arg\min_{\boldsymbol{\theta} \in \Theta} \widehat{\mathcal{R}}(\boldsymbol{\theta}, \widehat{\boldsymbol{\omega}}^{\text{opt}}(\boldsymbol{\theta})).$$

Prediction: Algorithm

- The optimal weight depends on $\boldsymbol{\theta}$ and should be optimized jointly.
- We recommend an iterative procedure that alternately updates the weight $\boldsymbol{\omega}^{\text{opt}}$ and the target parameter $\boldsymbol{\theta}$, as outlined in the following algorithm.

Algorithm 1 SADA for prediction

1: **Input:** Observations $(\mathbf{x}_i, y_i, \widehat{\mathbf{y}}_i)_{i=1}^n \cup (\mathbf{x}_i, \widehat{\mathbf{y}}_i)_{i=n+1}^N$, prediction function $f_{\boldsymbol{\theta}}$, loss function $\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y)$, number of iterations J.

2: Compute the naive estimator $\widehat{\boldsymbol{\theta}}^{\text{nv}} = \arg\min_{\boldsymbol{\theta}} n^{-1} \sum_{i=1}^n \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$.

3: Initialize $\widehat{\boldsymbol{\theta}}^{(0)} = \widehat{\boldsymbol{\theta}}^{\text{nv}}$.

4: **for** $j = 1, \ldots, J$ **do**

5: Compute the optimal weight

$$\widehat{\boldsymbol{\omega}}^{(j)} = \frac{N-n}{N} \left\{ \frac{1}{N} \sum_{i=1}^N \widetilde{\mathcal{L}}_N \left(f_{\widehat{\boldsymbol{\theta}}^{(j-1)}}(\mathbf{x}_i), \widehat{\mathbf{y}}_i \right) \widetilde{\mathcal{L}}_N \left(f_{\widehat{\boldsymbol{\theta}}^{(j-1)}}(\mathbf{x}_i), \widehat{\mathbf{y}}_i \right)^\top \right\}^{-1} \times \left\{ \frac{1}{n} \sum_{i=1}^n \mathcal{L} \left(f_{\widehat{\boldsymbol{\theta}}^{(j-1)}}(\mathbf{x}_i), \widehat{\mathbf{y}}_i \right) \ell \left(f_{\widehat{\boldsymbol{\theta}}^{(j-1)}}(\mathbf{x}_i), y_i \right) \right\},$$

where $\widetilde{\mathcal{L}}_N \left(f_{\boldsymbol{\theta}}(\mathbf{x}), \widehat{\mathbf{y}} \right) = \mathcal{L} \left(f_{\boldsymbol{\theta}}(\mathbf{x}), \widehat{\mathbf{y}} \right) - N^{-1} \sum_{i=1}^N \mathcal{L} \left(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \widehat{\mathbf{y}}_i \right)$.

6: Update the parameter estimate

$$\widehat{\boldsymbol{\theta}}^{(j)} = \arg\min_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \left\{ \frac{1}{N-n} \sum_{i=n+1}^N \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \widehat{\mathbf{y}}_i) - \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \widehat{\mathbf{y}}_i) \right\}^\top \widehat{\boldsymbol{\omega}}^{(j)}.$$

7: **end for**

8: **Output:** SADA estimator $\widehat{\boldsymbol{\theta}}^{\text{sada}} = \widehat{\boldsymbol{\theta}}^{(J)}$.

Prediction: Theoretical Guarantee

Theorem (Excess risk)

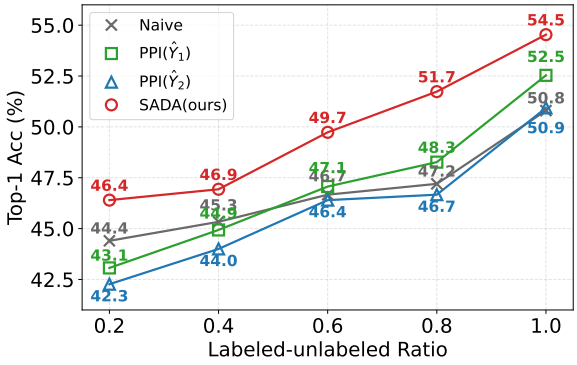
Under regularity conditions, we have that with probability at least $1-\delta$*,*

$$\mathcal{R}(\widehat{\boldsymbol{\theta}}^{\text{sada}}) - \mathcal{R}(\boldsymbol{\theta}^*) \leq \begin{cases} \mathcal{O}\left(\sqrt{\frac{1}{N}\log\frac{1}{\delta}} + \sqrt{\frac{1}{N}}\right), & \text{if } \widehat{Y}_k = Y \text{ for some } k \in \{1,\ldots,K\}, \\ \mathcal{O}\left(\sqrt{\frac{1}{n}\log\frac{1}{\delta}} + \sqrt{\frac{1}{n}} + \sqrt{\frac{1}{N}\log\frac{1}{\delta}} + \sqrt{\frac{1}{N}}\right), & \text{otherwise.} \end{cases}$$

- As a comparison, the excess risk of the naive (labeled-only) estimator, $\mathcal{R}(\widehat{\boldsymbol{\theta}}^{\text{nv}}) - \mathcal{R}(\boldsymbol{\theta}^*)$, converges at the order $\mathcal{O}(\sqrt{n^{-1}\log(\delta^{-1})} + \sqrt{n^{-1}})$ with high probability.

Application: Image Classification

- **Dataset and model**: ImageNet-5 with DaViT-T;
- **Teacher models**: EfficientNet-B0 and RegNetY-008;
- **Data structure**: 5000 training images, 750 validation images, and 750 test images;
- **Metric**: Top-1 accuracy.



R Package `sada`

- Install: `devtools::install_github("jw-shan/sada")`.
- Functions for inference (mean, M-estimation) and prediction.
- Accepts arbitrary K pseudo-label matrices.
- No hyperparameter tuning; weights are fully data-driven.

Key Takeaways

- SADA safely aggregates K predictions of unknown, possibly heterogeneous quality.
- Safety: never worse than labeled-only (vector parameters, any K).
- Adaptivity: attains semiparametric efficiency when any predictor is perfect.
- Applicable to inference *and* prediction tasks; software in R.
- Caveat: assumes same distribution across $\mathcal{L}$ and $\mathcal{U}$. The next sections drop this.

Main Paper

Shan, J., Chen, Z., Dong, Y., Wang, Y. & Zhao, J. (2025). SADA: Safe and adaptive aggregation of multiple black-box predictions in semi-supervised learning. *arXiv preprint arXiv:2509.21707*.

AOAS Special Issue

Annals of Applied Statistics

Special Issue: Statistics and Artificial Intelligence

Call for Papers

Lexin Li, Editor-in-Chief
2025-09-01

Outline

Refreshment at Room 152

1015-1030am

Outline

Distribution Shift?

- So far, the labeled data and unlabeled data follow the same distribution

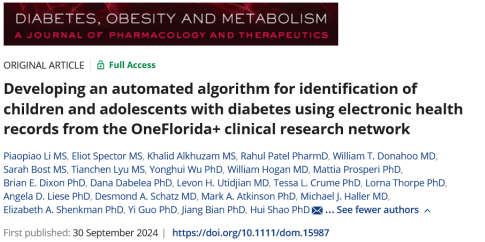
$$p(y \mid \mathbf{x}) = q(y \mid \mathbf{x}), \text{ and } p(\mathbf{x}) = q(\mathbf{x}).$$

- Is this reasonable?

Automated Computational Phenotype (ACP)

Table: Data structure (semi-supervised learning setting).

		SCENARIO I		
		R	Y	**X**
$\mathcal{L}$	1	1	$\surd$	$\surd$
	$\vdots$	$\vdots$	$\surd$	$\surd$
	n	1	$\surd$	$\surd$
$\mathcal{U}$	$n+1$	0		$\surd$
	$\vdots$	$\vdots$		$\surd$
	$n+N \equiv M$	0		$\surd$


DIABETES, OBESITY AND METABOLISM
A JOURNAL OF PHARMACOLOGY AND THERAPEUTICS

ORIGINAL ARTICLE | Full Access

Developing an automated algorithm for identification of children and adolescents with diabetes using electronic health records from the OneFlorida+ clinical research network

Piaopiao Li MS, Eliot Spector MS, Khalid Alkhuzam MS, Rahul Patel PharmD, William T. Donahoo MD, Sarah Bost MS, Tianchen Lyu MS, Yonghui Wu PhD, William Hogan MD, Mattia Prosperi PhD, Brian E. Dixon PhD, Dana Dabelea PhD, Levon H. Utidjian MD, Tessa L. Crume PhD, Lorna Thorpe PhD, Angela D. Liese PhD, Desmond A. Schatz MD, Mark A. Atkinson PhD, Michael J. Haller MD, Elizabeth A. Shenkman PhD, Yi Guo PhD, Jiang Bian PhD, Hui Shao PhD ... See fewer authors

First published: 30 September 2024 | https://doi.org/10.1111/dom.15987

- A stratified random sampling was applied to select individuals from the overall presumptive cases
- A group of reviewers were involved in the chart review process

Motivating Study: UF Diabetes

Li et al. (2025) – children/adolescents in UF OneFlorida+ network, 2012–2020.

- **Outcome** Y: binary diabetes status.
- **Labeled** $\mathcal{L}$ (manual chart review): $n = 297$, stratified sample by baseline.
- **Unlabeled** $\mathcal{U}$: $N = 3{,}000$ patients with ACP from a validated decision tree.
- **Covariates** X: age, visit counts, insurance type, BMI ≥ 30, CHF, comorbidity, family history, sex, race.

Stratified sampling $\Rightarrow$ covariate shift between labeled and unlabeled populations.

Covariate Shift

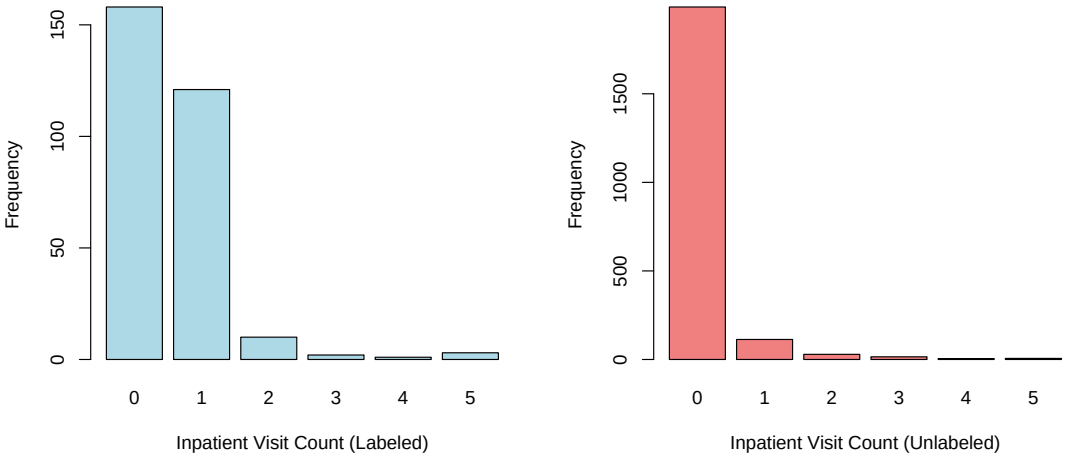


Figure: Difference in distributions of the inpatient visit count for labeled (left) and unlabeled (right) dataset.

Data Structure

Table: Data structure: Scenario I (w.o. ACP) vs Scenario II (w. ACP).

		Scenario I			Scenario II			
		R	Y	**X**	R	Y	**X**	$\widehat{Y}$
$\mathcal{L}$	1	1	$\surd$	$\surd$	1	$\surd$	$\surd$	$\surd$
	$\vdots$	$\vdots$	$\surd$	$\surd$	$\vdots$	$\surd$	$\surd$	$\surd$
	n	1	$\surd$	$\surd$	1	$\surd$	$\surd$	$\surd$
$\mathcal{U}$	$n+1$	0		$\surd$	0		$\surd$	$\surd$
	$\vdots$	$\vdots$		$\surd$	$\vdots$		$\surd$	$\surd$
	$n+N \equiv M$	0		$\surd$	0		$\surd$	$\surd$

Efficiency Analysis under Covariate Shift

- We[3] investigated whether the predicted data $\widehat{Y}$ would always yield an efficiency gain, and if so, where that gain originates.

Towards the Efficient Inference by Incorporating Automated Computational Phenotypes under Covariate Shift

---

Chao Ying [1 2] **Jun Jin** [3] **Yi Guo** [4] **Xiudi Li** [5] **Muxuan Liang** [6] **Jiwei Zhao** [1 2]

[3] Ying, C., Jin, J., Guo, Y., Li, X., Liang, M. & Zhao, J. (2025). Towards the efficient inference by incorporating automated computational phenotypes under covariate shift. *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR 267:72505-72534.

Efficiency Analysis?

- Will predicted data $\widehat{Y}$ always bring in efficiency gain?
- Where does the efficiency gain come from?

Covariate Shift Assumption

- We allow the marginal distributions of $\mathbf{X}$ to differ between $\mathcal{L}$ and $\mathcal{U}$, i.e.,

$$p(\mathbf{x}) \neq q(\mathbf{x}).$$

- This covariate shift assumption implies the independence between Y and R conditional on $\mathbf{X}$. That is,

$$\text{pr}(R = 1 \mid Y, \mathbf{X}) = \text{pr}(R = 1 \mid \mathbf{X}), \tag{1}$$

  which is less stringent than the assumption that the sampling of the labeled data is purely random; i.e., $\text{pr}(R = 1 \mid Y, \mathbf{X}) = \text{pr}(R = 1)$.

Objective and Metrics

- In the main paper, we focus on the d-dimensional parameter $\boldsymbol{\beta}$ for some characteristic in the unlabeled data population $\mathcal{U}$, defined as

$$\boldsymbol{\beta} = \text{argmin } \mathrm{E}_q\{\ell(y, \mathbf{x}; \boldsymbol{\beta})\},$$

  where E_q represents the expectation with respect to the distribution of $\mathcal{U}$, and $\ell(\cdot)$ denotes a generic loss function.
- Equivalently, with a smooth loss function, we write $\boldsymbol{\beta}$ as the solution of the estimating equation

$$\mathrm{E}_q\{\mathbf{s}(Y, \mathbf{X}; \boldsymbol{\beta})\} = \mathbf{0}. \tag{2}$$

Incorporating ACPs

- We assume that there exists a black-box prediction model, and we have $\widehat{y}_i$ for every subject $i \in \{1, \cdots, M\}$, where the input feature might contain not only $\mathbf{X}$ but also some other variable $\mathbf{Z}$.
- To understand the statistical benefits of these ACPs in the presence of covariate shift, we further assume

$$p(\widehat{y} \mid \mathbf{x}, y) \quad = \quad q(\widehat{y} \mid \mathbf{x}, y).$$

Together with the original covariate shift assumption in that $p(y \mid \mathbf{x}) = q(y \mid \mathbf{x})$, it implies

$$\text{pr}(R = 1 \mid Y, \widehat{Y}, \mathbf{X}) = \text{pr}(R = 1 \mid \mathbf{X}).$$

Again, this assumption is still less stringent than the assumption that the sampling of the labeled data is purely random.

Efficiency Gain Analysis

Proposition

With the ACP $\widehat{y}_i$'s, the EIF for estimating $\boldsymbol{\beta}$ defined in (2) is $\boldsymbol{\Omega}\boldsymbol{\phi}_w$ with $\boldsymbol{\phi}_w$ equals

$$\frac{r}{\pi} w_0(\mathbf{x})\{\mathbf{s}(y, \mathbf{x}; \boldsymbol{\beta}_0) - \widetilde{\mathbf{m}}_0(\mathbf{x}, \widehat{y})\} + \frac{1-r}{1-\pi}\mathbf{m}_0(\mathbf{x}) + \frac{w_0(\mathbf{x})}{\pi + (1-\pi)w_0(\mathbf{x})}\{\widetilde{\mathbf{m}}_0(\mathbf{x}, \widehat{y}) - \mathbf{m}_0(\mathbf{x})\}, \tag{3}$$

and the efficiency bound equals $\boldsymbol{\Omega}\mathbf{V}_w\boldsymbol{\Omega}$, where $\mathbf{V}_w$ is

$$\frac{1}{\pi}\mathrm{E}_p\left[w_0(\mathbf{X})^2\{\mathbf{s}(Y, \mathbf{X}; \boldsymbol{\beta}_0) - \widetilde{\mathbf{m}}_0(\mathbf{X}, \widehat{Y})\}^{\otimes 2}\right] + \frac{1}{(1-\pi)^2}\mathrm{E}\left[\{1 - \pi_0(\mathbf{X})\}^2\{\widetilde{\mathbf{m}}_0(\mathbf{X}, \widehat{Y}) - \mathbf{m}_0(\mathbf{X})\}^{\otimes 2}\right] + \frac{1}{1-\pi}\mathrm{E}_q\{\mathbf{m}_0(\mathbf{X})^{\otimes 2}\}.$$

Efficiency Gain Analysis (cont'd)

Proposition

Without $\widehat{y}_i$'s, the EIF for estimating $\boldsymbol{\beta}$ defined in (2) is $\boldsymbol{\Omega}\boldsymbol{\phi}_{wo}$ where

$$\boldsymbol{\phi}_{wo} = \frac{r}{\pi} w_0(\mathbf{x})\{\mathbf{s}(y, \mathbf{x}; \boldsymbol{\beta}_0) - \mathbf{m}_0(\mathbf{x})\} + \frac{1-r}{1-\pi}\mathbf{m}_0(\mathbf{x}),$$

and the efficiency bound equals $\boldsymbol{\Omega}\mathbf{V}_{wo}\boldsymbol{\Omega}$, where

$$\begin{aligned} \mathbf{V}_{wo} &= \frac{1}{\pi}\mathrm{E}_p\left[w_0^2(\mathbf{X})\{\mathbf{s}(Y, \mathbf{X}; \boldsymbol{\beta}_0) - \mathbf{m}_0(\mathbf{X})\}^{\otimes 2}\right] \\ &\quad + \frac{1}{1-\pi}\mathrm{E}_q\left\{\mathbf{m}_0(\mathbf{X})^{\otimes 2}\right\}. \end{aligned}$$

Efficiency Gain Analysis

Proposition

The difference of the two asymptotic variances, $\boldsymbol{\Omega}\mathbf{V}_{wo}\boldsymbol{\Omega} - \boldsymbol{\Omega}\mathbf{V}_{w}\boldsymbol{\Omega}$, equals

$$\frac{1}{(1-\pi)^2}\boldsymbol{\Omega}\mathrm{E}\left[\frac{\{1-\pi_0(\mathbf{X})\}^3}{\pi_0(\mathbf{X})}\{\widetilde{\mathbf{m}}_0(\mathbf{X},\widehat{Y}) - \mathbf{m}_0(\mathbf{X})\}^{\otimes 2}\right]\boldsymbol{\Omega},$$

which is always positive semidefinite.

- We denote $\pi \equiv n/M$, $w(\mathbf{x}) \equiv \frac{q(\mathbf{x})}{p(\mathbf{x})}$, $\pi(\mathbf{x}) \equiv \mathrm{pr}(R=1 \mid \mathbf{x})$, $\mathbf{m}(\mathbf{x}) \equiv \mathrm{E}\{\mathbf{s}(Y,\mathbf{x};\boldsymbol{\beta})|\mathbf{x}\}$ and $\widetilde{\mathbf{m}}(\mathbf{x},\widehat{y}) \equiv \mathrm{E}\{\mathbf{s}(Y,\mathbf{x};\boldsymbol{\beta})|\mathbf{x},\widehat{y}\}$. We write $\rho = (w(\mathbf{x}), \mathbf{m}(\mathbf{x}), \widetilde{\mathbf{m}}(\mathbf{x},\widehat{y}))$.
- We assume the $d \times d$ symmetric matrix $\mathrm{E}_q\{\partial\mathbf{s}(Y,\mathbf{X};\boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}\}$ evaluated at the true $\boldsymbol{\beta}_0$ is invertible and denote this inverse as $\boldsymbol{\Omega}$. Clearly, $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{\mathrm{T}}$.
- We use subscript $_0$ to denote the ground truth of nuisance functions and superscript * the best approximation of the truth within a possibly misspecified model, e.g., $\mathbf{m}_0(\mathbf{x})$ and $\mathbf{m}^*(\mathbf{x})$. For random vector $\boldsymbol{\phi}$, we write $\mathrm{E}(\boldsymbol{\phi}\boldsymbol{\phi}^{\mathrm{T}})$ as $\mathrm{E}(\boldsymbol{\phi}^{\otimes 2})$.

Source of the Efficiency Gain

Proposition

For the estimation efficiency bound, Scenario I = Scenario III.

Table: Data structure: Scenario III (intermediate scenario between Scenario I and Scenario II).

		SCENARIO III			
		R	Y	**X**	$\widehat{Y}$
	1	1	$\surd$	$\surd$	$\surd$
$\mathcal{L}$	$\vdots$	$\vdots$	$\surd$	$\surd$	$\surd$
	n	1	$\surd$	$\surd$	$\surd$
	$n+1$	0		$\surd$	
$\mathcal{U}$	$\vdots$	$\vdots$		$\surd$	
	$n+N \equiv M$	0		$\surd$	

Estimation

- To accommodate estimations of nuisance functions using nonparametric or machine learning methods, we adopt the cross-fitting technique.
- The estimator $\widehat{\boldsymbol{\beta}}(\widehat{\rho})$, shorthanded as $\widehat{\boldsymbol{\beta}}$, is the solution of the following equation:

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\widehat{E}_k\Bigg[&\frac{R}{\pi}\widehat{w}_k(\mathbf{X})\{\mathbf{s}(Y,\mathbf{X};\widehat{\boldsymbol{\beta}})-\widehat{\widehat{\mathbf{m}}}_k(\mathbf{X},\widehat{Y})\}\\&+\frac{\widehat{w}_k(\mathbf{X})}{\pi+(1-\pi)\widehat{w}_k(\mathbf{X})}\{\widehat{\widehat{\mathbf{m}}}_k(\mathbf{X},\widehat{Y})-\widehat{\mathbf{m}}_k(\mathbf{X})\}\\&+\frac{1-R}{1-\pi}\widehat{\mathbf{m}}_k(\mathbf{X})\Bigg]=\mathbf{0},\end{aligned}\tag{4}$$

  where $\widehat{E}_k$ denotes the sample average over the k-th fold, and we denote the left-hand side of the above equation as $\mathbf{N}(\widehat{\rho},\widehat{\boldsymbol{\beta}})$.
- For the estimations of nuisance functions, we adopted super learner (van der Laan et al. 2007), an ensemble learning approach that improves estimation by combining multiple machine learning models with data-adaptive weights.

Algorithm

Algorithm 1 Estimator and Confidence Interval

Input: Labeled dataset $D^l = \{(\mathbf{x}_i, y_i, \widehat{y}_i) : i = 1, \ldots, n\}$, unlabeled dataset $D^u = \{(\mathbf{x}_i, \widehat{y}_i) : i = n+1, \ldots, n+N\}$. To use cross-fitting, define $D^l = D^l_1 \cup \cdots \cup D^l_K$ and $D^u = D^u_1 \cup \cdots \cup D^u_K$, with $|D^l_1| = \cdots = |D^l_K| = n/K$ and $|D^u_1| = \cdots = |D^u_K| = N/K$. Define $D_k = D^l_k \cup D^u_k, k = 1, \cdots, K$.

Output: Obtain $\widehat{\boldsymbol{\beta}}$ and confidence interval of $\mathbf{v}^{\mathrm{T}}\boldsymbol{\beta}$ for any vector $\mathbf{v} \in \mathbb{R}^d$.

for $k = 1$ **to** K **do**

Estimate ρ_k using D^c_k, obtain $\widehat{\rho}_k$.

end for

Plug $\widehat{\rho}$ into equation (4) and solve the equation, obtain $\widehat{\boldsymbol{\beta}}$.

Plug $\widehat{\rho}$ and $\widehat{\boldsymbol{\beta}}$ into equation (5) and obtain $\widehat{\boldsymbol{\Omega}}\widehat{\mathbf{V}}_w\widehat{\boldsymbol{\Omega}}$.

Obtain the confidence intervals of $\mathbf{v}^{\mathrm{T}}\boldsymbol{\beta}$ by equation (7).

Assumption (Requirements on Nuisance Estimators)

For $k = 1, \ldots, K$, we assume positive density ratio models such that $\widehat{w}_k(\mathbf{x})$ and $w^*(\mathbf{x})$ are bounded away from zero. We assume the nuisance estimators $\widehat{\rho}_k = (\widehat{w}_k(\mathbf{x}), \widehat{\widetilde{\mathbf{m}}}_k(\mathbf{x}, \widehat{y}), \widehat{\mathbf{m}}_k(\mathbf{x}))$ converges to the limit $\rho^* = (w^*(\mathbf{x}), \widetilde{\mathbf{m}}^*(\mathbf{x}, \widehat{y}), \mathbf{m}^*(\mathbf{x}))$ in MSE at the following rates:

$$\begin{aligned}
\|\widehat{w}_k(\mathbf{x}) - w^*(\mathbf{x})\|_2 &= a_{1M}, \\
\|\widehat{\mathbf{m}}_k(\mathbf{x}) - \mathbf{m}^*(\mathbf{x})\|_2 &= a_{2M}, \\
\|\widehat{\widetilde{\mathbf{m}}}_k(\mathbf{x}, \widehat{y}) - \widetilde{\mathbf{m}}^*(\mathbf{x}, \widehat{y})\|_2 &= a_{3M},
\end{aligned}$$

where $\max\{a_{1M}, a_{2M}, a_{3M}\} \to 0$ as $M \to +\infty$.

Theoretical Investigations (cont'd)

Theorem (Double Robustness)

If either $w^*(\mathbf{x}) = w_0(\mathbf{x})$ *or* $(\widetilde{\mathbf{m}}^*(\mathbf{x}, \widehat{y}), \mathbf{m}^*(\mathbf{x})) = (\widetilde{\mathbf{m}}_0(\mathbf{x}, \widehat{y}), \mathbf{m}_0(\mathbf{x}))$, *then*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = o_p(1).$$

Theorem (Asymptotic Normality and Semiparametric Efficiency)

If we assume $a_{1M}a_{2M} = o(M^{-1/2})$, $a_{1M}a_{3M} = o(M^{-1/2})$, *and that all nuisance components are correctly specified, then as* $M \to \infty$,

$$\sqrt{M}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}\mathbf{V}_w\boldsymbol{\Omega}),$$

where $\boldsymbol{\Omega}\mathbf{V}_w\boldsymbol{\Omega} = \boldsymbol{\Omega}E(\boldsymbol{\phi}_w^{\otimes 2})\boldsymbol{\Omega}$ *is the semiparametric efficiency bound derived in Proposition 4.*

- Essentially $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \pi\boldsymbol{\Omega}\mathbf{V}_w\boldsymbol{\Omega})$.

Theorem (Construction of Confidence Interval)

If we assume $a_{1M}a_{2M} = o(M^{-1/2})$, $a_{1M}a_{3M} = o(M^{-1/2})$, *and that all nuisance components are correctly specified, then as* $M \to \infty$,

$$\widehat{\mathbf{\Omega}}\widehat{\mathbf{V}}_w\widehat{\mathbf{\Omega}} = \widehat{\mathbf{\Omega}}\frac{1}{K}\sum_{k=1}^{K}\widehat{E}_k\{\boldsymbol{\phi}_w^{\otimes 2}(\widehat{\rho};\widehat{\boldsymbol{\beta}})\}\widehat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}\mathbf{V}_w\mathbf{\Omega}. \tag{5}$$

Consequently, for any vector $\mathbf{v} \in \mathbb{R}^d$, *we have*

$$\frac{\sqrt{M}\mathbf{v}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(\mathbf{v}^{\mathrm{T}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{V}}_w\widehat{\mathbf{\Omega}}\mathbf{v})^{1/2}} \xrightarrow{d} N(0,1). \tag{6}$$

Thus, we can construct a $(1-\alpha)\times 100\%$ *confidence interval for* $\gamma = \mathbf{v}^{\mathrm{T}}\boldsymbol{\beta}_0$,

$$\mathrm{CI}_\alpha = \{\gamma : |\gamma - \mathbf{v}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}| \le \Phi^{-1}(1-\alpha/2)M^{-1/2}\widehat{\mathrm{sd}}\}, \tag{7}$$

where $\widehat{\mathrm{sd}} = (\mathbf{v}^{\mathrm{T}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{V}}_w\widehat{\mathbf{\Omega}}\mathbf{v})^{1/2}$. *Further, this CI satisfies*

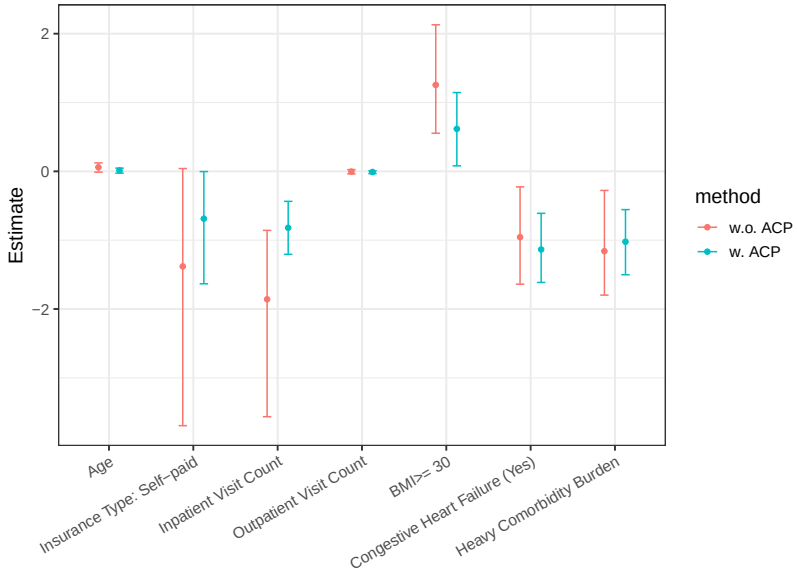
$$\lim_{M\to\infty} pr\left(\mathbf{v}^{\mathrm{T}}\boldsymbol{\beta}_0 \in \mathrm{CI}_\alpha\right) = 1-\alpha. \tag{8}$$

UF Diabetes Results

- Nuisances fit via R `SuperLearner` (GLM, RF, SVM, XGBoost), 10-fold CV.
- Perturbed bootstrap with Exp(1) weights for CIs.
- Covariate shift confirmed: inpatient visits, self-pay insurance, BMI≥ 30, CHF, comorbidity all $p < 0.001$ between $\mathcal{L}$ and $\mathcal{U}$.

Finding: method with ACP yields *shorter* 95% CIs; identifies self-paid insurance as a significant risk factor – consistent with stress / limited preventive care literature.

Application



Key Takeaways

- ACPs + covariate shift $\Rightarrow$ must *reweight* and *incorporate* carefully.
- Proposed estimator is doubly robust and semiparametrically efficient.
- Efficiency gain comes from ACPs on the unlabeled side only.
- Rate-double-robust: supports flexible ML nuisances without Donsker conditions.

Outline

From Covariate Shift to Label Shift

Two canonical distribution shifts between source $\mathbb{P}$ and target $\mathbb{Q}$:

Covariate shift (Section 3)

$p_X \neq q_X$, but $p_{Y|X} = q_{Y|X}$.
Causal setting: X causes Y. q_X/p_X estimable from X alone.

Label shift (this section)

$p_Y \neq q_Y$, but $p_{X|Y} = q_{X|Y}$.
Anticausal setting: Y causes X. Density ratio $\rho(y) = q_Y/p_Y$ is the culprit.

Examples of Label Shift

- **Pneumonia**: flu season shifts prevalence ($p_Y \neq q_Y$); symptom mechanism $g(x \mid y)$ stable.
- **Plankton counting**: species mix in 2013 vs 2014 at Woods Hole.
- **Medical imaging across hospitals**: disease prevalence differs; imaging physics stable.
- **Temperature forecasting**: Jan–Oct vs Nov–Dec distributions differ.

Key difficulty: Y is unobserved in $\mathbb{Q}$, so q_Y (and hence ρ) cannot be directly estimated.

Setup and Notation

- Source $\mathbb{P}$: (Y_i, X_i), $i = 1, \ldots, n_1$.
- Target $\mathbb{Q}$: X_j, $j = n_1 + 1, \ldots, n$. Let $n_0 = n - n_1$, $\pi = n_1/n$.
- Indicator $R = 1$ if from $\mathbb{P}$.
- Label shift: $p_Y(y) \neq q_Y(y)$, $p_{X|Y}(x, y) = q_{X|Y}(x, y) \equiv g(x, y)$.
- Three nuisances:
  - $\rho(y) = q_Y/p_Y$ – hardest; Y missing in $\mathbb{Q}$.
  - $p_{Y|X}$ or $\mathrm{E}_p(\cdot \mid x)$ – easy; ML on labeled $\mathbb{P}$.
  - $g(x, y)$ – high-dimensional, but only enters via 1-D expectations.
- Target: general θ solving $\mathrm{E}_q\{U(Y, X; \theta)\} = 0$. Mean: $\theta = \mathrm{E}_q(Y)$.

Side-by-Side with PPI under Label Shift

For $\theta = \mathrm{E}_q(Y)$:

PPI (assumes no shift):

$$\widehat{\theta}^{\mathrm{PPI}} = \tfrac{1}{N-n} \sum_{i=n+1}^{N} \widehat{y}_i + \tfrac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i).$$

Our efficient estimator:

$$\widehat{\theta} = \tfrac{1}{N-n} \sum_{i=n+1}^{N} \widehat{y}_i + \tfrac{1}{n} \sum_{i=1}^{n} \widetilde{\rho}(y_i)\,(y_i - \widehat{y}_i).$$

Differences:

- $\widetilde{\rho}$ in the rectifier (PPI implicitly sets $\rho \equiv 1$).
- $\widehat{y} = \widehat{b}(x)$ is engineered to satisfy $\mathrm{E}_q[\widehat{y} \mid y] = y$.
- Reaches the semiparametric efficiency bound for label shift.

“Doubly Flexible” vs “Doubly Robust”

Theory and Methods

Doubly Flexible Estimation under Label Shift

Seong-ho Lee, Yanyuan Ma & **Jiwei Zhao**

Pages 278-290 | Received 05 Dec 2022, Accepted 15 Feb 2024, Accepted author version posted online: 21 Feb 2024, Published online: 19 Mar 2024

Cite this article https://doi.org/10.1080/01621459.2024.2321653 Check for updates

Doubly robust (classic)	**Doubly flexible (this work)**
Exactly one of two models may be misspecified.	*Both* working models may be misspecified. Still consistent and asymptotically normal!

Working density-ratio model $\rho^*(y)$ and working conditional $p^{\star}_{Y|X}(y, x, \zeta)$ – both flexible.

The Doubly Flexible Estimator

Plug in $\rho^*(y)$ and $p^\star_{Y|X}$; compute $\widehat{b}^{\star\star}(x)$ via

$$\widehat{b}^{\star\star}(x) = \frac{\widehat{\mathrm{E}}^\star_p\{\widehat{a}^{\star\star}(Y)\rho^*(Y) \mid x\}}{\widehat{\mathrm{E}}^\star_p\{\rho^{*2}(Y) \mid x\} + \frac{\pi}{1-\pi}\widehat{\mathrm{E}}^\star_p\{\rho^*(Y) \mid x\}}.$$

$\widehat{a}^{\star\star}$ solves a Fredholm integral equation $\mathrm{E}[b(X) \mid y] = y$, approximated via kernel + Landweber iteration.

$$\widehat{\theta} \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{r_i}{\pi}\rho^*(y_i)\{y_i - \widehat{b}^{\star\star}(x_i)\} + \frac{1-r_i}{1-\pi}\widehat{b}^{\star\star}(x_i)\Big].$$

Proposition 2: $\widehat{\theta}$ consistent for θ even when *both* ρ^* and $p^\star_{Y|X}$ are misspecified.

Theorems 1 & Corollary 1

Theorem 1 (asymptotic normality). For any choice of working models,

$$\sqrt{n_1}(\widehat{\theta} - \theta) \;\Rightarrow\; \mathcal{N}(0, \sigma_\theta^2),$$

with σ_θ^2 an explicit variance – different regimes give different magnitudes.

Corollary 1 (efficiency). When *both* working models are correctly specified,

$$\sqrt{n_1}(\widehat{\theta}_{\text{eff}} - \theta) \;\Rightarrow\; \mathcal{N}\big(0,\, \text{var}\{\sqrt{\pi}\,\phi_{\text{eff}}\}\big),$$

attaining the semiparametric efficiency bound.

Three regimes unified: (i) both correct $\Rightarrow$ efficient; (ii) one correct $\Rightarrow$ still $\sqrt{n}$-CAN; (iii) both wrong $\Rightarrow$ still consistent and asymptotically normal.

DFE Application: Goal and Setting

1. We now illustrate the numerical performance of our proposed method through analyzing the Medical Information Mart for Intensive Care III (MIMIC-III), an openly available electronic health records database.
2. The outcome of interest in our analysis is the sequential organ failure assessment (SOFA) score, used to determine the extent of a patient's organ function or rate of failure during the stay in an intensive care unit.
3. We include 16 covariates from chart events and laboratory tests.
4. In our analysis, we define $\mathcal{P}$ as patients with private, government, and self-pay insurances ($R = 1, n_1 = 11,695$), and $\mathcal{Q}$ as patients whose insurance type is either Medicaid or Medicare ($R = 0, n_0 = 4,996$).

DFE Application: Working Models

1. The label shift assumption is the same as $\mathbf{X} \perp R|Y$. We test the conditional independence by the invariant environment prediction test in R package `CondIndTests`, and the p-values range from 0.460 to 0.628.
2. To identify a reasonable working model $\rho^*(y)$, we model the data $Y + 0.001$ from $\mathcal{P}$ as a gamma distribution $f(y, \alpha, \beta) = \Gamma(\alpha)^{-1}\beta^{-\alpha}y^{\alpha-1}\exp(-y/\beta)$. We estimate the unknown parameters α and β as $\widehat{\alpha}$ and $\widehat{\beta}$ using the MLE. Then we use the working model

$$\rho^*(y) = \frac{f(y + 0.001, \widehat{\alpha} + 1, \widehat{\beta})}{f(y + 0.001, \widehat{\alpha}, \widehat{\beta})}.$$

DFE Application: Results on Mean

Estimator	Diff	$\widehat{\text{SE}}$	CI
shift-dependent*	**.3120**	.0593	[3.9367, 4.1691]
doubly-flexible**	**.0170**	.0803	[3.6005, 3.9153]
singly-flexible*	**.0133**	.0496	[3.6570, 3.8514]
oracle	-	.0405	[3.6616, 3.8202]

1. **shift-dependent*(naive)**: $\check{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{r_i}{\pi}\rho(y_i)y_i$ with wrong ρ^*
2. **doubly-flexible**(proposed)**: $\widehat{\theta}$ with wrong ρ^* and $p^{\star}_{Y|\mathbf{X}}$
3. **singly-flexible*(proposed)**: $\widetilde{\theta}$ with wrong ρ^*
4. oracle: the $\sqrt{n_0}$-consistent estimator $\frac{1}{n_0}\sum_{i=n_1+1}^{n} y_i$.

DFE Application: Results on Quantiles

τ	Estimator	Diff	$\widehat{\text{SE}}$	CI
10%	shift-dependent*	.0000	.0102	[0.9801, 1.0199]
	doubly-flexible**	-.0002	.0095	[0.9812, 1.0185]
	singly-flexible*	-.0002	.0099	[0.9803, 1.0192]
	oracle	-	.0218	[0.9572, 1.0428]
25%	shift-dependent*	**.9995**	.0249	[1.9508, 2.0483]
	doubly-flexible**	**.0002**	.0276	[0.9460, 1.0543]
	singly-flexible*	**.0004**	.0196	[0.9620, 1.0389]
	oracle	-	.0315	[0.9383, 1.0617]
50%	shift-dependent*	.0002	.0408	[2.9203, 3.0801]
	doubly-flexible**	.0001	.0333	[2.9348, 3.0654]
	singly-flexible*	.0005	.0334	[2.9350, 3.0659]
	oracle	-	.0550	[2.8923, 3.1077]
75%	shift-dependent*	**.9999**	.0742	[5.8544, 6.1454]
	doubly-flexible**	**.0001**	.0759	[4.8512, 5.1489]
	singly-flexible*	**.0001**	.0381	[4.9254, 5.0748]
	oracle	-	.0591	[4.8842, 5.1158]
90%	shift-dependent*	**.9999**	.1380	[8.7295, 9.2703]
	doubly-flexible**	**.0004**	.1003	[7.8039, 8.1969]
	singly-flexible*	**.0004**	.0638	[7.8754, 8.1253]
	oracle	-	.1093	[7.7858, 8.2142]

Next Step: Efficient Estimation

- Doubly-flexible: efficient only if ρ^* happens to be correct.
- What if we want efficiency regardless of the working model?
- Solution: progressive three-stage estimation of $\rho(y)$.

Stage 1 (guess) $\Rightarrow$ Stage 2 (consistent $\widetilde{\rho}$) $\Rightarrow$ Stage 3 (efficient $\widehat{\rho}$).

Journal of Machine Learning Research 0 (2025) 1-47 Submitted 0/0; Revised 0/0; Published 0/0

Efficient Inference under Label Shift in Unsupervised Domain Adaptation

Seong-ho Lee SEONGHO@UOS.AC.KR
Department of Statistics
University of Seoul
Seoul, 02504, South Korea

Yanyuan Ma YZM63@PSU.EDU
Department of Statistics
Pennsylvania State University
University Park, PA 16802, USA

Jiwei Zhao* JIWEI.ZHAO@WISC.EDU
Departments of Statistics and of Biostatistics & Medical Informatics
University of Wisconsin-Madison
Madison, WI 53706, USA

Stage 1: Heuristic Guess

Pick any working density-ratio model $\rho^*(y)$.
Plug into the doubly-flexible algorithm ALGORITHM[θ, $\rho^*(y)$].

Guaranteed consistent; efficiency depends on whether ρ^* is close to ρ.

Stage 2: Consistent $\widetilde{\rho}$

Trick: approximate $q_Y(y_0)$ by a global feature of $\mathbb{Q}$:

$$\delta_0 \equiv \mathrm{E}_q\{K_h(Y - y_0)\}.$$

δ_0 is *estimable* via ALGORITHM[δ_0, ρ^*]:

$$\widetilde{\delta}_0 = \tfrac{1}{N-n} \sum_{i=n+1}^{N} \widetilde{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}^*(Y)\rho^*(Y) \mid x_i\}$$
$$+ \tfrac{1}{n} \sum_{i=1}^{n} \rho^*(y_i)\big[K_h(y_i - y_0) - \widetilde{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}^*(Y)\rho^*(Y) \mid x_i\}\big].$$

Then $\widetilde{\rho}(y_0) = \widetilde{\delta}_0 / \widehat{p}_Y(y_0)$.
Theorem 5: $\|\widetilde{\rho} - \rho\|_\infty = O(h^m) + O_p(\{\min(n, N-n)h\}^{-1/2}\log n) = o_p(1)$ – for any ρ^*.

Stage 3: Efficient $\widehat{\rho}$

Re-run ALGORITHM[δ_0, $\widetilde{\rho}$] – now with the consistent Stage-2 estimator as the "working model":

$$\widehat{\delta}_0 = \tfrac{1}{N-n} \sum_{i=n+1}^{N} \widehat{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}(Y)\widetilde{\rho}(Y) \mid x_i\} + \tfrac{1}{n} \sum_{i=1}^{n} \widetilde{\rho}(y_i)\big[K_h(y_i - y_0) - \widehat{w}_i \widehat{\mathrm{E}}_p\{\widehat{a}(Y)\widetilde{\rho}(Y) \mid x_i\}\big].$$

Theorem 6: If $\widetilde{\rho}$ uniformly consistent and $\widehat{\mathrm{E}}_p$ $o_p(n^{-1/4})$-consistent, $\widehat{\delta}_0$ achieves the semiparametric efficiency bound.

A "self-evolving" process: efficient estimator learned from its own consistent precursor.

Efficient $\widehat{\theta}$ (Algorithm 2)

Once $\widetilde{\rho}$ is in hand, solve

$$\tfrac{1}{N-n}\sum_{i=n+1}^{N}\widehat{b}(x_i) \;+\; \tfrac{1}{n}\sum_{i=1}^{n}\widetilde{\rho}(y_i)\{U(y_i,x_i;\theta)-\widehat{b}(x_i)\}=0,$$

with $\widehat{b}(x)=\widehat{w}\cdot\widehat{\mathrm{E}}_p\{U(Y,x;\theta)\widetilde{\rho}^2(Y)+\widehat{a}(Y)\widetilde{\rho}(Y)\mid x\}$.

Theorem 9: $\sqrt{n}(\widehat{\theta}-\theta)\Rightarrow\mathcal{N}\big(0,\mathrm{var}\{\sqrt{\pi}\,\phi_{\mathrm{eff}}\}\big)$.
Unlike Lee et al. (2025): efficiency does *not* require $\rho^*=\rho$.

Application 1: Plankton Counting

- Woods Hole Imaging FlowCytobot (same dataset as PPI).
- $\mathbb{P}$: 2013 labeled, $n = 421{,}238$; $\mathbb{Q}$: 2014, $N - n = 329{,}832$.
- Estimand: proportion of plankton in 2014.

Method	95% CI width
PPI* (no shift adjustment)	8466
shift-dependent*	578
singly-flexible*	662
efficient ~	433

$\sim 20\times$ narrower than PPI that ignores label shift.

Application 2: Temperature Forecast

- Szeged, Hungary 2006–2016; humidity $\to$ average daily temperature.
- $\mathbb{P}$: Jan–Oct ($n = 3{,}348$); $\mathbb{Q}$: Nov–Dec ($N - n = 671$).
- Estimand: $\mathrm{E}_q(Y)$ (avg temperature in target months).

Method	estimate	95% CI width
shift-dependent*	5.54	far off
doubly-flexible	4.27	1.98
efficient ~	4.09	0.48 (narrowest)

Key Takeaways

- Under label shift, $\rho(y)$ cannot be directly estimated – need clever leverage of $\mathbb{P}$-data.
- Doubly-flexible estimator (JASA): consistent + asymptotically normal under arbitrary working models.
- Progressive three-stage estimator (JMLR): self-evolves into a semiparametrically efficient estimator, *regardless* of the initial guess.
- Significant improvements over PPI with $\rho \equiv 1$, especially when the shift is severe.
- Applicable to general parameters: means, variances, quantiles, regression coefficients.

Outline

Overall Course Summary

- **Section 1 (PPI):** three-ingredient protocol for *any* black-box predictor.
- **Section 2 (SADA):** safe & adaptive aggregation of multiple predictions.
- **Section 3 (Covariate shift):** semiparametrically efficient inference with ACPs.
- **Section 4 (Label shift):** doubly flexible and efficient inference under label shift.

Trustworthy inference with black-box AI predictions — a unified perspective.

Acknowledgements

This research is partially supported by National Science Foundation through both NSF/NIGMS joint program (DMS-1953526/2122074) and statistics program (DMS-2310942, DMS-2610561), by National Institutes of Health (R01DC021431) as well as by the American Family Funding Initiative administered by the Data Science Institute of UW-Madison.

THANK YOU

jiwei.zhao@wisc.edu