Efficient Inference on High-Dimensional Linear Models with Missing Outcomes

Yikun Zhang†, Alexander Giessing, and Yen-Chi Chen

Department of Statistics, University of Washington

† yikun@uw.edu

REFERENCE

Problem of Interest

Statistical inference on the conditional mean $m_0(x) = \mathrm{E}(Y|X=x)$ **in the presence of high-dimensional covariates and missing outcomes is of practical importance. How can we conduct this inference efficiently?**

The efficiency should come from two aspects:

1. The final estimator of $m_0(x)$ is *semi-parametrically efficient* among a certain class of estimators.
2. The entire inference procedures are *computationally efficient*.

Basic Assumptions: Consider a random sample $\{(Y_i, R_i, X_i)\}_{i=1}^n$ drawn from the joint distribution of $(Y, R, X) \in \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ satisfying

(a) $Y = X^T\beta_0 + \epsilon$ with $\mathrm{E}(\epsilon|X) = 0$ and $\mathrm{E}(\epsilon^2|X) = \sigma_\epsilon^2$, where $||\beta_0||_0 = \sum_{k=1}^d \mathbb{1}_{\{\beta_{0k}\neq 0\}} = s_\beta < n \ll d$.

(b) The missingness indicator R is conditionally independent of Y given X (*i.e.*, missing at random; MAR).

Main Takeaway:

- **Semi-parametrically efficient debiased estimator of $m_0(x) = x^T\beta_0$ with high-dimensional covariates, MAR outcomes, and heavy-tailed noises.**
- **Computationally efficient procedures with both Python (`Debias-Infer`) and R (`DebiasInfer`) implementations.**
- **Clear motivation for the proposed debiasing program by the rationale of bias-variance trade-offs.**
- **Comparative simulations and real-world applications with finite-sample performance guarantees.**

Motivation: Bias-Variance Trade-off

The conditional mean squared error can be decomposed as:

$$\mathrm{E}\left[\left(\sqrt{n}\, m^{\mathrm{debias}}(x;\boldsymbol{w}) - \sqrt{n}\, m_0(x)\right)^2 \Big| \boldsymbol{X}\right]$$

$$= \underbrace{\sigma_\epsilon^2 \sum_{i=1}^n w_i^2 \pi(X_i)}_{\text{Main variance}} + \underbrace{\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i \pi(X_i) X_i - x\right)^T \sqrt{n}\,(\beta_0 - \beta)\right]^2}_{\text{Conditional bias}}$$

$$+ \underbrace{(\beta_0-\beta)^T\left[\sum_{i=1}^n w_i^2 \pi(X_i)\left(1-\pi(X_i)\right) X_i X_i^T\right](\beta_0-\beta)}_{\text{Asymptotic negligible variance}}.$$

Idea: Design a convex program that solves for the debiasing weights $w_i, i = 1, ..., n$ in order to

- Minimize the estimated "**Main variance**" $\sum_{i=1}^n w_i^2 \widehat{\pi}_i$.
- Constrain the estimated upper bound of "**Conditional bias**"

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i \widehat{\pi}_i X_i - x\right\|_\infty \sqrt{n}\left\|\beta_0 - \widehat{\beta}\right\|_1.$$

Methodology and Theory

To conduct statistical inference on $m_0(x) = x^T\beta_0$, we propose a debiasing method with the following procedures:

Step 1: Compute the **Lasso pilot estimate** $\widehat{\beta}$ with complete-case data

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^d}\left[\frac{1}{n}\sum_{i=1}^n R_i (Y_i - X_i^T\beta)^2 + \lambda\|\beta\|_1\right],$$

where $\lambda > 0$ is a regularization parameter.

Step 2: Obtain **consistent propensity score estimates** $\widehat{\pi}_i = \widehat{\mathrm{P}}(R_i = 1|X_i)$ for $i = 1, ..., n$ by any machine learning method (not necessarily a parametric model) on the data $\{(X_i, R_i)\}_{i=1}^n$.

Step 3: Solve for the debiasing weight vector $\widehat{\boldsymbol{w}} \equiv \widehat{\boldsymbol{w}}(x) = (\widehat{w}_1(x), ..., \widehat{w}_n(x))^T \in \mathbb{R}^n$ through a **debiasing program** defined as:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n}\left\{\sum_{i=1}^n \widehat{\pi}_i w_i^2 : \left\|x - \frac{1}{\sqrt{n}}\sum_{i=1}^n w_i \widehat{\pi}_i X_i\right\|_\infty \leq \frac{\gamma}{n}\right\}, \tag{1}$$

where $\gamma > 0$ is a tuning parameter.

Step 4: Define the **debiased estimator** for $m_0(x)$ as:

$$\widehat{m}^{\mathrm{debias}}(x;\widehat{\boldsymbol{w}}) = x^T\widehat{\beta} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{w}_i(x) R_i\left(Y_i - X_i^T\widehat{\beta}\right). \tag{2}$$

Step 5: Construct the **asymptotic** $(1-\tau)$**-level confidence interval** as:

$$\left[\widehat{m}^{\mathrm{debias}}(x;\widehat{\boldsymbol{w}}) \pm \Phi^{-1}\left(1-\frac{\tau}{2}\right)\cdot \sigma_\epsilon \cdot \sqrt{\frac{1}{n}\sum_{i=1}^n \widehat{\pi}_i \widehat{w}_i(x)^2}\right],$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of $\mathcal{N}(0,1)$.

Dual Formulation of the Debiasing Program (1):

$$\min_{\ell\in\mathbb{R}^d}\left\{\frac{1}{4n}\sum_{i=1}^n \widehat{\pi}_i\left[X_i^T\ell\right]^2 + x^T\ell + \frac{\gamma}{n}\|\ell\|_1\right\}. \tag{3}$$

- Relation between the solutions to the primal debiasing program $\widehat{\boldsymbol{w}}(x) \in \mathbb{R}^n$ and to the dual debiasing program $\widehat{\ell}(x) \in \mathbb{R}^d$:

$$\widehat{w}_i(x) = -\frac{1}{2\sqrt{n}}\cdot X_i^T\widehat{\ell}(x), \quad i = 1, ..., n.$$

  Plugging it into (2) is the key to deriving asymptotic normality.
- The tuning parameter $\gamma > 0$ of the debiasing program (1) can be selected via cross-validations (CV) on the dual program (3).

Consistency and Asymptotic Normality:

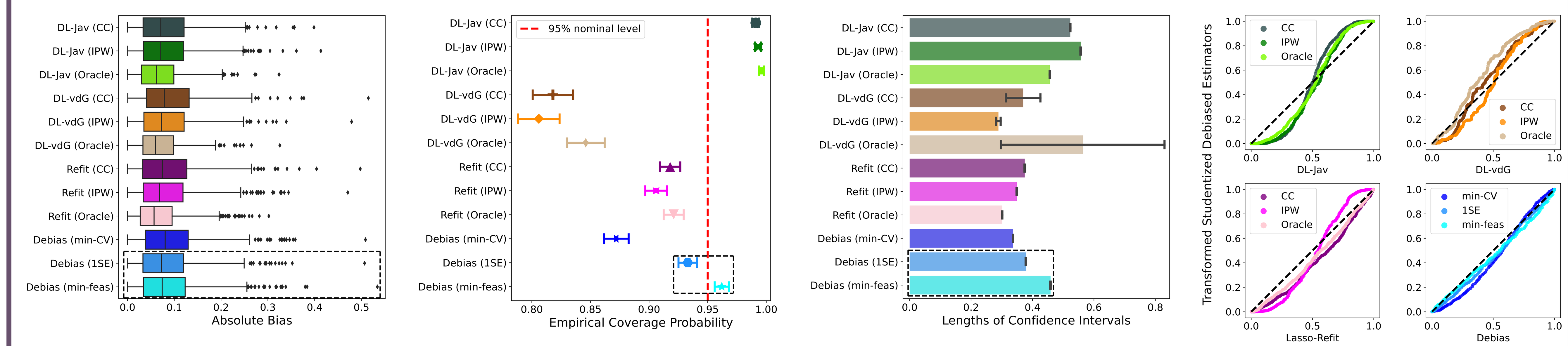
- Consistency of $\widehat{\beta}$ and the solution $\widehat{\ell}(x)$ to the dual program (3).
- Asymptotic normality of the debiased estimator (2):

$$\sqrt{n}\left[\widehat{m}^{\mathrm{debias}}(x;\widehat{\boldsymbol{w}}) - m_0(x)\right] \xrightarrow{d} \mathcal{N}\left(0, \sigma_m^2(x)\right),$$

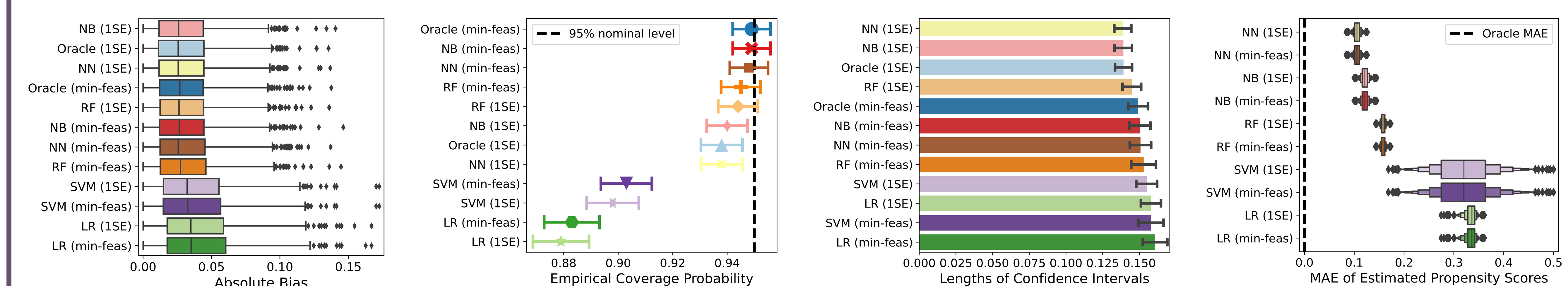
  where $\sigma_m^2(x) = \sigma_\epsilon^2 x^T\left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x$ attains the **semi-parametric efficiency bound** among all asymptotically linear estimators with MAR outcomes.

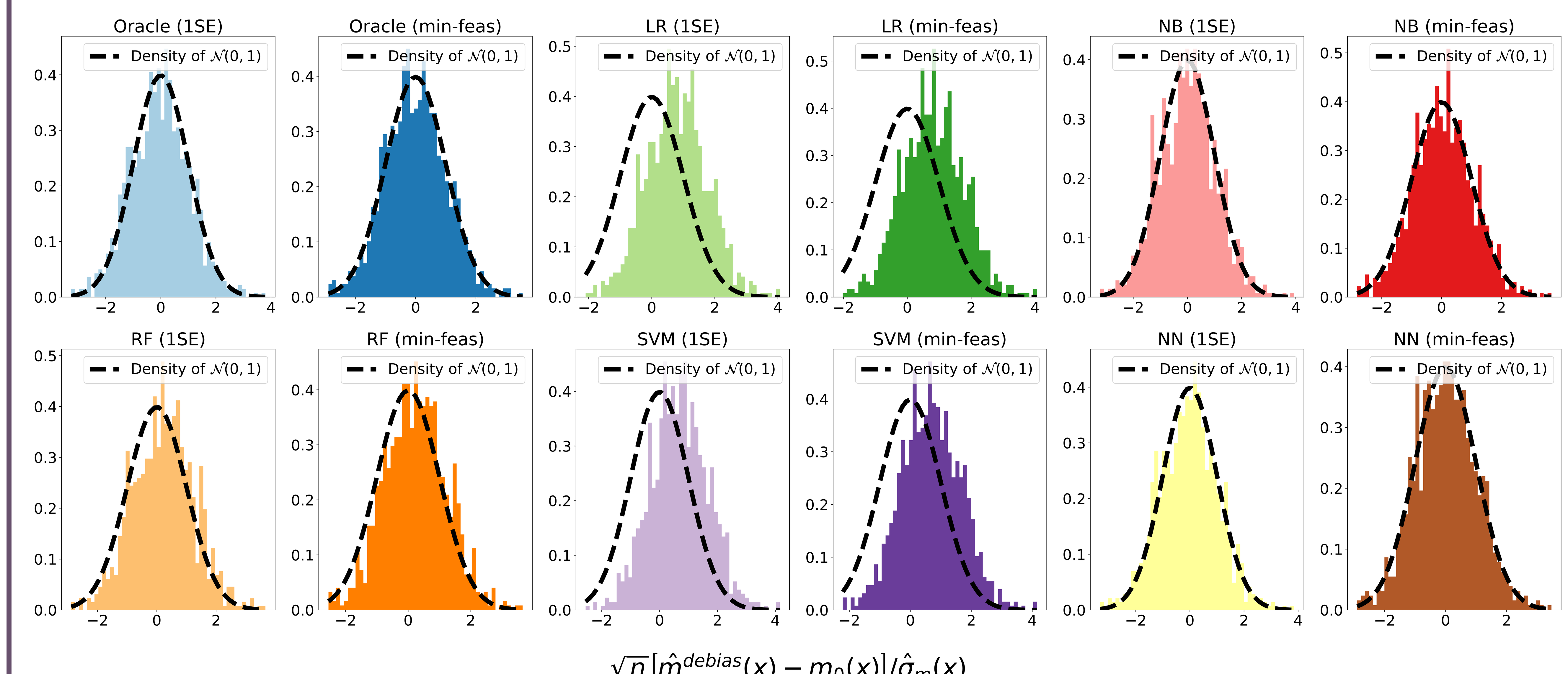
Simulation Studies and Real-World Applications

Comparisons With Existing Methods: DL-Jav, DL-vdG, and Refit are run on complete-case (CC), inverse probability weighting (IPW), and oracle data. Our proposed methods (Debias) under three CV-criteria are only run on the data with missing outcomes.

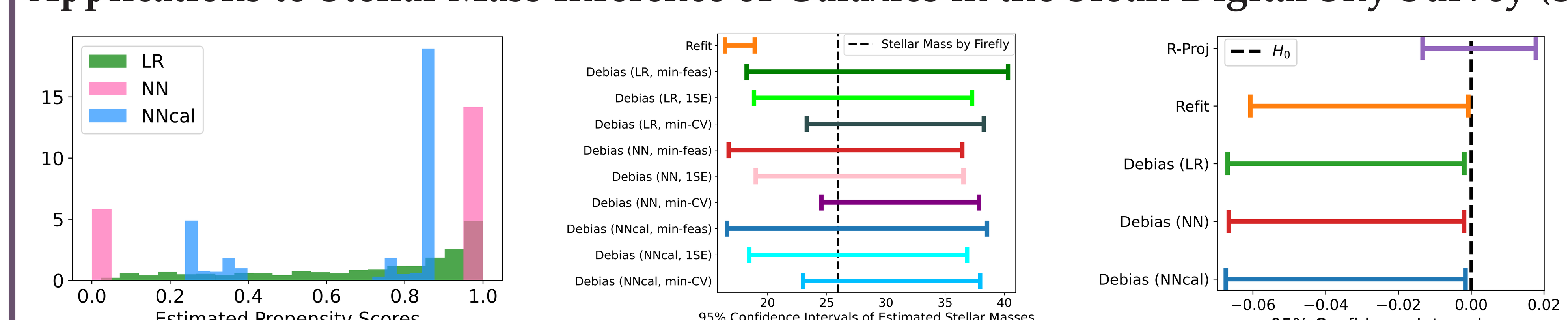


Proposed Debiasing Method With Nonparametrically Estimated Propensity Scores: LR (Lasso-type Logistic regression), NB (Gaussian Naive Bayes), RF (Random Forests), SVM (Support Vector Machine with Gaussian Radial Bases), and NN (Neural Networks).





Applications to Stellar Mass Inference of Galaxies in the Sloan Digital Sky Survey (SDSS-IV, DR16):



Middle Panel: Inference on the stellar mass of a newly observed galaxy based on its spectroscopic and photometric properties.
Right Panel: Testing the negative correlation between the stellar mass and the distance to nearby cosmic filament structures.