Efficient Inference on High-Dimensional Linear Models With Missing Outcomes

Yikun Zhang[1]

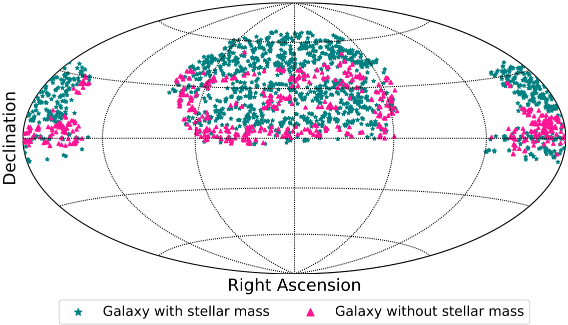
Joint Work with *Alexander Giessing*[2] and *Yen-Chi Chen*[1]

[1]Department of Statistics, University of Washington
[2]Department of Statistics and Data Science, National University of Singapore

2024 Joint Statistical Meetings
August 6, 2024

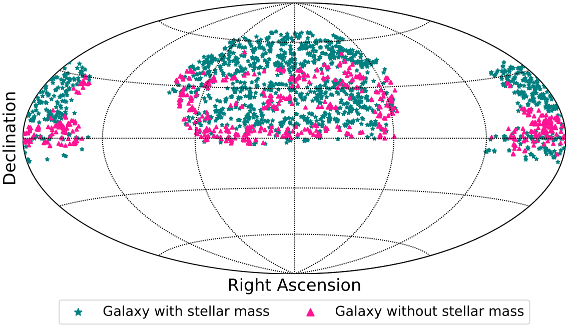
Department of STATISTICS

UNIVERSITY *of* WASHINGTON



Observed galaxies on the high redshift slice $0.4 \sim 0.401$.

▶ **Notes:** Sloan Digital Sky Survey (SDSS) observes millions of galaxies, but some (estimated) galactic stellar masses are missing in the associated value-added catalog (Comparat et al., 2017).



Observed galaxies on the high redshift slice $0.4 \sim 0.401$.

- **Scientific Question:**

How can we quantify the uncertainty of the (estimated) stellar mass of a newly observed galaxy based on the spectroscopic and photometric properties?

- **Notes:** Sloan Digital Sky Survey (SDSS) observes millions of galaxies, but some (estimated) galactic stellar masses are missing in the associated value-added catalog (Comparat et al., 2017).

- **High-dimensional Covariates:**

- Generate nonlinear features to capture complex patterns (Chang et al., 2015; Belloni et al., 2019).

- **High-dimensional Covariates:**
  - Generate nonlinear features to capture complex patterns (Chang et al., 2015; Belloni et al., 2019).
- **Reasons for Missingness:**
  - Limiting usage of the observational run in SDSS for galaxy targets;
  - Potential data contamination;
  - Misclassification of galaxies as stars.

- **High-dimensional Covariates:**
  - Generate nonlinear features to capture complex patterns (Chang et al., 2015; Belloni et al., 2019).
- **Reasons for Missingness:**
  - Limiting usage of the observational run in SDSS for galaxy targets;
  - Potential data contamination;
  - Misclassification of galaxies as stars.
- **Statistical Problem:**

How can we conduct valid and efficient inference on the regression function despite missing outcomes?

Model Assumptions

1. **Linearity:** The data $\{(Y_i, R_i, X_i)\}_{i=1}^n$ are i.i.d. observations from a sparse linear model

$$Y = X^T\beta_0 + \epsilon \quad \text{with} \quad \mathrm{E}\left(\epsilon \middle| X\right) = 0 \quad \text{and} \quad \mathrm{E}\left(\epsilon^2 \middle| X\right) = \sigma_\epsilon^2,$$

where $||\beta_0||_0 = s_\beta \ll d$ and $R \in \{0, 1\}$ when Y is missing or not.

Model Assumptions

1. **Linearity:** The data $\{(Y_i, R_i, X_i)\}_{i=1}^n$ are i.i.d. observations from a sparse linear model

$$Y = X^T \beta_0 + \epsilon \quad \text{with} \quad \mathrm{E}\left(\epsilon \middle| X\right) = 0 \quad \text{and} \quad \mathrm{E}\left(\epsilon^2 \middle| X\right) = \sigma_\epsilon^2,$$

where $||\beta_0||_0 = s_\beta \ll d$ and $R \in \{0, 1\}$ when Y is missing or not.

Relaxing the linearity assumption:

- Sparse additive model (Ravikumar et al., 2009);
- Partially linear model (Müller and van de Geer, 2015).

Our method can be generalized to handle heteroscedastic errors.

Model Assumptions

1. **Linearity:** The data $\{(Y_i, R_i, X_i)\}_{i=1}^n$ are i.i.d. observations from a sparse linear model

$$Y = X^T\beta_0 + \epsilon \quad \text{with} \quad \mathrm{E}\left(\epsilon \middle| X\right) = 0 \quad \text{and} \quad \mathrm{E}\left(\epsilon^2 \middle| X\right) = \sigma_\epsilon^2,$$

where $||\beta_0||_0 = s_\beta \ll d$ and $R \in \{0, 1\}$ when Y is missing or not.

Relaxing the linearity assumption:

- Sparse additive model (Ravikumar et al., 2009);
- Partially linear model (Müller and van de Geer, 2015).

Our method can be generalized to handle heteroscedastic errors.

2. **Missing At Random (MAR):** $Y_i \perp\!\!\!\perp R_i | X_i$ for $i = 1, ..., n$.

The existing works focus on the statistical inference on $\beta_0 \in \mathbb{R}^d$.

The existing works focus on the statistical inference on $\beta_0 \in \mathbb{R}^d$.

1. **Fully Observed Outcomes:** Debiased Lasso (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014):

$$\widehat{\beta}^{\text{debias}} = \widehat{\beta}_\lambda + \frac{1}{n}\widehat{\Theta}\sum_{i=1}^{n} X_i(Y_i - X_i^T\widehat{\beta}_\lambda),$$

   - $\widehat{\beta}_\lambda$ is a Lasso solution under the regularization parameter $\lambda > 0$;
   - $\widehat{\Theta} \in \mathbb{R}^{d\times d}$ is an approximation to the matrix inverse $\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1}$.

The existing works focus on the statistical inference on $\beta_0 \in \mathbb{R}^d$.

1. **Fully Observed Outcomes:** Debiased Lasso (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014):

$$\widehat{\beta}^{\text{debias}} = \widehat{\beta}_\lambda + \frac{1}{n}\widehat{\Theta}\sum_{i=1}^{n} X_i(Y_i - X_i^T\widehat{\beta}_\lambda),$$

   - $\widehat{\beta}_\lambda$ is a Lasso solution under the regularization parameter $\lambda > 0$;
   - $\widehat{\Theta} \in \mathbb{R}^{d\times d}$ is an approximation to the matrix inverse $\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1}$.

2. **MAR Outcomes:** M-estimation framework with a Lasso-type debiased and doubly robust estimator (Chakrabortty et al., 2019).

Our Contributions

- **Drawbacks of Existing Approaches:** Inference on $\beta_0 \in \mathbb{R}^d$.

1. Need to compute a $d \times d$ debiasing matrix $\widehat{\Theta}$.

2. Require sample splitting or cross fitting for valid inference.

Our Contributions

- **Drawbacks of Existing Approaches:** Inference on $\beta_0 \in \mathbb{R}^d$.

1. Need to compute a $d \times d$ debiasing matrix $\widehat{\Theta}$.
2. Require sample splitting or cross fitting for valid inference.

- **Our Focus:** Inference on $m_0(x) = x^T \beta_0$.

- *Computational efficiency:* Our debiasing program is convex and only needs to solve for an n-dimensional weight vector.
- *Statistical efficiency*: Our estimator is semi-parametrically efficient among all asymptotically linear estimators.

Methodology and Asymptotic Theory

Heuristics From Debiased Lasso

- The debiased Lasso estimator on the complete-case data is given by

$$\widehat{\beta}^{\text{debias}} = \widehat{\beta}_\lambda + \frac{1}{n}\sum_{i=1}^{n} R_i \widehat{\Theta} X_i \left(Y_i - X_i^T \widehat{\beta}_\lambda\right).$$

Heuristics From Debiased Lasso

- The debiased Lasso estimator on the complete-case data is given by

$$\widehat{\beta}^{\text{debias}} = \widehat{\beta}_\lambda + \frac{1}{n}\sum_{i=1}^{n} R_i \widehat{\Theta} X_i \left(Y_i - X_i^T \widehat{\beta}_\lambda \right).$$

- The candidate debiased estimator for $m_0(x) = x^T \beta_0$ is

$$\widehat{m}^{\text{debias}}(x) = x^T \widehat{\beta}^{\text{debias}} = x^T \widehat{\beta}_\lambda + \frac{1}{n} x^T \widehat{\Theta} \sum_{i=1}^{n} R_i X_i \left(Y_i - X_i^T \widehat{\beta}_\lambda \right).$$

- The debiased Lasso estimator on the complete-case data is given by

$$\widehat{\beta}^{\text{debias}} = \widehat{\beta}_\lambda + \frac{1}{n}\sum_{i=1}^{n} R_i \widehat{\Theta} X_i \left(Y_i - X_i^T \widehat{\beta}_\lambda\right).$$

- The candidate debiased estimator for $m_0(x) = x^T \beta_0$ is

$$\widehat{m}^{\text{debias}}(x) = x^T \widehat{\beta}^{\text{debias}} = x^T \widehat{\beta}_\lambda + \frac{1}{n} x^T \widehat{\Theta} \sum_{i=1}^{n} R_i X_i \left(Y_i - X_i^T \widehat{\beta}_\lambda\right).$$

► **Issue:** This naive estimator may not be asymptotically normal in general (van de Geer et al., 2014; Javanmard and Montanari, 2014)!

Heuristics From Debiased Lasso

$$\widehat{m}^{\text{debias}}(x) = x^T \widehat{\beta}^{\text{debias}} = x^T \widehat{\beta}_\lambda + \frac{1}{n} x^T \widehat{\Theta} \sum_{i=1}^{n} R_i X_i \left(Y_i - X_i^T \widehat{\beta}_\lambda \right).$$

- **Idea:** Introduce a weight vector $\boldsymbol{w} = (w_1, ..., w_n)^T \in \mathbb{R}^n$ to replace

$$\frac{1}{\sqrt{n}} x^T \widehat{\Theta} X_i \implies w_i \quad \text{for} \quad i = 1, ..., n$$

and formulate a generic debiased estimator

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

Heuristics From Debiased Lasso

$$\widehat{m}^{\text{debias}}(x) = x^T \widehat{\beta}^{\text{debias}} = x^T \widehat{\beta}_\lambda + \frac{1}{n} x^T \widehat{\Theta} \sum_{i=1}^{n} R_i X_i \left(Y_i - X_i^T \widehat{\beta}_\lambda \right).$$

- **Idea:** Introduce a weight vector $\boldsymbol{w} = (w_1, ..., w_n)^T \in \mathbb{R}^n$ to replace

$$\frac{1}{\sqrt{n}} x^T \widehat{\Theta} X_i \implies w_i \quad \text{for} \quad i = 1, ..., n$$

and formulate a generic debiased estimator

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

- **Question:** How do we estimate the weight vector $\boldsymbol{w} = (w_1, ..., w_n)^T$?

Conditional Mean Squared Error Decomposition

The conditional mean squared error of $\sqrt{n}\, m^{\text{debias}}(x; \boldsymbol{w})$ is

$$\mathrm{E}\left[\left(\sqrt{n}\, m^{\text{debias}}(x; \boldsymbol{w}) - \sqrt{n}\, m_0(x)\right)^2 \Big| X_1, ..., X_n\right]$$

Conditional Mean Squared Error Decomposition

The conditional mean squared error of $\sqrt{n}\, m^{\text{debias}}(x; \boldsymbol{w})$ is

$$
\begin{aligned}
&\mathrm{E}\left[\left(\sqrt{n}\, m^{\text{debias}}(x;\boldsymbol{w}) - \sqrt{n}\, m_0(x)\right)^2 \middle| X_1, \dots, X_n\right] \\
&= \underbrace{\sigma_\epsilon^2 \sum_{i=1}^n w_i^2 \pi_i}_{\text{Main Conditional Variance}} + \underbrace{\left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \pi_i X_i - x\right)^T \sqrt{n}\left(\beta_0 - \beta\right)\right]^2}_{\text{Conditional Bias}} \\
&\quad \underbrace{+ \left(\beta_0 - \beta\right)^T \left[\sum_{i=1}^n w_i^2 \pi_i \left(1 - \pi_i\right) X_i X_i^T\right] \left(\beta_0 - \beta\right)}_{\text{Asymptotically Negligible Conditional Variance}}.
\end{aligned}
$$

- **Notes:** $\pi_i := \mathrm{P}(R_i = 1 | X_i)$ is the propensity score under the MAR condition.

Bias-Variance Trade-off Optimization

$$\mathrm{E}\left[\left(\sqrt{n}\, m^{\text{debias}}(x;\boldsymbol{w}) - \sqrt{n}\, m_0(x)\right)^2 \middle| X_1, ..., X_n\right]$$

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

- By Hölder's inequality,

$$\text{“Conditional Bias”} \leq \left[\left\|\frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \pi_i X_i - x\right\|_\infty \sqrt{n}\,||\beta_0 - \beta||_1\right]^2.$$

Bias-Variance Trade-off Optimization

$$\mathrm{E}\left[\left(\sqrt{n}\, m^{\text{debias}}(x; \boldsymbol{w}) - \sqrt{n}\, m_0(x)\right)^2 \Big| X_1, \ldots, X_n\right]$$

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

- By Hölder's inequality,

$$\text{“Conditional Bias”} \leq \left[\left\|\frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \pi_i X_i - x\right\|_\infty \sqrt{n}\, ||\beta_0 - \beta||_1\right]^2.$$

- We design our debiasing program as:

$$\min_{\boldsymbol{w} \in \mathbb{R}^n} \sum_{i=1}^n w_i^2 \widehat{\pi}_i \quad \text{subject to} \quad \left\|x - \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \cdot \widehat{\pi}_i \cdot X_i\right\|_\infty \leq \frac{\gamma}{n}.$$

1. Compute the Lasso pilot estimate $\widehat{\beta}_\lambda$ on the complete-case data

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

Our Debiasing Inference Procedure

1. Compute the Lasso pilot estimate $\widehat{\beta}_\lambda$ on the complete-case data

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

2. Obtain consistent propensity score estimates $\widehat{\pi}_i, i = 1, ..., n$ by *any machine learning method*.

1. Compute the Lasso pilot estimate $\widehat{\beta}_\lambda$ on the complete-case data

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

2. Obtain consistent propensity score estimates $\widehat{\pi}_i, i = 1, ..., n$ by *any machine learning method*.

3. Solve the debiasing program defined as:

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

Our Debiasing Inference Procedure

1. Compute the Lasso pilot estimate $\widehat{\beta}_\lambda$ on the complete-case data

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

2. Obtain consistent propensity score estimates $\widehat{\pi}_i, i = 1, ..., n$ by *any machine learning method*.

3. Solve the debiasing program defined as:

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

4. Define the debiased estimator for $m_0(x) = x^T \beta$ as:

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

Our Debiasing Inference Procedure

1. How to select the tuning parameter $\gamma > 0$ for our debiasing program?

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

2. Is our debiased estimator asymptotically normal?

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

Our Debiasing Inference Procedure

1. How to select the tuning parameter $\gamma > 0$ for our debiasing program?

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

2. Is our debiased estimator asymptotically normal?

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

▶ **Answer:** The above two questions can be addressed by the *dual formulation* of our debiasing program!

- **Primal Program:**

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

Dual Formulation of Our Debiasing Program

- **Primal Program:**

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

- **Dual Program:**

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

Dual Formulation of Our Debiasing Program

- **Primal Program:**

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

- **Dual Program:**

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

- **Primal-Dual Relation:** Under the strong duality,

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

- **Dual Program:**

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

- It is an *unconstrained* optimization problem, and $\gamma > 0$ can be fine-tuned via cross-validation.

- **Dual Program:**

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

- It is an *unconstrained* optimization problem, and $\gamma>0$ can be fine-tuned via cross-validation.

- Primal-dual relation $\widehat{w}_i=-\frac{1}{2\sqrt{n}}\cdot X_i^T\widehat{\ell},\ i=1,...,n$ and dual consistency $\widehat{\ell}\xrightarrow{P}\ell_0$ reveal that

$$\sqrt{n}\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}})-m_0(x)\right]=\underbrace{-\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}R_i\epsilon_iX_i^T\ell_0}_{\text{i.i.d. sum!}}+\underbrace{\text{“Bias terms”}}_{o_P(1)}.$$

Consistency and Asymptotic Normality

Theorem (Theorem 7 in Zhang et al. 2023)

Under regularity conditions,

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

with $\sigma_m^2(x) = \lim_{n\to\infty} \sigma_\epsilon^2 \cdot x^T \left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x.$

Theorem (Theorem 7 in Zhang et al. 2023)

Under regularity conditions,

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

with $\sigma_m^2(x) = \lim_{n\to\infty} \sigma_\epsilon^2 \cdot x^T \left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x.$

1. For any fixed dimension $d > 0$, the asymptotic variance

$$\sigma_\epsilon^2 \cdot x^T \left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x$$

attains the *semi-parametric efficiency bound* among all asymptotically linear estimators under MAR outcomes (Müller and Keilegom, 2012).

Theorem (Theorem 7 in Zhang et al. 2023)

Under regularity conditions,

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

with $\sigma_m^2(x) = \lim_{n\to\infty} \sigma_\epsilon^2 \cdot x^T \left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x$.

1. For any fixed dimension $d > 0$, the asymptotic variance

$$\sigma_\epsilon^2 \cdot x^T \left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x$$

   attains the *semi-parametric efficiency bound* among all asymptotically linear estimators under MAR outcomes (Müller and Keilegom, 2012).

2. Under regularity conditions (Proposition 8 in Zhang et al. 2023),

$$\widehat{\sigma}_\epsilon^2 \sum_{i=1}^{n} \widehat{\pi}_i \widehat{w}_i^2 \xrightarrow{P} \sigma_\epsilon^2 \cdot x^T \left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x.$$

Sample splitting or cross fitting is often required in debiased inference via machine learning methods (Chernozhukov et al., 2018).

Sample splitting or cross fitting is often required in debiased inference via machine learning methods (Chernozhukov et al., 2018).

- Why don't we need sample splitting or cross fitting for estimating the propensity score by any machine learning method?

Theoretical Insights into Propensity Score Estimation

Sample splitting or cross fitting is often required in debiased inference via machine learning methods (Chernozhukov et al., 2018).

- Why don't we need sample splitting or cross fitting for estimating the propensity score by any machine learning method?

▶ **Answer:** Our asymptotic normality result depends on the *in-sample* estimation error r_π of the propensity score:

$$\max_{1\leq i\leq n} |\widehat{\pi}_i - \pi_i| = O_P\left(r_\pi\right) \quad \text{with} \quad \pi_i = \pi(X_i), i = 1, ..., n.$$

- Our debiased estimator performs even better when the estimated propensity scores on the training data are close to the true ones!!

Theoretical Insights into Propensity Score Estimation

Sample splitting or cross fitting is often required in debiased inference via machine learning methods (Chernozhukov et al., 2018).

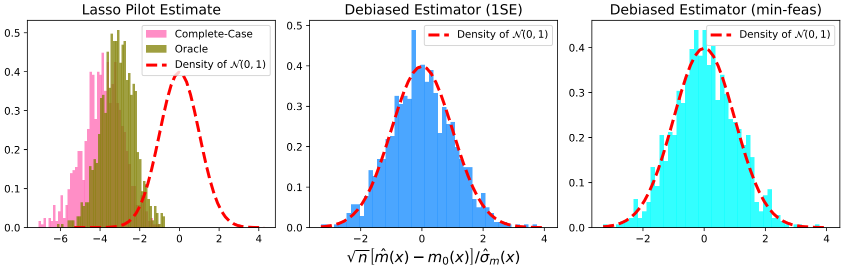
- Why don't we need sample splitting or cross fitting for estimating the propensity score by any machine learning method?

► **Answer:** Our asymptotic normality result depends on the *in-sample* estimation error r_π of the propensity score:

$$\max_{1\leq i\leq n} |\widehat{\pi}_i - \pi_i| = O_P(r_\pi) \quad \text{with} \quad \pi_i = \pi(X_i), i = 1, ..., n.$$

- Our debiased estimator performs even better when the estimated propensity scores on the training data are close to the true ones!!
- This permits the use of complex machine learning methods with high learnability (Steinwart, 2001; Farrell et al., 2021; Gao et al., 2022).

Simulation and Real-World Application



▶ **Effectiveness of Our Debiased Estimator:**

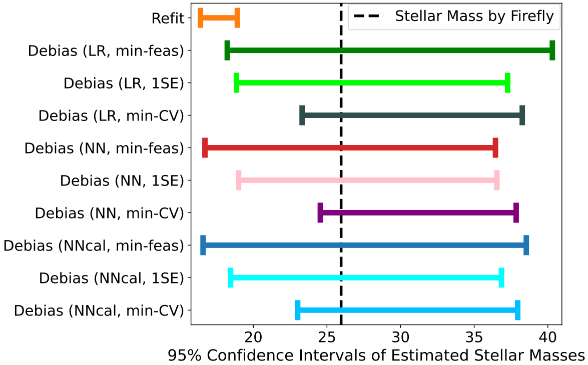
- Correct the bias of the Lasso pilot estimate.
- Asymptotically normal under a wide range of $\gamma > 0$.

▶ **Notes:** Our paper contains comprehensive comparisons with other existing methods.

Results on Galactic Stellar Mass Inference

How can we quantify the uncertainty of the (estimated) stellar mass of a galaxy based on the spectroscopic and photometric properties?

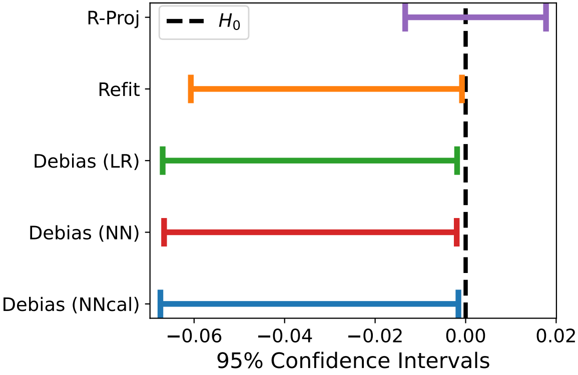
How can we quantify the uncertainty of the (estimated) stellar mass of a galaxy based on the spectroscopic and photometric properties?



- The 95% confidence intervals by our debiasing methods cover the true stellar mass of a new galaxy.

Results on Galactic Stellar Mass Inference

Is it statistically significant that the stellar mass of a galaxy is negatively correlated with its distance to the nearby cosmic filament structures?



- 95% confidence intervals by our debiasing methods exclude 0 and are all negative.

Summary of the Talk

We present an efficient debiasing method for conducting valid inference on high-dimensional linear models with MAR outcomes.

Summary of the Talk

We present an efficient debiasing method for conducting valid inference on high-dimensional linear models with MAR outcomes.

- The dual form explains its computational and statistical efficiencies.

Summary of the Talk

We present an efficient debiasing method for conducting valid inference on high-dimensional linear models with MAR outcomes.

- The dual form explains its computational and statistical efficiencies.
- The nuisance propensity score can be nonparametrically estimated without sample splitting or cross fitting.

Summary of the Talk

We present an efficient debiasing method for conducting valid inference on high-dimensional linear models with MAR outcomes.

- The dual form explains its computational and statistical efficiencies.
- The nuisance propensity score can be nonparametrically estimated without sample splitting or cross fitting.
- A novel application to the inference on galactic stellar mass.

Summary of the Talk

We present an efficient debiasing method for conducting valid inference on high-dimensional linear models with MAR outcomes.

- The dual form explains its computational and statistical efficiencies.
- The nuisance propensity score can be nonparametrically estimated without sample splitting or cross fitting.
- A novel application to the inference on galactic stellar mass.

More details can be found in

[1] Y. Zhang, A. Giessing, and Y.-C. Chen. Efficient Inference on High-Dimensional Linear Models with Missing Outcomes. *arXiv preprint*, 2023. https://arxiv.org/abs/2309.06429.

Python Package: Debias-Infer and R Package: DebiasInfer.

Summary of the Talk

We present an efficient debiasing method for conducting valid inference on high-dimensional linear models with MAR outcomes.

- The dual form explains its computational and statistical efficiencies.
- The nuisance propensity score can be nonparametrically estimated without sample splitting or cross fitting.
- A novel application to the inference on galactic stellar mass.

More details can be found in

[1] Y. Zhang, A. Giessing, and Y.-C. Chen. Efficient Inference on High-Dimensional Linear Models with Missing Outcomes. *arXiv preprint*, 2023. https://arxiv.org/abs/2309.06429.

Python Package: Debias-Infer and R Package: DebiasInfer.

Thank you!

Reference

A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

A. Belloni, V. Chernozhukov, and K. Kato. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association*, 114 (526):749–758, 2019.

L. Breiman, J. Friedman, C. J. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

A. Chakrabortty, J. Lu, T. T. Cai, and H. Li. High dimensional m-estimation with missing outcomes: A semi-parametric framework. *arXiv preprint arXiv:1911.11345*, 2019.

Y.-Y. Chang, A. van der Wel, E. da Cunha, and H.-W. Rix. Stellar masses and star formation rates for 1 m galaxies from sdss+ wise. *The Astrophysical Journal Supplement Series*, 219(1):8, 2015.

Y. Chen and Y. Yang. The one standard error rule for model selection: does it work? *Stats*, 4(4):868–892, 2021.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

J. Comparat, C. Maraston, D. Goddard, V. Gonzalez-Perez, J. Lian, S. Meneses-Goytia, D. Thomas, J. R. Brownstein, R. Tojeiro, A. Finoguenov, et al. Stellar population properties for 2 million galaxies from sdss dr14 and deep2 dr4 from full spectral fitting. *arXiv preprint arXiv:1711.06575*, 2017.

S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

A. Fu, B. Narasimhan, and S. Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020. doi: 10.18637/jss.v094.i14.

W. Gao, F. Xu, and Z.-H. Zhou. Towards convergence rate analysis of random forests for classification. *Artificial Intelligence*, 313:103788, 2022.

J. Jackson. A critique of rees's theory of primordial gravitational radiation. *Monthly Notices of the Royal Astronomical Society*, 156(1):1P–5P, 1972.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

N. Kaiser. Clustering in real space and in redshift space. *Monthly Notices of the Royal Astronomical Society*, 227(1):1–21, 1987.

U. Kuchner, A. Aragón-Salamanca, A. Rost, F. R. Pearce, M. E. Gray, W. Cui, A. Knebe, E. Rasia, and G. Yepes. Cosmic filaments in galaxy cluster outskirts: quantifying finding filaments in redshift space. *Monthly Notices of the Royal Astronomical Society*, 503(2):2065–2076, 2021.

P. Müller and S. van de Geer. The partial linear model in high dimensions. *Scandinavian Journal of Statistics*, 42(2):580–608, 2015.

U. U. Müller and I. V. Keilegom. Efficient parameter estimation in regression with missing responses. *Electronic Journal of Statistics*, 6(none):1200 – 1219, 2012.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Reference

S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76 (1):217–242, 2014.

Y. Zhang and Y.-C. Chen. Linear convergence of the subspace constrained mean shift algorithm: from euclidean to directional data. *Information and Inference: A Journal of the IMA*, 12(1):210–311, 2023.

Y. Zhang, R. S. de Souza, and Y.-C. Chen. Sconce: a cosmic web finder for spherical and conic geometries. *Monthly Notices of the Royal Astronomical Society*, 517(1):1197–1217, 2022.

Y. Zhang, A. Giessing, and Y.-C. Chen. Efficient inference on high-dimensional linear models with missing outcomes. *arXiv preprint arXiv:2309.06429*, 2023.

Implementation Details of the Proposed Debiasing Method

1. **Lasso pilot estimate:** We adopt the scaled Lasso (Sun and Zhang, 2012) with its universal regularization parameter $\lambda_0 = \sqrt{\frac{2\log d}{n}}$ as the initialization. Specifically, it iteratively updates $\widehat{\beta}(\widetilde{\lambda}), \widehat{\sigma}_\epsilon(\widetilde{\lambda}), \widetilde{\lambda}$ via the jointly convex optimization program:

$$\left(\widehat{\beta}(\widetilde{\lambda}), \widehat{\sigma}_\epsilon(\widetilde{\lambda})\right) = \mathop{\arg\min}_{\beta\in\mathbb{R}^d, \sigma_\epsilon>0} \left[\frac{1}{2n\sigma_\epsilon}\sum_{i=1}^n R_i\left(Y_i - X_i^T\beta\right)^2 + \frac{\sigma_\epsilon}{2} + \widetilde{\lambda}\,||\beta||_1\right].$$

2. **Debiasing program:** We solve the primal program by Python package "`CVXPY`" (Diamond and Boyd, 2016; Agrawal et al., 2018) or R package "`CVXR`" (Fu et al., 2020). For the dual program, we formulate a coordinate descent algorithm (Wright, 2015) as:

$$\left[\widehat{\ell}(x)\right]_j \leftarrow \frac{\mathcal{S}_{\frac{\gamma}{n}}\left(-\frac{1}{2n}\sum_{i=1}^n \widehat{\pi}_i\left(\sum_{k\neq j} X_{ik}X_{jk}\left[\widehat{\ell}(x)\right]_k\right) - x_j\right)}{\frac{1}{2n}\sum_{i=1}^n \widehat{\pi}_i X_{ij}^2} \quad \text{for } j = 1, \dots, d,$$

where $\mathcal{S}_{\frac{\gamma}{n}}(u) = \mathrm{sign}(u)\cdot\left(u - \frac{\gamma}{n}\right)_+$ is the soft-thresholding operator.

One Standard Error (1SE) Rule For Model Selection

- Suppose that we conduct a K-fold cross-validation on a candidate set $\Gamma = \{\gamma_1, ..., \gamma_m\}$ of the tuning parameter.

- For each $\gamma_i \in \Gamma$, we compute the cross-validated risk or error on each fold of the data as:
$$CV_k(\gamma_i), \quad k = 1, ..., K.$$
- For each $\gamma_i \in \Gamma$, we calculate the standard error of $CV_1(\gamma_i), ..., CV_K(\gamma_i)$ as:
$$SD(\gamma_i) = \sqrt{\mathrm{Var}\left(CV_1(\gamma_i), ..., CV_K(\gamma_i)\right)}, \quad SE(\gamma_i) = SD(\gamma_i)/\sqrt{K}.$$
- Let
$$CV(\gamma) = \frac{1}{K}\sum_{k=1}^{K} CV_k(\gamma) \quad \text{and} \quad \widehat{\gamma} = \arg\min_{\gamma \in \Gamma} CV(\gamma).$$
The 1SE rule (Breiman et al., 1984; Chen and Yang, 2021) selects $\gamma_{1SE} \in \Gamma$ with as the one with the smallest $CV(\gamma)$ such that
$$CV(\gamma_{1SE}) \geq CV(\widehat{\gamma}) + SE(\widehat{\gamma}).$$

One Standard Error (1SE) Rule For Model Selection

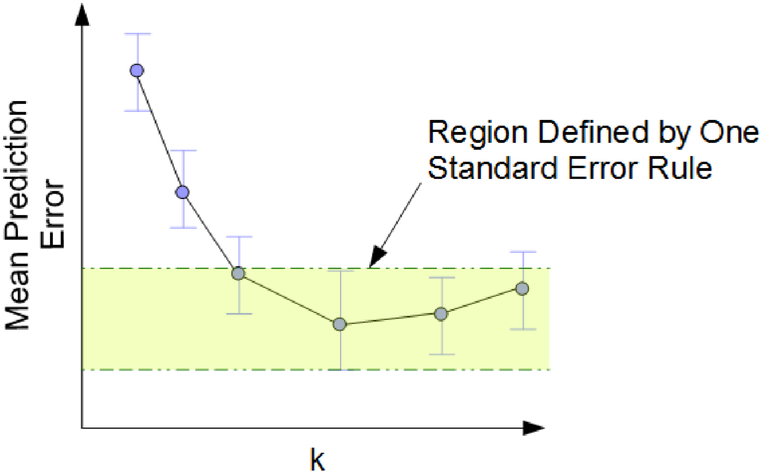


Figure: Illustration of the 1SE rule for selecting the model parameter.

Interpretations From Neyman Near-Orthogonalization

- Consider the regression function $m \equiv m(x) \in \mathbb{R}$ as the main parameter to be inferred and $\beta \in \mathbb{R}^d$ as the high-dimensional nuisance parameter.

- Our generic debiased estimator $m^{\text{debias}}(x, \boldsymbol{w})$ solves the sample-based estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \Xi_x(Y_i, R_i, X_i; m^{\text{debias}}, \beta) = m^{\text{debias}}(x; \boldsymbol{w}) - x^T\beta - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i \cdot R_i \left(Y_i - X_i^T\beta\right) = 0.$$

- The Neyman near-orthogonalization condition (Chernozhukov et al., 2018) given $\boldsymbol{X} = (X_1, ..., X_n)^T \in \mathbb{R}^{n\times d}$ at $(m_0, \beta_0) = (x^T\beta_0, \beta_0)$ requires

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \Xi_x(Y_i, R_i, X_i; m_0, \beta_0)\middle|\boldsymbol{X}\right] = 0, \\
&\sup_{\beta\in\mathcal{T}_n}\left|\left\{\frac{\partial}{\partial\beta}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \Xi_x(Y_i, R_i, X_i; m, \beta)\middle|\boldsymbol{X}\right]\Big|_{(m_0,\beta_0)}\right\}^T(\beta - \beta_0)\right| \le \frac{\delta_n}{\sqrt{n}},
\end{aligned} \tag{2}$$

where $\mathcal{T}_n$ is a properly shrinking neighborhood of β_0 and $\delta_n = o(1)$.

- Both conditions in (2) hold true, because for any $\beta \in \mathcal{T}_n$ and some convex set $\mathcal{B}$ containing β_0, we have that

$$
\begin{aligned}
&\left| \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \mathrm{E}\left[\Xi_x(Y_i, R_i, X_i; m, \beta) | X \right] \Big|_{(m_0, \beta_0)} \right\}^T (\beta - \beta_0) \right| \\
&= \left| \left[x - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \cdot \pi(X_i) X_i \right]^T (\beta_0 - \beta) \right| \\
&\text{“} \leq \text{”} \left\| x - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \cdot \widehat{\pi}_i \cdot X_i \right\|_\infty ||\beta - \beta_0||_1 \quad \text{by Hölder's inequality} \\
&\leq \frac{\gamma}{n} ||\beta - \beta_0||_1 \quad \text{by the box constraint in our debiasing program} \\
&\leq \frac{\delta_n}{\sqrt{n}} \quad \text{by setting } \mathcal{T}_n = \left\{ \beta \in \mathcal{B} \subset \mathbb{R}^d : ||\beta - \beta_0||_1 \leq \frac{\sqrt{n}\delta_n}{\gamma} \right\}.
\end{aligned}
$$

- Our debiasing program optimizes the (estimated) variance among all the estimators satisfying Neyman near-orthogonalization (2).
- (2) also allows our debiasing program to *de-correlate* the Lasso pilot regression from propensity score estimation and weight optimization.

Theoretical Implications of Our Dual Debiasing Program

- **Goal:** Establish the asymptotic normality of our debiased estimator

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

- Linearity assumption $Y_i = X_i^T\beta_0 + \epsilon_i$ for $i = 1, ..., n$ implies

$$\sqrt{n}\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) - m_0(x)\right] = \underbrace{\sum_{i=1}^{n}\widehat{w}_i R_i\epsilon_i}_{\text{Not an i.i.d. sum!}} + \left[x - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{w}_i R_i X_i\right]^T\sqrt{n}\left(\widehat{\beta} - \beta_0\right),$$

- Dual relation $\widehat{w}_i = -\frac{1}{2\sqrt{n}}\cdot X_i^T\widehat{\ell}$ for $i = 1, ..., n$ and dual consistency $\widehat{\ell} \xrightarrow{P} \ell_0$ reveal that

$$\begin{aligned}\sqrt{n}\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) - m_0(x)\right] &= -\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}R_i\epsilon_i X_i^T\widehat{\ell} + \left[x + \frac{1}{2n}\sum_{i=1}^{n}R_i X_i X_i^T\widehat{\ell}\right]^T\sqrt{n}\left(\beta_0 - \widehat{\beta}\right)\\ &= \underbrace{-\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}R_i\epsilon_i X_i^T\ell_0}_{\text{i.i.d. sum!}} + \underbrace{\text{``Bias terms''}}_{o_P(1)}.\end{aligned}$$

1. The covariate vector $X \in \mathbb{R}^d$ and the noise $\epsilon \in \mathbb{R}$ are sub-Gaussian.

2. There exists a constant $\kappa_R > 0$ such that
$$\inf_{v \in \mathbb{S}^{d-1}} \mathrm{E}\left[R(X^T v)^2\right] \geq \kappa_R^2 \quad \text{with} \quad \mathbb{S}^{d-1} = \left\{x \in \mathbb{R}^d : ||x||_2 = 1\right\}.$$

3. Given any $n \geq 1$ and $\delta \in (0,1)$, there exists $r_\pi \equiv r_\pi(n,\delta) > 0$ such that
$$\mathrm{P}\left(\max_{1 \leq i \leq n} |\widehat{\pi}_i - \pi_i| > r_\pi\right) < \delta \quad \text{with} \quad \pi_i = \pi(X_i), i = 1, ..., n.$$

4. Define the population dual program as:
$$\min_{\ell \in \mathbb{R}^d} \left\{\frac{1}{4}\mathrm{E}\left[R\left(X^T \ell\right)^2\right] + x^T \ell\right\},$$
whose exact solution is $\ell_0(x) = -2\left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x$. We assume that the r_ℓ-approximation $\widetilde{\ell}(x)$ to $\ell_0(x)$ is sparse with $r_\ell \in \left[0, \frac{1}{2}\right]$, *i.e.*,
$$s_\ell(x) = \left|\left|\widetilde{\ell}(x)\right|\right|_0 \ll \min\{n, d\} \;\text{ with }\; \widetilde{\ell}(x) = \arg\min_{u \in \mathbb{R}^d}\left\{||u||_0 : ||u - \ell_0(x)||_2 \leq r_\ell \,||\ell_0(x)||_2\right\}.$$

Methods to be compared:

- **"DL-Jav":** The debiased Lasso by Javanmard and Montanari (2014).
- **"DL-vdG":** The debiased Lasso by van de Geer et al. (2014).
- **"Refit":** Run the regular least-square regression on the support set of the Lasso pilot estimate (Belloni and Chernozhukov, 2013).

Implementation settings of the above methods:

- Complete-case (CC) data $\{(X_i, Y_i, R_i = 1)\}_{i=1}^n$;
- Inverse probability weighted (IPW) data $\left\{\left(\frac{X_i}{\sqrt{\widehat{\pi}_i}}, \frac{Y_i}{\sqrt{\widehat{\pi}_i}}, R_i = 1\right)\right\}_{i=1}^n$;
- Oracle fully observed data (X_i, Y_i) for $i = 1, ..., n$.

Evaluation metrics over 1000 Monte Carlo experiments:

- Average absolute bias $\left|\widehat{m}^{\text{debias}}(x) - m_0(x)\right|$;
- Average coverage and average length of the yielded 95% confidence intervals.

Simulation Results Under Gaussian Noises (I)

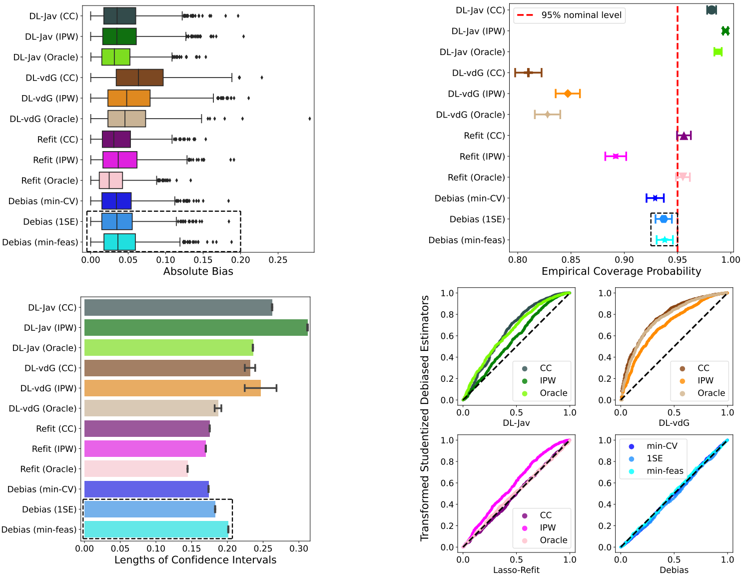


Figure: Sparse β_0^{sp} and sparse $x^{(2)}$ with $X_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma^{\mathrm{cs}}), i = 1, ..., n$.

Simulation Results Under Gaussian Noises (II)

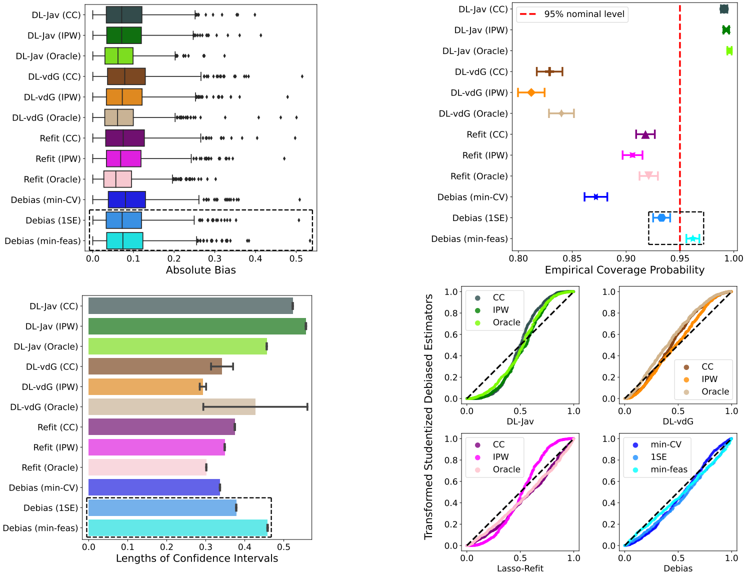


Figure: Pseudo-dense β_0^{pd} and sparse $x^{(2)}$ with $X_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma^{\text{ar}}), i = 1, ..., n$.

Simulation Results Under $\text{Laplace}(0, 1/\sqrt{2})$ Noises

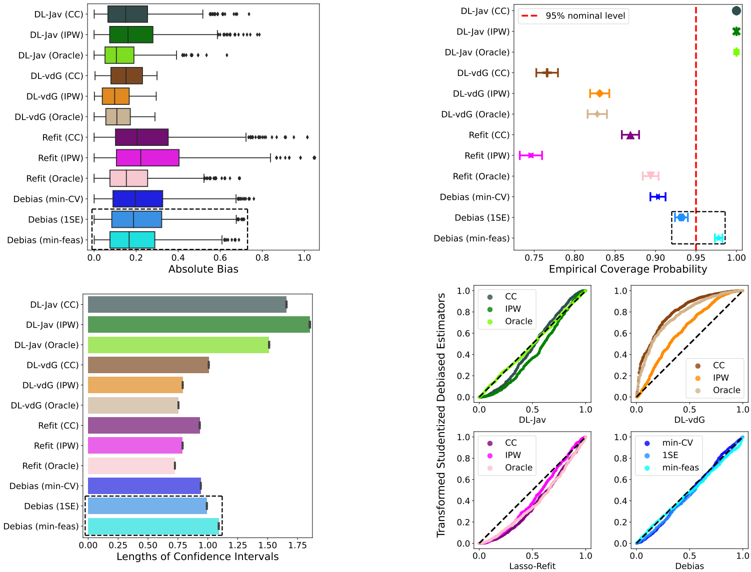


Figure: Dense β_0^{de} and sparse $x^{(2)}$ with $X_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma^{cs}), i = 1, ..., n$.

Simulation Results Under t_2-Distributed Noises

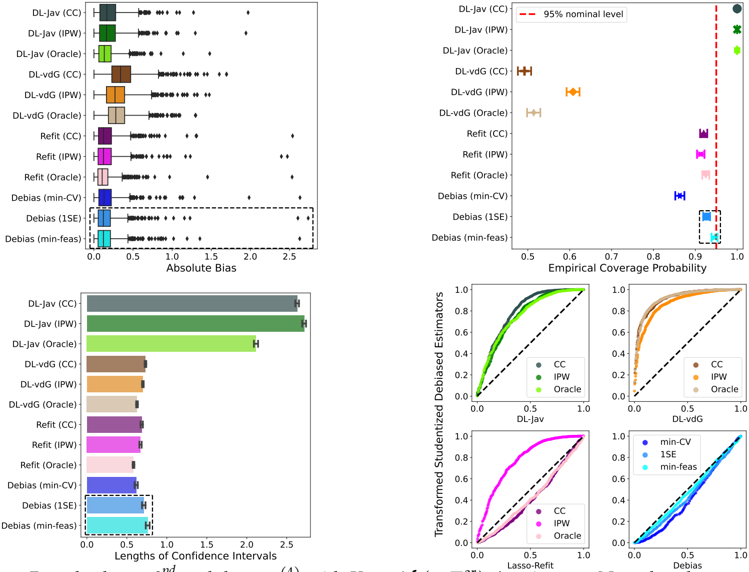


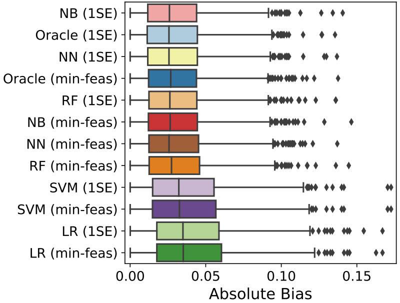
Figure: Pseudo-dense β_0^{pd} and dense $x^{(4)}$ with $X_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma^{\text{ar}}), i = 1, \ldots, n$. Note that the mean-zero t_2 distribution has *infinite* variance.

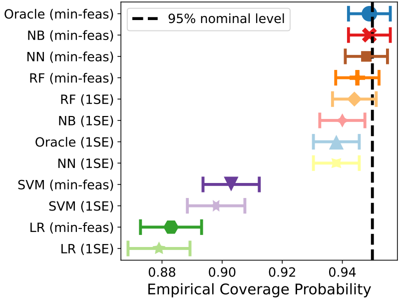
1. True propensity score model: $\mathrm{P}(R_i = 1|X_i) = \Phi\left(-4 + \sum_{k=1}^{K} Z_{ik}\right)$, where $(Z_{i1}, ..., Z_{iK})$ contains all polynomial combinations of the first eight components $X_{i1}, ..., X_{i8}$ of $X_i \in \mathbb{R}^{1000}$ with degrees ≤ 2.

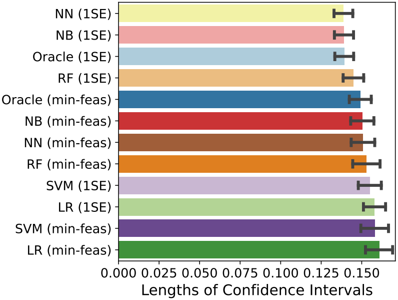
2. Estimate the propensity scores $\pi(X_i), i = 1, ..., n$ by the following nonlinear/nonparametric machine learning methods:
   - **Gaussian Naive Bayes ("NB")**.
   - **Random Forest ("RF"):** 100 trees, bootstrapping samples, and the Gini impurity.
   - **Support Vector Machine ("SVM"):** Gaussian radial basis function.
   - **Neural Network ("NN"):** Two hidden layers of size 80×50 and ReLU $h(x) = \max\{x, 0\}$ as the activation function.

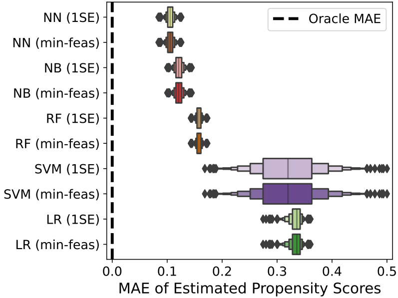
3. Include an extra evaluation metric as the average mean absolute error ("Avg-MAE") for the estimated propensity scores.

Simulation Results With Nonparametric Propensity Scores

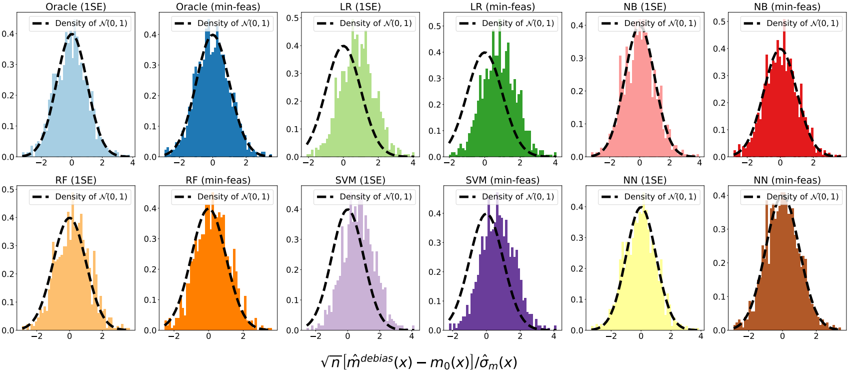








Simulation Results With Nonparametric Propensity Scores



Study Design For Stellar Mass Inference

1. Consider all the observed galaxies by SDSS-IV within a thin redshift slice $0.4 \sim 0.4005$, among which 30.2% of their stellar masses are missing in the Firefly value-added catalog.
2. Fetch their spectroscopic and photometric properties from SDSS-IV DR16 database similar to the input catalog of Chang et al. (2015).
3. Apply feature transformation, remove highly linearly correlated covariates, and generate univariate B-spline base covariates of polynomial order 3 with 40 knots.
4. Incorporate RA, DEC, and the angular diameter distances from the galaxies to the two-dimensional spherical cosmic filaments by Zhang and Chen (2023); Zhang et al. (2022).
5. Control for the confounding effects by including the distances from galaxies to candidate galaxy clusters.

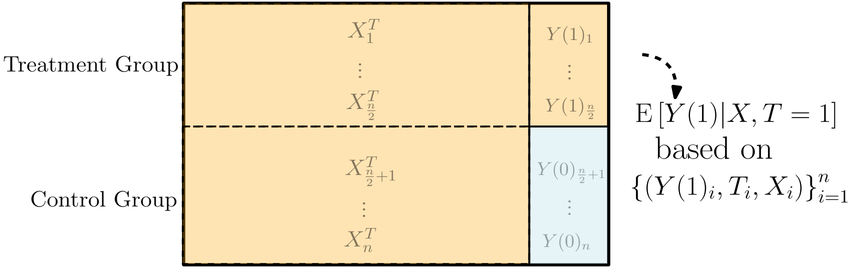
- **Final Dataset:** $n = 1185$ and $d = 1409$.

The observable data in causal inference are

$$\{(\mathbb{Y}_i, T_i, X_i)\}_{i=1}^n \subset \mathbb{R} \times \{0,1\} \times \mathbb{R}^d.$$

- $T_i \in \{0,1\}$ is a binary treatment assignment indicator;
- $\mathbb{Y}_i = T_i \cdot Y(1)_i + (1 - T_i) \cdot Y(0)_i$ with $Y(0), Y(1)$ as potential outcomes.

▶ **Objective:** Conduct valid inference on $\mathrm{E}\left[Y(1)|X, T=1\right]$.



Our debiasing method can be extended to valid inference on the high-dimensional linear average conditional treatment effect (ACTE)

$$\mathrm{E}[Y(1) - Y(0)|X].$$

- The modified debiasing program with tuning parameters $\gamma_1, \gamma_2 > 0$ is

$$\operatorname*{arg\,min}_{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(1)} \in \mathbb{R}^n} \sum_{i=1}^{n} \left[\widehat{\pi}_i w_{i(1)}^2 + (1 - \widehat{\pi}_i) w_{i(0)}^2 \right]$$

$$\text{s.t.} \quad \left\| x - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_{i(1)} \cdot \widehat{\pi}_i \cdot X_i \right\|_{\infty} \leq \frac{\gamma_1}{n} \quad \text{and} \quad \left\| x - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_{i(0)} \left(1 - \widehat{\pi}_i\right) X_i \right\|_{\infty} \leq \frac{\gamma_2}{n}.$$

- The extended debiased estimator becomes

$$\widehat{m}^{\text{debias}}(x; \widehat{\boldsymbol{w}}_{(1)}, \widehat{\boldsymbol{w}}_{(0)})$$
$$= x^T \left(\widehat{\beta}_{(1)} - \widehat{\beta}_{(0)} \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\widehat{w}_{i(1)} \cdot T_i \left(\mathbb{Y}_i - X_i^T \widehat{\beta}_{(1)} \right) - \widehat{w}_{i(0)} \cdot (1 - T_i) \left(\mathbb{Y}_i - X_i^T \widehat{\beta}_{(0)} \right) \right].$$

- The efficiency theory for this modified procedure is worth studying!

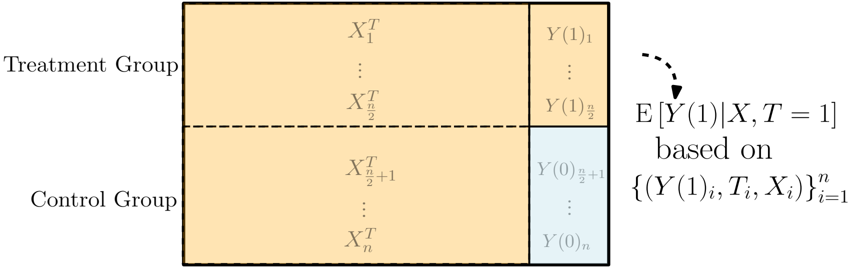
Potential Application to Causal Inference (I)

The observable data in causal inference are

$$\left\{(\mathbb{Y}_i, T_i, X_i)\right\}_{i=1}^n \subset \mathbb{R} \times \{0,1\} \times \mathbb{R}^d.$$

- $T_i \in \{0,1\}$ is a binary treatment assignment indicator;
- $\mathbb{Y}_i = T_i \cdot Y(1)_i + (1 - T_i) \cdot Y(0)_i$ with $Y(0), Y(1)$ as potential outcomes.

▶ **Objective:** Conduct valid inference on $\mathrm{E}\left[Y(1)|X, T=1\right]$.



Potential Application to Causal Inference (II)

Our debiasing method can be extended to valid inference on the high-dimensional linear average conditional treatment effect (ACTE)

$$\mathrm{E}[Y(1) - Y(0)|X].$$

- The modified debiasing program with tuning parameters $\gamma_1, \gamma_2 > 0$ is

$$\mathop{\arg\min}_{\boldsymbol{w}_{(0)}, \boldsymbol{w}_{(1)} \in \mathbb{R}^n} \sum_{i=1}^{n} \left[\widehat{\pi}_i w_{i(1)}^2 + (1 - \widehat{\pi}_i) w_{i(0)}^2 \right]$$

$$\text{s.t. } \left\| x - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_{i(1)} \cdot \widehat{\pi}_i \cdot X_i \right\|_\infty \leq \frac{\gamma_1}{n} \text{ and } \left\| x - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_{i(0)} \left(1 - \widehat{\pi}_i\right) X_i \right\|_\infty \leq \frac{\gamma_2}{n}.$$

- The extended debiased estimator becomes

$$\begin{aligned} &\widehat{m}^{\text{debias}}(x; \widehat{\boldsymbol{w}}_{(1)}, \widehat{\boldsymbol{w}}_{(0)}) \\ &= x^T \left(\widehat{\beta}_{(1)} - \widehat{\beta}_{(0)} \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\widehat{w}_{i(1)} \cdot T_i \left(\mathbb{Y}_i - X_i^T \widehat{\beta}_{(1)} \right) - \widehat{w}_{i(0)} \cdot (1 - T_i) \left(\mathbb{Y}_i - X_i^T \widehat{\beta}_{(0)} \right) \right]. \end{aligned}$$

- The efficiency theory for this modified procedure is worth studying!

The galaxy distribution is distorted along the line of sight due to the peculiar velocities of galaxies, *i.e.*, the so-called *finger-of-god* (Jackson, 1972) and *Kaiser* (Kaiser, 1987) effects.

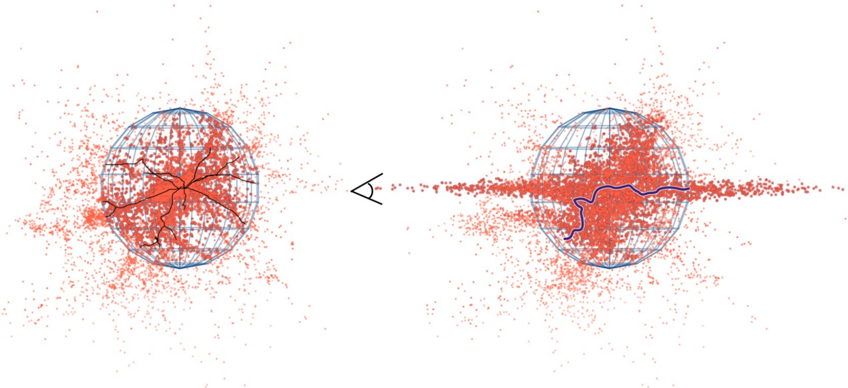


Figure: Redshift distortions along the line of sight (Kuchner et al., 2021).