Efficient Inference on High-Dimensional Linear Models With Missing Outcomes

Yikun Zhang

Joint Work with *Alexander Giessing* and *Yen-Chi Chen*

Department of Statistics,
University of Washington

November 8, 2023 at Casual Inference and Missing Data Reading Group

Department of STATISTICS

UNIVERSITY *of* WASHINGTON

Table of Contents

1. Introduction
2. Methodology: Efficient Debiasing Method
3. Theory: Consistency and Asymptotic Normality
4. Comparative Simulations
5. Real-World Applications: Stellar Mass Inference Problem
6. Conclusions and Future Works

Introduction

Consider a random sample $\{(Y_i, R_i, X_i)\}_{i=1}^n$ drawn from the joint distribution of (Y, R, X), where

- $Y \in \mathbb{R}$ is the outcome variable that could potentially be missing;
- $R \in \{0, 1\}$ is the indicator of Y being observed;
- $X \in \mathbb{R}^d$ is the high-dimensional covariate vector with $d \gg n$.

Consider a random sample $\{(Y_i, R_i, X_i)\}_{i=1}^n$ drawn from the joint distribution of (Y, R, X), where

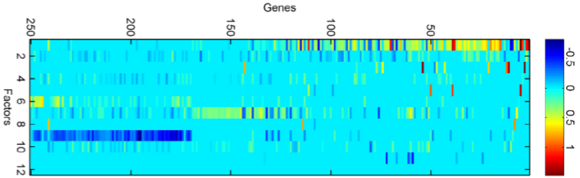
- $Y \in \mathbb{R}$ is the outcome variable that could potentially be missing;
- $R \in \{0, 1\}$ is the indicator of Y being observed;
- $X \in \mathbb{R}^d$ is the high-dimensional covariate vector with $d \gg n$.

▶ **Central Question of Interest:**

How can we conduct statistically and computationally efficient inference on $m_0(x) = \mathrm{E}(Y|X = x)$ despite missing outcomes?

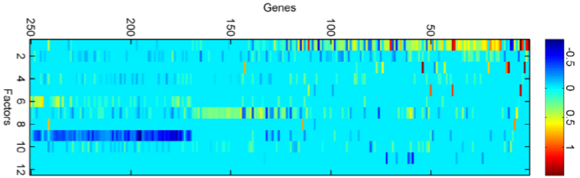
1. The covariates are easier to obtain within some population.

1. The covariates are easier to obtain within some population.



- Micro-array gene expression data in biology (Carvalho et al., 2008).
- Home-price data with cross-sectional effects (Fan et al., 2011).

1. The covariates are easier to obtain within some population.

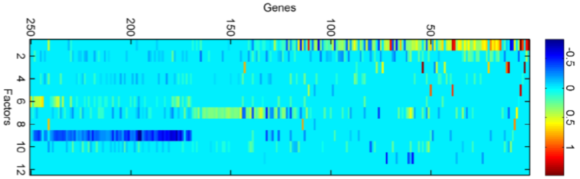


   - Micro-array gene expression data in biology (Carvalho et al., 2008).
   - Home-price data with cross-sectional effects (Fan et al., 2011).

2. Incorporating as many covariates as possible can control for potential confounders in causal inference (Wyss et al., 2022).

Motivations: High-Dimensional Data

1. The covariates are easier to obtain within some population.



   - Micro-array gene expression data in biology (Carvalho et al., 2008).
   - Home-price data with cross-sectional effects (Fan et al., 2011).

2. Incorporating as many covariates as possible can control for potential confounders in causal inference (Wyss et al., 2022).

3. Generating high-dimensional covariates with interaction terms or spline features enables the simple parametric (e.g., linear) model to capture complex patterns (Belloni et al., 2019).

Motivations: Missing Outcomes

The response/outcome variable in observational data could be missing.

The response/outcome variable in observational data could be missing.

1. Participants may drop out from the study in clinical trials (Higgins et al., 2008).

The response/outcome variable in observational data could be missing.

1. Participants may drop out from the study in clinical trials (Higgins et al., 2008).

2. The semi-supervised learning, where additional samples without labels are provided, is a missing-outcome problem (Chapelle et al., 2006).

The response/outcome variable in observational data could be missing.

1. Participants may drop out from the study in clinical trials (Higgins et al., 2008).
2. The semi-supervised learning, where additional samples without labels are provided, is a missing-outcome problem (Chapelle et al., 2006).

► **More Concrete Example:** Some (estimated) stellar masses of the observed galaxies in the Sloan Digital Sky Survey (SDSS-IV) are missing in the Firefly value-added catalog (Comparat et al., 2017).

Motivations: Stellar Mass Inference Problem

The missingness of (estimated) stellar masses is due to

- Limiting usage of the observational run in SDSS-IV for galaxy targets;
- Potential data contamination;
- Misclassification of galaxies as stars.

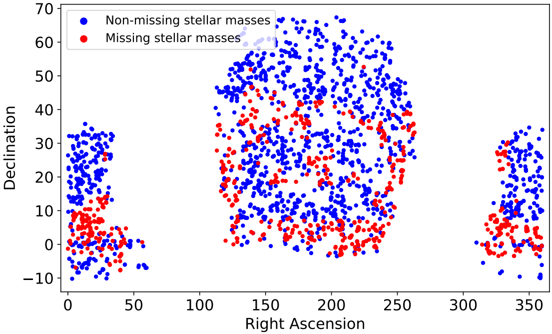


Figure 1: Galaxy distribution at a high redshift slice $0.4 \sim 0.401$.

The missingness of (estimated) stellar masses is due to

- Limiting usage of the observational run in SDSS-IV for galaxy targets;
- Potential data contamination;
- Misclassification of galaxies as stars.

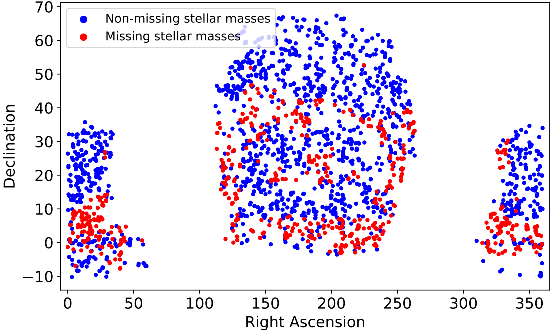


Figure 1: Galaxy distribution at a high redshift slice $0.4 \sim 0.401$.

► **Scientific Question:** *How can we conduct valid inference on the (estimated) stellar mass based on the spectroscopic and photometric properties?*

Challenges and Model Assumptions

To tackle the challenges of high-dimensional data with missing outcomes, we impose two basic assumptions.

To tackle the challenges of high-dimensional data with missing outcomes, we impose two basic assumptions.

1. *(Linearity)* The data $\{(Y_i, R_i, X_i)\}_{i=1}^n \subset \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ 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 $||\beta_0||_0 = \sum_{k=1}^d \mathbb{1}_{\{\beta_{0k} \neq 0\}} = s_\beta \ll d$.

Challenges and Model Assumptions

To tackle the challenges of high-dimensional data with missing outcomes, we impose two basic assumptions.

1. *(Linearity)* The data $\{(Y_i, R_i, X_i)\}_{i=1}^n \subset \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ 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 $||\beta_0||_0 = \sum_{k=1}^d \mathbb{1}_{\{\beta_{0k} \neq 0\}} = s_\beta \ll d$.

- **Notes:** The linearity assumption can be relaxed to
  - Sparse additive model (Ravikumar et al., 2009);
  - Partially linear model (Müller and van de Geer, 2015);
  - Approximately/weakly sparse linear model (Belloni et al., 2019).

Challenges and Model Assumptions

To tackle the challenges of high-dimensional data with missing outcomes, we impose two basic assumptions.

1. *(Linearity)* The data $\{(Y_i, R_i, X_i)\}_{i=1}^n \subset \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ 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 $||\beta_0||_0 = \sum_{k=1}^d \mathbb{1}_{\{\beta_{0k} \neq 0\}} = s_\beta \ll d$.

- **Notes:** The linearity assumption can be relaxed to
  - Sparse additive model (Ravikumar et al., 2009);
  - Partially linear model (Müller and van de Geer, 2015);
  - Approximately/weakly sparse linear model (Belloni et al., 2019).

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

Existing Works on High-Dimensional Inference

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

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

1. (*Fully observed outcomes*) Debiased Lasso is applicable (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014):

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

   - $\widehat{\beta}_\lambda = \mathop{\arg\min}_{\beta\in\mathbb{R}^d}\left[\frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\left|\left|\beta\right|\right|_1\right]$ is a Lasso solution with 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 mainly on the statistical inference on $\beta_0 \in \mathbb{R}^d$.

1. (*Fully observed outcomes*) Debiased Lasso is applicable (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014):

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

   - $\widehat{\beta}_\lambda = \arg\min_{\beta\in\mathbb{R}^d}\left[\frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\,||\beta||_1\right]$ is a Lasso solution with the regularization parameter $\lambda > 0$;
   - $\widehat{\Theta} \in \mathbb{R}^{d\times d}$ is an approximation to the matrix inverse $\left(\frac{1}{n}\sum_{i=1}^{n} X_iX_i^T\right)^{-1}$.

2. (*MAR outcomes*) Chakrabortty et al. (2019) proposed an M-estimation framework with a Lasso-type debiased and doubly robust estimator.

Drawback of Existing Works and Our Contributions

- **Drawbacks of the Existing Approaches:**

1. (*Computational issue*) They require a good approximation to the $d \times d$ debiasing matrix $\widehat{\Theta}$.

2. (*Loss of statistical efficiency*) Sample splitting or cross-fitting is necessary for the M-estimation framework.

Drawback of Existing Works and Our Contributions

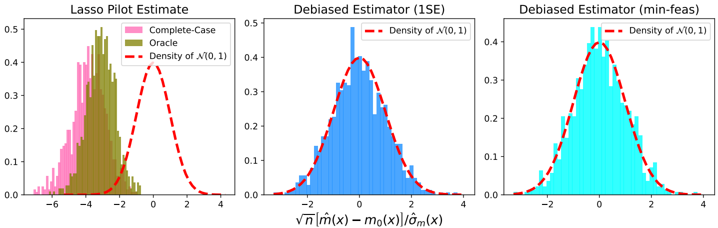
- **Drawbacks of the Existing Approaches:**

1. (*Computational issue*) They require a good approximation to the $d \times d$ debiasing matrix $\widehat{\Theta}$.
2. (*Loss of statistical efficiency*) Sample splitting or cross-fitting is necessary for the M-estimation framework.

- **Our Contributions:** Focus on the inference of $m_0(x) = x^T \beta_0$ instead.

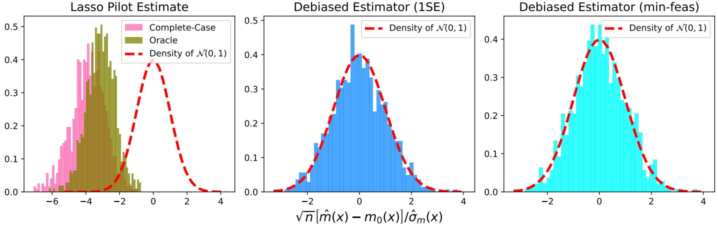
  - (*Computational efficiency*) Our core debiasing program is convex and only needs to solve for a n-dimensional weight vector.
  - (*Statistical efficiency*) Our debiased estimator is semi-parametrically efficient among all asymptotically linear estimators.

Highlights of Today's Presentation



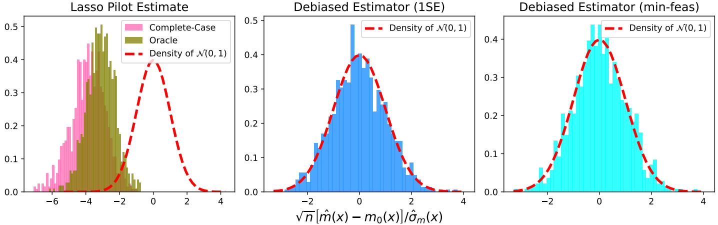
1. Introduce our efficient debiasing method for inferring $m_0(x) = x^T\beta_0$.

Highlights of Today's Presentation



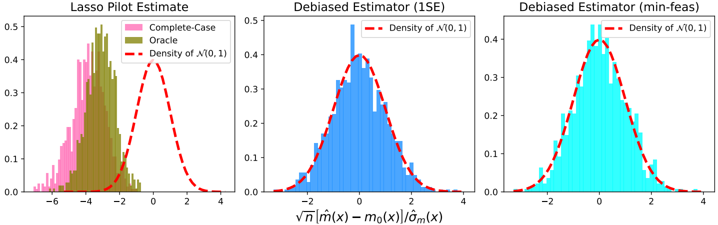
1. Introduce our efficient debiasing method for inferring $m_0(x) = x^T \beta_0$.
   - Estimate $\pi(X) = \mathrm{P}(R = 1|X)$ via any machine learning methods.
   - Design our debiasing program based on bias-variance trade-offs.
   - Fine-tune the program from its dual so as to debias the Lasso solution.

Highlights of Today's Presentation



1. Introduce our efficient debiasing method for inferring $m_0(x) = x^T\beta_0$.
   - Estimate $\pi(X) = \mathrm{P}(R = 1|X)$ via any machine learning methods.
   - Design our debiasing program based on bias-variance trade-offs.
   - Fine-tune the program from its dual so as to debias the Lasso solution.
2. Discuss the asymptotic normality and semi-parametric efficiency of our final debiased estimator.

Highlights of Today's Presentation



1. Introduce our efficient debiasing method for inferring $m_0(x) = x^T\beta_0$.
   - Estimate $\pi(X) = \mathrm{P}(R = 1|X)$ via any machine learning methods.
   - Design our debiasing program based on bias-variance trade-offs.
   - Fine-tune the program from its dual so as to debias the Lasso solution.
2. Discuss the asymptotic normality and semi-parametric efficiency of our final debiased estimator.
3. Demonstrate the finite-sample performances via simulations and present an application to the stellar mass inference problem.

Methodology

Heuristics From Debiased Lasso

For any fixed $\lambda > 0$, the Lasso solution (on the complete-case data) is a biased estimator of $\beta_0 \in \mathbb{R}^d$:

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

- **Question:** How can we correct for the bias in $\widehat{\beta}_\lambda$ or $\widehat{m}(x) = x^T \widehat{\beta}_\lambda$?

Heuristics From Debiased Lasso

For any fixed $\lambda > 0$, the Lasso solution (on the complete-case data) is a biased estimator of $\beta_0 \in \mathbb{R}^d$:

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

▶ **Question:** How can we correct for the bias in $\widehat{\beta}_\lambda$ or $\widehat{m}(x) = x^T \widehat{\beta}_\lambda$?

- Optimality/KKT condition reads

$$\frac{1}{n} \sum_{i=1}^{n} R_i X_i \left(Y_i - X_i^T \widehat{\beta}_\lambda \right) = \lambda \widehat{z} \quad \text{with} \quad \widehat{z} \in \partial \left|\left| \widehat{\beta}_\lambda \right|\right|_1 \in \mathbb{R}^d. \qquad (1)$$

For any fixed $\lambda > 0$, the Lasso solution (on the complete-case data) is a biased estimator of $\beta_0 \in \mathbb{R}^d$:

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

▶ **Question:** How can we correct for the bias in $\widehat{\beta}_\lambda$ or $\widehat{m}(x) = x^T \widehat{\beta}_\lambda$?

- Optimality/KKT condition reads

$$\frac{1}{n} \sum_{i=1}^n R_i X_i \left(Y_i - X_i^T \widehat{\beta}_\lambda \right) = \lambda \widehat{z} \quad \text{with} \quad \widehat{z} \in \partial \left|\left| \widehat{\beta}_\lambda \right|\right|_1 \in \mathbb{R}^d. \tag{1}$$

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

$$\frac{1}{n} \sum_{i=1}^n R_i X_i \epsilon_i + \widehat{\Sigma} \left(\beta_0 - \widehat{\beta}_\lambda \right) = \lambda \widehat{z} \quad \text{with} \quad \widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n R_i X_i X_i^T.$$

Heuristics From Debiased Lasso

- Given an approximation $\widehat{\Theta} \in \mathbb{R}^{d\times d}$ to $\widehat{\Sigma}^{-1}$, it becomes

$$\widehat{\beta}_\lambda - \beta_0 + \widehat{\Theta}\lambda\widehat{z} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} R_i \widehat{\Theta} X_i \epsilon_i}_{\text{Stochastic error} \sim \mathcal{N}_d(0,\widetilde{\Sigma})} + \underbrace{\left(\widehat{\Theta}\widehat{\Sigma} - I_d\right)\left(\beta_0 - \widehat{\beta}_\lambda\right)}_{\text{Asymptotically negligible bias}}.$$

Heuristics From Debiased Lasso

- Given an approximation $\widehat{\Theta} \in \mathbb{R}^{d\times d}$ to $\widehat{\Sigma}^{-1}$, it becomes

$$\widehat{\beta}_\lambda - \beta_0 + \widehat{\Theta}\lambda\widehat{z} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} R_i \widehat{\Theta} X_i \epsilon_i}_{\text{Stochastic error } \sim \mathcal{N}_d(0,\widetilde{\Sigma})} + \underbrace{\left(\widehat{\Theta}\widehat{\Sigma} - I_d\right)\left(\beta_0 - \widehat{\beta}_\lambda\right)}_{\text{Asymptotically negligible bias}}.$$

- By KKT condition (1), the debiased Lasso estimate is thus given by

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

Heuristics From Debiased Lasso

- Given an approximation $\widehat{\Theta} \in \mathbb{R}^{d\times d}$ to $\widehat{\Sigma}^{-1}$, it becomes

$$\widehat{\beta}_\lambda - \beta_0 + \widehat{\Theta}\lambda\widehat{z} = \underbrace{\frac{1}{n}\sum_{i=1}^n R_i \widehat{\Theta} X_i \epsilon_i}_{\text{Stochastic error } \sim \mathcal{N}_d(0,\widetilde{\Sigma})} + \underbrace{\left(\widehat{\Theta}\widehat{\Sigma} - I_d\right)\left(\beta_0 - \widehat{\beta}_\lambda\right)}_{\text{Asymptotically negligible bias}}.$$

- By KKT condition (1), the debiased Lasso estimate is thus given by

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

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

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

Heuristics From Debiased Lasso

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

- **Issue:** Fitting the debiasing matrix $\widehat{\Theta} \in \mathbb{R}^{d \times d}$ is computationally inefficient; see, *e.g.*, the nodewise regression (Meinshausen and Bühlmann, 2006; van de Geer et al., 2014).

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

- **Issue:** Fitting the debiasing matrix $\widehat{\Theta} \in \mathbb{R}^{d \times d}$ is computationally inefficient; see, *e.g.*, the nodewise regression (Meinshausen and Bühlmann, 2006; van de Geer et al., 2014).

- **Solution:** Introduce the weight vector $\widehat{\boldsymbol{w}} = (\widehat{w}_1, ..., \widehat{w}_n)^T \in \mathbb{R}^n$ with (Giessing and Wang, 2023)

$$\widehat{w}_i = \begin{cases} \frac{1}{\sqrt{n}} x^T \widehat{\Theta} X_i & R_i = 1, \\ 0 & R_i = 0, \end{cases}$$

for $i = 1, ..., n$ so that our final debiased estimator becomes

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

Heuristics From Debiased Lasso

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

▶ **Issue:** Fitting the debiasing matrix $\widehat{\Theta} \in \mathbb{R}^{d\times d}$ is computationally inefficient; see, *e.g.*, the nodewise regression (Meinshausen and Bühlmann, 2006; van de Geer et al., 2014).

▶ **Solution:** Introduce the weight vector $\widehat{\boldsymbol{w}} = (\widehat{w}_1, ..., \widehat{w}_n)^T \in \mathbb{R}^n$ with (Giessing and Wang, 2023)

$$\widehat{w}_i = \begin{cases} \frac{1}{\sqrt{n}} x^T \widehat{\Theta} X_i & R_i = 1, \\ 0 & R_i = 0, \end{cases}$$

for $i = 1, ..., n$ so that our final debiased estimator becomes

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

▶ **Question:** How do we estimate the weight vector $\widehat{\boldsymbol{w}} = (\widehat{w}_1, ..., \widehat{w}_n)^T$?

Optimize Over the Conditional Mean Square Error

Consider the generic debiased estimator $m^{\text{debias}}(x;\boldsymbol{w})$ from (2) as:

$$m^{\text{debias}}(x;\boldsymbol{w}) = x^T\beta + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i R_i \left(Y_i - X_i^T\beta\right). \tag{3}$$

Optimize Over the Conditional Mean Square Error

Consider the generic debiased estimator $m^{\text{debias}}(x;\boldsymbol{w})$ from (2) as:

$$m^{\text{debias}}(x;\boldsymbol{w}) = x^T\beta + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i R_i \left(Y_i - X_i^T\beta\right). \tag{3}$$

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

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

where $\pi(X) = \mathrm{P}(R = 1|X)$ is the propensity score under MAR condition.

Optimize Over the Conditional Mean Square Error

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

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

- By Hölder's inequality, the "Conditional Bias" is upper bounded by

$$\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i\pi(X_i)X_i - x\right\|_\infty\sqrt{n}\left|\left|\beta_0-\beta\right|\right|_1\right]^2.$$

Optimize Over the Conditional Mean Square Error

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

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

- By Hölder's inequality, the "Conditional Bias" is upper bounded by

$$\left[\left\|\frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \pi(X_i) X_i - x\right\|_\infty \sqrt{n}\, ||\beta_0 - \beta||_1\right]^2.$$

- We design our core debiasing program as:

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

where $\gamma > 0$ is a tuning parameter and $\widehat{\pi}_i$ is a consistent estimate of the propensity score $\pi(X_i)$ for $i = 1, ..., n$.

Proposed Debiasing Inference Procedure

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

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

Proposed 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* based on $\{(X_i, R_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{0, 1\}$.

Proposed 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* based on $\{(X_i, R_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{0, 1\}$.

3. Solve the debiasing program defined as:

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

Proposed 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* based on $\{(X_i, R_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{0,1\}$.
3. Solve the debiasing program defined as:

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

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

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

Proposed Debiasing Inference Procedure

1. Compute the Lasso pilot estimate $\widehat{\beta}_\lambda$ on the complete-case data
$$\widehat{\beta}_\lambda = \arg\min_{\beta \in \mathbb{R}^d} \left[\frac{1}{2n} \sum_{i=1}^n R_i (Y_i - X_i^T \beta)^2 + \lambda \|\beta\|_1 \right].$$
2. Obtain consistent propensity score estimates $\widehat{\pi}_i, i = 1, ..., n$ by *any machine learning method* based on $\{(X_i, R_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{0, 1\}$.
3. Solve the debiasing program defined as:
$$\min_{w \in \mathbb{R}^n} \left\{ \sum_{i=1}^n \widehat{\pi}_i w_i^2 : \left\| x - \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i \cdot \widehat{\pi}_i \cdot X_i \right\|_\infty \leq \frac{\gamma}{n} \right\}.$$
4. Define the debiased estimator for $m_0(x)$ as:
$$\widehat{m}^{\text{debias}}(x; \widehat{\boldsymbol{w}}) = x^T \widehat{\beta} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{w}_i R_i \left(Y_i - X_i^T \widehat{\beta} \right).$$
5. Construct the asymptotic $(1 - \tau)$-level confidence interval for $m_0(x)$ as:
$$\left[\widehat{m}^{\text{debias}}(x; \widehat{\boldsymbol{w}}) \pm \Phi^{-1} \left(1 - \frac{\tau}{2} \right) \cdot \widehat{\sigma}_\epsilon \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n \widehat{\pi}_i \widehat{w}_i^2} \right] \quad \text{with } \Phi(\cdot) \text{ being the CDF of } \mathcal{N}(0, 1).$$

Theory and Practice of Our Debiasing Program

There are two unanswered questions in our proposed debiasing inference procedure:

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

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

2. Why is the asymptotic $(1-\tau)$-level confidence interval for $m_0(x)$ valid?

$$\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) \pm \Phi^{-1}\left(1-\frac{\tau}{2}\right)\cdot\widehat{\sigma}_\epsilon\cdot\sqrt{\frac{1}{n}\sum_{i=1}^n \widehat{\pi}_i \widehat{w}_i^2}\right] \quad \text{with } \Phi(\cdot) \text{ being the CDF of } \mathcal{N}(0,1).$$

There are two unanswered questions in our proposed debiasing inference procedure:

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

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

2. Why is the asymptotic $(1-\tau)$-level confidence interval for $m_0(x)$ valid?

$$\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) \pm \Phi^{-1}\left(1-\frac{\tau}{2}\right)\cdot\widehat{\sigma}_\epsilon\cdot\sqrt{\frac{1}{n}\sum_{i=1}^n \widehat{\pi}_i \widehat{w}_i^2}\right] \quad \text{with } \Phi(\cdot) \text{ being the CDF of } \mathcal{N}(0,1).$$

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

Dual Formulation of Our Debiasing Program

The primal form of our debiasing program is a quadratic programming problem with a box constraint:

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

The primal form of our debiasing program is a quadratic programming problem with a box constraint:

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

Proposition (Proposition 1 in Zhang et al. 2023)

The dual form of our debiasing program is given by

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

If the strong duality holds, we further have that

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

where $\widehat{\boldsymbol{w}} \in \mathbb{R}^n$ and $\widehat{\ell} \in \mathbb{R}^d$ are the solutions to the primal and dual debiasing program, respectively.

Practical Implication of Our Dual Debiasing Program

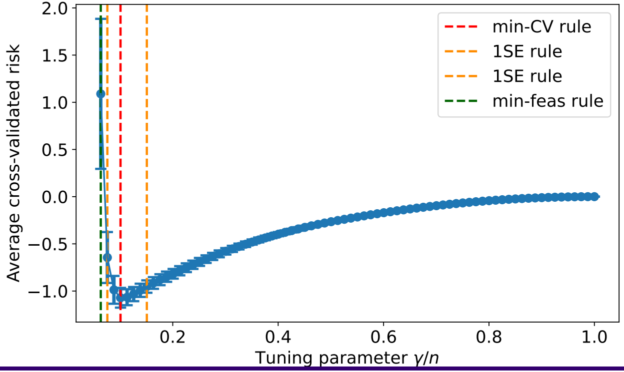
The dual form of our debiasing program is an *unconstrained* quadratic programming problem:

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

The dual form of our debiasing program is an *unconstrained* quadratic programming problem:

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

We can fine-tune $\gamma > 0$ by cross-validation.



- 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

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

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

$$\begin{aligned} &\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \Xi_x(Y_i, R_i, X_i; m_0, \beta_0)\middle|\boldsymbol{X}\right] = 0, \\ &\sup_{\beta\in\mathcal{T}_n}\left|\left\{\frac{\partial}{\partial\beta}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \Xi_x(Y_i, R_i, X_i; m, \beta)\middle|\boldsymbol{X}\right]\Bigg|_{(m_0,\beta_0)}\right\}^T(\beta-\beta_0)\right| \leq \frac{\delta_n}{\sqrt{n}}, \end{aligned} \tag{4}$$

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

- Both conditions in (4) hold true, because for any $\beta \in \mathcal{T}_n$ and some convex set $\mathcal{B}$ containing β_0, we have that

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

- Both conditions in (4) hold true, because for any $\beta \in \mathcal{T}_n$ and some convex set $\mathcal{B}$ containing β_0, we have that

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

- Our debiasing program optimizes the (estimated) variance among all the estimators satisfying Neyman near-orthogonalization (4).

- Both conditions in (4) hold true, because for any $\beta \in \mathcal{T}_n$ and some convex set $\mathcal{B}$ containing β_0, we have that

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

- Our debiasing program optimizes the (estimated) variance among all the estimators satisfying Neyman near-orthogonalization (4).
- (4) also allows our debiasing program to *de-correlate* the Lasso pilot regression from propensity score estimation and weight optimization.

Asymptotic Theory

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

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

Theoretical Implication of Our Dual Debiasing Program

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

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

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

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

Theoretical Implication of Our Dual Debiasing Program

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

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

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

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

- **Solution:** With the dual relation $\widehat{w}_i = -\frac{1}{2\sqrt{n}}\cdot X_i^T\widehat{\ell}, i = 1, ..., n$, we obtain

$$\begin{aligned}\sqrt{n}\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) - m_0(x)\right] &= -\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}R_i\epsilon_i X_i^T\widehat{\ell} + \left[x + \frac{1}{2n}\sum_{i=1}^{n}R_i X_i X_i^T\widehat{\ell}\right]^T\sqrt{n}\left(\beta_0 - \widehat{\beta}\right)\\ &= \underbrace{-\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}R_i\epsilon_i X_i^T\ell_0(x)}_{\text{i.i.d. sum!}} + \underbrace{\text{“Bias terms”}}_{o_P(1)}.\end{aligned}$$

Regularity Conditions For the Asymptotic Theory

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

Regularity Conditions For the Asymptotic Theory

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

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

Regularity Conditions For the Asymptotic Theory

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

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

Regularity Conditions For the Asymptotic Theory

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.*,
$$s_\ell(x) = \left|\left|\widetilde{\ell}(x)\right|\right|_0 \ll \min\{n, d\} \;\text{ with }\; \widetilde{\ell}(x) = \arg\min_{u \in \mathbb{R}^d}\left\{||u||_0 : ||u - \ell_0(x)||_2 \leq r_\ell \, ||\ell_0(x)||_2\right\}.$$

1. **Consistency of Lasso pilot estimate:** If $\lambda \asymp \sigma_\epsilon \sqrt{\frac{\log d}{n}}$ with $\log d = o(n)$, then $\left|\left|\widehat{\beta} - \beta_0\right|\right|_2 = O_P\left(\frac{1}{\kappa_R^2}\sqrt{\frac{s_\beta \log d}{n}}\right)$.

2. **Consistency of the solution to the dual debiasing program:** If r_ℓ shrinks to 0 in a certain rate and $\frac{\gamma}{n} \asymp \frac{||x||_2}{\kappa_R}\sqrt{\frac{\log d}{n}} + \frac{||x||_2}{\kappa_R^2} \cdot r_\pi$, then

$$\left|\left|\widehat{\ell}(x) - \ell_0(x)\right|\right|_2 = O_P\left(\frac{1}{\kappa_R^3}\sqrt{\frac{s_\ell(x)\log d}{n}} + \frac{r_\ell}{\kappa_R^4} + \frac{r_\pi\sqrt{s_\ell(x)}}{\kappa_R^4}\right).$$

Note: Under the same choice of $\gamma > 0$, the strong duality holds.

1. **Consistency of Lasso pilot estimate:** If $\lambda \asymp \sigma_\epsilon \sqrt{\frac{\log d}{n}}$ with $\log d = o(n)$, then $\left|\left|\widehat{\beta} - \beta_0\right|\right|_2 = O_P\left(\frac{1}{\kappa_R^2}\sqrt{\frac{s_\beta \log d}{n}}\right)$.

2. **Consistency of the solution to the dual debiasing program:** If r_ℓ shrinks to 0 in a certain rate and $\frac{\gamma}{n} \asymp \frac{||x||_2}{\kappa_R}\sqrt{\frac{\log d}{n}} + \frac{||x||_2}{\kappa_R^2} \cdot r_\pi$, then

$$\left|\left|\widehat{\ell}(x) - \ell_0(x)\right|\right|_2 = O_P\left(\frac{1}{\kappa_R^3}\sqrt{\frac{s_\ell(x)\log d}{n}} + \frac{r_\ell}{\kappa_R^4} + \frac{r_\pi \sqrt{s_\ell(x)}}{\kappa_R^4}\right).$$

Note: Under the same choice of $\gamma > 0$, the strong duality holds.

Theorem (Theorem 7 in Zhang et al. 2023)

If $\frac{(1+\kappa_R^2)s_{\max}\log(nd)}{\kappa_R^4} = o\left(\sqrt{n}\right)$, $\frac{(1+\kappa_R^4)\sqrt{s_{\max}\log(nd)}}{\kappa_R^6}(r_\ell + r_\pi) = o(1)$, *and* $||x||_2 = O(1)$ *with* $s_{\max} = \{s_\beta, s_\ell(x)\}$, *then*

$$\frac{\sqrt{n}\left[\widehat{m}^{\text{debias}}(x;\widehat{\boldsymbol{w}}) - m_0(x)\right]}{\sigma_m(x)} \xrightarrow{d} \mathcal{N}(0,\,1).$$

Remarks on Our Theoretical Results

1. Our growth requirement $s_{\max} = o\left(\frac{\sqrt{n}}{\log d}\right)$ on the sparsity level is a standard and *essentially necessary* condition for asymptotic normality; see Section 8.6 of Jankova and van de Geer (2018).

Remarks on Our Theoretical Results

1. Our growth requirement $s_{\max} = o\left(\frac{\sqrt{n}}{\log d}\right)$ on the sparsity level is a standard and *essentially necessary* condition for asymptotic normality; see Section 8.6 of Jankova and van de Geer (2018).

2. Given any dimension $d > 0$, the asymptotic variance of our debiased estimator
$$\sigma_m^2(x) = \sigma_\epsilon^2 \cdot x^T \left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x$$
attains the *semi-parametric efficiency bound* among all asymptotically linear estimators under MAR outcomes (Müller and Keilegom, 2012).

Remarks on Our Theoretical Results

1. Our growth requirement $s_{\max} = o\left(\frac{\sqrt{n}}{\log d}\right)$ on the sparsity level is a standard and *essentially necessary* condition for asymptotic normality; see Section 8.6 of Jankova and van de Geer (2018).

2. Given any dimension $d > 0$, the asymptotic variance of our debiased estimator
$$\sigma_m^2(x) = \sigma_\epsilon^2 \cdot x^T \left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x$$
attains the *semi-parametric efficiency bound* among all asymptotically linear estimators under MAR outcomes (Müller and Keilegom, 2012).

Proposition (Proposition 8 in Zhang et al. 2023)

If $\frac{(1+\kappa_R^3)}{\kappa_R^5}\sqrt{\frac{s_\ell(x)\log(nd)}{n}} = o(1)$, $\frac{(1+\kappa_R^4)}{\kappa_R^6}\left[r_\ell + r_\pi\sqrt{s_\ell(x)}\right] = o(1)$, *and* $||x||_2 = O(1)$, *then*
$$\left|\sum_{i=1}^n \widehat{\pi}_i \widehat{w}_i^2 - x^T\left[\mathrm{E}\left(RXX^T\right)\right]^{-1} x\right| = o_P(1).$$

Overfitting the Propensity Scores

Our theoretical results also provide insightful answers to the following two questions:

- Why don't we need sample splitting or cross fitting?
- Why can we estimate the propensity score by any machine learning methods without worrying about the overfitting issue?

Overfitting the Propensity Scores

Our theoretical results also provide insightful answers to the following two questions:

- Why don't we need sample splitting or cross fitting?
- Why can we estimate the propensity score by any machine learning methods without worrying about the overfitting issue?

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

$$\mathrm{P}\left(\max_{1\le i\le n} |\widehat{\pi}_i - \pi_i| > r_\pi\right) < \delta \quad \text{with} \quad \pi_i = \pi(X_i), i = 1, ..., n.$$

- In other words, our debiased estimator performs even better when we overfit the propensity scores $\pi(X_i) = \mathrm{P}(R_i = 1|X_i), i = 1, ..., n$.

Overfitting the Propensity Scores

Our theoretical results also provide insightful answers to the following two questions:

- Why don't we need sample splitting or cross fitting?
- Why can we estimate the propensity score by any machine learning methods without worrying about the overfitting issue?

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

$$\mathrm{P}\left(\max_{1\le i\le n} |\widehat{\pi}_i - \pi_i| > r_\pi\right) < \delta \quad \text{with} \quad \pi_i = \pi(X_i), i = 1, ..., n.$$

- In other words, our debiased estimator performs even better when we overfit the propensity scores $\pi(X_i) = \mathrm{P}(R_i = 1|X_i), i = 1, ..., n$.
- This coincides with *"benign overfitting"* in linear regression or neural networks (Bartlett et al., 2020; Li et al., 2021; Cao et al., 2022).

Comparative Simulations

Experimental Setups and Evaluation Metrics

We compare our debiasing method with L_1-penalized logistic regression for the propensity score estimation with several existing methods:

- **"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).

Experimental Setups and Evaluation Metrics

We compare our debiasing method with L_1-penalized logistic regression for the propensity score estimation with several existing methods:

- **"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).

These methods to be compared are implemented on

- 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 $i = 1, ..., n$.

Experimental Setups and Evaluation Metrics

We compare our debiasing method with L_1-penalized logistic regression for the propensity score estimation with several existing methods:

- **"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).

These methods to be compared are implemented on

- 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 on 1000 Monte Carlo experiments include

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

Simulation Results Under Gaussian Noises (I)

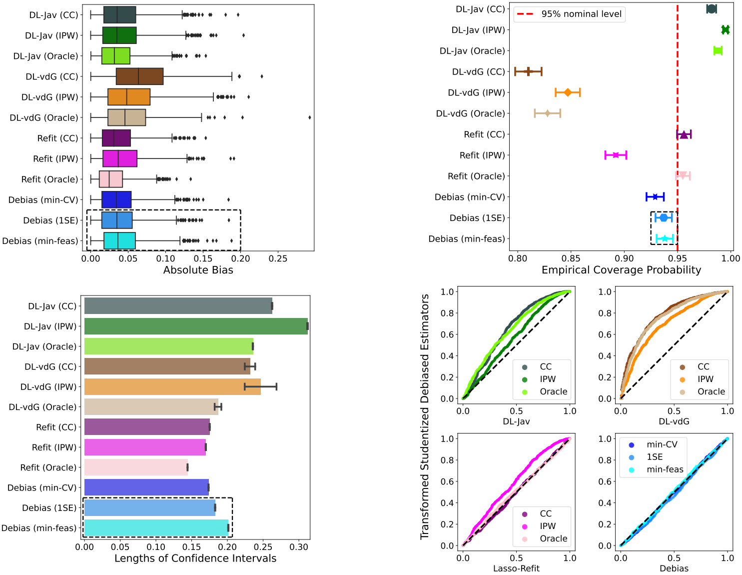


Figure 2: Sparse β_0^{sp} and sparse $x^{(2)}$ with $X_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma^{\mathrm{cs}}), i = 1, ..., n$.

Simulation Results Under Gaussian Noises (II)

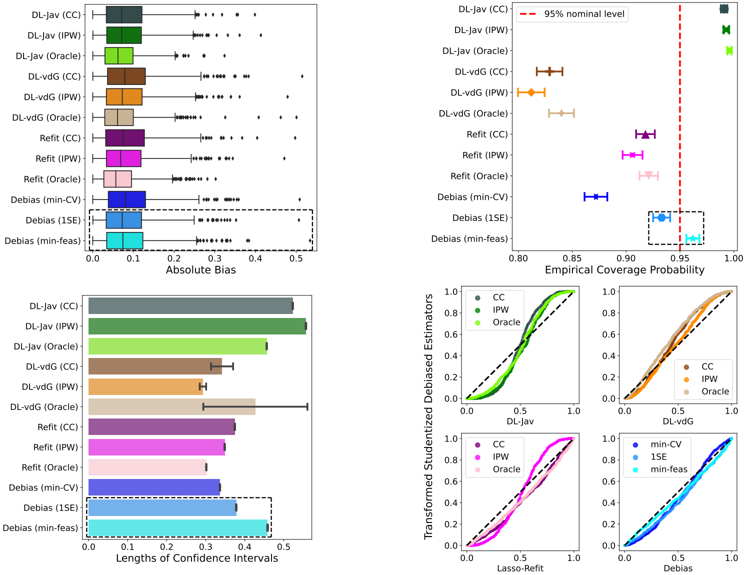


Figure 3: Pseudo-dense β_0^{pd} and sparse $x^{(2)}$ with $X_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma^{\text{ar}}), i = 1, ..., n$.

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

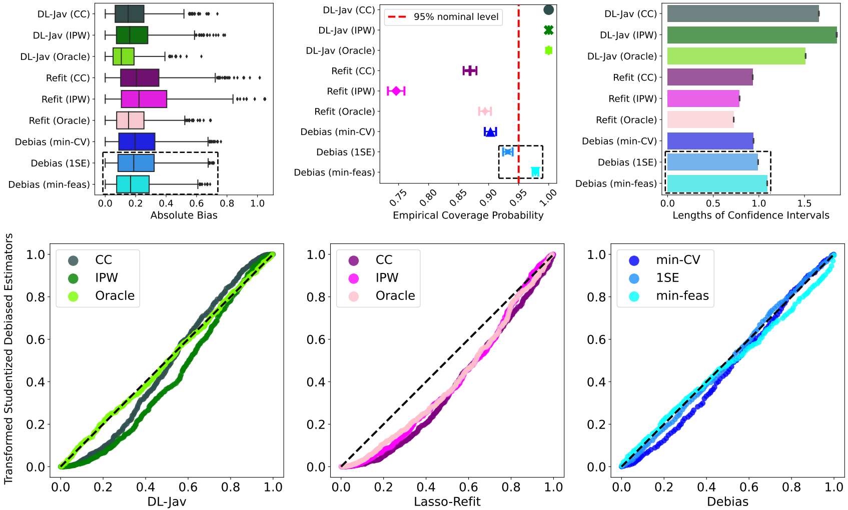


Figure 4: Dense β_0^{de} and sparse $x^{(4)}$ with $X_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma^{cs}), i = 1, ..., n$.

Simulation Results Under t_2-Distributed Noises

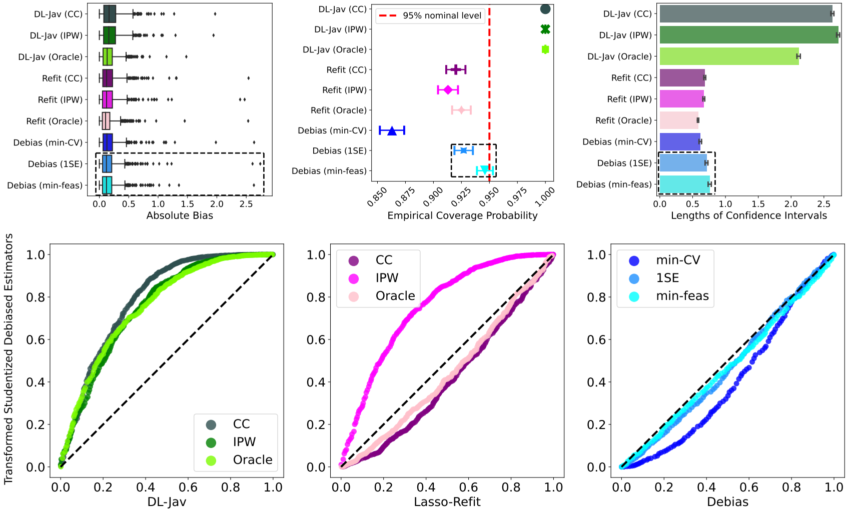


Figure 5: Pseudo-dense β_0^{pd} and dense $x^{(4)}$ with $X_i \sim \mathcal{N}_d(\mathbf{0}, \Sigma^{\mathrm{ar}}), i = 1, ..., n$. Note that the mean-zero t_2 distribution has *infinite* variance.

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.

Proposed Method With Nonparametric Propensity Scores

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.

Proposed Method With Nonparametric Propensity Scores

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

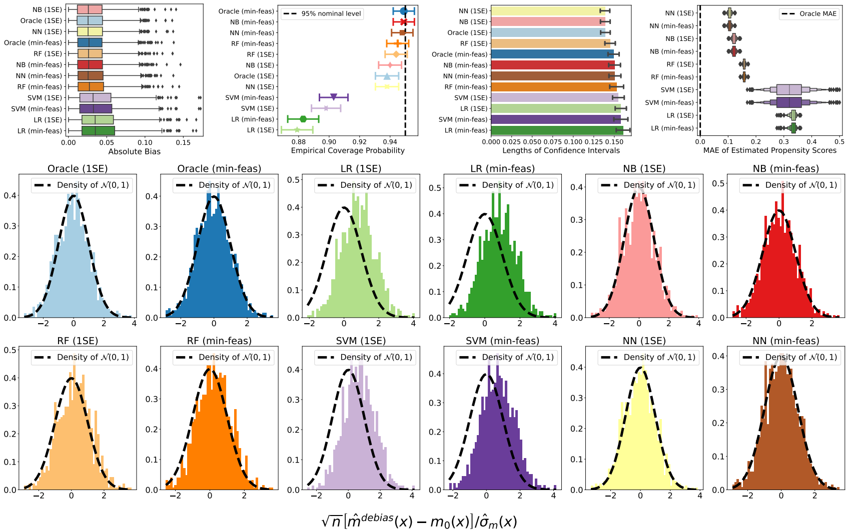
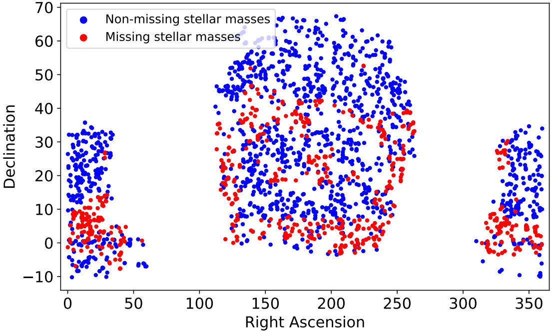


Figure 6: Sparse β_0^{sp} and (weakly) dense $x^{(4)}$.

Real-World Applications

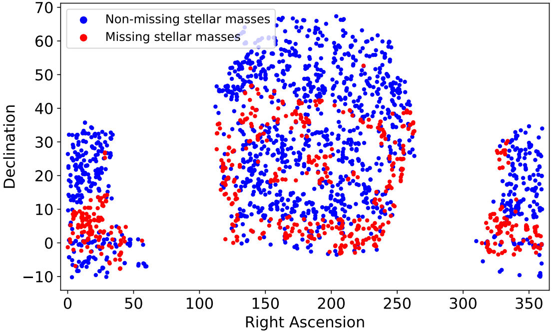
Background on Stellar Mass Inference

Recall that some estimated stellar masses of the observed galaxies in SDSS-IV are missing in the most recent Firefly value-added catalog.



Background on Stellar Mass Inference

Recall that some estimated stellar masses of the observed galaxies in SDSS-IV are missing in the most recent Firefly value-added catalog.



► **Scientific Questions:**

1. *How can we conduct valid inference on the (estimated) stellar mass based on the spectroscopic and photometric properties?*
2. *Is it statistically significant that the stellar mass of a galaxy is negatively correlated with its distance to the nearby cosmic filament structures?*

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.

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).

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.

Study Design For Stellar Mass Inference

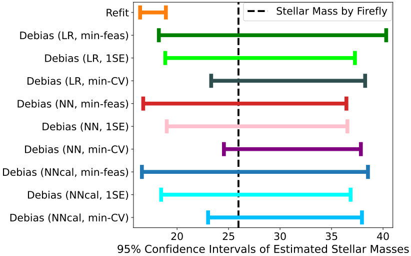
1. Consider all the observed galaxies by SDSS-IV within a thin redshift slice $0.4 \sim 0.4005$, among which 30.2% of their stellar masses are missing in the Firefly value-added catalog.
2. Fetch their spectroscopic and photometric properties from SDSS-IV DR16 database similar to the input catalog of 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).

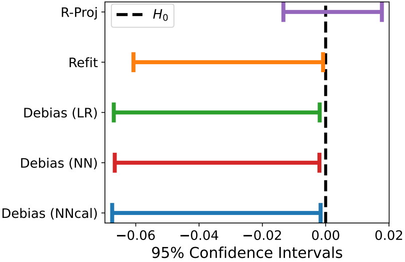
Study Design For Stellar Mass Inference

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 $d = 1409$.

Results on Stellar Mass Inference





- *Left Panel:* 95% confidence intervals by different debiasing methods for the estimated stellar mass of a new galaxy.
- *Right Panel:* 95% confidence intervals by different debiasing methods for the estimated regression coefficient associated with the distance to nearby cosmic filaments.

Conclusions and Future Works

Conclusions

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

Conclusions

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

- Its computational and statistical efficiencies follow from the dual formulation.
- Sample splitting and cross fitting are not required, and the nuisance propensity score can be estimated by any machine learning method.
- We provide interpretations to our debiasing method from the viewpoints of bias-variance trade-off and Neyman near-orthogonalization.
- Comprehensive simulation studies and motivating applications demonstrate the potential of our proposed debiasing method.

Potential Application to Causal Inference (I)

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 the regression function (or conditional mean outcome) of the treatment group.

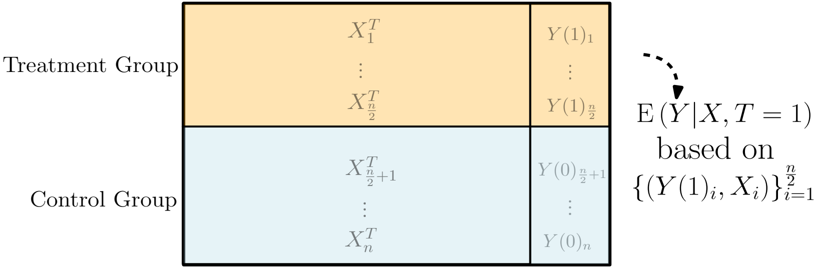


Figure 8: Traditional approaches for inferring $\mathrm{E}(Y|X, T=1)$.

Potential Application to Causal Inference (I)

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 the regression function (or conditional mean outcome) of the treatment group.

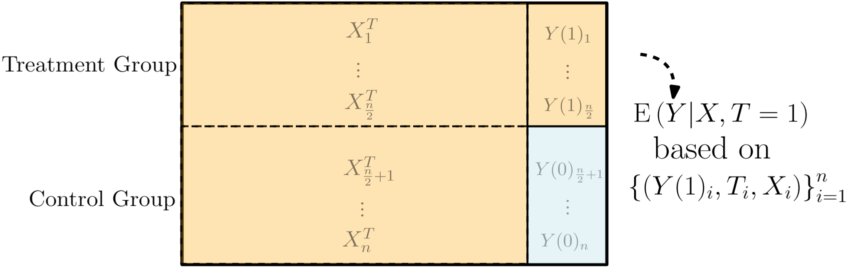


Figure 8: Our approach for inferring $\mathrm{E}(Y|X, T=1)$.

Potential Application to Causal Inference (II)

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

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

with no unmeasured confounding and high-dimensional covariates.

Potential Application to Causal Inference (II)

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

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

with no unmeasured confounding and high-dimensional covariates.

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

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

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

- The extended debiased estimator becomes

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

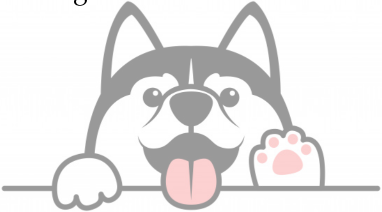
- The efficiency theory for this modified procedure is worth studying!

Thank you!

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.



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.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

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.

Y. Cao, Z. Chen, M. Belkin, and Q. Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.

C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

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.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 2006.

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.

J. Fan, J. Lv, and L. Qi. Sparse high-dimensional models in economics. *Annu. Rev. Econ.*, 3(1):291–317, 2011.

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.

A. Giessing and J. Wang. Debiased inference on heterogeneous quantile treatment effects with regression rank scores. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad075, 08 2023.

J. P. Higgins, I. R. White, and A. M. Wood. Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical trials*, 5(3):225–239, 2008.

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

J. Jankova and S. van de Geer. Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359, 2018.

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.

Z. Li, Z.-H. Zhou, and A. Gretton. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436 – 1462, 2006.

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.

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.

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

R. Wyss, C. Yanover, T. El-Hay, D. Bennett, R. W. Platt, A. R. Zullo, G. Sari, X. Wen, Y. Ye, H. Yuan, et al. Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database studies: An overview of the current literature. *Pharmacoepidemiology and Drug Safety*, 31(9): 932–943, 2022.

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.

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:

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

where $\mathcal{S}_{\frac{\gamma}{n}}(u) = \text{sign}(u)\cdot\left(u - \frac{\gamma}{n}\right)_+$ 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:

$$SD(\gamma_i) = \sqrt{\text{Var}\left(CV_1(\gamma_i), ..., CV_K(\gamma_i)\right)}, \quad SE(\gamma_i) = SD(\gamma_i)/\sqrt{K}.$$

- 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

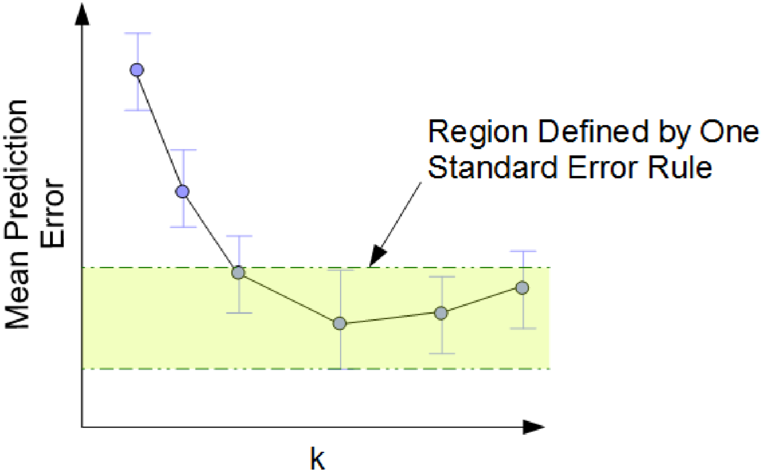


Figure 9: Illustration of the 1SE rule for selecting the model parameter.

Finger-of-God and Kaiser Effects

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.

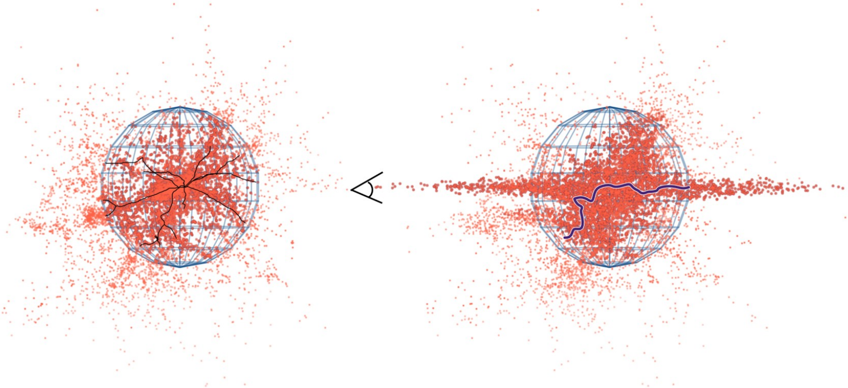


Figure 10: Redshift distortions along the line of sight (Kuchner et al., 2021).