Journal of Machine Learning Research 24 (2023) 1-52 Submitted 11/22; Revised 10/23; Published 11/23

Modular Regression: Improving Linear Models by Incorporating Auxiliary Data

Ying Jin YING531@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Dominik Rothenhäusler RDOMINIK@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Editor: Ilya Shpitser

Abstract

This paper develops a new framework, called modular regression, to utilize auxiliary information – such as variables other than the original features or additional data sets – in the training process of linear models. At a high level, our method follows the routine: (i) decomposing the regression task into several sub-tasks, (ii) fitting the sub-task models, and (iii) using the sub-task models to provide an improved estimate for the original regression problem. This routine applies to widely-used low-dimensional (generalized) linear models and high-dimensional regularized linear regression. It also naturally extends to missing-data settings where only partial observations are available. By incorporating auxiliary information, our approach improves the estimation efficiency and prediction accuracy upon linear regression or the Lasso under a conditional independence assumption for predicting the outcome. For high-dimensional settings, we develop an extension of our procedure that is robust to violations of the conditional independence assumption, in the sense that it improves efficiency if this assumption holds and coincides with the Lasso otherwise. We demonstrate the efficacy of our methods with simulated and real data sets.

Keywords: Data fusion, high dimensional statistics, missing data, regression, semiparametric efficiency, surrogates

1. Introduction

Suppose for a patient subject to a surgical procedure, we are interested in predicting their future health outcome $Y \in \mathbb{R}$ in two years using some features $X \in \mathbb{R}^{p_x}$ collected at the time of the surgery. The standard approach is supervised learning: The training data containing n observations $\{(X_i, Y_i)\}_{i=1}^n$ from previous patients is used to pick a predictor $\widehat{f}\colon \mathbb{R}^{p_x} \to \mathbb{R}$ that minimizes the average prediction error $\frac{1}{n}\sum_{i=1}^n \ell(f(X_i), Y_i)$ over $f \in \mathcal{F}$ for some loss function $\ell\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and some function class $\mathcal{F}$.

However, the long duration of the study can pose various practical challenges to this paradigm. The number of joint observations of (X, Y) may be very limited because they take at least two years to collect. The training data may also contain intermediate measurements Z, such as the health outcome one year after their surgery. In addition, there may be some

patients for whom we only observe (X, Z) which are easier to collect. It is also probable that another data set provides several (Z, Y) samples from patients who had surgery more than two years ago, but their features X were missing due to limited technology. In this case, the standard approach of empirical risk minimization only over (X, Y) samples falls short as it cannot make use of auxiliary variables or additional partial observations.

In the missing data literature, intermediate outcomes Z are often utilized to estimate treatment effects under a conditional independence assumption (i.e., $X \perp\!\!\!\perp Y \,|\, Z$ in our setting) when the outcome Y is difficult to measure (Prentice, 1989; Begg and Leung, 2000; Athey et al., 2016, 2019). Incorporating intermediate outcomes improves the estimation efficiency when joint observations of (X, Y, Z) are available. It also offers more flexibility in data usage because it allows to incorporate (X, Z) and (Z, Y) samples. However, the conditional independence assumption may not hold in practice, and the existing methods are highly specialized to estimating causal effects.

In this work, we aim to build a framework for flexibly incorporating auxiliary information into generic estimation and prediction procedures while maintaining rigorous guarantees. Such information may come from other variables Z in addition to the original features and responses in the training process; it may also be additional data sets that only cover a subset of variables. We will extend the ideas of leveraging independence from the missing data literature to generalized and high-dimensional linear models, and develop more robust methods for general dependence patterns. Let us begin with the tasks we consider.

A common algorithmic structure. We focus on statistical learning algorithms that take the form

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^{p_x}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \big\{ h(X_i, \theta) - Y_i X_i^\top \theta \big\}, \tag{1}$$

where $h \colon \mathbb{R}^{p_x} \times \mathbb{R}^{p_x} \to \mathbb{R}$ is a known function that is convex in θ, and $p_x \in \mathbb{N}$ is the dimension of the prediction features. The simplest example is ordinary least squares (OLS):

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^{p_x}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \Big\{ \frac{1}{2} \theta^\top X_i X_i^\top \theta - Y_i X_i^\top \theta \Big\}.$$

Logistic regression with $Y \in \{0, 1\}$ also satisfies (1):

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^{p_x}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \big\{ \log(1 + \exp(X_i^\top \theta)) - Y_i X_i^\top \theta \big\}.$$

We will later extend to other generalized linear models as we develop our method. The final example we consider is the ℓ_1-regularized linear regression (Tibshirani, 1996):

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^{p_x}}{\operatorname{argmin}} \big\{ -\tfrac{1}{n} \textstyle\sum_{i=1}^{n} Y_i X_i^\top \theta + \theta^\top \big(\frac{1}{2n} \sum_{i=1}^{n} X_i X_i^\top \big) \theta + \lambda \|\theta\|_1 \big\}. \tag{2}$$

In the situation we discuss at the beginning, the supervised learning estimators of the form (1) may not be able to utilize auxiliary data. Instead of minimizing equation (1), we propose an alternative risk minimization criterion:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^{p_x}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} h(X_i, \theta) - \widehat{C}^\top \theta \right\},$$

where $\widehat{C}$ is a proxy term for $C := \mathbb{E}[XY]$ computed from the data. We will see that the performance of $\widehat{\theta}$ is closely related to the estimation accuracy of $\widehat{C}$. At a high level, our idea is to replace $\frac{1}{n}\sum_{i=1}^n X_i Y_i$ by $\widehat{C}$ with negligible bias and a lower variance, which translates to the improved performance of $\widehat{\theta}$. We achieve this by developing a "modular" estimator $\widehat{C}$ – whose meaning will be made precise soon – that naturally allows for flexible incorporation of auxiliary information to improve the learning performance.

We emphasize that we still aim to minimize the same population risk (e.g., the population version of (1)) as when using $\frac{1}{n}\sum_{i=1}^n X_i Y_i$. As a result, the estimator still converges to the same limit θ^* that minimizes the population risk, that is, $\theta^* = \arg\min \mathbb{E}[h(X, \theta) - YX^\top \theta]$. Relatedly, our methods can be used for both estimation and prediction.

1.1 Leveraging the dependence structure

The driving force of our approach is to leverage the conditional independence structure among variables. Conditional independencies are often used to improve performance over saturated models. As a classical example, suppose we have i.i.d. copies of a vector (X, Y) obeying $X_{-S} \perp\!\!\!\perp Y \,|\, X_S$, that is, X_{-S} is conditionally independent of Y given X_S, where X_S denotes a subset S of features, and X_{-S} stores the remaining ones. It is known that running a regression of Y on X_S can have lower mean-squared error than regressing Y on X, since the latter may have very large variance (see, e.g., Hastie et al. (2009) for more intuition). However, the former estimate may be biased if the conditional independence relation is violated. Furthermore, even if the conditional independence holds, the set S is generally unknown. It is common to use model selection methods such as the best subset selection, Lasso, AIC, or BIC to navigate the bias-variance tradeoff.

Perhaps less widely known is that structures of the type $Y \perp\!\!\!\perp X_S \,|\, Z$ for auxiliary variables Z can also be leveraged to improve estimation and prediction. Analogous to classical model selection strategies, we want to derive a data-driven strategy to learn and exploit such structures. To fix ideas, let us consider an extreme case where $X \perp\!\!\!\perp Y \,|\, Z$.

A naive approach. If $X \perp\!\!\!\perp Y \,|\, Z$, we can re-write

$$\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[X \,|\, Z]\mathbb{E}[Y \,|\, Z]]. \tag{3}$$

Assuming access to i.i.d. copies $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, in view of (3), we can (i) estimate $\mathbb{E}[Y|Z = z]$ via $\widehat{\mu}_y(z)$, (ii) estimate $\mathbb{E}[X|Z = z]$ via $\widehat{\mu}_x(z)$ and (iii) combine these two estimates to compute $\widehat{C} = \frac{1}{n}\sum_{i=1}^n \widehat{\mu}_y(Z_i)\widehat{\mu}_x(Z_i)$ for $C = \mathbb{E}[XY]$. If $\widehat{\mu}_x$ and $\widehat{\mu}_y$ are accurate, then this estimate has smaller variance than $\frac{1}{n}\sum_{i=1}^n X_i Y_i$, as

$$\operatorname{Var}(\widehat{\mu}_y(Z)\widehat{\mu}_x(Z)) \approx \operatorname{Var}(\mathbb{E}[Y \,|\, Z]\mathbb{E}[X \,|\, Z]) = \operatorname{Var}(\mathbb{E}[XY \,|\, Z]) \leq \operatorname{Var}(XY).$$

However, this naive approach has a pressing issue that may hinder its performance: even if (3) holds, there can be considerable bias if the estimation of $\widehat{\mu}_y$ and $\widehat{\mu}_x$ has slow convergence rates. In this case, $\widehat{C}$ is not a good estimator for C because the bias $\mathbb{E}[\widehat{C}] - C$ can be comparable with, or even larger in large samples than, the variance of $\frac{1}{n}\sum_{i=1}^n X_i Y_i$.

Figure 1 illustrates this point via a simple numerical example, where $Z \in \mathbb{R}^{20}$, $X, Y \in \mathbb{R}$, and the conditional expectation functions $\mathbb{E}[Y \mid Z]$ and $\mathbb{E}[X \mid Z]$ only involve the first three entries in Z. The goal is to compute the best linear predictor for Y, and the default choice is OLS. We apply the above naive approach with $\widehat{\mu}_y$ and $\widehat{\mu}_x$ fitted by random forests in R, and show the bias and standard deviation scaled by $\sqrt{n}$ for various sample sizes n.

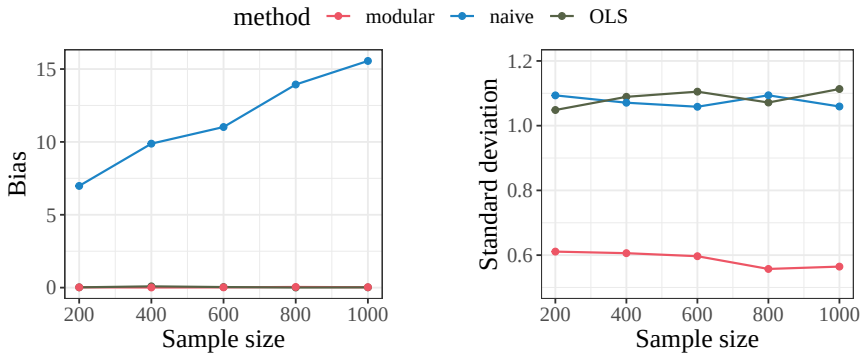


Figure 1: Scaled by $\sqrt{n}$, the bias (on the left) and the standard deviation (on the right) are shown for plain OLS, the `naive` approach, and our to-be-introduced `modular` approach, across various sample sizes n.

Even in this simple example where random forests should excel, the naive approach exhibits a significantly larger bias compared to plain OLS. This bias is larger than $n^{-1/2}$, and is due to the slow convergence of random forests. In addition, contrary to our prediction, the standard deviation of the naive approach is close to, instead of smaller than, the OLS, which is also due to the variability in training the random forests.

A better approach. The performance of our new approach is depicted in red in Figure 1: It has negligible bias and reduces the standard deviation by more than 40% compared with OLS at all sample sizes. Our new approach uses the same intuitions as mentioned earlier but additionally addresses bias using well-known semi-parametric techniques based on the mixed-bias property (Rotnitzky et al., 2021; Robins et al., 2008). We further rewrite (3) as

$$\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y \mid Z]] + \mathbb{E}[\mathbb{E}[X|Z]Y] - \mathbb{E}[\mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z]].$$

Again, an estimator can be obtained via three sub-tasks: (i) estimating $\mathbb{E}[Y|Z = z]$ via $\widehat{\mu}_y(z)$, (ii) estimating $\mathbb{E}[X|Z = z]$ via $\widehat{\mu}_x(z)$, and (iii) combining these two estimates via

$$\widehat{C} = \frac{1}{n}\sum_{i=1}^{n} \left\{\widehat{\mu}_y(Z_i)X_i + Y_i\widehat{\mu}_x(Z_i) - \widehat{\mu}_y(Z_i)\widehat{\mu}_x(Z_i)\right\}. \tag{4}$$

We will show that a slight variation of this approach has very favorable properties. Most importantly, even if $\widehat{\mu}_y$ and $\widehat{\mu}_x$ converge to the ground truth at slow speed, the bias of $\widehat{C}$ in (4) can still be negligible, and it has a lower variance than $\frac{1}{n}\sum_{i=1}^{n} X_iY_i$:

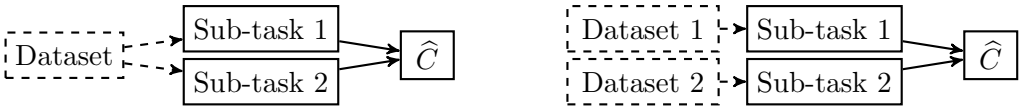
$$n \cdot \mathrm{Var}(\widehat{C}) \approx \mathrm{Var}(XY) - \mathrm{Var}\left((X - \mu_x(Z)) \cdot (Y - \mu_y(Z))\right) \leq \mathrm{Var}(XY).$$

Indeed, it leads to the most efficient estimator if $X \perp\!\!\!\perp Y \mid Z$ (see Section 2.2 for details). This approach also opens a door for data fusion: (4) only involve pairwise observations of (X, Z) or (Z, Y). Thus, it can flexibly use additional data to improve estimation accuracy, and works even when no joint observations of (X, Z, Y) are available.

1.2 A modular regression framework

We propose a modular regression framework that concretize the above ideas by carefully dealing with the nuances in obtaining estimators $\widehat{\mu}_x$, $\widehat{\mu}_y$ to eliminate bias, navigating bias-variance tradeoffs under potential violation of the conditional independence assumption, and developing principled algorithms for general estimators and data fusion scenarios.

As the name suggests, we decompose the estimation of $\mathbb{E}[XY]$ into smaller modules – each being a sub-task that involves a subset of variables such as (Y, X), (Y, Z) or (Z, X) – and then re-assemble the modules to construct $\widehat{C}$. This idea is visualized in Figure 2.



(a) Conditional independencies can be leveraged to decompose statistical tasks into subtasks.

(b) Modularity allows combining multiple datasets.

Figure 2: Decomposing the estimation of $C = \mathbb{E}[YX]$ tackles two settings that may seem different at first sight. In (a), one can leverage conditional independencies to reduce statistical uncertainty. In (b), one can use additional data to increase effective sample size.

As shown in panel (a), our method adapts to specific structures, such as conditional independence between variables, so as to ensure a properly chosen decomposition. Such structure can also be learned from data. The decomposition also allows data fusion; see panel (b). As long as a data set covers the variables in a sub-task, it can be incorporated to improve efficiency. We summarize our main results below.

- *Modular linear regression.* We develop a generic approach for improving the estimation and prediction in (generalized) linear models. This is achieved by rewriting the estimation equation to utilize auxiliary variables. Leveraging semi-parametric theory, we show the optimality of our approach under a conditional independence condition.

- *High-dimensional modular regression.* By adding ℓ_1-regularization, our tools extend to the high-dimensional setting. We develop a proxy empirical risk that has negligible bias and lower variance compared with the Lasso. We show that it improves the upper bounds for prediction accuracy.

- *Robustness under conditional dependence.* Since the conditional independence condition may be violated in practice, we develop an extension of our method that adapts to the dependency structures among variables, and we allow such structures to be learned from data. The proposed procedure interpolates between linear models (non-modular) and the fully-modular approach (which is efficient under conditional independence). The proposed procedure, bridging linear (non-modular) and fully-modular models, enhances robustness while still allowing for efficiency gains.

- *Data fusion.* We further extends method to combine multiple data sets. This formulation allows using additional data on sub-tasks to improve overall statistical efficiency.

The rest of the paper is organized as follows. Section 2 develops modular (generalized) linear regression in the fixed-p setting. Section 3 studies modular regression for penalized linear regression in high dimensions, and develops a robust variant that can adapt to different dependence structures. Applications to data fusion and partial observations are discussed in Section 4. Finally, we evaluate our methods on simulated datasets in Section 5 and apply them to real datasets in Section 6.

1.3 Related work

Modular regression combines evidence from sub-tasks, leveraging a "modular structure" provided by the conditional independence. One related strand is to use surrogates when the outcome of interest is expensive to measure (Fleming et al., 1994; Post et al., 2010). In particular, in causal inference, a series of work (Prentice, 1989; Lauritzen et al., 2004; Chen et al., 2007; VanderWeele, 2013; Athey et al., 2016, 2019; Kallus and Mao, 2020) have advocated using intermediate (short-term) outcome(s) as "surrogates" for long-term outcomes, and assumes various surrogate conditions such as conditional independence of the long-term outcome and the treatment given the surrogates. Similar to our approach, the surrogate condition allows to decompose treatment effect estimation into sub-tasks of estimating the effect of the treatment on the intermediate outcome and estimating the effect of the intermediate outcome on the long-term outcome. We work under a generic conditional independence condition and mainly focus on regression and prediction tasks. In addition, we propose a method that is robust to violations of conditional independence.

Our framework is related to missing data and in particular data fusion, which combines data sets that cover different subsets of variables. The preceding paragraph is also related; more generally, there has been a surge of interest in combining different data sets, such as combining experimental and observational data for treatment effect estimation (Rosenman et al., 2022, 2020; Colnet et al., 2020), generalizing inference to different populations (Dahabreh et al., 2019; Hartman et al., 2015; Jin and Rothenhäusler, 2021), etc. Many of these works require identifiability assumptions for the target estimand, which are similar to the conditional independence condition. Compared to these works, we study general tasks and enjoy robustness to failure of conditional independence.

In missing data scenarios, some other works study regression problem in settings that differ from ours. Those for high-dimensional data include Loh and Wainwright (2011) on penalized regression, Lounici (2014); Cai and Zhang (2016) on the estimation of covariance matrices, and Elsener and van de Geer (2019); Zhu et al. (2019) on sparse principal component analysis, which consider estimation when some covariates are missing at random. In contrast, we focus more on the data fusion aspect; we draw ideas from semiparametric statistics, and account for the dependence structure. A recent work of Cannings and Fan (2022) on U-statistics with low-dimensional data is related, which devises an estimator with smaller prediction MSE by incorporating a partially missing data set. Instead, our approach leverages conditional independence structures among variables instead of correlation between estimators; we focus on regression and consider the high dimensional regime.

Our estimator is inspired by recent progress in efficient regression adjustment (Henckel et al., 2019; Rotnitzky and Smucler, 2019) in low-dimensional graphical models. Our framework is more general as it applies to (generalized) linear models, works for high dimensional settings, does not require the graphical structure, and imposes minimal model assumptions; it is also more restricted in that we do not consider choosing the optimal covariate sets.

The modular estimator is optimal for linear regression under conditional independence, in the sense of semiparametric efficiency (Bickel et al., 1993; Tsiatis, 2006). Thus, this work is also connected to the literature of missing data. Furthermore, the double robustness of our estimator is similar to the AIPW estimator (Robins et al., 1994) and leverages the "mixed-bias" property (Rotnitzky et al., 2021; Robins et al., 2008).

The conditional independence structure has been used in a series of early works (Sargan, 1958; Hansen, 1982) to improve estimation efficiency; also related is Causeur and Dhorne (2003) on linear regression under conditional independence, which is closely related to our first modular regression algorithm in low dimensions. However, these works assume strong parametric models such as a joint Gaussian distribution and independent Gaussian noise, while our method does not require stringent model assumptions.

Broadly speaking, our variance reduction idea may also apply to other statistical problems where the 'modular' structure and conditional independence are present. For instance, sufficient dimensional reduction (SDR) (Li, 2018) imposes $Y \perp\!\!\!\perp X \mid X^\top \Gamma$ for some unknown matrix Γ, whose estimation often relies on inverse regression algorithms (Li, 1991; Adragni and Cook, 2009; Ma and Zhu, 2013) that demonstrate a modular structure. Our idea may be adapted to enable more efficient inverse regression given an approximate solution for Γ.

Notations. We use the standard $O_P(\cdot)$ and $o_P(\cdot)$ notation to denote smaller order in probability. For any two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we denote $a_n = \Omega(b_n)$ if $\lim_{n\to\infty} a_n/b_n < \infty$ and $\lim_{n\to\infty} b_n/a_n < \infty$; we denote $a_n = O(b_n)$ if $a_n/b_n \le C$ for some constant $C > 0$. We use $\mathbb{P}_{X,Y,Z}$ to denote the joint distribution of (X, Y, Z). For any functions $f, g\colon \mathcal{Z} \to \mathbb{R}^d$, we denote their L_2-distance as $\|f - g\|^2_{L_2(\mathbb{P}_Z)} = \mathbb{E}[\|f(Z) - g(Z))\|_2^2]$, where $\|\cdot\|_2$ is the Euclidean norm and the expectation is with respect to the distribution $\mathbb{P}_Z$ of Z. For any vector $v \in \mathbb{R}^l$ and positive definite matrix $\Sigma \in \mathbb{R}^{l\times l}$ we set $\|v\|^2_\Sigma = v^\top \Sigma v$.

2. Modular linear regression in low dimensions

We first introduce modular regression in low-dimensional linear models, where $X \in \mathbb{R}^{p_x}$, $Z \in \mathbb{R}^{p_z}$, and p_x, p_z do not grow with n asymptotically.

Assumption 1 *The joint distribution of (X, Z, Y) satisfies $X \perp\!\!\!\perp Y \mid Z$.*

In this section, we use this working assumption to show how conditional independencies can be leveraged to improve estimation accuracy. It will be relaxed later in Section 3.3, where we develop a method that is robust to violations of Assumption 1.

Intuitively, Assumption 1 ensures that Z contains all the information in Y that is relevant to X. While this assumption seems strong, it is reasonable in many situations and has been widely used in a variety of applications. For example, Z could be an intermediate outcome in the training data that is unavailable for the test samples, while Y is a long-term outcome of interest. Intermediate outcomes are widely used as a proxy in estimating long-term treatment effects (Athey et al., 2016, 2019), while we focus on regression problems.

2.1 Modular linear regression

For simplicity of exposition, we start with linear (OLS) regression to show the benefits of a modular structure. Our goal is to predict $Y \in \mathbb{R}$ by $\widehat{Y} = X^\top \theta$ for some $\theta \in \mathbb{R}^{p_x}$, but we do not necessarily assume a well-posed linear model. In the training process, we have observations of i.i.d. triples $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ for some auxiliary variables $Z_i \in \mathbb{R}^{p_z}$, and $X_i \in \mathbb{R}^{p_x}$ are random vectors (we view all vectors as column vectors throughout). The training and the target distributions are induced by the same distribution $(X, Y, Z) \sim \mathbb{P}$.

Modular regression proceeds in three steps to leverage conditional independence.

1. *Decompose into sub-tasks.* First, the conditional independence structure in Assumption 1 allows us to decompose the estimation of $\mathbb{E}[XY]$ into sub-tasks that only involve observations of (X, Z) or (Y, Z).

2. *Solve the sub-tasks.* We then learn the conditional mean functions. We use cross-fitting (Chernozhukov et al., 2018) to ensure desirable statistical properties: we randomly split $\mathcal{I} = \{1, \ldots, n\}$ into two disjoint folds $\mathcal{I}_1$, $\mathcal{I}_2$. Then for each fold $k = 1, 2$, we fit models $\widehat{\mu}_x^{(i)} \colon \mathcal{Z} \to \mathbb{R}^{p_x}$ and $\widehat{\mu}_y^{(k)} \colon \mathcal{Z} \to \mathbb{R}^{p_y}$ for $\mu_x(z) = \mathbb{E}[X \,|\, Z = z]$ and $\mu_y(z) = \mathbb{E}[Y \,|\, Z = z]$ using data in $\mathcal{I} \backslash \mathcal{I}_k$. With slight abuse of notation, we write $\widehat{\mu}_y(Z_i) = \widehat{\mu}_y^{(k)}(Z_i)$ and $\widehat{\mu}_x(Z_i) = \widehat{\mu}_x^{(k)}(Z_i)$, $i \in \mathcal{I}_k$.

3. *Assemble the sub-tasks.* Finally, we solve a modular least squares regression

$$\widehat{\theta}_n^{\mathrm{mod}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^{p_x}} \left\{ \frac{1}{n} \sum_{i=1}^n \theta^\top X_i X_i^\top \theta / 2 - \widehat{C}_{\mathrm{lm}}^\top \theta \right\}, \tag{5}$$

   where we use a proxy cross-term

$$\widehat{C}_{\mathrm{lm}} = \frac{1}{n} \sum_{i=1}^n \left[X_i \widehat{\mu}_y(Z_i) + \widehat{\mu}_x(Z_i) Y_i - \widehat{\mu}_x(Z_i) \widehat{\mu}_y(Z_i) \right]. \tag{6}$$

The procedure is summarized in Algorithm 1. When X contains an intercept, the corresponding entry of $\widehat{C}_{\mathrm{lm}}$ can be simply set as $\frac{1}{n} \sum_{i=1}^n Y_i$.

Algorithm 1 Modular linear regression

Input: Dataset $\mathcal{I} = \{(Y_i, X_i, Z_i\}_{i=1}^n$.

1: Randomly split $\mathcal{I}$ into equally-sized folds $\mathcal{I}_k$, $k = 1, 2$.
2: **for** $k = 1, 2$ **do**
3: Fit models $\widehat{\mu}_x^{(k)}(\cdot)$ for $\mathbb{E}[X \,|\, Z = \cdot]$ using observations (X_i, Z_i), $i \notin \mathcal{I}_k$;
4: Fit models $\widehat{\mu}_y^{(k)}(\cdot)$ for $\mathbb{E}[Y \,|\, Z = \cdot]$ using observations (Y_i, Z_i), $i \notin \mathcal{I}_k$;
5: Compute $\widehat{\mu}_x(Z_i) = \widehat{\mu}_x^{(k)}(Z_i)$ and $\widehat{\mu}_y(Z_i) = \widehat{\mu}_y^{(k)}(Z_i)$ for all $i \in \mathcal{I}_k$.
6: **end for**
7: Compute the proxy cross-term $\widehat{C}_{\mathrm{lm}}$ as in (6).

Output: The modular linear regression $\widehat{\theta}_n^{\mathrm{mod}}$ as in (5).

2.2 Double robustness and semiparametric efficiency

The modular estimator $\widehat{\theta}_n^{\text{mod}}$ is doubly robust with respect to the estimation error of $\widehat{\mu}_x$ and $\widehat{\mu}_y$. To be more precise, the estimator converges at $n^{-1/2}$ rate to θ^* and is semiparametrically efficient even when $\widehat{\mu}_x$ and $\widehat{\mu}_y$ are consistent with slow nonparametric rates, that is if $\|\widehat{\mu}_x^{(k)} - \mu_x\|_{L_2(\mathbb{P}_Z)} = o_P(n^{-1/4})$ and $\|\widehat{\mu}_y^{(k)} - \mu_y\|_{L_2(\mathbb{P}_Z)} = o_P(n^{-1/4})$. The proof of the next theorem is in Appendix C.1.

Theorem 2 (Double robustness and efficiency) *Suppose* $\|\widehat{\mu}_x^{(k)} - \mu_x\|_{L_2(\mathbb{P}_Z)} \cdot \|\widehat{\mu}_y^{(k)} - \mu_y\|_{L_2(\mathbb{P}_Z)} = o_P(1/\sqrt{n})$ *for* $k = 1, 2$, $\mathbb{E}[XX^\top] \succ 0$ *is finite, and* $X(Y - X^\top\theta^*)$ *has finite second moment. Then* $\sqrt{n}(\widehat{\theta}_n^{mod} - \theta^*) \xrightarrow{d} N(0, \Sigma^{mod})$ *as* $n \to \infty$, *where* $\Sigma^{mod} = \text{Cov}(\phi^{mod}(X_i, Y_i, Z_i))$ *for* $\phi^{mod}(x, y, z) = \mathbb{E}[XX^\top]^{-1}\big(x\mu_y(z) + \mu_x(z)y - \mu_x(z)\mu_y(z) - xx^\top\theta^*\big)$. *Further, suppose the distribution of* (X, Y, Z) *has density with respect to a base measure* μ. *Then* ϕ^{mod} *is the (semiparametrically) efficient influence function for estimating* θ^*.

A short discussion of the implications of Theorem 2 is in order. When $\widehat{\mu}_x$ and $\widehat{\mu}_y$ are estimated by nonparametric regression which typically comes with slow rates, Assumption 1 (conditional independence) is needed to ensure fast convergence of $\widehat{\theta}_n^{\text{mod}}$. However, when they are estimated by parametric regression (such as OLS) and converge at $\sqrt{n}$-rate to any μ_x^* and μ_y^* (not necessarily equal μ_x or μ_y), our estimator is consistent and asymptotically normal with $\sqrt{n}$-rate as long as $\mathbb{E}[(Y - \mu_y^*(Z))(X - \mu_x^*(Z))] = 0$, i.e., the residuals of Y and X are uncorrelated. This effect will also be visible in our simulations in Section 5.

Theorem 2 shows that modular regression has the lowest asymptotic variance among all regular estimators (Bickel et al., 1993). We now compare it with the OLS estimator $\widehat{\theta}_n^{\text{ols}} = \text{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^n (Y_i - X_i^\top\theta)^2$, which only uses the information in (X_i, Y_i) pairs. We know $\widehat{\theta}_n^{\text{ols}} - \theta^* = \frac{1}{n}\sum_{i=1}^n \phi^{\text{ols}}(X_i, Y_i) + o_P(1/\sqrt{n})$ with $\phi^{\text{ols}}(x, y) = \mathbb{E}[XX^\top]^{-1}(xy - xx^\top\theta^*)$ as $n \to \infty$. Under Assumption 1, the asymptotic variance of $\phi^{\text{mod}}(X_i, Y_i, Z_i)$ is dominated by that of $\phi^{\text{ols}}(X_i, Y_i)$. Indeed, one can check that

$$\text{Cov}\left(\phi^{\text{mod}}(X_i, Y_i, Z_i)\right) = \text{Cov}\left(\phi^{\text{ols}}(X_i, Y_i)\right) - \Sigma^{-1}\,\text{Cov}\left((Y - \mu_y(Z))(X - \mu_x(Z)\right)\Sigma^{-1},$$

where $\text{Cov}(A)$ denotes the covariance matrix for a random vector A, so that $\text{Cov}\left((Y - \mu_y(Z))(X - \mu_x(Z)\right) \succ 0$. More efficient parameter estimation also translates to more accurate prediction: The prediction mean squared error (MSE) for an independent test sample (X, Y) is $\mathbb{E}[(Y - X^\top\widehat{\theta})^2 \,|\, \widehat{\theta}] = \mathbb{E}[(Y - X^\top\theta^*)^2] + \|\widehat{\theta} - \theta^*\|_\Sigma^2$, where the second term is smaller when $\widehat{\theta}$ is the modular estimator instead of the OLS estimator. In our numerical experiments, we mostly focus on improving prediction in very noisy settings with low sample size because the improvement in $\|\widehat{\theta} - \theta^*\|_\Sigma^2$ is more pronounced in those cases.

Remark 3 (Alternative estimators) *There are several natural alternatives to our modular estimator. The first is outcome regression, i.e., running OLS of* $\widehat{\mu}_y(Z_i)$ *over* X_i, *where* $\widehat{\mu}_y(\cdot)$ *is an estimator for* $\mu_y(\cdot) = \mathbb{E}[Y \,|\, Z = \cdot]$. *The second is orthogonal regression, i.e., extracting the* X*-coefficients in OLS of* Y_i *over* $(X_i, Z_i^{\text{resid}})$, *where* $Z^{\text{resid}} := Z_i - \mu_z(X_i)$, *and* $\widehat{\mu}_z(\cdot)$ *is an estimator for* $\mu_z(\cdot) = \mathbb{E}[Z \,|\, X = \cdot]$. *However, both options can be less efficient than our proposals. When* $\widehat{\mu}_y$ *and* $\widehat{\mu}_z$ *are estimated by flexible nonparametric algorithms, both methods suffer from substantial bias that is comparable to or larger than the*

OLS variance (this is similar to the issue with our naive approach in Section 1); in contrast, our estimator achieves fast (product) convergence rate, and is the most efficient among all regular and asymptotically linear estimators. An extended discussion is in Appendix B.2.

Practitioners may be interested in quantifying the uncertainty of modular regression estimates. Wald-type confidence intervals can be derived using the asymptotic distribution in Theorem 2. The next corollary formalizes this result for the one-dimensional case; the multi-dimensional case follows similar ideas.

Corollary 4 (Confidence intervals) *When $p_x = 1$, set $\widehat{\mathrm{CI}} := \widehat{\theta}_n^{mod} \pm z_{1-\alpha/2}\widehat{\sigma}_n^{mod}/\sqrt{n}$, where $(\widehat{\sigma}_n^{mod})^2 = \frac{1}{n}\sum_{i=1}^n (d_i - \bar{d})^2$ for $\bar{d} = \frac{1}{n}\sum_{i=1}^n d_i$, and $d_i := \widehat{\mathbb{E}}[XX^\top]^{-1}\big(X_i\widehat{\mu}_y(Z_i) + \widehat{\mu}_x(Z_i)Y_i - \widehat{\mu}_x(Z_i)\widehat{\mu}_y(Z_i) - X_iX_i^\top\widehat{\theta}_n^{mod}\big)$ from Algorithm 1. Here, $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of a standard Gaussian random variable. Then, under the conditions of Theorem 2, $(\widehat{\sigma}_n^{mod})^2 \xrightarrow{P} \mathrm{Var}(\phi^{mod}(X,Y,Z))$, and thus $\mathbb{P}(\theta^* \in \widehat{\mathrm{CI}}) \to 1-\alpha$ as $n \to \infty$.*

One may also learn and adapt to the dependence structure from data (see Section 3.3 for a related discussion). In such cases, confidence intervals have to be adjusted to account for the variation induced by estimation of the structure. In practice, one can deal with this issue by data-splitting (i.e., use one fold of data for model selection and the second for estimation) and cross-fitting (Chernozhukov et al., 2018).

Remark 5 (Efficiency gain over OLS) *To gain more intuition on the improvement in parameter estimation, we consider a one-dimensional special case where $Z = \alpha X + \epsilon_z$, $Y = \beta Z + \epsilon_y$ for $\alpha, \beta \in \mathbb{R}$. Here $X \sim N(0, \sigma_x^2)$, and $\epsilon_z \sim N(0, \sigma_z^2)$, $\epsilon_y \sim N(0, \sigma_y^2)$ are independent random noise. The true parameter for OLS regression is $\theta^* = \alpha\beta$. The OLS estimator $\widehat{\theta}_n^{ols}$ has asymptotic variance $\sigma_{ols}^2 = \mathrm{Var}(X)^{-2}\,\mathrm{Var}(X\epsilon_y + X\beta\epsilon_z) = \mathrm{Var}(X)^{-1}(\sigma_y^2 + \beta^2\sigma_z^2)$, and our modular estimator $\widehat{\theta}_n^{mod}$ has asymptotic variance $\sigma_{mod}^2 = \mathrm{Var}(X)^{-2}\,\mathrm{Var}(\mu_x(Z)\epsilon_y + X\beta\epsilon_z) = \mathrm{Var}(X)^{-1}(\sigma_y^2\frac{\mathrm{Var}(\mu_x(Z))}{\mathrm{Var}(X)} + \beta^2\sigma_z^2)$. That is, the absolute improvement in asymptotic variance is $\sigma_{ols}^2 - \sigma_{mod}^2 = \sigma_y^2\frac{\mathrm{Var}(X-\mu_x(Z))}{\sigma_x^4}$; this quantity is large if σ_y^2 is large or Z explains a small proportion of the variance in X, i.e., $\mathrm{Var}(X - \mu_x(Z))$ is large compared to σ_x^2. The relative improvement in asymptotic variance is $1 - \frac{\sigma_{mod}^2}{\sigma_{ols}^2} = \frac{1}{1+\alpha^2\sigma_x^2/\sigma_z^2}\frac{1}{1+\beta^2\sigma_z^2/\sigma_y^2}$; this quantity is large if $\alpha^2\frac{\sigma_x^2}{\sigma_z^2}$ and $\beta^2\frac{\sigma_z^2}{\sigma_y^2}$ are small, which corresponds to (i) weak signal α, β, or (ii) large noise σ_z^2 compared to σ_x^2, and σ_y^2 compared to σ_z^2. In the most extreme case where $\alpha = \beta = 0$, the best prediction is 0, and we achieve the asymptotic variance $\sigma_{mod}^2 = 0$.*

Our discussion so far is mainly asymptotic, so that the bias incurred by the estimation error of $\widehat{\mu}_x$ and $\widehat{\mu}_y$ is negligible compared to the $O(1/\sqrt{n})$ statistical error. Next, we use a simple example to provide insights on the finite sample behavior of $\widehat{\theta}_n^{\mathrm{mod}}$ and $\widehat{\theta}_n^{\mathrm{ols}}$.

Remark 6 (Finite-sample efficiency) *We assume $(X, Y, Z) \in \mathbb{R}^3$ are jointly Gaussian, and $\widehat{\mu}_x$ and $\widehat{\mu}_y$ are estimated by OLS. Given OLS's $O_P(1/\sqrt{n})$ consistency, we forego cross-fitting to prevent cumbersome calculations. As a result, $\widehat{\mu}_x(z) = \widehat{\mathbb{E}}_n[Z^2]^{-1}\widehat{\mathbb{E}}_n[XZ]z$, and $\widehat{\mu}_y(z) = \widehat{\mathbb{E}}_n[Z^2]^{-1}\widehat{\mathbb{E}}_n[YZ]z$, where $\widehat{\mathbb{E}}_n[\cdot]$ denotes empirical mean over all the data. One can*

then check that $\widehat{\theta}_n^{mod} = \widehat{\mathbb{E}}_n[X^2]^{-1}\widehat{\mathbb{E}}_n[Z^2]^{-1}\widehat{\mathbb{E}}_n[YZ]\widehat{\mathbb{E}}_n[XZ]$ *and* $\widehat{\theta}_n^{ols} = \widehat{\mathbb{E}}_n[X^2]^{-1}\widehat{\mathbb{E}}_n[YX]$. *By joint Gaussianity, there exists some* $\alpha, \beta \in \mathbb{R}$ *such that* $X = \alpha Z + \epsilon_x$ *and* $Y = \beta Z + \epsilon_y$, *where* $\epsilon_x \sim N(0, \sigma_x^2)$ *and* $\epsilon_y \sim N(0, \sigma_y^2)$ *conditional on* Z. *Both* $\widehat{\theta}_n^{mod}$ *and* $\widehat{\theta}_n^{ols}$ *are unbiased for* θ^*, *while* $\mathrm{Var}(\widehat{\theta}_n^{ols}) - \mathrm{Var}(\widehat{\theta}_n^{mod}) = \sigma_y^2 \mathbb{E}\big[\frac{\widehat{\mathbb{E}}_n[Z^2]\cdot\widehat{\mathbb{E}}_n[\epsilon_x^2] - \widehat{\mathbb{E}}_n[Z\epsilon_x]^2}{n(\widehat{\mathbb{E}}_n[X^2])^2\cdot\widehat{\mathbb{E}}_n[Z^2]}\big] \geq 0$. *That is, in finite sample,* $\widehat{\theta}_n^{mod}$ *is always more efficient than* $\widehat{\theta}_n^{ols}$. *See detailed calculation and discussion in Appendix B.1.*

2.3 Modular regression in generalized linear models

The ideas outlined in the previous section also apply to generalized linear models (GLMs) as long as the estimation equation has a modular structure.

Following the general setup in Section 1, we suppose the true parameter $\theta^* \in \Theta \subseteq \mathbb{R}^{p_x}$ is the unique minimizer of $\mathbb{E}\big[\ell(X_i, Y_i, \theta^*)\big]$, where $\ell(x, y, \theta) = YX^\top\theta + h(x, \theta)$ for some function $h\colon \mathcal{X} \times \Theta \to \mathbb{R}^p$ that is convex in θ. A default estimator $\widehat{\theta}_n^{\mathrm{ml}}$ is the unique minimizer of the empirical loss (1). Maximum likelihood estimation for GLMs with log links satisfies this condition in general, such as logistic regression with $\ell(x, y, \theta) = -x^\top\theta y + \log(1 + \exp(x^\top\theta))$ for $y \in \{0, 1\}$, and Poisson regression with $\ell(x, y, \theta) = -x^\top\theta y + \exp(x^\top\theta)$ for $y \in \mathbb{N}$. More generally, for exponential regression with $\ell(x, y, \theta) = 2\log y \cdot \theta^\top x - (\theta^\top x)^2$ for $y \in \mathbb{R}^+$, one can use the transformed outcome $\widetilde{Y} := 2\log(Y)$ to make it a special case of equation (1).

To obtain a more accurate estimator for $C = \mathbb{E}[XY]$, the first two steps are exactly the same as in Section 2.1. In the third step, we use $\widehat{C}_{\mathrm{lm}}$ defined in (6), and compute

$$\widehat{\theta}_n^{\mathrm{mod}} = \operatorname*{argmin}_{\theta\in\Theta} \left\{ \frac{1}{n}\sum_{i=1}^n h(X_i, \theta) + \widehat{C}_{\mathrm{lm}}^\top\theta \right\}. \tag{7}$$

This estimator is again doubly-robust to the error of $\widehat{\mu}_x$ and $\widehat{\mu}_y$, and has smaller asymptotic variance than $\widehat{\theta}_n^{\mathrm{ml}}$ under mild conditions. Its theoretical justification follows similar ideas as before with slightly more involved technical conditions; we defer all the results to Appendix A.

3. Modular linear regression in high dimensions

Many prediction problems involve a huge number of predictive variables, entering the high-dimensional regime where p_x grows with, and can even be large than, the sample size n. A popular approach to dealing with high-dimensional data is penalized regression such as the Lasso (Tibshirani, 1996). By assuming a sparse linear regression function, high-dimensional regression has been shown to be consistent under various well-conditioning assumptions (Candes and Tao, 2005, 2007; Van de Geer, 2007; Zhang and Huang, 2008).

Our modular regression idea extends naturally to the high-dimensional setting. In this section, we show that by including an ℓ_1 penalty, modular regression improves upon the Lasso by seeking for a more efficient estimation equation. Throughout this section, we assume a sparse linear model to ensure the task is tractable.

Assumption 7 *There exists some* $\theta^* \in \mathbb{R}^{p_x}$ *with* $\|\theta^*\|_0 = s$, *such that* $\mathbb{E}[Y \mid X] = X^\top\theta^*$.

3.1 Regularized modular regression

We start with cross-fitting, and let $\widehat{\mu}_x^{(k)}$, $\widehat{\mu}_y^{(k)}$ be estimators for $\mu_x(\cdot) = \mathbb{E}[X \,|\, Z = \cdot]$ and $\mu_y(\cdot) = \mathbb{E}[Y \,|\, Z = \cdot]$ obtained from $\mathcal{I}\backslash\mathcal{I}_k$, and define the cross-terms as in (6). The only distinction from Section 2.1 is that we encourage sparsity by ℓ_1-regularization. We left

$$\widehat{\theta}_n^{\mathrm{mod}} = \underset{\theta\in\mathbb{R}^{p_x}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \theta^\top X_i X_i^\top \theta/2 - \widehat{C}_{\mathrm{lm}}^\top \theta + \lambda\|\theta\|_1 \right\} \tag{8}$$

for some regularization parameter $\lambda > 0$, where $\|\theta\|_1 = \sum_{j=1}^{p_x} |\theta_j|$. As (8) is convex in θ, the optimization can be efficiently solved similarly to the Lasso (e.g. coordinate descent). In practice, λ can be determined by cross validation.

3.2 Theoretical guarantee

With proper choice of λ, modular regression leads to a sharper upper bound on the estimation error. We assume a μ-RSC property for the design matrix $X \in \mathbb{R}^{n\times p_x}$ which is standard in the literature for the consistency of Lasso-type methods (Bickel et al., 2009; Negahban et al., 2012). Our theoretical analysis may be generalized to other conditions, which is beyond the scope of this work.

Assumption 8 *$X \in \mathbb{R}^{n\times p_x}$ obeys ζ-Restricted Strong Convexity (ζ-RSC), i.e., $\|X\Delta\|_2^2/n \geq \zeta\|\Delta\|_2^2$ for any $\Delta \in \mathbb{C}_3$, where $\mathbb{C}_3 := \left\{x \in \mathbb{R}^p \colon \|x_{S^c}\|_1 \leq 3\|x_S\|_1\right\}$, and $S = \{j \colon \theta_j^* \neq 0\}$.*

We assume entries of X and the response Y are bounded by constants. This condition is mild since the Lasso is often implemented after normalization.

Assumption 9 *$|X_j| \leq 1$ and $\|\widehat{\mu}_{x,j}^{(k)}(\cdot)\|_\infty \leq 1$ almost surely for $k = 1, 2$ and all $1 \leq j \leq p$. Also, $|Y_i| \leq c_0$ and $\|\widehat{\mu}_y^{(k)}(\cdot)\|_\infty \leq c_0$ almost surely for $k = 1, 2$ for some constant $c_0 > 0$.*

Finally, we assume $\widehat{\mu}_y^{(k)}$ and $\widehat{\mu}_x^{(k)}$ are consistent with $o(n^{-1/4})$ convergence rates.

Assumption 10 *Suppose $\|\widehat{\mu}_y^{(k)}(\cdot) - \mu_y(\cdot)\|_{L_2(\mathbb{P}_Z)} = o_P(n^{-1/4})$ for $k = 1, 2$. Also, there exists a sequence of constants $c_n \to 0$ such that for sufficiently large n,*

$$\max_{1\leq j\leq p_x, k=1,2} \mathbb{P}\left(\left\|\widehat{\mu}_{x,j}^{(k)}(\cdot) - \mu_{x,j}(\cdot)\right\|_{L_2(\mathbb{P}_Z)} \geq \frac{c_n \log(1/\delta)}{n^{1/4}} \right) \leq \delta. \tag{9}$$

Here $\widehat{\mu}_{x,j}^{(k)}(\cdot)$ is the j-th entry of $\widehat{\mu}_x^{(k)}$, an estimator for $\mu_{x,j}(\cdot) := \mathbb{E}[X_j \,|\, Z = \cdot]$.

Assumption 10 slightly differs from the commonly adopted consistency conditions, which is often of the form $\|\widehat{\mu}_{x,j}^{(k)} - \mu_{x,j}\|_{L_2(\mathbb{P}_Z)} = o_P(n^{-1/4})$, on the estimation of nuisance components. This is because we need an exponential tail bound to control all $2 \times p_x$ estimated functions $\{\widehat{\mu}_{x,j}^{(k)}\}_{j=1,k=1}^{p_x,2}$ simultaneously. Running many regressions may incur considerable computational cost in high dimensions; we provide a computational shortcut in Section 3.4.

Many estimators in the literature obey Assumption 10. When Z is high-dimensional (i.e., p_z may grow with n or be larger than n), if $\mathbb{E}[X_j \,|\, Z]$ is an s'-sparse linear function of Z,

then (9) holds for the Lasso estimator with $c_n \asymp \sqrt{s' \log(p_z)}/n^{1/4}$ when the regularization parameter is properly chosen. When Z is low-dimensional (i.e., p_z does not grow with n), under the standard assumption that $\mathbb{E}[X_j \mid Z]$ is sufficiently smooth, the well-established convergence results of sieve estimation can be turned into such bounds by exponential tail bounds for the concentration of empirical loss functions (Chen and Shen, 1998; Chen, 2007).

We show an improved bound for the estimation error of $\widehat{\theta}_n^{\mathrm{mod}}$ compared to the Lasso estimator using (X_i, Y_i). The proof of Theorem 11 is in Appendix C.2.

Theorem 11 (Finite-sample bound) *Suppose Assumptions 1, 8, 9 and 10 hold. Then there exists a sequence of constants $\{\bar{c}_n\}$ with $\bar{c}_n \to 0$ as $n \to \infty$, such that for any fixed $\delta \in (0,1)$ and any regularization parameter λ obeying*

$$\lambda \geq \frac{2}{n}\left\|\mu_y^\top X + Y^\top \mu_x - \mu_y^\top \mu_x - (X\theta^*)^\top X\right\|_\infty + \frac{\bar{c}_n(\log(3p_x/\delta))^2}{\sqrt{n}}, \tag{10}$$

it holds with probability at least $1-\delta$ that

$$\|\widehat{\theta}_n^{\mathrm{mod}} - \theta^*\|_2 \leq \frac{3\lambda\sqrt{s}}{2\zeta}.$$

Here we denote $\|z\|_\infty = \sup_j |z_j|$, $Y, \mu_y \in \mathbb{R}^n$ are vectors whose i-th entries are Y_i and $\mu_y(Z_i)$, and $X, \mu_x \in \mathbb{R}^{n \times p_x}$ are matrices whose (i,j)-th entries are X_{ij} and $\mu_{x,j}(Z_i)$.

The second term in (10) arises from the estimation error in $\widehat{\mu}_y$ and $\widehat{\mu}_x$. It enjoys a double robustness property that is similar to the low-dimensional case: under the slow convergence rate in Assumption 10, this error is negligible compared to the first term in (10) that is typically $\Omega(\sqrt{(\log p_x)/n})$ (Vershynin, 2018). The estimation error of $\widehat{\theta}_n^{\mathrm{mod}}$ is then characterized by the first term in (10).

When n is sufficiently large such that the second term in (10) is negligible, the deviation of the first term can be decided by the variance of each entry. We define the random vector $\epsilon^{\mathrm{mod}} := X^\top\theta^* X - \mu_y(Z)X - Y\mu_x(Z) + \mu_y(Z)\mu_x(Z) \in \mathbb{R}^{p_x}$. Then, in Theorem 11, choosing $\lambda \asymp L\sqrt{\max_j \mathrm{Var}(\epsilon_j^{\mathrm{mod}}) \cdot \log(p_x)/n}$ yields the estimation error

$$\|\widehat{\theta}_n^{\mathrm{mod}} - \theta^*\|_2 \leq \frac{L \cdot \sqrt{s\log(p_x)}}{\sqrt{n}\mu}\sqrt{\max_j \mathrm{Var}(\epsilon_j^{\mathrm{mod}})}, \tag{11}$$

where $L > 0$ is a universal constant. On the other hand, we let $\widehat{\theta}_n^{\mathrm{Lasso}}$ be the Lasso estimator (2). Under similar conditions like Assumption 8, existing results in the literature (Negahban et al., 2012) show that $\|\widehat{\theta}_n^{\mathrm{Lasso}} - \theta^*\|_2 \leq \frac{3\lambda\sqrt{s}}{2\mu}$ for any regularization parameter obeying $\lambda \geq \left\|\frac{2}{n}\sum_{i=1}^n (Y_i - X_i^\top\theta^*)X_i\right\|_\infty$. We define $\epsilon^{\mathrm{Lasso}} = X(Y - X^\top\theta^*)$. Similarly, choosing $\lambda \asymp L\sqrt{\max_j \mathrm{Var}(\epsilon_j^{\mathrm{Lasso}}) \cdot \log(p_x)/n}$ yields

$$\|\widehat{\theta}_n^{\mathrm{Lasso}} - \theta^*\|_2 \leq \frac{L \cdot \sqrt{s\log(p_x)}}{\sqrt{n}\mu}\sqrt{\max_j \mathrm{Var}(\epsilon_j^{\mathrm{Lasso}})}, \tag{12}$$

where $L > 0$ is the same universal scaling as above. The bounds in (11) and (12) distinguish our modular estimator from the Lasso, as one can check that $\mathrm{Var}(\epsilon_j^{\mathrm{mod}}) \leq \mathrm{Var}(\epsilon_j^{\mathrm{Lasso}})$ for all j. That is, modular regression reduces the uncertainty by a constant order.

Though this is only an upper bound analysis, our numerical experiments later on confirm the improvement in estimation accuracy. In particular, we will see that the regularization parameter λ chosen by cross-validation is smaller in modular regression. Intuitively, the reduction in $\mathrm{Var}(\epsilon_j^{\mathrm{mod}})$ allows cross-validation to choose a smaller λ than for the Lasso.

Interested readers may wonder whether the modular idea can be leveraged to construct more efficient confidence intervals following the 'debiasing' Lasso ideas (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; Van de Geer et al., 2014). While this may be feasible[1], this problem warrants a careful and rigorous investigation that is beyond the scope of this work. In addition, structure learning (which we introduce in the next part) may lead to irregular estimators that deserve extra care. In the current paper, we mainly focus on the estimation and prediction aspects, and leave this point for future work.

3.3 Robustness to the conditional independence condition

In practice, Assumption 1 may be violated, and the true dependence structure among the variables may be completely unknown. In this part, we generalize our modular regression framework to estimate and adapt to the conditional independence structure.

To be precise, Assumption 1 posits that $X_j \perp\!\!\!\perp Y \,|\, Z$ for every j. That is, Z captures all the predictive power of every X_j for Y. This condition can be violated in many ways. For instance, Z may capture all the information for a subset of variables in X, while others in X still have direct effects for Y. In this subsection, we assume that $X \perp\!\!\!\perp Y \,|\, (Z, X_{\mathcal{J}})$ for some subset $\mathcal{J} \subseteq \{1, \ldots, p_x\}$, and $X_{\mathcal{J}}$ is the vector containing X_j for all $j \in \mathcal{J}$. The choice of $\mathcal{J}$ can be from prior knowledge, or by estimating (consistently) the dependence structure among all the variables when joint observations of (X, Z, Y) are available. Learning the conditional independence structure is beyond our focus; popular methods in the literature include constraint-based (Spirtes et al., 2000; Margaritis and Thrun, 1999; Tsamardinos et al., 2006), score-based (Lam and Bacchus, 1994; Jordan, 1999; Friedman et al., 1999), and regression-based (Lee et al., 2006; Meinshausen and Bühlmann, 2006; Roth, 2004; Banerjee et al., 2006) ones assuming a high-dimensional graphical model, to name a few.

Remark 12 (Practical recommendation) *For high-dimensional linear regression, a heuristic idea is to use the Lasso for structure learning. For example, one can regress Y on (X, Z) via the Lasso, and use all features in X with nonzero estimated coefficient as $X_{\mathcal{J}}$. In our numerical experiments in Section 5, We find that this heuristic approach works well in improving estimation accuracy.*

In the following, we outline how to use the output from structure learning such as in Remark 12 in conjunction with modular regression. The condition $X \perp\!\!\!\perp Y \,|\, (Z, X_{\mathcal{J}})$ can be seen as a special case of Assumption 1 for a different choice of the conditioning set "Z":

$$X \perp\!\!\!\perp Y \,|\, Z^{\mathrm{full}}, \quad \text{where} \quad Z^{\mathrm{full}} = (Z, X_{\mathcal{J}}). \tag{13}$$

1. Note the 'modular' structure in the debiased estimator $\widehat{\beta}^{\mathrm{debias}} = \widehat{\beta} + \widehat{\Theta}(X^\top Y - X^\top X\widehat{\beta})/n$, where $\widehat{\beta}$ is the Lasso estimator, and $\widehat{\Theta}$ is the estimated precision matrix. A natural and heuristic idea is to apply modular regression to improve the estimation of $X^\top Y$ and $\widehat{\beta}$.

This again allows us to break the problem into sub-tasks. In learning μ_y, after data splitting, for each $k = 1, 2$, we aim $\widehat{\mu}_y^{(k)}$ for $\mathbb{E}[Y \,|\, Z^{\text{full}}]$ using data in $\mathcal{I}\backslash\mathcal{I}_k$, instead of $\mathbb{E}[Y \,|\, Z]$. We only learn $(\mu_x)_j$ for $j \notin \mathcal{J}$: we let $\widehat{\mu}_{x,j}^{(k)}(x)$ be an estimate for $\mathbb{E}[X_j \,|\, Z^{\text{full}}]$ using the data in $\mathcal{I}\backslash\mathcal{I}_k$ for $k = 1, 2$. Then for each $i \in \mathcal{I}_k$, we compute the cross-term C_i via

$$(C_i)_j = \begin{cases} (X_i)_j \widehat{\mu}_y^{(k)}(Z_i^{\text{full}}) + \widehat{\mu}_{x,j}^{(k)}(Z_i^{\text{full}}) Y_i - \widehat{\mu}_{x,j}^{(k)}(Z_i^{\text{full}}) \widehat{\mu}_y^{(k)}(Z_i^{\text{full}}), & \text{if } j \notin \mathcal{J}, \\ (X_i)_j Y_i, & \text{if } j \in \mathcal{J}. \end{cases} \tag{14}$$

Finally, we solve the same modular least squares (5) or penalized least squares (8) with $\widehat{C}_{\text{lm}} := \frac{1}{n}\sum_{i=1}^n C_i$ for the above C_i. When Assumption 1 is violated, the original $\widehat{C}_{\text{lm}}$ defined in (6) might be biased even if the estimators for conditional expectations are correct. In contrast, the new cross-terms we derive in (14) are unbiased for $(X_i)_j Y_i$ with potentially smaller variance under the generalized condition (13).

Let $\widehat{J}$ be the output of the structure learning step, and set $Z^{\widehat{\text{full}}} = (X_{\widehat{J}}, Z)$. We note that $Y \perp\!\!\!\perp X \,|\, Z^{\widehat{\text{full}}}$ holds with probability tending to 1 if $\widehat{J}$ converges in probability to some superset $\widetilde{J} \supseteq \mathcal{J}$, our asymptotic expansion in Theorems 2 and 11 holds on an event with probability tending to 1, and theoretical guarantees for this approach can be directly implied in view of (13) and the fact that the definition of C_i in (14) is equivalent to taking $\widehat{\mu}_{x,j}^{(k)}(Z_i^{\text{full}}) := (X_i)_j$, which we omit here. The convergence of $\widehat{J}$ can be obtained under, e.g., irrepresentability-type conditions (Zhao and Yu, 2006), and is typically compatible with Assumption 10; see Appendix B.3 for a more detailed discussion.

This variant can be viewed as a data-driven interpolation between the fully modular regression we introduce in the preceding part and the standard OLS or Lasso. If the conditional independence condition holds for X_j, then utilizing the information in Z^{full} reduces the estimation error in $[\widehat{\theta}_n^{\text{mod}}]_j$; otherwise, it reduces to $[\widehat{\theta}_n^{\text{mod}}]_j = [\widehat{\theta}_n^{\text{ols}}]_j$ in the low-dimensional case, and yields a similar bound as $\widehat{\theta}_n^{\text{Lasso}}$ for the high-dimensional regression.

In general, there is a trade-off between the robustness (related to accuracy of $\widehat{J}$) and efficiency gain of $\widehat{\theta}_n^{\text{mod}}$. When $\widehat{J} = J$, $\widehat{\theta}_n^{\text{mod}}$ achieves the most efficiency gain. When $\widehat{J} = \{1, \ldots, p\}$, then $\widehat{\theta}_n^{\text{mod}}$ reduces to $\widehat{\theta}_n^{\text{mod}}$ without any improvement.

Our simulation studies in Section 5 show that this approach robustly improves the estimation and prediction accuracy in cases where Assumption 1 is violated but the conditional independence structure (13) may be learned from the data.

3.4 Practical implementation via linear transformation

We now discuss a computational shortcut for high-dimensional modular regression. By using linear transformations for estimating the conditional mean functions, it reduces the computational costs and is readily compatible with standard implementation of the Lasso (e.g., `glmnet` R-package (Friedman et al., 2010)).

First, let us discuss why the estimator as defined in equation (8) is computationally demanding. Note that the modular estimator (8) minimizes

$$\frac{1}{n}\big\{\theta^\top X^\top X\theta/2 - \widehat{\mu}_y^\top X\theta - Y^\top \widehat{\mu}_x\theta + \widehat{\mu}_y^\top \widehat{\mu}_x\theta\big\} + \lambda\|\theta\|_1 \tag{15}$$

where $\widehat{\mu}_x \in \mathbb{R}^{n\times p_x}$ is a matrix whose (i,j)-th entry stores an estimator for $\mu_{x,j}(Z_i)$, and $\widehat{\mu}_y \in \mathbb{R}^n$ is a vector whose i-th entry is an estimator for $\mu_y(Z_i)$. In the cross-fitting approach we outline in Section 3.1, the estimators are specified as $[\widehat{\mu}_x]_{i,j} = \widehat{\mu}^{(k)}_{x,j}(Z_i)$ and $[\widehat{\mu}_y]_i = \widehat{\mu}^{(k)}_y(Z_i)$ for $i \in \mathcal{I}_k$. That is, we need to run $\Omega(p_x)$ times of regression to obtain $\widehat{\mu}^{(k)}_x$ and $\widehat{\mu}^{(k)}_y$.

Here, we take a different approach to estimating μ_x and μ_y:

$$\widehat{\mu}_x = \mathbf{\Pi}_x X, \quad \widehat{\mu}_y = \mathbf{\Pi}_y Y,$$

where $\mathbf{\Pi}_x, \mathbf{\Pi}_y \in \mathbb{R}^{n\times n}$ are symmetric matrices. Examples include OLS regression for $\Pi_x = \Pi_y = Z(Z^\top Z)^{-1}Z^\top$ and ridge regression (with ℓ_2-penalty parameter η) for $\Pi_x = Z(Z^\top Z + \eta\mathbf{I})^{-1}Z^\top$, where $Z \in \mathbb{R}^{n\times p_z}$ is the data matrix, $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identity matrix, and we call η the ridge penalty for clarity. We then compute the modular estimator by minimizing (15), or equivalently,

$$\frac{1}{n}\big\{\theta^\top X^\top X\theta/2 - Y^\top(\mathbf{\Pi}_y + \mathbf{\Pi}_x - \mathbf{\Pi}_y\mathbf{\Pi}_x)X\theta\big\} + \lambda\|\theta\|_1. \tag{16}$$

The objective (16) is equivalent to the Lasso estimator (2) applied to the design matrix X and the response vector $(\mathbf{\Pi}_y + \mathbf{\Pi}_x - \mathbf{\Pi}_x\mathbf{\Pi}_y)Y$. Our modular estimator could then be computed with standard libraries or packages for the Lasso (Friedman et al., 2010). The parameters in Π_x, Π_y (such as the ridge penalty) can also be chosen with cross-validation. In our real data experiments, this computation shortcut using ridge regression and cross-validated ridge penalty η achieves a prediction accuracy that is comparable to the fully modular approach (with entry-wise regression) and better than the Lasso.

As this shortcut combines ridge regression and the Lasso, it is related to the LAVA estimator (Chernozhukov et al., 2017) that is designed for recovering sums of dense and sparse signals. We develop a different estimator than theirs, which also serves a distinct goal of improving efficiency by exploiting the conditional independence structure.

4. Extension to missing data

Modular regression allows for flexible combination of data sets in various missing data settings. In this part, we discuss a general scenario where we may have access to a collection of pairwise observations (X_i, Z_i) or (Z_i, Y_i) and/or some tuples (X_i, Z_i, Y_i).

While our work finds deep connections to the missing data literature (see Section 1.3 for more discussion), a major distinction is that we mainly focus on linear models and their extensions. We leverage specific algorithmic structure (the product structure of expectation) and general independence patterns among variables, which differs from the missing data literature that usually draws upon independence between variables and the missing indicator. Furthermore, the modular algorithmic structure allows us to deal with various scenarios of data availability with a unified approach.

Throughout, we assume a 'completely-missing-at-random' (MCAR) mechanism, such that the probability of missing any variable is independent of (X, Y, Z). Formally, suppose we have access to data sets $\mathcal{I}^{xyz} = \{(Y_i, X_i, Z_i\}_{i=1}^{n_{xyz}}$, $\mathcal{I}^{xz} = \{(X_i, Z_i)\}_{i=1}^{n_{xz}}$, $\mathcal{I}^{yz} = \{(Y_i, Z_i)\}_{i=1}^{n_{yz}}$. For our exposition, it suffices to impose the following as a consequence of MCAR.

Assumption 13 *There exists a super-population $\mathbb{P}_{X,Y,Z}$, such that the joint observations obey $\{(X_i, Y_i, Z_i\}_{i=1}^{n_{xyz}} \overset{i.i.d.}{\sim} \mathbb{P}_{X,Y,Z}$, and the pairwise observations obey $\{(X_i, Z_i)\}_{i=1}^{n_{xz}} \overset{i.i.d.}{\sim} \mathbb{P}_{XZ}$ and $\{(Y_i, Z_i)\}_{i=1}^{n_{yz}} \overset{i.i.d.}{\sim} \mathbb{P}_{YZ}$, i.e., the marginal distributions are consistent with $\mathbb{P}_{X,Y,Z}$.*

We will see that the decomposition of the regression task allows us to flexibly modify our methods in Sections 2 and 3 according to the availability of data. We provide a general procedure in Algorithm 2 for low-dimensional linear regression. Extension to GLMs and high-dimensional setting can be similarly obtained by replacing $\theta^\top X_i X_i^\top \theta$ by $h(X_i, \theta)$ as in Section 2.3 or adding an ℓ_1-regularizer to the estimation equation as in Section 3.

Algorithm 2 Modular linear regression with missing data

Input: Datasets $\mathcal{I}^{xyz} = \{(Y_i, X_i, Z_i\}_{i=1}^{n_{xyz}}$, $\mathcal{I}^{xz} = \{(X_i, Z_i)\}_{i=1}^{n_{xz}}$, $\mathcal{I}^{yz} = \{(Y_i, Z_i)\}_{i=1}^{n_{yz}}$.

1: Randomly split $\mathcal{I}^{xyz}$, $\mathcal{I}^{xz}$ and $\mathcal{I}^{yz}$ into two equal-sized folds $\mathcal{I}_k^{xyz}$, $\mathcal{I}_k^{xz}$, $\mathcal{I}_k^{yz}$, $k = 1, 2$.

2: **for** $k = 1, 2$ **do**

3: Fit models $\widehat{\mu}_x^{(k)}(\cdot)$ for $\mathbb{E}[X \mid Z = \cdot]$ using $\{(X_i, Z_i) \colon i \in (\mathcal{I}^{xyz} \backslash \mathcal{I}_k^{xyz}) \cup (\mathcal{I}^{xz} \backslash \mathcal{I}_k^{xz})\}$;

4: Fit models $\widehat{\mu}_y^{(k)}(\cdot)$ for $\mathbb{E}[Y \mid Z = \cdot]$ using $\{(Y_i, Z_i) \colon i \in (\mathcal{I}^{xyz} \backslash \mathcal{I}_k^{xyz}) \cup (\mathcal{I}^{yz} \backslash \mathcal{I}_k^{yz})\}$;

5: Compute $\widehat{\mu}_x(Z_i) = \widehat{\mu}_x^{(k)}(Z_i)$ and $\widehat{\mu}_y(Z_i) = \widehat{\mu}_y^{(k)}(Z_i)$ for all $i \in \mathcal{I}_k$.

6: **end for**

7: Compute $\widehat{C}_{\text{miss}}^{zz} = \frac{1}{n_{xz}+n_{yz}+n_{xyz}} \sum_{i \in \mathcal{I}^{xz} \cup \mathcal{I}^{yz} \cup \mathcal{I}^{xyz}} \widehat{\mu}_y(Z_i)\widehat{\mu}_x(Z_i)$,
$\widehat{C}_{\text{miss}}^{xz} = \frac{1}{n_{xz}+n_{xyz}} \sum_{i \in \mathcal{I}^{xz} \cup \mathcal{I}^{xyz}} X_i \widehat{\mu}_y(Z_i)$, and $\widehat{C}_{\text{miss}}^{yz} = \frac{1}{n_{yz}+n_{xyz}} \sum_{i \in \mathcal{I}^{yz} \cup \mathcal{I}^{xyz}} Y_i \widehat{\mu}_x(Z_i)$.

8: Compute $\widehat{C}_{\text{miss}} = \widehat{C}_{\text{miss}}^{xz} + \widehat{C}_{\text{miss}}^{yz} - \widehat{C}_{\text{miss}}^{zz}$, $\widehat{\Sigma}_{\text{miss}} = \frac{1}{n_{xz}+n_{xyz}} \sum_{i \in \mathcal{I}^{xz}+\mathcal{I}^{xyz}} X_i X_i^\top$.

Output: The modular estimator $\widehat{\theta}_n^{\text{miss}} = \operatorname{argmin}_{\theta \in \mathbb{R}^{p_x}} \big\{\theta^\top \widehat{\Sigma}_{\text{miss}} \theta/2 - \widehat{C}_{\text{miss}}^\top \theta\big\}$.

Algorithm 2 is operable even when no joint observations of (X, Z, Y) are available, i.e., $\mathcal{I}^{xyz} = \varnothing$ and $n_{xyz} = 0$. This is inspired by the crucial fact that (6) only involves pairwise observations of (X, Z) and (Z, Y), and the same for learning $\mu_x(z)$ and $\mu_y(z)$. When $n_{xz} = n_{yz} = 0$, it reduces to Algorithm 1 that uses joint observations of (X, Y, Z).

The next proposition generalizes Theorem 2, whose proof is in Appendix C.3. Similar results can be derived for GLMs and the high-dimensional case, which we omit for brevity.

Proposition 14 *Suppose Assumptions 1 and 13 hold. Suppose $n_{xz}/(n_{xz} + n_{yz} + n_{xyz}) \to \rho_{xz}$, and $n_{yz}/(n_{xz} + n_{yz} + n_{xyz}) \to \rho_{yz}$ for some $\rho_{xz}, \rho_{yz} \in (0, 1)$. Denote $N_{xz} = n_{xz} + n_{xyz}$ and $N_{yz} = n_{yz} + n_{xyz}$. Assume $\|\widehat{\mu}_x^{(k)} - \mu_x\|_{L_2(\mathbb{P}_Z)} \cdot \|\widehat{\mu}_y^{(k)} - \mu_y\|_{L_2(\mathbb{P}_Z)} = o_P(1/\sqrt{N_{xz}}) + o_P(1/\sqrt{N_{yz}})$ for $k = 1, 2$, $\mathbb{E}[XX^\top] \succ 0$ is finite, and $X(Y - X^\top\theta^*)$ has finite second moments. Then $\widehat{\theta}_n^{mod} - \theta^* = \mathbb{E}[XX^\top]^{-1}\big\{\frac{1}{n_{xz}} \sum_{i \in \mathcal{I}^{xz}} \phi_{xz}(X_i, Z_i) + \frac{1}{n_{yz}} \sum_{i \in \mathcal{I}^{yz}} \phi_{yz}(Y_i, Z_i) + \frac{1}{n_{xyz}} \sum_{i \in \mathcal{I}^{xyz}} \phi_{xyz}(X_i, Y_i, Z_i)\big\} + o_P(1/\sqrt{N_{xz}} + 1/\sqrt{N_{yz}})$, where $\phi_{xz}(X, Z) = \frac{\rho_{xz}}{1-\rho_{yz}} X\mu_y(Z) - \rho_{xz}\mu_y(Z)\mu_x(Z) - \frac{\rho_{xz}}{1-\rho_{yz}} XX^\top\theta^*$, $\phi_{yz}(Y, Z) = \frac{\rho_{yz}}{1-\rho_{xz}} Y\mu_x(Z) - \rho_{yz}\mu_y(Z)\mu_x(Z)$, and $\phi_{xyz}(Y, Z) = \frac{X\mu_y(Z)}{1-\rho_{yz}} + \frac{1}{1-\rho_{xz}} Y\mu_x(Z) - \mu_y(Z)\mu_x(Z) - \frac{1}{1-\rho_{yz}} XX^\top\theta^*$. When $n_{xyz} = 0$, in the formulas above one should interpret $0/0$ as 0.*

4.1 Pairwise observations

We discuss a few consequences when only pairwise observations of (X, Z) and (Y, Z) are available. In general, predicting Y with X is impossible without identification assumptions (Assumption 1). Also, the structure learning approach in Section 3.3 is no longer feasible.

Since $n_{xyz} = 0$ and $\mathcal{I}^{xyz} = \varnothing$, we know $\rho_{xz} + \rho_{yz} = 1$. The asymptotic linear expansion in Proposition 14 reduces to $\widehat{\theta}_n^{\text{mod}} - \theta^* = \frac{1}{n_{xz}} \sum_{i \in \mathcal{I}^{xz}} \phi_{xz}(X_i, Z_i) + \frac{1}{n_{yz}} \sum_{i \in \mathcal{I}^{yz}} \phi_{yz}(Y_i, Z_i) + o_P(1/\sqrt{n_{xz}}) + 1/\sqrt{n_{yz}})$, where $\phi_{xz}(X, Z) = X\mu_y(Z) - \rho_{xz}\mu_y(Z)\mu_x(Z) - XX^\top\theta^*$, and $\phi_{yz}(Y, Z) = Y\mu_x(Z) - \rho_{yz}\mu_y(Z)\mu_x(Z)$. The modular estimator gets more efficient as n_{xz} and n_{yz} increases, while plain OLS is no longer feasible.

4.2 Partially pairwise observations

Modular prediction can also be adapted to settings where a limited number of triple observations are available in addition to (X, Z) and (Y, Z) pairs, i.e., $n_{xyz} > 0$.

In this case, one can run OLS on $\mathcal{I}^{xyz}$, whose asymptotic expansion is $\widehat{\theta}_n^{\text{ols}} - \theta^* = \frac{1}{n_{xyz}} \sum_{i \in \mathcal{I}^{xyz}} \mathbb{E}[XX^\top]^{-1}(Y_i - X_i^\top\theta^*) + o_P(1/\sqrt{n_{xyz}})$. The estimation error is of the scale $O_P(1/\sqrt{n_{xyz}})$. In contrast, by utilizing additional pairwise observations, Algorithm 2 achieves the rate of $O_P(1/\sqrt{n_{xz} + n_{xyz}}) + O_P(1/\sqrt{n_{yz} + n_{xyz}})$; this is a substantial improvement upon OLS if n_{xz} and n_{yz} are much larger than n_{xyz}, which may happen in the example at the beginning of this paper. Even if n_{xz} and n_{yz} are of the same order as n_{xyz}, i.e., $\rho_{xz} + \rho_{yz} < 1$, Algorithm 2 may still achieve a smaller asymptotic variance than OLS.

With a few (X, Z, Y) observations, one may leverage them to learn the conditional independence structure. Following the notations in Section 3.3, Z will then be replaced by Z^{full} which consist of both the original features in Z and some other features in X. Thus, in Line 7 of Algorithm 2, $\widehat{C}_{\text{miss}}^{yz}$ has to be computed with data in $\mathcal{I}^{xyz}$, and $\widehat{C}_{\text{miss}}^{zz}$ has to be computed with data in $\mathcal{I}^{xz} \cup \mathcal{I}^{xyz}$. That said, we do note that in our real data application (see Section 6.2), modular regression performs well without structure learning. Theoretically, the above procedure without structure learning may induce substantial bias in cases where conditional independence does not hold. However, in settings where we have recorded a rich set of intermediate covariates in Z (as often assumed in surrogate methods), the conditional independence assumption is often plausible or the bias is sufficiently small compared with the reduced variance.

5. Simulation studies

We evaluate our methods on simulated datasets to compare the bias, variance and overall estimation error of our methods to non-modular counterparts in both low and high dimensional settings. We also investigate the robustness to approximate conditional independence in the high dimensional setting.

5.1 Low-dimensional setting

We focus on parameter estimation in the low-dimensional setting. We are interested in estimating $\theta^* = \text{argmin}_{\theta \in \mathbb{R}^{p_x}} \mathbb{E}[(Y - X^\top\theta)^2]$, where $Y \in \mathbb{R}$ is the response, and $X \in \mathbb{R}^{p_x}$ are the covariates. We suppose in the training data we have access to $Z \in \mathbb{R}^{p_z}$ such that $X \perp\!\!\!\perp Y \mid Z$. We fix $p_x = 4$, $p_z = 6$, and a relatively small sample size at $n = 500$.

We design 2×2 data generating processes depending on (i) whether the true relation is linear and (ii) whether the data generating process follows the graphical model $X \to Z \to Y$ or $X \leftarrow Z \to Y$. The details are summarized in Table 1. The linear regression model is not necessarily well-specified for all settings, while the OLS parameter θ^* is always well-defined.

Setting	Data generating process	Comment
1	$Z \sim \text{Unif}[-1,1]$, $X = BZ + \epsilon_z$, $Y = Z^\top \gamma + \epsilon_y$	linear, $X \leftarrow Z \to Y$
2	$X \sim \text{Unif}[-1,1]$, $Z = BX + \epsilon_z$, $Y = Z^\top \gamma + \epsilon_y$	linear, $X \to Z \to Y$
3	$Z \sim \text{Unif}[-1,1]$, $X = f(Z) + \epsilon_z$, $Y = g(Z) + \epsilon_y$	nonlinear, $X \leftarrow Z \to Y$
4	$X \sim \text{Unif}[-1,1]$, $Z = f(X) + \epsilon_z$, $Y = g(Z) + \epsilon_y$	nonlinear, $X \to Z \to Y$

Table 1: Data generating processes in all settings, where $\epsilon_z \sim N(0, \sigma_z^2)$ and $\epsilon_y \sim N(0, \sigma_y^2)$ are independent noise, and f, g are nonlinear functions defined in the text.

In settings 1 and 2, we set $\gamma = (0.531, -0.126, 0.312, 0, 0, 0)^\top \in \mathbb{R}^{p_z}$, and $B \in \mathbb{R}^{p_x \times p_z}$ or $B \in \mathbb{R}^{p_z \times p_x}$ is a constant matrix where 8 out of 24 entries are randomly set to 0.5 then fixed for all configurations, while the other entries are zero. The f and g functions in Settings 3 and 4 are defined as follows. In setting 3, we set $X_4 = [BZ]_4$ with the same B as setting 1, and $X_1 = 0.5X_1 + \mathbb{1}\{Z_1 > 0\}$, $X_2 = -0.5Z_3 + \mathbb{1}\{Z_4 > 0\}$ and $X_3 = \mathbb{1}\{Z_4 > 0\}$. In setting 4, we set $Z_{4:6} = [BX]_{4:6}$ with the same B as setting 2, and $Z_1 = 0.5X_1 + \mathbb{1}\{X_1 > 0\}$, $Z_2 = -0.5X_3 + \mathbb{1}\{X_4 > 0\}$ and $Z_3 = \mathbb{1}\{X_4 > 0\}$. In all settings, $\epsilon_z \sim N(0, \sigma_z^2)$ and $\epsilon_y \sim N(0, \sigma_y^2)$ are independent noise, where we vary the noise strengths $\sigma_z, \sigma_y \in \{0.1, 0.5, 1, 2\}$, and $X \sim \text{Unif}[-1,1]$ means all entries in X are i.i.d from $\text{Unif}[-1,1]$.

We compute the OLS parameter $\widehat{\theta}_n^{\text{ols}}$ and our modular estimator with $\widehat{\mu}_y^{(k)}$ and $\widehat{\mu}_x^{(k)}$ estimated with (i) cross-validated Lasso (Friedman et al., 2010), (ii) cross-validated ridge regression (Friedman et al., 2010), (iii) regression random forest from `grf` R-package, and (iv) linear regression. All procedures are repeated for $N = 1000$ independent runs. For comprehensive illustration here, we aggregate all coefficients and evaluate the rooted mean square error (RMSE) $\mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2]^{1/2}$ (summation of) standard deviation (SD) $\sum_{j=1}^{p_x} \text{sd}(\widehat{\theta}_j - \theta_j^*)$ and bias $\sum_{j=1}^{p_x} |\mathbb{E}[\widehat{\theta}_j - \theta_j^*]|$ for the five estimators. We plot the aggregated RMSE in settings 2 and 3 in Figure 3. Bias, SD and RMSE in all settings (either aggregated or for each entry) are in Appendix E which convey similar messages.

Modular regression – no matter which machine learning regressor is used for $\widehat{\mu}_x$ and $\widehat{\mu}_y$ – achieves smaller RMSE than the plain OLS in almost all configurations. Modular regression has a larger bias than OLS due to the additional regression (Figure 14, Appendix E), yet much more substantial reduction in standard deviation (Figure 15, Appendix E) in all settings. Also, results from setting 1 confirms Remark 5, where the asymptotic variance reduction grows with both σ_z and σ_y.

A noteworthy exception is $\sigma_z = 0.1$ in setting 3, where the RMSE of modular regression with random forests is worse than that of the OLS estimator. This is because with small sample size ($n = 200$) and low noise (hence the uncertainty in $\widehat{\theta}_n^{\text{ols}}$ is very small), the bias introduced by the random forest regression is large compared to the reduction in variance. However, when σ_z becomes larger, the bias (Figure 14) has a smaller magnitude; this may be because we enter a signal-to-noise ratio regime that favors tree-based approaches.

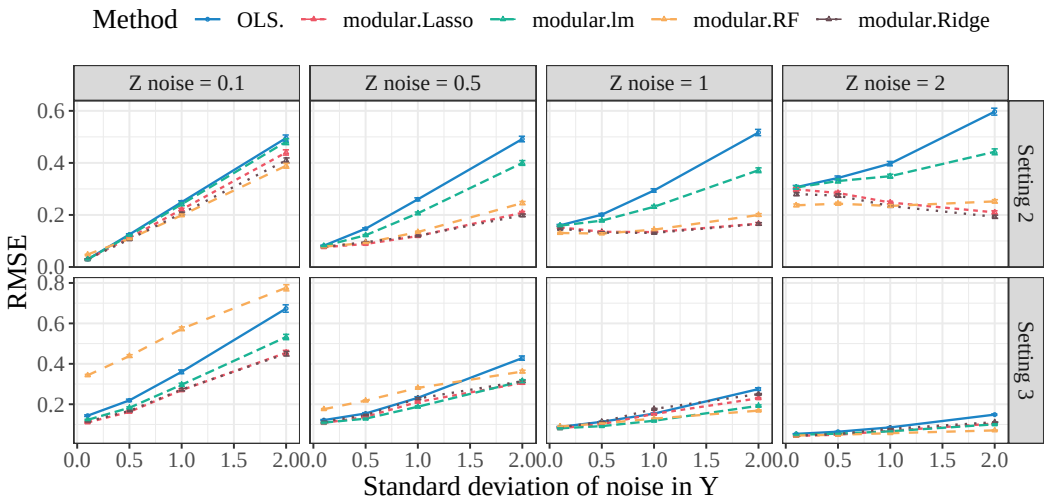


Figure 3: RMSE averaged over $N = 1000$ runs with $n = 200$. Each column corresponds to a value of σ_z. In each subplot, the x-axis (noise strength) equals σ_y, the standard deviation of noise in Y. Modular regression achieves smaller estimation error in almost all settings.

The impact of bias is also less substantial when we increase the sample size. Figure 16 in Appendix E plots the RMSE for estimating θ_1^* when $n = 2000$, where we see a much better performance of modular regression with random forests. Thus, we recommend using machine learning regressors for larger sample sizes.

We also note that modular regression with linear regression (green) performs better than the plain OLS in all settings (even when the true relation is nonlinear), although it is sometimes outperformed by other modular methods. This phenomenon is universal in our simulation (see Appendix E for RMSE for all entries in all settings), and verifies the relaxed consistency condition we observe after Theorem 2. Although this is not a general rule, we still recommend linear regression in practice especially for small sample size. One can also add a few transformed regressors into the linear regression to further adapt to nonlinearity.

Finally, to show the asymptotic behavior of modular regression, in Figure 4 we plot the aggregated RMSE for varying sample sizes $n \in \{200, 500, 1000, 2000\}$ with $\sigma_z = 0.5$ and $\sigma_y = 1$ in all settings. Modular regression outperforms OLS in nearly all settings, and linear models (lm, Lasso and Ridge) as base estimators show robust performance. The performance of random forest is less stable for small sample size ($n = 200$) in setting 3, hence we recommend using flexible machine learning models with larger sample sizes.

5.2 High-dimensional setting

We now consider two data generating processes in the high-dimensional setting, where the conditional independence structure holds only in one of them. We show that our methods outperform the Lasso in Setting 1 (with conditional independence), and is robust against the violation of conditional independence in Setting 2. Following the preceding notations, $X \in \mathbb{R}^{p_x}$ is the covariates available in the prediction phase, while $Z \in \mathbb{R}^{p_z}$ is only available

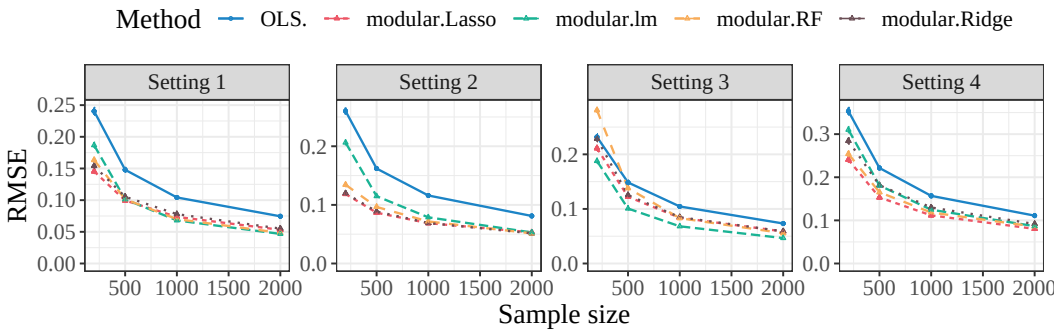


Figure 4: RMSE averaged over $N = 1000$ runs for various sample size, and $\sigma_z = 0.5$, σ_y=1. Modular regression outperforms OLS in all settings, and its performance improves with n.

for training data, and $Y \in \mathbb{R}$ is the outcome. The data generating processes are visualized in Figure 5.

(a) Setting 1: conditional independence

(b) Setting 2: approximate conditional independence

Figure 5: Illustration of the two data generating processes.

In settting 1, we generate $X \in \mathbb{R}^{p_x}$ where each entry is i.i.d. from $N(0,1)$; then we generate $Z = BX + \epsilon_z \in \mathbb{R}^{p_x}$ with i.i.d. noise $\epsilon_z \sim N(0,1)$ given a parameter matrix $B \in \mathbb{R}^{p_z \times p_x}$. Finally, we generate $Y = Z^\top \gamma + \epsilon_y$ for i.i.d. random noise $\epsilon_y \sim N(0,4)$ and some $\gamma \in \mathbb{R}^{p_z}$. To ensure sparsity, we let $B_{ij} = 0$ for all $j > s$, such that X_j's for $j > s$ are irrelevant for the prediction. Then for each $j \le s$, we randomly select $2s$ entries in the j-th column of B with $B_{ij} = 0.25$, the remainings with $B_{ij} = 0$. We set $\gamma_i = 0.5$ for $1 \le i \le s$, and $\gamma_i = 0$ for $i > s$. In this way, $Y = X^\top \theta^* + (\gamma^\top \epsilon_z + \epsilon_y)$ where $\theta^* = B^\top \gamma$, and $X \perp\!\!\!\perp Y \,|\, Z$.

In setting 2, we ensure a small subset of covariates in X to have direct impact on Y (the link from X to Y in Figure 5(b)). To be specific, we generate X and $Z = BX + \epsilon_z$, where $\{B_{ij} : j \le s\}$ and ϵ_z are the same as setting 1, and generate $Y = Z^\top \gamma + X^\top \widetilde{\gamma} + \epsilon_y$ for i.i.d. random noise $\epsilon_y \sim N(0,4)$; the direct coefficient $\widetilde{\gamma}$ satisfies $\widetilde{\gamma}_i = 0$ for $i \le s$ and $\sum_{i=1}^{p_x} \mathbb{1}\{\widetilde{\gamma}_i \ge 0\} = 5$. In this setting, the high dimensional model $Y = X^\top \theta^* + (\gamma^\top \epsilon_z + \epsilon_y)$ holds with $\theta^* = B^\top \gamma + \widetilde{\gamma}$, but $X \perp\!\!\!\perp Y \,|\, Z$ does not hold exactly.

We compare our method in Section 3.1 to the Lasso; to ensure fair comparison, we run the Lasso using our modular method while setting $\widehat{f}_y(Z_i) := Y_i$ and $\widehat{f}_x(Z_i) := X_i$ for all i, so that (8) reduces to the Lasso. We also evaluate an oracle modular regression algorith, that is, we set $\widehat{\mu}_y := \mathbb{E}[Y \,|\, Z]$ and $\widehat{\mu}_x := \mathbb{E}[X_j \,|\, Z]$ as the ground truth. The regularization parameter λ for ℓ_1-penalty is chosen by 5-fold cross validation on the training data for

all methods. In our modular regression algorithm, we fit $\widehat{\mu}_x$ and $\widehat{\mu}_y$ by ridge regression using the `cv.glmnet` function from the `glmnet` R-package (Friedman et al., 2010). The procedures are evaluated for $p_x = p_z = 100$ and $s = 10$. The training sample is $n = 500$ and we evaluate the prediction on $n_{\text{test}} = 1000$ test samples for a relatively accurate estimate.

5.2.1 Performance under conditional independence

In setting 1 with a well-specified high-dimensional linear model and exact conditional independence $X \perp\!\!\!\perp Y \mid Z$, we evaluate (i) the parameter estimation error $\|\widehat{\theta}_j - \theta_j^*\|_2$ where $\widehat{\theta}$ is the output of modular regression or the Lasso, as well as (ii) prediction performance: excess risk $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (X_i^\top \widehat{\theta} - X_i^\top \theta^*)^2$ and mean squared error (MSE) $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (X_i^\top \widehat{\theta} - Y_i)^2$ on the test samples.

Parameter estimation. For each $j \in [p_x]$, we evaluate the rooted mean squared error (RMSE) $(\widehat{\theta}_j - \theta_j^*)^2$ and bias $|\mathbb{E}[\widehat{\theta}_j] - \theta_j^*|$ over $N = 100$ replicates, and visualize the RMSE (left) and bias (right) via boxplots in Figure 12.

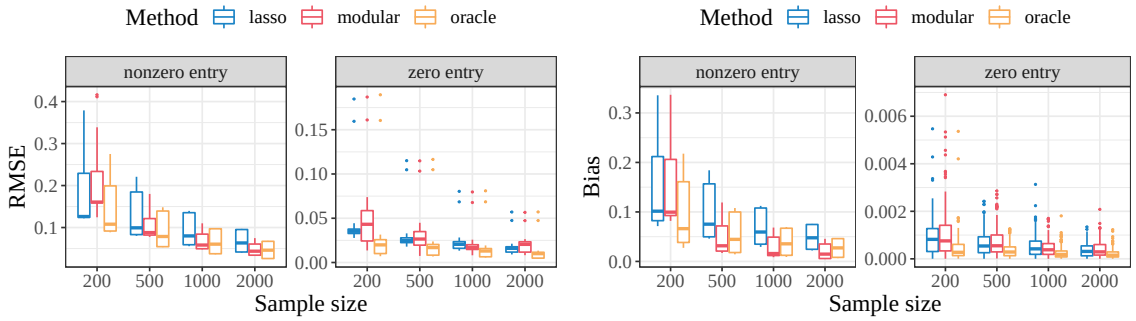


Figure 6: Boxplot of RMSE (left) and bias (right) of $\{\widehat{\theta}_j : j \in [p_x]\}$, averaged over all replicates in setting 1, separately for $\theta_j^* \neq 0$ (nonzero entry) and $\theta_j^* = 0$ (zero entry). The x-axis indicates the method to obtain $\widehat{\theta}_j$. Modular regression achieves smaller RMSE and smaller bias than the Lasso due to a different bias-variance trade-off.

For nonzero entries (the blue boxplots), we observe a significant reduction in RMSE compared to the Lasso under various sample sizes; despite the estimation error in fitting the conditional mean, our method is comparable to its oracle counterpart, even with a smaller overall RMSE (though less stable across entries). This might be due to the instability in cross-fitting with ridge regression or achieving a better bias-variance tradeoff by cross-validation. Also, as n increases, modular regression gets more and more stable. For zero entries, the RMSE are similar across three methods: while the oracle yields lower RMSE than the Lasso, the slight inflation of RMSE in modular prediction might be due to the estimation error of $\widehat{\mu}_x$ and $\widehat{\mu}_y$.

We find that the reduction in RMSE for nonzero entries mainly comes from the reduction in bias, as illustrated by the right panel of Figure 12. This is consistent with our theory in Theorem 11: The reduced variance of the proxy $\widehat{C}_{\text{lm}}$ allows the cross-validation step to choose a model with smaller bias. Due to limited space, we defer the corresponding plot of standard deviations to Figure 21 in Appendix E.

Prediction. We plot the $N = 1000$ excess risks averaged on the test samples for various training sample size n in Figure 7. It shows significant improvement in prediction accuracy of our method compared to plain Lasso; modular regression is slightly inferior to the oracle counterpart but the difference is relatively moderate. The error of modular regression also gets smaller as the sample size increases.

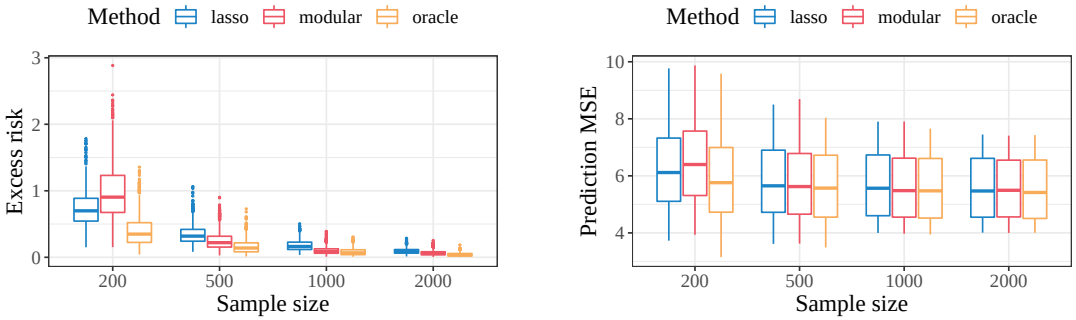


Figure 7: Boxplot of $N = 1000$ empirical excess risks $\mathbb{E}[(X_i^\top \widehat{\theta} - X_i^\top \theta^*)^2]$ and MSEs $\mathbb{E}[(X_i^\top \widehat{\theta} - Y_i)^2]$ in all replicates for setting 1. Modular regression achieves smaller excess risk than the Lasso. The improvement in prediction MSE is less visible due to the irreducible error.

5.2.2 Robustness to approximate conditional independence

In the following, we test the robustness of our method against potential violation of the conditional independence assumption. In setting 2 where the conditional independence only approximately holds, we additionally conduct a structure learning step using Lasso.

We first run a cross-validated Lasso of Y on (X, Z) using the `cv.glmnet` function (Friedman et al., 2010) for model selection; all X_j that are selected by this Lasso step is then merged into Z. For any selected X_j, we will skip the regression of $\mathbb{E}[X_j \mid Z]$ and directly set $\widehat{f}_{x,j}(Z_i) = X_{i,j}$ for all training samples. We also evaluate an oracle counterpart which uses the ground truth of the structure and the true conditional expectations for those $X_j \perp\!\!\!\perp Y \mid Z$; we skip the regression for those X_j with a direct impact on Y, as outlined in Section 3.3.

Figure 8 plots the RMSE (left) and bias (right) of all coefficients $\widehat{\theta}_j$ with various sample sizes, averaged over $N = 1000$ replicates. The plot for standard deviations of $\widehat{\theta}_j$ is in Figure 21 in Appendix E. We again see an improved RMSE especially for nonzero entries (blue). While the RMSE is less stable across different entries, perhaps because of the additional uncertainty introduced in the structure learning step, it is in general better than plain Lasso and comparable to the oracle. Similar to the previous setting, this reduction of RMSE mainly comes from a reduced bias as shown in the right panel of Figure 8. In general, our method is able to adapt to approximate conditional independence and maintain certain efficiency gain.

We summarize the prediction performance in Figure 9. The excess risk of modular prediction lies between that of the oracle counterpart and the Lasso. Still, the relative improvement in terms of prediction MSE is present yet smaller due to the irreducible noise.

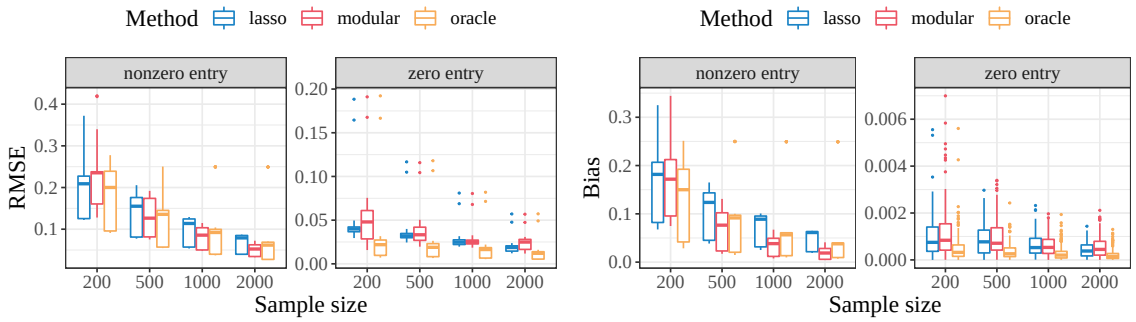


Figure 8: Boxplot of RMSE (left) and bias (right) of $\widehat{\theta}_j$, $j \in [p_x]$, averaged over all replicates, for $\theta_j^* \neq 0$ (nonzero entry) and $\theta_j^* = 0$ (zero entry) separately. Modular regression achieves smaller estimation RMSE and bias than the Lasso, due to a different bias-variance trade-off.

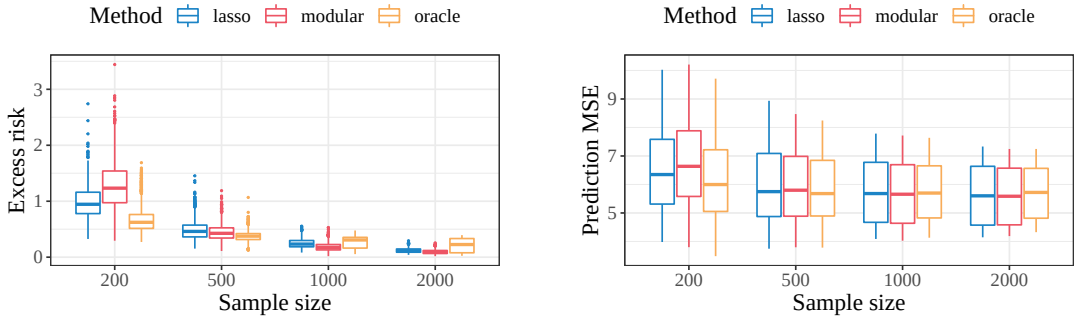


Figure 9: Boxplot of $N = 100$ empirical excess risks $(X_i^\top \widehat{\theta} - X_i^\top \theta^*)^2$ (left) and MSEs $(X_i^\top \widehat{\theta} - Y_i)^2$ (right) in all replicates for setting 2. Modular regression with structure learning achieves smaller prediction excess risk than the Lasso.

6. Real data analysis

6.1 Dataset overview

We apply our method to the English Longitudinal Study of Ageing dataset (Oldfield et al., 2021). The waves of data consistently measure several modules of features such as health trajectories, disability and healthy life expectancy, the economic and financial situations, cognition and mental health, etc. We use the Wave 7 and Wave 9 data, collected in 2014 and 2018, respectively, restricted to people who are present in both waves.

We consider predicting the future health outcomes of people based on their current available features. The Wave 7 data is used as covariates. We divide all variables into two categories: health (both mental and physical) and social conditions (including household demographics, financial, work and social situations). After pre-processing through one-hot encoding for categorial and string-valued variables and filtering out some highly imbalanced variables, the health and social categories contain 184 and 888 features, respectively. We take the `hehelf` variable from Wave 9 data as the response Y: the reported overall health situation ranging from 1 (excellent) to 5 (poor). We treat the outcome as continuous.

6.2 Real data application

We take the covariates in the social category as $X \in \mathbb{R}^{p_x}$ for $p_x = 888$, and those in the health category as $Z \in \mathbb{R}^{p_z}$ for $p_z = 184$. Our procedure is geared towards settings with high noise. To simulate such a setting, we smooth the discrete response Y by adding i.i.d. noise drawn from $N(0, 4)$. The task is to predict Y using X, while Z may be available during the training process. This is practical setting where health conditions (Z) may be more costly or difficult to evaluate, hence only available in pre-collected data.

We consider a scenario where one only has access to a limited number n of (X, Y, Z) triples as well as n_{xz} observations for (X, Z) and n_{yz} observation for (Z, Y) pairs in the training phase. This mimics a realistic scenario where it is difficult to obtain full observations simultaneously but modular data are more easily accessible. When $n_{xz} = n_{yz} = 0$, it reduces to the standard full-observation setting. While it is more difficult to test for conditional independence and the modeling assumptions with limited joint observations, we could still use our framework to merge the individual datasets and improve out-of-sample prediction. We focus on the prediction MSE on the test sample because no ground truth is available. We consider two scenarios:

(i) Fixed n and varying n_{xz} and n_{yz}. We fix $n = 200$, and the number of pair observations varies as $n_{xz} = n_{yz} = n \cdot \rho$ for $\rho \in \{0.5, 1, 5, 10\}$.

(ii) Fixed $n+n_{xz}+n_{yz}$ and varying proportion. We fix the total sample size $n+n_{xz}+n_{yz} = 1000$, while varying the proportion of joint observations by $n = 1000 \cdot \rho$, $n_{xz} = n_{yz}$ for $\rho \in \{0.05, 0.1, 0.2, 0.5, 0.8, 1\}$.

We evaluate our modular regression approach outlined in Section 4 that utilizes the partial observations, where we use 2-fold cross-fitting to obtain $\widehat{\mu}_x$ and $\widehat{\mu}_y$ from 1) the Lasso using `cv.glmnet` (10-fold cross-validation), 2) ridge regression using `cv.glmnet` (10-fold cross-validation), and 3) random forests using `grf` R-package. The parameter λ for ℓ_1-regularization is chosen by 10-fold cross-validation with the `1se` criterion, implemented in the same way as that in the `cv.glmnet` R function, i.e., we selects the largest λ within one `se` of CV error from the smallest CV error. We use the modular regression without structure learning. These implementations are compared to the default Lasso using `cv.glmnet` fitted over the (X, Y) joint observations, also using `1se` criterion for 10-fold validation. Here we omit the results for our methods and the Lasso using the `min` option (selecting minimum CV error) for cross-validation because the Lasso performs far less stable in this case. The average prediction MSEs over $N = 100$ independent runs are in Figure 10.

The left panel in Figure 10 shows the results for setting (i), where Lasso uses a fixed number of joint observations. As the number of auxiliary observations increases, our modular regression achieves smaller prediction error, showing quite substantial improvement due to incorporating auxiliary observations.

The right panel presents those for setting (ii). Naturally, the performance of the Lasso (blue, dashed line) improves as ρ, the proportion of joint observations, increases. Our modular regression, with all of the three regressors, outperforms the Lasso by utilizing auxiliary observations, including $\rho = 1$ without missing data. Surprisingly, keeping the total sample size fixed, we do not see much variation in the performance of modular regression

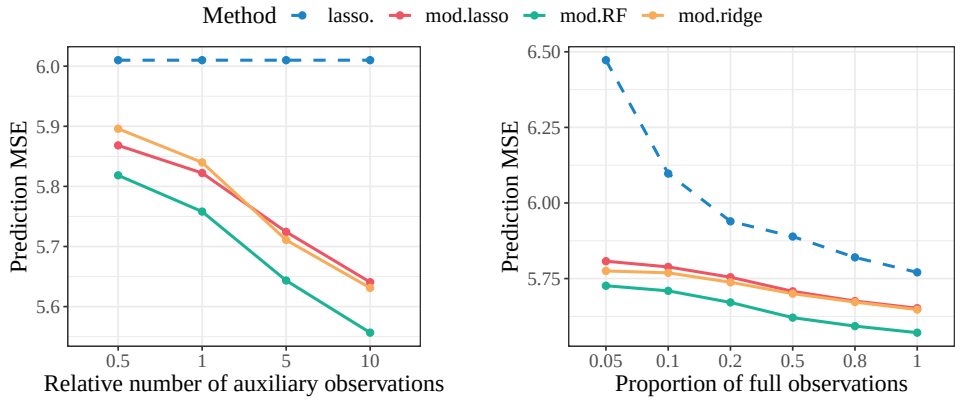


Figure 10: Average prediction MSEs on the test data in settings (i) left and (ii) right. All methods are `1se` cross-validated. The x-axis represents the value of ρ in both plots. Modular regression with all three base learners substantially reduce the prediction MSE.

(all solid lines) as ρ varies: the performance with only 5% joint observations is comparable to that with more than 50% joint observations. Our method achieves very similar effective sample size as full observations on this dataset. On the other hand, this phenomenon also indicates that we are in a regime where the irreducible error in Y is large compared to the learnable part. In the next part, we are to utilize semi-synthetic data to evaluate the performance of our method in a setting with slighly stronger signal.

6.3 Semi-synthetic data

As discussed in Section 3.4, a naive implementation of our procedure is computationally prohibitive in high-dimensional settings. Thus, in the following, we evaluate the shortcut described in Section 3.4 and compare it with the standard cross-fitting implementation. We keep the choice of X and Z as before, and randomly subsample without replacement the training and test folds, where we observe (X, Z, Y) for $n = 1000$ training data, but only X for $n_{\text{test}} = 1794$ test sample. This is a scenario where only those easier-to-measure social-related covariates are available at the time of prediction, while the pre-collected training data contain both health and social covariates.

Data generating process. As the signal-to-noise ratio in the original data (for both (X, Y) and (Z, Y) regression) is extremely low, we enhance the signal with the following synthetic data generating process to draw a more informative comparison. We standardize all features using the original dataset, from which we subsample a set of observations aside from the training and test data. On this set, we run the Lasso for Y given Z which finds 13 nonzero regression coefficients, and for Y given X which finds 12 nonzero coefficients; we then reorder the features so that $Z_{1:13}$ and $X_{1:12}$ have with nonzero coefficients. For each $j \in \{1, \ldots, 12\}$, we run a Lasso for X_j over $Z_{1:13}$ on this fold, and store all coefficients in the j-th column of a matrix $\widehat{B} \in \mathbb{R}^{13\times 12}$.

We generate the training and testing data by $Z_i = Z_i^{\text{org}} + \epsilon_i^z$, where Z_i^{org} is the original observation, and $\epsilon_i^z \sim N(0, \sigma_z^2)$ is independent noise. We then compute $\mu_x(Z_i) = 2.5 \cdot$

$\widehat{B}^\top Z_{i,1:13}$, and generate $X_{i,1:12} = \mu_x(Z_i) + \epsilon_i^x$, where $\epsilon_i^x \in \mathbb{R}^{12}$ is i.i.d. noise from $N(0, 0.25 \cdot \mathbf{1}_{12})$ to match the standardized variance in X, and $X_{13:p_x}$ are obtained by permuting each columns in the (standardized) original data matrices. Finally, we generate $Y_i = \mu_y(Z_i) + \epsilon_i^y$, where $\mu_y(Z_i) = Z_i^\top \gamma$ where the first 12 entires in $\gamma \in \mathbb{R}^{p_z}$ equals 0.5 while the others equal to zero, and $\epsilon_i^y \sim N(0, 4)$ is the i.i.d. random noise. This setup ensures $X \perp\!\!\!\perp Y \mid Z$, but the sparse linear model $\mathbb{E}[Y \mid X] = X^\top \theta^*$ does not necessarily hold, and the true parameters θ^* are not available. We thus focus on the prediction performance. We vary the signal-to-noise ratio by setting $\sigma_z \in \{1, 2\}$.

Methods. We evaluate two implementations of the modular regression:

(i) Cross fitting in Section 3.1. We use two-fold cross-fitting with cross-validated Lasso and ridge regression to form $\widehat{\mu}_x$ and $\widehat{\mu}_y$, and then use 10-fold cross-validation to decide the penalty parameter λ in (8) by either (a) `min`: minimal CV error or (b) `1se`: the same as the default implementation in `cv.glmnet` R function which selects the largest λ within one `se` of CV error for stable performance.

(ii) Projection shortcut in Section 3.4. We use ridge projection with regularization parameters (η_x, η_y) for $\mathbf{\Pi}_x, \mathbf{\Pi}_y$, and then run `cv.glmnet` (with both (a) `min` and (b) `1se` choice of cross-validation) for X and $(\mathbf{\Pi}_y + \mathbf{\Pi}_x - \mathbf{\Pi}_x\mathbf{\Pi}_y)Y$, where (η_x, η_y) are chosen by 10-fold cross-validation to minimize CV error.

The above two implementations are compared to two baselines:

(iii) Oracle modular regression. Set $\widehat{\mu}_x$ and $\widehat{\mu}_y$ as ground truth, and then the same as (i).

(iv) Lasso: Run `cv.glmnet` on (X, Y) with both (a) `min` and (b) `1se` cross-validation.

Results. The boxplots for all methods with $N = 100$ independent runs are in Figure 11, where we compare the performance under different cross-validation options.

The patterns across different values of σ_z are similar. Among the baselines, the Lasso with minimal CV error is more accurate than `1se` (see the blue boxplots in the last two columns versus in the first two columns), while the oracle modular regression always achieves smaller prediction MSE than the Lasso with the corresponding CV option (grey).

Our modular regression method performs reasonably well when the conditional mean functions $\mu_x(\cdot)$ and $\mu_y(\cdot)$ are estimated. When they are estimation by (i) ridge projection shortcut, modular regression with both `min` (see the last column) and `1se` (see the second column) improves upon the Lasso, although it is sometimes less accurate than the oracle. This shows the projection shortcut in Section 3.4 is a reliable alternative to the more computationally intensive cross-fitting approach. When they are estimated by (ii) cross fitting (see the first and third columns), our method improves upon the original Lasso with both Lasso (green) and ridge regression (yellow) as the regressor, and the performance is comparable to the oracle. The Lasso as the regressor is slightly better than the ridge regression; this may be due to the true sparse linear data generating process. In general, for the (ii) implementation, cross-validation with `min` CV error achieves smaller test MSE than `1se`, while the latter sees larger improvement upon the Lasso.

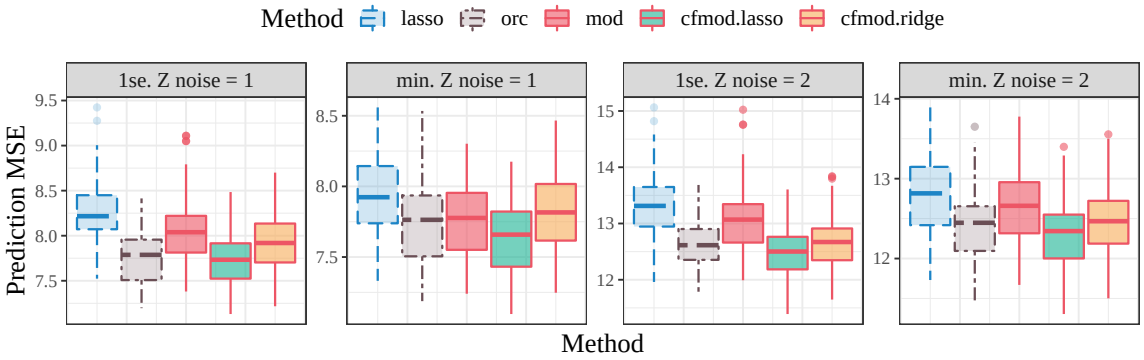


Figure 11: Boxplots of prediction MSEs on the test data. Each subplot summarizes methods with one cross validation option (`1se` or `min`) under one value of σ_z. `Method` stands for the projection shortcut (`mod`), and cross-fitting with Lasso (`cfmod.Lasso`) and ridge regression (`cfmod.ridge`). The Lasso (`Lasso`) and oracle modular regression (`orc`) with the same cv option are plotted for comparison. The oracle modular regression substantially reduces MSE; the cross-fitting implementation with Lasso and ridge regression is comparable to the oracle; the projection shortcut is slightly inferior but still improves upon the Lasso.

7. Discussion

In this work, we propose the modular regression framework and show that conditional independence structures between variables can be used to decompose statistical tasks into sub-tasks. We develop decomposition techniques for linear models in both low and high dimensional settings. We show that such decomposition can improve efficiency and allow to combine different datasets for a single estimation or prediction task with rigorous statistical guarantees. In practice, the conditional independence conditions for decomposition may be violated, leading to a bias-variance trade-off. We also develop a robust implementation of our method to adapt to potentially more complicated dependence structures.

Looking ahead, statistical tasks that allow for decomposition may go well beyond the cases studied in this work, and the assumptions for decomposition may vary with the nature of the tasks. For instance, in high dimensional graphical models, the edges between variables that indicate independence may be sparse, and conditional independence may not hold exactly for two disjoint sets (e.g., our $X \in \mathbb{R}^{p_x}$ and $Z \in \mathbb{R}^{p_z}$). In biological applications, there may exist several paths from X to Y rather than being fully mediated by Z. The dependence among the features and the response may still be sparse, but additional efforts are needed in order to leverage potential independence structures. In addition, the data fusion technique may be further extended: In practice, one may have access to many auxiliary datasets that cover different sets of features. Developing a framework to systematically combine multiple datasets may also be an interesting direction. Finally, our theoretical results in high dimensions only cover the sparse setting ($p \gg n \gg \log p$ with $s \leq \sqrt{n}$). In this regime, both the LASSO and the modular estimator are consistent, and the role of conditional independence is to lower the variance of residuals. Its role in modern asymptotic regimes such as $p/n \to \rho$ for some fixed $\rho > 0$ remains an interesting question.

Appendix A. Deferred theoretical results

This section presents the omitted theorem and proof for modular generalized linear regression in Section 2.3.

Assumption 15 *$\theta^* \in \Theta$ for a compact set Θ. Also, $\ell(x, y, \theta)$ is three-times-differentiable and convex in $\theta \in \Theta$. Let $\nabla_\theta h(x, \theta) \in \mathbb{R}^{p_x}$ denote the first-order derivative of $h(x, \theta)$ in θ, and similarly $\nabla^2_\theta h(x, \theta) \in \mathbb{R}^{p_x \times p_x}$ the second-order, and $\nabla^3_\theta h(x, \theta) \in \mathbb{R}^{p_x \times p_x \times p_x}$ the third-order ones. There exists some $L\colon \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}[L(X)] < \infty$, and $\|\nabla_\theta h(x, \theta) - \nabla_\theta h(x, \theta')\| \le L(x) \cdot \|\theta - \theta'\|$ for any $\theta, \theta' \in \Theta$. Also, there exists some $m\colon \mathcal{X} \times \Theta \to \mathbb{R}$ such that $\mathbb{E}[m(X)] < \infty$ and $\|\nabla^3_\theta h(x, \theta) - \nabla^3_\theta h(x, \theta')\|_{\mathrm{Fro}} \le m(x) \cdot \|\theta - \theta'\|$ for any $\theta, \theta' \in \Theta$.*

Theorem 16 *Suppose Assumptions 1 and 15 hold, and $\|\widehat{\mu}_x^{(k)} - \mu_x\|_{L_2(\mathbb{P}_Z)} \|\widehat{\mu}_y^{(k)} - \mu_y\|_{L_2(\mathbb{P}_Z)} = o_P(1/\sqrt{n})$ for $k = 1, 2$. Let $\widehat{\theta}_n^{mod}$ be the unique minimizer of (7), and θ^* be the unique minimizer of $\mathbb{E}\big[\ell(X_i, Y_i, \theta^*)\big]$. Let $\mu_x(\cdot) = \mathbb{E}[f(X_i) \,|\, Z_i = \cdot]$ and $\mu_y(\cdot) = \mathbb{E}[Y_i \,|\, Z_i = \cdot]$. Define the influence function $\phi(x, y, z) = \mu_x(z)g(y) + f(x)\mu_y(z) - \mu_x(z)\mu_y(z) + \nabla_\theta h(x, \theta^*)$. Then $\sqrt{n}(\widehat{\theta}_n^{mod} - \theta^*) \xrightarrow{d} N(0, \mathrm{Cov}(\phi(X_i, Y_i, Z_i)))$ and $\sqrt{n}(\widehat{\theta}_n^{mod} - \theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(X_i, Y_i, Z_i) + o_P(1/\sqrt{n})$ as $n \to \infty$. Furthermore, $\phi(X_i, Y_i, Z_i)$ is the efficient influence function for estimating θ^* under the model space obeying Assumption 1.*

Proof [Proof of Theorem 16] The proof idea is similar to that of Jin and Rothenhäusler (2021, Proposition E.2 and Theorem 3.11). Write $\widehat{\ell}(X_i, Y_i, Z_i, \theta) = \big[\widehat{\mu}_x(Z_i)g(Y_i) + f(X_i)\widehat{\mu}_y(Z_i) - \widehat{\mu}_x(Z_i)\widehat{\mu}_y(Z_i)\big]^\top \theta + h(X_i, \theta)$. We first show that $\widehat{\theta}_n^{\mathrm{mod}} \xrightarrow{P} \theta^*$ as $n \to \infty$. Since the score function s is differentiable and convex in θ, equivalently, $\widehat{\theta}_n^{\mathrm{mod}} \in \Theta$ is the unique solution to

$$\widehat{L}_n^{\mathrm{mod}}(\theta) := \frac{1}{n} \sum_{i=1}^n \big[\widehat{\mu}_x(Z_i)g(Y_i) + f(X_i)\widehat{\mu}_y(Z_i) - \widehat{\mu}_x(Z_i)\widehat{\mu}_y(Z_i) + \nabla_\theta h(X_i, \theta)\big] = 0,$$

while the population parameter θ^* is the unique solution to

$$L(\theta) := \mathbb{E}\big[\nabla_\theta \ell(X_i, Y_i, \theta)\big] = \mathbb{E}\big[g(Y_i)f(X_i) + \nabla_\theta h(X_i, \theta)\big] = 0.$$

We also define the empirical score at any $\theta \in \Theta$ as

$$\widehat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \big[g(Y_i)f(X_i) + \nabla_\theta h(X_i, \theta)\big].$$

Then for any fixed $\theta \in \Theta$, we note that

$$\widehat{L}_n^{\mathrm{mod}}(\theta) - \widehat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \big[\widehat{\mu}_x(Z_i)g(Y_i) + f(X_i)\widehat{\mu}_y(Z_i) - \widehat{\mu}_x(Z_i)\widehat{\mu}_y(Z_i) - g(Y_i)f(X_i)\big]$$

$$= \frac{1}{n} \sum_{k=1}^2 \sum_{i \in \mathcal{I}_k} \big[\widehat{\mu}_x^{(k)}(Z_i)g(Y_i) + f(X_i)\widehat{\mu}_y^{(k)}(Z_i) - \widehat{\mu}_x^{(k)}(Z_i)\widehat{\mu}_y^{(k)}(Z_i) - g(Y_i)f(X_i)\big].$$

Similar to the proof of Theorem 2 (see equation (19)), under the given consistency condition for $\widehat{\mu}_x^{(k)}$ and $\widehat{\mu}_y^{(k)}$, we know $\sup_{\theta \in \Theta} \big|\widehat{L}_n^{\mathrm{mod}}(\theta) - \widehat{L}_n(\theta)\big| = o_P(1)$. Furthermore, the law of

large numbers implies $\frac{1}{n}\sum_{i=1}^n |\widehat{L}_n(\theta) - L(\theta)| = o_P(1)$ for any fixed $\theta \in \Theta$. Therefore, $\big|\widehat{L}_n^{\mathrm{mod}}(\theta) - L(\theta)\big| = o_P(1)$ for any fixed $\theta \in \Theta$.

Since Θ is compact and $\nabla_\theta h(x,\theta)$ is $L(x)$-Lipschitz in $\theta \in \Theta$ for any $x \in \mathcal{X}$, we know that for any $\theta, \theta' \in \Theta$, by the triangular inequality,

$$\big|\widehat{L}_n(\theta) - \widehat{L}_n(\theta')\big| \le \frac{1}{n}\sum_{i=1}^n L(X_i)\cdot\|\theta - \theta'\|,$$

where $\frac{1}{n}\sum_{i=1}^n L(X_i) = O_P(1)$ since $\mathbb{E}[L(X)] < \infty$. Thus, for any fixed $\epsilon > 0$, there exists a finite subset $\{\theta_1, \ldots, \theta_{N_\epsilon}\}$ of Θ such that $\sup_\theta \inf_{1\le i\le N_\epsilon} |\widehat{L}_n(\theta) - \widehat{L}_n(\theta_i)| \le \epsilon$. Similar arguments applied to $L(\theta)$, together with the convergence, implies $\sup_{\theta\in\Theta} |\widehat{L}_n(\theta) - L(\theta)| = o_P(1)$. Thus we have $\sup_{\theta\in\Theta} |\widehat{L}_n^{\mathrm{mod}}(\theta) - L(\theta)| = o_P(1)$. On the other hand, the compactness of Θ and the uniqueness of θ^* as a solution to $L(\theta) = 0$ implies the well-separatedness condition (c.f. Van der Vaart (2000, Theorem 5.9)): for any $\epsilon > 0$, one could find some $\delta > 0$ such that $\inf_{\theta:\ \|\theta-\theta^*\|\ge\delta} \big|L(\theta)\big| > 2\epsilon$. Since $\sup_{\theta\in\Theta} |\widehat{L}_n^{\mathrm{mod}}(\theta) - L(\theta)| = o_P(1)$, for any fixed $\epsilon > 0$, one could find some $\delta > 0$ such that for n sufficiently large, $\mathbb{P}\big(\inf_{\theta:\ \|\theta-\theta^*\|\ge\delta} \frac{1}{n}|\widehat{L}_n^{\mathrm{mod}}(\theta)| > \epsilon\big) \ge 1 - \epsilon$, i.e., $\mathbb{P}(\|\widehat{\theta}_n^{\mathrm{mod}} - \theta^*\| \le \delta) \ge 1 - \epsilon$. This proves $\widehat{\theta}_n^{\mathrm{mod}} - \theta^* = o_P(1)$.

We now proceed to show the asymptotic normality of $\widehat{\theta}_n^{\mathrm{mod}}$. Recall that $\widehat{L}_n^{\mathrm{mod}}(\widehat{\theta}_n^{\mathrm{mod}}) = 0$. Taylor expansion around θ^* then gives

$$0 = \widehat{L}_n^{\mathrm{mod}}(\theta^*) + \nabla_\theta \widehat{L}_n^{\mathrm{mod}}(\theta^*)(\widehat{\theta}_n^{\mathrm{mod}} - \theta^*) + 1/2\cdot(\widehat{\theta}_n^{\mathrm{mod}} - \theta^*)\nabla_\theta^2 \widehat{L}_n^{\mathrm{mod}}(\widetilde{\theta}_n)(\widehat{\theta}_n^{\mathrm{mod}} - \theta^*) \quad (17)$$

for some $\widetilde{\theta}_n$ that lies on the segment between $\widehat{\theta}_n^{\mathrm{mod}}$ and θ^*, which implies $\widetilde{\theta}_n = \theta^* + o_P(1)$. Here $\nabla_\theta \widehat{L}_n^{\mathrm{mod}} \in \mathbb{R}^{p_x\times p_x}$ and $\nabla_\theta^2 \widehat{L}_n^{\mathrm{mod}} \in \mathbb{R}^{p_x\times p_x\times p_x}$ are second and third order derivative of $\sum_{i=1}^n \widehat{\ell}(X_i, Y_i, Z_i, \theta)$. By definition

$$\nabla_\theta \widehat{L}_n^{\mathrm{mod}}(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 h(X_i,\theta), \quad \nabla_\theta^2 \widehat{L}_n^{\mathrm{mod}}(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta^3 h(X_i,\theta).$$

The law of large numbers imply $\nabla_\theta \widehat{L}_n^{\mathrm{mod}}(\theta^*) = \mathbb{E}[\nabla_\theta^2 h(X_i,\theta^*)] + o_P(1)$ where $o_P(1)$ means a matrix whose every entry is $o_P(1)$, hence $\nabla_\theta \widehat{L}_n^{\mathrm{mod}}(\theta^*)$ is invertible for sufficiently large n. Also, by the Lipschitz condition of $\nabla_\theta^3 h(x,\theta)$ and the triangular inequality, we have

$$\big|\nabla_\theta^2 \widehat{L}_n^{\mathrm{mod}}(\theta^*) - \nabla_\theta^2 \widehat{L}_n^{\mathrm{mod}}(\widetilde{\theta}_n)\big| \le \frac{1}{n}\sum_{i=1}^n m(X_i)\|\theta^* - \widetilde{\theta}_n\| = o_P(1)$$

since $\mathbb{E}[m(X_i)] < \infty$. Hence $\nabla_\theta^2 \widehat{L}_n^{\mathrm{mod}}(\widetilde{\theta}_n) = O_P(1)$. Returning to (17), we have

$$\begin{aligned} 0 &= \widehat{L}_n^{\mathrm{mod}}(\theta^*) + \big\{\mathbb{E}[\nabla_\theta^2 h(X_i,\theta^*)] + o_P(1)\big\}(\widehat{\theta}_n^{\mathrm{mod}} - \theta^*) + O_P\big(\|\widehat{\theta}_n^{\mathrm{mod}} - \theta^*\|^2\big) \\ &= \widehat{L}_n^{\mathrm{mod}}(\theta^*) + \big\{\mathbb{E}[\nabla_\theta^2 h(X_i,\theta^*)] + o_P(1)\big\}(\widehat{\theta}_n^{\mathrm{mod}} - \theta^*), \end{aligned}$$

and thus

$$\widehat{\theta}_n^{\mathrm{mod}} - \theta^* = \mathbb{E}\big[\nabla_\theta^2 h(X_i,\theta^*)\big]^{-1} \widehat{L}_n^{\mathrm{mod}}(\theta^*) + o_P\big(\widehat{L}_n^{\mathrm{mod}}(\theta^*)\big). \quad (18)$$

Finally, we note that under the consitency condition of $\widehat{\mu}_x^{(k)}$ and $\widehat{\mu}_y^{(k)}$, similar to the proof of Theorem 2, one could show that

$$\widehat{L}_n^{\mathrm{mod}}(\theta^*) = \frac{1}{n}\sum_{i=1}^{n}\big[\mu_x(Z_i)g(Y_i) + f(X_i)\mu_y(Z_i) - \mu_x(Z_i)\mu_y(Z_i) + \nabla_\theta h(X_i,\theta^*)\big] + o_P(1/\sqrt{n}).$$

Because $X \perp\!\!\!\perp Y \mid Z$, we have $\mathbb{E}\big[\mu_x(Z_i)g(Y_i) + f(X_i)\mu_y(Z_i) - \mu_x(Z_i)\mu_y(Z_i) + \nabla_\theta h(X_i,\theta^*)\big] = \mathbb{E}[\nabla_\theta s(X_i,Y_i,\theta^*)] = 0$, hence $\widehat{L}_n^{\mathrm{mod}}(\theta^*) = O_P(1/\sqrt{n})$, which, combined with (18), implies

$$\begin{aligned}
&\widehat{\theta}_n^{\mathrm{mod}} - \theta^* = \mathbb{E}\big[\nabla_\theta^2 h(X_i,\theta^*)\big]^{-1}\widehat{L}_n^{\mathrm{mod}}(\theta^*) + o_P(1/\sqrt{n}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\nabla_\theta^2 h(X_i,\theta^*)\big]^{-1}\big[\mu_x(Z_i)g(Y_i) + f(X_i)\mu_y(Z_i) - \mu_x(Z_i)\mu_y(Z_i) + \nabla_\theta h(X_i,\theta^*)\big] + o_P(1/\sqrt{n}).
\end{aligned}$$

We thus complete the proof of the asymptotic expansion of $\widehat{\theta}_n^{\mathrm{mod}}$ in Theorem 16.

Finally, the efficient influence function for estimating θ^* among all models obeying $X \perp\!\!\!\perp Y \mid Z$ can be obtained by, similar to Theorem 2, projecting the influence function of $\widehat{\theta}_n^{\mathrm{ml}}$ onto the tangent space $\mathcal{T} = \mathcal{T}_1 \oplus \mathcal{T}_2 \oplus \mathcal{T}_3$ where $\mathcal{T}_j$ is defined as in (20). Standard M-estimator theory (Van der Vaart, 2000) gives $\widehat{\theta}_n^{\mathrm{ml}} - \theta^* = \frac{1}{n}\sum_{i=1}^{n}\phi^{\mathrm{ml}}(X_i,Y_i) + o_P(1/\sqrt{n})$ where

$$\phi^{\mathrm{ml}}(x,y) = \mathbb{E}\big[\nabla_\theta^2 h(X_i,\theta^*)\big]^{-1}\big[g(Y_i)f(X_i) + \nabla_\theta h(X_i,\theta^*)\big].$$

The projection follows exactly the same idea as Theorem 16 and one could see ϕ is the efficient influence function. We thus complete the proof of Theorem 16. ■

Appendix B. Deferred discussion

B.1 Deferred discussion for Remark 6

Below, we provide a (simplified) finite-sample analysis of the two methods: (i) OLS, and (ii) modular OLS, where $Y \mid Z$ and $X \mid Z$ are estimated using OLS, under a joint Gaussian distribution. This discussion elaborates on the computation details for Remark 6.

To be more specific, the cross-term $C = \mathbb{E}[XY]$ is estimated via $\widehat{C}_{\mathrm{ols}} := \frac{1}{n}\sum_{i=1}^{n} X_iY_i$ or $\widehat{C}_{\mathrm{mod}} = \frac{1}{n}\sum_{i=1}^{n} Y_i\widehat{\mu}_x(Z_i) + X_i\widehat{\mu}_y(Z_i) - \widehat{\mu}_x(Z_i)\widehat{\mu}_y(Z_i)$. Hereafter, we override the notations and use X, Y, Z to denote data vectors in $\mathbb{R}^n$. With OLS regression, we have $\widehat{\mu}_x(z) = \widehat{\theta}_x^\top z$, where $\widehat{\theta}_x = (Z^\top Z)^{-1}Z^\top X$ is the fitted OLS coefficient of X on Z, and similarly $\widehat{\mu}_y(z) = \widehat{\theta}_y^\top z$, where $\widehat{\theta}_y = (Z^\top Z)^{-1}Z^\top Y$. Then, our modular estimator for the cross-term $\widehat{C}_{\mathrm{lm}}$ is

$$\begin{aligned}
\widehat{C}_{\mathrm{mod}} &= \frac{1}{n}\sum_{i=1}^{n} Y_iZ_i^\top\widehat{\theta}_x + X_iZ_i^\top\widehat{\theta}_y - Z_i^\top\widehat{\theta}_xZ_i^\top\widehat{\theta}_y \\
&= \frac{1}{n}\Big(Y^\top Z\widehat{\theta}_x + X^\top Z\widehat{\theta}_y - \widehat{\theta}_x^\top Z^\top Z\widehat{\theta}_y\Big) \\
&= \frac{1}{n}\Big(Y^\top Z(Z^\top Z)^{-1}Z^\top X + X^\top Z(Z^\top Z)^{-1}Z^\top Y - X^\top Z(Z^\top Z)^{-1}Z^\top Z(Z^\top Z)^{-1}Z^\top Y\Big)
\end{aligned}$$

$$= \frac{1}{n}\Big(Y^\top Z(Z^\top Z)^{-1}Z^\top X + X^\top Z(Z^\top Z)^{-1}Z^\top Y - X^\top Z(Z^\top Z)^{-1}Z^\top Y\Big)$$
$$= \frac{1}{n}Y^\top Z(Z^\top Z)^{-1}Z^\top X,$$

where the third row plugs in the definition of $\widehat{\theta}_x$ and $\widehat{\theta}_y$, and the last line uses the fact that $Y^\top Z(Z^\top Z)^{-1}Z^\top X = X^\top Z(Z^\top Z)^{-1}Z^\top Y$. The two estimators are $\widehat{\theta}_n^{\mathrm{ols}} = (X^\top X/n)^{-1}\widehat{C}_{\mathrm{ols}}$, and $\widehat{\theta}_n^{\mathrm{mod}} = (X^\top X/n)^{-1}\widehat{C}_{\mathrm{mod}}$, respectively.

Suppose X, Y, Z are all one-dimensional and jointly Gaussian. Then, there exists a decomposition $X = \alpha Z + \epsilon_x$ and $Y = \beta Z + \epsilon_y$, where conditional on Z, ϵ_x and ϵ_y are normal with mean zero and mutually independent under the assumption $X \perp\!\!\!\perp Y \,|\, Z$. We further assume $\epsilon_x \,|\, Z \sim N(0, \sigma_x^2)$ and $\epsilon_y \,|\, Z \sim N(0, \sigma_y^2)$. Then, the population OLS parameter is $\theta^* = \alpha\beta/(\alpha^2 + \sigma_x^2)$. Also,

$$\mathbb{E}[\widehat{\theta}_n^{\mathrm{mod}}] = \mathbb{E}\left[\mathbb{E}\left[\frac{Y^\top Z \cdot Z^\top X}{X^\top X \cdot Z^\top Z}\,\middle|\, X, Z\right]\right] = \beta\mathbb{E}\left[\frac{Z^\top Z \cdot Z^\top X}{X^\top X \cdot Z^\top Z}\right] = \beta\mathbb{E}\left[\mathbb{E}\left[\frac{Z^\top X}{X^\top X}\,\middle|\, X\right]\right],$$

where we repeatedly used the tower property. Using the joint Gaussianity of (X, Z), we know that $\mathbb{E}[Z \,|\, X] = \frac{\alpha}{\alpha^2+\sigma_x^2}X$, which gives $\mathbb{E}[\widehat{\theta}_n^{\mathrm{mod}}] = \theta^*$, hence $\widehat{\theta}_n^{\mathrm{mod}}$ is unbiased. Similarly one can show $\widehat{\theta}_n^{\mathrm{ols}}$ is unbiased.

We now study the variances of the two estimators. By the decomposition of variance,

$$\begin{aligned}\mathrm{Var}(\widehat{\theta}_n^{\mathrm{mod}}) &= \mathbb{E}[\mathrm{Var}(\widehat{\theta}_n^{\mathrm{mod}} \,|\, X, Z)] + \mathrm{Var}(\mathbb{E}[\widehat{\theta}_n^{\mathrm{mod}} \,|\, X, Z]) \\ &= \mathbb{E}\left[\frac{(Z^\top X)^2 \sum_{i=1}^n Z_i^2\sigma_y^2}{(X^\top X \cdot Z^\top Z)^2}\right] + \mathrm{Var}\left(\frac{\beta Z^\top X}{X^\top X}\right) \\ &= \sigma_y^2\mathbb{E}\left[\frac{(Z^\top X)^2}{(X^\top X)^2 \cdot Z^\top Z}\right] + \beta^2\,\mathrm{Var}\left(\frac{Z^\top X}{X^\top X}\right).\end{aligned}$$

Similarly, using the fact that $Y \perp\!\!\!\perp X \,|\, Z$,

$$\begin{aligned}\mathrm{Var}(\widehat{\theta}_n^{\mathrm{ols}}) &= \mathbb{E}[\mathrm{Var}(\widehat{\theta}_n^{\mathrm{ols}} \,|\, X, Z)] + \mathrm{Var}(\mathbb{E}[\widehat{\theta}_n^{\mathrm{ols}} \,|\, X, Z]) \\ &= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i^2\sigma_y^2}{(X^\top X)^2}\right] + \mathrm{Var}\left(\frac{\beta Z^\top X}{X^\top X}\right) = \sigma_y^2\mathbb{E}\left[\frac{1}{X^\top X}\right] + \beta^2\,\mathrm{Var}\left(\frac{Z^\top X}{X^\top X}\right).\end{aligned}$$

This gives (using $X = \alpha Z + \epsilon_x$)

$$\begin{aligned}\mathrm{Var}(\widehat{\theta}_n^{\mathrm{ols}}) - \mathrm{Var}(\widehat{\theta}_n^{\mathrm{mod}}) &= \sigma_y^2\mathbb{E}\left[\frac{Z^\top Z \cdot X^\top X - (Z^\top X)^2}{(X^\top X)^2 \cdot Z^\top Z}\right] \\ &= \sigma_y^2\mathbb{E}\left[\frac{Z^\top Z \cdot \epsilon_x^\top \epsilon_x - (Z^\top \epsilon_x)^2}{(X^\top X)^2 \cdot Z^\top Z}\right] \geq 0\end{aligned}$$

by the Cauchy-Schwarz inequality (it is a strict inequality as long as ϵ_x is not deterministic given Z). That is, in finite sample, both estimators are unbiased, while the modular estimator is non-inferior to OLS in terms of variance.

B.2 Deferred discussion on alternative estimators

In this part, we elaborate our discussion on the alternative estimators in Remark 3. In the following, we consider using powerful machine learning algorithms (black-boxes) to estimate the nuisance components μ_x and μ_y. They may achieve a smaller MSE than the OLS estimator, potentially at the cost of some bias. The semiparametric efficiency in Theorem 2 of our manuscript does not account for such cases. In the following, we conduct additional experiments to explore this regime.

Surprisingly, we found that (a) outcome regression can achieve a smaller MSE than the OLS, and that (b) orthogonal regression is often less accurate. Nonetheless, our modular estimator often achieves the smallest (or nearly smallest) MSE.

In Figure 12, we show results for OR (standing for outcome regression), PS (standing for orthogonal regression), OLS, and modular regression under several simple data-generating processes, where the true regression functions only involve ≤ 4 entries. We vary the signal-to-noise ratio and the smoothness of the underlying regression functions across settings. We label the first row as (a-c), and the second as (d-f). Settings (a) and (b) has $Z \in \mathbb{R}^{p_{20}}$, and the true regression functions are a combination of indicator functions and smooth exponential functions. Setting (c) involves a larger number of combinations of polynomial and exponential functions for $Z \in \mathbb{R}^{20}$. Settings (d) and (e) adds a term $\sqrt{n}Z_3Z_4$ to fixed regression functions that only involve the first 4 variables in $Z \in \mathbb{R}^{100}$. Setting (f) uses a similar model as (a-c) but kept $Z \in \mathbb{R}^{100}$. (d-f) has larger noise level. We omit details on the regression functions for brevity.

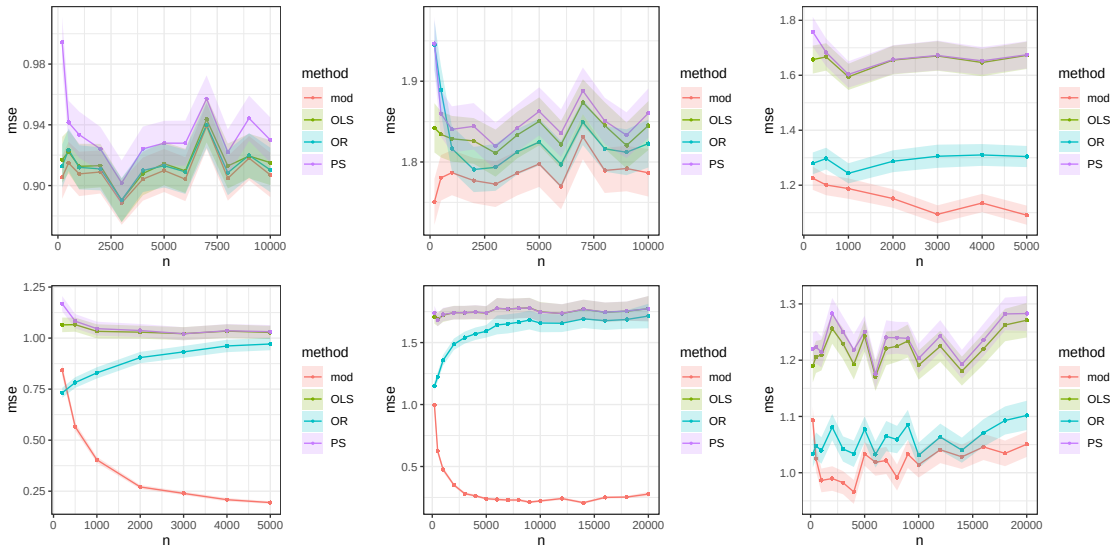


Figure 12: Root-mean-squared error scaled by $\sqrt{n}$, for all methods under different sample sizes.

We see that OR often performs better in terms of MSE; we conjecture that this is because the tuned ML regressor achieves a good bias-variance tradeoff. The PS approach is usually not very accurate, but surprisingly comparable to OLS. A potential explanation is that we are in the regime of "undersmoothing", i.e., the ML estimator has extremely small

bias but larger variance, so that it is still possible to achieve good performance for both OR and PS approaches. Finally, however, the modular estimator always performs the best across all settings; we conjecture that this is because it takes the best of the two worlds, powerful prediction machines and rigorous debiasing ideas from semiparametric statistics.

B.3 Discussion on the estimation of $\widehat{J}$

Here, we discuss the conditions for estimating $\widehat{J}$ and its compatibility with Assumption 10.

Consistent estimation of J require strong conditions such as the beta-min conditions or the irrepresentable conditions Zhao and Yu (2006). However, it is also known that Lasso selects a superset of "relevant" features, i.e., entries whose coefficient is sufficently bounded away from zero (roughly, these are coefficients with absolute values larger than $O(\sqrt{s_0/n \cdot \log p}$, where s_0 is the number of nonzero coefficients) under less restrictive eigenvalue conditions of the covariates X (see, e.g., Van De Geer and Bühlmann (2009); Bühlmann and Van De Geer (2011); Bühlmann (2010)). When the entries in J are sufficiently large, then a superset of J can be recovered with high probability. Otherwise, entries with small coefficients would not incur too much bias in this process.

On the other hand, consistency estimation of J typically asks for $\log p_z/\sqrt{n} \to 0$, while Assumption 10 asks for $\|\widehat{\mu}^{(k)}_{x,z}(\cdot) - \mu^{(k)}_{x,z}(\cdot)\|_{L_2(\mathbb{P}_Z)} \preceq n^{-1/4}$. When the regression function $\widehat{\mu}_{x,z}$ is obtained from high-dimensional regression such as the Lasso, the convergence rate is typically $\sqrt{s_{x,z} \log p_z/n}$, where $s_{x,z}$ is the sparsity level for an entry in X over Z. These conditions on p_z are typically compatible for the two parts (as they are both satisfied for relatively small p_z). For example, if the sparsity $s_{x,z}$ is bounded, and $\log p_z = o(n^{1/4})$, then both conditions are satisfied.

Appendix C. Technical proofs

C.1 Proof of Theorem 2

Proof [Proof of Theorem 2] Recall that θ^* is the least-squares estimator $\theta^* = \operatorname{argmin}_{\theta \in \mathbb{R}^{p_x}} \mathbb{E}[(Y - X^\top \theta)^2]$. The estimation equation gives

$$\widehat{\theta}^{\text{ols}}_n - \theta^* = \Big(\sum_{i=1}^n X_i X_i^\top\Big)^{-1} \sum_{i=1}^n X_i Y_i - \theta^* = \Big(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\Big)^{-1} \frac{1}{n} \sum_{i=1}^n X_i (Y_i - X_i^\top \theta^*).$$

Since $X_i(Y_i - X_i^\top \theta^*)$ has finite second moment and we know $\mathbb{E}[X(Y - X^\top \theta^*)] = 0$ by the optimality of θ^*, we have $\frac{1}{n}\sum_{i=1}^n X_i(Y_i - X_i^\top \theta^*) = O_P(1/\sqrt{n})$, where $O_P(1/\sqrt{n})$ represents a random vector whose each entry is $O_P(1/\sqrt{n})$. Since $X_i X_i^\top$ has finite expectation $\mathbb{E}[XX^\top] \succ 0$, we have

$$\Big(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\Big)^{-1} = \Big(\mathbb{E}[XX^\top] + o_P(1)\Big)^{-1} = \big(\mathbb{E}[XX^\top]\big)^{-1} + o_P(1),$$

where $o_P(1)$ stands for a random matrix whose all entries converge in probability to zero. We thus have

$$\widehat{\theta}^{\text{ols}}_n - \theta^* = \Big(\big(\mathbb{E}[XX^\top]\big)^{-1} + o_P(1)\Big) \frac{1}{n} \sum_{i=1}^n X_i (Y_i - X_i^\top \theta^*)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \big(\mathbb{E}[XX^\top]\big)^{-1} X_i(Y_i - X_i^\top\theta^*) + o_P(1) \times O_P(1/\sqrt{n})$$
$$= \frac{1}{n}\sum_{i=1}^{n} \phi^{\text{ols}}(X_i, Y_i) + o_P(1/\sqrt{n}),$$

where $\phi^{\text{ols}}(x, y) = \mathbb{E}[XX^\top]^{-1}x(y - x^\top\theta^*)$. For our modular estimator, we first note the closed form solution

$$\widehat{\theta}_n^{\text{mod}} - \theta^* = \Big(\sum_{i=1}^{n} X_i X_i^\top\Big)^{-1} \sum_{i=1}^{n} C_i - \theta^* = \Big(\sum_{i=1}^{n} X_i X_i^\top\Big)^{-1} \sum_{i=1}^{n} \big(C_i - X_i X_i^\top\theta^*\big).$$

We now show that under the conditions in Theorem 2, one has

$$\frac{1}{n}\sum_{i=1}^{n} \big(C_i - X_i X_i^\top\theta^*\big) = \frac{1}{n}\sum_{i=1}^{n} \big(C_i^* - X_i X_i^\top\theta^*\big) + o_P(1/\sqrt{n}), \tag{19}$$

where $C_i^* = X_i\mu_y(Z_i) + \mu_x(Z_i)Y_i - \mu_x(Z_i)\mu_y(Z_i)$. To see this, by rearranging the terms, we have $\frac{1}{n}\sum_{i=1}^{n}(C_i - C_i^*) = \text{(i)} + \text{(ii)} + \text{(iii)}$, where we define

$$\text{(i)} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k} \{X_i - \mu_x(Z_i)\}\{\widehat{\mu}_y^{(k)}(Z_i) - \mu_y(Z_i)\},$$
$$\text{(ii)} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k} \{\widehat{\mu}_x^{(k)}(Z_i) - \mu_x(Z_i)\}\{Y_i - \mu_y(Z_i)\},$$
$$\text{(iii)} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k} \{\widehat{\mu}_x^{(k)}(Z_i) - \mu_x(Z_i)\}\{\widehat{\mu}_y^{(k)}(Z_i) - \mu_y(Z_i)\}.$$

We first bound the summation in (i). For each k, because $\widehat{\mu}_y^{(k)}$ is obtained from the independent fold $\mathcal{I}\backslash\mathcal{I}_k$, for any $i \in \mathcal{I}_k$, by the tower property,

$$\mathbb{E}\big[\{X_i - \mu_x(Z_i)\}\{\widehat{\mu}_y^{(k)}(Z_i) - \mu_y(Z_i)\} \,\big|\, \mathcal{I}\backslash\mathcal{I}_k\big] = \mathbb{E}\big[\mathbb{E}[X_i - \mu_x(Z_i) \,|\, Z_i]\{\widehat{\mu}_y^{(k)}(Z_i) - \mu_y(Z_i)\} \,\big|\, \mathcal{I}\backslash\mathcal{I}_k\big] = 0,$$

and they are i.i.d. copies conditional on $\mathcal{I}\backslash\mathcal{I}_k$ with conditional variance

$$\mathbb{E}\big[\{X_i - \mu_x(Z_i)\}^2\{\widehat{\mu}_y^{(k)}(Z_i) - \mu_y(Z_i)\}^2 \,\big|\, \mathcal{I}\backslash\mathcal{I}_k\big] = o_P(1).$$

The Markov's inequality thus implies

$$\frac{1}{n}\sum_{i\in\mathcal{I}_k} \{X_i - \mu_x(Z_i)\}\{\widehat{\mu}_y^{(k)}(Z_i) - \mu_y(Z_i)\} = o_P(1/\sqrt{n}),$$

hence $\text{(i)} = o_P(1/\sqrt{n})$. Similar arguments also apply to (ii) and yield $\text{(ii)} = o_P(1/\sqrt{n})$. Finally, conditional on $\mathcal{I}\backslash\mathcal{I}_k$, for any $i \in \mathcal{I}_k$, by the Cauchy-Schwarz inequality,

$$\Big|\sum_{i\in\mathcal{I}_k} \{\widehat{\mu}_x^{(k)}(Z_i) - \mu_x(Z_i)\}\{\widehat{\mu}_y^{(k)}(Z_i) - \mu_y(Z_i)\}\Big|$$

$$\leq \Big[\sum_{i\in\mathcal{I}_k}\{\widehat{\mu}_x^{(k)}(Z_i)-\mu_x(Z_i)\}^2\Big]^{1/2}\cdot\Big[\sum_{i\in\mathcal{I}_k}\{\widehat{\mu}_y^{(k)}(Z_i)-\mu_y(Z_i)\}^2\Big]^{1/2}.$$

By the Markov's inequality, we have

$$\begin{aligned}\sum_{i\in\mathcal{I}_k}\{\widehat{\mu}_x^{(k)}(Z_i)-\mu_x(Z_i)\}^2 &= O_P\big(|\mathcal{I}_k|\cdot\mathbb{E}\big[\{\widehat{\mu}_x^{(k)}(Z_i)-\mu_x(Z_i)\}^2\,\big|\,\mathcal{I}\backslash\mathcal{I}_k\big]\big)\\ &= O_P\big(n\cdot\|\widehat{\mu}_x^{(k)}-\mu_x\|^2_{L_2(\mathbb{P}_Z)}\big).\end{aligned}$$

With similar arguments applied to the second summation, we arrive at

$$\Big|\sum_{i\in\mathcal{I}_k}\{\widehat{\mu}_x^{(k)}(Z_i)-\mu_x(Z_i)\}\{\widehat{\mu}_y^{(k)}(Z_i)-\mu_y(Z_i)\}\Big| \leq O_P\big(n\cdot\|\widehat{\mu}_x^{(k)}-\mu_x\|_{L_2(\mathbb{P}_Z)}\|\widehat{\mu}_y^{(k)}-\mu_y\|_{L_2(\mathbb{P}_Z)}\big).$$

Combining $k=1,2$, we have

$$\text{(iii)} = O_P\big(\|\widehat{\mu}_x^{(k)}-\mu_x\|_{L_2(\mathbb{P}_Z)}\|\widehat{\mu}_y^{(k)}-\mu_y\|_{L_2(\mathbb{P}_Z)}\big) = o_P(1/\sqrt{n})$$

by the conditions in Theorem 2. Putting together our bounds on the three terms, we prove the claim in (19). Note that $\mathbb{E}[C_i^*-X_iX_i^\top\theta^*]=0$, hence $\frac{1}{n}\sum_{i=1}^n(C_i^*-X_iX_i^\top\theta^*)=O_P(1/\sqrt{n})$, thus

$$\begin{aligned}\widehat{\theta}_n^{\text{mod}}-\theta^* &= \Big[\frac{1}{n}\sum_{i=1}^n X_iX_i^\top\Big]^{-1}\cdot\frac{1}{n}\sum_{i=1}^n\big(C_i-X_iX_i^\top\theta^*\big)\\ &= \big[\mathbb{E}[XX^\top]^{-1}+o_P(1)\big]\cdot\Big[\frac{1}{n}\sum_{i=1}^n\big(C_i^*-X_iX_i^\top\theta^*\big)+o_P(1/\sqrt{n})\Big]\\ &= \frac{1}{n}\sum_{i=1}^n\mathbb{E}[XX^\top]^{-1}\big(C_i^*-X_iX_i^\top\theta^*\big)+o_P(1/\sqrt{n}).\end{aligned}$$

That is, we obtain the asymptotic linear expansion $\widehat{\theta}_n^{\text{mod}}-\theta^*=\frac{1}{n}\sum_{i=1}^n\phi^{\text{mod}}(X_i,Y_i,Z_i)+o_P(1/\sqrt{n})$, where the influence function is given by

$$\phi^{\text{mod}}(x,y,z)=\mathbb{E}[XX^\top]^{-1}\big(x\mu_y(z)+\mu_x(z)y-\mu_x(z)\mu_y(z)-xx^\top\theta^*\big).$$

The Central Limit Theorem thus gives the asymptotic distribution $\sqrt{n}(\widehat{\theta}_n^{\text{mod}}-\theta^*)\xrightarrow{d}N(0,\Sigma^{\text{mod}})$ where $\Sigma^{\text{mod}}=\text{Cov}(\phi^{\text{mod}}(X_i,Y_i,Z_i))$.

We now proceed to show that $\phi^{\text{mod}}(X_i,Y_i,Z_i)$ is the efficient influence function for estimating θ^* under the current distribution $\mathbb{P}$. Our argument is similar to that of (Rotnitzky and Smucler, 2019, Lemma 9) and (Tsiatis, 2006, Theorem 4.5). We let $\mathcal{P}$ denote the collection of all distributions obeying Assumption 1, so $\mathbb{P}\in\mathcal{P}$. For any $P\in\mathcal{P}$, let p denote the density of P with respect to the base measure μ. Then the joint density $p(x,y,z)$ decomposes as $p(x,y,z)=p(z)p(x\,|\,z)p(y\,|\,z)$, where $p(z)$ is the marginal density of Z, $p(x\,|\,z)$ is the conditional density of $P_{X\,|\,Z}$, and $p(y\,|\,z)$ is the conditional density of $P_{Y\,|\,Z}$. By (Van der Laan and Robins, 2003, Lemma 1.6), the tangent space of $\mathcal{P}$ at P is given by $\mathcal{T}=\mathcal{T}_1\oplus\mathcal{T}_2\oplus\mathcal{T}_3$, where $\{\mathcal{T}_j\}_{j=1}^3$ are orthogonal spaces. More specifically, $\mathcal{T}_1$ is the closed linear span of socres

of one-dimensional regular parametric submodel $\gamma \mapsto p(z;\gamma)p(x \mid z)p(y \mid z)$, $\mathcal{T}_2$ is that of the parametric submodel $\gamma \mapsto p(x \mid z;\gamma)p(z)p(y \mid z)$, and $\mathcal{T}_3$ is that of the parametric submodel $\gamma \mapsto p(y \mid z;\gamma)p(z)p(x \mid z)$. Similar to (Tsiatis, 2006, Theorem 4.5), these subspaces can be equivalently represented as

$$\begin{aligned}
\mathcal{T}_1 &= \big\{ f(Z) \mapsto \mathbb{R}^{p_x} : \mathbb{E}[f(Z)] = 0 \big\}, \\
\mathcal{T}_2 &= \big\{ f(X,Z) \mapsto \mathbb{R}^{p_x} : \mathbb{E}[f(X,Z) \mid Z] = 0 \big\} \\
\mathcal{T}_3 &= \big\{ f(Y,Z) \mapsto \mathbb{R}^{p_x} : \mathbb{E}[f(Y,Z) \mid Z] = 0 \big\},
\end{aligned} \tag{20}$$

where all the functions in them are additionally square integrable.

By standard semiparametric theory (Tsiatis, 2006), the efficient influence function, denoted as ϕ^*, can be obtained by the projection of ϕ^{ols} onto the tangent space $\mathcal{T}$. That is, $\phi^*(x,y,z) = \phi_1^*(z) + \phi_2^*(x,z) + \phi_3^*(y,z)$, where ϕ_j^* is the projection of ϕ^{ols} onto $\mathcal{T}_j$ defined in (20). Hence

$$\begin{aligned}
\phi^*(X,Y,Z) &= \mathbb{E}\big[\phi^{\text{ols}}(X,Y) \,\big|\, Z\big] - \mathbb{E}\big[\phi^{\text{ols}}(X,Y)\big] + \mathbb{E}\big[\phi^{\text{ols}}(X,Y) \,\big|\, Y,Z\big] \\
&\qquad - \mathbb{E}\big[\phi^{\text{ols}}(X,Y) \,\big|\, Z\big] + \mathbb{E}\big[\phi^{\text{ols}}(X,Y) \,\big|\, X,Z\big] - \mathbb{E}\big[\phi^{\text{ols}}(X,Y) \,\big|\, Z\big] \\
&= \mathbb{E}[XX^\top]^{-1}\big\{ \mathbb{E}[X(Y - X^\top\theta^*) \mid X,Z] + \mathbb{E}[X(Y - X^\top\theta^*) \mid Y,Z] \\
&\qquad\qquad - \mathbb{E}[X(Y - X^\top\theta^*) \mid Z] - \mathbb{E}[X(Y - X^\top\theta^*)] \big\} \\
&= \mathbb{E}[XX^\top]^{-1}\big\{ X\mathbb{E}[Y \mid Z] - XX^\top\theta^* + \mathbb{E}[X \mid Z]Y \\
&\qquad\qquad - \mathbb{E}[XX^\top\theta^* \mid Z] - \mathbb{E}[XY \mid Z] + \mathbb{E}[XX^\top\theta^* \mid Z] \big\} \\
&= \mathbb{E}[XX^\top]^{-1}\big\{ X\mathbb{E}[Y \mid Z] - XX^\top\theta^* + \mathbb{E}[X \mid Z]Y - \mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z] \big\} \\
&= \phi^{\text{mod}}(X,Y,Z),
\end{aligned}$$

where the first and second equalities are from the projection onto subsapces, the third equality uses the fact that $\mathbb{E}[X(Y - X^\top\theta^*)] = 0$ and the tower property, and the fourth equality uses the conditional independence in Assumption 1. We thus conclude the proof of Theorem 2. ■

C.2 Proof of Theorem 11

Proof [Proof of Theorem 11] For notational simplicity, throughout the proof we write $\widehat{\theta}_n^{\text{mod}}$ as $\widehat{\theta}$. Denote $\Delta = \widehat{\theta} - \theta^*$, and let $X \in \mathbb{R}^{n \times p_x}$ be the design matrix. We define the vector $G = \sum_{i=1}^n C_i \in \mathbb{R}^{p_x}$, where we recall the definition of C_i in (6). Since $\widehat{\theta}$ is the minimizer of (8), we have

$$\frac{1}{2n}\widehat{\theta}^\top X^\top X\widehat{\theta} - \frac{1}{n}G^\top\widehat{\theta} + \lambda\|\widehat{\theta}\|_1 \le \frac{1}{2n}(\theta^*)^\top X^\top X\theta^* - \frac{1}{n}G^\top\theta^* + \lambda\|\theta^*\|_1,$$

which implies

$$\frac{1}{2n}\Delta^\top X^\top X\Delta \le \frac{1}{n}\big\langle (\theta^*)^\top X^\top X - G, \Delta \big\rangle + \lambda\|\theta^*\|_1 - \lambda\|\widehat{\theta}\|_1.$$

Let S be the set of indices for nonzero entries of θ^*, and θ_S be the subvector containing all entries with indices in S. Then

$$\|\theta^*\|_1 - \|\widehat{\theta}\|_1 = \|\theta^*_S\|_1 - \|\widehat{\theta}_S\|_1 - \|\widehat{\theta}_{S^c}\|_1 \le \|\Delta_S\|_1 - \|\Delta_{S^c}\|_1,$$

where the last inequality uses triangular inequality and the fact that $\theta^*_{S^c} \equiv 0$. We thus have

$$\begin{aligned}\frac{1}{2n}\Delta^\top X^\top X\Delta &\le \frac{1}{n}\big\langle (\theta^*)^\top X^\top X - G, \Delta\big\rangle + \lambda\|\Delta_S\|_1 - \lambda\|\Delta_{S^c}\|_1 \\ &\le \frac{1}{n}\|\Delta\|_1\big\|(\theta^*)^\top X^\top X - G\big\|_\infty + \lambda\|\Delta_S\|_1 - \lambda\|\Delta_{S^c}\|_1. \end{aligned} \tag{21}$$

The first implication is that, as the left-handed side is non-negative, once $\lambda \ge \frac{2}{n}\|(\theta^*)^\top X^\top X - G\|_\infty$, we have $0 \le \frac{\lambda}{2}\|\Delta\|_1 + \lambda\|\Delta_S\|_1 - \lambda\|\Delta_{S^c}\|_1$, hence $\Delta \in \mathbb{C}_3 = \big\{x \in \mathbb{R}^p\colon \|x_{S^c}\|_1 \le 3\|x_S\|_1\big\}$. By Assumption 8, (21) further implies

$$\zeta\|\Delta\|_2^2 \le \frac{\lambda}{2}\|\Delta\|_1 + \lambda\|\Delta_S\|_1 - \lambda\|\Delta_{S^c}\|_1 \le \frac{3\lambda}{2}\|\Delta_S\|_1 \le \frac{3\lambda}{2}\sqrt{k}\|\Delta\|_2,$$

where $k = |S| = \|\theta^*\|_0$ is the sparsity level. To summarize, under Assumption 8, the solution $\widehat{\theta}$ satisfies

$$\|\widehat{\theta} - \theta^*\|_2 \le \frac{3\lambda\sqrt{k}}{2\zeta} \tag{22}$$

for any regularization parameter $\lambda \ge \frac{2}{n}\|(\theta^*)^\top X^\top X - G\|_\infty$. We now write $\widehat{\mu}_y \in \mathbb{R}^n$ as the vector whose i-th entry is $\widehat{\mu}_y^{(k)}(Z_i)$ for $i \in \mathcal{I}_k$, and $\widehat{\mu}_x \in \mathbb{R}^{n\times p_x}$ whose (i,j)-th entry is $\widehat{\mu}_{x,j}^{(k)}(Z_i)$ for $i \in \mathcal{I}_k$. Similarly, $\mu_y \in \mathbb{R}^n$ and $\mu_x \in \mathbb{R}^{n\times p_x}$ record the ground truth of these regression functions. The error term is then

$$\begin{aligned}(\theta^*)^\top X^\top X - G &= (\theta^*)^\top X^\top X - \widehat{\mu}_y^\top X - Y^\top\widehat{\mu}_x + \widehat{\mu}_y^\top\widehat{\mu}_x \\ &= (Y-\mu_y)^\top(X-\mu_x) + (X\theta^* - Y)^\top X \\ &\quad + (\mu_y - \widehat{\mu}_y)^\top(X-\mu_x) + (Y-\mu_y)^\top(\mu_x - \widehat{\mu}_x) + (\mu_y-\widehat{\mu}_y)^\top(\mu_x-\widehat{\mu}_x).\end{aligned} \tag{23}$$

For any constant $\delta \in (0,1)$, we define the event

$$\mathcal{E}^{n,\delta}_{\text{est}} = \left\{\big\|\widehat{\mu}_{x,j}^{(k)} - \mu_{x,j}\big\|_{L_2(\mathbb{P}_Z)},\ \big\|\widehat{\mu}_y^{(k)} - \mu_y\big\|_{L_2(\mathbb{P}_Z)} \le \frac{4c_n\log(3p_x/\delta)}{n^{1/4}},\ \forall 1\le j\le p_x,\ \forall k = 1,2\right\}. \tag{24}$$

Then under Assumption 10, for any constant $\delta \in (0,1)$, taking a union bound over $2p_x+2 \le 3p_x$ estimated functions for $1 \le j \le p_x$ and $k = 1,2$, we know that $\mathbb{P}(\mathcal{E}^{n,\delta}_{\text{est}}) \ge 1-\delta$.

We now proceed to analyze each entry $j \in \{1,\dots,p_x\}$ of the above error term. First,

$$\frac{1}{n}\big[(\mu_y - \widehat{\mu}_y)^\top(X-\mu_x)\big]_j = \frac{1}{n}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\big(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i)\big)\big(X_{i,j} - \mu_{x,j}(Z_i)\big)$$

For each $k = 1, 2$, conditional on $\mathcal{I}\backslash\mathcal{I}_k$, the terms in the summation $(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i))(X_{i,j} - \mu_{x,j}(Z_i))$, $i \in \mathcal{I}_k$ are mutually independent with mean zero, since

$$\begin{aligned}&\mathbb{E}\Big[\big(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i)\big)\big(X_{i,j} - \mu_{x,j}(Z_i)\big)\,\Big|\,\mathcal{I}\backslash\mathcal{I}_k\Big]\\ &= \mathbb{E}\Big[\big(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i)\big)\cdot\mathbb{E}\big[X_{i,j} - \mu_{x,j}(Z_i)\,\big|\,Z_i, \mathcal{I}\backslash\mathcal{I}_k\big]\,\Big|\,\mathcal{I}\backslash\mathcal{I}_k\Big] = 0\end{aligned}$$

by the tower property. Also, the absolute value of each term is bounded below $4c_0$ by the boundedness in Assumption 9. We now define the event

$$\mathcal{E}_{\mathrm{cr}}^{1,k}(\delta) = \left\{\Big|\sum_{i\in\mathcal{I}_k}\big(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i)\big)\big(X_{i,j} - \mu_{x,j}(Z_i)\big)\Big| \le \max\Big\{16c_0\log(2/\delta)/3,\ 2\widehat{\epsilon}_1\sqrt{|\mathcal{I}_k|\log(2/\delta)}\Big\}\right\},$$

where we define

$$\widehat{\epsilon}_1^2 = 4\sup_{k=1,2}\big\|\mu_y - \widehat{\mu}_y^{(k)}\big\|_{L_2(\mathbb{P}_Z)}^2 \ge \mathbb{E}\Big[\big(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i)\big)^2\big(X_{i,j} - \mu_{x,j}(Z_i)\big)^2\,\Big|\,\mathcal{I}\backslash\mathcal{I}_k\Big].$$

The Bernstein's inequality in Lemma 17, implies $\mathbb{P}\big(\mathcal{E}_{\mathrm{cr}}^{1,k}\,\big|\,\mathcal{I}\backslash\mathcal{I}_k\big) \le \delta$. Marginalizing out the conditional probability and taking a union bound over $k = 1, 2$, we know that $\mathbb{P}(\mathcal{E}_{\mathrm{cr}}^1(\delta)) \ge 1 - \delta$ for any constant δ, where we define $\mathcal{E}_{\mathrm{cr}}^1(\delta) = \mathcal{E}_{\mathrm{cr}}^{1,1}(\delta/2) \cup \mathcal{E}_{\mathrm{cr}}^{1,2}(\delta/2)$. That is, with probability at least $1 - \delta$,

$$\Big|\frac{1}{n}\big[(\mu_y - \widehat{\mu}_y)^\top(X - \mu_x)\big]_j\Big| \le \max\left\{\frac{16c_0\log(4/\delta)}{3n},\ \widehat{\epsilon}_1\sqrt{\frac{2\log(4/\delta)}{n}}\right\}. \tag{25}$$

For each fixed j, applying exactly the same arguments to

$$\frac{1}{n}\big[(Y - \mu_y)^\top(\mu_x - \widehat{\mu}_x)\big]_j = \frac{1}{n}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\big(Y_i - \mu_y(Z_i)\big)\big(\mu_{x,j}(Z_i) - \widehat{\mu}_{x,j}^{(k)}(Z_i)\big),$$

with probability at least $1 - \delta$,

$$\Big|\frac{1}{n}\big[(Y - \mu_y)^\top(\mu_x - \widehat{\mu}_x)\big]_j\Big| \le \max\left\{\frac{16c_0\log(4/\delta)}{3n},\ c_0\widehat{\epsilon}_{2,j}\sqrt{\frac{2\log(4/\delta)}{n}}\right\}, \tag{26}$$

where we define $\widehat{\epsilon}_{2,j} = 2\sup_{k=1,2}\big\|\mu_{x,j} - \widehat{\mu}_{x,j}^{(k)}\big\|_{L_2(\mathbb{P}_Z)}$. The third term is

$$\frac{1}{n}\big[(\mu_y - \widehat{\mu}_y)^\top(\mu_x - \widehat{\mu}_x)\big]_j = \frac{1}{n}\sum_{k=1}^{2}\sum_{i\in\mathcal{I}_k}\big(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i)\big)\big(\mu_{x,j}(Z_i) - \widehat{\mu}_{x,j}^{(k)}(Z_i)\big).$$

Here for each $k = 1, 2$, conditional on $\mathcal{I}\backslash\mathcal{I}_k$, each term $(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i))(\mu_{x,j}(Z_i) - \widehat{\mu}_{x,j}^{(k)}(Z_i))$, $i \in \mathcal{I}_k$ in the above summation is i.i.d. whose expectation is bounded as

$$\Big|\mathbb{E}\Big[\big(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i)\big)\big(\mu_{x,j}(Z_i) - \widehat{\mu}_{x,j}^{(k)}(Z_i)\big)\,\Big|\,\mathcal{I}\backslash\mathcal{I}_k\Big]\Big|$$

$$\leq \big\|\mu_y - \widehat{\mu}_y^{(k)}\big\|_{L_2(\mathbb{P}_Z)} \big\|\mu_{x,j} - \widehat{\mu}_{x,j}^{(k)}\big\|_{L_2(\mathbb{P}_Z)} \leq \widehat{\epsilon}_1 \cdot \widehat{\epsilon}_{2,j}/4.$$

Meanwhile, their second moments are bounded as

$$\mathbb{E}\Big[\big(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i)\big)^2 \big(\mu_{x,j}(Z_i) - \widehat{\mu}_{x,j}^{(k)}(Z_i)\big)^2 \,\Big|\, \mathcal{I}\backslash\mathcal{I}_k\Big] \leq 4\big\|\mu_y - \widehat{\mu}_y^{(k)}\big\|_{L_2(\mathbb{P}_Z)}^2 \leq \widehat{\epsilon}_1^2$$

according to the bounededness in Assumption 9. Combining the above bounds and invoking the Bernstein's inequality in Lemma 17, it holds with probability at least $1-\delta$ that

$$\begin{aligned}
&\Big|\sum_{i\in\mathcal{I}_k} \big(\mu_y(Z_i) - \widehat{\mu}_y^{(k)}(Z_i)\big)\big(\mu_{x,j}(Z_i) - \widehat{\mu}_{x,j}^{(k)}(Z_i)\big)\Big| \\
&\leq |\mathcal{I}_k|\widehat{\epsilon}_1\widehat{\epsilon}_{2,j}/4 + \max\Big\{16c_0\log(2/\delta)/3,\ 2\widehat{\epsilon}_1\sqrt{|\mathcal{I}_k|\log(2/\delta)}\Big\}.
\end{aligned}$$

Taking a union bound for $k=1,2$ implies that with probability at least $1-\delta$,

$$\Big|\frac{1}{n}\big[(\mu_y - \widehat{\mu}_y)^\top(\mu_x - \widehat{\mu}_x)\big]_j\Big| \leq \frac{\widehat{\epsilon}_1\widehat{\epsilon}_{2,j}}{4} + \max\left\{\frac{16c_0\log(4/\delta)}{3n},\ \widehat{\epsilon}_1\sqrt{\frac{2\log(4/\delta)}{n}}\right\}. \tag{27}$$

Putting together (25), (26) and (27), we know that with probability at least $1-\delta/2$,

$$\begin{aligned}
&\big\|(\mu_y - \widehat{\mu}_y)^\top(X-\mu_x) + (Y-\mu_y)^\top(\mu_x - \widehat{\mu}_x) + (\mu_y - \widehat{\mu}_y)^\top(\mu_x - \widehat{\mu}_x)\big\|_\infty \\
&\leq \frac{\widehat{\epsilon}_1\widehat{\epsilon}_{2,j}}{4} + 2\max\left\{\frac{16c_0\log(24/\delta)}{3n},\ \widehat{\epsilon}_1\sqrt{\frac{2\log(24/\delta)}{n}}\right\} \\
&\qquad + \max\left\{\frac{16c_0\log(24/\delta)}{3n},\ c_0\widehat{\epsilon}_{2,j}\sqrt{\frac{2\log(24/\delta)}{n}}\right\}.
\end{aligned}$$

Further taking a union bound over the above event and (24) for $\mathcal{E}_{\text{est}}^{n,\delta/2}$, and using the fact that $\max\{a,b\} \leq a+b$ for $a,b\geq 0$, we know it holds with probability at least $1-\delta$ that

$$\begin{aligned}
&\big\|(\mu_y - \widehat{\mu}_y)^\top(X-\mu_x) + (Y-\mu_y)^\top(\mu_x - \widehat{\mu}_x) + (\mu_y - \widehat{\mu}_y)^\top(\mu_x - \widehat{\mu}_x)\big\|_\infty \\
&\leq \frac{4c_n^2(\log(3p_x/\delta))^2}{\sqrt{n}} + \frac{16c_0\log(24/\delta)}{n} + \frac{(2+c_0)\log(3p_x/\delta)\sqrt{2\log(24/\delta)}}{n^{3/4}}.
\end{aligned}$$

Recalling (23) and letting

$$\bar{c}_n = 4c_n^2 + 16c_0/\sqrt{n} + (2+c_0)/n^{1/4},$$

we have $\bar{c}_n \to 0$ as $n\to\infty$ and with probability at least $1-\delta$,

$$\frac{1}{n}\big\|(\theta^*)^\top X^\top X - G\big\|_\infty \leq \frac{1}{n}\big\|(Y-\mu_y)^\top(X-\mu_x) + (X\theta^* - Y)^\top X\big\|_\infty + \frac{\bar{c}_n(\log(3p_x/\delta))^2}{\sqrt{n}}.$$

Combining this bound with (22) and writing $D := (Y-\mu_y)^\top(X-\mu_x) + (X\theta^*-Y)^\top X$,

$$\begin{aligned}
&\mathbb{P}\left(\|\widehat{\theta} - \theta^*\|_2 \leq \frac{3\lambda\sqrt{k}}{2\zeta},\ \forall\ \lambda \geq \frac{2\|D\|_\infty}{n} + \frac{2\bar{c}_n(\log(3p_x/\delta))^2}{\sqrt{n}}\right) \\
&\geq \mathbb{P}\left(\frac{\|(\theta^*)^\top X^\top X - G\|_\infty}{n} \leq \frac{\|D\|_\infty}{n} + \frac{\bar{c}_n(\log(3p_x/\delta))^2}{\sqrt{n}}\right) \geq 1-\delta,
\end{aligned}$$

which proves (10). ■

C.3 Proof of Proposition 14

Proof [Proof of Proposition 14] Similar to the proof of Theorem 2, we are to show that $\widehat{C}_{\text{miss}} = C_{\text{miss}} + o_P(1/\sqrt{n_{xz}+n_{xyz}}) + o_P(1/\sqrt{n_{yz}+n_{xyz}})$, where $C_{\text{miss}} = C^{xz}_{\text{miss}} + C^{yz}_{\text{miss}} - C^{zz}_{\text{miss}}$,

$$
\begin{aligned}
C^{xz}_{\text{miss}} &= \frac{1}{n_{xz}+n_{xyz}} \sum_{i \in \mathcal{I}^{xz} \cup \mathcal{I}^{xyz}} X_i \mu_y(Z_i), \\
C^{yz}_{\text{miss}} &= \frac{1}{n_{yz}+n_{xyz}} \sum_{i \in \mathcal{I}^{yz} \cup \mathcal{I}^{xyz}} Y_i \mu_x(Z_i), \\
C^{zz}_{\text{miss}} &= \frac{1}{n_{xz}+n_{yz}+n_{xyz}} \sum_{i \in \mathcal{I}^{xz} \cup \mathcal{I}^{yz} \cup \mathcal{I}^{xyz}} \mu_y(Z_i)\mu_x(Z_i).
\end{aligned}
$$

To see this, we note that $\widehat{C}_{\text{miss}} - C_{\text{miss}} = \text{(i)} + \text{(ii)} + \text{(iii)}$, where

$$
\begin{aligned}
\text{(i)} &= \frac{1}{n_{xz}+n_{xyz}} \sum_{k=1}^{2} \sum_{i \in \mathcal{I}^{xz}_k \cup \mathcal{I}^{xyz}_k} \{X_i - \mu_x(Z_i)\}\{\widehat{\mu}^{(k)}_y(Z_i) - \mu_y(Z_i)\}, \\
\text{(ii)} &= \frac{1}{n_{yz}+n_{xyz}} \sum_{k=1}^{2} \sum_{i \in \mathcal{I}^{yz}_k \cup \mathcal{I}^{xyz}_k} \{Y_i - \mu_y(Z_i)\}\{\widehat{\mu}^{(k)}_x(Z_i) - \mu_x(Z_i)\}, \\
\text{(iii)} &= -\frac{1}{n_{xz}+n_{yz}+n_{xyz}} \sum_{i \in \mathcal{I}^{xz} \cup \mathcal{I}^{yz} \cup \mathcal{I}^{xyz}} \{\widehat{\mu}^{(k)}_x(Z_i) - \mu_x(Z_i)\}\{\widehat{\mu}^{(k)}_y(Z_i) - \mu_y(Z_i)\}, \\
\text{(iv)} &= \frac{1}{n_{xz}+n_{xyz}} \sum_{k=1}^{2} \sum_{i \in \mathcal{I}^{xz}_k \cup \mathcal{I}^{xyz}_k} \mu_x(Z_i)\{\widehat{\mu}^{(k)}_y(Z_i) - \mu_y(Z_i)\} \\
&\quad - \frac{1}{n_{xz}+n_{yz}+n_{xyz}} \sum_{k=1}^{2} \sum_{i \in \mathcal{I}^{xz}_k \cup \mathcal{I}^{yz}_k \cup \mathcal{I}^{xyz}_k} \mu_x(Z_i)\{\widehat{\mu}^{(k)}_y(Z_i) - \mu_y(Z_i)\}, \\
\text{(v)} &= \frac{1}{n_{yz}+n_{xyz}} \sum_{k=1}^{2} \sum_{i \in \mathcal{I}^{yz}_k \cup \mathcal{I}^{xyz}_k} \mu_y(Z_i)\{\widehat{\mu}^{(k)}_x(Z_i) - \mu_x(Z_i)\} \\
&\quad - \frac{1}{n_{xz}+n_{yz}+n_{xyz}} \sum_{k=1}^{2} \sum_{i \in \mathcal{I}^{xz}_k \cup \mathcal{I}^{yz}_k \cup \mathcal{I}^{xyz}_k} \mu_y(Z_i)\{\widehat{\mu}^{(k)}_x(Z_i) - \mu_x(Z_i)\}.
\end{aligned}
$$

Note that each data in the k-th fold is independent of the fitted functions applied to them. Thus, each summation term for $i \in \mathcal{I}^{xz}_k \cup \mathcal{I}^{xyz}_k$ in (i) is i.i.d. and mean zero conditional on $\widehat{\mu}^{(k)}_y$, whose conditional variance is $\mathbb{E}[(\widehat{\mu}^{(k)}_y(Z) - \mu_y(Z))^2 (X - \mu_x(Z))^2 \,|\, \widehat{\mu}^{(k)}_y] = o_P(1)$. The Markov's inequality thus implies $\text{(i)} = o_P(1/\sqrt{n_{xz}+n_{xyz}})$. Similarly, we know $\text{(ii)} = o_P(1/\sqrt{n_{yz}+n_{xyz}})$. Also, since $\|\widehat{\mu}^{(k)}_x - \mu_x\|_{L_2(\mathbb{P}_Z)} \cdot \|\widehat{\mu}^{(k)}_y - \mu_y\|_{L_2(\mathbb{P}_Z)} = o_P(1/\sqrt{n})$, we have $\text{(iii)} = o_P(1/\sqrt{n_{xz}+n_{xyz}} + 1/\sqrt{n_{yz}+n_{xyz}})$. In addition, let $D_i := \mu_x(Z_i)\{\widehat{\mu}^{(k)}_y(Z_i) - \mu_y(Z_i)\} - \mathbb{E}\big[\mu_x(Z_i)\{\widehat{\mu}^{(k)}_y(Z_i) - \mu_y(Z_i)\}\big]$ for $i \in \mathcal{I}^{xz}_k \cup \mathcal{I}^{xyz}_k$, where the expectation is condi-

tional on data out of the k-th fold. Note that

$$\text{(iv)} = \frac{1}{n_{xz}+n_{xyz}} \sum_{k=1}^{2} \sum_{i \in \mathcal{I}_k^{xz} \cup \mathcal{I}_k^{xyz}} D_i - \frac{1}{n_{xz}+n_{yz}+n_{xyz}} \sum_{k=1}^{2} \sum_{i \in \mathcal{I}_k^{xz} \cup \mathcal{I}_k^{yz} \cup \mathcal{I}_k^{xyz}} D_i.$$

For each k, conditional on $\{(Y_i, Z_i) \colon i \in (\mathcal{I}^{xyz} \backslash \mathcal{I}_k^{xyz}) \cup (\mathcal{I}^{yz} \backslash \mathcal{I}_k^{yz})\}$, we know that $\{D_i\}$ are i.i.d. with mean zero. This implies

$$\begin{aligned} \big|\text{(iv)}\big| &\le O_P(\|D_i\|_{L_2(\mathbb{P})}/\sqrt{n_{xz}+n_{xyz}}) + O_P(\|D_i\|_{L_2(\mathbb{P})}/\sqrt{n_{xz}+n_{yz}+n_{xyz}}) \\ &\le o_P(1/\sqrt{n_{xz}+n_{xyz}}) + o_P(1/\sqrt{n_{xz}+n_{yz}+n_{xyz}}). \end{aligned}$$

The same bounds hold for (v). Putting them together, we obtain

$$\widehat{C}_{\text{miss}} = C_{\text{miss}} + o_P(1/\sqrt{n_{xz}+n_{xyz}} + 1/\sqrt{n_{yz}+n_{xyz}}).$$

Recall that $n_{xz}/(n_{xz}+n_{yz}+n_{xyz}) \to \rho_{xz}$, and $n_{yz}/(n_{xz}+n_{yz}+n_{xyz}) \to \rho_{yz}$. Rearranging the terms in C_{miss} gives $C_{\text{miss}} = C^*_{\text{miss}} + o_P(1/\sqrt{n_{xz}+n_{xyz}} + 1/\sqrt{n_{yz}+n_{xyz}})$, where

$$\begin{aligned} C^*_{\text{miss}} &= \sum_{i \in \mathcal{I}^{xz}} \frac{X_i \mu_y(Z_i)}{n_{xz}+n_{xyz}} - \frac{\mu_y(Z_i)\mu_x(Z_i)}{n_{xz}+n_{yz}+n_{xyz}} + \sum_{i \in \mathcal{I}^{yz}} \frac{Y_i \mu_x(Z_i)}{n_{yz}+n_{xyz}} - \frac{\mu_y(Z_i)\mu_x(Z_i)}{n_{xz}+n_{yz}+n_{xyz}} \\ &\quad + \sum_{i \in \mathcal{I}^{xyz}} \frac{X_i \mu_y(Z_i)}{n_{xz}+n_{xyz}} + \frac{Y_i \mu_x(Z_i)}{n_{yz}+n_{xyz}} - \frac{\mu_y(Z_i)\mu_x(Z_i)}{n_{xz}+n_{yz}+n_{xyz}} \\ &= \frac{1}{n_{xz}} \sum_{i \in \mathcal{I}^{xz}} \Big(\frac{\rho_{xz}}{1-\rho_{yz}} X_i \mu_y(Z_i) - \rho_{xz} \mu_y(Z_i)\mu_x(Z_i) \Big) \\ &\quad + \frac{1}{n_{yz}} \sum_{i \in \mathcal{I}^{yz}} \Big(\frac{\rho_{yz}}{1-\rho_{xz}} Y_i \mu_x(Z_i) - \rho_{yz} \mu_y(Z_i)\mu_x(Z_i) \Big) \\ &\quad + \frac{1}{n_{xyz}} \sum_{i \in \mathcal{I}^{xyz}} \Big(\frac{X_i \mu_y(Z_i)}{1-\rho_{yz}} + \frac{1}{1-\rho_{xz}} Y_i \mu_x(Z_i) - \mu_y(Z_i)\mu_x(Z_i) \Big). \end{aligned}$$

Furthermore, we note that $C^*_{\text{miss}} = \mathbb{E}[XY] + O_P(1/\sqrt{n_{xz}+n_{xyz}} + 1/\sqrt{n_{yz}+n_{xyz}})$. Similar to the arguments in the proof of Theorem 2, we have

$$\widehat{\theta}^{\text{mod}}_n - \theta^* = \mathbb{E}[XX^\top]^{-1}(C^*_{\text{miss}} - \widehat{\Sigma}_{\text{miss}}\theta^*).$$

This concludes the proof of Proposition 14. ∎

Appendix D. Supporting lemmas

Lemma 17 (Bernstein's inequality) *Suppose $X_1, \ldots, X_n$ are independent zero-mean random variables such that $|X_i| \le M$ almost surely for some constant $M > 0$. Then for any constant $t > 0$,*

$$\mathbb{P}\Big(\Big| \sum_{i=1}^{n} X_i \Big| \ge t \Big) \le 2 \exp\Big(- \frac{t^2}{2\sum_{i=1}^{n} \mathbb{E}[X_i^2] + 2Mt/3} \Big).$$

That is, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that $\big|\sum_{i=1}^{n} X_i\big| \le \max\big\{2\sqrt{\sum_{i=1}^{n}\mathbb{E}[X_i^2]\log(2/\delta)}, 4M\log(2/\delta)/3\big\}$.

Appendix E. Deferred simulation results

E.1 Low-dimensional simulation

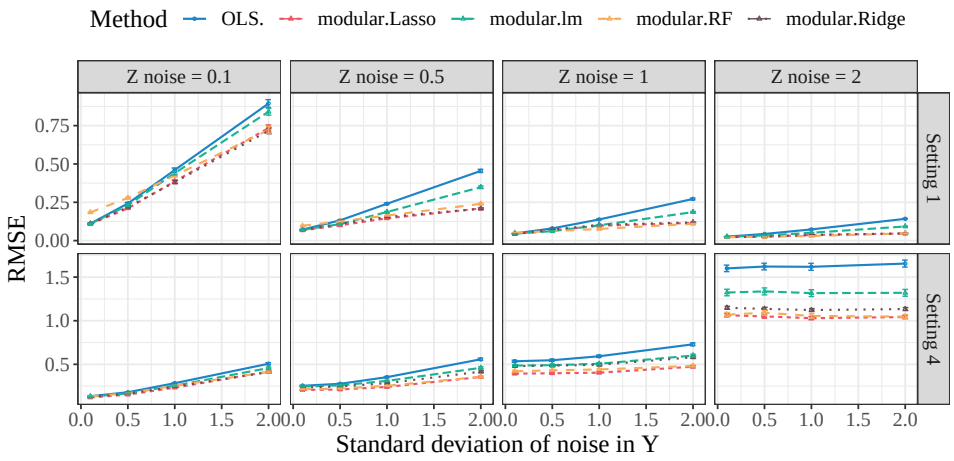


Figure 13: Aggregate RMSE in settings 1 and 4. Details are otherwise the same as Figure 3.

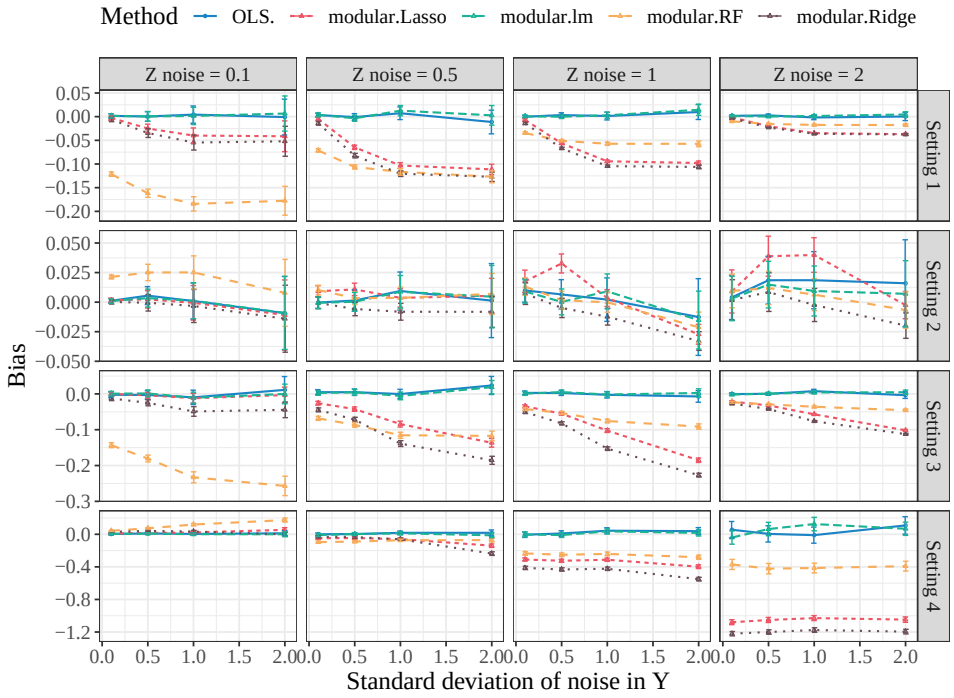


Figure 14: Bias summed over all entries in all settings. Details are otherwise the same as Figure 3. Modular regression incurs a slightly larger bias than OLS when the sub-tasks are learned by flexible machine learning algorithms.

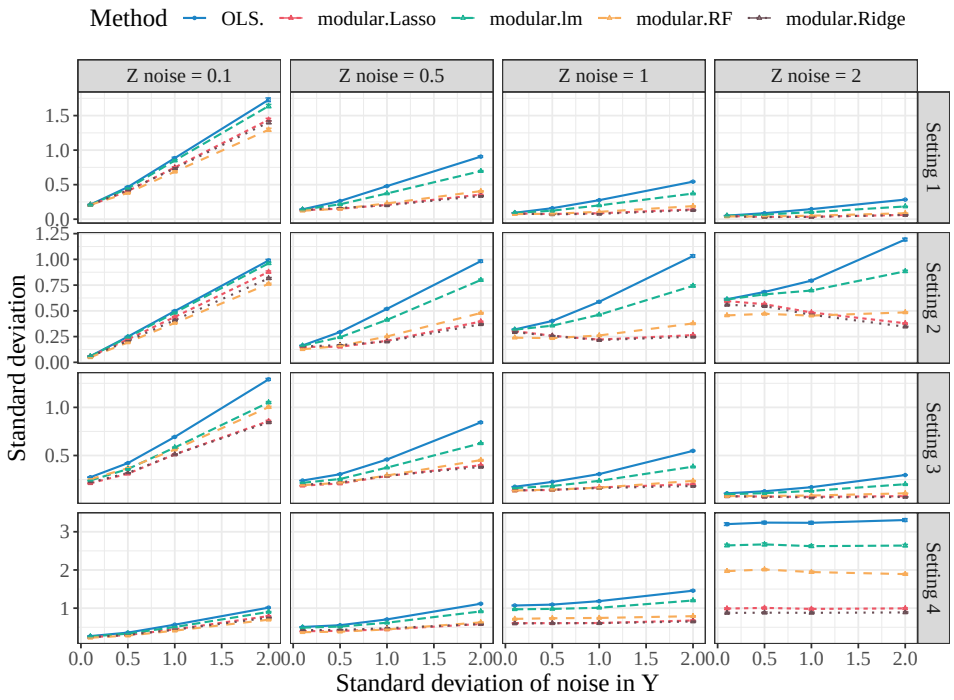


Figure 15: Aggregated SD in all settings. Details are otherwise the same as Figure 3.

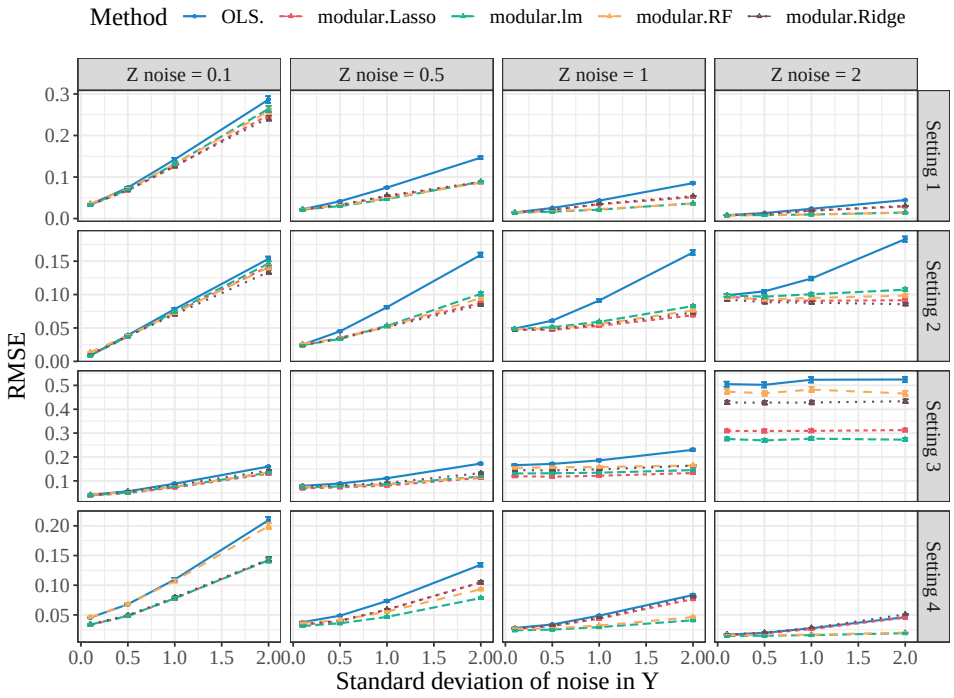


Figure 16: Aggregated RMSE when $n = 2000$. Details are otherwise the same as Figure 3.

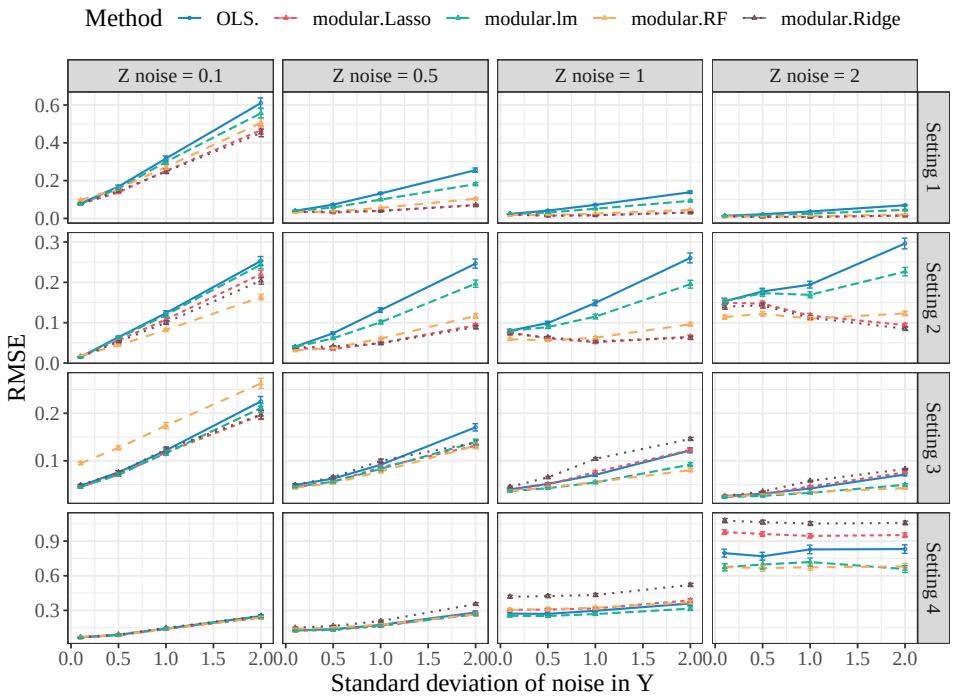


Figure 17: RMSE for estimating θ_1^* in all settings. Details are the same as Figure 3.

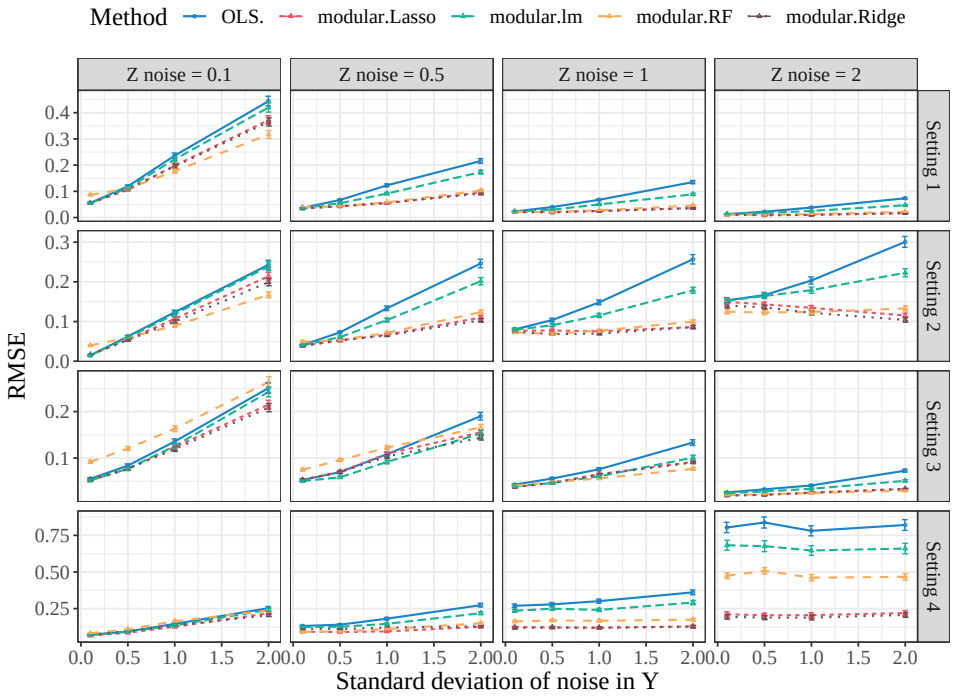


Figure 18: RMSE for estimating θ_2^* in all settings. Details are the same as Figure 3.

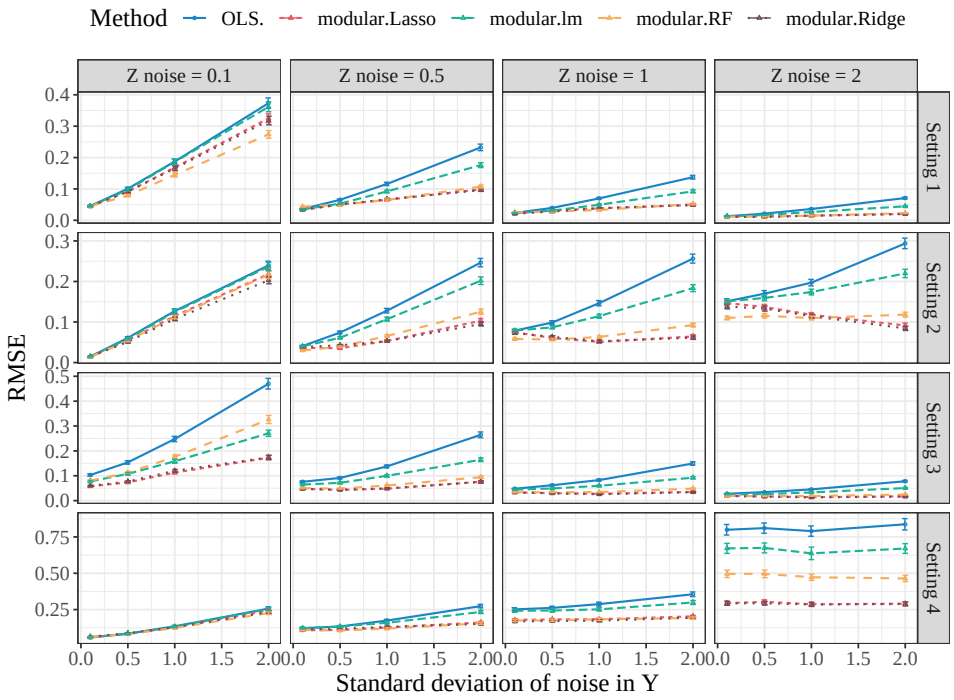


Figure 19: RMSE for estimating θ_3^* in all settings. Details are the same as Figure 3.

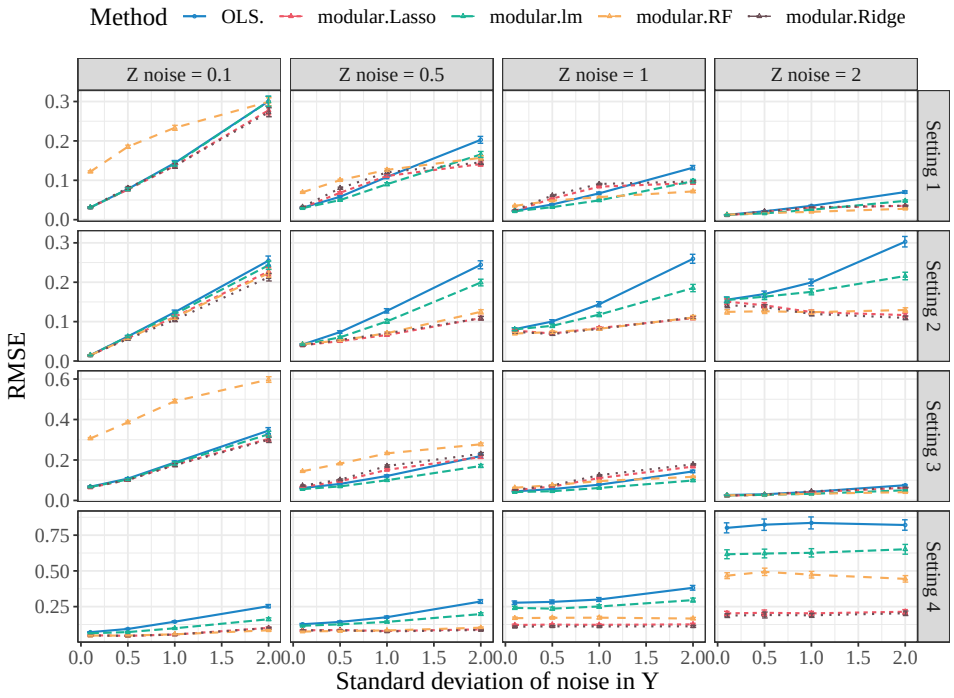


Figure 20: RMSE for estimating θ_4^* in all settings. Details are the same as Figure 3.

E.2 High-dimensional simulation

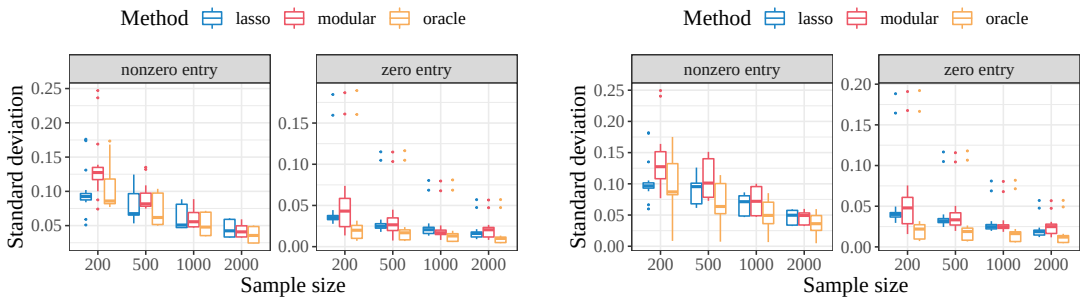


Figure 21: Boxplot of standard deviations for $\{\widehat{\theta}_j : j \in [p_x]\}$, averaged over $N = 1000$ replicates in setting 1 (left) and setting 2 (right). Other details are the same as Figure 12.

References

Kofi P Adragni and R Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 2009.

Susan Athey, Raj Chetty, Guido Imbens, and Hyunseung Kang. Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. *arXiv preprint arXiv:1603.09326*, 2016.

Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

Onureena Banerjee, Laurent El Ghaoui, Alexandre d'Aspremont, and Georges Natsoulis. Convex optimization techniques for fitting sparse gaussian graphical models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 89–96, 2006.

Colin B Begg and Denis HY Leung. On the use of surrogate end points in randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1): 15–28, 2000.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Springer, 1993.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Peter Bühlmann. Proposing the vote of thanks: Regression shrinkage and selection via the lasso: a retrospective by robert tibshirani, 2010.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

T Tony Cai and Anru Zhang. Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of Multivariate Analysis*, 150:55–74, 2016.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Timothy I Cannings and Yingying Fan. The correlation-assisted missing data estimator. *Journal of Machine Learning Research*, 23:41–1, 2022.

David Causeur and Thierry Dhorne. Linear regression models under conditional independence restrictions. *Scandinavian Journal of Statistics*, 30(3):637–650, 2003.

Hua Chen, Zhi Geng, and Jinzhu Jia. Criteria for surrogate end points. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):919–932, 2007.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Xiaohong Chen and Xiaotong Shen. Sieve extremum estimates for weakly dependent data. *Econometrica*, pages 289–314, 1998.

Victor Chernozhukov, Christian Hansen, and Yuan Liao. A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL `https://doi.org/10.1111/ectj.12097`.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

Andreas Elsener and Sara van de Geer. Sparse spectral estimation with missing and corrupted measurements. *Stat*, 8(1):e229, 2019.

Thomas R Fleming, Ross L Prentice, Margaret S Pepe, and David Glidden. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and aids research. *Statistics in Medicine*, 13(9):955–968, 1994.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `https://www.jstatsoft.org/v33/i01/`.

Nir Friedman, Iftach Nachman, and Dana Peér. Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 206–215, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606149.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):757–778, 2015.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Leonard Henckel, Emilija Perković, and Marloes H Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*, 2019.

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

Ying Jin and Dominik Rothenhäusler. Attribute-adaptive statistical inference for finite populations under distribution shift. *arXiv preprint arXiv:2104.04565*, 2021.

Michael Irwin Jordan. *Learning in graphical models*. MIT Press, 1999.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*, 2020.

Wai Lam and Fahiem Bacchus. Learning bayesian belief networks: An approach based on the mdl principle. *Computational Intelligence*, 10(3):269–293, 1994.

Steffen L Lauritzen, Odd O Aalen, Donald B Rubin, and Elja Arjas. Discussion on causality [with reply]. *Scandinavian Journal of Statistics*, 31(2):189–201, 2004.

Su-In Lee, Varun Ganapathi, and Daphne Koller. Efficient structure learning of markov networks using l_1-regularization. *Advances in Neural Information Processing Systems*, 19, 2006.

Bing Li. *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC, 2018.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in Neural Information Processing Systems*, 24, 2011.

Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

Yanyuan Ma and Liping Zhu. Efficient estimation in sufficient dimension reduction. *The Annals of Statistics*, 2013.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. *Advances in Neural Information Processing Systems*, 12, 1999.

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

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Z. Oldfield, N. Rogers, A. Phelps, M. Blake, A. Steptoe, A. Oskala, M. Marmot, S. Clemens, J. Nazroo, and J. Banks. English longitudinal study of ageing: Waves 0-9, 1998-2019, 2021. URL `https://beta.ukdataservice.ac.uk/datacatalogue/doi/?id=5050#21`.

Wendy J Post, Ciska Buijs, Ronald P Stolk, Elisabeth GE de Vries, and Saskia Le Cessie. The analysis of longitudinal quality of life measures with informative drop-out: a pattern mixture approach. *Quality of Life Research*, 19(1):137–148, 2010.

Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989.

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2, pages 335–422. Institute of Mathematical Statistics, 2008.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *arXiv preprint arXiv:2002.06708*, 2020.

Evan TR Rosenman, Art B Owen, Mike Baiocchi, and Hailey R Banack. Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine*, 41 (1):65–86, 2022.

Volker Roth. The generalized lasso. *IEEE Transactions on Neural Networks*, 15(1):16–28, 2004.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *arXiv preprint arXiv:1912.00306*, 2019.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415, 1958.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Sara Van de Geer. The deterministic lasso. *ETH Zürich research collection*, 2007.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 2014.

Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. 2009.

Mark J Van der Laan and James M Robins. *Unified methods for censored longitudinal data and causality*, volume 5. Springer, 2003.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Tyler J VanderWeele. Surrogate measures and consistent surrogates. *Biometrics*, 69(3): 561–565, 2013.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242, 2014.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.