Inference of nonlinear causal effects with application to TWAS with GWAS summary data

Ben Dai* BENDAI@CUHK.EDU.HK
Department of Statistics, The Chinese University of Hong Kong

Chunlin Li* CHUNLIN@IASTATE.EDU
Department of Statistics, Iowa State University

Haoran Xue XUEXX268@UMN.EDU
Department of Biostatistics, City University of Hong Kong

Wei Pan PANXX014@UMN.EDU
Division of Biostatistics, The University of Minnesota

Xiaotong Shen XSHEN@UMN.EDU
School of Statistics, The University of Minnesota

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Large-scale genome-wide association studies (GWAS) have offered an exciting opportunity to discover putative causal genes or risk factors associated with diseases by using SNPs as instrumental variables (IVs). However, conventional approaches assume linear causal relations partly for simplicity and partly for the availability of GWAS summary data. In this work, we propose a novel model for transcriptome-wide association studies (TWAS) to incorporate nonlinear relationships across IVs, an exposure/gene, and an outcome, which is robust against violations of the valid IV assumptions, permits the use of GWAS summary data, and covers two-stage least squares (2SLS) as a special case. We decouple the estimation of a marginal causal effect and a nonlinear transformation, where the former is estimated via sliced inverse regression and a sparse instrumental variable regression, and the latter is estimated by a ratio-adjusted inverse regression. On this ground, we propose an inferential procedure. An application of the proposed method to the ADNI gene expression data and the IGAP GWAS summary data identifies 18 causal genes associated with Alzheimer's disease, including APOE and TOMM40, in addition to 7 other genes missed by 2SLS considering only linear relationships. Our findings suggest that nonlinear modeling is required to unleash the power of IV regression for identifying potentially nonlinear gene-trait associations. The source code and accompanying software `nl-causal` can be accessed through the link: https://github.com/statmlben/nonlinear-causal.

Keywords: nonlinear causal effect, sliced inverse regression, GWAS, TWAS

1. Introduction

Causal inference methods in transcriptome-wide association studies (TWAS) have successfully discovered numerous (putative) *causal genes* associated with complex traits and diseases (Gusev et al., 2016), using genetic variants, typically single nucleotide polymorphisms (SNPs), as instrumental variables (IVs) (Yang et al., 2010). Understanding these gene-to-disease associations has considerable ramifications in the field of genomics, possibly spearheading a much-anticipated revolution in personalized and precision medicine.

* Co-first authorship.

2SLS in TWAS. Conventional TWAS applies two-sample two-stage least squares (2SLS; Kang et al. (2016b)) to integrate expression quantitative trait locus (eQTL) data for gene expression and genome-wide association study (GWAS) summary data for a trait of interest, thereby pinpointing potential causal genes for disease risk, such as Alzheimer's Disease (AD). Specifically, we denote instrumental variables as $\boldsymbol{z} \in \mathbb{R}^p$, a scalar exposure as $x \in \mathbb{R}$, and a scalar outcome as $y \in \mathbb{R}$. For example, SNPs ($\boldsymbol{z}$) are used as instrumental variables for a gene's expression (x) to identify its causal association with AD risk (y). 2SLS assumes that $(\boldsymbol{z}, x, y)$ satisfy a two-stage *linear* model:

$$x = \boldsymbol{z}^\top \boldsymbol{\theta} + w, \qquad y = \beta x + \boldsymbol{z}^\top \boldsymbol{\alpha} + \varepsilon, \tag{1}$$

where (w, ε) are the error terms independent of the instruments $\boldsymbol{z}$, however, w and ε may be correlated due to underlying confounders, and $\beta \in \mathbb{R}$, $\boldsymbol{\alpha} \in \mathbb{R}^p$, $\boldsymbol{\theta} \in \mathbb{R}^p$ are unknown parameters.

The primary objective of 2SLS is for statistical inference on the causal effect β of the exposure x on the outcome y based on (1). The estimation of β via 2SLS can be executed in two stages: (Stage 1) 2SLS utilizes IVs $\boldsymbol{z}$ to predict the exposure x via linear regression, subsequently providing an estimate $\widehat{\boldsymbol{\theta}}$; (Stage 2) the estimated "debiased" exposure (obtained as $\widehat{x} = \widehat{\boldsymbol{\theta}}^\top \boldsymbol{z}$) is used to estimate the causal effect β via a regression from $\widehat{x}$ to y. Consequently, 2SLS produces *unbiased* estimation of the causal effect from exposure to the outcome by mitigating confounder-induced bias. Another key benefit of 2SLS is its ability to infer based solely on the summary statistics of x-$\boldsymbol{z}$ and y-$\boldsymbol{z}$ correlations. This feature is particularly beneficial for *privacy-constrained* datasets, such as SNP genotype data. In the content of TWAS, for each gene being treated as an exposure, 2SLS first builds a predictive model using its cis-SNPs around this gene as IVs for the expression level with the eQTL data. Then the predicted gene expression is obtained with the GWAS summary data and tested for association with the trait to determine whether the gene is putatively causal to the trait.

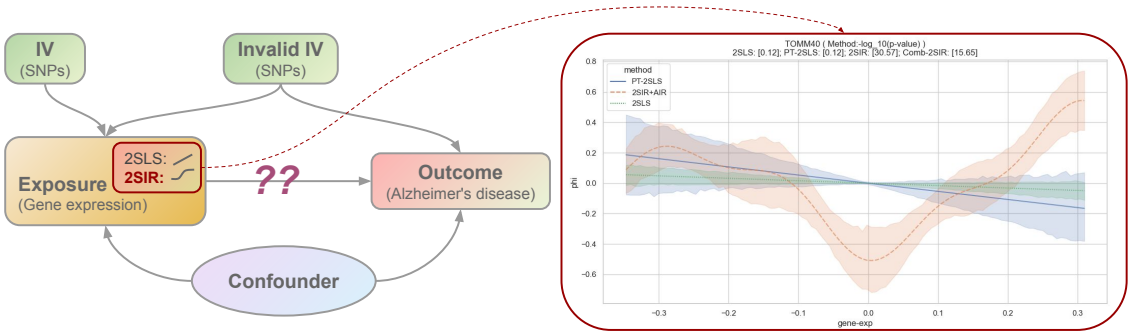

Figure 1: **Left.** A structure plot of the proposed 2SIR model, which admits a nonlinear causal effect from exposure to outcome. **Right.** Estimated transformations of TOMM40 (a well-known AD gene) based on 2SLS, PT-2SLS, and our 2SIR+AIR, and the resulting p-values are included in the title, yielding that TOMM40 is only identified by our method. Moreover, the R^2s for the stage one model on $\widehat{\phi}(x) \sim \widehat{\boldsymbol{\theta}}^\top \boldsymbol{z}$ are 0.230 (2SLS), 0.230 (PT-2SLS), and 0.253 (2SIR), suggesting that the nonlinear model (2) is suited for this data.

Despite the substantial advantages of the TWAS using 2SLS in causal inference, a primary limitation surfaces due to its inherent assumption of linearity. Previous TWAS studies (Gamazon et al.,

2015; Gusev et al., 2016; Zhu et al., 2016) generally propose a linear relationship between cis-SNPs and gene expression in the first stage and between gene expression and a GWAS trait/outcome in the subsequent stage. This framework overlooks the likely existence of nonlinear effects (Mackay, 2014). On the other hand, to our knowledge, none of the existing non-parametric IV regression methods are applicable to GWAS summary data, while individual-level GWAS data are usually unavailable due to privacy and logistic issues, presenting challenges to incorporating flexible nonlinear models into TWAS with GWAS summary data. In our motivating example, the individual-level AD GWAS data from many sub-studies are unavailable, but its meta-analyzed summary data are available. Some recently proposed methods (Zhang and Ghosh, 2017; Okoro et al., 2021) relax the linear assumption in stage 1, while others do so in stage 2 (He et al., 2023), which however requires the use of individual-level data. Misspecification of a nonlinear effect as a linear (or other specific) one may distort subsequent causal inference, damping the statistical power of the TWAS method. For illustration, we consider the eQTL data for a well-known AD-related gene, TOMM40, from our real data example; see Section 4 for more details. Figure 1 provides some compelling evidence for the nonlinear effects in both stages of TWAS. In the first stage, it displays a nonlinear relationship between the cis-SNPs and the gene expression level of TOMM40, as evidenced by a higher R^2 value of the nonlinear model over those of its linear competitors. In the second stage, a nonlinear causal association of TOMM40 with the AD risk is strongly corroborated by the highly significant p-value obtained with our method. Consequently, this well-known AD gene is successfully identified by our proposed method (2SIR+AIR) but missed by both 2SLS and its power-transformed extension (PT-2SLS), suggesting the necessity of nonlinear modeling in TWAS.

Moreover, as an IV regression method, conventional TWAS relies on three key IV assumptions to remove the hidden confounding effects: (IV1) the IVs are associated with the exposure, (IV2) the IVs only affects the outcome through the exposure, and (IV3) the IVs are not associated with the unmeasured confounders. While (IV1) is straightforward to handle, (IV2) and (IV3) are fragile in practice due to the widespread pleiotropy of SNPs (Solovieff et al., 2013). This phenomenon refers to the situation when an SNP affects the GWAS trait/disease not mediated through exposure, violating (IV2) and/or (IV3) and causing severe bias in causal inference. A line of recent works (Kang et al., 2016a; Windmeijer et al., 2019; Guo et al., 2018) has been focusing on the violation of (IV2) and/or (IV3). Of note, these methods use linear models, and their nonlinear counterparts remain unexplored.

Other methods. Besides TWAS, Mendelian Randomization (MR) is another important and popular subject in genetics that uses SNPs as IVs to infer a causal relationship between an exposure and an outcome, typically two complex traits (Morrison et al., 2020; Xue et al., 2021). Both TWAS and conventional MR are two-stage IV regression methods for causal inference, and they share many similarities, yet their implementations are different due to distinct types of data being used. Although both TWAS and conventional MR use GWAS summary data in the second stage, in the first stage MR uses GWAS summary data of sample size typically in tens of thousands or even larger, while TWAS typically uses individual-level eQTL data of sample size in a few hundreds or at most one or two thousands. Usually, SNPs being used in TWAS are around the target gene (i.e. cis-SNPs) and are correlated, while most MR methods use independent SNPs from the whole genome. Due to these distinctions, the existing typical MR methods do not fit the TWAS analysis.

In a nutshell, nonlinear modeling that is robust to the violation of IV assumptions and at the same time leverages large-scale GWAS summary data lacks for TWAS analysis. To addressing the limitations of existing methods, we develop an approach with the following novel aspects.

- We propose a flexible model to admit an *arbitrary unknown nonlinear* causal relationship between an exposure and an outcome. Importantly, the proposed model is applicable to GWAS summary data while being robust to invalid IVs, and *covers 2SLS* as a special case.

- Based on the proposed model, we decouple the estimation of a *causal effect* and a *nonlinear causal transformation*. The inference of the causal effect are established by the proposed 2SIR based on sliced inverse regression. Then, the unknown nonlinear transformation can be estimated by the proposed AIR. The validity of the proposed hypothesis testing and interval estimation is ensured by our theoretical result, and verified by extensive simulation study.

- The ADNI data and the IGAP GWAS summary data confirm the efficacy of our approach. The results (Section 4) indicate that our method successfully replicates the significant AD genes identified by 2SLS, while uniquely identifying 7 additional causal genes. Our real data analysis suggests that nonlinear modeling is suited to unleash the power of TWAS.

2. Nonlinear modeling of TWAS data

We denote a vector of IVs as $\boldsymbol{z} \in \mathbb{R}^p$, a scalar exposure as $x \in \mathbb{R}$, and a scalar outcome as $y \in \mathbb{R}$. In our TWAS case study (cf. Section 4), SNPs are used as instrumental variables for a gene's expression to identify its causal association with the AD risk. Without loss of generality, we assume $(\boldsymbol{z}, x, y)$ has mean zero. Suppose $(\boldsymbol{z}, x, y)$ satisfy a nonlinear model

$$\phi(x) = \boldsymbol{z}^\intercal \boldsymbol{\theta} + w, \qquad y = \beta\phi(x) + \boldsymbol{z}^\intercal \boldsymbol{\alpha} + \varepsilon, \tag{2}$$

where (w, ε) are the error terms independent of the instruments $\boldsymbol{z}$, and $\beta \in \mathbb{R}$, $\boldsymbol{\alpha} \in \mathbb{R}^p$, $\boldsymbol{\theta} \in \mathbb{R}^p$ are unknown parameters, and $\phi(\cdot)$ is an unknown transformation.

The following provides some in-depth motivations for the proposed model (2). First, as shown by others (Lin et al., 2022; He et al., 2023) and to be shown here, there is empirical evidence to support the existence of non-linear effects that certain genes have on various traits, thus the possible non-linear function $\phi(x)$ in (2). Second, it is well known that, due to the small effect sizes of SNPs on complex traits, linear models for the effects of SNPs perform well in practice, hence we adopt the widely-used linearity assumption of $\boldsymbol{z}$, which (implicitly) connects the two-stage models in (2). An alternative, and perhaps more popular, non-linear model as used in Hartford et al. (2017); He et al. (2023) would be a linear model of the effects of SNPs $\boldsymbol{z}$ on the gene expression x in Stage 1 but a similar non-linear Stage 2 model as in (2), which however would imply a non-linear model for the effects of SNPs $\boldsymbol{z}$ on trait y. This perhaps is debatable: since the causal pathway is likely to be from SNPs $\boldsymbol{z}$ to gene x then to trait y, the effect sizes of SNPs (i.e. their heritabilities) are expected to be smaller on y than on x, suggesting that if a linear model of $\boldsymbol{z}$ on x is reasonable, another linear model of $\boldsymbol{z}$ on y should approximately hold. In fact, it was shown empirically that, even if a linear model of the effects of SNPs $\boldsymbol{z}$ on a gene's expression level x was reasonable in Stage 1, assuming a linear model of $\boldsymbol{z}$ on x^2 (Stage 1 in our model) performed better than a non-linear model (as implied by the linearity of $\boldsymbol{z}$ on x), again likely due to the small effect sizes of SNPs and the parsimony of linear models (see Remarks subsection in Materials and Methods section of Lin et al. (2022)). Importantly, the implicit linear structure allows the use of GWAS summary data of our method, in contrast to requiring individual-level data by the other non-linear models.

Furthermore, our model (2) holds two significant advantages over 2SLS (1). First, the assumptions of (2) are weaker than the classical 2SLS. Specifically, (2) admits an *arbitrary* nonlinear transformation $\phi(\cdot)$ across $\boldsymbol{z}$, x and y, relaxing the linearity assumption in the standard TWAS/2SLS. Second, it includes 2SLS and Yeo-Johnson power transformation 2SLS (PT-2SLS) (Yeo and Johnson, 2000) as special cases. It is worth mentioning that the proposed method remains competitive against 2SLS/PT-2SLS even if the linear assumption or normality assumption holds; see Section 3. Overall, the proposed model (2) is a natural extension of 2SLS.

In (2), $\beta\phi(\cdot)$ represents the influence of the exposure on the outcome, which is our primary focus, while $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are nuisance parameters. In particular, $\boldsymbol{\alpha} \neq \mathbf{0}$ indicates the violation of the second and/or third IV assumptions. Generally, the effect $\beta\phi(\cdot)$ may not be identifiable with the presence of invalid IVs. In the literature, additional structural constraints are imposed to avoid this issue. For example, if $\|\boldsymbol{\alpha}\|_0 < p/2$ is known a priori, then $\beta\phi(\cdot)$ becomes well-defined (Kang et al., 2016b). Furthermore, note that β and ϕ are only identifiable up to a multiplicative scalar, even if $\beta\phi(\cdot)$ is well-defined in (2). Thus, we fix $\|\boldsymbol{\theta}\|_2 = 1$ and $\beta \geq 0$ in the subsequent discussion so that β and ϕ are identifiable.

On this ground, Definition 1 summarizes the quantities of interest.

Definition 1 (Causal effect and transformation) *In (2), let $\|\boldsymbol{\theta}\|_2 = 1$ and $\beta \geq 0$. Then,*
(i) β is called the marginal causal effect;
(ii) $\phi(\cdot)$ is called the nonlinear transformation (of the exposure);
(iii) $\beta\phi(\cdot)$ is called the nonlinear effect function.

Specifically, β summarizes the marginal effect of the causal influence of the exposure x on the outcome y, in that $\beta > 0$ indicates the presence of the causal relation, and the corresponding hypothesis testing and confidence interval are developed in Sections 2.1. It is worth noting that β in (2) only represents the magnitude of the causal effect, which does not imply a positive/negative relation as in 2SLS, due to the nonlinear transformation $\phi(\cdot)$. If the model (2) is well-specified, the nonlinear effect function $\beta\phi(\cdot)$ in (iii) can be used to measure the average treatment effect (ATE) between two exposure/treatment levels. In our case study, $\beta > 0$ indicates the presence of the causal influence of a gene on the AD risk, and if the model (2) is well-specified, $\phi(\cdot)$ represents the potentially nonlinear pattern of a putative causal association.

Let $(\boldsymbol{Z}_\nu, \boldsymbol{X}_\nu, \boldsymbol{Y}_\nu)$ be $n_\nu \times (p+2)$ matrix, where each row $(\boldsymbol{z}_{\nu i}, x_{\nu i}, y_{\nu i})$, $1 \leq i \leq n_\nu$, $\nu = 1, 2$, represents an independent observation from (2). In what follows, assume that we have two independent samples $\mathcal{D}_1 = \{\boldsymbol{Z}_1, \boldsymbol{X}_1\}$ and $\mathcal{D}_2 = \{n_2^{-1}\boldsymbol{Z}_2^\intercal\boldsymbol{Z}_2, n_2^{-1}\boldsymbol{Z}_2^\intercal\boldsymbol{Y}_2, n_2^{-1}\boldsymbol{Y}_2^\intercal\boldsymbol{Y}_2\}$ from (2). Without loss of generality, we assume that $\boldsymbol{Y}_2$ is pre-normalized as $n_2^{-1}\boldsymbol{Y}_2^\intercal\boldsymbol{Y}_2 = 1$. Importantly, we require neither that all variables $(\boldsymbol{z}, x, y)$ are observed simultaneously, nor the availability of individual-level data $(\boldsymbol{Z}_2, \boldsymbol{X}_2, \boldsymbol{Y}_2)$, allowing the application to summary statistics, like GWAS summary data, for the second sample. Our goal is to infer β and $\beta\phi(\cdot)$ from the observed data $\mathcal{D}_1, \mathcal{D}_2$. In the sequel, we propose estimating the marginal causal effect β and the nonlinear transformation ϕ separately.

2.1. Estimation and inference of marginal causal effect

The proposed procedure for estimating β consists of two stages. In the first stage, note that $x \perp\!\!\!\perp \boldsymbol{z} \mid \boldsymbol{z}^\intercal\boldsymbol{\theta}$ in (2), which coincides with a single index model (Duan and Li, 1991; Cook, 2009), and the sliced inverse regression (SIR; Li (1991)) can be used to estimate $\boldsymbol{\theta}$. Specifically, given the dataset $\mathcal{D}_1$, SIR divides the range of x_i into S non-overlapping slices $\text{Slice}_s (s = 1, \cdots, S)$, and estimates

$\boldsymbol{\theta}$ as the eigenvector of $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Gamma}}$ associated with the largest eigenvalue:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\max}\, \boldsymbol{\theta}^\intercal \widehat{\boldsymbol{\Gamma}}\boldsymbol{\theta}, \quad \text{s.t.} \quad \boldsymbol{\theta}^\intercal\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} = 1, \tag{3}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix of $\boldsymbol{z}$, and $\widehat{\boldsymbol{\Gamma}}$ is the between slice covariance matrix, with n_{1s} being the number of samples in the s-th slice Slice_s:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n_1}\sum_{i=1}^{n_1} \boldsymbol{z}_{1i}\boldsymbol{z}_{1i}^\intercal, \quad \widehat{\boldsymbol{\Gamma}} = \sum_{s=1}^{S} \frac{n_{1s}}{n_1}\bar{\boldsymbol{z}}_{(s)}\bar{\boldsymbol{z}}_{(s)}^\intercal, \quad \bar{\boldsymbol{z}}_{(s)} = \frac{1}{n_{1s}}\sum_{x_i\in\text{Slice}_s} \boldsymbol{z}_{1i},$$

In the second stage, we estimate β via a sparse instrumental variable regression using the data $\mathcal{D}_2$. Specifically, note that the second equation in (2) can be rewritten as

$$y = \boldsymbol{z}^\intercal\boldsymbol{\theta}\beta + \boldsymbol{z}^\intercal\boldsymbol{\alpha} + e, \qquad e = w\beta + \varepsilon, \qquad \mathrm{E}(e) = 0, \qquad \mathrm{E}(e^2) = \sigma_e^2. \tag{4}$$

Recall that $\alpha_j \neq 0$ indicate z_j violates (IV2) and/or (IV3). Motivated by Xue et al. (2021), we separate the potential bias due to invalid IVs from the causal effect β via a sparse regression:

$$\min_{\boldsymbol{\alpha},\beta} (\widehat{\boldsymbol{\theta}}\beta + \boldsymbol{\alpha})^\intercal \boldsymbol{Z}_2^\intercal\boldsymbol{Z}_2(\widehat{\boldsymbol{\theta}}\beta + \boldsymbol{\alpha}) - 2\boldsymbol{Y}_2^\intercal\boldsymbol{Z}_2(\widehat{\boldsymbol{\theta}}\beta + \boldsymbol{\alpha}) \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_0 \leq K, \tag{5}$$

where $\|\boldsymbol{\alpha}\|_0 = \sum_{j=1}^p \mathrm{I}(\alpha_j \neq 0)$ and $K \geq 0$ is an integer tuning parameter indicating the number of invalid IVs. For implementation, $\|\cdot\|_0$ penalty can be replaced by a sparsity-inducing surrogate penalty, such as SCAD (Fan and Li, 2001), TLP (Shen and Huang, 2010), and MCP (Zhang, 2010). In our data analysis, we use the SCAD as a computational surrogate; see Appendix B.1 for details.

Taken together, the proposed procedure consists of the estimation of $\boldsymbol{\theta}$ via a Sliced Inverse Regression, and that of β via a Sparse Instrumental Regression. This methodology is named Two-Stage Instrumental Regression (2SIR), as summarized in Algorithm 1.

Algorithm 1: Two-stage instrumental regression (2SIR) for β estimation

Input: Datasets $\mathcal{D}_1$ and $\mathcal{D}_2$

1 (Stage 1: Sliced inverse regression) Estimate $\widehat{\boldsymbol{\theta}}$ via (3) with $\mathcal{D}_1$;

2 (Stage 2: Sparse instrumental regression) Estimate $\widehat{\beta}$ via (5) with $\mathcal{D}_2$ and $\widehat{\boldsymbol{\theta}}$;

3 (Sign adjustment for identifiability) $\widehat{\boldsymbol{\theta}} \leftarrow \text{sign}(\widehat{\beta})\widehat{\boldsymbol{\theta}}, \qquad \widehat{\beta} \leftarrow |\widehat{\beta}|$;

4 **return** *Estimated causal effect* $(\widehat{\beta}, \widehat{\boldsymbol{\theta}})$

Next, we turn to present inferential procedures for the marginal causal effect β, including hypothesis testing and confidence intervals. Before proceeding, Theorem 2 summarizes the asymptotic properties of the 2SIR estimator.

Theorem 2 *Let $\widehat{\beta}$ be the 2SIR estimator produced by Algorithm 1 with $\|\cdot\|_0$ penalty, SCAD, TLP, or MCP being used in (5). Assume Conditions C.1 and C.2 in Appendix C.2. If $K = |A|$ in (5) and (w, ε) is normally distributed, then*

$$n_2^{1/2}(\widehat{\beta} - \beta) = |n_2^{1/2}\beta + \zeta - \eta| - n_2^{1/2}\beta + o_p(1), \quad \zeta \perp\!\!\!\perp \eta,$$
$$\zeta \sim N(0, \Omega_X\sigma_e^2), \quad \eta \sim \sqrt{r}\beta\Omega_X\boldsymbol{\theta}^\intercal\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\xi},$$

where $A = \{j : \alpha_j \neq 0\}$, $n_2/n_1 \to r$ and $n_1^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{\xi}$, $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{*A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A*}$, $\Omega_X = (\boldsymbol{\theta}^\intercal\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\theta})^{-1}$, and $\boldsymbol{\Sigma}_{*A}, \boldsymbol{\Sigma}_{A*}$ denote the columns and rows of $\boldsymbol{\Sigma}$ indexed by A, respectively.

In Theorem 2, (w, ε) is assumed to be normally distributed for simplicity, which is not critical to large-sample inference. Now, we infer β based on Theorem 2. First, consider the hypotheses:

$$H_0 : \beta = 0 \text{ versus } H_a : \beta > 0,$$

where rejecting the null hypothesis H_0 indicates evidence for causal influence of the exposure x on the outcome y. Define the pivotal test statistic

$$\widehat{T} = \frac{n_2^{1/2}\widehat{\beta}}{\widehat{\sigma}_e(\widehat{\boldsymbol{\theta}}^\intercal\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^\intercal\widehat{\boldsymbol{\Sigma}}_{*A}(\widehat{\boldsymbol{\Sigma}}_{AA})^{-1}\widehat{\boldsymbol{\Sigma}}_{A*}\widehat{\boldsymbol{\theta}})^{1/2}}. \tag{6}$$

Given a significance level $\alpha \in (0, 1)$, the null hypothesis H_0 is rejected if and only if $\widehat{T} > \Phi^{-1}_{N(0,1)}(1 - \alpha/2)$, where $\Phi^{-1}_{N(0,1)}(\cdot)$ denotes the quantile function of $N(0, 1)$. As a consequence of Theorem 2, Corollary 3 justifies the proposed test.

Corollary 3 *Assume the conditions in Theorem 2. The following statements are true.*
(i) Under the null hypothesis $H_0 : \beta = 0$, we have

$$\limsup_{n_2\to\infty} P_{H_0}\left(\widehat{T} > \Phi^{-1}_{N(0,1)}(1 - \alpha/2)\right) \le \alpha. \tag{7}$$

(ii) Under the alternative hypothesis $H_a : \beta = n_2^{-1/2}h$, we have

$$\liminf_{n_2\to\infty} P_{H_a}\left(\widehat{T} > \Phi^{-1}_{N(0,1)}(1 - \frac{\alpha}{2})\right) \ge P\left(|N(\Omega_X^{-1/2}\sigma_e^{-1}h, 1)| > \Phi^{-1}_{N(0,1)}(1 - \frac{\alpha}{2})\right).$$

Empirically, Section 3 shows that the proposed test can control the Type I error under the null hypothesis H_0 while possessing desirable power under H_a. Moreover, we developed a combined test over a different number of slices for 2SIR, see Appendix B.2.

Next, we consider constructing a valid CI for β. Indeed, this can be challenging, since the asymptotics of the SIR estimator depends on an unknown distribution $\boldsymbol{z} \mid x$ (Zhu and Ng, 1995), which is intractable. To overcome this difficulty, we propose a resampling-based CI in light of Theorem 2. Specifically, by the triangle inequality, $n_2^{1/2}|\widehat{\beta} - \beta| \le |\zeta - \eta| + o_p(1)$. Therefore, the CI of β can be produced by resampling $|\zeta - \eta|$.

For implementation, we first compute $(\widehat{\boldsymbol{\theta}}, \widehat{\beta})$ via Algorithm 1, denote $\widehat{\boldsymbol{\Sigma}}_R = \widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}_{*A}\widehat{\boldsymbol{\Sigma}}_{AA}^{-1}\widehat{\boldsymbol{\Sigma}}_{A*}$, $\widehat{\Omega}_X = (\widehat{\boldsymbol{\theta}}^\intercal\widehat{\boldsymbol{\Sigma}}_R\widehat{\boldsymbol{\theta}})^{-1}$, and $\widehat{\sigma}_e^2 = n_2^{-1}(\boldsymbol{Y}_2^\intercal\boldsymbol{Y}_2 - \boldsymbol{Y}_2^\intercal\boldsymbol{Z}_2(\boldsymbol{Z}_2\boldsymbol{Z}_2)^{-1}\boldsymbol{Z}_2^\intercal\boldsymbol{Y}_2)$. Then the bootstrap estimates $\widehat{\boldsymbol{\theta}}_l^*$s are computed as $\widehat{\boldsymbol{\theta}}_l^* = \text{sign}(\widehat{\boldsymbol{\theta}}^\intercal\widetilde{\boldsymbol{\theta}}_l^*)\widetilde{\boldsymbol{\theta}}_l^*$, where $\widetilde{\boldsymbol{\theta}}_l^*$ is computed via Step 1 (SIR) in Algorithm 1 based on resampling $\mathcal{D}_1$, and $\zeta_l^* \sim N(0, \widehat{\Omega}_X\widehat{\sigma}_e^2)$ is generated according to its asymptotic distribution; $l = 1, \ldots, M$, where M is the Monte-Carlo size. In this way, we approximate the distribution of η by the Monte-Carlo sample: for $l = 1, \ldots, M$, $\eta_l^* = \frac{1}{2}n_2^{1/2}\widehat{\beta}\widehat{\Omega}_X\big((\widehat{\boldsymbol{\theta}_l^*})^\intercal\widehat{\boldsymbol{\Sigma}}_R\widehat{\boldsymbol{\theta}}_l^* - \widehat{\boldsymbol{\theta}}^\intercal\widehat{\boldsymbol{\Sigma}}_R\widehat{\boldsymbol{\theta}}\big)$. Hence, the $(1 - \alpha)$-confidence interval is:

$$\beta \in \left[\max(0, \widehat{\beta} - n_2^{-1/2}\widehat{Q}^*(1 - \alpha)),\ \widehat{\beta} + n_2^{-1/2}\widehat{Q}^*(1 - \alpha)\right], \tag{8}$$

where $\widehat{Q}^*(\cdot)$ is the quantile function of $(|\zeta_l^* - \eta_l^*|)_{l=1}^M$. It is worth noting that if $\beta = 0$, then $|\zeta - \eta| = |\zeta|$ and $n_2^{1/2}\widehat{\beta} = |\zeta| + o_p(1)$ by Theorem 2, and the test based on (6) is thus optimal. However, in the case of $\beta \approx 0$, the bootstrap quantile $\widehat{Q}^*(1 - \alpha)$ is usually larger than the quantile

$Q(1-\alpha)$ of $|\zeta-\eta|$, due to additional variations of η^*. Hence, the CI in (8) is less efficient than the test (7) when detecting a small signal $\beta \approx 0$. Finally, for interval estimation of $\beta\phi(\cdot)$, there is ample literature devoted to constructing nonparametric confidence bands; see (Hall et al., 2013).

Section 3 indicates that the proposed method yields peak performance in the estimation and inference of the marginal causal effect in various simulated examples. Yet, in practice, visualization of ϕ may shed light on the specific relationship between the exposure and outcome. In the next section, we develop an algorithm to estimate the nonlinear transformation ϕ.

2.2. Estimation of nonlinear transformation

The challenge of estimating $\phi(\cdot)$ is twofold. First, individual-level data of (z, x, y) are usually unavailable, preventing the estimation of ϕ from the second equation of (2). Second, w is correlated with $\phi(x)$ in (2), rendering a biased estimator when for example a least-squares regression of $\boldsymbol{z}^\intercal\boldsymbol{\theta}$ is conducted over x. To address these issues, we propose an Adjusted Inverse Regression (AIR) for consistent estimation of ϕ. An important observation is made in Proposition 4, showing that the transformation ϕ is proportional to the least-squares estimator.

Proposition 4 *Suppose* $\mathrm{E}(\boldsymbol{z}^\intercal\boldsymbol{\theta} \mid x) = \mathrm{E}(\boldsymbol{z}^\intercal\boldsymbol{\theta} \mid \phi(x))$ *and* $(\boldsymbol{z}^\intercal\boldsymbol{\theta}, w)$ *has an elliptically symmetric distribution. Then there exists a constant* ρ *such that* $\phi(x) = \rho\,\mathrm{E}(\boldsymbol{z}^\intercal\boldsymbol{\theta} \mid x)$.

In light of Proposition 4, ϕ can be estimated by a two-stage procedure. First, we estimate the conditional mean $\mathrm{E}(\boldsymbol{z}^\intercal\boldsymbol{\theta} \mid x)$ via the least-squares regression:

$$\widehat{m} = \underset{m\in\mathcal{F}}{\arg\min}\ \frac{1}{2n_1}\sum_{i=1}^{n_1}\left(\boldsymbol{z}_{1i}^\intercal\widehat{\boldsymbol{\theta}} - m(x_{1i})\right)^2, \tag{9}$$

where $\mathcal{F}$ is a class of functions, and (9) includes various nonparametric methods, such as spline regression (Wahba, 1990), and gradient boosting regression (Friedman, 2001). Then, $\widehat{\rho}$ is estimated base on the uncorrelatedness between $\boldsymbol{z}^\intercal\boldsymbol{\theta}$ and w, that is,

$$\frac{1}{n_1}\sum_{i=1}^{n_1}(\boldsymbol{z}_{1i}^\intercal\widehat{\boldsymbol{\theta}})\left(\boldsymbol{z}_{1i}^\intercal\widehat{\boldsymbol{\theta}} - \widehat{\rho}\widehat{m}(x_{1i})\right) = 0, \qquad \widehat{\rho} = \frac{\widehat{\boldsymbol{\theta}}^\intercal\sum_{i=1}^{n_1}(\boldsymbol{z}_{1i}\boldsymbol{z}_{1i}^\intercal)\widehat{\boldsymbol{\theta}}}{\widehat{\boldsymbol{\theta}}^\intercal\sum_{i=1}^{n_1}\widehat{m}(x_{1i})\boldsymbol{z}_{1i}}. \tag{10}$$

Finally, the AIR estimator is $\widehat{\phi} = \widehat{\rho}\widehat{m}$. It is worth noting that AIR allows the estimation of a non-invertible transformation ϕ, this is in contrast to the existing literature on data transformation (see Yeo and Johnson (2000)), where only invertible transformations are considered. In Section 3, the numerical results demonstrate the advantages of our method in detecting a quadratic relationship. For interval estimation of $\beta\phi(\cdot)$, there is ample literature devoted to constructing nonparametric confidence bands; see Hall et al. (2013) and references therein.

2.3. Robustness to misspecified nonlinearity

The proposed model (2) considerably relaxes the linearity assumption in 2SLS. Nevertheless, it is possible that the nonlinear transformation $\phi(\cdot)$ in (2) could be misspecified in practice, especially when two structural equations do not share the same transformation for the exposure:

$$\phi(x) = \boldsymbol{z}^\intercal\boldsymbol{\theta} + w, \quad y = \beta\psi(x) + \boldsymbol{z}^\intercal\boldsymbol{\alpha} + \varepsilon, \tag{11}$$

where $\phi \neq \psi$ are two different nonlinear functions. In TWAS, it is generally impossible to consistently estimate ψ from the summary statistics. Yet, testing in Section 2.1 remains valid.

Corollary 5 *Assume the conditions in Theorem 2, then under* H_0, (7) *still holds for the model* (11).

As a result, in our TWAS analysis, the p-values of the putative causal genes produced by 2SIR remain reliable regardless of whether the transformations are correctly specified. The simulation indicates that the proposed test enables control of the Type I error and outperforms its competitors in power in the misspecified cases; see Example 6 in Appendix B.7.

3. Simulations

This section examines the performance of the proposed 2SIR and AIR methods. Moreover, for hypothesis testing, we propose to combine tests based on different slices, denoted as Comb-2SIR. Let $\mathcal{S}$ be a collection of candidate slices, we combine p-values based on different slices $S \in \mathcal{S}$ using the Cauchy combining method (Liu and Xie, 2020). More discussion about the Cauchy combining version of 2SIR over the number of slices is included in Appendix B.2. Specifically, the results are compared against 2SLS and PT-2SLS. For PT-2SLS, the optimal parameter λ for minimizing skewness is estimated using maximum likelihood, c.f., Section 3 in Yeo and Johnson (2000).

The performance for both β and $\phi(\cdot)$ are considered. Due to space constraints, this section only reports the performance of controlling Type I and II errors, coverage, and effectiveness of confidence intervals of β, details and results about $\phi(\cdot)$ estimation are provided in the Appendix A.

The simulated data $\mathcal{D} = (\boldsymbol{z}_i, x_i, y_i)_{i=1}^n$ is generated as follows. First, $\boldsymbol{z}_i$ is generated independently from $N(\mathbf{0}_p, \boldsymbol{\Sigma})$, and $w_i = u_i^2 + \gamma_i$, where u_i and γ_i are independently generated from $N(0,1)$. Second, x_i is generated as $x_i = \phi^{-1}(\boldsymbol{\theta}^\intercal \boldsymbol{z}_i + w_i)$ when ϕ is invertible, and x_i is randomly selected from the solution set $\{x : \phi(x) = \boldsymbol{\theta}^\intercal \boldsymbol{z}_i + w_i\}$ when ϕ is non-invertible. Third, $y_i = \beta\phi(x_i) + \varepsilon_i$, where $\varepsilon_i = u_i + \zeta_i$, and $\zeta_i \sim N(0,1)$, thus u_i acts as a confounder, and w_i is dependent with ε_i. Finally, the first half of the data is provided as $\mathcal{D}_1$, and the summary data $\mathcal{D}_2$ is produced by the second half of the data to mimic the GWAS data. Six transformations are considered: (1) linear: $\phi(x) = x$; (2) logarithm: $\phi(x) = \log(x)$; (3) cube root: $\phi(x) = x^{1/3}$; (4) inverse: $\phi(x) = 1/x$; (5) piecewise linear: $\phi(x) = xI(x \leq 0) + 0.5xI(x > 0)$; (6) quadratic: $\phi(x) = x^2$.

For Type I error and power analysis, we compute the proportions of rejecting out of 1,000 simulations under H_0 and out of 100 simulations under H_a, respectively. For constructing the CI, we report the averaged coverage and CI length out of 1,000 simulations. Note that the CIs for 2SLS and PT-2SLS are generated based on the asymptotic variance in Inoue and Solon (2010), the CIs for 2SIR are generated based on (8), and all CIs are left truncated at 0 since $\beta \geq 0$.

Example 1 (Standard setting). In this example, we examine the proposed method under a standard setting. Specifically, we set $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ and normalize it by its norm. We examine four cases: (i) $\beta = 0$, (ii) $\beta = .05$, (iii) $\beta = .10$, (iv) $\beta = .15$. Note that case (i) is for Type I error analysis, while $\beta > 0$ in (ii) - (iv), suggests power analysis. Moreover, the CI is produced based on (ii) $\beta = 0.05$. All empirical results are summarized in Figure 2 (testing) and Table 1 (CI).

Examples 2-6. Additional examples, including Example 2 (Invalid IVs), Example 3 (Categorical IVs), Example 4 (Weak IVs), Example 5 (Non-additive effects), Example 6 (Misspecified models) can be found in Appendix B to assess the performance of our methods under various data situations.

In summary, the simulation suggests the efficacy of the proposed 2SIR in managing all types of nonlinear transformations across various scenarios. The key conclusions are itemized below.

- For testing, as suggested in Figure 2, the proposed 2SIR and its combined test yield competitive performance for "linear", "cube-root" and "PL" cases compared with 2SLS and PT-2SLS; and superior performance for "log", "inverse", and "quad" cases.

- For CI, as indicated in Table 1, 2SLS and PT-2SLS fail to provide valid CIs when "inverse" and "quad" transformations are used. For other cases, the proposed 2SIR yields competitive performance. In general, 2SIR is the only one that can provide a valid CI under an unknown nonlinear transformation.

- As suggested in Figures B.2 - B.6, and Tables B.1 - B.5, the proposed 2SIR continues to perform well with invalid, weak or categorical IVs. As indicated in Figure A.4 and Table A.4, 2SIR is also the most robust method against dominance and epistatic effects. As indicated in Figure B.7, the proposed methods can control Type I errors and are more powerful than the competitors when the transformation is misspecified.

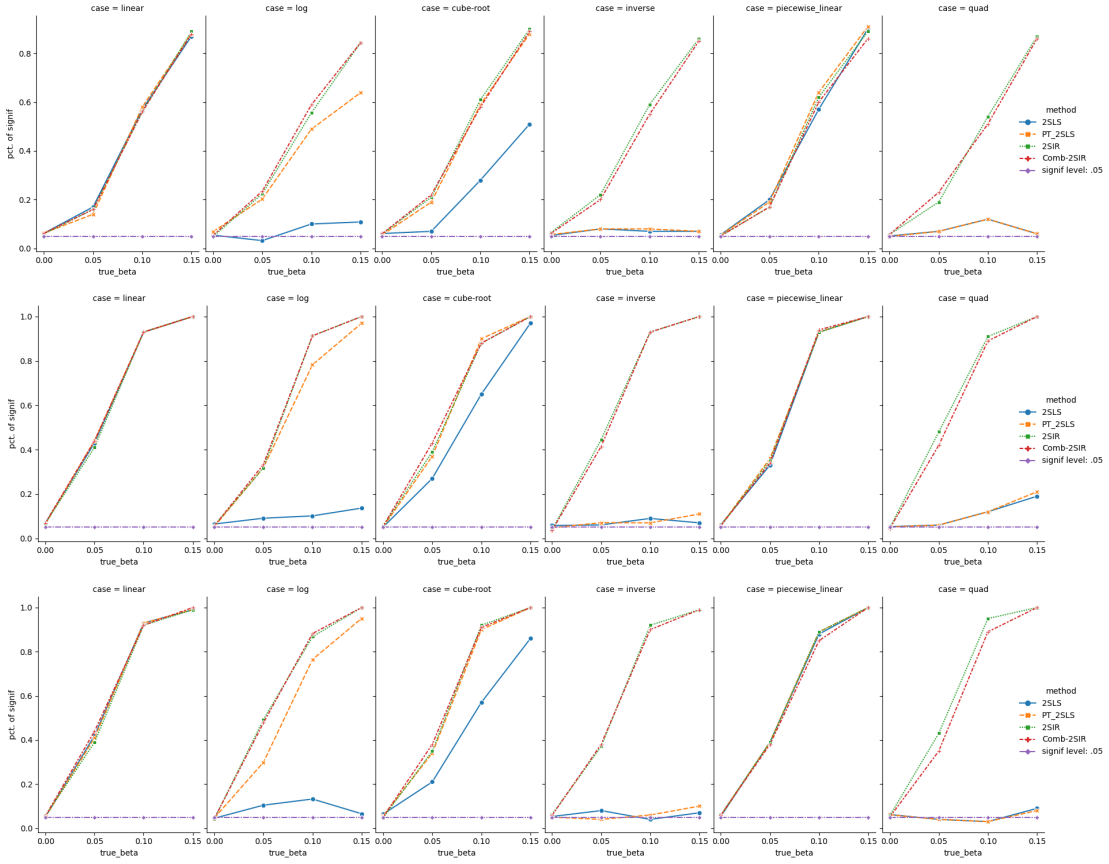

Figure 2: Empirical Type I error ($\beta_0 = 0$) and power ($\beta_0 > 0$) of marginal effect inference in Example 1 of Section 3. $(n, p) = (2000, 50), (5000, 50), (5000, 100)$ from up to bottom.

(n, p)		*2SLS* coverage	*2SLS* length	*PT-2SLS* coverage	*PT-2SLS* length	*2SIR* (proposed) coverage	*2SIR* (proposed) length
$(2000, 10)$	linear	0.944	0.132	0.943	0.132	0.967	0.138
	log	0.946	156.422	0.946	0.133	0.925	0.136
	cube-root	1.000	0.390	1.000	0.436	0.975	0.138
	inverse	0.964	0.522	0.930	0.134	0.979	0.138
	PL	0.950	0.134	0.949	0.134	0.971	0.138
	quad	0.831	0.093	0.823	0.092	0.951	0.139
$(2000, 50)$	linear	0.941	0.128	0.943	0.129	0.974	0.136
	log	1.000	176.916	0.913	0.123	0.935	0.136
	cube-root	1.000	0.328	0.940	0.132	0.976	0.136
	inverse	0.990	0.149	0.882	0.096	0.979	0.131
	PL	0.943	0.126	0.944	0.127	0.982	0.134
	quad	0.743	0.084	0.743	0.083	0.976	0.134
$(5000, 50)$	linear	0.950	0.094	0.952	0.095	0.978	0.095
	log	1.000	95.559	1.000	0.090	0.972	0.097
	cube-root	1.000	0.215	0.999	0.095	0.982	0.097
	inverse	0.801	0.209	0.640	0.060	0.972	0.096
	PL	0.951	0.096	0.960	0.096	0.977	0.096
	quad	0.522	0.052	0.523	0.051	0.976	0.095
$(10000, 50)$	linear	0.952	0.074	0.955	0.074	0.958	0.074
	log	1.000	76.293	0.925	0.072	0.955	0.074
	cube-root	0.949	0.174	0.960	0.074	0.944	0.074
	inverse	0.532	0.126	0.380	0.042	0.969	0.075
	PL	0.954	0.075	0.956	0.075	0.959	0.075
	quad	0.287	0.040	0.291	0.043	0.931	0.074

Table 1: Empirical coverage and length of the CI for in Example 1 of Section 3.

4. Real data analysis

In this section, we implement the proposed method for an analysis of the AD Neuroimaging Initiative (ADNI) dataset and the International Genomics of Alzheimer's Project (IGAP; Lambert et al. (2013)) GWAS summary dataset to identify putative causal AD genes. Specifically, the ADNI dataset consists of 819 individual-level subjects, 17,201 genes, and 620,901 SNPs. The IGAP dataset consists of summary statistics of about 7 million SNPs to AD based on 54,162 samples.

Data preprocessing. To facilitate the analysis, we pre-process the dataset and refine the candidate SNPs as follows. For the ADNI dataset, we first exclude SNPs with MAF ≤ 0.05, with missing values, or failing the Hardy-Weinberg equilibrium test at the significant level of 0.001. Next, we further prune the SNPs to ensure that any of their pairwise Pearson correlations in absolute values were no more than 0.6. For the IGAP GWAS dataset, we conduct imputation for missing SNPs by using the software `ImpG` (Pasaniuc et al., 2014), based on 489 unrelated individuals with European ancestry from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015), yet remove the imputed SNPs with imputation accuracy smaller than 0.3. Finally, we define the cis-region of the gene by expanding 100kb upstream and downstream of its coding region, and take the top 50 intersecting SNPs (available both on the ADNI dataset and imputed IGAP dataset), with the largest absolute correlations with the gene's expression level. Taken together, the pre-processed dataset consists of 712 individual-level genotypes and gene expression with 50 SNPs and independent summary statistics for the associated SNPs based on 54,162 samples.

Results. Next, all methods are applied to the pre-processed data. As indicated in Figure 3, with the Bonferroni adjusted significance cutoff $0.05/17201$, 20 genes are identified as significantly related to AD by at least one method. Specifically, among them 12 were significant by 2SLS and/or PT-2SLS, 18 are significant by Comb-2SIR. Two genes, APOE and TOMM40 on chromosome 19, are well-known to be related to AD (Bu, 2009; Mise et al., 2017; Lyall et al., 2014); the former is identified by all three methods while the latter is only identified by Comb-2SIR. Besides TOMM40,

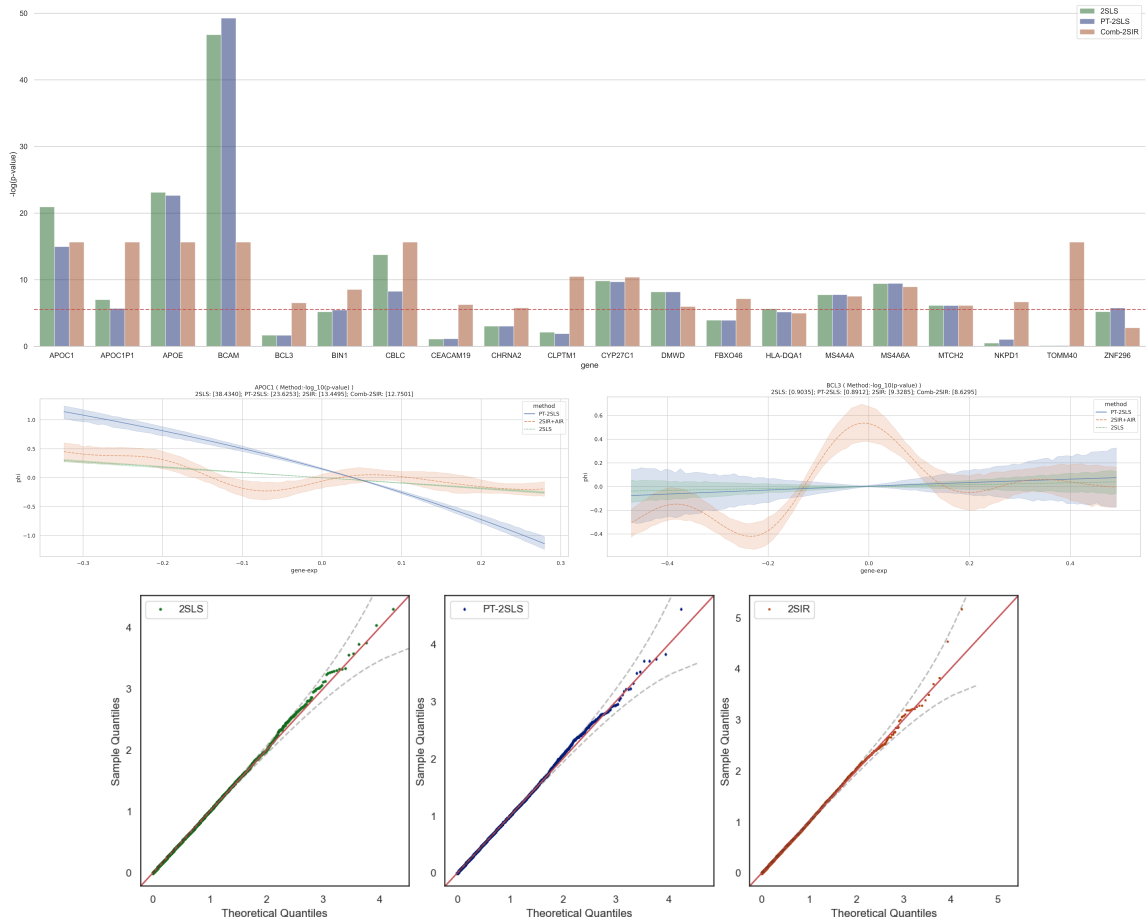

Figure 3: **Upper.** The bar-plot for significant AD genes, where the x-axis represents genes, the y-axis represents $-\log_{10}(p)$. **Middle.** Fitted transformations of two illustrative genes. APOC1 (*left*) is identified by all methods. BCL3 (*right*) is only identified by our method. **Lower.** QQ-plots for 2SLS, PT-2SLS, and 2SIR on ADNI negative control outcomes.

7 genes, BCL3, BIN1, CEACAM19, CHRNA2, CLPTM1, FBXO46, NKPD1, are only identified by Comb-2SIR. We searched these 7 genes in large-scale GWAS results and found all of them except FBXO46 contained genetic variants that have been reported to be significantly associated with AD (Jansen et al., 2019; Marioni et al., 2018; Beecham et al., 2014). A further literature search gives more supporting evidence about their associations with AD. Specifically, BCL3 has been discovered to be associated with late-onset familial AD (Nho et al., 2017; Pericak-Vance et al., 1991); in AD brains, BIN1 has increased expression levels (De Jager et al., 2014; Chapuis et al., 2013); CEACAM19 has been suggested as a candidate gene related to human aging (Evans and Cummings, 2019); CHRNA2 has been implicated in potentially contributing to learning and memory functions (Nichol, 2015) and as a potential target of clinical AD drugs (Cummings et al., 2019).

For illustration, Figure 3 (middle panel) shows the fitted transformations for two genes: APOC1 and BCL3 (others are included in Supplementary). For APOC1, which is successfully detected by

2SLS, the estimated transformation by our method is roughly in agreement with the linear pattern estimated by 2SLS. For BCL3, in contrast, the estimated transformation by our method is largely different from that of 2SLS and PT-2SLS, indicating that the linear pattern might be invalid here. This may be a reason for less significance given by 2SLS and PT-2SLS, offering practical and empirical evidence for nonlinear causal effects in a real dataset.

Negative control outcomes. We also demonstrate Type I error control based on the ADNI dataset with negative control outcomes. Specifically, we implement the methods based on individual-level SNPs and gene expressions while generating negative control outcomes by simulating random noises so that *no gene is causal to the outcome*. In this case, the p-value is expected to follow a uniform distribution. Figure 3 exhibits the QQ plots of the methods, suggesting that the p-values provided by 2SLS, PT-2SLS, and 2SIR are appropriately distributed in this negative control dataset.

5. Discussion and conclusions

Nonlinear modeling in TWAS has potential significance in identifying causal gene-trait associations. However, it is plagued by the lack of individual-level GWAS data (with only summary statistics for the outcome available). In this paper, we have proposed a flexible causal model for summary data while allowing an arbitrary nonlinear causal effect, substantially relaxing the assumption of linearity in the current practice of TWAS. A novel method called 2SIR+AIR is developed to estimate the marginal causal effect and the nonlinear transformation, covering 2SLS as a special case. In addition, we have developed inferential tools to assess exposure-outcome associations, including hypothesis testing and interval estimation; in particular, our test is robust to model misspecification.

We have demonstrated the applicability of the proposed model and methods by studying the ADNI gene expression and the IGAP GWAS datasets to identify putative causal genes for AD. Our results suggest that the proposed method agrees with two existing methods (2SLS and PT-2SLS) in 10 of 12 putative causal genes, but it additionally identifies 7 other potential AD genes. We also observe higher R^2's for the stage one model of our method than existing models, offering another source of evidence that nonlinear causal effects are likely to be present in real data. Our finding reasonably suggests potential nonlinearity in gene-trait causal associations based on GWAS data. We believe that the proposed method has great potential and could further advance research in TWAS, including nonlinear treatment effect analysis, subgroup analysis, and robustness analysis. Finally, in addition to TWAS, the proposed method can be equally applied to study other exposure-outcome causal relationships in a more general context.

Acknowledgments

We thank the reviewers and the area chair for many insightful comments and suggestions. This research was supported by HK GRF grants 14304823, 24302422 and CUHK Science Direct Grant for Research, NSF grant DMS-1952539, NIH grants R01 GM113250, R01 GM126002, RF1 AG067924, U01 AG073079, R01 AG074858, and R01 AG065636.

References

Claudia Becker and Ursula Gather. A note on the choice of the number of slices in sliced inverse regression. Technical report, Technical Report, 2007.

Gary W Beecham, Kara Hamilton, Adam C Naj, Eden R Martin, Matt Huentelman, Amanda J Myers, Jason J Corneveaux, John Hardy, Jean-Paul Vonsattel, Steven G Younkin, et al. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. *PLoS Genetics*, 10(9):e1004606, 2014.

Guojun Bu. Apolipoprotein E and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. *Nature Reviews Neuroscience*, 10(5):333–344, 2009.

J Chapuis, F Hansmannel, Marc Gistelinck, A Mounier, C Van Cauwenberghe, KV Kolen, F Geller, Y Sottejeau, D Harold, P Dourlen, et al. Increased expression of BIN1 mediates Alzheimer genetic risk by modulating tau pathology. *Molecular Psychiatry*, 18(11):1225–1234, 2013.

R Dennis Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*, volume 482. John Wiley & Sons, 2009.

R. Dennis Cook and Sanford Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991. ISSN 01621459. URL http://www.jstor.org/stable/2290564.

Jeffrey Cummings, Garam Lee, Aaron Ritter, Marwan Sabbagh, and Kate Zhong. Alzheimer's disease drug development pipeline: 2019. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5:272–293, 2019.

Philip L De Jager, Gyan Srivastava, Katie Lunnon, Jeremy Burgess, Leonard C Schalkwyk, Lei Yu, Matthew L Eaton, Brendan T Keenan, Jason Ernst, Cristin McCabe, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature Neuroscience*, 17(9):1156–1163, 2014.

Naihua Duan and Ker-Chau Li. Slicing regression: a link-free regression method. *Annals of Statistics*, 19(2):505–530, 1991.

Daniel S Evans and Steven R Cummings. Identification of ADAMTS7 and CEACAM19 as candidate healthy aging associated genes. *Innovation in Aging*, 3(Supplement_1):S102–S102, 2019.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

Zijian Guo, Hyunseung Kang, T Tony Cai, and Dylan S Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815, 2018.

Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 2016.

Peter Hall and Ker-Chau Li. On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics*, 21(2):867 – 889, 1993. doi: 10.1214/aos/1176349155. URL https://doi.org/10.1214/aos/1176349155.

Peter Hall, Joel Horowitz, et al. A simple bootstrap method for constructing nonparametric confidence bands for functions. *Annals of Statistics*, 41(4):1892–1921, 2013.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

Ruoyu He, Mingyang Liu, Zhaotong Lin, Zhong Zhuang, Xiaotong Shen, and Wei Pan. DeLIVR: a deep learning approach to iv regression for testing nonlinear causal effects in transcriptome-wide association studies. *Biostatistics*, 2023.

Atsushi Inoue and Gary Solon. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561, 2010.

Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics*, 51(3):404–413, 2019.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American statistical Association*, 111(513):132–144, 2016a.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016b.

Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452–1458, 2013.

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

Zhaotong Lin, Haoran Xue, Mykhaylo M Malakhov, Katherine A Knutson, and Wei Pan. Accounting for nonlinear effects of gene expression identifies additional associated genes in transcriptome-wide association studies. *Human molecular genetics*, 31(14):2462–2470, 2022.

Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529): 393–402, 2020.

Donald M Lyall, Sarah E Harris, Mark E Bastin, Susana Muñoz Maniega, Catherine Murray, Michael W Lutz, Ann M Saunders, Allen D Roses, Maria del C Valdés Hernández, Natalie A Royle, et al. Alzheimer's disease susceptibility genes APOE and TOMM40, and brain white matter integrity in the Lothian Birth Cohort 1936. *Neurobiology of Aging*, 35(6):1513–e25, 2014.

Trudy FC Mackay. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics*, 15(1):22–33, 2014.

Riccardo E Marioni, Sarah E Harris, Qian Zhang, Allan F McRae, Saskia P Hagenaars, W David Hill, Gail Davies, Craig W Ritchie, Catharine R Gale, John M Starr, et al. GWAS on family history of Alzheimer's disease. *Translational Psychiatry*, 8(1):1–7, 2018.

Ayano Mise, Yuta Yoshino, Kiyohiro Yamazaki, Yuki Ozaki, Tomoko Sao, Taku Yoshida, Takaaki Mori, Yoko Mori, Shinichiro Ochi, Jun-ichi Iga, et al. TOMM40 and APOE gene expression and cognitive decline in Japanese Alzheimer's disease subjects. *Journal of Alzheimer's Disease*, 60 (3):1107–1117, 2017.

Jean Morrison, Nicholas Knoblauch, Joseph H Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 52(7):740–747, 2020.

Kwangsik Nho, Sungeun Kim, Emrin Horgusluoglu, Shannon L Risacher, Li Shen, Dokyoon Kim, Seunggeun Lee, Tatiana Foroud, Leslie M Shaw, John Q Trojanowski, et al. Association analysis of rare variants near the APOE region with CSF and neuroimaging biomarkers of Alzheimer's disease. *BMC Medical Genomics*, 10(1):45–52, 2017.

Heather Nichol. *Optogenetic Investigation of Chrna2 Cells in The Subiculum and Their Role in Modulating Entorhinal Cortex Input*. McGill University (Canada), 2015.

Paul C Okoro, Ryan Schubert, Xiuqing Guo, W Craig Johnson, Jerome I Rotter, Ina Hoeschele, Yongmei Liu, Hae Kyung Im, Amy Luke, Lara R Dugas, et al. Transcriptome prediction performance across machine learning models and diverse ancestries. *Human Genetics and Genomics Advances*, 2(2):100019, 2021.

David Pacini and Frank Windmeijer. Robust inference for the two-sample 2SLS estimator. *Economics Letters*, 146:50–54, 2016.

Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P Strachan, Nick Patterson, and Alkes L Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20): 2906–2914, 2014.

MA Pericak-Vance, JL Bebout, PC Gaskell, LH Yamaoka, W-Y Hung, MJ Alberts, AP Walker, RJ Bartlett, CA Haynes, KA Welsh, et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *American Journal of Human Genetics*, 48(6):1034, 1991.

Xiaotong Shen and Hsin-Cheng Huang. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105(490):727–739, 2010.

Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009.

Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019.

Haoran Xue, Xiaotong Shen, and Wei Pan. Constrained maximum likelihood-based mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *The American Journal of Human Genetics*, 108(7):1251–1269, 2021.

Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7): 565–569, 2010.

In-Kwon Yeo and Richard A Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010.

Weiming Zhang and Debashis Ghosh. On the use of kernel machines for Mendelian randomization. *Quantitative Biology*, 5(4):368–379, 2017.

Li-Xing Zhu and Kai W Ng. Asymptotics of sliced inverse regression. *Statistica Sinica*, 5(2): 727–736, 1995.

Lixing Zhu, Baiqi Miao, and Heng Peng. On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474):630–643, 2006.

Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, 48(5):481–487, 2016.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509, 2008.

Appendix A. Simulation for transformation estimation

This subsection examines the proposed adjusted inverse regression (*2SIR+AIR*) in (A.1) under various nonlinear transformations, and the estimation accuracy is measured by mean square error (MSE) and uniform error (UE):

$$\mathrm{MSE}(\widehat{\phi},\phi_0)=\mathrm{E}\left(\left(\widehat{\phi}(x)-\phi_0(x)\right)^2\right),\qquad \mathrm{UE}(\widehat{\phi},\phi_0)=\mathrm{E}\sup_{x\in\mathcal{X}}\left|\widehat{\phi}(x)-\phi_0(x)\right| \tag{A.1}$$

where $\mathcal{X}$ is a region of causal interest, which is replaced as 100 grid points of [5%-quantile, 95%-quantile] of x for evaluation. We also compare the results with a conditional mean function to highlight the role of the ratio correction in (10).

Specifically, we set $\boldsymbol{\theta}=(p^{-1/2},\cdots,p^{-1/2})^\intercal$ and $\beta=1$ in (2). Note that $\mathcal{D}_1$ and $\mathcal{D}_2$ are generated with the same setting in Example 1 in Section 3 with $w_i=u_i+\gamma_i$, u_i and γ_i are independently generated from $N(0,1)$. Five nonlinear transformations are considered: (1) linear: $\phi(x)=x$; (2) logarithm: $\phi(x)=\log(x)$; (3) cube root: $\phi(x)=x^{1/3}$, (4) piecewise linear (PL): $\phi(x)=xI(x\le 0)+0.5xI(x>0)$, (5) quadratic (quad): $\phi(x)=x^2$. Note that the conditional mean regression (9) is conducted based on a KNN model with the number of neighbors as 100. The simulation is replicated 100 times with $n=2000, p=10,50,100$, the resulting MSEs and UEs are summarized in Table A.1, and the fitted transformations for $p=10$ is illustrated in Figure A.1.

It is evident that the proposed 2SIR+AIR method substantially outperforms 2SLS and PT-2SLS in most cases, except that 2SLS yields better performance in the "linear" case where the proposed model in (2) becomes a linear structural equation model. For other cases, the amount of improvement is significant, with the largest improvement of (MSE: 99.9%, UE: 96.6%) and (MSE: 92.2%, UE: 64.4%) over 2SLS and PT-2SLS, respectively.

Appendix B. Implementation and additional simulations

B.1. Computation and hyperparameter tuning

To solve (4), we first approximate the $\|\cdot\|_0$ penalty by the SCAD (Fan and Li, 2001), and then consider the corresponding regularized problem:

$$\min_{\boldsymbol{\alpha},\beta}(\widehat{\boldsymbol{\theta}}\beta+\boldsymbol{\alpha})^\intercal\boldsymbol{Z}_2^\intercal\boldsymbol{Z}_2(\widehat{\boldsymbol{\theta}}\beta+\boldsymbol{\alpha})-2\boldsymbol{Y}_2^\intercal\boldsymbol{Z}_2(\widehat{\boldsymbol{\theta}}\beta+\boldsymbol{\alpha})+\lambda p_a(\boldsymbol{\alpha}), \tag{B.1}$$

p		*2SLS* MSE	*2SLS* UE	*PT-2SLS* MSE	*PT-2SLS* UE
10	linear	0.000(.000)	0.000(.000)	0.525(.005)	1.216(.005)
	log	363.405(48.756)	9.892(0.045)	0.619(.004)	1.362(.004)
	cube-root	346.575(6.023)	21.777(0.042)	1.293(.009)	1.737(.009)
	PL	1.026(.005)	2.130(.002)	0.540(.004)	1.284(.005)
	quad	2.461(.009)	3.073(.004)	2.083(.009)	2.824(.004)
50	linear	0.000(.000)	0.000(.000)	0.535(.005)	1.171(.004)
	log	223.565(22.881)	12.106(.028)	0.616(.004)	1.342(.003)
	cube-root	355.761(5.317)	19.961(.038)	1.302(.010)	1.738(.008)
	PL	1.022(.004)	2.134(.002)	0.546(.005)	1.256(.005)
	quad	2.474(.009)	3.287(.003)	2.095(.008)	3.033(.004)
100	linear	0.000(.000)	0.000(.000)	0.526(.004)	1.204(.004)
	log	615.467(32.895)	7.429(.044)	0.623(.005)	1.580(.004)
	cube-root	354.663(5.198)	20.571(.023)	1.300(.009)	1.740(.010)
	PL	1.018(.005)	2.103(.002)	0.541(.004)	1.176(.004)
	quad	2.468(.009)	3.097(.004)	2.092(.008)	2.851(.004)

p		*Cond-mean(KNN)* MSE	*Cond-mean(KNN)* UE	*2SIR+AIR* (proposed) MSE	*2SIR+AIR* (proposed) UE
10	linear	3.530(.179)	3.076(.090)	0.117(.003)	0.615(.012)
	log	3.471(.205)	2.945(.094)	0.118(.002)	0.589(.016)
	cube-root	3.336(.205)	2.766(.099)	0.113(.002)	0.584(.016)
	PL	2.853(.207)	2.614(.104)	0.123(.003)	0.645(.016)
	quad	1.323(.060)	1.568(.042)	0.123(.004)	0.638(.013)
50	linear	3.305(.214)	3.022(.096)	0.125(.003)	0.598(.015)
	log	3.273(.214)	2.829(.104)	0.124(.002)	0.534(.016)
	cube-root	3.408(.216)	2.922(.100)	0.121(.003)	0.561(.013)
	PL	3.113(.214)	2.965(.100)	0.119(.003)	0.583(.016)
	quad	1.162(.069)	1.581(.055)	0.163(.006)	0.837(.020)
100	linear	3.203(.217)	3,019(.095)	0.142(.003)	0.570(.010)
	log	3.591(.220)	2.741(.111)	0.148(.003)	0.539(.011)
	cube-root	3.818(.217)	3.157(.104)	0.140(.003)	0.565(.012)
	PL	3.638(.219)	3.057(.107)	0.142(.003)	0.572(.015)
	quad	1.201(.076)	1.492(.057)	0.232(.009)	1.015(.023)

Table A.1: Mean square error (MSE) and uniform error (UE) (standard errors in parentheses) for the simulated example in Section A. Here cond-mean(KNN), and 2SIR+AIR denote nonparametric regression in (9), and the proposed method in (10), respectively.

where $p_a(\boldsymbol{\alpha}) = \sum_{j=1}^{p} p_a(\alpha_j)$ is the SCAD penalty, $\lambda > 0$ is a tuning parameter controlling the sparsity of the solution, and $a > 0$ is a parameter in the SCAD, c.f. (C.1). For each choice of λ, the solution $(\widehat{\boldsymbol{\alpha}}_\lambda, \widehat{\beta}_\lambda)$ of (B.1) can be efficiently computed by the local linear approximation algorithm (Zou and Li, 2008). Next, fixing K, we refit an ordinary least squares (OLS) regression with $\widehat{\boldsymbol{\theta}}^\intercal \boldsymbol{z}$ and the top K variables in $\widehat{\boldsymbol{\alpha}}_\lambda$ for each λ. Let $(\widehat{\boldsymbol{\alpha}}_{\lambda,K}, \widehat{\beta}_{\lambda,K})$ be the resulting OLS estimate. Then define

$$(\widehat{\boldsymbol{\alpha}}_K, \widehat{\beta}_K) = \underset{(\widehat{\boldsymbol{\alpha}}_{\lambda,K}, \widehat{\beta}_{\lambda,K})}{\arg\min} \ \mathrm{RSS}_2(\widehat{\boldsymbol{\alpha}}_{\lambda,K}, \widehat{\beta}_{\lambda,K})$$

as the solution to (4), where $\mathrm{RSS}_2(\widehat{\boldsymbol{\alpha}}_{\lambda,K}, \widehat{\beta}_{\lambda,K})$ is the residual sum of squares.

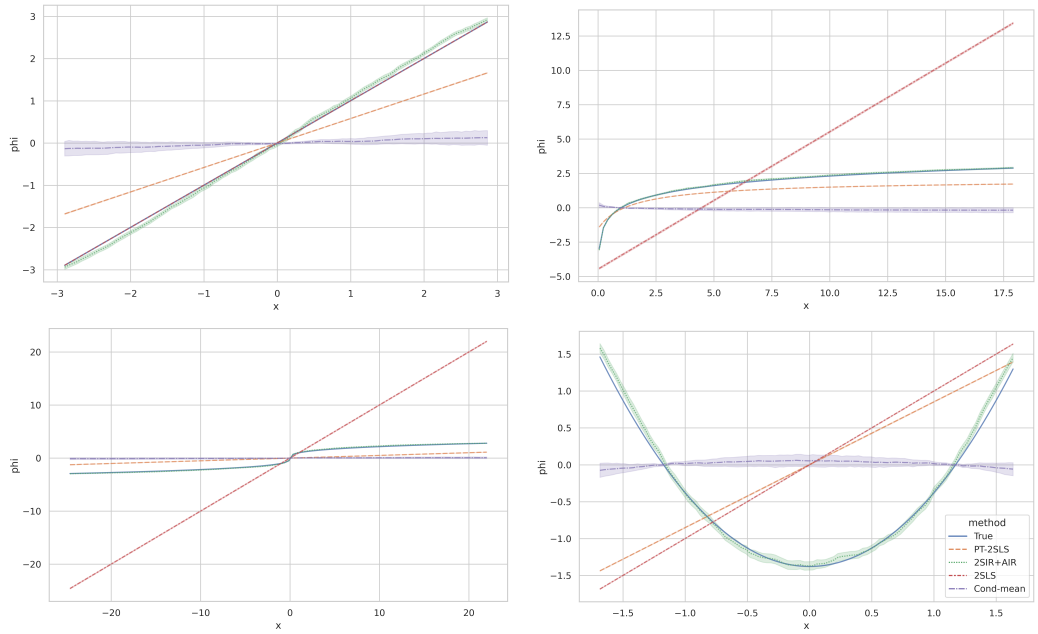

Figure A.1: Fitted transformations of the simulated example in Section A, where the true transformations are: (1,1) linear; (1,2) logarithm; (2,1) cubic root; (2,2) quadratic.

To choose the best performing K, we use BIC for tuning criteria. Specifically, define

$$\widehat{\mathrm{BIC}}(K) = \frac{\mathrm{RSS}_2(\widehat{\boldsymbol{\alpha}}_K, \widehat{\beta}_K)}{\widehat{\sigma}_e^2} + \log(n_2)(K+1),$$

where $\widehat{\sigma}_e^2 = \mathrm{RSS}_2(\widehat{\boldsymbol{\alpha}}_{\mathrm{ols}}, 0)/n_2$ is an estimate of σ_e^2 in (4) and $\widehat{\boldsymbol{\alpha}}_{\mathrm{ols}} = (\boldsymbol{Z}_2^\intercal \boldsymbol{Z}_2)^{-1}\boldsymbol{Z}_2^\intercal \boldsymbol{Y}_2$. Then we choose K that minimizes $\widehat{\mathrm{BIC}}(K)$, and use $(\widehat{\boldsymbol{\alpha}}_K, \widehat{\beta}_K)$ for the subsequent data analysis.

B.2. Stability combination of p-values

In (8), the slicing scheme is treated as fixed. Although the number of slices S has been regarded as a hyperparameter of minor importance (Li, 1991; Cook, 2009), our experiments and existing literature (Becker and Gather, 2007) suggest that the numerical results may vary greatly as S changes. Specifically, we produce p-values for significant genes in Section 4 with a different number of slices based on the proposed method. Figure B.1 clearly suggests that p-values significantly affected by the number of slices ($S = 2, 3, 5, 10$). Hence, a gap in the choice of S exists between theory and practice.

To bridge this gap, we propose to combine the tests based on different slicing schemes. Specifically, let $\mathcal{S}$ be a collection of candidate slicing schemes. We combine p-values based on different slices $S \in \mathcal{S}$ based on the Cauchy combination method (Liu and Xie, 2020):

$$p_* = 0.5 - (\arctan t_0)/\pi, \quad t_0 = \sum_{S\in\mathcal{S}} w_i \tan\Big(\big(0.5 - P_W(|W| > \widehat{T}_S)\big)\pi\Big), \tag{B.2}$$

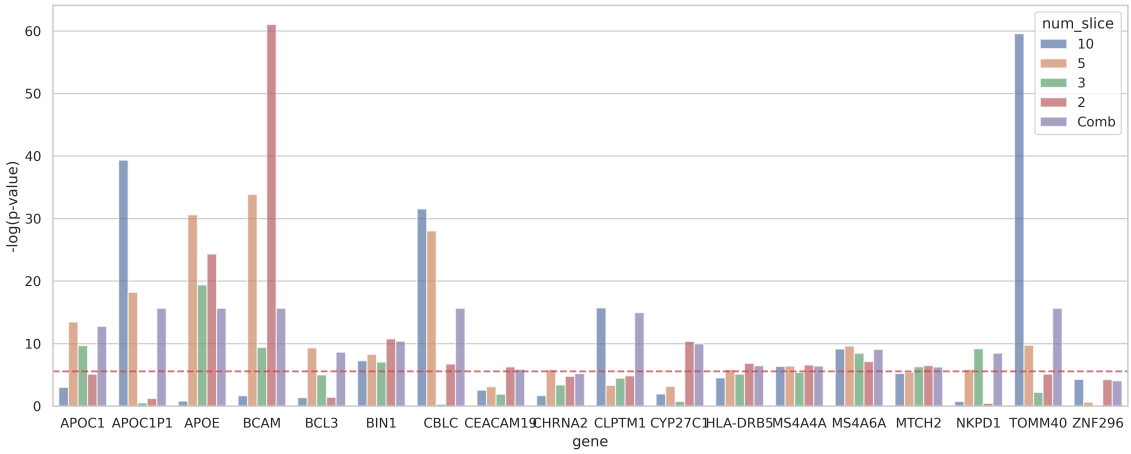

Figure B.1: The bar-plot for the negative logarithm of p-values of significant genes in Section 4 with different numbers of slices ($S = 2, 3, 5, 10$) based on the proposed method.

where the weights w_is are nonnegative and $\sum_{i=1}^{|\mathcal{S}|} w_i = 1$, $\widehat{T}_S$ is the test statistic in (8) with the subscript emphasizing its dependence on S, and $W \sim N(0,1)$ is a standard normal variable independent of the data. For illustration, we focus on a combined version of the proposed method with $w_i = 1/|\mathcal{S}|$. Note that we could apply other types of combining such as order statistics of the p-values, and corrected arithmetic and geometric means (Vovk and Wang, 2020).

B.3. Simulation results for Invalid IVs with or without correlated pleiotropy

Example 2 (Invalid IVs). In this example, we examine the proposed method with invalid IVs. Specifically, $\boldsymbol{z}_i$ is generated with $\Sigma_{ij} = \nu^{|i-j|}$. Then, x_i is generated based on the same procedure in Example 1. Finally, y_i is generated as $y_i = \beta\phi(x_i) + \boldsymbol{\alpha}^\intercal \boldsymbol{z}_i + \epsilon_i$. Here $\boldsymbol{\alpha} = (1, 1, 1, 1, 1, 0, \cdots, 0)$ indicates that the first five elements are invalid IVs. We examine four cases: (i) $\beta = 0$, (ii) $\beta = .03$, (iii) $\beta = .05$, (iv) $\beta = .10$. We construct CIs for (iii) $\beta = .05$. All empirical results are summarized in Figure B.2 (testing) and Table B.1 (CI) based on $(n, p) = (10000, 50)$, and $\nu = 0.0, 0.5$. Moreover, we further consider invalid IVs with correlated pleiotropy, where $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{\mu}$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \boldsymbol{\mu}$ where $\boldsymbol{\theta}_0$ and $\boldsymbol{\alpha}_0$ are simulated with the same procedure in Example 1, and $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_5, 0, \cdots)^\intercal$ with $\mu_j \sim N(0, 1)$. All empirical results are summarized in Figure B.3 (testing) based on $n = 10000, p = 50$, and $\nu = 0.5$.

B.4. Simulation results for categorical IVs

Example 3 (Categorical IVs). Note that the proposed method requires that the IVs follow an elliptically symmetric distribution, which is usually invalid for categorical data. Yet, in practice, a categorical IV is often involved in causal inference, such as SNP data. In this example, we examine if the proposed method can be applied to categorical IVs. Specifically, the IVs $(\boldsymbol{z}_i)_{i=1,\cdots,n}$ are generated as $\boldsymbol{z}_i = \boldsymbol{\tau}_i + \boldsymbol{\tau}_i'$ to mimic the SNP data, where $\boldsymbol{\tau}_i$ and $\boldsymbol{\tau}_i'$ are independent Bernoulli trials, each with a probability of success 0.3. Moreover, we set $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{I}_p)$ and normalize it by its norm,

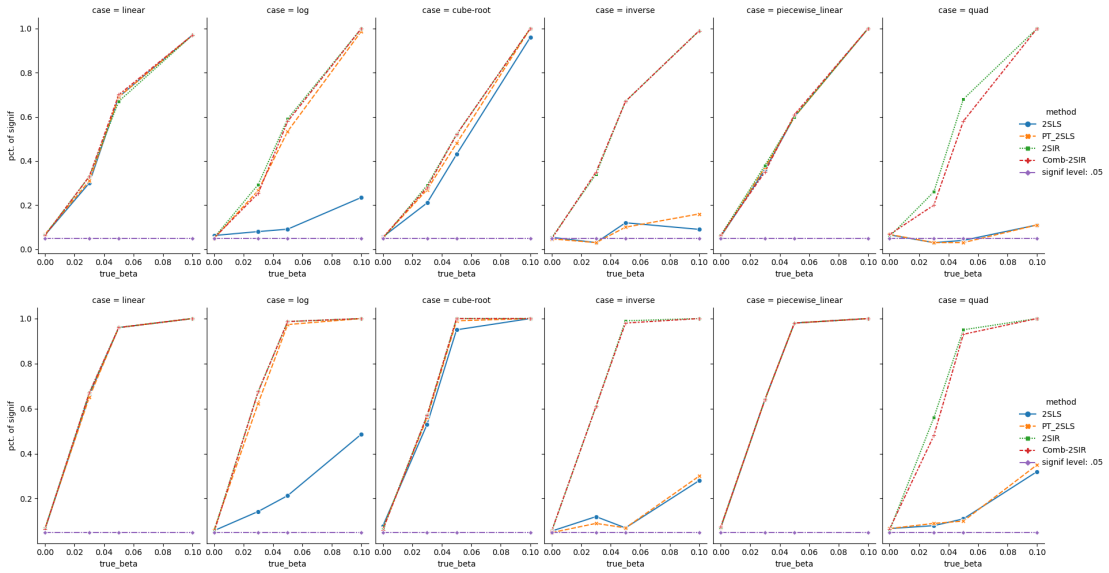

Figure B.2: Empirical Type I error ($\beta_0 = 0$) and power ($\beta = 0.05, 0.10, 0.15$) for the simulated example (invalid IVs) in Example 2. $\nu = 0.0, 0.5$ from top to bottom.

ν		*2SLS* coverage	*2SLS* length	*PT-2SLS* coverage	*PT-2SLS* length	*2SIR* (proposed) coverage	*2SIR* (proposed) length
0.0	linear	0.945	0.078	0.945	0.078	0.948	0.078
	log	0.999	79.988	0.928	0.078	0.952	0.078
	cube-root	0.965	0.190	0.972	0.079	0.949	0.079
	inverse	0.598	0.159	0.510	0.050	0.954	0.078
	PL	0.951	0.079	0.950	0.079	0.951	0.079
	quad	0.443	0.043	0.456	0.043	0.964	0.079
0.5	linear	0.951	0.050	0.951	0.050	0.945	0.050
	log	1.000	213.678	0.948	0.056	0.946	0.050
	cube-root	1.000	0.216	0.945	0.051	0.943	0.049
	inverse	0.827	0.210	0.645	0.062	0.940	0.050
	PL	0.942	0.050	0.912	0.049	0.936	0.050
	quad	0.541	0.055	0.514	0.055	0.946	0.049

Table B.1: Empirical coverage and length of the CI for in Example 2 (invalid IVs).

then x_i and y_i are generated following the same procedure in Example 1. All empirical results are summarized in Table B.2 (testing), Table B.3 (CI), and Figure 2 (boxplot).

B.5. Simulation results for weak IVs

Example 4 (Weak IVs). In this example, we examine the performance and stability of the proposed method with weak IVs. Specifically, we set $\boldsymbol{\theta} \sim N(\mathbf{0}, \boldsymbol{I}_p)$, $\theta_j = 0; j = 1, \cdots, \lfloor \pi p \rfloor$, and normalize it by its norm, then x_i and y_i are generated following the same procedure in Example 1 based on $(n = 5000, p = 50)$, and $\pi = 0.0, 0.1, 0.3$. All empirical results are summarized in Figure B.5 (testing), Table B.4 (CI).

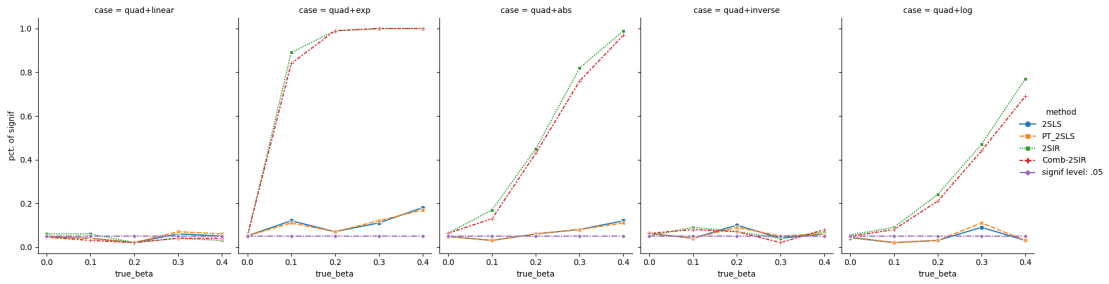

Figure B.3: Empirical Type I error ($\beta_0 = 0$) and power ($\beta = 0.05, 0.10, 0.15$) of the proposed nonlinear causal test for the simulated example (invalid IVs with correlated pleiotropy) in Example 2 of Section 3.

		2SLS		*PT-2SLS*		*2SIR* (proposed)		*Comb-2SIR* (proposed)	
(n, p)		Type I	Power	Type I	Power	Type I	Power	Type I	Power
(2000, 10)	linear	.040	(0.20, 0.37, 0.47)	.040	(0.20, 0.39, 0.51)	.048	(0.20, 0.40, 0.52)	.046	(0.18, 0.40, 0.54)
	log	.050	(0.03, 0.14, 0.14)	.058	(0.07, 0.25, 0.52)	.055	(0.06, 0.23, 0.60)	.057	(0.09, 0.22, 0.60)
	cube-root	.052	(0.08, 0.16, 0.36)	.055	(0.10, 0.32, 0.54)	.057	(0.13, 0.35, 0.53)	.054	(0.13, 0.35, 0.55)
	inverse	.050	(0.05, 0.07, 0.15)	.044	(0.03, 0.08, 0.12)	.060	(0.14, 0.24, 0.57)	.066	(0.14, 0.25, 0.58)
	PL	.058	(0.13, 0.30, 0.52)	.059	(0.15, 0.29, 0.52)	.055	(0.15, 0.28, 0.51)	.055	(0.14, 0.30, 0.52)
	quad	.051	(0.08, 0.02, 0.14)	.053	(0.08, 0.03, 0.15)	.040	(0.10, 0.23, 0.58)	.043	(0.10, 0.17, 0.56)
(2000, 50)	linear	.070	(0.11, 0.19, 0.52)	.069	(0.13, 0.21, 0.56)	.060	(0.10, 0.18, 0.59)	.062	(0.11, 0.18, 0.59)
	log	.065	(0.04, 0.08, 0.12)	.062	(0.08, 0.15, 0.28)	.061	(0.08, 0.18, 0.46)	.063	(0.06, 9.18, 0.47)
	cube-root	.059	(0.05, 0.09, 0.14)	.061	(0.04, 0.18, 0.46)	.042	(0.07, 0.21, 0.54)	.045	(0.06, 0.24, 0.49)
	inverse	.050	(0.05, 0.06, 0.06)	.055	(0.04, 0.11, 0.09)	.059	(0.08, 0.25, 0.49)	.069	(0.10, 0.21, 0.45)
	PL	.050	(0.09, 0.26, 0.50)	.053	(0.08, 0.33, 0.51)	.061	(0.08, 0.30, 0.49)	.059	(0.07, 0.33, 0.51)
	quad	.061	(0.05. 0.06, 0.06)	.062	(0.05, 0.06, 0.06)	.064	(0.13, 0.20, 0.58)	.069	(0.09, 0.16, 0.52)
(5000, 50)	linear	.058	(0.24, 0.59, 0.86)	.054	(0.23, 0.59, 0.88)	.053	(0.26, 0.62, 0.89)	.060	(0.27, 0.61, 0.88)
	log	.062	(0.06, 0.13, 0.11)	.051	(0.17, 0.46, 0.68)	.066	(0.20, 0.64, 0.86)	.068	(0.22, 0.66, 0.84)
	cube-root	.053	(0.16, 0.26, 0.31)	.056	(0.22, 0.61, 0.88)	.046	(0.24, 0.57, 0.88)	.042	(0.26, 0.58, 0.89)
	inverse	.047	(0.02, 0.09, 0.03)	.040	(0.03, 0.11, 0.05)	.056	(0.24, 0.58, 0.90)	.058	(0.22, 0.58, 0.87)
	PL	.058	(0.21, 0.52, 0.86)	.054	(0.23, 0.55, 0.88)	.058	(0.22, 0.56, 0.89)	.054	(0.22, 0.56, 0.89)
	quad	.053	(0.10, 0.09, 0.06)	.053	(0.11, 0.09, 0.06)	.043	(0.24, 0.59, 0.86)	.051	(0.22, 0.52, 0.83)
(5000, 100)	linear	.052	(0.15, 0.54, 0.85)	.049	(0.17, 0.56, 0.87)	.050	(0.14, 0.62, 0.89)	.056	(0.17, 0.60, 0.89)
	log	.044	(0.07, 0.09, 0.03)	.064	(0.16, 0.32, 0.61)	.064	(0.21, 0.57, 0.86)	.069	(0.25, 0.58, 0.86)
	cube-root	.050	(0.06, 0.14, 0.35)	.055	(0.22, 0.52, 0.85)	.047	(0.21, 0.58, 0.86)	.053	(0.26, 0.56, 0.88)
	inverse	.053	(0.07, 0.08, 0.08)	.061	(0.03, 0.07, 0.11)	.048	(0.17, 0.55, 0.88)	.063	(0.18, 0.49, 0.83)
	PL	.055	(0.21, 0.50, 0.80)	.058	(0.25, 0.56, 0.87)	.055	(0.25, 0.59, 0.88)	.060	(0.22, 0.58, 0.85)
	quad	.056	(0.07, 0.04, 0.10)	.057	(0.06, 0.04, 0.10)	.050	(0.26, 0.61, 0.87)	.051	(0.23, 0.51, 0.84)

Table B.2: Empirical Type I error and power of the proposed nonlinear causal test for the simulated example (categorical instrument variables) in Example 3 of Section 3.

B.6. Simulation results for non-additive and epistatic effects

Example 5 (Non-additive and epistatic effects). In this example, we examine the performance and stability of the proposed method under non-additive and epistatic genetic effects. First, $(\boldsymbol{z}_i)_{i=1,\cdots,n}$ are generated based on the same setting in Example 3. To incorporate the non-additive and epistatic effects, $x_i = \phi^{-1}(\boldsymbol{\theta}_a^\intercal I(\boldsymbol{z}_i = 1) + \boldsymbol{\theta}_d^\intercal I(\boldsymbol{z}_i = 2) + \sum_{(j,j')\in\mathcal{J}} \delta_{j,j'} z_{ij} z_{ij'} + w_i)$. Here, we set $\boldsymbol{\theta}_a \sim N(\mathbf{0}, \boldsymbol{I}_p)$, and $\boldsymbol{\theta}_d = \lambda\boldsymbol{\theta}_a$ presents non-additive effects when $\lambda \neq 2$. Besides, $\boldsymbol{\delta} \sim N(\mathbf{0}, 0.1\boldsymbol{I}_{|\mathcal{J}|})$ presents epistatic (i.e. interaction) effects, and $\mathcal{J}$ is a set of randomly selected pairs, where each pair is uniformly sampled. Finally, y_i is generated following the same procedure in Example 3. In

(n, p)		*2SLS* coverage	*2SLS* length	*PT-2SLS* coverage	*PT-2SLS* length	*2SIR* (proposed) coverage	*2SIR* (proposed) length
$(2000, 10)$	linear	0.939	0.179	0.940	0.179	0.980	0.198
	log	1.000	331.983	0.936	0.184	0.974	0.199
	cube-root	1.000	0.694	0.930	0.181	0.979	0.201
	inverse	0.998	0.503	0.810	0.130	0.980	0.199
	PL	0.959	0.183	0.957	0.184	0.965	0.199
	quad	0.863	0.126	0.834	0.122	0.986	0.197
$(2000, 50)$	linear	0.951	0.178	0.955	0.179	0.975	0.197
	log	0.953	370.745	0.937	0.158	0.982	0.195
	cube-root	1.000	0.610	0.954	0.166	0.980	0.198
	inverse	0.996	0.802	0.922	0.134	0.991	0.198
	PL	0.964	0.173	0.953	0.175	0.974	0.197
	quad	0.892	0.143	0.893	0.140	0.964	0.202
$(5000, 50)$	linear	0.960	0.129	0.961	0.129	0.979	0.133
	log	1.000	250.982	0.910	0.112	0.978	0.133
	cube-root	1.000	0.312	0.945	0.125	0.975	0.134
	inverse	0.968	0.461	0.781	0.085	0.980	0.133
	PL	0.956	0.129	0.961	0.130	0.972	0.135
	quad	0.783	0.086	0.785	0.085	0.957	0.135
$(10000, 50)$	linear	0.960	0.102	0.954	0.102	0.977	0.104
	log	1.000	145.694	0.964	0.101	0.976	0.104
	cube-root	1.000	0.218	0.932	0.101	0.971	0.103
	inverse	0.892	0.589	0.670	0.060	0.979	0.104
	PL	0.958	0.103	0.956	0.103	0.982	0.103
	quad	0.673	0.067	0.666	0.066	0.987	0.103

Table B.3: Empirical coverage and length of the CI for the simulated example (marginal effect inference) with categorical instrument variables in Example 3 of Section 3.

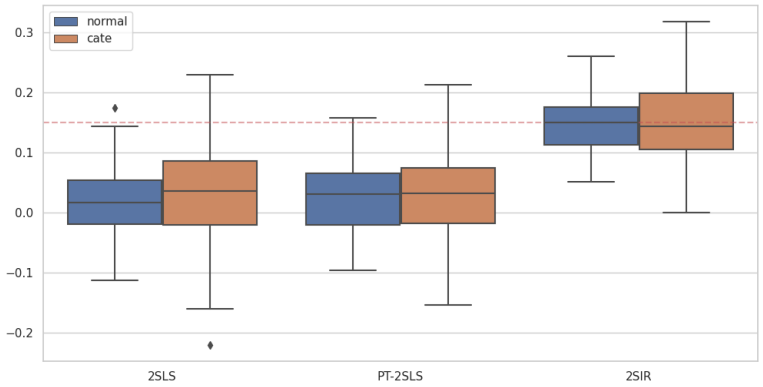

Figure B.4: The boxplot for estimated marginal causal effect β for both normal distributed and categorical instrument variables based on an "inverse" transformation function in Example 3 of Section 3 with $n = 2000, p = 10, \beta = .15$.

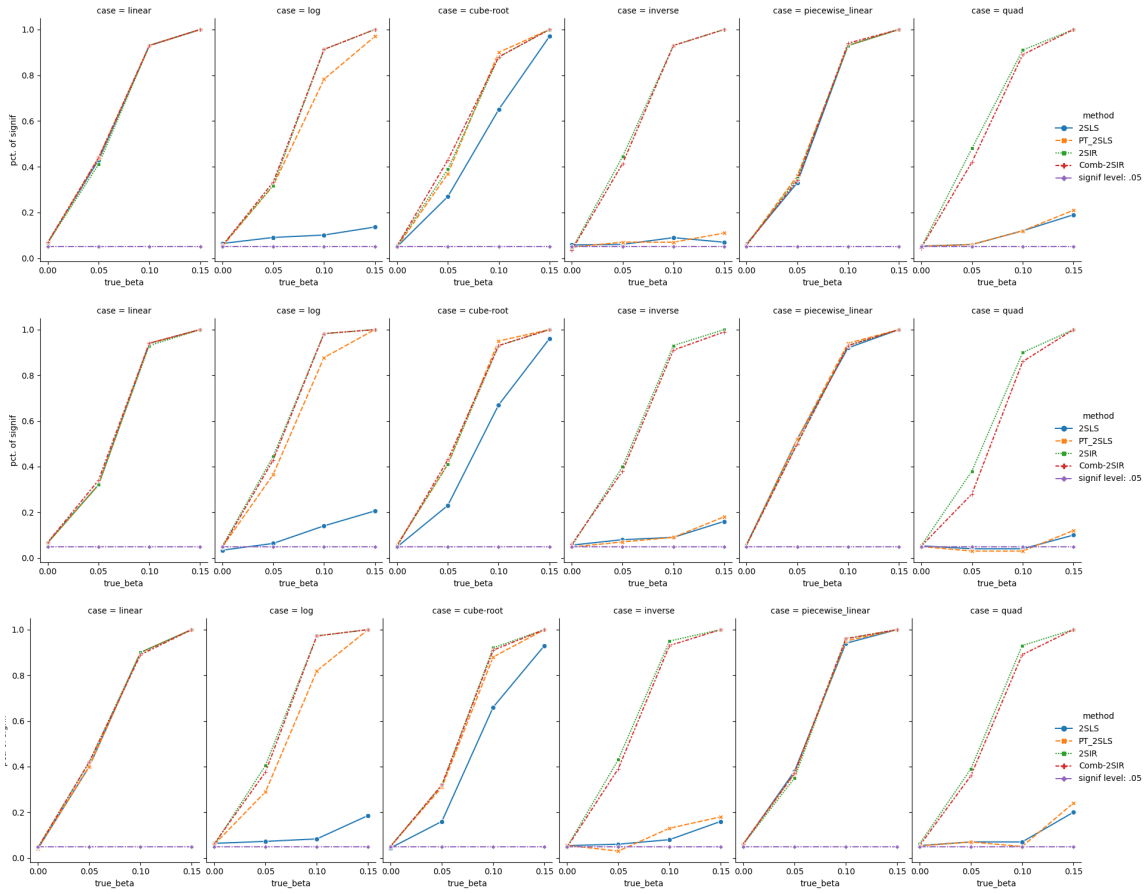

Figure B.5: Empirical Type I error (for $\beta_0 = 0$) and power (for $\beta_0 = 0.05, 0.10, 0.15$) of the proposed nonlinear causal test for the simulated example (marginal effect inference) in Example 4 (weak IVs) of Section 3, $\pi = 0.0, 0.1, 0.3$ from up to bottom. Here 2SLS, PT-2SLS, 2SIR, and Comb-2SIR denote two-stage least square, Yeo-Johnson power transformed two-stage least square, the proposed method, and the Cauchy combined proposed method, respectively.

this example, we set $n = 5000, p = 50$, $\lambda = 0.3, 0.5$ and $|\mathcal{J}| = \lfloor 0.1p \rfloor, \lfloor 0.3p \rfloor$. All empirical results are summarized in Figure B.6 (testing), Table B.5 (CI).

B.7. Simulation results for misspecified models

Example 6 (Misspecified models). We examine the performance and stability of the proposed method for misspecified models. Specifically, $(\boldsymbol{z}_i, x_i)_{i=1,\cdots,n}$ are generated with the same procedure in Example 1. In Stage 2, we consider misspecified models: $y_i = \beta\psi(x_i) + \epsilon_i$ with $\psi(x) = x$, $\psi(x) = \exp(x)$, $\psi(x) = |x|$, $\psi(x) = 1/x$, and $\psi(x) = \log(|x|)$. According to the simulation results in Example 1, we mainly consider $\phi(x) = x^2$ and $\phi(x) = 1/x$ to highlight the differences

π		*2SLS* coverage	*2SLS* length	*PT-2SLS* coverage	*PT-2SLS* length	*2SIR* (proposed) coverage	*2SIR* (proposed) length
0.0	linear	0.950	0.094	0.952	0.095	0.978	0.095
	log	1.000	95.559	1.000	0.090	0.972	0.097
	cube-root	1.000	0.215	0.999	0.095	0.982	0.097
	inverse	0.801	0.209	0.640	0.060	0.972	0.096
	PL	0.951	0.096	0.960	0.096	0.977	0.096
	quad	0.522	0.052	0.523	0.051	0.976	0.095
0.1	linear	0.952	0.096	0.952	0.096	0.972	0.095
	log	1.000	104.281	0.947	0.090	0.965	0.095
	cube-root	1.000	0.206	0.947	0.094	0.968	0.094
	inverse	0.775	0.485	0.607	0.059	0.960	0.094
	PL	0.952	0.095	0.955	0.095	0.971	0.094
	quad	0.584	0.054	0.578	0.054	0.969	0.094
0.3	linear	0.945	0.096	0.947	0.096	0.968	0.095
	log	1.000	118.202	0.936	0.091	0.955	0.095
	cube-root	1.000	0.222	0.958	0.097	0.966	0.096
	inverse	0.775	1.026	0.597	0.054	0.964	0.094
	PL	0.936	0.095	0.943	0.095	0.971	0.094
	quad	0.566	0.054	0.567	0.054	0.975	0.094

Table B.4: Empirical coverage and length of the confidence interval for the simulated example (marginal effect inference) in Example 4 (weak IVs) of Section 3.

between the proposed methods and other competitors. All empirical results are summarized in Figure B.7 (testing).

B.8. R-squared values for the estimated equation

This subsection includes the R-squared values for the estimated equation ($\mathbf{z}$-x) based on the ADNI dataset. The numerical results are summarized in the folder "app_S11_r2".

Appendix C. Supplementary results and technical proofs

C.1. Selection bias

In the real data application, we pre-screen SNPs based on multiple criteria. This subsection analyzes the potential selection bias in our procedure. To this end, consider the following situation. Suppose $(\boldsymbol{z}, x, y)$ comes from the model

$$\phi(x) = \boldsymbol{z}^\intercal \boldsymbol{\theta} + w, \qquad y = \beta\phi(x) + \boldsymbol{z}^\intercal \boldsymbol{\alpha} + \varepsilon,$$

where $\boldsymbol{z} \in \mathbb{R}^d$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_M, \mathbf{0})$, $|M| \ll d$, and the other settings remain the same as model (1). The prescreening procedure based on $(\boldsymbol{Z}_1, \boldsymbol{X}_1) \in \mathbb{R}^{n_1 \times (d+1)}$ selects a model $\widehat{M}$ with cardinality $|\widehat{M}| = p \ll d$ being fixed. Assume the prescreening procedure satisfies the sure screening property (Fan and Lv, 2008) in that $P(\widehat{M} \supseteq M) \to 1$. Moreover, $A = \{j : \alpha_j \neq 0\} \subseteq M$. For any $M' \supseteq M$ with $|M'| = p$, we have a submodel

$$\phi(x) = \boldsymbol{z}_{M'}^\intercal \boldsymbol{\theta}^{(M')} + w, \qquad y = \beta\phi(x) + \boldsymbol{z}_{M'}^\intercal \boldsymbol{\alpha}^{(M')} + \varepsilon,$$

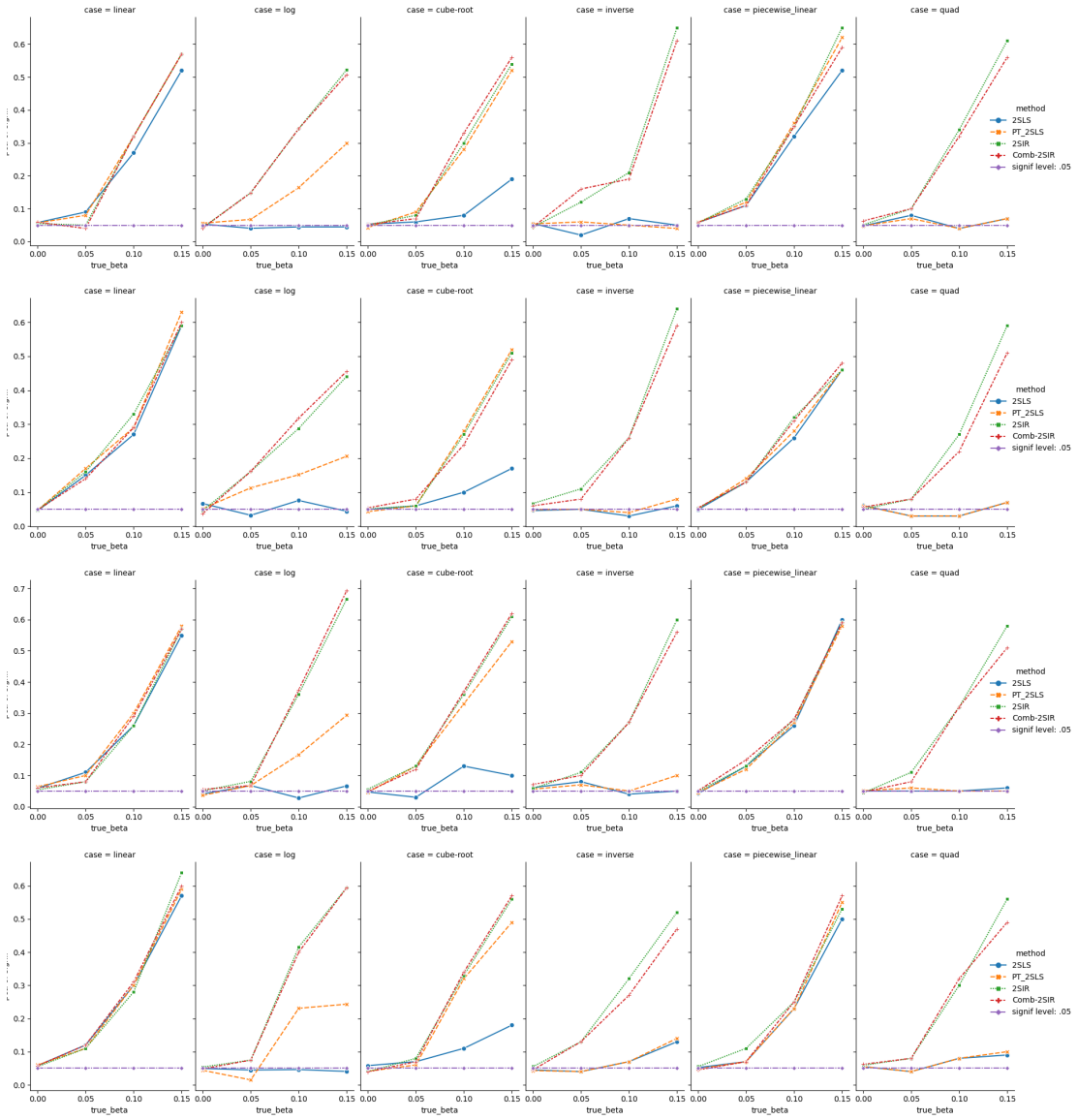

Figure B.6: Empirical Type I error (for $\beta_0 = 0$) and power (for $\beta_0 = 0.05, 0.10, 0.15$) of the proposed nonlinear causal test for the simulated example (marginal effect inference) in Example 5 (non-additive and epistatic effects) of Section 3, $(\lambda, |\mathcal{J}|) = (0.3, \lfloor 0.1p \rfloor), (0.5, \lfloor 0.1p \rfloor), (0.3, \lfloor 0.3p \rfloor), (0.5, \lfloor 0.3p \rfloor)$ from up to bottom.

where $\boldsymbol{\theta}^{(M')} = (\boldsymbol{\theta}_M, \mathbf{0}) \in \mathbb{R}^{|M'|}$ and $\boldsymbol{\alpha}^{(M')} = (\boldsymbol{\alpha}_A, \mathbf{0}) \in \mathbb{R}^{|M'|}$. Let $\widehat{\boldsymbol{\theta}}^{(M')}$ be the SIR estimator based on M'. Then by Theorem 4 of (Zhu and Ng, 1995), we have $\sqrt{n_1}(\widehat{\boldsymbol{\theta}}^{(M')} - \boldsymbol{\theta}^{(M')}) \xrightarrow{d} \boldsymbol{\xi}^{(M')}$ for a subgaussian random variable $\boldsymbol{\xi}^{(M')}$. Since there are $\binom{d}{p-|M|} \leq (ed/(p-|M|))^{p-|M|}$ possible

$(\lambda, \lvert\mathcal{J}\rvert)$		*2SLS* coverage	*2SLS* length	*PT-2SLS* coverage	*PT-2SLS* length	*2SIR* (proposed) coverage	*2SIR* (proposed) length
(1.3, 0.1p)	linear	0.945	0.115	0.942	0.116	0.992	0.123
	log	1.000	221.321	0.889	0.103	0.989	0.124
	cube-root	1.000	0.281	0.915	0.115	0.987	0.125
	inverse	0.979	0.331	0.796	0.087	0.992	0.123
	PL	0.937	0.115	0.932	0.116	0.986	0.125
	quad	0.787	0.084	0.780	0.084	0.998	0.125
(1.3, 0.3p)	linear	0.923	0.113	0.925	0.113	0.989	0.123
	log	1.000	234.423	0.892	0.103	0.992	0.124
	cube-root	1.000	0.275	0.907	0.112	0.989	0.123
	inverse	0.977	1.920	0.772	0.087	0.982	0.123
	PL	0.918	0.112	0.914	0.113	0.985	0.123
	quad	0.785	0.084	0.784	0.083	0.990	0.123
(1.5, 0.1p)	linear	0.948	0.115	0.952	0.116	0.992	0.124
	log	1.000	188.532	0.893	0.105	0.991	0.121
	cube-root	1.000	0.268	0.917	0.112	0.982	0.123
	inverse	0.977	0.334	0.801	0.088	0.989	0.124
	PL	0.944	0.115	0.945	0.116	0.991	0.124
	quad	0.776	0.082	0.773	0.082	0.988	0.124
(1.5, 0.3p)	linear	0.931	0.115	0.942	0.114	0.989	0.123
	log	1.000	210.321	0.882	0.104	0.991	0.123
	cube-root	1.000	0.281	0.942	0.114	0.995	0.124
	inverse	0.974	0.583	0.799	0.085	0.986	0.124
	PL	0.926	0.114	0.928	0.115	0.986	0.123
	quad	0.789	0.085	0.788	0.085	0.990	0.123

Table B.5: Empirical coverage and length of the CI for the simulated example (marginal effect inference) in Example 5 (non-additive and epistatic effects) of Section 3.

M', we have

$$\max_{M' \supseteq M : |M'| = p} \sqrt{n_1} |\widehat{\boldsymbol{\theta}}^{(M')} - \boldsymbol{\theta}^{(M')}| = O_p\Big(\sqrt{(p - |M|)\log(d)}\Big).$$

Thus, $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{(\widehat{M})}$ is consistent provided that $n_1 \gg (p - |M|)\log(d)$. It follows that $\widehat{\beta}$ is also consistent in this situation. In view of Theorem 1, $\sqrt{n_2}\widehat{\beta} \xrightarrow{d} |N(0, (\boldsymbol{\theta}^\intercal \widetilde{\boldsymbol{\Sigma}} \boldsymbol{\theta})^{-1}\sigma_e^2)|$ when $\beta = 0$. Consequently, the test (8) of $H_0 : \beta = 0$ remains valid after a sure screening procedure. To conclude, our procedure seems largely immune to the potential selection bias provided that the sample size $n_1 \gg (p - |M|)\log(d)$ and a sure screening method is used.

C.2. Regularity conditions and supplementary results

We impose the following regularity conditions for 2SIR and AIR. In particular, Condition C.1 is used to establish the asymptotic distribution of SIR estimate $\widehat{\boldsymbol{\theta}}$, Conditions C.1 and C.2 are used to derive the asymptotic properties of 2SIR estimate $\widehat{\beta}$, and Conditions C.1 and C.3 are used to quantify the convergence rate of AIR estimate $\widehat{\phi}$.

Condition C.1 *Assume the following conditions for sliced inverse regression.*

(i) $\mathrm{E}(\boldsymbol{z} \mid \boldsymbol{z}^\intercal\boldsymbol{\theta})$ *is linear in* $\boldsymbol{z}^\intercal\boldsymbol{\theta}$*;*

(ii) $c_- \le \lambda_{\min}(\boldsymbol{\Sigma}) \le \lambda_{\max}(\boldsymbol{\Sigma}) \le c_+$*, where* $\boldsymbol{\Sigma} = \mathrm{E}\,\boldsymbol{z}\boldsymbol{z}^\intercal$*;*

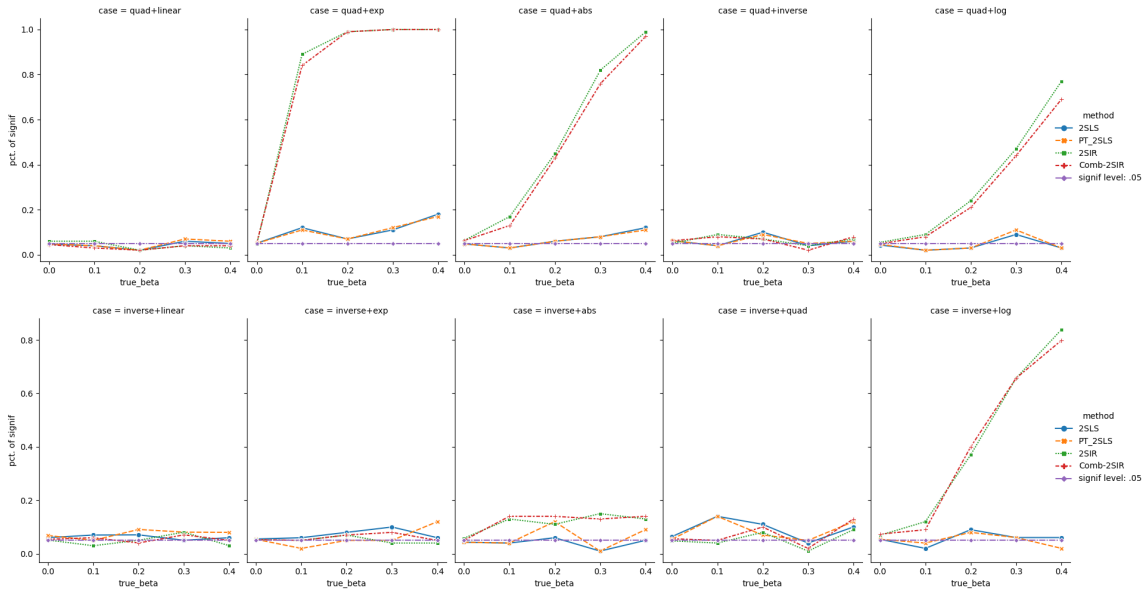

Figure B.7: Empirical Type I error ($\beta_0 = 0$) and power ($\beta_0 = 0.1, 0.2, 0.3, 0.4$) of the proposed nonlinear causal test for the simulated example with misspecified causal transformation in Example 6 of Section 3. Here $\phi(x) = x^2$ and $\phi(x) = 1/x$ are specified for two rows, respectively; and $\psi(x) = x$, $\psi(x) = e^x$, $\psi(x) = |x|$, $\psi(x) = 1/x$, $\psi(x) = \log(|x|)$ are specified for five columns.

(iii) $\mathrm{E}\,\|\boldsymbol{z}\|^4 < \infty$*;*

(iv) $\mathrm{E}(\boldsymbol{z} \mid x)$ *has a total variation of order* $1/4$ *in that*

$$\lim_{n_1 \to \infty} \frac{1}{n_1^{1/4}} \sup_{\Pi_{n_1}(D)} \sum_{i=1}^{n_1 - 1} \| \mathrm{E}(\boldsymbol{z} \mid x^*_{(i+1)}) - \mathrm{E}(\boldsymbol{z} \mid x^*_{(i)}) \| = 0,$$

where $\Pi_{n_1}(D)$ *is the collection of all* n_1*-point partitions,* $-D \le x^*_{(1)} \le \cdots \le x^*_{(n_1)} \le D$ *of the interval* $[-D, D]$*,* $D > 0$ *and* $\|\cdot\|$ *is the Euclidean norm;*

(v) There exist a nondecreasing real-valued function M *and a real number* $D_0 > 0$ *such that for any two points* $x_1, x_2 < -D_0$ *or* $x_1, x_2 > D_0$*,*

$$\| \mathrm{E}(\boldsymbol{z} \mid x_1) - \mathrm{E}(\boldsymbol{z} \mid x_2) \| \le |M(x_1) - M(x_2)|,$$

and $M^4(t)P(x > t) \to 0$ *as* $t \to \infty$*, as* $n_1 \to \infty$*;*

(vi) Let $\mathrm{Cov}(\boldsymbol{u} \mid x)$ *has a total variation of order* 1 *in that*

$$\lim_{n_1 \to \infty} \frac{1}{n_1} \sup_{\Pi_{n_1}(D)} \sum_{i=1}^{n_1 - 1} \| \mathrm{Cov}(\boldsymbol{u} \mid x^*_{(i+1)}) - \mathrm{Cov}(\boldsymbol{u} \mid x^*_{(i)}) \|_F = 0,$$

where $\boldsymbol{u} = \boldsymbol{z} - \mathrm{E}(\boldsymbol{z} \mid x)$ *and* $\|\cdot\|_F$ *is the Frobenius norm.*

Condition C.1 is common in the sufficient dimension reduction literature (Zhu and Ng, 1995; Zhu et al., 2006). Note that (i) and (ii) impose distributional assumptions on $\boldsymbol{z}$, where (i) is equivalent to that $\boldsymbol{z}$ has an elliptically symmetric distribution (Cook and Weisberg, 1991). However, it can be approximately extended to categorical IVs as indicated in (Hall and Li, 1993). Moreover, the numerical performance in Example 3 of Section 3 also suggests that the proposed method can apply to categorical IVs. Condition C.1 (iii)-(vi) are used to derive the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$; see (Zhu and Ng, 1995) for details. Under Condition 1, we have $n_1^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{\xi}$, where the distribution of $\boldsymbol{\xi}$ is given in Theorem 4 of (Zhu and Ng, 1995).

For estimating β, we aim to solve a sparse regression problem in (4) of the main text. In (4), $\|\cdot\|_0$ penalty is used. For theoretical analysis, we also consider its surrogates SCAD, TLP, and MCP, defined as follows:

- (SCAD)

$$p_a(t) = \begin{cases} 4|t|/3a & |t| \le a/2, \\ 1 - 4(|t| - a)^2/3a^2 & a/2 < |t| \le a \\ 1 & |t| > a, \end{cases} \tag{C.1}$$

- (TLP)

$$p_a(t) = \begin{cases} |t|/a & |t| \le a, \\ 1 & |t| > a, \end{cases} \tag{C.2}$$

- (MCP)

$$p_a(t) = \begin{cases} 2|t|/a - |t|^2/a^2 & |t| \le a, \\ 1 & |t| > a, \end{cases} \tag{C.3}$$

where $a > 0$ is a hyperparameter.

Condition C.2 *Assume the following conditions are satisfied.*
(i) $|A| < p/2$, *where* $A = \{j : \alpha_j \neq 0\}$;
(ii) $\|\boldsymbol{\alpha}\|_{\min} = \min_{j \in A} |\alpha_j| \ge 32\sigma_e \sqrt{c_+} c_-^{-1} \sqrt{\log(n)/n}$;
(iii) $n_2/n_1 \to r \in (0, \infty)$.
(iv) $0 < a < \sqrt{c_+ \log(n)/(2p\lambda_{\max}(\boldsymbol{Z}_2^{\mathrm{T}} \boldsymbol{Z}_2))}$ *when SCAD, TLP, or MCP is used.*

Condition C.2 (i) is an assumption for identifiability of β, cf. Corollary 1 of (Kang et al., 2016b), while (ii) is nearly necessary for the consistent selection of invalid instruments (Wainwright, 2009). Condition C.2 (iii) is a common assumption in two-sample inference (Pacini and Windmeijer, 2016).

For the estimation of nonlinear transformation ϕ, we impose the following condition.

Condition C.3 *Assume* $\widehat{m}$ *satisfies the following properties.*
(i) $\mathrm{E}\,\|\widehat{m} - m\|_\infty \le c_1 n_1^{-\kappa_1}$, *where* $\|m\|_\infty = \sup_{x \in \mathcal{X}} |m(x)|$ *and* $\kappa_1 > 0$;
(ii) $\mathrm{E}\, n_1^{-1} \sum_{i=1}^{n_1} |\widehat{m}(x_{1i}) - m(x_{1i})|^2 \le c_2 n_1^{-\kappa_2}$, *where* $\kappa_2 > 0$.

In Condition C.3, (i) specifies the local estimation quality via the sup-norm convergence rate over the treatment region of interest $\mathcal{X}$, while (ii) specifies the global estimation quality via the convergence rate in the empirical L_2-norm. The convergence results of various nonparametric regressions have been extensively studied; see (Tsybakov, 2008) for an overview.

Theorem 6 presents the convergence rate for estimating nonlinear transformation $\phi(\cdot)$ and nonlinear causal effect $\beta\phi(\cdot)$.

Theorem 6 *Assume Conditions C.1 and C.3 in Section C.2 and conditions in Proposition 1. Then* $\|\widehat{\phi} - \phi\|_\infty \leq O_p(\max(\sqrt{p/n_1}, n_1^{-\min(\kappa_1,\kappa_2)}))$. *If in addition Condition C.2 in Section C.2 holds, then* $\|\widehat{\beta}\widehat{\phi} - \beta\phi\|_\infty = O_p(\max(\sqrt{p/n_1}, \sqrt{1/n_2}, n_1^{-\min(\kappa_1,\kappa_2)}))$, *where* $\|\cdot\|_\infty$ *is sup-norm given in Condition C.3, and* $\kappa_1, \kappa_2 > 0$ *are convergence rates of* $\widehat{m}$ *in (5).*

Theorem 6 shows that the convergence rate of $\widehat{\beta}\widehat{\phi}$ is determined by the slowest rate of estimating $\boldsymbol{\theta}$, β, and $m(\cdot)$. Note that the estimation of $\boldsymbol{\theta}$ and β possesses a parametric root-n rate. Hence, the overall convergence rate $\|\widehat{\beta}\widehat{\phi} - \beta\phi\|_\infty$ is usually determined by that of the nonparametric function estimation.

C.3. Technical proofs

Proof [Proof of Proposition 1] Note that $\mathrm{E}(\boldsymbol{z}^\intercal\boldsymbol{\theta} \mid x) = \mathrm{E}(\boldsymbol{z}^\intercal\boldsymbol{\theta} \mid \phi(x)) = \mathrm{E}(\boldsymbol{z}^\intercal\boldsymbol{\theta} \mid \boldsymbol{z}^\intercal\boldsymbol{\theta} + w)$. By the property of elliptical symmetry, $\mathrm{E}(\boldsymbol{z}^\intercal\boldsymbol{\theta} \mid \boldsymbol{z}^\intercal\boldsymbol{\theta} + w) = \rho(\boldsymbol{z}^\intercal\boldsymbol{\theta} + w) = \rho\phi(x)$. ■

Lemma 7 (Theorem 1 of (Zhu and Ng, 1995)) *Assume Condition C.1 is satisfied, then we have* $n_1^{-1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{\xi}$, *where the distribution of* $\boldsymbol{\xi}$ *is given in Theorem 1 of (Zhu and Ng, 1995).*

Lemma 8 *Under Condition C.2, if* $K = |A|$, *then* $P(\widehat{A} \neq A) \leq 4n_2^{-3}$, *where* $\widehat{A} = \{j : \widehat{\alpha}_j \neq 0\}$.

Proof [Proof of Lemma 8] Denote $\widehat{\boldsymbol{X}}_2 = (\boldsymbol{z}_{21}^\intercal\widehat{\boldsymbol{\theta}}, \ldots, \boldsymbol{z}_{2n_2}^\intercal\widehat{\boldsymbol{\theta}})^\intercal$, and let $\widetilde{\boldsymbol{Z}} = (\boldsymbol{Z}_2, \widehat{\boldsymbol{X}}_2) \in \mathbb{R}^{n_2\times(p+1)}$ be the augmented data matrix and $\widehat{\boldsymbol{\gamma}}^\circ = (\widehat{\boldsymbol{\alpha}}^\circ, \widehat{\beta}^\circ)$ be the oracle estimator. Let $B = \{j : \widehat{\gamma}_j^\circ \neq 0\} = A \cup \{p+1\}$ and $\widehat{B} = \{j : \widehat{\gamma}_j \neq 0\} = \widehat{A} \cup \{p+1\}$.

First, suppose $\|\cdot\|_0$ penalty is used in (4). Since $\widehat{\boldsymbol{\gamma}}$ is the solution of (4), we have $n_2^{-1}\|\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\gamma}}\|_2^2 \leq n_2^{-1}\|\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\gamma}}^\circ\|_2^2$, which, after rearrangement, yields that

$$\frac{1}{n_2}\|\widetilde{\boldsymbol{Z}}(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^\circ)\|_2^2 \leq \frac{2}{n_2}\widehat{\boldsymbol{e}}^\intercal\widetilde{\boldsymbol{Z}}(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^\circ),$$

where $\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\gamma}}^\circ$ is the residual vector of the oracle estimator. By the first-order optimality condition of the oracle estimator $\widehat{\boldsymbol{\gamma}}^\circ$, we have $\widehat{\boldsymbol{e}}^\intercal\widetilde{\boldsymbol{Z}}_B = \boldsymbol{0}$. Moreover, we have $\widehat{\boldsymbol{\gamma}}_{(\widehat{B}\cup B)^c} - \widehat{\boldsymbol{\gamma}}^\circ_{(\widehat{B}\cup B)^c} = \boldsymbol{0}$. Hence, we have

$$\begin{aligned}\frac{1}{n_2}\|\widetilde{\boldsymbol{Z}}_{\widehat{B}\cup B}(\widehat{\boldsymbol{\gamma}}_{\widehat{B}\cup B} - \widehat{\boldsymbol{\gamma}}^\circ_{\widehat{B}\cup B})\|_2^2 &\leq \frac{2}{n_2}\widehat{\boldsymbol{e}}^\intercal\widetilde{\boldsymbol{Z}}_{\widehat{B}\setminus B}(\widehat{\boldsymbol{\gamma}}_{\widehat{B}\setminus B} - \widehat{\boldsymbol{\gamma}}^\circ_{\widehat{B}\setminus B}) \\ &\leq \frac{2}{n_2}\|\widehat{\boldsymbol{e}}^\intercal\widetilde{\boldsymbol{Z}}_{\widehat{B}\setminus B}\|_2\|\widehat{\boldsymbol{\gamma}}_{\widehat{B}\setminus B} - \widehat{\boldsymbol{\gamma}}^\circ_{\widehat{B}\setminus B}\|_2.\end{aligned} \tag{C.4}$$

Further,

$$\frac{1}{n_2}\|\widetilde{\boldsymbol{Z}}_{\widehat{B}\cup B}(\widehat{\gamma}_{\widehat{B}\cup B}-\widehat{\gamma}^{\circ}_{\widehat{B}\cup B})\|_2^2 \geq \lambda_{\min}(n_2^{-1}\widetilde{\boldsymbol{Z}}^{\intercal}_{\widehat{B}\cup B}\widetilde{\boldsymbol{Z}}_{\widehat{B}\cup B})\|\widehat{\gamma}_{\widehat{B}\cup B}-\widehat{\gamma}^{\circ}_{\widehat{B}\cup B}\|_2^2.$$

Note that $\|\widehat{\gamma}_{\widehat{B}\cup B}-\widehat{\gamma}^{\circ}_{\widehat{B}\cup B}\|_2 \geq \sqrt{|\widehat{B}\setminus B|}\|\boldsymbol{\alpha}\|_{\min}$ and $\|\widehat{\gamma}_{\widehat{B}\cup B}-\widehat{\gamma}^{\circ}_{\widehat{B}\cup B}\|_2 \geq \|\widehat{\gamma}_{\widehat{B}\setminus B}-\widehat{\gamma}^{\circ}_{\widehat{B}\setminus B}\|_2$. Combining the above results, we obtain

$$\lambda_{\min}(n_2^{-1}\widetilde{\boldsymbol{Z}}^{\intercal}_{\widehat{B}\cup B}\widetilde{\boldsymbol{Z}}_{\widehat{B}\cup B})\sqrt{|\widehat{B}\setminus B|}\|\boldsymbol{\alpha}\|_{\min} \leq \frac{2}{n}\sup_{\{S:|S\setminus B|=|\widehat{B}\setminus B|\}}\|\widehat{\boldsymbol{e}}^{\intercal}\widetilde{\boldsymbol{Z}}_{S\setminus B}\|_2.$$

Now, let $\mathcal{E}=\mathcal{E}_1\cap\mathcal{E}_2$, where

$$\mathcal{E}_1=\left\{\lambda_{\min}(n^{-1}\widetilde{\boldsymbol{Z}}^{\intercal}_{\widehat{B}\cup B}\widetilde{\boldsymbol{Z}}_{\widehat{B}\cup B})>\frac{c_-}{2}\right\},$$
$$\mathcal{E}_2=\left\{\sup_{\{S:|S\setminus B|=k\}}\|\widehat{\boldsymbol{e}}^{\intercal}\widetilde{\boldsymbol{Z}}_{S\setminus B}\|_2\leq 4\sqrt{c_+}\sigma_e\sqrt{k\log(n)/n},\ 1\leq k\leq|B|\right\}.$$

Then on event $\mathcal{E}$, we have $2^{-1}c_-\sqrt{|\widehat{B}\setminus B|}\|\boldsymbol{\alpha}\|_{\min} < 8\sigma_e\sqrt{c_+}\sqrt{|\widehat{B}\setminus B|}\sqrt{\log(n)/n}$. However, by Condition C.2 (iii), we have $\|\boldsymbol{\alpha}\|_{\min}\geq 32\sigma_e\sqrt{c_+}c_-^{-1}\sqrt{\log(n)/n}$. This implies that $|\widehat{A}\setminus A|=|\widehat{B}\setminus B|=0$, and hence $\widehat{A}=A$ on event $\mathcal{E}$.

Next, suppose a surrogate penalty (SCAD, TLP, or MCP) is used in (4). Let $\widehat{A}_1^*=\{j:|\widehat{\alpha}_j|>a\}$ and $\widehat{A}_2^*=\{j:|\widehat{\alpha}_j|\leq a\}$. Then (C.4) needs a modification,

$$\begin{aligned}\frac{1}{n_2}\|\widetilde{\boldsymbol{Z}}_{\widehat{B}\cup B}(\widehat{\gamma}_{\widehat{B}\cup B}-\widehat{\gamma}^{\circ}_{\widehat{B}\cup B})\|_2^2 &\leq \frac{2}{n_2}\widehat{\boldsymbol{e}}^{\intercal}\widetilde{\boldsymbol{Z}}_{\widehat{A}_1^*\setminus A}(\widehat{\gamma}_{\widehat{A}_1^*\setminus A}-\widehat{\gamma}^{\circ}_{\widehat{A}_1^*\setminus A})\\ &\quad+\frac{2}{n_2}\widehat{\boldsymbol{e}}^{\intercal}\widetilde{\boldsymbol{Z}}_{\widehat{A}_2^*\setminus A}(\widehat{\gamma}_{\widehat{A}_2^*\setminus A}-\widehat{\gamma}^{\circ}_{\widehat{A}_2^*\setminus A})\\ &\leq \frac{2}{n_2}\|\widehat{\boldsymbol{e}}^{\intercal}\widetilde{\boldsymbol{Z}}_{\widehat{A}_1^*\setminus A}\|_2\|\widehat{\gamma}_{\widehat{A}_1^*\setminus A}-\widehat{\gamma}^{\circ}_{\widehat{A}_1^*\setminus A}\|_2\\ &\quad+\frac{2}{n_2}\|\widehat{\boldsymbol{e}}\|_2\|\widetilde{\boldsymbol{Z}}_{\widehat{A}_2^*\setminus A}(\widehat{\gamma}_{\widehat{A}_2^*\setminus A}-\widehat{\gamma}^{\circ}_{\widehat{A}_2^*\setminus A})\|_2\\ &\leq \frac{2}{n_2}\|\widehat{\boldsymbol{e}}^{\intercal}\widetilde{\boldsymbol{Z}}_{\widehat{A}_1^*\setminus A}\|_2\|\widehat{\gamma}_{\widehat{A}_1^*\setminus A}-\widehat{\gamma}^{\circ}_{\widehat{A}_1^*\setminus A}\|_2\\ &\quad+\frac{3\sigma_e}{\sqrt{n_2}}\sqrt{p\lambda_{\max}(\boldsymbol{Z}^{\intercal}\boldsymbol{Z})}a.\end{aligned}\tag{C.5}$$

We have $\|\widehat{\gamma}_{\widehat{B}\cup B}-\widehat{\gamma}^{\circ}_{\widehat{B}\cup B}\|_2\geq\sqrt{|\widehat{A}_1^*\setminus A|}\|\boldsymbol{\alpha}\|_{\min}$ and $\|\widehat{\gamma}_{\widehat{B}\cup B}-\widehat{\gamma}^{\circ}_{\widehat{B}\cup B}\|_2\geq\|\widehat{\gamma}_{\widehat{A}_1^*\setminus A}-\widehat{\gamma}^{\circ}_{\widehat{A}_1^*\setminus A}\|_2$. Thus,

$$\begin{aligned}\lambda_{\min}(n_2^{-1}\widetilde{\boldsymbol{Z}}^{\intercal}_{\widehat{B}\cup B}\widetilde{\boldsymbol{Z}}_{\widehat{B}\cup B})\sqrt{|\widehat{A}_1^*\setminus A|}\|\boldsymbol{\alpha}\|_{\min} &\leq \frac{2}{n}\sup_{\{S:|S\setminus A|=|\widehat{A}_1^*\setminus A|\}}\|\widehat{\boldsymbol{e}}^{\intercal}\widetilde{\boldsymbol{Z}}_{S\setminus B}\|_2\\ &\quad+\frac{3\sigma_e}{\sqrt{n_2}}\sqrt{p\lambda_{\max}(\boldsymbol{Z}^{\intercal}\boldsymbol{Z})}a\\ &\leq \frac{2}{n}\sup_{\{S:|S\setminus A|=|\widehat{A}_1^*\setminus A|\}}\|\widehat{\boldsymbol{e}}^{\intercal}\widetilde{\boldsymbol{Z}}_{S\setminus B}\|_2+\sqrt{c_+}\sigma_e\sqrt{\frac{\log(n)}{n}},\end{aligned}$$

where the second inequality follows from Condition C.2 (iv). Similarly, on event $\mathcal{E}$, we have $2^{-1}c_-\sqrt{|\widehat{A}_1^* \setminus A|}\|\boldsymbol{\alpha}\|_{\min} < 8\sigma_e\sqrt{c_+}(\sqrt{|\widehat{A}_1^* \setminus A|}+1)\sqrt{\log(n)/n}$.

However, $\|\boldsymbol{\alpha}\|_{\min} \geq 32\sigma_e\sqrt{c_+}c_-^{-1}\sqrt{\log(n)/n}$. This implies that $|\widehat{A}_1^* \setminus A| = 0$, and hence $\widehat{A} = A$ on event $\mathcal{E}$.

Finally, note that $P(\widehat{A} \neq A) \leq P(\mathcal{E}^c) \leq P(\mathcal{E}_1^c) + P(\mathcal{E}_2^c)$, where the Gaussian tail bounds yields that

$$P(\mathcal{E}_1^c) \leq n^{-3}, \quad P(\mathcal{E}_2^c) \leq \sum_{k=1}^{|A|} 2\exp(-3k\log(n)) \leq 3n^{-3}.$$

The proof is completed. ■

Proof [Proof of Theorem 1] By Lemma 8, it suffices to consider the event $\{\widehat{A} = A\}$. Denote the oracle estimator by $(\widehat{\beta}^\circ, \widehat{\boldsymbol{\alpha}}_A^\circ)$, namely the OLS estimator with A known. Then

$$\begin{pmatrix} \widehat{\beta}^\circ \\ \widehat{\boldsymbol{\alpha}}_A^\circ \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{X}}^\intercal\widehat{\boldsymbol{X}} & \widehat{\boldsymbol{X}}^\intercal\boldsymbol{Z}_A \\ \boldsymbol{Z}_A^\intercal\widehat{\boldsymbol{X}} & \boldsymbol{Z}_A^\intercal\boldsymbol{Z}_A \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\boldsymbol{X}}^\intercal\boldsymbol{Y} \\ \boldsymbol{Z}_A^\intercal\boldsymbol{Y} \end{pmatrix}.$$

It follows from matrix algebra that

$$\begin{aligned} \widehat{\beta}^\circ &= \widehat{\Omega}_X\widehat{\boldsymbol{X}}^\intercal\boldsymbol{Y} - \widehat{\Omega}_X\widehat{\boldsymbol{X}}^\intercal\boldsymbol{Z}_A(\boldsymbol{Z}_A^\intercal\boldsymbol{Z}_A)^{-1}\boldsymbol{Z}_A^\intercal\boldsymbol{Y}, \\ \widehat{\Omega}_X &= (\widehat{\boldsymbol{X}}^\intercal\widehat{\boldsymbol{X}} - \widehat{\boldsymbol{X}}^\intercal\boldsymbol{Z}_A(\boldsymbol{Z}_A^\intercal\boldsymbol{Z}_A)^{-1}\boldsymbol{Z}_A^\intercal\widehat{\boldsymbol{X}})^{-1}. \end{aligned}$$

By Lemma 7, $\sqrt{n_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{\xi}$. Then by direct calculation,

$$\begin{aligned} &\sqrt{n_2}(\widehat{\beta}^\circ - \beta) \\ =&\sqrt{n_2}\Omega_X^{-1}\boldsymbol{\theta}^\intercal\boldsymbol{Z}^\intercal(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}_A})\boldsymbol{e} - \sqrt{r}\beta\Omega_X^{-1}(\boldsymbol{\theta}^\intercal\boldsymbol{Z}^\intercal\boldsymbol{Z} - \boldsymbol{\theta}^\intercal\boldsymbol{Z}^\intercal\boldsymbol{Z}_A(\boldsymbol{Z}_A^\intercal\boldsymbol{Z}_A)^{-1}\boldsymbol{Z}_A^\intercal\boldsymbol{Z})\boldsymbol{\xi} + o_p(1) \\ =&\zeta - \eta + o_p(1), \end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{Z}_A} = \boldsymbol{Z}_A(\boldsymbol{Z}_A^\intercal\boldsymbol{Z}_A)^{-1}\boldsymbol{Z}_A^\intercal$. Since two samples are independent, we have $\zeta \perp\!\!\!\perp \eta$. Finally, note that by Lemma 8, $P(\widehat{\beta} = |\widehat{\beta}^\circ|) \to 1$. Hence, we have $n_2^{1/2}(\widehat{\beta} - \beta) = |n_2^{1/2}\beta + \zeta - \eta| - n_2^{1/2}\beta + o_p(1)$, which completes the proof. ■

Proof [Proof of Corollaries 1 and 2] The desired results follow immediately from Theorem 1. ■

Proof [Proof of Theorem 6] Let

$$a = n_1^{-1}\sum_{i=1}^{n_1}(\boldsymbol{z}_{1i}^\intercal\widehat{\boldsymbol{\theta}})^2, \qquad \widetilde{a} = n_1^{-1}\sum_{i=1}^{n_1}(\boldsymbol{z}_{1i}^\intercal\boldsymbol{\theta})^2,$$

$$b = n_1^{-1}\sum_{i=1}^{n_1}\widehat{m}(x_{1i})\boldsymbol{z}_{1i}^\intercal\widehat{\boldsymbol{\theta}}, \qquad \widetilde{b} = n_1^{-1}\sum_{i=1}^{n_1}m_0(x_{1i})\boldsymbol{z}_{1i}^\intercal\boldsymbol{\theta}.$$

Let $\widehat{\boldsymbol{\Sigma}} = n_1^{-1}\sum_{i=1}^{n_1} \boldsymbol{z}_{1i}\boldsymbol{z}_{1i}^\intercal$ and note that $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \le c_1\sqrt{p/n_1}$. Let $\widehat{C}_{zw} = n_1^{-1}\sum_{i=1}^{n_1} \boldsymbol{z}_{1i}^\intercal\boldsymbol{\theta} w_{1i}$ and note that $|\widehat{C}_{zw}| \le c_3\sqrt{1/n_1}$. Then we have

$$
\begin{aligned}
|a - \widetilde{a}| &= |(\widehat{\boldsymbol{\theta}} + \boldsymbol{\theta})^\intercal\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)| \le c_2\sqrt{p/n_1},\\
|\widetilde{a} - \boldsymbol{\theta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\theta}| &= |\boldsymbol{\theta}^\intercal(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\theta}| \le c_1\sqrt{p/n_1},\\
|b - \widetilde{b}| &= |n_1^{-1}\sum_{i=1}^{n_1}(\widehat{m}(x_{1i}) - m(x_{1i}))\boldsymbol{z}_{1i}^\intercal\widehat{\boldsymbol{\theta}} + n_1^{-1}\sum_{i=1}^{n_1} m(x_{1i})\boldsymbol{z}_{1i}^\intercal(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)|\\
&\le \sqrt{n_1^{-1}\sum_{i=1}^{n_1}(\widehat{m}(x_{1i}) - m(x_{1i}))^2}\sqrt{\widehat{\boldsymbol{\theta}}^\intercal\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\theta}}} + C\sqrt{p/n_1}\\
&\le cn_1^{-\kappa_2} + C\sqrt{p/n_1},\\
|\widetilde{b} - \rho^{-1}\boldsymbol{\theta}^\intercal\boldsymbol{\Sigma}\boldsymbol{\theta}| &= |\rho^{-1}\boldsymbol{\theta}^\intercal(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\theta} + \widehat{C}_{zw}| \le (c_1 + c_3)\sqrt{p/n_1}.
\end{aligned}
$$

Thus,

$$
|\widehat{\rho} - \rho| \le \left|\frac{a}{b} - \frac{\widetilde{a}}{\widetilde{b}}\right| + \left|\frac{\widetilde{a}}{\widetilde{b}} - \rho\right| \le c_4\max(\sqrt{p/n_1}, n_1^{-\kappa_2}).
$$

Taken together, we have

$$
\|\widehat{\phi} - \phi\|_\infty \le |\widehat{\rho} - \rho|\|\widehat{m}\|_\infty + |\rho|\|\widehat{m} - m\|_\infty \le c_5\max(\sqrt{p/n_1}, n_1^{-\min(\kappa_1,\kappa_2)}).
$$

This completes the proof. ■