Sparse Causal Effect Estimation using Two-Sample Summary Statistics in the Presence of Unmeasured Confounding

Shimeng Huang
Institute of Science and Technology Austria

Niklas Pfister
University of Copenhagen

Jack Bowden
Novo Nordisk
University of Exeter

Abstract

Observational genome-wide association studies are now widely used for causal inference in genetic epidemiology. To maintain privacy, such data is often only publicly available as summary statistics, and often studies for the endogenous covariates and the outcome are available separately. This has necessitated methods tailored to two-sample summary statistics. Current state-of-the-art methods modify linear instrumental variable (IV) regression—with genetic variants as instruments—to account for unmeasured confounding. However, since the endogenous covariates can be high dimensional, standard IV assumptions are generally insufficient to identify all causal effects simultaneously. We ensure identifiability by assuming the causal effects are sparse and propose a sparse causal effect two-sample IV estimator, `spaceTSIV`, adapting the `spaceIV` estimator by Pfister and Peters (2022) for two-sample summary statistics. We provide two methods, based on L0- and L1-penalization, respectively. We prove identifiability of the sparse causal effects in the two-sample setting and consistency of `spaceTSIV`. The performance of `spaceTSIV` is compared with existing two-sample IV methods in simulations. Finally, we showcase our methods using real proteomic and gene-expression data for drug-target discovery.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

The use of observational data to study the causal effects of covariate interventions on an outcome has seen a surge in popularity in many scientific areas. A primary example is genetic epidemiology, where a common research topic is to study the causal effects of genetically predictive phenotypic traits, such as a person's body mass index or their low density lipoprotein cholesterol, on downstream disease outcomes. This is often based on Mendelian randomization (MR)—that is, instrumental variable estimation (IV) with genetic variants being the instruments—to account for unmeasured confounding between the endogenous covariates and the outcome. However, due to privacy concerns, access to individual-level genetic data is highly regulated. To both preserve privacy and enable data sharing, public data repositories of genetic summary statistics are made available by various international genome-wide association study (GWAS) consortia. These summary statistics usually contain estimates for the marginal effect of single nucleotide polymorphisms (SNPs) on the phenotypic traits and disease outcomes, along with their standard errors, which are often themselves obtained from two separate GWAS. This is referred to as the "two-sample summary statistics" setting. Zhao et al. (2019) discuss sufficient assumptions that enable consistent estimation under two-sample IV, specifically the homogeneity of the two samples. When the number of endogenous covariates under investigation is high dimensional or the instruments are highly correlated, a case in point being human gene expression phenotypes and genetic expression quantitative trait loci, there may be an insufficient number of strong and valid instruments to ensure the identifiability of the multivariable causal effects.

Lack of identifiability leads to poor estimation, or weak instrument bias. In the univariable two-sample summary statistics setting, Bowden et al. (2019) develop heuristic weak-instrument robust inference

strategies based on heterogeneity statistic estimating equations. Under the same setting, Wang and Kang (2022) further clarify the connection between these approaches and summary statistics analogues of the Anderson-Rubin (AR) test statistic (Anderson and Rubin, 1949) and Limited Information Maximum Likelihood (LIML). Wang et al. (2021a) further extend weak-instrument robust models to the multivariable case. Another way to circumvent the weak instrument problem is to employ principal component analysis (PCA). Building on the work of Batool et al. (2022), Patel et al. (2024) show how many individually weak variants could be fashioned into PCA scores with improved instrument strength.

An alternative strategy to tackle the lack of identifiability is to introduce sparsity assumption on the causal effects. This is often a reasonable assumption in MR studies, as it is usually the case that many endogenous traits do not have direct causal effects on the outcome. Under the assumptions of independent instruments and the number of instruments is no less than the number of covariates, Grant and Burgess (2022) consider the use of L1 penalization on the causal effects in multivariable MR models where one covariate is of special interest but the others are allowed to be penalized. In related works, Kang et al. (2016), Rees et al. (2019), and Grant and Burgess (2021) consider L1 penalization for individual instruments suspected to be invalid due to exclusion restriction violation, rather than penalization on the number of causal effects. From a variable selection perspective, another related approach is to employ Bayesian model averaging to select the best set of covariates in multivariable MR models (Zuber et al., 2020). However, all of the above-mentioned methods require that the number of instruments is at least as large as the number of covariates.

In the one sample individual-level data setting, the identifiability conditions for sparse causal effects have been studied by Pfister and Peters (2022), and they propose a sparse causal effect estimator, `spaceIV`. Tang et al. (2023) also consider sparse causal effect identification and estimation under assumptions on the sparsity level and propose a synthetic two-stage regularized regression approach.

We propose `spaceTSIV`, adapting the `spaceIV` estimator for two-sample summary statistics. To the best of our knowledge, our work proposes the first method that uses sparse causal effects for identifiability and applies to GWAS summary statistics, which is often the only data source in genetics research. We allow the IVs to be correlated by extending the adjustment method in Wang and Kang (2022). Two specific approaches based on L0- and L1-penalization, respectively, are provided. We prove identifiability of the sparse causal effects under the two-sample summary statistics setting, prove consistency of the proposed estimator, and provide theoretical guarantees that the adapted test using two-sample summary statistics is uniformly asymptotically level. Moreover, the proof of our Theorem 3.1 extends the consistency results in spaceIV and does not rely on the assumption of Gaussian noise. We evaluate the performance of `spaceTSIV` with simulated data and compare it with existing (non-sparse) methods that work with two-sample summary statistics. Finally, we showcase our methods using proteomic and gene-expression data within the context of a drug-target discovery analysis. Notation is summarized below and all proofs are provided in Supplementary Material C.

Notation. For all $k \in \mathbb{N}$, we define $[k] := \{1, \dots, k\}$ and for all $\beta \in \mathbb{R}^d$, we denote by $\text{supp}(\beta) := \{j \in [d] : \beta^j \neq 0\}$ the set of non-zero components of β. For an arbitrary matrix $A \in \mathbb{R}^{n \times m}$, we denote for all $i \in [n]$ and $j \in [m]$, the i-th row of A by A_i, the j-th column of A by A^j, and the ij-th entry of A by A_i^j. If A is a square block matrix containing $k \times k$ square matrices of dimension $l \times l$, then $A^{[ij]}$ for all $i, j \in [k]$ denotes the ij-th block of A.

2 REDUCED FORM IV MODEL AND SUMMARY STATISTICS

We start from the conventional one-sample individual-level data setting and assume we observe n independently and identically distributed (iid) observations $\{(X_i, Y_i, Z_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^m$, where Y is a response variable, X a vector of endogenous covariates, and Z a vector of instruments. The IV model assumptions can then be expressed as a linear structural causal model (SCM) over these variables.[1] Formally, for all $i \in [n]$, we assume,

$$\begin{aligned} X_i &:= AZ_i + BX_i + g(H_i, \nu_i^X) \\ Y_i &:= X_i^\top \beta^* + h(H_i, \nu_i^Y), \end{aligned} \tag{1}$$

where $H_i \in \mathbb{R}^q$ is a vector of unobserved variables, g and h are arbitrary measurable functions, Z_i, $h(H_i, \nu_i^X)$, and $g(H_i, \nu_i^Y)$ have mean 0 and finite variance, and $\left\{Z_i, H_i, \nu_i^X, \nu_i^Y\right\}_{i=1}^n$ are jointly independent. The coefficient $\beta^* \in \mathbb{R}^d$ denotes the true causal effect of the covariates on the response, and the matrices $A \in \mathbb{R}^{d \times m}$ and $B \in \mathbb{R}^{d \times d}$ encode the causal relationships between the variables Z and X, where we additionally assume that B induces a DAG over X. The

[1] The required assumptions can also be expressed via other causal models (e.g., potential outcomes). Not all causal implications of the model introduced here are strictly necessary, but to keep the presentation concise we avoid presenting the most general assumptions.

matrix $I_d - B$ is assumed to be invertible, where I_d is the identity matrix of dimension d. Finally, we call the support of β^* the parent set of Y and denote it as $\mathrm{PA}(Y)$, that is, $\mathrm{PA}(Y) := \mathrm{supp}(\beta^*)$. The SCM (1) can also be expressed in what is called its reduced form by only considering how the instruments affect the covariates and the response. Formally, for all $i \in [n]$ the reduced form is given by

$$\begin{aligned} X_i &:= Z_i^\top \Pi + u_i^X \\ Y_i &:= Z_i^\top \pi + u_i^Y, \end{aligned} \tag{2}$$

where $\Pi := A^\top (I_d - B)^{-\top} \in \mathbb{R}^{m \times d}$, $\pi := \Pi\beta^* \in \mathbb{R}^m$, $u_i^X := g(H_i, \nu_i^X)^\top (I_d - B)^{-\top}$, and $u_i^Y := (u_i^X)^\top \beta^* + h(H_i, \nu_i)$.

In this work, we assume that we do not directly observe the individual-level data and instead only have access to summary statistics of partially observed paired data from two independent samples $\{(Y_{ai}, Z_{ai})\}_{i=1}^{n_a}$ and $\{(X_{bi}, Z_{bi})\}_{i=1}^{n_b}$ of the SCM (1).

As discussed in Section 1, this is often the case in MR studies utilizing summary statistics from two GWAS: one contains the associations between genetic variants and endogenous traits (such as gene expression levels), and the other contains the associations between genetic variants and an outcome trait (such as a disease). These summary statistics are used to study the causal relationship between the endogenous traits and the outcome trait, with genetic variants being the IVs.

There are two types of summary statistics that we focus on here. Firstly, the two-sample joint OLS summary statistics, which consist of estimates of the reduced form parameters in (2) and are formally defined as follows.

Definition 2.1 (Two-sample joint OLS summary statistics)**.** Given two independent samples of observations $\{(Y_{ai}, Z_{ai})\}_{i=1}^{n_a}$ and $\{(X_{bi}, Z_{bi})\}_{i=1}^{n_b}$, the *two-sample joint OLS summary statistics* (joint summary statistics) are defined as the set of estimates

$$\mathcal{D}_{a,b}^{\text{joint}} := \left\{ \hat{\pi}, \hat{\Sigma}_\pi, \hat{\Pi}, \hat{\Sigma}_\Pi \right\},$$

where $\hat{\pi} := (\mathbf{Z}_a^\top \mathbf{Z}_a)^{-1} \mathbf{Z}_a^\top \mathbf{Y}_a \in \mathbb{R}^m$, $\hat{\Sigma}_\pi := \hat{\varepsilon}_a^\top \hat{\varepsilon}_a (\mathbf{Z}_a^\top \mathbf{Z}_a)^{-1} \in \mathbb{R}^{m \times m}$ with $\hat{\varepsilon}_a := \mathbf{Y}_a - \mathbf{Z}_a \hat{\pi}$, $\hat{\Pi} := (\mathbf{Z}_b^\top \mathbf{Z}_b)^{-1} \mathbf{Z}_b^\top \mathbf{X}_b \in \mathbb{R}^{m \times d}$, and $\hat{\Sigma}_\Pi \in \mathbb{R}^{md \times md}$ consists of $d \times d$ blocks of dimension $m \times m$ defined for all $k, l \in [d]$ by $\hat{\Sigma}_\Pi^{[kl]} := (\hat{\varepsilon}_b^k)^\top \hat{\varepsilon}_b^l (\mathbf{Z}_b^\top \mathbf{Z}_b)^{-1}$ with $\hat{\varepsilon}_b^k := \mathbf{X}_b^k - \mathbf{Z}_b \hat{\Pi}^k$.

Secondly, the two-sample marginal OLS summary statistics, which instead of capturing the joint effects described by the parameters in (2), only contain marginal univariate effects.

Definition 2.2 (Two-sample marginal OLS summary statistics)**.** Given two independent samples of observations $\{(Y_{ai}, Z_{ai})\}_{i=1}^{n_a}$ and $\{(X_{bi}, Z_{bi})\}_{i=1}^{n_b}$, the *two-sample marginal OLS summary statistics* (marginal summary statistics) are defined as the set of estimates

$$\mathcal{D}_{a,b}^{\text{marginal}} := \left\{ \hat{\eta}, \hat{\sigma}_\eta^2, \hat{H}, \hat{\sigma}_H^2, \widehat{M}_{Z_a}, \widehat{M}_{Z_b}, \widehat{M}_X \right\},$$

where $\hat{\eta} \in \mathbb{R}^m$, $\hat{\sigma}_\eta^2 \in \mathbb{R}^m$, $\hat{H} \in \mathbb{R}^{m \times d}$, $\hat{\sigma}_H^2 \in \mathbb{R}^{m \times d}$, and for all $j \in [m]$ and all $k \in [d]$, $\hat{\eta}_j := (\mathbf{Z}_a^j)^\top \mathbf{Y}_a / (\mathbf{Z}_a^j)^\top \mathbf{Z}_a^j$, $\hat{\sigma}_{\eta,j}^2 := (\hat{\varepsilon}_a^j)^\top \hat{\varepsilon}_a^j / \left((\mathbf{Z}_a^j)^\top \mathbf{Z}_a^j \right)$ with $\hat{\varepsilon}_a^j := \mathbf{Y}_a - \hat{\eta}_j \mathbf{Z}_a^j$, $\hat{H}_j^k := (\mathbf{Z}_b^j)^\top \mathbf{X}_b^k / (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j$, and $(\hat{\sigma}_{H,j}^k)^2 := (\hat{\varepsilon}_{bj}^k)^\top \hat{\varepsilon}_{bj}^k / \left((\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j \right)$ with $\hat{\varepsilon}_{bj}^k := \mathbf{X}_b^k - \hat{H}_j^k \mathbf{Z}_b^j$. For both $s \in \{a, b\}$, let D_{Z_s} be the diagonal matrix containing the diagonal elements of $\mathbf{Z}_s^\top \mathbf{Z}_s$, then $\widehat{M}_{Z_s} := D_{Z_s}^{-1/2} \mathbf{Z}_s^\top \mathbf{Z}_s D_{Z_s}^{-1/2} \in \mathbb{R}^{m \times m}$ are the sample correlation matrices of Z_s respectively. Similarly, let D_X be the diagonal matrix containing the diagonal elements of $\mathbf{X}_b^\top \mathbf{X}_b$, then $\widehat{M}_X := D_X^{-1/2} \mathbf{X}_b^\top \mathbf{X}_b D_X^{-1/2} \in \mathbb{R}^{d \times d}$ is the sample correlation matrix of X_b.

Due to the close relation of the joint summary statistics with the reduced form model (2) it is easier to develop methods for the joint summary statistics. However, in most publicly available data (e.g., UK Biobank and GWAS Catalog) only the marginal summary statistics are available. Fortunately, it is possible to transform marginal summary statistics into joint summary statistics. This means that any theoretical argument that applies to one also applies to the other. The exact correspondence is given in the following proposition.

Proposition 2.1 (Marginal to joint summary statistics)**.** Assume we are given $\mathcal{D}_{a,b}^{\text{marginal}} = \{\hat{\eta}, \hat{\sigma}_\eta^2, \hat{H}, \hat{\sigma}_H^2, \widehat{M}_{Z_a}, \widehat{M}_{Z_b}, \widehat{M}_X\}$. Define diagonal matrices $D_a, D_b^{(1)}, \ldots, D_b^{(m)} \in \mathbb{R}^{m \times m}$ such that for all $k, i \in [m]$, $(D_a)_i^i := (\hat{\sigma}_{\eta,i}^2 + (\hat{\eta}_i)^2)^{1/2}$ and $(D_b^{(k)})_i^i := ((\hat{\sigma}_{H,i}^k)^2 + (\hat{H}_i^k)^2)^{1/2}$. Then it holds for all $k, l \in [d]$ that

- $\hat{\pi} = D_a (D_a \, \widehat{M}_{Z_a})^{-1} \, \hat{\eta}$,
- $\hat{\Sigma}_\pi = (1 - \hat{\eta}^\top D_a \widehat{M}_{Z_a}^{-1} D_a \hat{\eta}) D_a \widehat{M}_{Z_a}^{-1} D_a$,
- $\hat{\Pi}^k = D_b^{(k)} (D_b^{(k)} \, \widehat{M}_{Z_b})^{-1} \, \hat{H}^k$, and
- $\hat{\Sigma}_\Pi^{[kl]} = (\widehat{M}_{X,k}^l - \hat{H}^{k\top} D_b^{(k)} \widehat{M}_{Z_b}^{-1} D_b^{(l)} \hat{H}^l) D_b^{(k)} \widehat{M}_{Z_b}^{-1} D_b^{(l)}$.

In practice, one often does not observe both $\widehat{M}_{Z_a}$ and $\widehat{M}_{Z_b}$ and instead only observes a single estimate that converges to the correlation of Z. In such cases, it can be shown that using the same transformation as in Proposition 2.1 is asymptotically equivalent to working with the joint summary statistics.

2.1 Identifiability via Sparsity under the Reduced IV Model

For the causal effect β^* to be identified, the number of instruments is usually required to be no less than the number of covariates. In the one-sample individual-level data setting, this can be seen from the solution space based on the IV moment condition under the SCM (1),

$$\mathcal{B}^{\text{ind}} = \left\{\beta \in \mathbb{R}^d : \mathbb{E}(ZY) = \mathbb{E}(ZX^\top)\beta)\right\}. \quad (3)$$

This space is in general non-degenerate if the dimension of the instruments is larger than the number of covariates. When the causal effect is sparse, however, it is possible to allow more covariates than instruments.

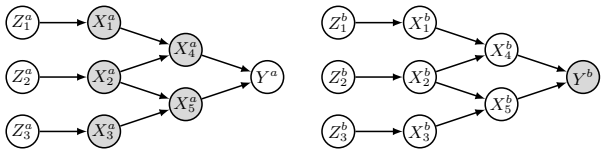

Figure 1: An example of two-sample IV scenario that is considered as underidentified in the usual sense. Hidden confounders between X and Y are omitted for clarity. While the two DAGs have the same structure, in sample a (left) the covariates X are not observed and in sample b (right) the outcome Y is not observed, these unobserved variables are represented by gray nodes.

Pfister and Peters (2022) study in detail the identifiability conditions under the SCM (1). In the following, we describe the identifiability conditions under the reduced model (2) which is compatible with the two-sample summary statistics scenario. Lemma 2.1 describes the solution space of the causal effects with the reduced form model.

Lemma 2.1. If $\mathbb{E}[ZZ^\top]$ has full rank, the solution space of the causal effects based on the IV moment condition can be written as

$$\mathcal{B}^{\text{sum}} = \left\{\beta \in \mathbb{R}^d : \pi - \Pi\beta = 0\right\}. \quad (4)$$

We will focus on the case where the instruments do not have direct effects on the response, which is implied by the SCMs (1) and its reduced form (2). This is usually referred to as the exclusion restriction criteria of IV. In genetics research, such a direct effect is also referred to as pleiotropy (see e.g., Hemani et al., 2018). We demonstrate empirically with additional simulations in Supplementary Material E.2 that the proposed methods still perform well under small violations of this assumption. An example of the possible scenario represented by directed acyclic graphs (DAGs) is given in Figure 1. As we will see shortly, although the number of instrument is less than the number of covariates in this case, the causal effect from X to Y may still be identified.

Under the two-sample summary statistic setting, the identifiability conditions in Pfister and Peters (2022) can be written as follows.[2]

Assumption 2.1. For all $S \subseteq [d]$, let Π^S be the submatrix of Π containing the columns in S. We assume the following regarding the true parameter Π

(a) $\text{rank}(\Pi^{\text{PA}(Y)}) = |\text{PA}(Y)|$.

(b) $\forall S \subseteq [d]$, it holds that $\text{rank}(\Pi^S) \leqslant \text{rank}(\Pi^{\text{PA}(Y)})$ and $\text{Im}(\Pi^S) \neq \text{Im}(\Pi^{\text{PA}(Y)})$ imply that $\forall w \in \mathbb{R}^{|S|}$, $\Pi^S w \neq \Pi^{\text{PA}(Y)}(\beta^*)^{\text{PA}(Y)}$.

(c) $\forall S \subseteq [d]$ with $|S| = |\text{PA}(Y)|$ and $S \neq \text{PA}(Y)$, $\text{Im}(\Pi^S) \neq \text{Im}(\Pi^{\text{PA}(Y)})$.

Assumption 2.1 serves as a key assumption in our setting to ensure identifiability with more covariates than instruments. In graphical terms, (a) implies that there should be at least $|\text{PA}(Y)|$ paths that are strictly disjoint from the IVs to Y, making sure that there is enough heterogeneity introduced by the instruments, (c) further ensures that subsets of the covariates that have the same cardinality as the parent set can be distinguished, and (b) assumes no exact cancellation in the causal paths from the instruments to response. We give an example in Supplementary Material E.2 where Assumption 2.1 is violated, and show empirically that our procedure is no longer consistent in that case.

To obtain a sparse solution, it is natural to consider the following optimization problem

$$\min_{\beta \in \mathcal{B}^{\text{sum}}} ||\beta||_0. \quad (5)$$

Theorem 2.1 shows that under Assumption 2.1, β^* is a unique solution to (5). The proof follows similarly as in Pfister and Peters (2022, Theorem 3).

Theorem 2.1 (Identifiability of sparse causal effect with reduced form model)**.** If Assumption 2.1 (a) and (b) hold, then β^* is a solution to (5). If in addition Assumption 2.1 (c) holds, then β^* is the unique solution.

2.2 Anderson-Rubin Test for Two-Sample Summary Statistics

The AR test is a well-known weak-instrument robust test for the causal effect, and the LIML estimator is known to minimize the AR statistic (e.g., Dhrymes, 2012). Wang and Kang (2022) consider the two-sample

[2] For a matrix A, $\text{Im}(A)$ denotes the image of A.

summary statistic version of the AR test when there is a single covariate, which can be seen as a a generalization of the the modified Q statistic proposed by Bowden et al. (2019) for independent instruments. The following result is a generalization of Wang and Kang (2022) in the presence of multiple covariates[3], which will be referred to as the Q statistic.

Theorem 2.2 (Q statistic)**.** Assume Assumption B.1 holds. For all $\beta \in \mathbb{R}^d$, define the Q statistic as

$$Q(\beta) := (\widehat{\pi} - \widehat{\Pi}\beta)^\top (\tfrac{1}{n_a}\widehat{\Sigma}_\pi + \tfrac{1}{n_b}\widehat{\Sigma}_\Pi(\beta))^{-1}(\widehat{\pi} - \widehat{\Pi}\beta), \quad (6)$$

where $\widehat{\Sigma}_\Pi(\beta) := \xi(\beta)\widehat{\Sigma}_\Pi\xi^\top(\beta)$ with $\xi(\beta) := \beta^\top \otimes I_m$. Then it holds for all $\beta \in \mathbb{R}^d$ and all $r \in (0, \infty)$ that

$$\lim_{\substack{n_a, n_b \to \infty \\ n_a/n_b \to r}} \sup_{t \in \mathbb{R}} \sup_{\substack{P \in \mathcal{P}: \\ \beta \in \mathcal{B}^{\text{sum}}(P)}} |\mathbb{P}_P\left(Q(\beta) \leqslant t\right) - \kappa_m(t)| = 0,$$

where κ_m is the CDF of the chi-squared distribution with m-degrees of freedom.

The Q statistic is the two-sample counterpart of the one-sample AR statistic, we present their connections in Supplementary Material A.1. Its minimizer can also be viewed as a generalized method of moments (GMM) estimator (Hansen, 1982), and it is related to the J statistic in economics literature. See Remark D.1 in the Supplementary Material for additional comments on the definition of $\widehat{\Sigma}_\Pi(\beta)$.

3 ESTIMATING SPARSE CAUSAL EFFECTS WITH spaceTSIV

We describe two estimation procedures to the optimization problem (5). The first procedure is the two-sample summary statistics counterpart of `spaceIV` by Pfister and Peters (2022), and the second procedure employs L1-penalization to replace subset selection which has the advantage of faster computational speed. For both procedures, we will use the following estimator, which is the minimizer of the Q statistic constrained on a specific support. For all $S \subseteq \{1, \dots, d\}$, define

$$\widehat{\beta}^Q(S) := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d:\ \text{supp}(\beta) = S} Q(\beta). \quad (7)$$

In order to provide precise theoretical results, we further let $\mathcal{P}$ denote a family of distributions for (X, Y, Z) generated by (1) which is assumed to be sufficiently regular (see Assumption B.1 for details). For all $P \in \mathcal{P}$, we let $\beta^*(P)$ denote the causal effect and $\mathcal{B}^{\text{sum}}(P)$ be the subset $\mathcal{B}^{\text{sum}}$ induced by the distribution P (both of which are fully identified from the observational distribution P).

[3] A related result is also considered by Patel et al. (2024) where a dispersion parameter is included and the principal components of the instruments are used. Here we focus on the case where the instruments are valid.

3.1 Sparsity by subset selection

For all $s \in [d]$, let

$$\widehat{\beta}^Q(s) := \widehat{\beta}^Q\Big(\operatorname{arg\,min}_{S \subseteq \{1,\dots,d\}:|S|=s} Q\left(\widehat{\beta}^Q(S)\right)\Big).$$

Moreover, following Theorem 2.2, for all $s \in [d]$ and for all $\alpha \in (0, 1)$, the hypothesis test

$$\varphi_s^\alpha(\mathcal{D}_{a,b}^{\text{joint}}) := \mathbb{1}\left(Q(\widehat{\beta}^Q(s)) > \kappa_m^{-1}(1-\alpha)\right)$$

has uniform asymptotic level for the null hypothesis

$$\mathcal{H}_0(s) := \{P \in \mathcal{P} \mid \exists \beta \in \mathcal{B}^{\text{sum}}(P) :\ \|\beta\|_0 = s\},$$

that is, for $\alpha \in (0, 1)$, it holds that

$$\lim_{n_a, n_b \to \infty} \sup_{P \in \mathcal{H}_0(s)} \mathbb{P}_P(\varphi_s^\alpha(\mathcal{D}_{a,b}^{\text{joint}}) = 1) \leqslant \alpha.$$

An algorithm defining the `spaceTSIV` estimator using subset selection is given in Algorithm 1. Theorem 3.1 shows that it is consistent.

Theorem 3.1. Assume Assumption B.1 holds. Let $\mathcal{D}_{a,b}^{\text{joint}}$ be the joint summary statistics based on two independent samples of size n_a and n_b respectively. Let $P \in \mathcal{P}$ and $s_{\max} \in \mathbb{N}$ such that $s_{\max} \geqslant \|\beta^*(P)\|_0$. If Assumption 2.1 (a) and (b) holds, then for all $r \in (0, \infty)$

$$\lim_{\substack{n_a, n_b \to \infty \\ n_a/n_b \to r}} \mathbb{P}_P\left(\|\widehat{\beta}^{\leqslant s_{\max}}\|_0 = \|\beta^*\|_0\right) \geqslant 1 - \alpha;$$

if in addition Assumption 2.1 (c) also holds, then for all $\varepsilon > 0$ and all $r \in (0, \infty)$

$$\lim_{\substack{n_a, n_b \to \infty \\ n_a/n_b \to r}} \mathbb{P}_P\left(\|\widehat{\beta}^{\leqslant s_{\max}} - \beta^*\|_2 < \varepsilon\right) \geqslant 1 - \alpha.$$

3.2 Sparsity by L1 penalty

The subset selection approach introduced in Section 3.1 becomes computationally infeasible when the number of covariates is large. We therefore propose a faster approach that uses L1 penalization to estimate the support of β^* and then adapt the testing procedure from the previous section. More specifically, for a penalty parameter $\lambda > 0$, we first minimize the following L1-loss

$$\mathcal{L}_\lambda^{\text{TSIV-L1}}(\beta) = \frac{1}{2}\|\widehat{\pi} - \widehat{\Pi}\beta\|_2^2 + \lambda\|\beta\|_1. \quad (8)$$

Algorithm 1: `spaceTSIV` with L0 penalization

Input: Joint summary statistics $\mathcal{D}_{a,b}^{\text{joint}}$, maximum support size $s_{\max}$, significance level $\alpha \in (0,1)$

Initialize $s \leftarrow 1$ and $\phi \leftarrow 1$
while $s \leqslant s_{max}$ *and* $\phi = 1$ **do**
  Set $\mathbf{S}_s$ to the set of all subsets of $[d]$ of size s
  for $S \in \mathbf{S}_s$ **do**
    Compute $\widehat{\beta}^Q(S)$
    Compute $Q(\widehat{\beta}^Q(S))$
  end
  $S_{\text{best}} \leftarrow \arg\min_{S \in \mathbf{S}_s} Q(\widehat{\beta}^Q(S))$
  $\widehat{\beta}(s) \leftarrow \widehat{\beta}^Q(S_{\text{best}})$
  $\phi \leftarrow \varphi_s^\alpha(\mathcal{D}_{a,b}^{\text{joint}})$
  $s \leftarrow s + 1$
end
$\widehat{\beta}_{\leqslant s_{\max}} \leftarrow \widehat{\beta}(s)$

Output: Final estimate $\widehat{\beta}_{\leqslant s_{\max}}$ and test result ϕ

Algorithm 2: `spaceTSIV` with L1 penalization

Input: Joint summary statistics $\mathcal{D}_{a,b}^{\text{joint}}$, a vector of penalty values in decreasing order $\{\lambda_1, \ldots, \lambda_\ell\}$, significance level $\alpha \in (0,1)$

Initialize $l \leftarrow 1$ and $\phi \leftarrow 1$
while $l \leqslant \ell$ *and* $\phi = 1$ **do**
  $\lambda \leftarrow \lambda_l$
  $\widehat{S}_\lambda \leftarrow \text{supp}\left(\arg\min_{\beta \in \mathbb{R}^d} \mathcal{L}_\lambda^{\text{TSIV-L1}}(\beta)\right)$
  Compute $\widehat{\beta}^Q(\widehat{S}_\lambda)$
  $\phi \leftarrow \varphi_\lambda^\alpha(\mathcal{D}_{a,b}^{\text{joint}})$
  $l \leftarrow l + 1$
end
$\widehat{\beta}_{\leqslant \lambda_{\max}} \leftarrow \widehat{\beta}(\widehat{S}_\lambda)$

Output: Final estimate $\widehat{\beta}_{\leqslant \lambda_{\max}}$ and test result ϕ

Define $\widehat{\beta}(\lambda) := \arg\min_{\beta \in \mathbb{R}^d} \mathcal{L}_\lambda^{\text{TSIV-L1}}(\beta)$ and $\widehat{S}_\lambda := \text{supp}(\widehat{\beta}(\lambda))$. We then propose to refit the parameter as in (7) using the set $\widehat{S}_\lambda$ and performing the hypothesis test defined by

$$\varphi_\lambda^\alpha(\mathcal{D}_{n_a,n_b}^{\text{joint}}) := \mathbb{1}\left(Q(\widehat{\beta}^Q(\widehat{S}_\lambda)) > \kappa_m^{-1}(1-\alpha)\right).$$

By similar arguments as in Section 3.1 this test for $S = \widehat{S}_\lambda$ has uniform asymptotic level for the null hypothesis

$$H_0(S) := \{P \in \mathcal{P} \mid \exists \beta \in \mathcal{B}^{\text{sum}}(P) : \text{supp}(\beta) = S\}.$$

Under sufficient regularity conditions and assuming that β^* is indeed sparse, one can hope—based on similar results for high-dimensional linear models (e.g., Bühlmann and Van De Geer, 2011)—that for appropriately chosen λ it holds that $\widehat{S}_\lambda$ converges to $\text{supp}(\beta^*)$. This motivates the following estimator, `spaceTSIV` with L1 penalization, defined in Algorithm 2.

Intuitively, if the subset selection is indeed correct (i.e., it recovers the support of β^*) for the first accepted set, then this procedure should correctly estimate β^*. A full theoretical analysis, however, goes beyond the scope of this work and we propose this procedure only as a heuristic computational speed up.

3.3 Practical considerations

When using the subset selection approach in practice, it can happen that there are multiple estimates with different support of the same (smallest) size not being rejected by φ_s^α. This indicates, that at least in finite sample, the causal effect β^* is not fully identified. We recommend reporting all subsets of the smallest size that are not rejected by φ_s^α as possible effects.

Moreover, since the estimator `spaceTSIV` is based on optimizing a test statistics, one immediate approach to construct confidence intervals (CIs) is by inverting the test. In the real application, we construct the CIs for the non-zero causal effects by inverting φ_s^α or φ_λ^α and projecting onto each non-zero coordinate. We choose this approach for its practicality, but other approaches exist which may be more suitable (e.g., Londschien and Bühlmann, 2024), and one should also take into account the effect of post-selection inference (e.g., Lee et al., 2016). One of the advantages of inverting the test is that it takes into account the strength of the instruments (and hence identifiability). So if the resulting CIs are unbounded this generally indicates that there is limited identifiability. This is a well-known property for the AR test (e.g., Dufour, 1997; Davidson and MacKinnon, 2014).

Lastly, in genetics applications, high linkage disequilibrium (LD) can render the correlation matrix of instruments non-invertible, complicating the conversion of marginal to joint summary statistics (Proposition 2.1). A common solution is to apply LD clumping to remove highly correlated genetic variants. Future research could explore using a ridge penalty to achieve this conversion, though it lies beyond the scope of this work.

4 EXPERIMENTS

Code for reproducing the simulations and the real-data application along with the data are available in the GitHub repository `https://github.com/shimenghuang/spacetsiv`. All experiments were run on a MacBook Pro laptop with M1 chip.

4.1 Simulations

We present simulation results for two data-generating processes (DGPs) summarized below in this section. Further simulation results are provided in Supplementary Material E. The first, DGP1, is a low-dimensional example taken from Pfister and Peters (2022, Figure 3). We compare the subset selection and the L1-penalization versions of `spaceTSIV`, denoted as `spaceTSIV-L0` and `spaceTSIV-L1` respectively, as well as the `TSIV` estimator (defined as the minimizer of (8) with $\lambda = 0$, in which case the generalized inverse is used). The second, DGP2, illustrates the scenario with higher dimensional covariates, sparser causal effects, and correlated instruments. In this setting, we omit `spaceTSIV-L0` from the comparison due to its high computational cost. Since there are no obvious comparators (see discussions in Section 1), we benchmark our methods against the most commonly used method for a straightforward comparison to a standard analysis. An overview of the simulation setup is given below and more details can be found in Supplementary Material E.1.

DGP1 overview: $m = 3$ and $d = 5$ and $\|\beta^*\|_0 = 2$. For increasing $n = n_a = n_b$, we generate iid $\{(Y_i, Z_i)\}_{i=1}^{n_a}$ and $\{(X_i, Z_i)\}_{i=1}^{n_b}$ according to a linear SCM with Gaussian errors and then compute the summary statistics using seemingly unrelated regression.
DGP2 overview: $m = 5$, $d = 100$, and $\|\beta^*\|_0 = 2$. With fixed values of π, Π, Σ_π, and Σ_Π, and increasing $n = n_a = n_b$, we generate $\widehat{\pi}_{n_a} \sim \mathcal{N}(\pi, \frac{1}{n_a}\Sigma_\pi)$ and $\widehat{\Pi}_{n_b} \sim \mathcal{N}(\Pi, \frac{1}{n_b}\Sigma_\Pi)$, and set $\widehat{\Sigma}_{\pi,n_a} = \Sigma_\pi$ and $\widehat{\Sigma}_{\Pi,n_b} = \Sigma_\Pi$.

We evaluate `spaceTSIV` based on both its variable selection and estimation performances. The results are shown in Figure 2. We can see that the bias and root mean square error (rmse) of `spaceTSIV` shrinks with increasing sample size with either L0 or L1 penalization, which is not the case for the non-sparse estimator `TSIV`. In terms of variable selection, we see that for both DGPs as the sample sizes increase, the Jaccard similarity[4] increases to around 1, and the percentage of estimates having the correct support size also increases to around 100%, empirically confirming the consistency results in Theorem 3.1. The performance of `spaceTSIV-L0` and `spaceTSIV-L1` are similar in terms of both estimation and variable selection for DGP1.

Supplementary Material E.2 includes three additional simulations. The first shows that bias and rmse decrease with increasing sample size even when the exclusion restriction is violated, although the estimated causal parents tend to form a superset of the true ones. The second compares `spaceIV` with individual-level data from a complete DGP1 sample to `spaceTSIV` using two-sample summary statistics from two subsets of the same sample. While `spaceIV` outperforms `spaceTSIV` in bias and rmse for a given total sample size, the ratio of the two subsets has little impact on `spaceTSIV`. The third demonstrates that `spaceTSIV` is inconsistent when Assumption 2.1 is violated and causal parents are unidentifiable, but the estimated parents still tend to block paths to the true causal parents.

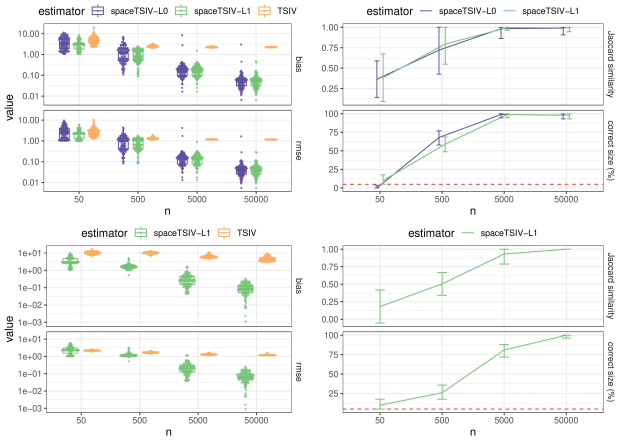

Figure 2: Results using data generated by DGP1 (top) and DGP2 (bottom) based on 100 repetitions. Left: Bias and rmse of the estimators. The y-axis is on log scale for clarity. Right: Average Jaccard similarity between the selected covariates and the true causal covariates (error bars indicate confidence intervals constructed by mean plus/minus one standard error), and percentage of estimates that have the correct support size (error bars indicate 95% binomial confidence intervals).

4.2 Application

We apply our methods to summary statistics of SNP-level associations where the covariates and the outcome come from two separate GWAS sources. The covariates' summary statistics come from the GTEX consortium, which measure levels of expression of protein coding genes across multiple tissue types in the human body. Gene expression is a convenient and reliable upstream marker of protein production, which would be the natural target of a future drug. We specifically focus on expression of the GLP1R gene in 10 tissue types that are relevant to the treatment of cardio-metabolic disease. These are brain caudate,

[4]For two sets A and B, the Jaccard similarity is defined as $\text{Jaccard}(A, B) := \dfrac{|A \cap B|}{|A \cup B|}$.

hypothalamus, atrial appendage, left ventricle, lung, nerve, pancreas, stomach, testis, and thyroid. The SNP-outcome summary statistics measure the genetic association with coronary artery disease (CAD) risk, and are obtained from the CARDIoGRAMplusC4D consortium. These data were first analysed in in Patel et al. (2024), who proposed a novel principle component analysis (PCA) method for constructing near-orthogonal composite instruments from 851 SNPs in the GLP1R gene region. For this analysis, they use 23 principle components (PCs) as IVs for the 10 covariates. The analysis by Patel et al. (2024) suggests that GLP1R expression only has a significant effect on CAD risk in 2 of the 10 tissues, although this was based on 95% confidence intervals using a normal approximation which, unlike the test-inversion method we use, does not always reliably capture the true uncertainty of IV estimates when the instruments are weak.

Based on the analysis of Patel et al. (2024), it is reasonable to believe that the causal effects are sparse in this application. Rather than opting for PCA pre-processing of the genetic summary statistics, we consider the selection of individual SNPs instruments based on the more conventional approach using the first-stage F-statistics[5] of the gene expression summary statistics. We keep the top two genetic variants with the largest first-stage F-statistics for each of the 10 covariates. Since some SNPs are most strongly associated with multiple covariates, we eventually keep 17 of the 851 genetic variants in the original data. Moreover, since the summary statistic data contains only the marginal associations along with their standard errors, we use the adjustment method in Proposition 2.1 to obtain the estimated joint effects and variance-covariance matrices.

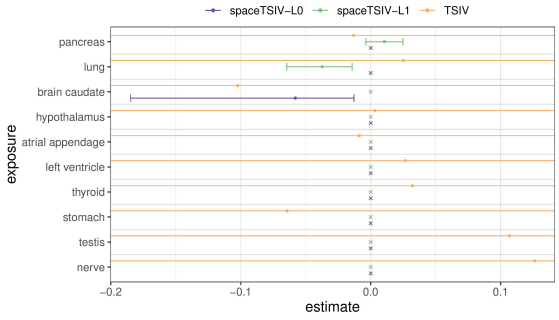

Figure 3: Estimated effects of the GLP1R expression in 10 tissues using the selected 17 genetic variants as instruments. Error bars represent 90% confidence intervals (CIs) constructed by inverting φ_s^α and φ_λ^α respectively, and projecting onto each coordinate.

[5]Given a marginal OLS coefficient $\hat{\gamma} \in \mathbb{R}$ and its corresponding standard error $\hat{\sigma} \in \mathbb{R}$, the first-stage F-statistic is defined as $\hat{\gamma}^2/\hat{\sigma}^2$.

The analysis results based on `spaceTSIV` with L0 and L1 penalization and regular `TSIV` are reported in Figure 3, at the significance level $\alpha = 0.1$. The 90% CIs are obtained from inverting φ_s^α and φ_λ^α as described in Section 3.3. They show that the CIs for `TSIV` are all of infinite length. This demonstrates that, even though there are more instruments than covariates, the causal effects are still under-identified due to weak instruments. Moreover, the `spaceTSIV` with L0 penalization yields a single set of size 1 while with L1 penalization we obtain a set of size 2. The significant negative effect of brain caudate aligns with the analysis result in Patel et al. (2024) and is biologically meaningful. The different result from `spaceTSIV-L1` could be due to the high correlation of the SNPs, which may result in the L1 relaxation of the L0 minimization problem not achieving the same estimate. In general, we recommend using the L0 procedure whenever computationally feasible as it comes with clear theoretical guarantees.

5 DISCUSSION

We propose `spaceTSIV` for sparse multivariable causal effect estimation under unobserved confounding, which is applicable to the two-sample summary statistics setting. Two methods using subset selection and L1-penalization respectively are provided. We prove consistency for the subset selection approach and illustrate the results in simulations. We also show in simulations that the L1-penalization approach, which is much more computationally efficient, can achieve similar performance as the subset selection approach in terms of bias and consistency. To focus on the main idea of this work, we have assumed that the summary statistics utilized in the analysis are obtained from two independent and homogeneous samples, which is commonly assumed in genetic epidemiology. However, it would be interesting to generalize the methods to heterogeneous samples similar to results by Zhao et al. (2019) in the non-sparse setting. Moreover, if the summary statistics are obtained from two samples with overlapping observations, additional correlations should be taken into account, and the ideas from Wang et al. (2021b) can possibly be applied. Lastly, weakening linearity in IV models is generally non-trivial, but it may be interesting to apply similar arguments in Zhao et al. (2019) and Zhao et al. (2020) to extend to non-linear IV models.

Acknowledgement

Most of this work was done while SH was at the Department of Mathematical Sciences, University of Copenhagen. The authors would like to thank Stephen

Burgess and Ashish Patel for helpful discussions at the start of this research project, and Anton Rask Lundborg for helpful discussions on the uniform asymptotic results. SH and NP are supported by a research grant (0069071) from Novo Nordisk Fonden. JB is funded at the University of Exeter by research grant MR/X011372/1.

References

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63, 1949.

F. Batool, A. Patel, D. Gill, and S. Burgess. Disentangling the effects of traits with shared clustered genetic predictors using multivariable Mendelian randomization. *Genetic Epidemiology*, 46(7):415–429, 2022.

J. Bowden, F. Del Greco M, C. Minelli, Q. Zhao, D. A. Lawlor, N. A. Sheehan, J. Thompson, and G. Davey Smith. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the nome assumption. *International Journal of Epidemiology*, 48(3):728–742, 2019.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

R. Davidson and J. G. MacKinnon. Confidence sets based on inverting Anderson–Rubin tests. *The Econometrics Journal*, 17(2):S39–S58, 2014.

P. J. Dhrymes. *Econometrics: Statistical foundations and applications*. Springer, 2012.

J.-M. Dufour. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica: Journal of the Econometric Society*, 65(6):1365–1387, 1997.

A. J. Grant and S. Burgess. Pleiotropy robust methods for multivariable Mendelian randomization. *Statistics in Medicine*, 40(26):5813–5830, 2021.

A. J. Grant and S. Burgess. An efficient and robust approach to mendelian randomization with measured pleiotropic effects in a high-dimensional setting. *Biostatistics*, 23(2):609–625, 2022.

GWAS Catalog. URL `https://www.ebi.ac.uk/gwas/`, accessed 2024-10-02.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 4(4):1029–1054, 1982.

G. Hemani, J. Bowden, and G. Davey Smith. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics*, 27(R2):R195–R208, 2018.

H. Kang, A. Zhang, T. T. Cai, and D. 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, 2016.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

M. Londschien and P. Bühlmann. Weak-instrument-robust subvector inference in instrumental variables regression: A subvector lagrange multiplier test and properties of subvector Anderson-Rubin confidence sets. *arXiv preprint arXiv:2407.15256*, 2024.

A. Patel, D. Gill, D. Shungin, C. S. Mantzoros, L. B. Knudsen, J. Bowden, and S. Burgess. Robust use of phenotypic heterogeneity at drug target genes for mechanistic insights: Application of cis-multivariable Mendelian randomization to GLP1R gene region. *Genetic Epidemiology*, 48(4):151–163, 2024.

N. Pfister and J. Peters. Identifiability of sparse causal effects using instrumental variables. In J. Cussens and K. Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1613–1622. PMLR, 2022.

J. M. Rees, A. M. Wood, F. Dudbridge, and S. Burgess. Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PLOS One*, 14(9):e0222362, 2019.

D. Tang, D. Kong, and L. Wang. The synthetic instrument: From sparse association to sparse causation. *arXiv preprint arXiv:2304.01098*, 2023.

UK Biobank. URL `https://www.ukbiobank.ac.uk/`, accessed 2024-10-02.

J. Wang, Q. Zhao, J. Bowden, G. Hemani, G. Davey Smith, D. S. Small, and N. R. Zhang. Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLOS Genetics*, 17(6):1–24, 06 2021a.

J. Wang, Q. Zhao, J. Bowden, G. Hemani, G. Davey Smith, D. S. Small, and N. R. Zhang. Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLoS genetics*, 17(6):e1009575, 2021b.

S. Wang and H. Kang. Weak-instrument robust tests in two-sample summary-data Mendelian randomization. *Biometrics*, 78(4):1699–1713, 2022.

Q. Zhao, J. Wang, W. Spiller, J. Bowden, and D. S. Small. Two-sample instrumental variable analyses using heterogeneous samples. *Statistical Science*, 34 (2):317–333, 2019.

Q. Zhao, J. Wang, G. Hemani, J. Bowden, and D. S. Small. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, 48 (3):1742–1769, 2020.

V. Zuber, J. M. Colijn, C. Klaver, and S. Burgess. Selecting likely causal risk factors from high-throughput experiments using multivariable mendelian randomization. *Nature communications*, 11(1):29, 2020.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material

A DETAILS OF TEST STATISTICS AND TEST-BASED ESTIMATORS

A.1 Connection between Anderson-Rubin statistic and Q statistic

Suppose we observe one set of iid samples $\{(X_i, Y_i, Z_i)\}_{i=1}^n$. The AR statistic is given by

$$AR(\beta) \coloneqq \frac{n-m}{m} \cdot \frac{(\mathbf{Y} - \mathbf{X}\beta)^\top P_Z (\mathbf{Y} - \mathbf{X}\beta)}{(\mathbf{Y} - \mathbf{X}\beta)^\top M_Z (\mathbf{Y} - \mathbf{X}\beta)}, \tag{S.1}$$

where $P_Z = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$ and $M_Z = I_d - P_Z$. When the true causal effect β^* is identified, $m \cdot AR(\beta^*) \xrightarrow{d} \chi^2_m$ (Anderson and Rubin, 1949; Londschien and Bühlmann, 2024).

We can rewrite the AR statistic in terms of (joint) OLS estimates and their respective estimated variance-covariance matrices. Specifically, let $\widehat{\pi} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}\mathbf{Y}$ and $\widehat{\Pi} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}\mathbf{X}$, we have that

$$(\mathbf{Y} - \mathbf{X}\beta)^\top P_Z (\mathbf{Y} - \mathbf{X}\beta) = (\widehat{\pi} - \widehat{\Pi}\beta)^\top (\mathbf{Z}^\top\mathbf{Z})(\widehat{\pi} - \widehat{\Pi}\beta).$$

Moreover, for all $\beta \in \mathbb{R}^d$ define

$$\begin{aligned}
\widehat{\Sigma}_\pi &= (\mathbf{Y} - \mathbf{Z}\widehat{\pi})^\top (\mathbf{Y} - \mathbf{Z}\widehat{\pi})(\mathbf{Z}^\top\mathbf{Z})^{-1}, \\
\widehat{\Sigma}_\Pi(\beta) &= \beta^\top (\mathbf{X} - \mathbf{Z}\widehat{\Pi})^\top (\mathbf{X} - \mathbf{Z}\widehat{\Pi})\beta(\mathbf{Z}^\top\mathbf{Z})^{-1}, \text{ and} \\
\widehat{\Sigma}_{\pi,\Pi}(\beta) &= (\mathbf{Y} - \mathbf{Z}\widehat{\pi})^\top (\mathbf{X} - \mathbf{Z}\widehat{\Pi})\beta(\mathbf{Z}^\top\mathbf{Z})^{-1}.
\end{aligned}$$

Then, we can expand the denominator in (S.1) as follows

$$\begin{aligned}
&(\mathbf{Y} - \mathbf{X}\beta)^\top M_Z (\mathbf{Y} - \mathbf{X}\beta) \\
&\quad = (M_Z\mathbf{Y} - M_Z\mathbf{X}\beta)^\top (M_Z\mathbf{Y} - M_Z\mathbf{X}\beta) \\
&\quad = (\mathbf{Y} - \mathbf{Z}\widehat{\pi} - (\mathbf{X} - \mathbf{X}\widehat{\Pi}\beta))^\top (\mathbf{Y} - \mathbf{Z}\widehat{\pi} - (\mathbf{X} - \mathbf{X}\widehat{\Pi})\beta) \\
&\quad = (\mathbf{Y} - \mathbf{Z}\widehat{\pi})^\top (\mathbf{Y} - \mathbf{Z}\widehat{\pi}) + \beta^\top (\mathbf{X} - \mathbf{X}\widehat{\Pi})^\top (\mathbf{X} - \mathbf{X}\widehat{\Pi})\beta - 2(\mathbf{Y} - \mathbf{Z}\widehat{\pi})^\top (\mathbf{X} - \mathbf{X}\widehat{\Pi})\beta,
\end{aligned}$$

which implies

$$(\mathbf{Y} - \mathbf{X}\beta)^\top M_Z (\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Z}^\top\mathbf{Z})^{-1} = \widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta) - 2\widehat{\Sigma}_{\pi,\Pi}(\beta).$$

Therefore,

$$\begin{aligned}
AR(\beta) &= \frac{(\widehat{\pi} - \widehat{\Pi}\beta)^\top (\mathbf{Z}^\top\mathbf{Z})(\widehat{\pi} - \widehat{\Pi}\beta)}{(\mathbf{Y} - \mathbf{X}\beta)^\top M_Z (\mathbf{Y} - \mathbf{X}\beta)} \\
&= \frac{1}{m}(\widehat{\pi} - \widehat{\Pi}\beta)^\top \left(\frac{1}{n-m}\widehat{\Sigma}_\pi + \frac{1}{n-m}\widehat{\Sigma}_\Pi(\beta) - \frac{2}{n-m}\widehat{\Sigma}_{\pi,\Pi}(\beta) \right)^{-1} (\widehat{\pi} - \widehat{\Pi}\beta).
\end{aligned} \tag{S.2}$$

From this expression, we can see the connection and differences between the AR statistic and the Q statistic. Specifically, the AR statistic can also be written in terms of summary statistics, for a fixed m and large n, the difference between $mAR(\beta)$ and $Q(\beta)$ is the term $\widehat{\Sigma}_{\pi,\Pi}$, which is related to the covariance between the residuals of a Y on Z and a X on Z regression. This is because the summary statistics in AR are based on the same sample, which requires that the covariance between $\widehat{\pi}$ and $\widehat{\Pi}$ to be accounted for, as shown in (S.2). In contrast, the summary statistics used in the Q statistic are derived from two independent samples, so this covariance is zero in population.

A.2 Coordinate descent for minimizing the TSIV-L1 loss

We describe the coordinate descent procedure for minimizing (8). Let

$$\mathcal{L}^{\text{TSIV}}(\beta) = \frac{1}{2}\|\widehat{\pi} - \widehat{\Pi}\beta\|_2^2.$$

For a matrix A, denote A^{-j} as the matrix removing A's j-th column. The derivative of $\mathcal{L}^{\text{TSIV}}(\beta)$ w.r.t β_j is

$$\begin{aligned}\frac{\partial \mathcal{L}^{\text{TSIV}}(\beta)}{\beta_j} &= -(\widehat{\Pi}^j)^\top(\widehat{\pi} - \widehat{\Pi}\beta) \\ &= -(\widehat{\Pi}^j)^\top\widehat{\pi} + (\widehat{\Pi}^j)^\top\widehat{\Pi}^{-j}\beta_{-j} + (\widehat{\Pi}^j)^\top\widehat{\Pi}^j\beta_j \\ &= -\rho_j + \eta_j\beta_j\end{aligned} \tag{S.3}$$

where $\rho_j \coloneqq (\widehat{\Pi}^j)^\top\widehat{\pi} + (\widehat{\Pi}^j)^\top\widehat{\Pi}^{-j}\beta_{-j}$ and $\eta_j \coloneqq (\widehat{\Pi}^j)^\top\widehat{\Pi}^j$. The subgradient of $\lambda||\beta||_1$ w.r.t β_j is

$$\frac{\partial \lambda||\beta||_1}{\beta_j} = \frac{\partial \lambda|\beta_j|}{\beta_j}\begin{cases}\{-\lambda\} & \beta_j < 0 \\ [-\lambda, \lambda] & \beta_j = 0 \\ \{-\lambda\} & \beta_j > 0\end{cases} \tag{S.4}$$

Combining (S.3) and (S.4), we have that the subgradient of $\mathcal{L}_\lambda^{\text{TSIV-L1}}(\beta)$ w.r.t. β_j is

$$\frac{\partial \mathcal{L}_\lambda^{\text{TSIV}}(\beta)}{\partial \beta_j} = \begin{cases}-\rho^j + \eta^j\beta_j - \lambda & \beta_j < 0 \\ [-\rho^j - \lambda, -\rho^j + \lambda] & \beta_j = 0 \\ -\rho^j + \eta^j\beta_j + \lambda & \beta_j > 0.\end{cases} \tag{S.5}$$

Starting from an initial value of $\widehat{\beta}$, we loop through $j \in [J]$ and update the value of $\widehat{\beta}_j$ by solving the equation resulting from setting (S.5) to 0, which gives

$$\widehat{\beta}_j = \begin{cases}\frac{\rho^j + \lambda}{\eta^j} & \rho^j < -\lambda \\ 0 & -\lambda < \rho^j < \lambda \\ \frac{\rho^j - \lambda}{\eta^j} & \rho^j > \lambda.\end{cases}$$

B REGULARITY CONDITIONS

Assumption B.1 (Regularity conditions)**.** Let $\mathcal{P}$ be a family of distributions for $(X, Y, Z) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^m$ generated by (1) and additionally satisfies that there exists $C_1, C_2, c, \eta > 0$ such that

- $\sup_{P\in\mathcal{P}}\left(\mathbb{E}_P[\|X\|^{4+\eta}] + \mathbb{E}[\|Y\|^{4+\eta}] + \mathbb{E}_P[\|Z\|^{4+\eta}]\right) \leqslant c$
- $\inf_{P\in\mathcal{P}} \min\left(\lambda_{\min}(\mathbb{E}_P[ZZ^\top]), \lambda_{\min}(\mathbb{E}_P[XX^\top])\right) \geqslant C_1.$
- $\sup_{P\in\mathcal{P}} \max\left(\lambda_{\max}(\mathbb{E}_P[ZZ^\top]), \lambda_{\max}(\mathbb{E}_P[XX^\top])\right) \leqslant C_2.$

The first bullet point ensures that all variables have sufficient moments (the $+\eta$-term is required for uniform convergence). The second bullet point is a uniform minimum eigenvalue condition that ensures the invertibility of the design matrices (this invertibility is also needed for the estimators to even exist). Finally, bullet point three is an upper bound on the maximum eigenvalue which ensures a lower bound on the eigenvalue of the inverse. All three are mild conditions and substantially more general than assuming a Gaussian linear model, for example.

C PROOFS

C.1 Proof of Lemma 2.1

Proof. The following equivalences hold

$$\begin{aligned}
\beta \in \mathcal{B}^{\text{ind}} &\iff \mathbb{E}(ZY) = \mathbb{E}(ZX^\top)\beta \\
&\iff \mathbb{E}(ZZ^\top)\pi = \mathbb{E}(ZZ^\top)\Pi\beta \quad \text{since following (2), we have } \mathbb{E}(ZY) = \mathbb{E}(ZZ^\top)\pi \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{and } \mathbb{E}(ZX^\top)\beta = \mathbb{E}(ZZ^\top)\Pi\beta \\
&\iff \pi = \Pi\beta \qquad \text{since } \mathbb{E}[ZZ^\top] \text{ is full rank.}
\end{aligned}$$

□

C.2 Proof of Proposition 2.1

Proof. We only prove the result for $\widehat{\Pi}$ and $\widehat{\Sigma}_\Pi$. $\widehat{\pi}$ and $\widehat{\Sigma}_\pi$ can be viewed as a special case of the former, with $d = 1$.

We first express $D_b^{(k)}$ in terms of the design matrices. To this end, observe that for all $j \in [m]$ and all $k \in [d]$, we have from the marginal OLS summary statistics that

$$\begin{aligned}
\left(\widehat{\sigma}^k_{\eta,j}\right)^2 &= \frac{(\mathbf{X}_b^k - \widehat{H}_j^k \mathbf{Z}_b^j)^\top (\mathbf{X}_b^k - \widehat{H}_j^k \mathbf{Z}_b^j)}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} \\
&= \frac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k - 2\widehat{H}_j^k (\mathbf{X}_b^k)^\top \mathbf{Z}_b^j + (\widehat{H}_j^k)^2 (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} \\
&= \frac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k - 2(\widehat{H}_j^k)^2 (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j + (\widehat{H}_j^k)^2 (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} \\
&= \frac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k - (\widehat{H}_j^k)^2 (\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} \\
&= \frac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k}{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j} - (\widehat{H}_j^k)^2.
\end{aligned}$$

This further implies that $(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j = \dfrac{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k}{\left(\widehat{\sigma}^k_{\eta,j}\right)^2 + (\widehat{H}_j^k)^2}$, and hence $\left(\left(\widehat{\sigma}^k_{\eta,j}\right)^2 + (\widehat{H}_j^k)^2\right)^{-1} \widehat{H}_j^k = \dfrac{(\mathbf{Z}_b^j)^\top \mathbf{Z}_b^j}{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k} \widehat{H}_j^k = \dfrac{(\mathbf{X}_b^k)^\top \mathbf{Z}_b^j}{(\mathbf{X}_b^k)^\top \mathbf{X}_b^k}$. Therefore, if we define the diagonal matrix D_{Z_b} for all $i \in [m]$ by $(D_{Z_b})_i^i := (\mathbf{Z}_b^i)^\top \mathbf{Z}_b^i$, it holds that

$$D_b^{(k)} = D_{Z_b}^{-1/2} \left((\mathbf{X}_b^k)^\top \mathbf{X}_b^k\right)^{1/2}.$$

Using this result, for all $k \in [d]$, we can expand the joint OLS estimate $\widehat{\Pi}^k$ as follows

$$\begin{aligned}
\widehat{\Pi}^k &= \left(\mathbf{Z}_b^\top \mathbf{Z}_b\right)^{-1} \mathbf{Z}_b^\top \mathbf{X}_b^k \\
&= \left(\mathbf{Z}_b^\top \mathbf{Z}_b\right)^{-1} D_{Z_b} \widehat{H}^k \\
&= D_{Z_b}^{-1/2} \widehat{M}_{Z_b}^{-1} D_{Z_b}^{-1/2} D_{Z_b} \widehat{H}^k \\
&= D_{Z_b}^{-1/2} \widehat{M}_{Z_b}^{-1} D_{Z_b}^{1/2} \widehat{H}^k \\
&= D_b^{(k)} \left(\widehat{D}_b^{(k)} M_{Z_b}\right)^{-1} \widehat{H}^k
\end{aligned}$$

Similarly, for all $k, l \in [d]$, the variance-covariance matrix between $\widehat{\Pi}^k$ and $\widehat{\Pi}^l$, can be expanded as follows.

$$
\begin{aligned}
&\widehat{\Sigma}_{\Pi}^{[kl]} \\
&= \left(\mathbf{X}_b^k - \mathbf{Z}_b\widehat{\Pi}^k\right)^\top \left(\mathbf{X}_b^l - \mathbf{Z}_b\widehat{\Pi}^l\right)(\mathbf{Z}_b^\top\mathbf{Z}_b)^{-1} \\
&= \left((\mathbf{X}_b^k)^\top\mathbf{X}_b^l - (\mathbf{X}_b^k)^\top\mathbf{Z}_b\widehat{\Pi}^l - (\widehat{\Pi}^k)^\top(\mathbf{Z}_b)^\top\mathbf{X}_b^l + (\widehat{\Pi}^k)^\top\mathbf{Z}_b^\top\mathbf{Z}_b\widehat{\Pi}^l\right)(\mathbf{Z}_b^\top\mathbf{Z}_b)^{-1} \\
&= \left((\mathbf{X}_b^k)^\top\mathbf{X}_b^l - 2(\widehat{\Pi}^k)^\top\mathbf{Z}_b^\top\mathbf{Z}_b\widehat{\Pi}^l + (\widehat{\Pi}^k)^\top\mathbf{Z}_b^\top\mathbf{Z}_b\widehat{\Pi}^l\right)(\mathbf{Z}_b^\top\mathbf{Z}_b)^{-1} \\
&= \left((\mathbf{X}_b^k)^\top\mathbf{X}_b^l - (\widehat{\Pi}^k)^\top\mathbf{Z}_b^\top\mathbf{Z}_b\widehat{\Pi}^l\right)(\mathbf{Z}_b^\top\mathbf{Z}_b)^{-1} \\
&= \left(\widehat{M}_{X,k}^l - \frac{\mathbf{Z}_b^\top\mathbf{X}_b^k}{(\mathbf{X}_k)^\top\mathbf{X}_b^k}\left(\frac{\mathbf{Z}_b^\top\mathbf{Z}_b}{\sqrt{(\mathbf{X}_k)^\top\mathbf{X}_b^k}\sqrt{(\mathbf{X}_b^l)^\top\mathbf{X}_b^l}}\right)^{-1}\frac{\mathbf{Z}_b^\top\mathbf{X}_b^k}{(\mathbf{X}_b^l)^\top\mathbf{X}_b^l}\right)\left(\frac{\mathbf{Z}_b^\top\mathbf{Z}_b}{\sqrt{(\mathbf{X}_b^k)^\top\mathbf{X}_b^k}\sqrt{(\mathbf{X}_b^l)^\top\mathbf{X}_b^l}}\right)^{-1} \\
&= \left(\widehat{M}_{X,k}^l - (\widehat{H}^k)^\top D_b^{(k)}\widehat{M}_{Z_b}^{-1}D_b^{(l)}\widehat{H}^l\right)D_b^{(k)}\widehat{M}_{Z_b}^{-1}D_b^{(l)}.
\end{aligned}
$$

□

C.3 Proof of Theorem 2.1

Proof. (First statement) Assume Assumption 2.1 (a) and (b) hold. We would like to show that

$$\beta^* \in \underset{\beta\in\mathcal{B}^{\text{sum}}}{\arg\min}\|\beta\|_0.$$

Since $\beta^* \in \mathcal{B}^{\text{sum}}$, it suffices to show that for all $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$, we have $\|\tilde{\beta}\|_0 \geqslant |\text{PA}(Y)|$. Fix a $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$. Since $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$, it holds that $\pi = \Pi\tilde{\beta} = \Pi\beta^*$. Let $S \coloneqq \text{supp}(\tilde{\beta})$. Since $\forall j \in [d]\backslash\text{PA}(Y)$, $(\beta^*)^j = 0$, $\Pi\tilde{\beta} = \Pi\beta^*$ implies that

$$\Pi^S\tilde{\beta}^S = \Pi^{\text{PA}(Y)}(\beta^*)^{\text{PA}(Y)}. \tag{S.6}$$

For the sake of contradiction, suppose that $|S| < |\text{PA}(Y)|$. Then by Assumption 2.1 (a), we have that

$$\text{rank}(\Pi^{\text{PA}(Y)}) = \dim\left(\text{Im}(\Pi^{\text{PA}(Y)}\right) = |\text{PA}(Y)| > |S| \geqslant \dim\left(\text{Im}(\Pi^S)\right) = \text{rank}(\Pi^S).$$

This gives $\text{rank}(\Pi^{\text{PA}(Y)}) > \text{rank}(\Pi^S)$ which implies $\text{Im}(\Pi^{\text{PA}(Y)}) \neq \text{Im}(\Pi^S)$. Then by Assumption 2.1 (b), we have that $\forall w \in \mathbb{R}^{|S|}$, $\Pi^S w \neq \Pi^{\text{PA}(Y)}(\beta^*)^{\text{PA}(Y)}$, but this contradicts (S.6). This concludes the proof of the first statement.

(Second statement) It remains to show that there is no other solutions than β^* when Assumption 2.1 (c) holds. Suppose for the sake of contradiction that there exists $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$ with $S \coloneqq \text{supp}(\tilde{\beta}) = |\text{PA}(Y)|$ and $S \neq \text{PA}(Y)$. Similarly as above, since $\tilde{\beta} \in \mathcal{B}^{\text{sum}}$, (S.6) holds. Then by Assumption 2.1 (c) we have $\text{Im}(\Pi^S) \neq \text{Im}(\text{PA}(Y))$. Moreover, by Assumption 2.1 (a) it holds that

$$\text{rank}(\Pi^{\text{PA}(Y)}) = |\text{PA}(Y)| = |S| \geqslant \text{rank}(\Pi^S).$$

Therefore, by Assumption 2.1 (b) $\forall w \in \mathbb{R}^{|S|}$, $\Pi^S w \neq \Pi^{\text{PA}(Y)}(\beta^*)^{\text{PA}(Y)}$, which again contradicts (S.6). This concludes the proof of the second statement. □

C.4 Proof of Theorem 2.2

Proof. First, observe that using S_{n_a,n_b} as defined in Lemma D.1, we can express the Q statistic for all $\beta \in \mathbb{R}^d$ as

$$Q(\beta) = S_{n_a,n_b}(\beta)^\top S_{n_a,n_b}(\beta).$$

Moreover, for all $\beta \in \mathcal{B}^{\text{sum}}$ it holds by definition that $\mu_{n_a,n_b} = 0$, hence Lemma D.1 implies that $S_{n_a,n_b}(\beta)$ converges uniformly to a standard Gaussian distribution as n_a, n_b tend to infinity and $n_a/n_b \to r$ for $r \in (0,\infty)$. Hence, by the continuous mapping theorem it holds that

$$\lim_{\substack{n_a,n_b\to\infty \\ n_a/n_b\to r}} \sup_{\substack{P\in\mathcal{P}: \\ \beta\in\mathcal{B}^{\text{sum}}(P)}} \sup_{t\in\mathbb{R}} |\mathbb{P}_P(Q(\beta) \leqslant t) - \kappa_m(t)| = 0,$$

which completes the proof of Theorem 2.2. □

C.5 Proof of Theorem 3.1

Proof. Let $r \in (0, \infty)$ and assume that $n_a/n_b \to r$ throughout the proof. Using S_{n_a,n_b} as defined in Lemma D.1, we can express the Q statistic for all $\beta \in \mathbb{R}^d$ as

$$Q(\beta) = S_{n_a,n_b}(\beta)^\top S_{n_a,n_b}(\beta).$$

Furthermore, let $\overline{\mathcal{B}} \subseteq \mathbb{R}^d$ be a compact set and choose $\overline{\beta} \in \overline{\mathcal{B}}$ such that $\inf_{\beta\in\overline{\mathcal{B}}} \|S_{n_a,n_b}(\beta)\|_2^2 = \|S_{n_a,n_b}(\beta^*)\|_2^2$. Then, using standard probability bounds and dropping the n_a, n_b from the notation for simplicity, we get for all $P \in \mathcal{P}$ and all $t \in [0, \infty)$ that

$$\begin{aligned}
&\mathbb{P}_P\left(\inf_{\beta\in\overline{\mathcal{B}}} \|S(\beta)\|_2^2 \leqslant t\right) \qquad \text{(S.7)}\\
&= \mathbb{P}_P\left(\|S(\overline{\beta}) - \mu(\overline{\beta}) + \mu(\overline{\beta})\|_2 \leqslant \sqrt{t}\right)\\
&\leqslant \mathbb{P}_P\left(\big|\|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 - \|\mu(\overline{\beta})\|_2\big| \leqslant \sqrt{t}\right)\\
&= \mathbb{P}_P\left(\|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 - \|\mu(\overline{\beta})\|_2 \leqslant \sqrt{t}, \|S(\overline{\beta}) - \mu(\overline{\beta})\| \geqslant \|\mu(\overline{\beta})\|\right)\\
&\qquad + \mathbb{P}_P\left(\|\mu(\overline{\beta})\|_2 - \|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 \leqslant \sqrt{t}, \|S(\overline{\beta}) - \mu(\overline{\beta})\| \leqslant \|\mu(\overline{\beta})\|\right)\\
&\leqslant \mathbb{P}_P\left(\|S(\overline{\beta}) - \mu(\overline{\beta})\| \geqslant \|\mu(\overline{\beta})\|\right)\\
&\qquad + \mathbb{P}_P\left(\|\mu(\overline{\beta})\|_2 - \|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 \leqslant \sqrt{t}\right)\\
&\leqslant 2\mathbb{P}_P\left(\|S(\overline{\beta}) - \mu(\overline{\beta})\|_2 \geqslant \|\mu(\overline{\beta})\|_2 - \sqrt{t}\right)\\
&\leqslant 2\mathbb{P}_P\left(\sup_{\beta\in\overline{\mathcal{B}}} \|S(\beta) - \mu(\beta)\|_2 \geqslant \inf_{\beta\in\overline{\mathcal{B}}} \|\mu(\beta)\|_2 - \sqrt{t}\right). \qquad \text{(S.8)}
\end{aligned}$$

Next, observe that

$$\begin{aligned}
S(\beta) &= \sqrt{n_b}\left(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1}\right)^{-1/2}(\pi - \Pi\beta)\\
&= \sqrt{n_b}\left(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + (\beta/\|\beta\|_2)^\top\widehat{\Sigma}_X(\beta/\|\beta\|_2)\widehat{\Sigma}_{Z_b}^{-1}\right)^{-1/2}(\pi - \Pi(\beta/\|\beta\|_2)),
\end{aligned}$$

where $\widehat{\Sigma}_X := \frac{1}{n_b}\sum_{i=1}^{n_b}(X_{bi} - \widehat{\Pi}^\top Z_{bi})(X_{bi} - \widehat{\Pi}^\top Z_{bi})^\top$. This in particular implies that S and hence Q does not depend on the norm of β. Moreover, for all $\beta \in \mathbb{R}^d$ with $\|\beta\|_2 = 1$ it holds that

$$\begin{aligned}
\|S(\beta) - \mu(\beta)\|_2 &= \sqrt{n_b}\|(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1})^{-1/2}((\pi - \Pi\beta) - (\widehat{\pi} - \widehat{\Pi}\beta))\|_2\\
&\leqslant \sqrt{n_b}\|\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1}\|_{\mathrm{op}}^{-1/2}(\|\pi - \widehat{\pi}\|_2 + \|\Pi\beta - \widehat{\Pi}\beta\|_2)\\
&\leqslant \left(\lambda_{\min}(\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi) + \lambda_{\max}(\beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1})\right)^{-1/2}(\sqrt{n_b}\|\pi - \widehat{\pi}\|_2 + \sqrt{n_b}\|\Pi - \widehat{\Pi}\|_{\mathrm{op}})\\
&\leqslant \left(\lambda_{\min}(\widehat{\Sigma}_X)\lambda_{\max}(\widehat{\Sigma}_{Z_b}^{-1})\right)^{-1/2}(\sqrt{n_b}\|\pi - \widehat{\pi}\|_2 + \sqrt{n_b}\|\Pi - \widehat{\Pi}\|_{\mathrm{op}})\\
&\leqslant \left(\frac{\lambda_{\min}(\widehat{\Sigma}_{Z_b})}{\lambda_{\min}(\widehat{\Sigma}_X)}\right)^{1/2}(\sqrt{n_b}\|\pi - \widehat{\pi}\|_2 + \sqrt{n_b}\|\Pi - \widehat{\Pi}\|_{\mathrm{op}}).
\end{aligned}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm, and we used Weyl's inequality for the second inequality and that β has norm one for the last inequality. Hence, using the bounds on the minimal eigenvalues of Σ_X and Σ_{Z_b} in Assumption B.1, it holds that

$$\sup_{\beta\in\mathbb{R}^d:\|\beta\|_2=1} \|S(\beta) - \mu(\beta)\|_2 = \mathcal{O}_{\mathcal{P}}(1) \qquad \text{(S.9)}$$

as n_a, n_b tend to infinity, where $\mathcal{O}_{\mathcal{P}}(1)$ denotes a uniformly bounded random variable with respect to $\mathcal{P}$. Finally, for all $s \in [d]$ define $\overline{\mathcal{B}}_s := \{\beta \in \mathbb{R}^d \mid \|\beta\|_0 = s \text{ and } \|\beta\|_2 = 1\}$. Then, using that Q does not depend on the scale

of β and (S.8) we get that

$$\mathbb{P}_P\left(\inf_{\beta:\|\beta\|_0=s} Q(\beta) \leqslant t\right) = \mathbb{P}_P\left(\inf_{\beta\in\overline{\mathcal{B}}_s} Q(\beta) \leqslant t\right)$$
$$\leqslant 2\mathbb{P}_P\left(\sup_{\beta\in\mathbb{R}^d:\|\beta\|_2=1} \|S(\beta)-\mu(\beta)\|_2 \geqslant \inf_{\beta\in\overline{\mathcal{B}}_s} \|\mu(\beta)\|_2 - \sqrt{t}\right). \tag{S.10}$$

Now for the first statement of Theorem 3.1, fix $s \in \mathbb{N}$ such that $s < \|\beta^*\|_0 = |\mathrm{PA}(Y)|$. It follows from Theorem 2.1 that for all $\beta \in \mathbb{R}^d$ with $\|\beta\|_0 = \mathrm{s}$, $\pi - \Pi\beta \neq 0$. Therefore, there exists $\epsilon > 0$ such that for all $\beta \in \mathbb{R}^d$ with $\|\beta\|_0 = s$, it holds that $\|\pi - \Pi\beta\|_2 > \epsilon$. Therefore, by (S.10) it holds that

$$\begin{aligned}
&\lim_{n_a,n_b\to\infty} \mathbb{P}_P\left(\varphi_s(\mathcal{D}_{a,b}^{\text{joint}}) = 1\right)\\
&= \lim_{n_a,n_b\to\infty} \mathbb{P}_P\left(\inf_{\beta:\|\beta\|_0=s} Q(\beta) > \kappa_m(1-\alpha)\right)\\
&\geqslant 1 - \lim_{n_a,n_b\to\infty} 2\mathbb{P}_P\left(\sup_{\beta\in\mathbb{R}^d:\|\beta\|_2=1} \|S(\beta)-\mu(\beta)\|_2 \geqslant \inf_{\beta\in\overline{\mathcal{B}}_s} \|\mu_{n_a,n_b}(\beta)\|_2 - \sqrt{\kappa_m(1-\alpha)}\right)\\
&= 1,
\end{aligned}$$

where we used (S.9) together with

$$\begin{aligned}
\lim_{n_a,n_b\to\infty} \inf_{\beta\in\overline{\mathcal{B}}_s} \|\mu_{n_a,n_b}(\beta)\|_2 &\geqslant \lim_{n_a,n_b\to\infty} \inf_{\beta\in\overline{\mathcal{B}}_s} \sqrt{n_b}\|\tfrac{n_b}{n_a}\widehat{\Sigma}_\pi + \beta^\top\widehat{\Sigma}_X\beta\widehat{\Sigma}_{Z_b}^{-1}\|_{\text{op}}^{-1/2}\epsilon\\
&\geqslant \lim_{n_a,n_b\to\infty} \inf_{\beta\in\overline{\mathcal{B}}_s} \left(\frac{1}{n_a}\lambda_{\max}(\widehat{\Sigma}_\pi) + \tfrac{1}{n_b}\tfrac{\lambda_{\max}(\widehat{\Sigma}_X)}{\lambda_{\min}(\widehat{\Sigma}_{Z_b})}\right)^{-1/2}\epsilon\\
&= \infty,
\end{aligned}$$

where we again used the bounds on the minimal eigenvalues of Σ_X and Σ_{Z_b} in Assumption B.1. Since this holds for all $s \in [d]$ with $s < \|\beta^*\|_0$, we further get

$$\begin{aligned}
\lim_{n_a,n_b\to\infty} \mathbb{P}_P\left(\|\beta^{\leqslant s_{\max}}\|_0 = \|\beta^*\|_0\right) &= \lim_{n_a,n_b\to\infty} \mathbb{P}_P\left(\min_{s<\|\beta^*\|_0} \varphi_s = 1 \text{ and } \varphi_{\|\beta^*\|_0} = 0\right)\\
&= \lim_{n_a,n_b\to\infty} \mathbb{P}_P\left(\varphi_{\|\beta^*\|_0} = 0\right)\\
&\geqslant 1-\alpha.
\end{aligned}$$

For the second statement of Theorem 3.1, we can use the same argument. In this case, Theorem 2.1 implies that for all $c > 0$ there exists $\epsilon > 0$ such that for all $\beta \in \mathbb{R}^d$ with either $\|\beta\|_0 < \|\beta^*\|_0$ or $\|\beta-\beta^*\| > \epsilon$ and $\|\beta\|_0 = \|\beta^*\|_0$ it holds that $\|\pi - \Pi\beta\|_2 > \epsilon$. Therefore, $\mu_{n_a,n_b}(\beta)$ again diverges and the arguments above remain valid. This completes the proof of Theorem 3.1. □

D Additional results

Remark D.1. In the definition of the (empirical) Q statistic in Theorem 2.2, we used

$$\widehat{\Sigma}_\Pi(\beta) \coloneqq \xi(\beta)\widehat{\Sigma}_\Pi\xi^\top(\beta) \tag{S.11}$$

where $\xi(\beta) \coloneqq \beta^\top \otimes I_m$. It follows from the properties of Kronecker product that (S.11) is equivalent to

$$\widehat{\Sigma}_\Pi(\beta) \coloneqq (\beta^\top(\mathbf{X}_b - \mathbf{Z}_b\widehat{\Pi})^\top(\mathbf{X}_b - \mathbf{Z}_b\widehat{\Pi})\beta)(\mathbf{Z}_b^\top\mathbf{Z}_b)^{-1}, \tag{S.12}$$

which aligns with its population quantity $\Sigma_\Pi(\beta) \coloneqq (\beta^\top\mathbb{E}[u_b^X(u_b^X)^\top]\beta)\mathbb{E}[Z_bZ_b^\top]^{-1}$ used in Lemma D.1, where u_b^X is the population residual in (2). The reason why (S.11) is used instead of (S.12) in the Q statistic is that (S.11) only relies on the joint summary statistics, as the individual-level data is not available under the two-sample summary statistics setting.

Lemma D.1. Assume Assumption B.1. Let $\mathcal{D}_{a,b}^{\text{joint}} = \left\{\widehat{\pi}, \widehat{\Sigma}_\pi, \widehat{\Pi}, \widehat{\Sigma}_\Pi\right\}$ be the joint summary statistics based on two independent samples of sizes n_a and n_b, respectively. For all $\beta \in \mathbb{R}^d$, define

$$S_{n_a,n_b}(\beta) \coloneqq \left(\tfrac{1}{n_a}\widehat{\Sigma}_\pi + \tfrac{1}{n_b}\widehat{\Sigma}_\Pi(\beta)\right)^{-1/2}(\widehat{\pi} - \widehat{\Pi}\beta)$$

and

$$\mu_{n_a,n_b}(\beta) \coloneqq \left(\tfrac{1}{n_a}\widehat{\Sigma}_\pi + \tfrac{1}{n_b}\widehat{\Sigma}_\Pi(\beta)\right)^{-1/2}(\pi - \Pi\beta).$$

Then, for all $\beta \in \mathbb{R}^d$ and all $r \in (0, \infty)$ it holds that

$$\lim_{\substack{n_a,n_b\to\infty \\ n_a/n_b\to r}} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^m} \left|\mathbb{P}_P\left(S_{n_a,n_b}(\beta) - \mu_{n_a,n_b}(\beta) \leqslant t\right) - \Phi_m(t)\right| = 0.$$

Proof. Fix an arbitrary $\beta \in \mathbb{R}^d$. Using by standard uniform convergence results for the OLS estimator (e.g., Lundborg et al., 2022, Lemma S10) it holds that

$$\sqrt{n_a}\Sigma_\pi^{-1/2}(\widehat{\pi} - \pi)$$

with $\Sigma_\pi \coloneqq \mathbb{E}[(u_a^Y)^2]\mathbb{E}[Z_a Z_a^\top]^{-1}$ (where u_a^Y are the population residuals in (2) for sample a) converges uniformly w.r.t. $\mathcal{P}$ to a standard m-variate Gaussian distribution as n_a tends to infinity. Similarly, when considering the regression of $\beta^\top X$ on Z, it holds that

$$\sqrt{n_b}\Sigma_\Pi(\beta)^{-1/2}(\widehat{\Pi} - \Pi)\beta$$

with $\Sigma_\Pi(\beta) \coloneqq (\beta^\top\mathbb{E}[u_b^X(u_b^X)^\top]\beta)\mathbb{E}[Z_b Z_b^\top]^{-1}$ (where u_b^X are the residuals in (2) for sample b) converges uniformly w.r.t. $\mathcal{P}$ to a standard m-variate Gaussian distribution as n_b tends to infinity. Combining these results and using that $n_a/n_b \to r$ and $\widehat{\pi}$ and $\widehat{\Pi}$ are estimated based on independent samples, we further have that

$$\sqrt{n_b}\left(\tfrac{1}{r}\Sigma_\pi + \Sigma_\Pi(\beta)\right)^{-1/2}\left((\widehat{\pi} - \widehat{\Pi}\beta) - (\pi - \Pi\beta)\right) \tag{S.13}$$

converges uniformly w.r.t. $\mathcal{P}$ to a standard m-variate Gaussian distribution as n_a and n_b tend to infinity.

Next, we show for all $\epsilon > 0$ that

$$\lim_{n_a\to\infty}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left(\|\widehat{\Sigma}_\pi - \Sigma_\pi\|_{\text{op}} > \epsilon\right) = 0 \quad \text{and} \quad \lim_{n_b\to\infty}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left(\|\widehat{\Sigma}_\Pi(\beta) - \Sigma_\Pi(\beta)\|_{\text{op}} > \epsilon\right) = 0. \tag{S.14}$$

As the proofs for both results are the same we only show it for $\widehat{\Sigma}_\pi$. First, we express the estimator as

$$\widehat{\Sigma}_\pi = \frac{1}{n_a}\sum_{i=1}^{n_a}(Y_{ai} - \widehat{\pi}^\top Z_{ai})^2\left(\frac{1}{n_a}\sum_{i=1}^{n_a} Z_{ai}Z_{ai}^\top\right)^{-1}.$$

We now consider the two product terms separately. Using the uniform law of large numbers (e.g., Klyne and Shah, 2023, Lemma 9) on each component, it holds for all $\epsilon > 0$ that

$$\lim_{n_a\to\infty}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left(\left\|\frac{1}{n_a}\sum_{i=1}^{n_a} Z_{ai}Z_{ai}^\top - \mathbb{E}[Z_a Z_a^\top]\right\|_{\text{op}} > \epsilon\right) = 0. \tag{S.15}$$

Moreover, we can expand the residual variance part as follows

$$\begin{aligned}\frac{1}{n_a}\sum_{i=1}^{n_a}(Y_{ai} - \widehat{\pi}^\top Z_{ai})^2 = \frac{1}{n_a}\sum_{i=1}^{n_a}(Y_{ai} - \pi^\top Z_{ai})^2 + \frac{1}{\sqrt{n_a}}\left(\frac{2}{n_a}\sum_{i=1}^{n_a}(Y_{ai} - \pi^\top Z_{ai})\sqrt{n_a}(\widehat{\pi} - \pi)^\top Z_{ai}\right) \\ + \frac{1}{n_a}\left(\sqrt{n_a}(\widehat{\pi} - \pi)\left(\frac{1}{n_a}\sum_{i=1}^{n_a} Z_{ai}Z_{ai}^\top\right)\sqrt{n_a}(\widehat{\pi} - \pi)\right).\end{aligned}$$

Then, by the uniform asymptotic normality, the bounded moments of Z and Y and a further application of the law of large numbers (e.g., Klyne and Shah, 2023, Lemma 9) it follows for all $\epsilon > 0$ that

$$\lim_{n_a \to \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left(\left| \frac{1}{n_a} \sum_{i=1}^{n_a} (Y_{ai} - \widehat{\pi}^\top Z_{ai})^2 - \mathbb{E}[(u_a^Y)^2] \right| > \epsilon \right) = 0. \tag{S.16}$$

Finally, denote $W_n \coloneqq \frac{1}{n_a} \sum_{i=1}^{n_a} (Y_{ai} - \widehat{\pi}^\top Z_{ai})^2$, $W \coloneqq \mathbb{E}[(u_a^Y)^2]$, $V_n \coloneqq \frac{1}{n_a} \sum_{i=1}^{n_a} Z_{ai} Z_{ai}^\top$ and $V \coloneqq \mathbb{E}[Z_a Z_a^\top]$. Then, by combining (S.15) and (S.16) it follows for all $\epsilon > 0$ that

$$\begin{aligned}
&\sup_{P \in \mathcal{P}} \mathbb{P}_P(\|W_n V_n^{-1} - W V^{-1}\|_{\mathrm{op}} > \epsilon) \\
&\quad \leqslant \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|W_n V_n^{-1} - W_n V^{-1}\|_{\mathrm{op}} > \tfrac{\epsilon}{2}) + \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|W_n V^{-1} - W V^{-1}\|_{\mathrm{op}} > \tfrac{\epsilon}{2}) \\
&\quad \leqslant \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|W_n V_n^{-1} - W_n V^{-1}\|_{\mathrm{op}} > \tfrac{\epsilon}{2}) + \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|W_n V^{-1} - W V^{-1}\|_{\mathrm{op}} > \tfrac{\epsilon}{2}) \\
&\quad \leqslant \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|V_n^{-1} - V^{-1}\|_{\mathrm{op}} \|W_n\|_{\mathrm{op}} > \tfrac{\epsilon}{2}) + \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|W_n - W\|_{\mathrm{op}} \|V^{-1}\|_{\mathrm{op}} > \tfrac{\epsilon}{2}) \\
&\quad \leqslant \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|V_n^{-1} - V^{-1}\|_{\mathrm{op}} \|W_n\|_{\mathrm{op}} > \tfrac{\epsilon}{2}) + \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|W_n - W\|_{\mathrm{op}} C > \tfrac{\epsilon}{2})
\end{aligned}$$

By standard arguments and using the lower bound on the minimal eigenvalue of $V = \mathbb{E}[ZZ^\top]$ from Assumption B.1, this proves (S.14) (left).

Combining the two convergence results in (S.14) and using that $n_a/n_b \to r$ shows that for all $\epsilon > 0$ it holds that

$$\lim_{\substack{n_a, n_b \to \infty \\ n_a/n_b \to r}} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left(\|(\tfrac{n_b}{n_a} \widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta)) - (\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))\|_{\mathrm{op}} > \epsilon \right) = 0. \tag{S.17}$$

Furthermore, we can apply Johnson and Horn (1985, eq. (7.2.13)) to get that

$$\begin{aligned}
&\|(\tfrac{n_b}{n_a} \widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta))^{1/2} - (\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))^{1/2}\|_{\mathrm{op}} \\
&\quad \leqslant \|(\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))^{-1/2}\|_{\mathrm{op}} \|(\tfrac{n_b}{n_a} \widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta)) - (\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))\|_{\mathrm{op}},
\end{aligned}$$

which together with (S.17) and since Assumption B.1 implies that $\inf_{P \in \mathcal{P}} \lambda_{\min}(\frac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta)) > 0$, implies for all $\epsilon > 0$ that

$$\lim_{\substack{n_a, n_b \to \infty \\ n_a/n_b \to r}} \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|(\tfrac{n_b}{n_a} \widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta))^{1/2} - (\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))^{1/2}\|_{\mathrm{op}} > \epsilon) = 0.$$

Together with (S.13) this implies by Klyne and Shah (2023, Lemma 10 (b)) that

$$\lim_{\substack{n_a, n_b \to \infty \\ n_a/n_b \to r}} \sup_{P \in \mathcal{P}} \sup_{t \in \mathbb{R}^m} |\mathbb{P}_P \left(S_{n_a, n_b}(\beta) - \mu_{n_a, n_b}(\beta) \leqslant t \right) - \Phi_m(t)| = 0,$$

where we in particular used that

$$\begin{aligned}
&S_{n_a, n_b}(\beta) - \mu_{n_a, n_b}(\beta) \\
&\quad = (\tfrac{1}{n_a} \widehat{\Sigma}_\pi + \tfrac{1}{n_b} \widehat{\Sigma}_\Pi(\beta))^{-1/2} ((\widehat{\pi} - \widehat{\Pi}\beta) - (\pi - \Pi\beta)) \\
&\quad = \left((\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))^{-1/2} (\tfrac{n_b}{n_a} \widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta))^{1/2} \right)^{-1} \sqrt{n_b} (\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))^{-1/2} ((\widehat{\pi} - \widehat{\Pi}\beta) - (\pi - \Pi\beta)) \\
&\quad = \Big(I + (\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))^{-1/2} \{ (\tfrac{n_b}{n_a} \widehat{\Sigma}_\pi + \widehat{\Sigma}_\Pi(\beta))^{1/2} - (\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))^{1/2} \} \Big)^{-1} \\
&\qquad\qquad \cdot \sqrt{n_b} (\tfrac{1}{r} \Sigma_\pi + \Sigma_\Pi(\beta))^{-1/2} ((\widehat{\pi} - \widehat{\Pi}\beta) - (\pi - \Pi\beta)).
\end{aligned}$$

This completes the proof of Lemma D.1. □

E EXPERIMENT DETAILS AND ADDITIONAL SIMULATION RESULTS

E.1 Details of the simulated experiments in Section 4.1

DGP1: The individual-level data are generated from an SCM (1) with the following parameters

$$A \coloneqq \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad B \coloneqq \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix},$$

$Z \overset{\text{iid}}{\sim} \mathcal{N}_m(0, I_m)$, $H \overset{\text{iid}}{\sim} \mathcal{N}_d(0, I_d)$ and $\nu^X, \nu^Y \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ with $g(H, \nu^X) \coloneqq H + \nu^X$ and $h(H, \nu^Y) \coloneqq H^\top \mathbf{1}_d + \nu^Y$. The true causal effect $\beta^* = (1, 2, 0, 0, 0)$.

DGP2: Let

$$A \coloneqq \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad B \coloneqq \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0, \end{pmatrix}, \quad \mathrm{Var}(Z) \coloneqq \begin{pmatrix} 1 & 0.05 & -0.1 & 0.075 & 0.025 \\ 0.05 & 1 & 0 & 0 & 0 \\ -0.1 & 0 & 1 & 0 & 0 \\ 0.075 & 0 & 0 & 1 & 0 \\ 0.025 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$\mathrm{Var}(\nu^X) \coloneqq WW^\top + I_d$ where $W \in \mathbb{R}^{d \times d}$ with $W_i^j \overset{\text{iid}}{\sim} \mathrm{Unif}(-0.3, 0.5)$ for all $i, j \in [d]$, $\mathrm{Var}(\nu^Y) \coloneqq 1$, and $\mathrm{Cov}(\nu^X, \nu^Y) \in \mathbb{R}^{100}$ such that $\mathrm{Cov}(\nu^X, \nu^Y)^j$ is uniformly sampled from the set $\{0.2, 0.4, 0.6, 0.8\}$ for all $j \in \{1, \ldots, 100\}$.

Then using $\beta^* \in \mathbb{R}^{100}$ with $(\beta^*)^1 \coloneqq 1$, $(\beta^*)^2 \coloneqq 2$, and $(\beta^*)^j \coloneqq 0$ for all $j \in \{3, \ldots, 100\}$, we define $\Pi \coloneqq A^\top (I_d - B)^{-1}$ and $\pi \coloneqq \Pi \beta^*$. Moreover, based on the linear SCM and with $V \coloneqq (I_d - B)^{-1} \beta^*$ we have

$$\begin{aligned} \Sigma_\pi &= \left(V^\top \mathrm{Var}(\nu^X) V + \mathrm{Var}(\nu^Y) + 2 V^\top \mathrm{Cov}(\nu^X, \nu^Y) \right) \mathrm{Var}(Z)^{-1} \text{ and} \\ \Sigma_\Pi &= \mathrm{Var}(\nu^X)^\top (I_d - B)^{-1} \otimes \mathrm{Var}(Z)^{-1}. \end{aligned}$$

We then generated $\widehat{\pi}$, $\widehat{\Pi}$ from the following multivariate Gaussian distributions for a specific sample size n:

$$\widehat{\pi}_n \sim \mathcal{N}(\pi, \frac{1}{n} \Sigma_\pi) \qquad \text{and} \qquad \widehat{\Pi}_n \sim \mathcal{N}(\Pi, \frac{1}{n} \Sigma_\Pi).$$

E.2 Additional simulated experiments

E.2.1 An example where the exclusion restriction criterion is violated

We first provide additional simulation results of a setting where the exclusion restriction criterion of IV is violated. The DGP is described below and the corresponding DAG is given in Figure E1.

DGP3: $m = 5$ and $d = 5$ and $\|\beta^*\|_0 = 2$. For increasing $n \coloneqq n_1 = n_2$, we generate iid $\{(Y_i, Z_i)\}_{i=1}^{n_1}$ and $\{(X_i, Z_i)\}_{i=1}^{n_2}$ according to the following SCM

$$\begin{aligned} X_i &\coloneqq AZ_i + BX_i + H_i + \nu_i^X \\ Y_i &\coloneqq X_i^\top \beta^* + Z_i^\top \gamma + H_i^\top \mathbf{1}_5 + \nu_i^Y, \end{aligned} \tag{S.18}$$

with the following parameters:

$$A = I_5, \quad B \coloneqq \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \quad \gamma = (0.1, 0.1), \quad \beta^* = (1, 2, 0, 0, 0),$$

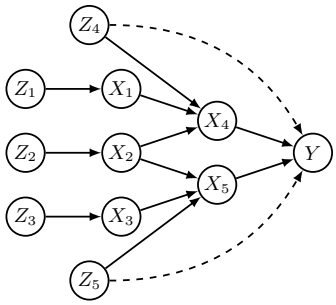

Figure E1: DAG for DGP3 which contains two invalid instruments violating the exclusion restriction criteria (dashed arrows).

$H_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_5)$, and $\nu_i^X, \nu_i^Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. Then we compute the summary statistics using seemingly unrelated regression. The results are shown in Figure E2. The γ parameter in (S.18) represents the violation of the exclusion restriction criteria. We see that as sample size goes larger, the bias and root mean squared error (rmse) continue to decrease. Although the Jaccard similarity and percentage of correct size start to decline, the average true positive rate (tpr) still stays around 100%. In this example, due to the invalid instruments, the estimated causal parents tend to be a superset of the true causal parent, but the estimated effects of the non-parent covaraites are relatively small.

E.2.2 Comparison with `spaceIV`

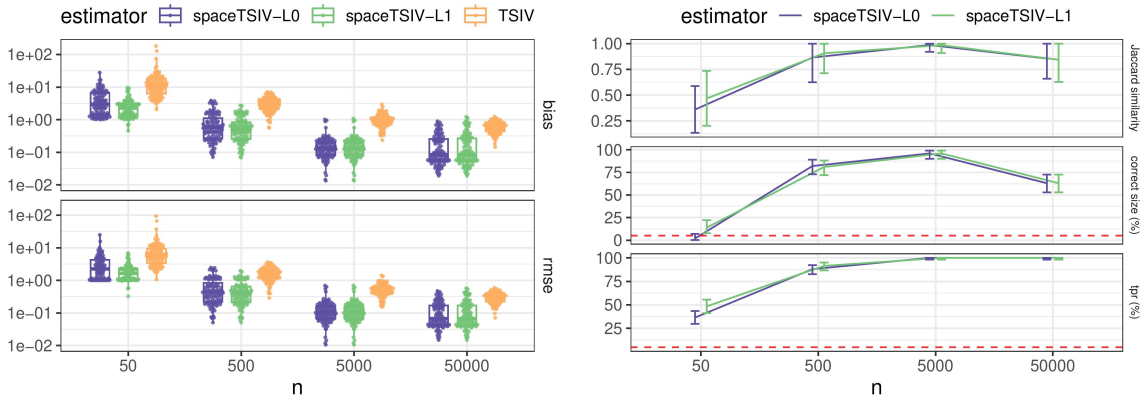

Figure E2: Results using data generated by DGP3 based on 100 repetitions. Left: Bias and rmse of the estimators. The y-axis is on log scale for clarity. Right: Average Jaccard similarity between the selected covariates and the true causal covariates (error bars indicate confidence intervals constructed by mean plus/minus one standard error), percentage of estimates that have the correct support size, and tpr (error bars indicate 95% binomial confidence intervals). This DGP contains 2 invalid instruments among the 5 instruments.

We also compare `spaceTSIV` with `spaceIV` which uses individual-level data, to assess the difference in their finite sample performance. To do this, we simulate individual-level data from DGP1 with different total numbers of observations, we then split this original sample into two sub-samples (s1 and s2) without overlap. We apply `spaceIV` both on the original sample (spaceIV-all) and on the two sub-samples s1 and s2 (spaceIV-s1 and spaceIV-s2). We compute joint summary statistics using Z and X in s1, and using Z and Y in s2, then apply `spaceTSIV` to the two-sample joint summary statistics. The results are shown in Figure E3 to Figure E7. The total number of observations in the original sample is shown on the x-axis, and we split the original sample into sizes of ratios 1:1, 2:3, 3:2, 1:4, and 4:1.

We can see that for each total sample size, `spaceIV` with individual-level data performs better than `spaceTSIV` using two-sample summary statistics. However, bias and rmse both reduce, the Jaccard distance approaches one, and the percentage of estimates having the correct support size goes to 100% for all methods other than `TSIV`

as the sample size increases. Moreover, it seems that there are very small differences in whether s1 or s2 is the larger sample for `spaceTSIV`.

The differences in the finite sample performance of spaceTSIV and spaceIV can be attributed to two reasons: (1) The two-sample summary statistics contain less information than their corresponding full individual-level data; (2) The AR statistic and Q statistic are different statistics, as discussed in Supplementary Material A.1, which means that the objective function being minimized is different even for the same subset of covariates.

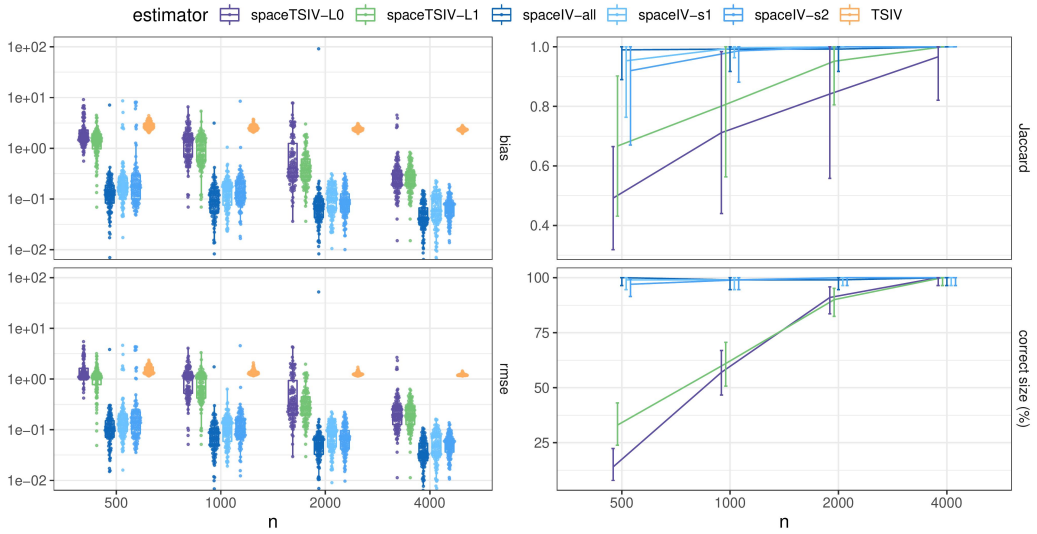

Figure E3: The ratio of the number of observations in s1 and s2 is 1:1.

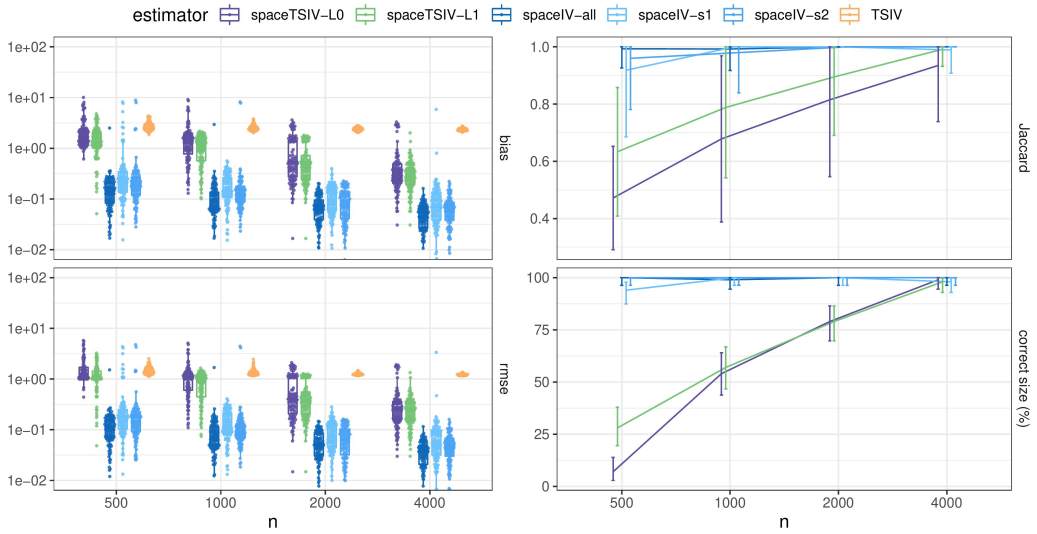

Figure E4: The ratio of the number of observations in s1 and s2 is 2:3.

E.2.3 An example where Assumption 2.1 is violated

We simulate data based on the following DGP which violates Assumption 2.1 (c). Its corresponding DAG is given in Figure E8. Estimation results are provided in Figure E9 and Figure E10. We see from Figure E9 that `spaceTSIV` is indeed not consistent in this case. However, with a large enough sample size, we see in Figure E10 that the estimated sets are most likely the ones containing two variables, one on each path from an instrument to the response.

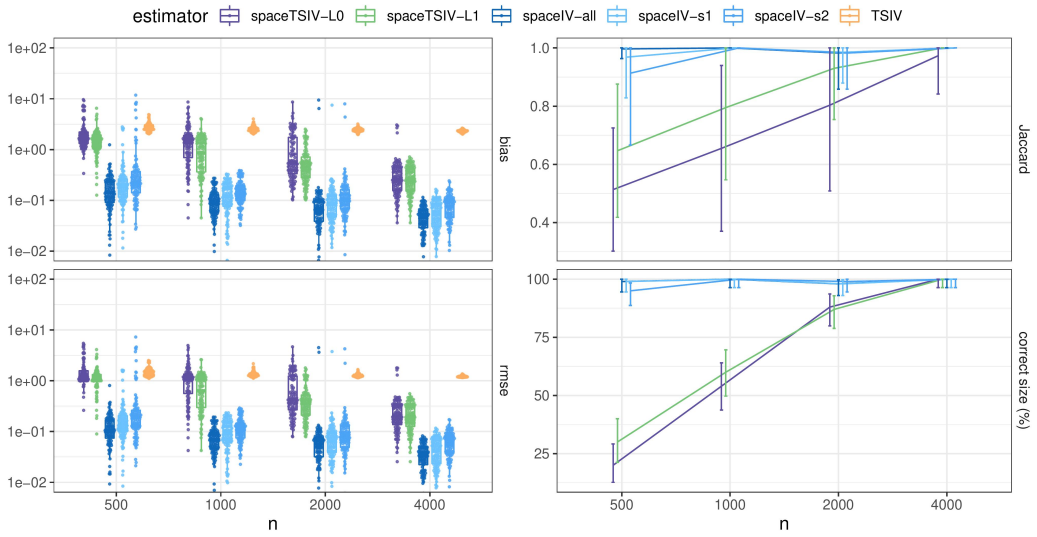

Figure E5: The ratio of the number of observations in s1 and s2 is 3:2.

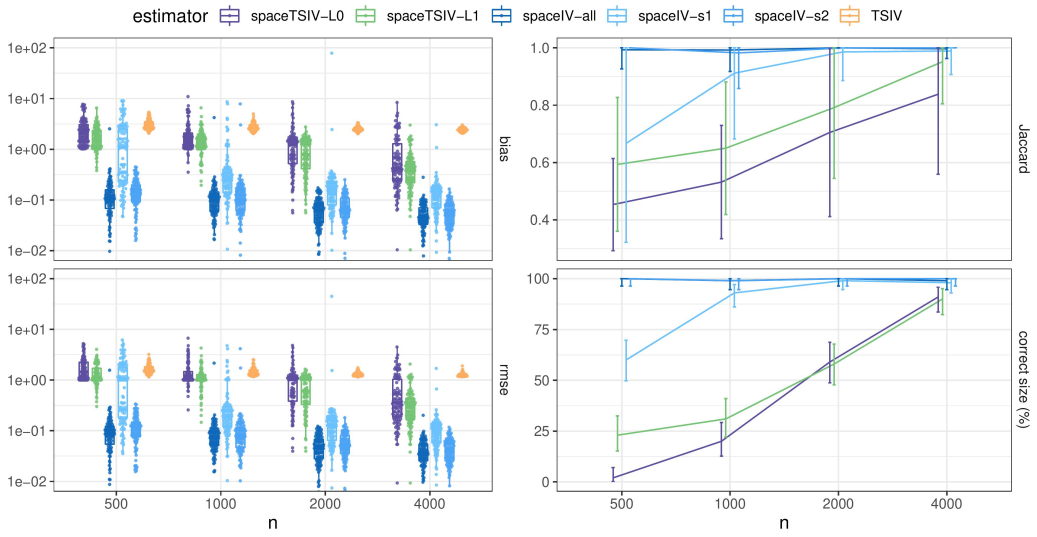

Figure E6: The ratio of the number of observations in s1 and s2 is 1:4.

DGP4: $m = 2$, $d = 4$, and $\|\beta^*\|_0 = 2$. For increasing $n := n_1 = n_2$, we generate iid samples from an SCM (1) with the following parameters

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$Z \overset{\text{iid}}{\sim} \mathcal{N}_m(0, I_m)$, $H \overset{\text{iid}}{\sim} \mathcal{N}_d(0, I_d)$ and $\nu^X, \nu^Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ with $g(H, \nu^X) := H + \nu^X$ and $h(H, \nu^Y) := H^\top \mathbf{1}_d + \nu^Y$. The true causal effect $\beta^* = (0, 0, 1, 1)$.

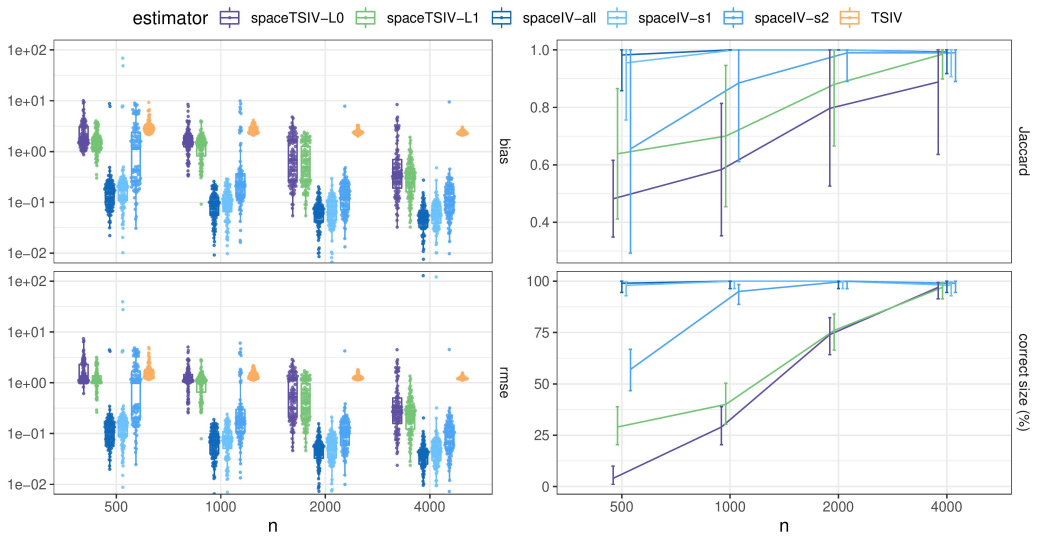

Figure E7: The ratio of the number of observations in s1 and s2 is 4:1.

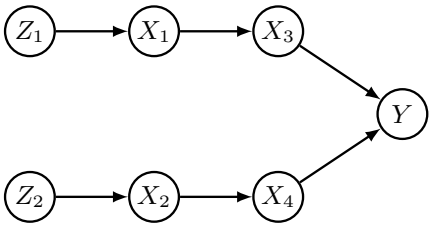

Figure E8: DAG for DGP4 which violates Assumption 2.1 and the causal effect is not identifiable. The hidden variable H is omitted.

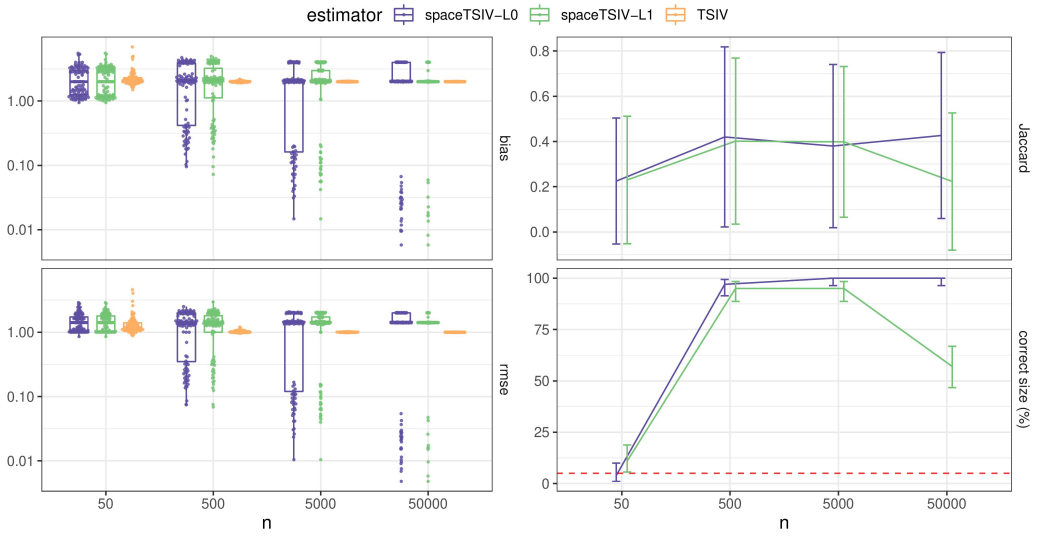

Figure E9: Results using data generated by DGP4 based on 100 repetitions. Left: Bias and rmse of the estimators. The y-axis is on log scale for clarity. Right: Average Jaccard similarity between the selected covariates and the true causal covariates (error bars indicate confidence intervals constructed by mean plus/minus one standard error), percentage of estimates that have the correct support size, and tpr (error bars indicate 95% binomial confidence intervals). This DGP violates Assumption 2.1(c) and the causal effects are not identifiable.

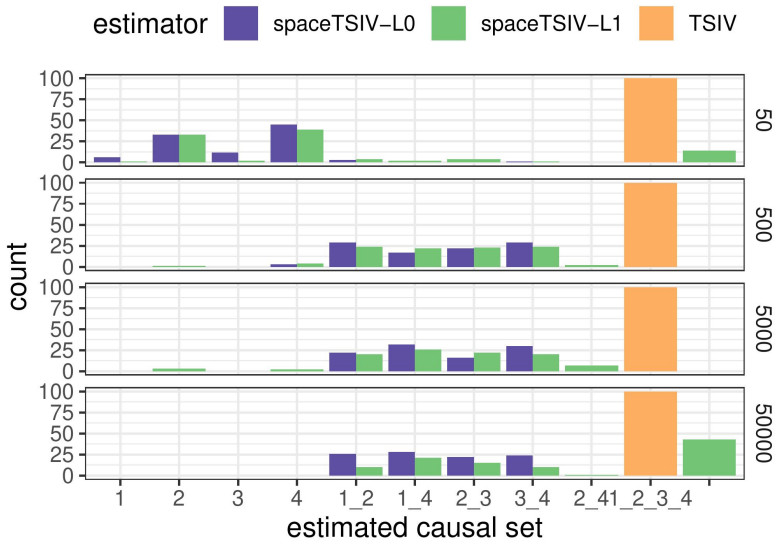

Figure E10: Counts of estimated causal sets using data generated by DGP4 based on 100 repetitions. This DGP violates Assumption 2.1(c) and the causal effects are not identifiable.

References

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63, 1949.

C. R. Johnson and R. A. Horn. *Matrix analysis*. Cambridge university press Cambridge, 1985. doi: 10.1017/CBO9780511810817.

H. Klyne and R. D. Shah. Average partial effect estimation using double machine learning. *arXiv preprint arXiv:2308.09207*, 2023.

M. Londschien and P. Bühlmann. Weak-instrument-robust subvector inference in instrumental variables regression: A subvector lagrange multiplier test and properties of subvector Anderson-Rubin confidence sets. *arXiv preprint arXiv:2407.15256*, 2024.

A. R. Lundborg, I. Kim, R. D. Shah, and R. J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *arXiv preprint arXiv:2304.01098*, 2022.