Correcting Confounding via Random Selection of Background Variables

You-Lin Chen\textsuperscript{1}, Lenon Minorics\textsuperscript{2}, Dominik Janzing\textsuperscript{2}
\textsuperscript{1) Amazon Research Seattle, USA}
\textsuperscript{2) Amazon Research Tübingen, Germany}
\{cyoulin, minorics,janzind\}@amazon.com
February 7, 2022

Abstract

We propose a method to distinguish causal influence from hidden confounding in the following scenario: given a target variable $Y$, potential causal drivers $X$, and a large number of background features, we propose a novel criterion for identifying causal relationships based on the stability of regression coefficients of $X$ on $Y$ with respect to selecting different background features. To this end, we propose a statistic $V$ measuring the coefficient’s variability. We prove, subject to a symmetry assumption for the background influence, that $V$ converges to zero if and only if $X$ contains no causal drivers. In experiments with simulated data, the method outperforms state of the art algorithms. Further, we report encouraging results for real-world data. Our approach aligns with the general belief that causal insights admit better generalization of statistical associations across environments, and justifies similar existing heuristic approaches from the literature.

1 Introduction

Understanding causal relations is crucial for many if not all scientific disciplines. Data-driven inference of causal relations is commonly based on observing statistical dependencies. According to Reichenbach (1956), every statistical dependence between two random variables $X$ and $Y$ is due to $X$ influencing $Y$, $Y$ influencing $X$, or a common cause influencing both $X$ and $Y$. To distinguish between these alternatives is challenging and identification of causal directions from passive observations alone requires strong assumptions on the data generating process, see e.g., Proposition 4.1 in Peters et al. (2017). If $X$ and $Y$ are observed together with additional variables, Markov condition and faithfulness admit identification up to Markov equivalence classes Spirtes et al. (2000). In linear acyclic models with non-Gaussian errors, cause and effect can be identified by independent component analysis Shimizu et al. (2006). In linear structural equation models, identifiability of the exact causal structure (rather than only the Markov equivalence class) can be proved when error terms have equal variances Peters & Bühlmann (2014); Chen et al. (2019). In the setting of a non-linear data-generating process with additive noise, a causal system becomes identifiable under mild conditions Hoyer et al. (2009).

In many practical applications one of the causal directions, say $Y \rightarrow X$, can be excluded (e.g. the patient’s health condition after a medical treatment cannot be the cause of the latter) due to time order, and the remaining challenge consists in inferring to what extent the dependences between $X$ and $Y$ are due to the influence of $X$ on $Y$ and which part is due to a common cause of both. If a complete list of common causes is given, the problem boils down to a statistical problem whose challenge comes from the large number of confounders Imbens & Rubin (2015);
Chernozhukov et al. (2018); Shalit et al. (2017); Raj et al. (2020). Our work considers the same scenario, but does not assume that the full list of confounders is known. This way, identification of the effect of $X$ on $Y$ requires additional assumptions and methodology beyond statistics. Here we assume there are $m$ datasets \{$(Y_i, X_i, E_i)$\}$_{i=1}^m$, possibly from different environments, where \{$E_i$\}$_{i=1}^m$ is a set of background features. For example, data $(Y, X, E_i)$ are collected from different countries, and $E_i$ are population and several economic indicators. For each $(Y, X, E_i)$, we obtain different regression coefficients of $X$. The key insight (provable subject to a symmetry assumption) is that the stability of the regression coefficients with respect to different background covariates $E_i$ is a reasonable criterion for the regression coefficients to be causal. Lu et al. (2021) used a similar heuristic idea to recover a directed gene network in a time series framework but without theoretical justification. This idea is in the spirit of an increasingly popular belief about the following relationship:  
\[
\text{relation causal} \iff \text{robust and invariant statistical association}
\]

Already Haavelmo (1943) argued that causal statements should be robust regarding different perturbations and environments. Recent approaches to causal inference exploit the opposite direction of this relationship. For example, Peters et al. (2016); Bühlmann (2020); Heinze-Deml et al. (2018); Bengio et al. (2020) argued that the predictions from a causal model are more robust under different environments, while predictions from a non-causal model generalize poorly. They further showed that the causal relationship could be identified given samples from different experimental settings. See, however, also critical remarks in Rosenfeld et al. (2021); Kamath et al. (2021).

This work proposes a method to distinguish causal influence from hidden confounding based on the stability of the regression coefficients. Our scenario differs from other theoretical results exploiting the relation between invariance and causality in the literature as follows. First, our environments differ with respect to different background features $E_i$, while the above-mentioned literature typically considers the same variables with distributional shifts. Second, we explicitly allow for unobserved confounders and propose a method to correct for confounding effects. Our method, however, requires an assumption for the prior distribution of the true causal coefficients of the background variables. While we are aware that our assumption is rather restrictive and debatable, correcting for unobserved confounding is an extremely hard problem which is only solvable subject to strong assumptions. As valid for many methods in ML, and causality in particular, the essential question is whether the method shows some robustness regarding violations. Our empirical study with real datasets in Section 6 suggests some robustness in this regard. Moreover, similarly strong assumptions relying on symmetries (e.g. Janzing (2019); Janzing & Schölkopf (2018)) or a confounder that affect multiple measured features (Wang & Blei (2019)) are commonly used in confounder correction. Our contributions can be summarized as follows:

1. We propose a procedure of detecting causal relations based on the stability of regression coefficients with respect to different background features. We examine an example of structural equation models in detail and provide sufficient assumptions under which coefficients’ stability can be used to detect the true causal drivers even in the presence of unobserved confounding.

2. We propose a test statistic to quantify the stability of regression coefficients and show that under a mild condition on the strength of the confounding effect, the test statistic is zero if and only if $X$ is the causal driver of $Y$.

3. We derive a permutation test for inferring the causal influence of $X$ on $Y$ based on our model assumption and proposed statistic.

2 Related work

**Stability of regression coefficients:** Leamer & Leamer (1978); Lu & White (2014); Chen & Pearl (2015) related coefficient robustness to unobserved confounding and causal inference. Oster (2017) built a theory connecting omitted variable bias to coefficient stability and used coefficient’s movement to
estimate a confounding effect. However, their method required the strong assumption that unobservable and observable confounders are independent, which we do not need. Likewise, Imbens (2003) estimated the amount of variation of treatment and variable required to generate significant confounding bias and discuss plausibility of strong bias based on these insights. However, also these estimations assume observable and unobservable confounders to be independent, and also specific parametric models. Cinelli & Hazlett (2019) was based on a similar idea as Imbens (2003), with a different model fit method and relaxing some parametric assumptions.

Correcting hidden confounding: Instrumental variables (White 1982; Muandet et al. 2020b) is a classic approach to correcting for hidden confounding and estimating causal effects. However, the existence of instrumental variables is always debatable. Several recent works proposed new methods utilizing different causal structures and assumptions. Mastakouri et al. (2019) proposed a constraint-based causal feature selection method for identifying potential causes under selection method for identifying potential causes under the setting that there can also be hidden variables acting as common causes, provided that a cause for each candidate cause could be observed. Wang & Blei (2019) used unsupervised machine learning to learn the representation of unobserved confounders in multiple-cause settings. Chernozhukov et al. (2017) assumed the hidden confounding effect is sparse and combined lasso and ridge regression to estimate the bias and the true structure coefficient, and Čevid et al. (2020; Guo et al. 2020) established its asymptotic normality and efficiency in the Gauss-Markov sense. Janzing & Schölkopf (2018) suggested confounder detection methods that are inspired by the intuition that $P_X$ and the conditional $P_{Y|X}$ do not contain information about each other.

Conditional independence tests: Given that the association between $X$ and $Y$ is only confounded by observed variables, high-dimensional conditional independence (which is, however, hard to test Shah & Peters 2020) detects whether $X$ influences $Y$. For linear and parametric methods, Zhang & Zhang (2014; van de Geer et al. 2014) constructed confidence intervals for individual coefficients or linear combinations of some variables and showed asymptotically optimality under certain assumptions. Dezeure et al. (2017) proposed a nonlinear generalization of the partial $R^2$ statistic and showed (without any distributional assumption) that the limit of the proposed statistic is zero if and only if conditional independence holds. Barber & Janson (2020) generalized Goodness-of-fit tests to testing conditional independence.

3 Notions and Problem Setup

This section presents our notions, model assumptions, and the definition of causality via structural equation models.

Notations. Let $S$ be an ordered subset of $\{1, \ldots, q\}$ and $S^c = \{1, \ldots, q\} - S$. $|S|$ denotes the cardinality of $S$. Given a vector $v = (v_1, \ldots, v_q)\top$ and a matrix $A = (a_1, \ldots, a_d)\top$, denote $v_S = (v_i)_{i \in S}$ as the sub-vector of $v$ with respect to $S$ and $A_S = (a_i)_{i \in S}$ as the sub-matrix of $A$. $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution with the mean vector $\mu$ and the covariance matrix $\Sigma$. $X_1 \sim X_2$ means $X_1, X_2$ are equal in law. $I_d$ and $I_{d^2}$ are the identity matrix of unspecified size and of size $d$, respectively. Similarly, $\mathbf{0}$ is zero matrix of unspecified size. Given $a \in \mathbb{R}^d$, $\text{diag}(a) \in \mathbb{R}^{dxd}$ is a matrix in which the entries outside the main diagonal are all zero and diagonal elements are equal to $a$. $\| \cdot \|$ denotes the Euclidean norm. For a real-valued function $f(x)$ and $g(x)$, we write $f(x) = o(g(x))$ and $f(x) = O(g(x))$ if $\lim_{x \to \infty} f(x)/g(x) = 0$ and there exist a constant $c$ such that $\lim_{x \to \infty} f(x)/g(x) = c$, respectively. $X \overset{d}{\rightarrow} Y$ means $X$ converges to $Y$ in probability.

Problem Setup. To understand why regression coefficients’ stability is a reasonable criterion for detecting causal relations, we utilize an example of structural equation models (SEMs) to examine this idea in detail. To be specific, we consider the following linear
Figure 1: Left: An example of our scenario is provided. An arrow indicates a causal effect in the arrow’s direction, e.g., $X$ causes $Y$. $Z$ is a latent variable and introduced for creating a non-trivial relationship between $X$ and $W$. Suppose no dataset contains all variables in $Y, X, W$. Instead, we have three datasets: $\{Y, X, W_{S_1}\}, \{Y, X, W_{S_2}\}, \{Y, X, W_{S_3}\}$ where $S_1 = \{1, 2, 3\}, S_2 = \{5, 6, 7, 8\}, S_3 = \{10, 11, 12\}$. Note that $\{W_{1}, W_{2}\}$ are not contained in any dataset. The existence of the unobserved confounders $\{Z_1, Z_2, W_4, W_5\}$ implies conditional independence tests cannot directly solve our scenario. Right: This figure visualizes the model (1) where the latent variables $Z$ are independent sources that influence $X$ and $W$ at the same time.

$\text{SEM.}$

$$
\begin{align*}
Z & \leftarrow N_Z \in \mathbb{R}^r, \\
W & \leftarrow AZ + N_w \in \mathbb{R}^q, \\
X & \leftarrow BZ + N_x \in \mathbb{R}^d, \\
Y & \leftarrow \beta^\top X + \gamma^\top W + N_Y \in \mathbb{R},
\end{align*}
$$

where $A = (a_1, \ldots, a_r) \in \mathbb{R}^{q \times r}, B = (b_1, \ldots, b_r) \in \mathbb{R}^{d \times r}, N_Z, N_w, N_x, N_Y$ are independent noises such that $\text{Cov}(N_Z) = \text{diag}(\sigma_Z^2)$ for $\square = z, y, w, z$. In SEM (1), $Y$ is an outcome of interest, $X$ represents the potential cause, $W$ are the auxiliary background features, $Z$ is a hidden confounder, and $d, q, r$ are their dimensionality, respectively. The corresponding directed acyclic graph is shown in Figure 1 right.

The causal interpretation of (1) implies that $X$ with $\|\beta\| \neq 0$ is called a cause of $Y$. Our work considers two crucial tasks:

(i) qualitative causal analysis: identify the cause $X$, that is, determine whether $\|\beta\| \neq 0$  

(ii) quantitative causal analysis: estimate $\beta$.

Throughout, we assume that the data $(y_i, x_i, w_i, n_{y,i})_{i=1}^\ell$ consist of $\ell$ independent and identically distributed (i.i.d.) samples that are drawn from the SEM (1) and write $Y = (y_1, \ldots, y_\ell)^\top \in \mathbb{R}^\ell$, $X = (x_1, \ldots, x_\ell)^\top \in \mathbb{R}^{d \times \ell}$, $W = (w_1, \ldots, w_\ell)^\top \in \mathbb{R}^{q \times \ell}$ and $N_y = (n_{y,1}, \ldots, n_{y,\ell})^\top \in \mathbb{R}^\ell$.

Our scenario supposes that there are $m$ nonempty subsets $S_1, \ldots, S_m$ of $\{1, \ldots, q\}$ such that the background feature $E_j$ correspond to possibly different selections $W_{S_j}$, i.e. $E_j = W_{S_j}$. It is worth emphasizing that our idea of detecting causes does not highly depend on the correlation structure between $X$ and $Y$. In fact, our idea only requires the following high-level causal assumptions.

**Assumption 3.1.** (i) All confounding paths between $X$ and $Y$ are blocked by $W$ (ii) $W$ does not contain an effect of $Y$ and its descendants. (iii) $X$ does not influence $W$. (iv) The structural equations are invariant across different $E_j$. (v) There is no systematic bias caused by hidden confounders.

It can be shown that SEM (1) satisfies (i) - (iv) in our causal assumption. (i) and (ii) ensure that interventional probabilities $P(Y|\text{do}(X = x))$ can be computed from the joint distribution of $Y, X, W$. For the simple case of a linear regression model, the linear influence of $X$ on $Y$ is given by regressing $Y$ on $X, W$, and ignoring the regression coefficients corresponding to $W$. However, we would like to emphasize that not all $W$ need to be observed in our scenario. That is, some background features may be hidden (i.e., $\cup_j S_j \neq \{1, \ldots, q\}$). See the left of
Figure 1 for an example). Assumption (iii) rules out the scenario where \(X\) influences \(Y\) indirectly via the mediator \(Y^*\). Assumption (iv) means that the causal relation between \(X\) and \(Y\) remains constant across environments and is a typical assumption in the literature of discussing the relation between invariance and causality. Assumption (v) may be vague at this point, but it is crucial for our method and is the key insight we need to leverage multiple environments and is a typical assumption in the literature.

We have three datasets \(\{S_j\}_{j=1}^3\) for each dataset and obtain the regression coefficients \(\hat{\beta}(S)\). Further, we explain why the stability of regression coefficients can only be a necessary and sufficient condition to detect causal influence if certain assumptions about the data generating process are made. All proofs are deferred to Appendix C and D.

### 4 Identifying causal drivers in the presence of unobserved confounders

In this section, we discuss our proposed method that shows how the variability of regression coefficients can be used to identify causal relations in the presence of unobserved confounders. Before we get to the main result of this section, Theorem 4.3, we first explain the difficulty of identifying causal drivers if we only consider the corresponding regression coefficients. Further, we explain why the stability of regression coefficients can only be a necessary and sufficient property to detect causal influence if certain assumptions about the data generating process are made. All proofs are deferred to Appendix C and D.

#### The problem of biased regression coefficients

Recall the scenario in Figure 1 under which we have infinite samples for each dataset and obtain the regression coefficients \(\hat{\beta}(S_j)\) of \(Y\) on \(X\) given the background features \(W_{S_j}\) for \(j = 1, 2, 3\). A straightforward approach to test whether \(X\) causally influences \(Y\) would be to test whether \(\hat{\beta}(S_j)\) is zero or not. However, due to latent confounders \(\{Z_1, Z_2, W_4, W_9\}\), \(\hat{\beta}(S_j)\) is biased even if we have infinite samples unless \(W_{S_j}\) blocks the confounding path entirely. To see this, for any subset \(S\), \(\hat{\beta}(S)\) is given by

\[
\hat{\beta}(S) = (I_d, 0) \Var (X, W_S)^{-1} \Cov ((X, W_S), Y) = \beta + C(S)\gamma
\]

where \(D_S = \text{diag}(\sigma_{wS})\) and

\[
(C(S)^\top)_S = (D_S^{-1}B(I + B^\top D_S^{-1}B + A_S^\top D_S A_S)^{-1} A_S^\top)^\top,
\]

\[
(C(S)^\top)_{S^c} = 0.
\]

Hence, biased regression coefficients seemingly render the problem of identifying causal drivers by testing \(\hat{\beta} \neq 0\) impossible due to hidden confounders. However, a key observation is that \(\hat{\beta}(S_j)\) is ‘relatively’ close to \(\beta\) (on the scale of \(\|\beta\|\)) for all \(j = 1, 2, 3\) as long as \(\|\beta\|\) is large enough. That is, all vectors \(\hat{\beta}(S_1), \hat{\beta}(S_2), \hat{\beta}(S_3)\) are similar when \(X\) is the cause of \(Y\), which may reveal the causal relation between \(X\) and \(Y\) by virtue of (2). As a result, we might suspect that if \(\hat{\beta}(S_j)\) are close to each other for all \(j\) and are not zero, then \(X\) indeed causally influences \(Y\).

#### The problem of asymmetric hidden confounding

Although we have argued that causal influence implies stable regression coefficients, we do not have the reverse implication. That is, stability of orientation will indicate that \(X\) causally influences \(Y\). In fact, all vectors \(\hat{\beta}(S_1), \hat{\beta}(S_2), \hat{\beta}(S_3)\) can be similar even when \(\|\beta\|\) is small, which leads to stable regression coefficients even if the relation between \(X\) and \(Y\) is purely due to hidden confounding. To see this, consider the scenario in which all elements of \(A\) and \(\gamma\) are one. In that case, the confounders \(C(S_j)\) are similar for all \(S_j\), and so \(\hat{\beta}(S_1), \hat{\beta}(S_2), \hat{\beta}(S_3)\) are close to each other even if \(\|\beta\| = 0\). This argument shows that we need additional assumptions to ensure that stable regression coefficients imply true causal relation.

The example above shows that, in general, stable regression coefficients are not enough to ensure true causal relation. Hence, we need to restrict the confounding behavior in the sense of Assumption 3.1(v).
Formally, we use the notion of spherical symmetry to obtain non-biased hidden confounding.

**Definition 4.1.** (Bryc 2012) A random vector $\nu = (\nu_1, \ldots, \nu_q)^\top$ is spherically symmetric if $e^\top \nu \sim \nu_1$, for all $e$ provided that $\|e\| = 1$.

Examples of spherically symmetric distributions include multivariate normal distributions with zero mean and identity covariance matrix and uniform distributions on the unit sphere. The former is a typical prior distribution in Bayesian statistics, and the latter provides an example where the components $\{\nu_j\}_{j=1}^q$ are not necessarily independent. Spherical symmetry is a popular notion in multivariate analysis and has a wide range of applications in Bayesian statistics and shape analysis in object recognition (Hamsici 2014). As we will see later, this problem can be solved as vectors $v_j$ in $\mathbb{R}^d$.

**Definition 4.2** (causal links between coefficients and relates its asymptotic properties to statistic that relies on the variability of the regression remain stable. In particular, we develop a stability formalize what it means that regression coefficients after we discussed our assumption on the prior distribution, we now require strong assumptions. It is common in causal inference that it is possible to establish that $\parallel \beta \parallel = 0$ and (ii) estimate $\beta$. Note that all results are established in the infinite sample limit, and the randomness here only comes from the prior.

**Theorem 4.3.** Let $S_1, \ldots, S_n \subseteq \{1, \ldots, q\}$ and $\|\beta\| = 0$. Assume that $\gamma$ is a spherically symmetric random vector such that $E\|\gamma\|^4 = O(\rho_\gamma^4/q^2)$ and that

$$V := V(v_1, \ldots, v_m) = 1 - \frac{\|m^{-1} \sum_{j=1}^m v_j\|^2}{m^{-1} \sum_{j=1}^m \|v_j\|^2};$$

Note that if $v_1 = \cdots = v_m$, then $V = 0$. Further, if $v_1, \ldots, v_m$ are uniformly distributed on the unit sphere, we get $V \to 1$ as $m \to \infty$. The closer $v_1, \ldots, v_n$ are, the smaller $V$ gets. Hence, $V$ quantifies the variability of $v_1, \ldots, v_n$.

We are ready to present our main result addressing the equivalence between causation and stable regression coefficients. In particular, our main theorem enables us to solve our crucial tasks: (i) infer if $\parallel \beta \parallel \neq 0$ and (ii) estimate $\beta$. We admit that one can easily construct real-world examples where spherical symmetry is violated (for instance, if many features have a positive influence on the target). However, the qualitative result regarding the relation between regression stability weak confoundedness persists for many other priors. Further, it is common in causal inference that *provable* results require strong assumptions.

**The equivalence between causation and stable regression coefficients:** After we discussed our assumption on the prior distribution, we now formalize what it means that regression coefficients remain stable. In particular, we develop a stability statistic that relies on the variability of the regression coefficients and relates its asymptotic properties to causal links between $X$ and $Y$.

**Definition 4.2 (Stability statistic).** Given some vectors $\{v_j\}_{j=1}^m$ in $\mathbb{R}^d$, the stability statistic is defined as

$$V := V(v_1, \ldots, v_m) = 1 - \frac{\|m^{-1} \sum_{j=1}^m v_j\|^2}{m^{-1} \sum_{j=1}^m \|v_j\|^2};$$

where $C(S)$ is defined in (3). Then, it holds that (i) $\frac{1}{m} \sum_{j=1}^m \beta(S_j) \stackrel{p}{\to} \beta$ and (ii) $V(\hat{\beta}(S_1), \ldots, \hat{\beta}(S_m)) \sim \begin{cases} 0 & \text{if } \rho_\beta > 0 \\ c & \text{if } \rho_\beta = 0 \end{cases}$.

where $c > 0$ is a deterministic constant depending on $\{C(S_j)\}_{j=1}^m$.

(i) in Theorem 4.3 shows that the confounding effect can be corrected by averaging when the strength of confounding increases slowly enough in $q$, and (ii) in Theorem 4.3 justifies that $V$ indeed is a reasonable test statistic for:

$$H_0 : \parallel \beta \parallel = 0, \quad H_A : \parallel \beta \parallel \neq 0.$$
This observation means that some unknown model variables \( W \), i.e., we can have \( \bigcup_{j=1}^m S_j \neq \{1, \ldots, q\} \). Thus, Theorem 4.3 addresses also the case of hidden confounding. In Appendix C.3 we give two examples where the conditions of Theorem 4.3 are satisfied.

We would like to emphasize that our idea of detecting causes and its analysis does not highly depend on the generative process (1), and our simplified models are needed only to allow us to study and present our idea for the relationship between causality and regression coefficients’ stability in a simple manner. The goal of this paper is to understand why regression coefficients’ stability is a reasonable criterion for detecting causal relations instead of deriving theorems in the most general form.

5 Statistical testing for causal influence

This section discusses how to conduct statistical tests for hypothesis (6). As discussed, Theorem 4.3 shows that \( V \) converges a positive number under the null hypothesis \( H_0 \) and converges to zero under the alternative hypothesis when the sample size is infinite and \( q \) goes to infinity. Given a realization of \( \gamma \), we do not have any finite sample statements about \( V \), which depend on the causal structure parameters and the noise \( N_y \). In particular, when the noise \( N_y \) is normally distributed, the numerator and denominator of \( V \) follows two correlated noncentral chi-square distributions whose covariance depends on \( \{S_j\} \) and the model parameters \( A \) and \( B \). In other words, \( V \) is not a pivotal statistic, i.e., \( V \) is a statistic whose null distribution depends on some unknown parameters. This observation means that some unknown model parameters are needed to compute the null distribution of \( V \) as well as \( p \)-values. However, it may be intractable to infer the parameters of the model due to high dimensionality. Even worse, the model may be misspecified. To address those problems, we recommend permutation testing.

### Algorithm 1 Permutation Test

**input:** \( m \) datasets \( \{(Y, X, W_j)\}_{j=1}^m \), an estimation of \( \gamma \), the number of permutations \( M \).

1. Compute \( V_0 = V(\hat{\beta}(S_1), \ldots, \hat{\beta}(S_m)) \) via (4) where \( \hat{\beta}(S) \) are the regression coefficients of \( X \) regressing \( Y \) on \( X, W \).
2. for \( i = 1 \) to \( M \) do
   1. Generate permutations \( \pi_i \) uniformly.
   2. Compute
      \[
      \hat{Y}(\pi_i) = W \hat{\gamma} + \hat{N}^{\pi_i},
      \]
      where \( \hat{N} = Y - W \hat{\gamma} = (n_{y,1}, \ldots, n_{y,q})^\top \) and \( \hat{N}^{\pi_i} = (n_{y,\pi(1)}, \ldots, n_{y,\pi(q)})^\top \).
   3. Compute \( V_i = V(\hat{\beta}^{(i)}(S_1), \ldots, \hat{\beta}^{(i)}(S_m)) \) via (4) where \( \hat{\beta}^{(i)}(S) \) is regression coefficients of \( X \) from regressing \( \hat{Y}(\pi_i) \) on \( X, W \).
3. end for

**output:** \( p \)-value \( p = (\sum_{i=1}^M \mathbb{1}\{V_i \geq V_0\} + 1)/(M + 1) \).

The following theorem shows how the type I error can be controlled, which thus justifies Algorithm 1. In particular, it proves that \( (\mathbb{1}\{V \}, \hat{Y}(\pi), X, W) \) is approximately exchangeable, where \( \hat{Y}(\pi) \) is defined in (7). A similar argument...
for permutation tests also appears in Berrett et al. (2019); Barber & Janson (2020).

**Theorem 5.1.** Suppose that \( N_y \sim \mathcal{N}(0, \sigma_y^2) \) and \( M \) permutations \( \{\pi_i\}_{i=1}^M \) are generated uniformly. Given an estimation \( \hat{\gamma} \) of \( \gamma \) and the \( p \)-value \( p \) obtained by Algorithm 1, under the null hypothesis \( H_0 \) in (6), the type I error is bounded from above by

\[
P(p \leq \alpha \mid X, W, \gamma, \hat{\gamma}) \leq \alpha + \frac{\sqrt{M} \|W(\gamma - \hat{\gamma})\|}{2\sigma_y}.
\]

Hence, the type I error is small whenever the prediction error is small. Lasso, principal components regression, variable subset selection, and partial least squares all can be used to get the residuals in the high dimensional setting. In particular, ridge regression and model averaging are more desirable due to their superior ability to minimize the prediction error.

6 Numerical studies

In this section, synthetic and real datasets are used to gain additional insight into the performance of our method. Further, we compare our proposed method against existing tests for hypothesis (6). The code for all experiments is given in the supplementary material.

6.1 Simulations

**Data generating procedure:** The data \((Y, X, W)\) is generated via SEM (1) with the following parameters: (1) Draw the entries of \( A, B \) from independent Gaussian distributions \( \mathcal{N}(0, 1/d) \) and \( \mathcal{N}(0, 1/q) \), respectively. (2) Draw the parameters \( \beta, \gamma \) from the uniform distribution on the sphere with radius \( \rho_\beta, \rho_\gamma \), respectively. (3) Draw \( \ell \) samples of each \( N_\square \) independently from the standard multivariate Gaussian distribution \( \mathcal{N}(0, I) \) for \( \square = z, w, x, y \).

**Random selection procedure (RS):** If we only have a single dataset with a large number of background features \( W \) rather than multiple datasets, we artificially generate multiple datasets \( \{(Y, X, W_{S_j})\}_{j=1}^m \) by uniformly sampling \( m \) subsets \( S_j \subset \{1, \ldots, q\} \) with size \( |S_j| = k \). We call this method random selection procedure.

**Algorithm 1+2+RS:** To apply the permutation test in the numerical studies, we (i) use the random selection procedure to create \( \{(Y, X, W_{S_j})\}_{j=1}^m \), (ii) execute Algorithm 2 in Appendix B to estimate \( \gamma \), and (iii) run Algorithm 1 to test the hypothesis in (6). For the ease of notation, we often refer to RS when we mean the procedure Algorithm 1+2+RS.

**Basis results:** We generate data \((Y, X, W)\) by using the data generating procedure with the following parameters: \( d = 1, q = 1000, r = 5, \ell = 100, \rho_\gamma = 10 \) and \( \rho_\beta = 0, 0.01, 0.1, 1.0 \). For each \( \rho_\beta \), we repeated the experiment 1000 times. Note that, \( \beta \) and \( \gamma \) are only drawn once and were then fixed for all 1000 repetitions. The only randomness comes from generating new noises \( N_y \) for each experiment. The result of running Algorithm 1+2+RS with \( m = 100, M = 1000, k = 10 \) is shown in Figure 2 (left and middle). We can see in the left figure that the distribution of \( V \) concentrates at zero for \( \rho_\beta = 1 \), which aligns with our findings in Theorem 5.1. The middle figure in Figure 2 shows that the null distribution generated by the permutation test is similar to the true null distribution for different \( \rho_\beta \), which verifies Theorem 5.1 empirically.

**Results of varying parameters:** Again, we use the data generating procedure but with the parameters \( d = 3, r = 5, \ell = 300, \rho_\gamma = 2.5 \), and Algorithm 1+2+RS with \( m = 40, M = 400, k = 3 \) to investigate the impact of different choices of \( q \). The result is shown in Figure 2 (right), thereby we repeat the experiments 500 times and take the average for each point in the figure. We can see that increasing the number of background features can increase the power that approaches one, which aligns with Theorem 4.3. We also conduct experiments with varying \( \rho_\beta, d, \) and \( k \) in Appendix F.

**Comparisons:** Next, we compare our proposed approach (Algorithm 1+2+RS) to existing methods. Thereby, we use the data generating procedure to produce 1000 data for the null and alternative hypothesis with \( q = 1000 \) (see caption in Table 1 for all hyperparameters), where we consider two settings:

**Setting 1:** All \( W_1, \ldots, W_{1000} \) are accessible.

**Setting 2:** \( W_1, \ldots, W_{700} \) are latent, and only \( W_{701}, \ldots, W_{1000} \) are accessible.

In this comparison, we run Algorithm 1+2+RS with \( m = 1000, k = 10 \) and \( M = 1000 \), referred to as "RS".
Figure 2: Left: the true distribution of the variation $V$. Middle: the null distribution generated by the permutation test. Right: The Power and Type I error with varying the number of background features.

Table 1: $q = 1000, r = 5, \ell = 100, k = 10, m = 1000$ and $\rho_\gamma = 10, \rho_\beta = 0, 1.5$.  

|               | Setting 1 | Setting 2 |
|---------------|-----------|-----------|
| $\alpha$      | RS JS BM FL DR | RS JS BM FL DR |
| type I        | 0.05 0.01 | 0.05 0.01 |
| error         | 0.001 0.000 | 0.001 0.000 |
| power         | 0.029 0.017 | 0.025 0.005 |
|               | 0.009 0.005 | 0.024 0.005 |

The following methods are included in our experiment:

(i) "JS" is a test for non-confounding proposed in Janzing & Schölkopf (2018) and particularly designed for the case that all elements of $W$ are hidden. Under an ICA-based model, the authors derive a test statistic for the null hypothesis $\rho_\beta/\rho_\gamma = 0$ and a simple approximation of the null distribution of their proposed statistic.

(ii) "BM" is a high-dimensional test based on ridge projections, proposed in Bühlmann (2013). This test is based on a bias-corrected estimation and an asymptotic upper bound of its distribution.

(iii) "FL" is the Freedman-Lane HD test proposed by Hemerik et al. (2020) with the test statistic based on the generalized partial correlation.

(iv) "DR" is the Double Residualization method in Hemerik et al. (2020), which residualizes both $Y$ and $X$ and tests the sample correlation.

Note that FL and DR use ridge regression for $\gamma$, while our method uses model averaging (Algorithm 2). BM, FL, and DR are designed for the conditional independence tests and do not consider the presence of unobserved confounders. They only use the assumption of linear models and do not restrict any generating processes of $\gamma$. In contrast, RS and JS aim to correct or detect hidden confounding and rely on strong assumptions.

The summary is shown in Table 1. We can see that RS shows the highest power and its type I error is closest to $\alpha$ in both settings. It may be counter-intuitive that FL and DR work better in Setting 2, but since $\gamma$ is drawn from the prior with zero mean, the confounding effect will be offset when $q$ is large. Thus, the prediction error increases in Setting 1 due to this particular prior. We would like to remind readers that BM, FL, and DR are designed for more general model settings. JS does not use any information about $W$ and does not take into account noise terms $N_y$, which may explain the bad performance of JS. This observation also shows the importance of permutation tests that do not rely on the model’s parameters. Due to the limit of space, we include additional experiments in Appendix F where we also examine more settings and model parameters.
6.2 Real datasets

This section presents an analysis for a real dataset about the educational attainment of teenagers [Rouse 2015]. We work with simplified data as provided in Stock et al. [2012]. In this data, for 4739 pupils from approximately 1100 US high schools in different states, 13 attributes were recorded in 1980. We split the whole dataset into several sub-datasets based on the attribute of the state hourly wage in manufacturing ("stwmfg80"). In this experiment, we want to examine how different attributes of pupils affect the composite test score denoted by $Y$. The attribute "years of education completed", the total years that students get their final degree, and "tuition", the average intuition in four-year state college for each state, are dropped due to the time order and collinearity, respectively. Based on Assumption 3.1, our method requires that causal influences of $X$ on $Y$ are invariant across different environments. Thus, the attributes we choose to test for potential causal influence are: (1) The boolean information of whether the parents are college graduates, the corresponding features are denoted by "momcoll" and "dadcoll", and (2) The distance from a four-year college to students’ school, denoted by "dist". We only use those sub-datasets whose sample size is more than 70. As a result, 11 attributes, 3908 samples, and 20 sub-dataset remain in this analysis. Other attributes include as background features in our analysis are gender ("Female"), race ("black", "Hispanic”), scores on relevant achievement tests ("bytest"), whether the family income is larger than 25,000 ("income"), whether the family owns a house ("own home"), whether the student’s school is in an urban area ("urban") and the unemployment rate of the country where the school is located ("cue80").

We perform our Algorithm 1+2 with $M = 1000$ and run the ordinary least squares (OLS) regression for baseline comparison. For the complete result of the ordinary linear regression, see Appendix G. The result is summarized in Table 2. We can see for OLS that all features are statistically significant as those p-values are extremely small. This result is not reasonable since "dist" should not causally influence students' composite test scores. Indeed, it is highly possible that some hidden confounders exist between "dist", and another attribute as "dist" is a proxy for the type of schools. In contrast to OLS, our proposed method produces a large p-value for "dist". On the other hand, "momcoll" and "dadcoll" can influence composite test scores in many aspects. For example, the parents who are college graduates may emphasize education more. Both methods yield a small p-value for "momcoll" and "dadcoll", which verifies our reasoning. Note that we test "momcoll" and "dadcoll" together in our proposed method. Even though we do not know the ground truth in this case, the results from Algorithm 1 are more plausible than OLS regression because of potentially hidden common causes.

Table 2: p-values conducted by OLS and our proposed procedure

| Method       | momcoll | dadcoll | dist  |
|--------------|---------|---------|-------|
| Algorithm 1  | 0.0     | 0.970   |       |
| OLS          | 3.10e-09| 1.36e-18| 1.08e-06|

7 Conclusions and discussions

We proposed a procedure to identify the linear influence of a feature on a target variable in a scenario with multiple features as background conditions. We showed that stability of regression coefficients with respect to different selections of background variables (formalized by our stability statistic $V$) indicates that the statistical association relies on a direct causal influence, as opposed to hidden confounders. Although assumption 3.1 might seem very restrictive and needs to be carefully examined in practice, one should be aware of the difficulty of inferring causal relations from observational data in the presence of hidden confounders. Hence, it is natural to postulate assumptions about the underlying generation process to be able to prove correctness of methods which hopefully work also under more general circumstances. Since our model is based on a large number of possibly cor-
related confounders it complements previous studies using stability of regression coefficients as indicators for associations to be causal.

References

Anderson, M. J. and Legendre, P. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, 62(3):271–303, feb 1999. doi: 10.1080/00949659908811936.

Azadkia, M. and Chatterjee, S. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6), dec 2021. doi: 10.1214/21-aos2073.

Barber, R. F. and Janson, L. Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *arXiv preprint arXiv:2007.09857*, 2020.

Bengio, Y., Deleu, T., Rahaman, N., Ke, N. R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxWIgBFPS.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, oct 2019. doi: 10.1111/rssb.12340.

Bryc, W. *The normal distribution: characterizations with applications*, volume 100. Springer Science & Business Media, 2012.

Bühlmann, P. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4), sep 2013. doi: 10.3150/12-bejsp11.

Bühlmann, P. Invariance, causality and robustness. *Statistical Science*, 35(3), aug 2020. doi: 10.1214/19-sts721.

Čevid, D., Bühlmann, P., and Meinshausen, N. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21(232):1–41, 2020. URL http://jmlr.org/papers/v21/19-545.html.

Chen, B. and Pearl, J. Exogeneity and robustness. Technical report, Tech. Rep, 2015.

Chen, W., Drton, M., and Wang, Y. S. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, sep 2019. doi: 10.1093/biomet/asz049.

Chernozhukov, V., Hansen, C., and Liao, Y. A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1), feb 2017. doi: 10.1214/16-aos1434.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, jan 2018. doi: 10.1111/ectj.12097.

Cinelli, C. and Hazlett, C. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, dec 2019. doi: 10.1111/rssb.12348.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. High-dimensional simultaneous inference with the bootstrap. *TEST*, 26(4):685–719, oct 2017. doi: 10.1007/s11749-017-0554-2.

Durrett, R. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Eberhardt, F. and Scheines, R. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, dec 2007. doi: 10.1086/525638.

Fan, J., Feng, Y., and Xia, L. A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *Journal of Econometrics*, 218(1):119–139, sep 2020. doi: 10.1016/j.jeconom.2019.12.016.
Fang, K.-T., Kotz, S., and Ng, K. W. *Symmetric multivariate and related distributions*. Chapman and Hall/CRC, 2018.

Feng, Y., Liu, Q., and Okui, R. On the sparsity of mallows model averaging estimator. *Economics Letters*, 187:108916, feb 2020. doi: 10.1016/j.econlet.2019.108916.

Fisher, R. A. ”the coefficient of racial likeness” and the future of craniometry. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 66:57, jan 1936. doi: 10.2307/2844116.

Fourdrinier, D. and Strawderman, W. E. Robust generalized Bayes minimax estimators of location vectors for spherically symmetric distributions with unknown scale. Institute of Mathematical Statistics, 2010. doi: 10.1214/10-imscoll617.

Freedman, D. and Lane, D. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, oct 1983. doi: 10.1080/07350015.1983.10509354.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2007/file/3a0772443a0739141292a5429b952fe6-Paper.pdf.

Guo, Z., Čević, D., and Bühlmann, P. Doubly debiased lasso: High-dimensional inference under hidden confounding and measurement errors. *arXiv preprint* [arXiv:2004.03758] 2020.

Haavelmo, T. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1, jan 1943. doi: 10.2307/1905714.

Hamsici, O. and Martinez, A. Rotation invariant kernels and their application to shape analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1985–1999, nov 2009. doi: 10.1109/tpami.2008.234.

Hansen, B. E. Leasts squares model averaging. *Econometrica*, 75(4):1175–1189, 2007. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/4502024.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), sep 2018. doi: 10.1515/jci-2017-0016.

Hemerik, J., Thoresen, M., and Finos, L. Permutation testing in high-dimensional linear models: an empirical investigation. *Journal of Statistical Computation and Simulation*, 91(5):897–914, nov 2020. doi: 10.1080/00949655.2020.1836183.

Hjort, N. L. and Claeskens, G. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003. ISSN 01621459. URL http://www.jstor.org/stable/30045339.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2008/file/f7664060cc52bc56f3d20bcedc94a4b6-Paper.pdf.

Imbens, G. W. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, apr 2003. doi: 10.1257/000282803321946921.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Janzing, D. Causal regularization. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/2172fde4930104727062897085e4319d-Paper.pdf.
Janzing, D. and Schölkopf, B. Detecting non-causal artifacts in multivariate linear regression models. In Dy, J. and Krause, A. (eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 2245–2253. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/janzing18a.html

Kamath, P., Tangella, A., Sutherland, D., and Srebro, N. Does invariant risk minimization capture invariance? In Banerjee, A. and Fukumizu, K. (eds.), Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pp. 4069–4077. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/kamath21a.html.

Leamer, E. E. and Leamer, E. E. Specification searches: Ad hoc inference with nonexperimental data, volume 53. Wiley New York, 1978.

Lu, J., Dumitrascu, B., McDowell, I. C., Jo, B., Barrera, A., Hong, L. K., Leichter, S. M., Reddy, T. E., and Engelhardt, B. E. Causal network inference from gene transcriptional time-series response to glucocorticoids. PLOS Computational Biology, 17(1):e1008223, jan 2021. doi: 10.1371/journal.pcbi.1008223.

Lu, X. and White, H. Robustness checks and robustness tests in applied economics. Journal of Econometrics, 178:194–206, jan 2014. doi: 10.1016/j.jeconom.2013.08.016.

Maruyama, Y. and Strawderman, W. E. Robust bayesian variable selection in linear models with spherically symmetric errors. Biometrika, 101(4):992–998, 2014. ISSN 00063444. doi: 10.2307/43304704. URL http://www.jstor.org/stable/43304704

Maruyama, Y. and Takemura, A. Admissibility and minimaxity of generalized bayes estimators for spherically symmetric family. Journal of Multivariate Analysis, 99(1):50–73, jan 2008. doi: 10.1016/j.jmva.2007.01.002.

Mastakouri, A., Schölkopf, B., and Janzing, D. Selecting causal brain features with a single conditional independence test per feature. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R. (eds.), Advances in Neural Information Processing Systems, pp. 12553–12564. Curran Associates, Inc., 2019.

Muandet, K., Jitkrittum, W., and Kübler, J. M. Kernel conditional moment test via maximum moment restriction. In Proceedings of the 36th International Conference on Uncertainty in Artificial Intelligence (UAI), volume 124 of Proceedings of Machine Learning Research, pp. 41–50. PMLR, August 2020a. URL http://proceedings.mlr.press/v124/muandet20a.html.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. Dual instrumental variable regression. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), Advances in Neural Information Processing Systems, volume 33, pp. 2710–2721. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/file/1c383cd30b7c298ab50293adfeceb7b18-Paper.pdf.

Oster, E. Unobservable selection and coefficient stability: Theory and evidence. Journal of Business & Economic Statistics, 37(2):187–204, jun 2017. doi: 10.1080/07350015.2016.1227711.

Pearl, J. Causality. Cambridge university press, 2009.

Peters, J. and Bühlmann, P. Identifiability of gaussian structural equation models with equal error variances. Biometrika, 101(1):219–228, 2014. ISSN 00063444. URL http://www.jstor.org/stable/43305605

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.
Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Raj, A., Bauer, S., Soleymani, A., Besserve, M., and Schölkopf, B. Causal feature selection via orthogonal search. *arXiv preprint arXiv:2007.02938*, 2020.

Ramsey, J. D. A scalable conditional independence test for nonlinear, non-gaussian data. *arXiv preprint arXiv:1401.5031*, 2014.

Reichenbach, H. *The direction of time*, volume 65. Univ of California Press, 1956.

Rencher, A. C. and Schaalje, G. B. *Linear models in statistics*. John Wiley & Sons, 2008.

Rosenfeld, E., Ravikumar, P. K., and Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BbNIbVPJ-42

Rouse, C. E. Democratization or diversion? the effect of community colleges on educational attainment. *Journal of Business & Economic Statistics*, 13(2): 217–224, 1995. ISSN 07350015. URL http://www.jstor.org/stable/1392376.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), jun 2020. doi: 10.1214/19-aos1857.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/shalit17a.html.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(2):2003–2030, 2006. URL http://jmlr.org/papers/v7/shimizu06a.html

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Stock, J. H., Watson, M. W., et al. *Introduction to econometrics*, volume 3. Pearson New York, 2012.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), jun 2014. doi: 10.1214/14-aos1221.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wang, Y. and Blei, D. M. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, oct 2019. doi: 10.1080/01621459.2019.1686987.

White, H. Instrumental variables regression with independent observations. *Econometrica*, 50(2): 483–499, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912639

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, may 2014. doi: 10.1016/j.neuroimage.2014.01.060.

Zhang, C.-H. and Zhang, S. S. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1): 217–242, 2014. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/24772752

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Zhang, X. Consistency of model averaging estimators. *Economics Letters*, 130:120–123, may 2015. doi: 10.1016/j.econlet.2015.03.017.
A Permutation Tests for the conditional independence test

This section briefly introduces permutation tests [Fisher (1936); Anderson & Legendre (1999)] with a focus on conditional independence. Permutation tests provide exact type I error control and are robust to misspecified assumptions such as normality and homoscedasticity.

Exchangeability is the crucial concept for permutation tests. Formally, given a finite or infinite sequence of random variables $V = (V_1, V_2, \ldots)^\top$, we say that $V$ is exchangeable if for any finite permutation $\pi$ of the set of indices $\{1, 2, \ldots\}$ we have

$$V \sim V^{\pi} = (V_{\pi(1)}, V_{\pi(2)}, \ldots).$$

For example, $N_y$ is exchangeable since the assumption of i.i.d. samples implies $N_y \sim N_y^{\pi} = (n_{y,\pi(1)}, \ldots, n_{y,\pi(\ell)})^\top$.

Before discussing the conditional independence test, we first recap how to conduct the independence test in the simplest case as follows. Recall that SEM (1) is used for the data generating process. Consider $q = 0$ and the case $H_0: Y \perp \perp X$. Given $M$ permutations $\{\pi_i\}_{i=1}^M$ of $\{1, \ldots, \ell\}$ generated uniformly, $((Y, X), (Y, X^{\pi_1}), \ldots, (Y, X^{\pi_M}))$ is exchangeable by the assumption of i.i.d. samples. Given any valid statistic $T(Y, X)$ such as the correlation between $Y$ and $X$, $p$-values are of the form

$$p_0 = \frac{\sum_{i=1}^M \mathbb{I}\{T(Y, X^{\pi_i}) \geq T(Y, X)\} + 1}{M + 1},$$

where $\mathbb{I}$ is the indicator function. Then it holds for permutation tests that the exact type I error can be bounded as follows by the exchangeability, i.e.,

$$\mathbb{P}(p_0 < \alpha | H_0, Y, X) < \alpha$$

However, if $q > 0$ and $X, W$ are correlated, $((Y, X^{\pi_1}, W), \ldots, (Y, X^{\pi_M}, W))$ is not exchangeable due to the correlation between $X$ and $W$. The confounding effect must be considered or removed to fix this issue, and several different permutation tests are proposed. The following sections discuss two types of permutation tests for high-dimensional conditional independence tests. See [Winkler et al. (2014)] for more different types of permutation tests.

A.1 Freedman-Lane permutation method

Even if we cannot simply permute indices of $X$, we note that $N_y$ is exchangeable. As a result, if the null hypothesis $H_0: Y \perp \perp X \mid W$ holds, $((Y, X, W), (\hat{Y}(\pi_1), X, W), \ldots, (\hat{Y}(\pi_M), X, W))$ is exchangeable where $\hat{Y}^{\pi_i} = W\gamma + N_y^{\pi_i}$ and $\pi_i$ is drawn uniformly. This is called Freedman-Lane permutation method [Freedman & Lane (1983)] and requires the estimation of the nuisance parameters $\gamma$. Every high-dimensional regression method can be combined with this permutation method. For instance, [Hemerik et al. (2020)] proposed the ridge regression to estimate $\gamma$:

$$\hat{\gamma}_{\text{ridge}} = \arg\min_{\gamma} \|Y - W\gamma\|^2 + \lambda\|\gamma\|^2.$$  \hspace{1cm} (8)

Then $\hat{N}_y = Y - W\hat{\gamma}_{\text{ridge}}$ is approximated exchangeable, so we can generate copies of $Y$ by $\hat{Y}(\pi_i) = W\hat{\gamma}_{\text{ridge}} + \hat{N}_y^{\pi_i}$ with $M$ permutations $\{\pi_i\}_{i=1}^M$. Given any valid statistic $T(Y, X, W)$ such as the partial
correlation between $Y$ and $X$ given $W$, $p$-values are of the form

$$p = \frac{\sum_{i=1}^{M} \mathbb{I}(T(\hat{Y}(\pi_i)), X, W) \geq T(Y, X, W))}{M + 1}.$$ 

Then we have the approximated type I error control:

$$P(p < \alpha | H_0, Y, X, W) < \alpha + \text{an approximation error},$$

where the approximation error is due to the fact that we used $\hat{\gamma}_{\text{ridge}}$ instead of $\gamma$. See Proposition 5.1 for an example of the approximation error.

### A.2 The conditional permutation tests

To correct the confounding effect, Berrett et al. (2019) proposed the conditional permutation test (CPT), which utilizes the conditional distribution of $X$ given $W$ that is assumed to be known. They argued that in semi-supervised learning, unlabelled data $(X, W)$ are plentiful, and labeled data $(Y, X, W)$ are rare. As a result, to accurately estimate the conditional distribution of $X$ given $W$ is more likely but to test for independence with $Y$ remains challenging due to the limited sample size of the labeled data. Instead of using independent noises as the Freedman-Lane permutation method, CPT generates permutations $\{\pi_i\}_{i=1}^{M}$ non-uniformly with the probability

$$P(\pi_j = \pi | Y, X, W) = \frac{q(X^\pi | W)}{\sum_{\pi' \in S_n} q(X^{\pi'} | W)} \quad (9)$$

where $q(X|W)$ is the conditional density of $X$ given $W$ and $S_n$ is the permutation group over $\{1, \ldots, \ell\}$. Then Theorem 1 in Berrett et al. (2019) showed that $(Y, X, W), (Y, X^{\pi_1}, W), \ldots, (Y, X^{\pi_M}, W))$ is exchangeable. For sampling permutations from the distribution $(9)$, they implemented Markov chain Monte Carlo sampler.

### B Model Averaging

This section introduces model averaging within the frequentist paradigm. Model averaging is an approach of averaging over the estimators for a set of candidate models using predefined or data-adaptive weights. Formally, under the null hypothesis $(6)$, we are concerned with this model:

$$y_i \leftarrow \gamma^\top w_i + n_{y,i}, \text{ for } i = 1, \ldots, \ell. \quad (10)$$

Here $n_{y,i}$ is the independent noise such that $\mathbb{E} n_{y,i} = 0$ and $\mathbb{E} n_{y,i}^2 = \sigma_y^2$ for all $i$. Given data $(y_i, x_i, w_i)$ consist of $\ell$ i.i.d. samples generated from $(10)$. We write $Y = (y_1, \ldots, y_\ell) \in \mathbb{R}^\ell$, $X = (x_1, \ldots, x_\ell) \in \mathbb{R}^{\ell \times d}$, $W = (w_1, \ldots, w_\ell) \in \mathbb{R}^{\ell \times q}$. Given $m$ subsets $\{S_i\}_{i=1}^{m}$, the estimator $\hat{\gamma}(S_j)$ for $\gamma$ via the $j$-th submodel is

$$\hat{\gamma}(S_j) = \left( W_{S_j}^\top W_{S_j} \right)^{-1} W_{S_j}^\top Y_j, \quad (11)$$

and the averaging estimator of $\gamma$ with respect to a weight vector $w = (w_1, \ldots, w_m)^\top$ is

$$\hat{\gamma}(w) = \sum_{i=1}^{m} w_i \hat{\gamma}(S_i),$$
where the weight vector $w$ lies in the unit simplex $W$ in $\mathbb{R}^m$, i.e.

$$W = \left\{ w \in [0, 1]^M : \sum_{i=1}^m w_i = 1 \right\}.$$  

In our paper, we only consider the uniform weighting, i.e. $w_i = w_j$ for all $i, j$ and state our model averaging procedure in Algorithm 2.

Algorithm 2 Model Averaging

**input:** $m$ datasets $\{(Y, X, W_{S_j})\}_{j=1}^m$, a estimation of $\gamma$

**for** $j = 1$ **to** $M$ **do**

Compute the estimation $\hat{\gamma}(S_j)$ of coefficients $\gamma$ by regressing $Y$ on $W_{S_j}$ via (11)

**end for**

**output:** $\hat{\gamma} = \frac{1}{m} \sum_{j=1}^m \hat{\gamma}(S_j)$

However, it is possible to implement other data-adaptive weights to improve the accuracy of the estimator of $\hat{\gamma}(w)$. As we stated in Theorem 5.1, an accurate prediction can decrease the type I error. We will discuss some popular model averaging strategies. See Zhang (2015) for some consistency results for those strategies.

B.1 Smoothed AIC and smoothed BIC

In the $j$-th candidate model $W_{S_j}$ with the size $k_j = |S_j|$, the AIC score is

$$\text{AIC}_j = \ell \log \hat{\sigma}_j^2 + 2k_j$$

and the BIC score is

$$\text{BIC}_j = \ell \log \hat{\sigma}_j^2 + 2k_j \log(\ell),$$

where $\hat{\sigma}_j^2 = \ell^{-1}||Y - W\hat{\gamma}(S_j)||^2$. We define the weights from AIC and BIC Hjort & Claeskens (2003):

$$\hat{w}_{\text{xIC},i} = \frac{\exp(-\text{xC}_{i}/2)}{\sum_{j=1}^m \exp(-\text{xC}_{j}/2)}, \quad i = 1, \ldots, m,$$

(12)

where xIC$_i$ is the AIC or BIC score under the $i$-th candidate model. The averaging estimators combined by weights in (12) are commonly called smoothed AIC (S-AIC) or smoothed BIC (S-BIC) estimators.

B.2 MMA

Let $H(w) = \sum_{j=1}^m w_j W_{S_j}(W_{S_j}^T W_{S_j})^{-1} W_{S_j}^T$, and $k = (|S_1|, \ldots, |S_m|)^T$. Hansen (2007) proposed choosing weights by minimizing the Mallow criterion

$$C(w) = ||I_n - H(w)||^2 + 2\hat{\sigma} w^T k$$

where

$$\hat{\sigma} = (\ell - |S_{j^{*}}|)^{-1}||Y - W\hat{\gamma}(S_{j^{*}})||,$$

and $j^{*} = \arg \max_j |S_j|$. 17
Define the empirical Mallow weight as 
\[ \hat{w}_{\text{MMA}} = \arg \min_w C(w) \] 
which has several good theoretical properties. For example, it can be shown that the mallow criterion is the unbiased of squared error plus a constant (Lemma 3, Hansen (2007)), i.e.
\[ \mathbb{E}C(w) = (W\hat{\gamma}(w) - W\gamma)^\top (W\hat{\gamma}(w) - W\gamma) + \ell \sigma_y^2. \] (13)
Thus, the empirical Mallow weight \( \hat{w}_{\text{MMA}} \) asymptotically minimizes the squared error (Theorem 1, Hansen (2007)). Moreover, the weight vector obtained by Mallow’s criterion has a sparsity property in the sense that a subset of models receives exactly zero weights Feng et al. (2020).

C Normally distributed Prior

This section our main results for normally distributed prior. For the sake of simplicity, we first impose two conditions, then show general case in the next section. To begin with, let \( r = 1, \) and \( A \) and \( B \) reduce to \( a := A \) and \( b := B, \) respectively. Then, assume \( \gamma \) is normally distributed, i.e.
\[ \gamma \sim \mathcal{N}(0, \rho^2_\beta / q I_q). \] (14)

Now we are ready to present our first main result addressing the equivalence between causation and stable regression coefficients. In particular, we prove that the cause \( X \) can be identified as long as \( \|\beta\| \) or \( q \) is large enough. Note that all results are established in the infinite sample limit. That is, the randomness here only comes from the prior (14).

**Theorem C.1.** Let \( S_1, \ldots, S_m \subseteq \{1, \ldots, q\}. \) Define \( \Sigma = [\Sigma_{ij}]_{i,j=1}^{m} \in \mathbb{R}^{m \times m} \) by
\[ \Sigma_{ij} = \frac{|a_{S_i} \cap a_{S_j}|^2}{(1 + ||b||_2^2 + ||a_{S_i}||_2^2)(1 + ||b||_2^2 + ||a_{S_i}||_2^2)}. \] (15)

where \( ||b||_2^2 = b^\top D b, ||a_{S_i}||_2^2 = a_{S_i}^\top D a_{S_i}. \) Assume that [14] holds, and \( \Sigma \) has full rank with eigendecomposition \( \Sigma = \text{diag}(\lambda_1, \ldots, \lambda_m)P \) \( \top \). Let \( e_m = (1/\sqrt{m}, \ldots, 1/\sqrt{m})^\top. \)

(a) We have two cases: if \( \rho_\beta = 0, \) it holds that
\[ \mathbb{V}(\hat{\beta}(S_1), \ldots, \hat{\beta}(S_m)) \sim 1 - \frac{\sum_{i=1}^{m} c_i \lambda_i^{1/2} g_i^2}{\sum_{i=1}^{m} \lambda_i g_i^2}; \] (16)
if \( \rho_\beta^2 / \rho_{\beta}^2 \to 0, \) we have \( \mathbb{E}(\hat{\beta}(S_1), \ldots, \hat{\beta}(S_m)) = o(1). \) \( (c_1, \ldots, c_m) = (e_m)^\top P \) and \( (g_1, \ldots, g_m)^\top \sim \mathcal{N}(0, I_m). \)

(b) Further, assume that
\[ \text{tr}(\Sigma)/m = o(q), \quad \text{as} \ q \to \infty, \] (17)
then,
\[ \mathbb{V}(\hat{\beta}(S_1), \ldots, \hat{\beta}(S_m)) \overset{p}{\to} \begin{cases} 0 & \text{if} \ \rho_\beta > 0 \\ 1 - (e_m)^\top \Sigma_e e_m / \text{tr}(\Sigma) & \text{if} \ \rho_\beta = 0 \end{cases}, \]
and \( \frac{1}{m} \sum_{j=1}^{m} \hat{\beta}(S_j) \overset{p}{\to} \beta. \)
The first result states that the upper bound of the mean of $V$ is decreasing as $\rho_\beta/\rho_\alpha$ increases, while if $\|\beta\| = 0$ we get some positive number drawn from some known distributions. The second result says as $q$ increasing, $V$ converges to zero if and only if $\rho_\beta > 0$. As a result, since the distribution of $V$ under the null alternative hypothesis are well separated from each other given large enough $\rho_\beta/\rho_\alpha$ or $q$. Theorem 4 proves the idea shows that $V$ is a reasonable test statistic for testing whether $\|\beta\|$ is zero. $\Sigma$ in (15) relates to the covariance matrix of $\hat{\beta}(S_j)$, so the condition (17) implies mild correlation among $\{\hat{\beta}(S_j)\}_{j=1}^m$ caused by confounders between $X$ and $W$. We provide two examples that show the sufficient condition for (17) is $\|a\|^2 = o(q)$ provided (25) holds in Appendix C.3.

To show our main results, we first provide some useful calculations. First, we define

$$v = (v_1, \ldots, v_m) := \left( \frac{a_{S_1}^\top \gamma_{S_1}}{1 + \|b\|^2 + \|a_{S_1}\|^2_{S_1}}, \ldots, \frac{a_{S_m}^\top \gamma_{S_m}}{1 + \|b\|^2 + \|a_{S_m}\|^2_{S_m}} \right).$$

and $e^m := (1/\sqrt{m}, \ldots, 1/\sqrt{m})^\top$ is a $m$-dimensional unit vector. Then, $\gamma \sim \mathcal{N}\left(0, \frac{\rho_2^2}{q} I_q\right)$ yields $v \sim \mathcal{N}(0, \frac{\rho_2^2}{q} \Sigma)$ where

$$\Sigma_{ij} = \frac{\|a_S^j \cap S_i\|^2}{(1 + \|b\|^2 + \|a_S^j\|^2_{S_j})(1 + \|b\|^2 + \|a_S^j\|^2_{S_j})}.$$ 

Recall the regression coefficients $\hat{\beta}(S)$ of $Y$ on $X, W_S$ is

$$\hat{\beta}(S) = \beta + \frac{a_S^\top \gamma_S}{1 + \|b\|^2 + \|a_S\|^2_S} D_x^{-1} b,$$

implying

$$\left\| \frac{1}{m} \sum_{j=1}^m \hat{\beta}(S_j) \right\|^2 = \left\| \beta + \frac{1}{\sqrt{m}} (e^m)^\top v D_x^{-1} b \right\|^2 = \|\beta\|^2 + \frac{2}{\sqrt{m}} (e^m)^\top v \beta^\top D_x^{-1} b + \frac{1}{m} (e^m)^\top v \|D_x^{-1} b\|^2,$$

$$\frac{1}{m} \sum_{j=1}^m \|\hat{\beta}(S_j)\|^2 = \frac{1}{m} \sum_{j=1}^m \|\beta + v_j D_x^{-1} b\|^2 = \|\beta\|^2 + \frac{2}{\sqrt{m}} (e^m)^\top v \beta^\top D_x^{-1} b + \frac{1}{m} \|v\|^2 \|D_x^{-1} b\|^2.$$

Using a well-known result of normal distributions Rencher & Schaalje (2008) that

$$\mathbb{E}[g^\top H g] = \text{tr}[HG], \quad \text{Var}[g^\top H g] = 2 \text{tr}[HG H]$$

where $H$ is a symmetric matrix and $g \sim \mathcal{N}(0, \Sigma)$, we know

$$\mathbb{E} \left[ \frac{1}{m} \|e^m\|^2 \right] = \rho_2^2 \frac{1}{q/m} \mathbb{E} \left[ \frac{1}{m} \|e^m\|^2 \right] = \rho_2^2 \frac{1}{q/m} \text{tr}(\Sigma),$$

and

$$\text{Var} \left[ \frac{1}{m} \|e^m\|^2 \right] = \rho_2^2 \frac{2}{q^2/m^2} \mathbb{E} \left[ \frac{1}{m} \|e^m\|^2 \right] = \rho_2^2 \frac{2}{q^2/m^2} \text{tr}(\Sigma^2).$$

In what follows, we will prove Theorem C.1.
C.1 Proof of (a) in Theorem C.1

If $\rho_\beta = 0$, we have $\beta = 0$ almost surely. Then, from (19), we get

$$
\left\| \frac{1}{m} \sum_{j=1}^{m} \hat{\beta}(S_j) \right\|^2 = \frac{1}{m} (e^m)^T v \| D_x^{-1} b \|^2, \quad \frac{1}{m} \sum_{j=1}^{m} \| \hat{\beta}(S_j) \|^2 = \frac{1}{m} \| v \|^2 \| D_x^{-1} b \|^2.
$$

Since $\Sigma$ has eigenvalue decomposition $\mathbf{P} \Lambda \mathbf{P}^T$ such that $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$, the above equations yield that

$$
1 - \nu = \frac{|(e^m)^T v|^2}{\|v\|^2} = \frac{\nu \| \Sigma^{1/2} (e^m)^T \Sigma^{1/2} \Sigma^{-1/2} v \|^2}{\nu \| \Sigma^{-1/2} \Sigma^{-1/2} v \|^2} \sim \frac{\mathbf{g}^T \mathbf{P} \Lambda^{1/2} \mathbf{P}^T \mathbf{e}^m (e^m)^T \mathbf{P} \Lambda^{1/2} \mathbf{P}^T \mathbf{g}}{\mathbf{g}^T \Lambda \mathbf{g}} \sim \frac{\sum_{i=1}^{m} c_i \lambda_i^{1/2} g_i^2}{\sum_{i=1}^{m} \lambda_i g_i^2},
$$

where $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I})$ and we have used rotation invariance of $\mathbf{g}$, and $c = (c_1, \ldots, c_m) = (e^m)^T \mathbf{P}$.

For $\rho_\beta > 0$, using (19), first we can write

$$
1 - \nu = \frac{\left\| \frac{1}{m} \sum_{j=1}^{m} \hat{\beta}(S_j) \right\|^2}{\frac{1}{m} \sum_{j=1}^{m} \| \hat{\beta}(S_j) \|^2}
= 1 - \frac{\frac{1}{m} \| v \|^2 \| D_x^{-1} b \|^2 - \frac{1}{m} \| (e^m)^T v \|^2 \| D_x^{-1} b \|^2}{\| \hat{\beta} \|^2 + \frac{2}{\sqrt{m}} (e^m)^T \nu \beta^T D_x^{-1} b + \frac{1}{m} \| v \|^2 \| D_x^{-1} b \|^2}
= 1 - \frac{1}{\rho_\beta^2} \frac{\| v \|^2 \| D_x^{-1} b \|^2 - \frac{1}{m \rho_\beta^2} [(e^m)^T v]^2 \| D_x^{-1} b \|^2}{\| \hat{\beta} \|^2 + \frac{2}{\sqrt{m \rho_\beta^2}} (e^m)^T \nu \beta^T D_x^{-1} b + \frac{1}{m \rho_\beta^2} \| v \|^2 \| D_x^{-1} b \|^2}.
$$

Combining (20) and (21) implies

$$
\mathbb{E} \left[ \frac{1}{m \rho_\beta^2} \| v \|^2 \| D_x^{-1} b \|^2 \right] \to 0, \quad \text{Var} \left[ \frac{1}{m \rho_\beta^2} \| v \|^2 \| D_x^{-1} b \|^2 \right] \to 0,
$$

and

$$
\mathbb{E} \left[ \frac{1}{m \rho_\beta^2} [(e^m)^T v]^2 \| D_x^{-1} b \|^2 \right] \to 0, \quad \text{Var} \left[ \frac{1}{m \rho_\beta^2} [(e^m)^T v]^2 \| D_x^{-1} b \|^2 \right] \to 0,
$$

20
as $\rho_\gamma/\rho_\beta^2 \to 0$, which also yields $\frac{2}{\sqrt{m\rho_\beta^2}}(e^m)^\top v\beta\top D_x^{-1}b \overset{p}{\to} 0$ as $\rho_\gamma^2/\rho_\beta^2 \to 0$.

Then, continuous mapping theorem [Durrett (2019)] implies

$$1 - \frac{1}{\sqrt{m\rho_\beta^2}}\|v\|^2\|D_x^{-1}b\|^2 - \frac{2}{\sqrt{m\rho_\beta^2}}(e^m)^\top v\beta\top D_x^{-1}b + \frac{1}{m\rho_\beta^2}\|v\|^2\|D_x^{-1}b\|^2 \overset{p}{\to} 1.$$  

Moreover, we know

$$0 \leq \left\|\frac{1}{m} \sum_{j=1}^m \hat{\beta}(S_j)\right\|^2 \leq 1,$$

so \{1 - \mathcal{V}(\hat{\beta}(S_1), \ldots, \hat{\beta}(S_m)) : \rho_\alpha/\rho_\beta \geq 0\} is uniformly integrable and $\mathbb{E}[1 - \mathcal{V}] \to 1$ as $\rho_\alpha/\rho_\beta \to \infty$. As a result, the first part of Theorem C.1 follows.

### C.2 Proof (b) in Theorem C.1

If $\text{tr}(\Sigma)/m = o(q)$, we have

$$0 < (e^m)^\top \Sigma e < \frac{1}{m} \max\left\{\text{tr}(\Sigma), \sqrt{\text{tr}(\Sigma^2)}\right\} = o(q),$$

which implies

$$\frac{1}{m}|(e^m)^\top v|^2 \to 0, \quad \frac{1}{m}\|v\|^2 \to 0,$$

in probability. Moreover, it holds that

$$\text{Var}\left[\frac{2}{\sqrt{m}}(e^m)^\top v\beta\top D_x^{-1}b\right] = \frac{4\rho_\beta^2}{mdq}(e^m)^\top \Sigma e\|D_x^{-1}b\|^2 \to 0,$$

yielding

$$\frac{2}{\sqrt{m}}(e^m)^\top v\beta\top D_x^{-1}b \to 0,$$

in probability. Consequently, we get

$$\frac{1}{m} \sum_{j=1}^m \hat{\beta}(S_j) = \beta + \frac{1}{\sqrt{m}}(e^m)^\top vD_x^{-1}b \to \beta,$$

in probability. Furthermore, if $\rho_\beta > 0$ and $q \to \infty$, we have

$$1 - \mathcal{V} \to \frac{\|\beta\|^2}{\|\hat{\beta}\|^2} = 1,$$

in probability by Theorem 2.7 in [Van der Vaart (2000)]. On the other hand, if $\beta = 0$, we have

$$1 - \mathcal{V} = \left\|\frac{1}{m} \sum_{j=1}^m \hat{\beta}(S_j)\right\|^2 \overset{\text{tr}(\Sigma)}{\to} \frac{(e^m)^\top \Sigma e}{\text{tr}(\Sigma)},$$

in probability as $q \to \infty$. Then this theorem follows.
C.3 More details of Theorem C.1

Two propositions are given to illustrate Theorem C.1. The latter provides an example where our method can identify the cause when there is a hidden confounder, while the partial correlation has no information for finding causal drivers. However, note that the partial correlation is not designed for detecting causes and can be used without any strong assumption on confounding.

Proposition C.2. Suppose \( \text{Cov}(N_w) = I \) and we have \( m \) subsets \( S_1, \ldots, S_m \subset \{1, \ldots, q\} \) such that
\[
\bigcup_{j=1}^{m} S_j = \{1, \ldots, q\}, \quad S_i \cap S_j = \emptyset \quad \text{for} \quad i \neq j.
\]
If \( \|a\|^2 = o(q) \), the condition in Theorem C.1 holds, i.e.
\[
\frac{\text{tr}(\Sigma)}{m} = o(q),
\]
as \( q \to \infty \).

Proof of Proposition C.2. we know
\[
\frac{1}{m} \text{tr}(\Sigma) = \frac{1}{m} \sum_{j=1}^{m} \|a_{S_j}\|^2 \leq \frac{(m-1)\|a\|^2}{m(1 + \|b\|_2^2)^2}
\]
where \( \Sigma \) is defined in (15). Thus, \( \|a\|^2 = o(q) \) implies \( \text{tr}(\Sigma)/m = o(q) \), which completes the proof.

Proposition C.3. Assume \( d_1 = b, \rho_3 = 0, \text{Cov}(N_w) = I, \) and \( W_1 \) is the unobserved confounder with the non-trivial coefficient \( a_1 \neq 0 \). Then the partial correlation does not converge to a constant even when \( q \to \infty \). On the other hand, suppose that we have subsets \( S_1, \ldots, S_m \subset \{2, \ldots, q\} \) satisfying (25). If \( \|a\|^2 = o(q) \), the condition (26) in Theorem C.1 holds.

Proof of Proposition C.3. Define \( W_{-1} = (W_2, \ldots, W_q)^\top, a_{-1} = (a_2, \ldots, a_q)^\top, \) and \( \gamma_{-1} = (\gamma_2, \ldots, \gamma_q)^\top \). Then the residuals are
\[
R_Y := Y - E[Y|W_{-1}, \gamma] = \frac{\gamma_1 a_1 Z + N_y - \gamma_1 a_1 a_{-1}^\top N_{w_{-1}}}{1 + \|a_{-1}\|^2} + N_y - \frac{\gamma_1 a_1 a_{-1}^\top N_{w_{-1}}}{1 + \|a_{-1}\|^2}
\]
\[
R_X := X - E[X|W_{-1}, \gamma] = \frac{Z}{1 + \|a_{-1}\|^2} + N_x - \frac{a_{-1}^\top N_{w_{-1}}}{1 + \|a_{-1}\|^2}
\]
Therefore, we have
\[
E[R_Y R_X | \gamma] = a_1 \gamma_1
\]
which implies
\[
\frac{E[R_Y R_X | \gamma]}{\text{Var}[R_Y | \gamma] \text{Var}[R_X | \gamma]} \neq 0
\]
On the other hand, we know \( m = q/k \) and
\[
\frac{1}{m} \text{tr}(\Sigma) \leq \frac{(m-1)\|a\|^2 + a_1^2}{m(1 + \|b\|_2^2)^2}.
\]
Thus, \( \|a\|^2 = o(q) \) implies \( \text{tr}(\Sigma)/m = o(q) \), which completes the proof.
D Rotation invariant distributions

In this section, we generalize our result in Theorem C.1 to $r > 1$ and rotation invariant prior. Recall the definition of rotation invariant distributions:

**Definition D.1.** A random vector $\nu = (\nu_1, \ldots, \nu_q)^T$ is spherically symmetric if the distribution of every linear form is the same as the distribution of $\nu_1$, i.e., $e^T \nu \sim \nu_1$ for all $e$ provided that $\|e\| = 1$.

From the definition, we get two important properties.

**Lemma D.2.** Suppose that $\nu$ is spherically symmetric. Then, it holds that

1. $O\nu \sim \nu$, where $O$ is an orthogonal matrix,
2. $\nu = \tau U$, where $U$ is uniformly distributed on the unit sphere in $\mathbb{R}^d$, $\tau \geq 0$ is a real value with distribution $\tau \sim \|\nu\|$, and $\tau, U$ are independent.

**Proof.** See Theorem 4.1.2 in Bryc (2012).

From the first property, we can see that the spherically symmetric random variable is rotation invariant. Moreover, the second property states rotation invariant distributions can be decomposed by a uniform distribution on the unit sphere and a positive random variable. Lemma D.2 is so powerful since we know the exact marginal distribution of uniformly distributed on the unit sphere, which facilitates many calculations and is presented as follows:

**Proposition D.3.** Suppose $U$ is uniformly distributed on the unit sphere in $\mathbb{R}^d$. Then, it holds that $e^T U \sim \|e\|(2 - \text{Beta}(\frac{d-1}{2}, \frac{d-1}{2}))$, where Beta is beta distribution.

**Proof.** Since $U$ is spherically symmetric, we get $e^T U \sim \|e\|U_1$. From example 4.1.2 in Bryc (2012), we know the density of the marginal distribution of $U_1$ is

$$f_{U_1}(u) \propto (1 - u^2)^{\frac{d-1}{2}-1} = (1 - u)^{\frac{d-1}{2}-1}(1 + u)^{\frac{d-1}{2}-1} \text{ for } -1 \leq u \leq 1. \tag{28}$$

On the other hand, the density of $2 \text{Beta}(\frac{d-1}{2}, \frac{d-1}{2}) - 1$ is

$$f_{2 \text{Beta}(\frac{d-1}{2}, \frac{d-1}{2})}(u) \propto f_{\text{Beta}(\frac{d-1}{2}, \frac{d-1}{2})}\left(\frac{u + 1}{2}\right)$$

$$\propto \left(\frac{u + 1}{2}\right)^{\frac{d-1}{2}-1}\left(1 - \frac{u + 1}{2}\right)^{\frac{d-1}{2}-1} \tag{29}$$

$$\propto (1 - u)^{\frac{d-1}{2}-1}(1 + u)^{\frac{d-1}{2}-1} \text{ for } -1 \leq u \leq 1.$$

The proof is completed by comparing two densities.

The next key result shows upper bounds of the second moment of quadratic forms for rotation invariant random variables.

**Proposition D.4.** Suppose that $\nu$ is spherically symmetric. Then, it holds that $E(\nu^T H \nu)^2 \leq (\text{tr } H)^2 E\|\nu\|^4$. 

23
Proof. Let the eigendecomposition of \( H \) is \( Q \text{diag}(\lambda_H)Q^\top \). Then we get
\[
\mathbb{E}(\nu^\top H \nu)^2 = \mathbb{E}(\nu^\top Q \text{diag}(\lambda_H)Q^\top \nu)^2
\]
1 in Lemma \[D.2\]
\[
= \mathbb{E}(\|\text{diag}(\lambda_H)^{1/2} \nu\|^2)^2
\]
2 in Lemma \[D.2\]
\[
= \mathbb{E}(\|\text{diag}(\lambda_H)^{1/2} U\|^2)^2 \mathbb{E} \tau^4
\]
2 in Lemma \[D.2\]
\[
= \mathbb{E} \left[ \sum_{i=1}^d \langle \lambda_H \rangle U_i^2 \right]^2 \mathbb{E} \|\nu\|^4
\]
|\(U\)|\(\leq 4\) (\(\text{tr} H\))^2 \(\mathbb{E} \|\nu\|^4\).

Therefore, this proposition follows. \(\square\)

Now we are ready to state our generalized results which includes Theorem \[4.3\]

**Theorem D.5.** Let \( S_1, \ldots, S_m \subset \{1, \ldots, q\} \). Assume that \( \gamma \) is a spherically symmetric random vector such that \( \mathbb{E} \|\gamma\|^4 = O(\rho_1^4/q^2) \). Define \( C_m = \frac{1}{m} \sum_{j=1}^m C(S_j) \) and \( C_m = \frac{1}{m} \sum_{j=1}^m C(S_j)^\top C(S_j) \).

(a) We have two cases: if \( \rho_\beta = 0 \), it holds that
\[
\mathcal{V} \left( \hat{\beta}(S_1), \ldots, \hat{\beta}(S_m) \right) \sim 1 - \frac{\gamma^\top C_m^\top C_m \gamma}{\gamma^\top \gamma}
\]

If \( \rho_\beta^2/\rho_\beta^2 \to 0 \), we have \( \mathbb{E} \mathcal{V}(\hat{\beta}(S_1), \ldots, \hat{\beta}(S_m)) = o(1) \).

(b) Further, assume that
\[
\text{tr}(C_m) = o(q), \quad \text{as} \ q \to \infty,
\]
then,
\[
\mathcal{V}(\hat{\beta}(S_1), \ldots, \hat{\beta}(S_m)) \overset{p}{\to} \begin{cases} 0 & \text{if } \rho_\beta > 0 \\ 1 - \frac{\text{tr}(C_m^\top C_m)}{\text{tr}(C_m)} & \text{if } \rho_\beta = 0 \end{cases},
\]
and \( \frac{1}{m} \sum_{j=1}^m \hat{\beta}(S_j) \overset{p}{\to} \beta \), where \( \rho_\beta = \|\beta\| \).

We can see that Theorem \[C.1\] is actually a special case of Theorem \[D.5\]. Again, the condition that \( \beta \) is drawn from normal distributions is not necessary to relate stable regression coefficients to causal relations and id imposed due to technical consideration. The assumption (31) is similar to (17) and means mild correlation among \( \{\hat{\beta}(S_j)\}_{j=1}^m \) caused by confounders between \( X \) and \( W \).

To prove Theorem \[D.5\], we need some calculations. Recall the regression coefficients \( \hat{\beta}(S) \) on \( X, W_S \) is
\[
\hat{\beta}(S) = (I_d, 0) \text{Var} (X, W_S)^{-1} \text{Cov} ((X, W_S), Y)
\]
\[
= \beta + C(S)\gamma
\]
(33)
where \( D_S = \text{diag}(\sigma_{w_S}) \) and
\[
(C(S)^\top)_S = (D_x^{-1} B (I + B^\top D_x^{-1} B + A_S D_S A_S)^{-1} A_S^\top )^\top, \quad \text{and} \quad (C(S)^\top)_S = 0.
\]
(34)

\[24\]
Then we have

$$\left\| \frac{1}{m} \sum_{j=1}^{m} \hat{\beta}(S_j) \right\|^2 = \left\| \beta + \frac{1}{m} \sum_{j=1}^{m} C(S_j) \gamma \right\|^2 = \|\beta\|^2 + 2\beta^T C_m \gamma + \gamma^T C_m^T C_m \gamma,$$

(35)

$$\frac{1}{m} \sum_{j=1}^{m} \|\hat{\beta}(S_j)\|^2 = \frac{1}{m} \sum_{j=1}^{m} \|\beta + C(S_j) \gamma\|^2 = \|\beta\|^2 + 2\beta^T C_m \gamma + \gamma^T \tilde{C}_m \gamma$$

where $C_m = \frac{1}{m} \sum_{j=1}^{m} C(S_j)$ and $\tilde{C}_m = \frac{1}{m} \sum_{j=1}^{m} C(S_j)^T C(S_j)$.

**D.1 Proof of (a) in Theorem D.5**

If $\rho_\beta = 0$, we have $\beta = 0$ almost surely. Then, from (35), we get

$$1 - V = \frac{\gamma^T C_m^T C_m \gamma}{\gamma^T \tilde{C}_m \gamma}.$$

(36)

For $\rho_\beta > 0$ and $\rho_\gamma^2/\rho_\beta^2 \to 0$, following the same argument in Appendix C.1 it suffices to show

$$E\left( \frac{1}{\rho_\beta^2} \gamma^T C_m^T C_m \gamma \right) \to 0, \quad E\left( \frac{1}{\rho_\beta^2} \gamma^T \tilde{C}_m \gamma \right) \to 0,$$

(37)

since $E(\gamma^T C_m^T C_m \gamma)^2 \to 0$ implies $E(2\beta^T C_m \gamma)^2 \to 0$.

By Proposition D.4 and assumptions, we have

$$E\left( \frac{1}{\rho_\beta^2} \gamma^T C_m^T C_m \gamma \right)^2 \leq \frac{\|C_m\|_1^4 E\|\nu\|^4}{\rho_\beta^4} = O\left( \frac{\rho_\gamma^2 [tr(C_m^T C_m)]^2}{\rho_\beta^4 q^2} \right),$$

(38a)

$$E\left( \frac{1}{\rho_\beta^2} \gamma^T C_m^T C_m \gamma \right)^2 \leq \frac{\|\tilde{C}_m\|_1^4 E\|\nu\|^4}{\rho_\beta^4} = O\left( \frac{\rho_\gamma^2 [tr(\tilde{C}_m)]^2}{\rho_\beta^4 q^2} \right).$$

(38b)

Both of them converges to zero as $\rho_\gamma/\rho_\beta \to 0$. As a result, the first part of Theorem D.5 follows.

**D.2 Proof (b) in Theorem D.5**

First, we know two facts:

1. $E\gamma^T H \gamma = tr(H E(\gamma \gamma^T))$.

2. If $U$ is uniformly distributed on the unit sphere in $\mathbb{R}^q$, then $EUU^T = \frac{1}{q} I$.

Then by 2. in Lemma D.2 we get

$$E\left( \gamma^T C_m^T C_m \gamma \right) = E\|\gamma\|^2 tr(C_m^T C_m)/q,$$

$$E\left( \gamma^T \tilde{C}_m \gamma \right) = E\|\gamma\|^2 tr(\tilde{C}_m)/q$$
Moreover, the assumption (31) implies (38a) and (38b) converge to zero. Consequently, following the same argument in Appendix C.2, we get the following results: If $\rho_\beta > 0$ and $q \to \infty$, it holds that
\[
1 - \mathcal{V} \mathbb{P} \frac{||\beta||^2}{||\beta||^2} = 1, \tag{39}
\]
while if $\beta = 0$, we have
\[
1 - \mathcal{V} = \mathbb{P} \frac{1}{m} \sum_{j=1}^{m} ||\hat{\beta}(S_j)||^2 \to \frac{\operatorname{tr}(C_m^T C_m)}{\operatorname{tr}(C_m)} \tag{40}
\]
as $q \to \infty$. Then this theorem follows.

## E Proof of Theorem 5.1

Drawing $M$ permutations $\{\pi_i\}_{i=1}^{M}$ uniformly, we know that
\[
((\hat{Y}(\pi_1), X, W), \ldots, (\hat{Y}(\pi_M), X, W))
\]
is exchangeable under the null hypothesis $H_0$ in (6) where $\hat{Y}^{\pi_i} = W\gamma + N_i^\pi$ since $N_i$ is an independent noise. Given any valid statistic $T(Y, X, W)$ such as $T = \mathcal{V}$ in our case, $p$-values are of the form
\[
\frac{\sum_{i=1}^{M} \mathbb{I}(T(\hat{Y}^{\pi_i}, X, W) \geq T(Y, X, W))}{M + 1}
\]
By exchangeability, we get
\[
\mathbb{P}(p((Y, X, W), (\hat{Y}(\pi_1), X, W), \ldots, (\hat{Y}(\pi_M), X, W)) \leq \alpha \mid (X, W), \gamma) \leq \alpha \tag{41}
\]
For the permutation $\pi_i$ and the corresponding permutation matrix $P_i$ and an estimation $\hat{\gamma}$, we get
\[
\hat{Y}(\pi_i) \mid H_0, X, W, \gamma = W\hat{\gamma} + P_i(Y - W\hat{\gamma}) \mid H_0, X, W, \gamma
\]
\[
=W\hat{\gamma} + P_iW(\gamma - \hat{\gamma}) + P_iN_i \mid H_0, X, W, \gamma
\]
\[
\sim N(W\hat{\gamma} + P_iW(\gamma - \hat{\gamma}), \sigma^2_i I) \tag{42}
\]
Recall that $e^M = (1/\sqrt{M}, \ldots, 1/\sqrt{M})^T$ is a $M$-dimensional unit vector. Since $P_i$ and $P_j$ are drawn uniformly and independently, it holds that
\[
\text{Cov}(\hat{Y}^{\pi_i}, \hat{Y}^{\pi_j} \mid H_0, X, W, \gamma) = \mathbb{E}[P_i N_i N_i^T P_j] = \sigma^2_i e_{M} e_{M}^T,
\]
and
\[
\text{Cov}(\hat{Y}^{\pi_i}, \hat{Y}^{\pi_j} \mid H_0, X, W, \gamma) = \mathbb{E}[P_i N_i N_i^T P_j] = \sigma^2_i e_{M} e_{M}^T.
\]
Therefore, we know eigenvalues of the covariance matrices $(\hat{Y}(\pi_1), \ldots, \hat{Y}(\pi_M))$ and $(\hat{Y}(\pi_1), \ldots, \hat{Y}(\pi_M))$ conditional on $H_0, X, W, \gamma$ are $(M\sigma_y^2, \sigma_y^2, \ldots, \sigma_y^2)$. It is well-known that if $g_1 \sim N(\mu_1, G)$ and $g_2 \sim N(\mu_2, G)$, then
\[
d_{KL}(g_1, g_2) = \frac{1}{2}(\mu_1 - \mu_2)^T G^{-1}(\mu_1 - \mu_2).
\]
Then we get

\[
d_{KL}((\bar{Y}(\pi_1), \ldots, \bar{Y}(\pi_M)), (\hat{Y}(\pi_1), \ldots, \hat{Y}(\pi_M)))
\]

\[
\leq \frac{1}{2\sigma^2_y} \sum_{i=1}^{M} \|W(\gamma - \hat{\gamma}) + P_i(W(\gamma - \hat{\gamma}))\|^2
\]

\[
\leq \frac{M}{2\sigma^2_y} \|W(\gamma - \hat{\gamma})\|^2,
\]

where \(d_{KL}\) is the Kullback–Leibler divergence.

Note that the total variation distance \(d_{TV}\) is defined as \(d_{TV}(Q_1, Q_2) = \sup_E |Q_1(E) - Q_2(E)|\) where the supremum is taken over all measurable sets. Then by the definition of total variation, it follows that

\[
P(\{(Y, X, W), (\hat{Y}(\pi_1), X, W), \ldots, (\hat{Y}(\pi_M), X, W)\} | \alpha, X, W, \gamma)
\]

\[
\leq P(\{(Y, X, W), (\hat{Y}(\pi_1), X, W), \ldots, (\hat{Y}(\pi_M), X, W)\} | \alpha, X, W, \gamma)
\]

\[
+ d_{TV}\{[(\hat{Y}(\pi_1), \ldots, \hat{Y}(\pi_M)) | X, W, \gamma], [(\bar{Y}(\pi_1), \ldots, \bar{Y}(\pi_M)) | X, W, \gamma]\}.
\]

By Pinsker’s inequality relating total variation distance to the Kullback Leibler divergence \(d_{KL}\), we get

\[
d_{TV}^{\alpha} \{[(\hat{Y}(\pi_1), \ldots, \hat{Y}(\pi_M)) | X, W, \gamma], [(\bar{Y}(\pi_1), \ldots, \bar{Y}(\pi_M)) | X, W, \gamma]\}
\]

\[
\leq \frac{1}{2} d_{KL} \{[(\hat{Y}(\pi_1), \ldots, \hat{Y}(\pi_M)) | X, W, \gamma], [(\bar{Y}(\pi_1), \ldots, \bar{Y}(\pi_M)) | X, W, \gamma]\}
\]

\[
\leq \frac{M}{2\sigma^2_y} \|W(\gamma - \hat{\gamma})\|^2,
\]

which completes the proof.

\section{Additional Experiments}

This section includes additional experiments.

**Results of varying parameters:** First, we vary parameters \(\rho_\beta, d,\) and \(|S_j|\) to examine more details of our approach. Note that the basic setting is that \(d = 3, r = 5, \ell = 300, \rho_\gamma = 2.5\) for data generating procedure and \(m = 40, M = 400, k = 3\) for Algorithm 2+RS. We repeated the experiment 500 times for each point in Figure 3.

We can see the power increasing exponentially with respect to \(\rho_\beta\) and increasing the dimension of \(X\) decreases the power. This is only slightly changed when varying \(k\). Type I error is small in all settings.

**Results of varying numbers of hidden components:** Next, we conduct experiments with varying numbers of hidden components for the various competing approaches as well as our random selection method. We used \(d = 1, q = 500, r = 5, \ell = 100, \rho_\gamma = 1, m = 100, M = 1000, |S_j| = 10,\) and generated data for \(\rho_\beta = 0, 1\). We consider two ways to generate \(\gamma:\)

1. **Generating model 1:** Draw \(\gamma'\) from \(\mathcal{N}(0, I)\) and \(\gamma = \rho_\alpha \gamma'/||\gamma'||\),
2. **Generating model 2:** Draw \(\gamma'\) from Student’s t-distribution with degree of freedom 2.2 and \(\gamma = \rho_\alpha \gamma'/||\gamma'||\),

and three settings:

- **Setting 1:** \((W_1, \ldots, W_{500})\) is accessible,
- **Setting 2:** We set \((W_1, \ldots, W_{250})\) as latent variables. That is, only \((W_{251}, \ldots, W_{500})\) is accessible,
Setting 3: We set \((W_1, \ldots, W_{450})\) as latent variables. That is, only \((W_{451}, \ldots, W_{500})\) is accessible. 

Note that in generating procedure 1, \(\gamma'\) is drawn from a rotation invariant and light-tailed distribution, i.e., a normal distribution. Thus, with high probability, the magnitude of all elements of \(\gamma\) are similar, and there is no outlier. On the other hand, in generating procedure 2, \(\gamma'\) is drawn from a rotation invariant and heavy-tailed distribution, i.e., a student’s t distribution with a low degree of freedom. Hence, with high probability, the magnitude of elements of \(\gamma\) are dominated by some outliers. In other words, most coefficients of \(\gamma\) are very close to zero, while only a few elements’ magnitudes are large.

The results are summarized in Table 3 and Figure 4. We first discuss the result for the alternative hypothesis.

Table 3: This table shows the proportion p-values that are less than the significant level \(\alpha\) under the alternative hypothesis (6) for \(\rho^\beta = 1\).

| \(H_A\) \(\rho_\beta = 1\) \(\alpha\) Generating model 1 | Generating model 2 |
|--------------------------------|-------------------|
| \(\alpha\)                  | RS    | BM    | FL    | DR    | RS    | BM    | FL    | DR    |
| Setting 1 0.01               | 0.985 | 0.98  | 0.3425| 0.395 | 0.985 | 0.9825| 0.295 | 0.335 |
| Setting 2 0.01               | \textbf{0.9775}| 0.9625| 0.8175| 0.8775| \textbf{0.98}| 0.9725| 0.8175| 0.87   |
| Setting 3 0.01               | 0.9675| 0.945 | 0.995 | \textbf{0.9975}| 0.955 | 0.9625| 0.995 | \textbf{0.9975}| 

Table 3 shows the similar pattern as the comparison result in Section 6.1. We can see that RS shows the highest power in settings 1 and 2 for all generating models. As we mentioned in Section 6.1, it may be counter-intuitive that FL and DR work better as the number of hidden confounders increase, but since \(\gamma\) is drawn from the prior with zero mean, the confounding effect will be offset when \(q\) is large. There is no significant difference between generating models 1 and 2. Overall, RS is most stable and performing well in all settings.

From Figure 4, we can see distributions of p-values of all methods are not uniform under the null hypothesis except our proposed method. Interestingly, increasing the number of hidden variables eliminates this phenomenon. On the other hand, the performance of RS and BM decreases, and the performance of FL and DR increases as the number of hidden variables increases. Again, this result is because \(\gamma\) is drawn from some prior with zero mean, and the confounding effect is cancelled, which aligns with our conclusion in Section 6. The distribution of \(\gamma\) has no significant effect on all methods. DR outperforms RS in setting 3. However, we would like to point out additional experiments are conducted in the case \(\rho_\alpha/\rho_\beta = 1\), which is a simpler case compared to \(\rho_\alpha/\rho_\beta = 20/3\) in Section 6. Moreover, our proposed method is most robust in all settings.
G Additional Result for the real dataset

In this section, we include the full analysis result for OLS in Section 6.2. The summary is stated in Table 4. We can see that even though R-squared is not high, the probability of F-statistics is very small, which indicates we reject the hypothesis that all the variables have zero regression coefficients. Moreover, there is no significant evidence that the error assumption is wrong, and the sample size is large enough, which validates the use of linear regression. All covariates are statistically significant except “urban” and “cue80”. The reason may be “urban” and “cue80” are highly correlated to other features like “stwmfg80”.

Figure 4: The histogram of p-values under the null hypothesis (6).
Table 4: OLS Regression Results

| Dep. Variable: | bytest | R-squared: | 0.193 |
| Model: | OLS | Adj. R-squared: | 0.191 |
| Method: | Least Squares | F-statistic: | 102.4 |
| No. Observations: | 4739 | Prob (F-statistic): | 5.83e-210 |
| Df Residuals: | 4727 | Log-Likelihood: | -16470. |
| Df Model: | 11 | AIC: | 3.296e+04 |
| Covariance Type: | nonrobust | BIC: | 3.304e+04 |

|            | coef   | std err | t     | P>|t| | [0.025 | 0.975 |
|------------|--------|---------|-------|-----|--------|-------|
| Intercept  | 49.0890| 0.906   | 54.198| 0.000| 47.313 | 50.865|
| female     | -1.0427| 0.229   | -4.544| 0.000| -1.493 | -0.593|
| black      | -6.8793| 0.330   | -20.851| 0.000| -7.526 | -6.232|
| hispanic   | -4.0855| 0.309   | -13.243| 0.000| -4.690 | -3.481|
| dadcoll    | 2.8857 | 0.327   | 8.837 | 0.000| 2.246  | 3.526 |
| momcoll    | 2.1927 | 0.369   | 5.938 | 0.000| 1.469  | 2.917 |
| ownhome    | 1.1346 | 0.303   | 3.746 | 0.000| 0.541  | 1.728 |
| urban      | -0.3818| 0.289   | -1.321| 0.187| -0.948 | 0.185 |
| cue80      | 0.0222 | 0.045   | 0.491 | 0.624| -0.067 | 0.111 |
| stwmfg80   | 0.2795 | 0.090   | 3.094 | 0.002| 0.102  | 0.457 |
| dist       | -0.2674| 0.055   | -4.883| 0.000| -0.375 | -0.160|
| incomehi   | 0.7163 | 0.275   | 2.604 | 0.009| 0.177  | 1.256 |
| Omnibus:   | 145.874|         |       |     |        |       |
| Durbin-Watson: | 1.781 |         |       |     |        |       |
| Prob(Omnibus): | 0.000 |         |       |     |        |       |
| Skew:      | -0.047 |         |       |     |        |       |
| Kurtosis:  | 2.411  |         |       |     |        |       |
| Cond. No.  |        |         |       |     |        | 102.  |