Identifiability of Sparse Causal Effects using Instrumental Variables

Niklas Pfister\textsuperscript{1,*}
Jonas Peters\textsuperscript{1,*}

\textsuperscript{1}Department of Mathematical Sciences, University of Copenhagen, Denmark
*Authors contributed equally.

Abstract

Exogenous heterogeneity, for example, in the form of instrumental variables can help us learn a system’s underlying causal structure and predict the outcome of unseen intervention experiments. In this paper, we consider linear models in which the causal effect from covariates $X$ on a response $Y$ is sparse. We provide conditions under which the causal coefficient becomes identifiable from the observed distribution. These conditions can be satisfied even if the number of instruments is as small as the number of causal parents. We also develop graphical criteria under which identifiability holds with probability one if the edge coefficients are sampled randomly from a distribution that is absolutely continuous with respect to Lebesgue measure and $Y$ is childless. As an estimator, we propose \textit{spaceIV} and prove that it consistently estimates the causal effect if the model is identifiable and evaluate its performance on simulated data. If identifiability does not hold, we show that it may still be possible to recover a subset of the causal parents.

1 INTRODUCTION

Instrumental variables [Wright 1928; Imbens and Angrist 1994; Newey 2013] allow us to consistently estimate causal effects from covariates $X$ on a response $Y$ even if the covariates and response are connected through hidden confounding. These approaches usually rely on identifying moment equations such as $\text{Cov}[I, Y - X^\top \beta] = 0$ with $I$ being the instrumental variable (IV). Under some assumptions such as the exclusion restriction, this equation is satisfied for the true causal coefficient $\beta = \beta^*$; in a linear setting, for example, this is the case if we can write $Y = X^\top \beta^* + q(H, \varepsilon_Y)$ with $H, \varepsilon_Y$ being independent of $X$ and $I$ and $\varepsilon_Y$ independent of $X$. Identifiability of $\beta^*$, however, requires that the moment equation is not satisfied for any other $\beta \neq \beta^*$. Formally, this condition is often written as a rank condition on the covariance between $I$ and $X$, which implies that the dimension of $I$ must be at least as large as the number of components of $X$.

In this work, we consider the case where the causal coefficient $\beta^*$ is assumed to be sparse. This assumption allows us to relax existing identifiability conditions: it is, for example, possible to identify $\beta^*$ even if there are much less instruments than covariates. Our results are proved in the context of linear structural causal models (SCMs) [Pearl 2009; Bongers et al. 2021], that is, we also assume linearity among the $X$ variables. We prove sufficient conditions for identifiability of $\beta^*$ that are based on rank conditions of the matrix of causal effects from $I$ on the parents of $Y$. We then investigate for which graphical structures we can expect such conditions to hold. Consider, for example, the graph shown in Figure 1. Square nodes represent instruments, and hidden variables between variables in $X \cup \{Y\}$ can exist but are not drawn (we formally introduce such graphs in Section 2.1). Sparse identifiability in this graph is not obvious: Is the causal effect from the parents of $Y$ to $Y$ generically identifiable if the true underlying and unknown graph is the one shown (including the two dashed edges)? And what about the graph excluding the two dashed edges? We translate the rank conditions for identifiability to structural SCMs whose coefficients are drawn randomly from a distribution that is absolutely continuous with respect to Lebesgue measure. This allows us to develop graphical criteria that can answer these questions.

If identifiability holds, the causal effect can be estimated from data. We propose an estimator called \textit{spaceIV} (‘sparse causal effect IV’). It is based on the limited information maximum likelihood (LIML) estimator [Anderson and Rubin 1949; Amemiya 1985]. This estimator has similar properties as the two stage least squares estimator and has the same asymptotic normal distribution, for example [Mariano 2001]. But as it minimizes the Anderson-Rubin...
test statistic, it allows us to prove theoretical guarantees. We evaluate the performance of spaceIV on simulated data. If identifiability does not hold, we prove that it may still be possible to identify a subset of the causal parents of \( Y \).

Numerous extensions to the classical linear instrumental variable setting have been proposed. For example, nonlinear effects \cite{Imbens2022,Mckeigue2010,Bennett2019,Muandet2020,Saengkyongam2022}. Furthermore, \cite{Belloni2012,McKeigue2010} assume that the effect from the instruments on the covariates is sparse. For example, it has been shown that consistent estimators exist if at least half of all instruments are valid \cite{Kang2016}. To the best of our knowledge, while existing work considers sparsity constraints between the instruments and the covariates (“first stage”), the assumption of a sparse causal effect (“second stage”) and its benefits has not yet been analyzed.

Our paper is structured as follows. Section 2 introduces the formal setup. Section 3 presents the main identifiability result for sparse causal effect models and Section 4 develops the corresponding graphical criteria. Section 5 introduces the estimator spaceIV and Section 6 includes simulation experiments. Code is attached as supplementary material.

\section{IV MODELS WITH SHIFT INTERVENTIONS}

Consider the following structural causal model (SCM)

\begin{align}
X & := BX + AI + h(H, \varepsilon^X) \\
Y & := X^\top \beta^* + g(H, \varepsilon^Y),
\end{align}

where \( h \) and \( g \) are arbitrary measurable functions and \( \text{Id} - H^\top \) is invertible. Here, \( X \in \mathbb{R}^d \) denotes the observed variables, \( H \in \mathbb{R}^q \) the unobserved variables, \( I \in \mathbb{R}^m \) the instrumental variables (following an \( m \)-dimensional distribution, which is not modelled explicitly), \( Y \in \mathbb{R} \) the response and \( I, H, \varepsilon^X \) and \( \varepsilon^Y \) are jointly independent and assume that the covariates are non-descendants of \( Y \) (see also Remark \ref{rem:IV}). In contrast to classical IV settings, we thus explicitly model the causal effects of the instruments \( I \) on the predictor variables \( X \). Throughout the paper, we assume that \( \text{Cov}[I] \) is invertible. We assume that we have access to an i.i.d. data set \((X_1, Y_1, I_1), \ldots, (X_n, Y_n, I_n)\) sampled from the induced distribution and are interested in estimating the causal effect \( \beta^* \). We call the set of non-zero components of \( \beta^* \) the parents of \( Y \) and denote it by \( \text{PA}(Y) \).

Our model covers the case, where we observe data from \( m \) different experiments, each of which corresponds to a fixed intervention shift. More precisely, we can choose \( I \) such that for all \( k \in \{1, \ldots, m\} \), we have \( P(I = e_k) = 1/m \), with \( e_k, k \in \{1, \ldots, m\} \), being the \( k \)-th unit vector in \( \mathbb{R}^m \). Here, each column in the matrix \( A \) specifies a different experiment in which (a subset of) the \( X \) variables is shifted by the amount specified in that column.

\subsection{GRAPHICAL REPRESENTATION}

Given a data generating process of the form \cite{Imbens2022}, we represent it graphically as follows: Each of the \( d \) components of \( X \) is represented by a node, which we call a prediction node. There is a directed edge from \( X^j \) to \( X^j \) if and only if \( B_{jj} \neq 0 \). In addition, we represent the \( k \)th component of \( I \) by a square node with label ‘\( k \)’, which we call instrument node. There is a directed edge from \( k \) to \( X^j \) if and only if \( A_{jk} \neq 0 \). (There are no connections between instrument nodes, even though they may be dependent.) Finally, we represent the response \( Y \) with the same node style as is used for the predictors and include a directed edge from \( X^j \) to \( Y \) if and only if \( \beta^j \neq 0 \). In the graph, we do not represent hidden variables (even though they are allowed to exist). Consequently, such graphs do not satisfy the Markov condition \cite{Lauritzen1996}.

\footnote{Here \( \text{Id} \) denotes the identity matrix.}

\footnote{In a slight abuse of notation, we sometimes identify each component with its index.}

\vspace{1cm}

\textit{Accepted (UPDATED VERSION) for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).}
we define the coefficients. The matrix \( C \) will play a central role when analyzing identifiability. For example, using the matrix \( C \), Proposition 2 characterizes settings under which individual components of the causal coefficient \( \beta^* \) are identifiable. This result does not require any additional assumptions on the underlying model. In Section 3.1 we then show that if the causal coefficient \( \beta^* \) is sparse (i.e., it contains many zeros) it can still be identifiable even if \( B \) is non-degenerate.

**Proposition 2** (Partial identifiability of causal coefficient). Consider a data generating process of the form (1). Then, for all \( j \in \{1, \ldots, d\} \) it holds that

\[
\beta^*_j \text{ is identifiable by } (3) \iff \text{Null}(C)_j = \{0\},
\]

where \( \text{Null}(C)_j \) denotes the \( j \)-th coordinate of the null space of \( C \). Moreover, whenever \( \text{Null}(C)_j = \{0\} \) it holds that \( \beta^*_j = (\text{Cov}(I, X)^{\top} \text{Cov}(I, Y))_j \), where \( (\cdot)^{\top} \) denotes the Moore-Penrose inverse.

The proof can be found in Appendix A.

### 3 IDENTIFIABILITY IN SPARSE-EFFECT IV MODELS

Consider a data generating process of the form (1). Because the intervention \( I \) does not directly enter the structural assignment of \( Y \) and \((H, \varepsilon^Y, I)\) are jointly independent, the causal coefficient \( \beta^* \) satisfies the moment condition

\[
\text{Cov}(I, Y - X^\top \beta^*) = 0.
\]

The solution space of the moment condition is given by

\[
B := \{ \beta \in \mathbb{R}^d \mid \text{Cov}(I, X)\beta = \text{Cov}(I, Y) \}.
\]

As we will see below, under mild conditions on the interventions \( I \), the causal coefficient \( \beta^* \) is a unique solution to this problem.

We now make the following assumptions:

(A1) It holds that \( \text{Rank}(C_{PA(Y)}) = |PA(Y)| \).

(A2) For all \( S \subseteq \{1, \ldots, d\} \) it holds that

\[
\text{Rank}(C_S) \leq \text{Rank}(C_{PA(Y)}) \quad \text{and} \quad \text{Im}(C_S) \neq \text{Im}(C_{PA(Y)})
\]

implies

\[
\{ \forall w \in \mathbb{R}^{|S|} : C_{SW} \neq C_{PA(Y)}\beta^*_{PA(Y)} \}.
\]

(A3) For all \( S \subseteq \{1, \ldots, d\} \) with \(|S| = |PA(Y)|\) and \( S \neq PA(Y) \) we have \( \text{Im}(C_S) \neq \text{Im}(C_{PA(Y)}) \).

(A1) is a necessary assumption in order to identify \( \beta^* \); it guarantees that an IV regression based on the true parent set \( PA(Y) \) identifies the correct coefficients. (A2) is an assumption on the underlying causal model that ensures

---

**Figure 2:** Graphical representation of Example 1 as described in Section 2.1 (the hidden variable \( H \) is omitted).
that certain types of cancellation cannot occur. It is a rather
mild assumption in the following sense: if one considers
the true causal parameter $\beta^*$ as randomly drawn from a
distribution absolutely continuous with respect to Lebesgue
measure it would almost surely lead to a system that satisfies
(A2) (see Proposition 9 in Appendix B). As shown in the
following theorem, (A1) and (A2) are sufficient to ensure that
$\beta^*$ solves (5). Additionally assuming (A3) ensures that
the solution is unique: it can be seen as requiring an extra
level of heterogeneity in how the interventions affect the
system (see also Section 4).

**Theorem 3** (Identifiability of sparse causal parameters).
Consider a data generating process of the form (1). If (A1)
and (A2) hold, then $\beta^*$ is a solution to (5). Moreover, if, in
addition, (A3) holds, then $\beta^*$ is the unique solution.

**Proof.** We use the notation $\xi^X := h(H,\varepsilon^X)$ and $\xi^Y :=
g(H,\varepsilon^X)$. Then (1) and the assumption of joint independence
of $I$, $\xi^X$, and $\xi^Y$ imply that
\[
\text{Cov}[I, X] = \text{Cov} \left[ I, (\text{Id} - B)^{-1}(AI + \xi^X) \right] = \text{Cov}[I]A^\top(\text{Id} - B)^{-\top}.
\]
(6)

Similarly, we get that
\[
\text{Cov}[I, Y] = \text{Cov} \left[ I, (AI + \xi^X)^\top(\text{Id} - B)^{-\top}\beta^* + \xi^Y \right] = \text{Cov}[I, \beta^*\top(\text{Id} - B)^{-1}(AI + \xi^X) + \xi^Y] = \text{Cov}[I]A^\top(\text{Id} - B)^{-\top}\beta^*.
\]
(7)

Hence, for any $\tilde{\beta} \in B$, using the definition of $B$ and combining (6) and (7) we get that
\[
\text{Cov}[I]C\beta = \text{Cov}[I]C\beta^*.
\]

Here, we used the definition of $C$ in (4). As $\text{Cov}[I]$ is in-
vertible, we get
\[
C\beta = C\beta^*,
\]
(8)

Furthermore, it holds for $S = \text{supp}(\tilde{\beta})$ that
\[
C_S\beta_S = C_{PA(Y)}\beta^*_{PA(Y)}.
\]
(9)

We now prove the first part of the theorem. Assume (A1)
and (A2) are satisfied. We want to show that
\[
\beta^* \in \arg\min_{\beta \in B}\|\beta\|_0.
\]
(10)

Since $\beta^* \in B$, it is sufficient to show that for all $\tilde{\beta} \in B$
it holds that $\|\tilde{\beta}\|_0 \geq |PA(Y)|$. To this end, fix $\beta \in B$ and
set $S = \text{supp}(\beta)$. For the sake of contradiction assume
$|S| < |PA(Y)|$, then using (A1) we get that
\[
\text{Rank}(C_{PA(Y)}) = \dim(\text{Im}(C_{PA(Y)})) = |PA(Y)| > S \geq \dim(\text{Im}(C_S)) = \text{Rank}(C_S).
\]

This implies $\text{Rank}(C_{PA(Y)}) \geq \text{Rank}(C_S)$ and
$\text{Im}(C_{PA(Y)}) \neq \text{Im}(C_S)$. Thus, by (A2), this contradicts (9). This completes the first part of the proof.

Next, we prove the second part of the theorem. Assume that
(A1), (A2) and (A3) are satisfied. By the previous part of
the proof, we have seen that $\beta^*$ satisfies (10). It therefore
only remains to show that there is no other solution. Assume
for the sake of contradiction that there exists $\tilde{\beta} \in B$ with
$S := \text{supp}(\tilde{\beta})$ such that $|S| = |PA(Y)|$ and $S \neq PA(Y)$. Then
by (A3) we have $\text{Im}(C_S) \neq \text{Im}(C_{PA(Y)})$. By (A1) it holds that
\[
\text{Rank}(C_{PA(Y)}) = |PA(Y)| = |S| \geq \text{Rank}(C_S).
\]

Hence, together with the condition $\text{Im}(C_S) \neq \text{Im}(C_{PA(Y)})$
we can use (A2) to get a contradiction to (9). This completes
the proof of Theorem 3.

Section 5.1 shows how one can identify a subset of the
causal parents under even milder conditions. Remark 5 dis-
cusses the case where the covariates can also be descendents
of $Y$.

## 4 Graphical Characterization

We now formulate the identifiability result from Section 3.1 in
graphical terms. Suppose we are given a data generating
process of the form (1) with corresponding graph $G$ (as
described in Section 2.1), which in this section is assumed
be acyclic. The parents of $Y$ are denoted by $PA(Y)$ and
 correspon-
d to the non-zero entries of $\beta^*$. Moreover, for any
set $S \subseteq \{1, \ldots, d\}$, we define the set of all intervention
ancestors of variables in $S$ as
\[
AN_t[S] := \{ j \in \{1, \ldots, m\} \mid j \in AN(S) \}\]
4

This set contains the instrument nodes that are ancestors of
$S$.

We can now state the following graphical assumptions.

(B1) There are at least $|PA(Y)|$ disjoint directed paths (not sharing
any node) from $I$ to $PA(Y)$.

(B2) The non-zero coefficients of the causal coefficient $\beta^*_{PA(Y)} \in \mathbb{R}^{\text{PA}(Y)}$ and the non-zero entries of $A$ and
$B$ are randomly drawn from a distribution $\mu$ which is
absolutely continuous with respect to Lebesgue measure
(and are independent of the other variables).

(B3) For all $S \subseteq \{1, \ldots, d\}$ with $|S| = |PA(Y)|$ and
$S \neq PA(Y)$ at least one of the following conditions is satisfied

4$AN(S)$ denotes the ancestor set of $S$ consisting of all nodes
with a directed path to a node in $S$. Throughout the paper, we use
the convention that a node is contained in the set of its ancestors.
We will see in Theorem 5 below that the causal effect becomes identifiable if (B1)–(B3) hold. Let us discuss these assumptions using a few examples.

**Example 4.** (i) The example from Example 7 and Figure 2 is discussed in Figure 7 in Appendix E.

(ii) We now come back to the example graphs shown in Figure 2 Consider an SCM with the graph structure shown in Figure 2 (dashed edges not included). Here, (B1) holds as there are two distinct paths from $I$ to PA($Y$). (B3) is satisfied, too: $\{X^1, X^5\}$ and $\{X^3, X^2\}$ are the only sets $\mathcal{S}$ violating (B3) (i) (because $\text{AN}_I[\{X^1, X^5\}] = \text{AN}_I[\{X^1, X^2\}] = \text{AN}_I[\{X^3, X^2\}]$), but they satisfy (B3) (ii). Thus, given (B2), the causal effect from $X^1, X^2$ on $Y$ is identifiable – even though there are less instruments than covariates.

(iii) We now come back to the example graphs shown in Figure 1. Consider an SCM with the graph structure shown in Figure 1. Figure 2 is discussed in Figure 7 in Appendix E.

(iv) Consider an SCM with the graph structure shown in Figure 2 (dashed edges not included). Here, (B1) holds. (B3) is satisfied, too: e.g., for the set $\mathcal{S} := \{X^3, X^4\}$, we have $\text{AN}_I[\{X^3, X^4\}] = \text{AN}_I[\{X^3, X^2\}]$, so (B3) (i) is violated, but (B3) (ii) holds (which implies $\text{Im}(\mathcal{C}_S) \neq \text{Im}(\text{PA}(Y))$, see proof of Theorem 5): there is no set of size two such that all directed paths go through this set (note that $|\text{AN}(\{X^3, X^4\})| = 3$, for example). Thus, if additionally (B2) holds, then $\beta^*$ is identifiable.

A graphical marginalization of graphs similar to the latent projection [Richardson, 2003; Verma, 1993] may help to gain further intuition about the assumptions. Consider a subset $V \subseteq \{1, \ldots, d\}$ of the covariates. The marginalized graph $\mathcal{G}^V$ is then constructed from $\mathcal{G}$ by the following procedure: (i) $\mathcal{G}^V$ consists of all instrument nodes $k$ from $\mathcal{G}$, all predictor nodes $X^j$ from $\mathcal{G}$ for which $j \in V$, and node $Y$; (ii) $\mathcal{G}^V$ contains a directed edge from $X^i$ to $X^j$ if and only if $\mathcal{G}$ contains a directed path from $X^i$ to $X^j$ that does not have any intermediate nodes in $V$ (e.g., because there are no intermediate nodes); (iii) $\mathcal{G}^V$ contains a directed edge from $k$ to $X^j$ if and only if $\mathcal{G}$ contains a directed path from $k$ to $X^j$ that does not have any intermediate nodes in $V$ (e.g., because there are no intermediate nodes). The set $\text{PA}_I^V[U]$ denotes the $\mathcal{G}^V$-parents of $U$ that are intervention nodes: $\text{PA}_I^V[U] := \text{PA}_{\mathcal{G}^V} \cap I$. Figure 7 in Appendix E shows the marginalized graph corresponding to Example 7 and Figure 2.

(B1) ensures that there is sufficient heterogeneity coming from instruments. In particular, there need to be as many instruments as parents of $Y$ and for all $S \subseteq \text{PA}(Y)$, we have $|\text{PA}_I^\text{PA}(Y)[S]| \geq |S|$. In particular, this implies that for all $k \in \text{PA}(Y)$, we have $\text{PA}_I^\text{PA}(Y)[k] \neq \emptyset$. In general, however, this is not sufficient for identifiability (see Section 5).

We can now state the graphical version of Theorem 3.

**Theorem 5 (Identifiability of sparse causal coefficients (graph version)).** Consider a data generating process of the form (1). If (B1) and (B2) hold, then (A1) and (A2) hold $\mu$-almost surely. Moreover, if, in addition, (B3) holds, then (A3) holds $\mu$-almost surely.

Together with Theorem 3, this implies that under (B1) and (B2), $\beta^*$ is $\mu$-almost surely a solution to (5) and (B1), (B2), and (B3), it is $\mu$-almost surely the unique solution.

**Proof.** Regarding (A1): With respect to (A1), consider first the marginalization of model (1) over $\text{PA}(Y)$. To do so, we repeatedly substitute $X^j$, $j \in \{1, \ldots, d\}$ with its assignment, that is, the corresponding right-hand side of (1) and obtain

$$X^{\text{PA}(Y)} := C^{\times^T} I + h^{\times}(H, \varepsilon^{\times}).$$

We then have

$$C_{i, \text{PA}(Y)} = C^{\times},$$

where $C_{i, \text{PA}(Y)}$ is the matrix constructed from the columns of $C$ corresponding to $\text{PA}(Y)$. Equality (12) holds by construction: The element of $C_{i, \text{PA}(Y)}$ in row $i$ and the column corresponding to $X^j \in \text{PA}(Y)$ equals the $i$-th component of the total causal effect from $I$ on $X^j$; this is exactly the same in the marginalized model (1). We now argue that $C^{\times}$ has full rank $\mu$-almost surely. To do so, we perform a more careful replacement scheme that allows us to write

$$X^{\text{PA}(Y)} = C_1 \cdot C_2 \cdot \ldots \cdot C_f \cdot I + h^{\times}(H, \varepsilon^{\times}).$$

Figure 3: Graphical representation of an example SCM, as described in Section 2.1 (there may be hidden variables between all predictor variables). Here, (B1) holds as there are two distinct paths from $I$ to $\text{PA}(Y)$. (B3) is satisfied, too: $\{X^1, X^5\}$ and $\{X^3, X^2\}$ are the only sets $\mathcal{S}$ violating (B3) (i) (because $\text{AN}_I[\{X^1, X^5\}] = \text{AN}_I[\{X^1, X^2\}] = \text{AN}_I[\{X^3, X^2\}]$), but they satisfy (B3) (ii). Thus, given (B2), the causal effect from $X^1, X^2$ on $Y$ is identifiable – even though there are less instruments than covariates.
It then holds that $C^X = (C_1, C_2, \ldots, C_f)^\top$. As a first step of the replacement scheme, consider all $X$ nodes on directed paths from $I$ to $PA(Y)$, that is $W := AN(PA(Y)) \cap DE(I)$. Among these nodes we consider a causal ordering on the induced graph, that is, we choose $i_1, \ldots, i_f$ such that for all $k, \ell \in \{1, \ldots, f\}$ with $k < \ell$, we have $X_{i_k} \in N_D g_{iw}(X_{i_\ell})$ (ND denotes the "non-descendants"), where $G_{iw}$ is the subgraph of $G$ over nodes in $W$. We now start from the equation $X^\text{PA(Y)} = X^\text{PA}C^\text{PA} = A^\text{PA(Y)} + A^\text{PA} \cdot h_1(H, W^c, \varepsilon_{X^i})$, where $A^\text{PA} = \{A^\text{PA}Y, A^\text{PA}W, \{X_{i_k}\}\}$ and $A^\text{PA} \cdot h_1(H, W^c, \varepsilon_{X^i})$ is the term collecting error terms and variables not in $W$. $A^\text{PA} \cdot h_1$ is a matrix with dimension $|PA(Y)| \times |PA(Y)|$. We did not replace the variables in $PA(Y)$ \{\{X_{i_k}\}\}, the corresponding submatrix in $C^\text{PA}$ is the identity. All directed paths from $I$ to $PA(Y)$ go through $PA_1$, Condition (B1) therefore implies $|PA_1| \geq |PA(Y)|$. The row corresponding to $X_{i_k}$ contains the path coefficients from $X^\text{PA(Y)}$ to $X_{i_k}$, which are $\mu$-almost surely non-zero. Thus, $C^\text{PA}$ has $\mu$-almost surely rank $|PA(Y)|$. We now repeatedly (for $k \in \{2, \ldots, f\}$) substitute the variable $X_{i_k}$ in $X^\text{PAk-1}$ with its structural equation yielding $X^\text{PA(Y)} = C_1 \cdot C_2 \cdot \ldots \cdot C_{f-1} \cdot h_1(H, W^c, \varepsilon_{X^i})$, where $PA_{k} = \{PA_{k-1} \setminus \{X_{i_k}\}\} \cup PA_{iw}(X_{i_k})$ and $C_k$ contains an identity matrix for the submatrix, corresponding to $PA_{k} \cap PA_{k-1}$ and in the row corresponding to $X_{i_k}$ a vector of coefficients. With the same arguments as above, we have that $|PA_{k}| \geq |PA(Y)|$. (Indeed, otherwise, all directed path would go through a set of nodes of size strictly smaller than $|PA(Y)|$.) Furthermore, $C_k$ has rank at least $|PA(Y)|$. (Indeed, if $|PA_{k-1}| > |PA(Y)|$, then $C_k$ contains a $|PA(Y)| \times |PA(Y)|$ submatrix that is equal to the identity; if $|PA_{k-1}| = |PA(Y)|$, then $C_k$ contains a $(|PA(Y)| - 1) \times (|PA(Y)| - 1)$ submatrix that is equal to the identity, $PA_k \setminus PA_{k-1} \neq \emptyset$, and the entry corresponding to one of the new parents will be non-zero $\mu$-almost surely.) As $W^c$ can be written as a function of $\varepsilon_X$, the above replacement scheme yields the desired form $[13]$. Since $C^X = (C_1, C_2, \ldots, C_f)^\top$ and all non-zero entries are independent realizations from $\mu$, this proves that $C^X \mu$-almost surely of rank at least $|PA(Y)|$, that is, (A1) holds $\mu$-almost surely.

Regarding (A2): Proposition 9 shows that (A2) holds $\mu$-almost surely.

Regarding (A3): Consider a set $S \subseteq \{1, \ldots, d\}$ with $|S| = |PA(Y)|$ and $S \neq PA(Y)$. First, we argue that (B3) (i) implies (A3). To see this, assume $AN_I(S) \neq AN_I(PA(Y))$. Without loss of generality assume that there is an $i^* \in AN_I(S) \setminus AN_I(PA(Y))$. This implies that the $i^*$th row of $C_{PA(Y)}$ is entirely zero. Moreover, there is a node $X_j \in S$ such that $i^* \in AN_I(\{j\})$, and therefore the entry of the $i^*$th row of $C_S$ that corresponds to $X_j$ must be non-zero $\mu$-almost surely $(C_{i,j} \mu$-corresponds to the $i$-th component of the total causal effect from $I$ on $X_j$ in the SCM given in [1]). It therefore follows that $\mu$-almost surely it holds that

$$\text{Im}(C_S) \neq \text{Im}(C_{PA}(Y)) \cdot$$

Now consider a set $S$ and assume that (B3) (i) does not hold but (B3) (ii) holds. To argue that (A3) holds, we proceed similarly as in the part of the proof showing that (B1) implies (A1). We consider the graph $G$ over the nodes $W_{S \cup PA(Y)} := AN(PA(Y) \cup S) \cap DE(I)$. As before, we construct a causal order and substitute the nodes one after each other. This time, we obtain the equation $X^{S \cup PA(Y)} = C_1 \cdot C_2 \cdot \ldots \cdot C_{f-1} \cdot I + h^X(H, \varepsilon_X)$ and $C_{S \cup PA(Y)} = (C_1 \cdot C_2 \cdot \ldots \cdot C_f)^\top$. With the same argument as above, we conclude that $\mu$-almost surely, the rank of $C_{S \cup PA(Y)}$ is strictly larger than $|PA(Y)|$. This implies that $\text{Im}(C_S) \neq \text{Im}(C_{PA}(Y)) \mu$-almost surely. (Indeed, if $\text{Im}(C_S) = \text{Im}(C_{PA}(Y))$, then each column of $C_S$ can be written as a linear combination of the columns of $C_{PA(Y)}$, which implies that $C_{S \cup PA(Y)}$ is of rank at most $|PA(Y)|$.) This completes the proof of Theorem 5.

5 ALGORITHM AND CONSISTENCY

The theoretical identifiability results from the previous sections highlight that the causal coefficient $\beta^*$ can be identifiable even in cases that are considered non-identifiable in classical IV literature. We now propose an estimation procedure called spaceIV (sparse causal effect IV) that allows us to infer $\beta^*$ from a finite data set $(X, Y) \in \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times m} \times \mathbb{R}^n$. The procedure is based on the optimization problem $\min_{\beta \in B} \|\beta\|_0$. It iterates over the sparsity level $s$ and searches over all subsets $S \subseteq \{1, \ldots, S\}$ of predictors for that sparsity level to check whether there is a $\beta \in \mathbb{R}^d$ with $\text{supp}(\beta) = S$ that solves [3]. We motivate our estimator by considering a hypothesis test. To obtain finite sample guarantees for the test, we assume that the error term is normally distributed (to obtain asymptotic results [Anderson and Rubin [1950]], such assumptions can be relaxed).

Let us consider a fixed sparsity level $s \in \{1, \ldots, d\}$ and the null hypothesis

$$H_0(s) : \exists \beta \in \mathbb{R}^d \text{ with } \|\beta\|_0 = s \text{ such that } \beta \in B.$$

This hypothesis can be tested using the Anderson-Rubin test [Anderson and Rubin [1949]]. Let $T^ I := (I^\top I)^{-1/2} I^\top$, then the Anderson-Rubin test statistic is defined as

$$T(\beta) := \frac{(Y - \bar{X})^\top P_I(Y - \bar{X})}{(Y - \bar{X})^\top (I - P_I)(Y - \bar{X})} \frac{n - m}{m}, \quad (15)$$
and (still, for Gaussian error variables) satisfies $T(\beta) \sim F_{n-m,m}$, i.e. an $F$ distribution with $n-m$ and $m$ degrees of freedom, for all $\beta \in \mathcal{B}$, see also Jakobsen and Peters [2022]. It is known [e.g., Dhrymes, 2012] that the limited maximum likelihood estimator (LIML) minimizes this test statistic. For any set $S \subseteq \{1, \ldots, d\}$, denote by $\hat{\beta}_{\text{LIML}}(S) \in \mathbb{R}^d$ the LIML estimator based on the subset of predictors $X^S$ (adding zeros in the other coordinates). It then holds for all $\beta \in \mathbb{R}^d$ with $\text{supp}(\beta) = S$ that

$$T(\hat{\beta}_{\text{LIML}}(S)) \leq T(\beta).$$  \hfill (16)

Next, for each sparsity level $s \in \{1, \ldots, d\}$ define

$$\hat{\beta}(s) := \hat{\beta}_{\text{LIML}} \left( \arg\min_{S \subseteq \{1, \ldots, d\}, |S| = s} T(\hat{\beta}_{\text{LIML}}(S)) \right),$$  \hfill (17)

which can be computed by iterating over all subsets with sparsity level $s$. Then, by (16), the hypothesis test $\phi_s : \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times m} \times \mathbb{R}^d \rightarrow \{0, 1\}$ defined by

$$\phi_s(X, I, Y) = \mathbb{1}(T(\hat{\beta}(s)) > F_{n-m,m}(1 - \alpha))$$

has valid level for the null hypothesis $H_0(s)$ (again, if the error variables are Gaussian; otherwise it has uniform asymptotic level given sufficient regularity).

Motivated by this test, we now define our estimator \text{spaceIV}. It iterates over $s \in \{1, \ldots, s_{\text{max}}\}$ and in each step computes $\hat{\beta}(s)$ by exhaustively searching over all subsets of size $s$. Then, either $\phi_s$ is accepted and \text{spaceIV} returns $\hat{\beta}_{\leq s_{\text{max}}} := \hat{\beta}(s)$ as its final estimator or it continues with $s + 1$. If none of the tests are accepted, the procedure outputs $\hat{\beta}_{\leq s_{\text{max}}} := \hat{\beta}(s_{\text{max}})$ and a warning indicating that the model assumptions may be violated. The detailed procedure is presented in Algorithm 1.

The proposed \text{spaceIV} estimator $\hat{\beta}_{\leq s_{\text{max}}}$ satisfies the following guarantees.

**Theorem 6.** Consider i.i.d. data from a data generating process of the form (1) for which $g(H, \varepsilon^Y)$ is Gaussian, $(I, X, Y)$ has mean zero and (A1) and (A2) hold. Let $s_{\text{max}} \in \mathbb{N}$ be such that $s_{\text{max}} \geq \|\beta^*\|_0$. Then, the following two statements hold. (i) We have

$$\lim_{n \rightarrow \infty} P(||\hat{\beta}_{\leq s_{\text{max}}}||_0 = \|\beta^*\|_0) \geq 1 - \alpha.$$  

(ii) If, in addition, (A3) holds, we have, for all $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(||\hat{\beta}_{\leq s_{\text{max}}} - \beta^*\|_2 < \varepsilon) \geq 1 - \alpha.$$  

The proof can be found in Appendix C.

### 5.1 CAUSAL SUBSET IDENTIFIABILITY

It is possible to identify a subset of the causal parents under even weaker conditions. This can be done in an idea similar to invariant causal prediction [Peters et al., 2016]. Define the hypothesis

$$H_0(S) : \exists \beta \in \mathbb{R}^d \text{ such that } \text{supp}(\beta) = S \text{ and } \beta \in \mathcal{B}$$

and the corresponding Anderson-Rubin test

$$\mathbb{1}(T(\hat{\beta}_{\text{LIML}}(S)) > F_{n-m,m}(1 - \alpha)).$$

We then have the following guarantees.

**Proposition 7.** (i) Consider i.i.d. data of $(I, X, Y)$ from a data generating process of the form

$$Y := X^\top \beta^* + g(H, \varepsilon^Y),$$

with $I \perp (H, \varepsilon^Y)$, $g(H, \varepsilon^Y)$ Gaussian and $(I, X, Y)$ mean zero. Then,

$$\lim_{n \rightarrow \infty} P \left( \bigcap_{S : |S| = |\text{PA}[Y]| \text{ and } S \subseteq \text{PA}[Y]} H_0(S) \text{ accepted} \right) \geq 1 - \alpha,$$  \hfill (18)

where we define the intersection over an empty index set as the empty set.

(ii) Consider now i.i.d. data from a data generating process of the form (1) such that $g(H, \varepsilon^Y)$ is Gaussian and $(I, X, Y)$ has mean zero. If (A1) and (A2) hold, then

$$\lim_{n \rightarrow \infty} P \left( \bigcap_{S : |S| = M \text{ and } S \subseteq \text{PA}[Y]} H_0(S) \text{ accepted} \right) \geq 1 - \alpha,$$  \hfill (19)

where $M := \min\{|S| : H_0(S) \text{ accepted}\}$.  

---

*Accepted (UPDATED VERSION) for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).*
The first statement requires the sparsity \( \|\beta^*\|_0 \) of \( \beta^* \) to be known. It still holds when replacing \( |PA[Y]| \) by any \( k \) such that \( |PA[Y]| \leq k \leq d \). The second statement does not require knowledge of \( \|\beta^*\|_0 \) and provides a guarantee when increasing the subset size until one has found a set that is accepted. The proof of Proposition \[4] can be found in Appendix \[D].

Remark 8 (Allowing for children of \( Y \)). We now discuss the scenario where some of the covariates are causal descendants of the response \( Y \). More precisely, we extend the model in (1) to

\[
X := BX + \gamma Y + AI + h(H, \epsilon^X) \\
Y := \beta^*^\top X + g(H, \epsilon^Y),
\]

where we assume that the matrix

\[
B_{ext} := \begin{pmatrix} B & 0 \\ (\beta^*)^\top & 0 \end{pmatrix} \in \mathbb{R}^{(d+1) \times (d+1)}
\]

is invertible. Theorem \[3\] and therefore also the results in Sections \[5\] and \[5.1\] still hold when using \( C_{ext} := (A^\top, 0)(\text{Id} - B_{ext})_{1:d} \) instead of \( C \). Assumption (A2), however, becomes rather restrictive: If there is a child of \( Y \) such that all directed paths from \( I \) to that child go through \( Y \), (A2) is not satisfied as the exact intervention effect on \( Y \) is recoverable from that child. In particular, in this generalized setting, Proposition \[5\] or (B1) and (B2) no longer imply that (A2) holds almost surely.

6 NUMERICAL EXPERIMENTS

For the numerical experiments, we consider models of the form (1) with \( h(H, \epsilon^X) = H + \epsilon^X, g(H, \epsilon^Y) = H + \epsilon^Y \) and dimensions \( d = 20, q = 1 \) and \( m = 10 \). We generate 2000 random models of this form using the following procedure:

- Generate a random matrix \( B \in \mathbb{R}^{20 \times 20} \) by sampling a random causal order over \( X^1, \ldots, X^{20} \). \( B \) then has a zero-structure that corresponds to a fully connected graph with this causal order. Each non-zero entry in \( B \) is drawn independently and uniformly from \((-1.5, -0.5) \cup (0.5, 1.5) \). Finally, each row of \( B \) is rescaled by the maximal value in each row (using one if it is a zero row).
- Generate a random matrix \( A \in \mathbb{R}^{20 \times 10} \) by sampling each entry independently with distribution \( \text{Bernoulli}(1/10) \) and setting all diagonal entries to 1.
- Generate the parameter \( \beta^* \in \mathbb{R}^{20} \) by sampling two random coordinates uniformly from \( \{1, \ldots, d\} \) and setting them to 1. All remaining coordinates are set to zero.
- The random variables \( I, H, \epsilon^X \) and \( \epsilon^Y \) are all drawn as i.i.d. standard normal.

For each random model we sample 6 data sets with sample sizes \( n \in \{50, 100, 200, 400, 800, 1600\} \). For each data set, we apply the following four methods: (i) \text{spaceIV} \; ; this is our proposed method described in Algorithm \[1\] with \( s_{\text{max}} = 3 \). (ii) \text{OLS-sparse} \; ; this method goes over all subsets of size at most \( s_{\text{max}} \), fits a linear OLS and then selects the subset with the smallest AIC. We also compare our estimator to two oracle methods. (iii) \text{oracle-|PA|} \; ; this method iterates over all subsets with size 2 (correct parent size), fits the moment equation \[3\] and selects the best subset in terms of a squared loss based on the moment equation. (iv) \text{oracle-PA} \; ; this method considers the correct parent set and fits the moment equation \[3\]. Each method results in a sparse estimate \( \hat{\beta} \) of \( \beta^* \) based on which we compute the root mean squared error (RMSE) given by \( \|\hat{\beta} - \beta\|_2 \).

For each random model, we explicitly check whether the assumptions (A1) and (A3) are satisfied by computing \( C \) and verifying the conditions \[6\]. The results, considering only the random models for which assumptions (A1)–(A3) are satisfied, are given in Figure \[4\]. As expected, \text{spaceIV} indeed seems to consistently estimate the causal parameter \( \beta^* \), while \text{OLS-sparse} does not. Furthermore, \text{spaceIV} performs worse as the two oracle methods, illustrating that the estimation in \text{spaceIV} contains three parts: estimating the correct sparsity, estimating the correct parents set and finally estimating the correct parameters. A mistake in any of these three steps may result in substantial RMSE, which explains the outliers in the plot.

To investigate the consistency of estimating the correct sparsity level in more detail, we consider the fraction of times the correct sparsity level was selected by \text{spaceIV}.

\[4\] Assumption (A2) is satisfied by construction because we pick random coefficients for the \( B \)-matrix, see also (B2).
result is given in Figure 5. It suggests that the sparsity level is consistently estimated by \text{spaceIV}. 

Figure 5: Expected fraction of random models for which \text{spaceIV} estimated the correct sparsity level. Only random models that satisfy (A1)-(A3) are considered (in total 1867 models). As the sample size increases the estimation of the sparsity level becomes more accurate.

Finally, to investigate the performance of \text{spaceIV} based on the assumptions (A1)–(A3), we compared the performance of all methods at sample size $n = 1600$ depending on which assumptions are satisfied. (Assumption (A2) is satisfied with probability one, see Proposition 9.) The results are shown in Figure 6. As expected given the theoretical results presented in Section 3.1, \text{spaceIV} only performs well if all assumptions are satisfied. If only assumption (A1) is satisfied, there are multiple sets with sparsity 2 for which the moment equation (3) can be satisfied. Therefore, while the oracle with the correct parent sets is able to estimate the causal parameter, \text{spaceIV} and the oracle that only uses the sparsity level may select wrong sets leading to a larger error. Moreover, if none of the assumptions are satisfied the causal parameter is not even identifiable if the true parent set is known.

Figure 6: Results for all 2000 random models with $n = 1600$. We split the models into three cases depending on which of the assumptions (A1) and (A3) are satisfied (the group ‘(A1)’ contains 83 models, the group ‘(A1) & (A3)’ contains 1867 models and the group ‘none’ contains 50 models). If none of the assumptions are satisfied, not even the oracle with known parent set works. If only (A1) is satisfied, multiple sets of size 2 are able to satisfy the moment equation (3) and \text{spaceIV} may not estimate the correct set. These findings are in par with Theorem 5.

may yield ways for choosing a significance level for finite samples. Furthermore, it could be interesting to investigate to which extent our results generalize to nonlinear models.

Acknowledgements

NP was supported by a research grant (0069071) from Novo Nordisk Fonden. JP was supported by a research grant (18968) from VILLUM FONDEN.

References

T. Amemiya. Advanced Econometrics. Harvard University Press, Cambridge, MA, 1985.

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.

T. W. Anderson and H. Rubin. The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. Annals of Mathematical Statistics, 21:570–582, 1950.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6): 2369–2429, 2012.
A. Bennett, N. Kallus, and T. Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*. Curran Associates, Inc., 2019.

S. Bongers, P. Forre, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *Annals of Statistics*, 49(5):2885–2915, 2021.

R. Christiansen, N. Pfister, M. Jakobsen, N. Gnecco, and J. Peters. The difficult task of distributional generalization in nonlinear models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (accepted); ArXiv e-prints (2006.07433), 2020.

P. J. Dhrymes. *Econometrics: Statistical foundations and applications*. Springer Science & Business Media, 2012.

F. Dunker, J.-P. Florens, T. Hohage, J. Johannes, and E. Mammen. Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *Journal of Econometrics*, 178:444–455, 2014.

J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1414–1423. PMLR, 2017.

S. Huang, N. Pfister, and J. Bowden. Sparse causal effect estimation using two-sample summary statistics in the presence of unmeasured confounding. *ArXiv e-prints* (2410.12300), 2024.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62 (2):467–75, 1994.

G. W. Imbens and W. K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

M. Jakobsen and J. Peters. Distributional robustness of k-class estimators and the PULSE. *The Econometrics Journal*, 25(2):404–432, 2022.

H. Kang, A. Zhang, 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.

S. Lauritzen. *Graphical Models*. Oxford University Press, New York, NY, 1996.

I. Loh. Nonparametric identification and estimation with independent, discrete instruments. *ArXiv e-prints* (1906.05231), 2019.

R. S. Mariano. Simultaneous equation model estimators: Statistical properties and practical implications. In Badi H. Baltagi, editor, *A companion to theoretical econometrics*, pages 122–43. Blackwell, Malden, MA, 2001.

P. Mckeigue, J. Krohn, A. J. Storkey, and F. Agakov. Sparse instrumental variables (spiv) for genome-wide studies. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23. Curran Associates, Inc., 2010.

K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. Dual instrumental variable regression. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*. Curran Associates, Inc., 2020.

W. K. Newey. Nonparametric instrumental variables estimation. *American Economic Review*, 103(3):550–556, 2013.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, 2nd edition, 2009.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B* (with discussion), 78(5):947–1012, 2016.

T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

S. Saengkyongam, L. Henckel, N. Pfister, and J. Peters. Exploiting independent instruments: Identification and distribution generalization. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, accepted, ArXiv e-prints (2202.01864), 2022.

R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*. Curran Associates, Inc., 2019.

A. Torgovitsky. Identification of nonseparable models using instruments with small support. *Econometrica*, 83(3):1185–1197, 2015.

T. S. Verma. Graphical aspects of causal models. Technical Report R-191, Computer Science Department, University of California, Los Angeles, CA, 1993.

P. G. Wright. *The Tariff on Animal and Vegetable Oils*. Investigations in International Commercial Policies. Macmillan, New York, NY, 1928.

*Accepted (UPDATED VERSION) for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).*
A PROOF OF PROPOSITION 2

Proof. Fix \( j \in \{1, \ldots, d\} \), then it holds that \( \beta_j^* \) is identifiable by (3) if and only if the space \( B \) is degenerate in the \( j \)-th coordinate, that is, \( B_j = \{ \beta_j^* \} \). Next, define \( M := \text{Cov}(I, X) \) and \( v := \text{Cov}(I, Y) \). Then, denoting the Moore-Penrose inverse of \( M \) by \( M^\dagger \), we get that for any solution \( \beta \in B \) there exists \( w \in \text{Null}(M) \subseteq \mathbb{R}^d \) such that

\[
\beta = M^\dagger v + w.
\]  
(20)

Therefore, the space \( B \) has a degenerate \( j \)-th coordinate if and only if \( \text{Null}(M) \subseteq \{0\} \). Using \( M^\dagger \), the null space of \( M \) can be expressed as

\[
\text{Null}(M) = \{(\text{Id} - M^\dagger M)w \mid w \in \mathbb{R}^d\}.
\]

Next, \((1)\) and the assumption of joint independence of \( I, \xi^X := h(H, e^X) \) and \( \xi^Y := g(H, e^X) \) imply that

\[
M = \text{Cov}[I, X] = \text{Cov}[I, (\text{Id} - B)^{-1}(AI + \xi^X)] = \text{Cov}[I, A^\top (\text{Id} - B)^{-\top}] = \text{Cov}[I, C].
\]

Therefore, using the properties of the Moore-Penrose inverse and that \( \text{Cov}[I] \) is invertible we get that

\[
M^\dagger M = C^\dagger \text{Cov}[I]^{-1} \text{Cov}[I] C.
\]

Hence, we get that \( M^\dagger M = C^\dagger C \) which implies that \( \text{Null}(M) = \text{Null}(C) \). This proves the first part of the statement. The second part of the proposition uses (20) together with \( \text{Null}(M) = \text{Null}(C) \). This completes the proof of Proposition 2. \( \square \)

B FURTHER RESULTS

Proposition 9. Let \( A \in \mathbb{R}^{n \times m} \) and \( B \in \mathbb{R}^{n \times p} \) be two matrices satisfying

\[
\text{Rank}(B) \leq \text{Rank}(A) \quad \text{and} \quad \text{Im}(A) \neq \text{Im}(B)
\]

and let \( W \in \mathbb{R}^m \) be a random variable with a distribution on \( \mathbb{R}^m \) that is absolutely continuous with respect to Lebesgue measure. Then it holds that

\[
P(AW \in \text{Im}(B)) = 0.
\]

Proof. We begin by showing that

\[
\text{Im}(B)^\perp \cap \text{Im}(A) \neq \emptyset.
\]

(22)

Assume for the sake of contradiction this is not true. Then it would hold that \( \text{Im}(A) \subseteq \text{Im}(B) \). Moreover, since by assumption \( \text{Rank}(B) \leq \text{Rank}(A) \) this would imply that \( \text{Im}(A) = \text{Im}(B) \), which contradicts the assumptions on \( A \) and \( B \). Hence, (22) is true.

Next, let \( b_1, \ldots, b_n \in \mathbb{R}^n \) be an orthogonal basis of \( \mathbb{R}^n \) such that

\[
\text{span}(b_1, \ldots, b_k) = \text{Im}(B)^\perp
\]

and

\[
\text{span}(b_{k+1}, \ldots, b_n) = \text{Im}(B).
\]

Then, for every \( \ell \in \{1, \ldots, m\} \) there exists unique \( \alpha^\ell_1, \ldots, \alpha^\ell_n \in \mathbb{R} \) such that

\[
A^\ell = \sum_{i=1}^n \alpha^\ell_i b_i.
\]

Furthermore, by (22), it holds that there exists at least one \( \ell^* \in \{1, \ldots, k\} \) and \( \ell^* \in \{1, \ldots, m\} \) such that \( \alpha^\ell^*_i \neq 0 \). Furthermore, for every \( w \in \mathbb{R}^m \) it holds that

\[
Aw = \sum_{\ell=1}^m w^\ell A^\ell = \sum_{\ell=1}^m \sum_{i=1}^n w^\ell_i \alpha^\ell_i b_i = \sum_{i=1}^n \left( \sum_{\ell=1}^m w^\ell_i \alpha^\ell_i \right) b_i.
\]

This implies that \( Aw \in \text{Im}(B) \) if and only if \( \sum_{\ell=1}^m w^\ell_i \alpha^\ell_i = 0 \) for all \( i \in \{1, \ldots, k\} \). Using this we get

\[
P(AW \in \text{Im}(B)) = P(\forall i \in \{1, \ldots, k\} : \sum_{\ell=1}^m w^\ell_i \alpha^\ell_i = 0) \leq P(\sum_{\ell \neq \ell^*} w^\ell_{\ell^*} \alpha^\ell_{\ell^*} = 0) = 0,
\]

where for the last step we used that the distribution of \( W \) is absolutely continuous with respect to Lebesgue measure. This completes the proof of Proposition 9. \( \square \)

C PROOF OF THEOREM 6

Proof. \footnote{Theorem 6 and this proof have been updated to fix a mistake in the accepted version of this manuscript. The new version uses similar arguments as in the proof of [Huang et al. (2024) Theorem 3.1].}

First, for any collection of i.i.d. mean zero random variables \( (V_1, W_1), \ldots, (V_n, W_n) \) define \( \text{Cov}(V) := 1/n \sum_{i=1}^n V_i^2 \) and \( \text{Cov}(V, W) := 1/n \sum_{i=1}^n V_i W_i \). Next define for all \( \beta \in \mathbb{R}^d \) the following quantities

\[
\pi := \text{Cov}(I)^{-1} \text{Cov}(I, Y)
\]

\[
\tilde{\pi} := \text{Cov}(I)^{-1} \text{Cov}(I, Y)
\]

\[
\Pi := \text{Cov}(I)^{-1} \text{Cov}(I, X)
\]

\[
\tilde{\Pi} := \text{Cov}(I)^{-1} \text{Cov}(I, X)
\]

\[
\Sigma(\beta) := [\text{Cov}(Y - \pi^T I) + \beta^T \text{Cov}(X - \Pi^T I) \beta - 2 \text{Cov}(Y - \pi^T I, X - \Pi^T I) \beta] \text{Cov}(I)^{-1}
\]

\[
\tilde{\Sigma}(\beta) := [\text{Cov}(Y - \tilde{\pi}^T I) + \beta^T \text{Cov}(X - \tilde{\Pi}^T I) \beta - 2 \text{Cov}(Y - \tilde{\pi}^T I, X - \tilde{\Pi}^T I) \beta] \text{Cov}(I)^{-1}
\]

\[
t(\beta) := \sqrt{n-m} \tilde{\Sigma}(\beta)^{-1/2} (\tilde{\pi} - \tilde{\Pi} \beta)
\]

\[
\mu(\beta) := \sqrt{n-m} \Sigma(\beta)^{-1/2} (\pi - \Pi \beta).
\]
Then, we can reformulate $T(β)$ such that

$$T(β) = \frac{n-m}{m}(\widehat{π} - \widehat{Π}β)^{T}\Sigma(β)^{-1}(\widehat{π} - \widehat{Π}β) = ||t(β)||^2.$$ 

Moreover, let $C \subseteq \mathbb{R}^d$ be a compact set and denote by $β^* \in C$ a minimizer of $\inf_{β \in C}||t(β)||^2$. Then, we can use standard probability bounds to get for all $x \in [0, \infty)$ that

$$P\left( \inf_{β \in C}||t(β)||^2 \leq x \right) = P\left( ||t(β^*) - μ(β^*)||^2 \leq \sqrt{x} \right) \leq P\left( ||t(β^*) - μ(β^*)||^2 \leq \sup_{β \in C}||μ(β)||^2 \right) \leq P\left( ||μ(β)||^2 \leq \sup_{β \in C}||μ(β)||^2, ||t(β^*) - μ(β^*)||^2 \leq \sqrt{x} \right) \leq 2P\left( \sup_{β \in C}||t(β) - μ(β)||^2 \geq \inf_{β \in C}||μ(β)||^2 - \sqrt{x} \right).$$

(23)

Furthermore, it holds that

$$\sup_{β \in C}||t(β) - μ(β)||^2 = \|\Sigma(β)^{-1/2}\|_2 \sqrt{\frac{n-m}{m}}\|π - \widehat{Π}β\|_2 \leq \|\Sigma(β)^{-1/2}\|_{\text{op}} \sqrt{\frac{n-m}{m}}\|π - \widehat{Π}β\|_2 \leq \left(λ_{\min}(\Sigma(β))\right)^{-1/2} \sqrt{m}\|π - \widehat{Π}β\|_2 + \|\Piβ - \widehat{Π}β\|_2.$$ 

By classical asymptotic theory it can be shown that $\widehat{Σ}(β)$ converges to $Σ(β)$, which by assumptions on the model is invertible. Hence, we get that $\sup_{β \in C}||t(β) - μ(β)||^2$ is asymptotically bounded in probability. Furthermore, using that $T$ does not depend on the scale of $β$ it follows together with (24) that

$$P\left( \inf_{β:||β||_0 = s} T(β) \leq x \right) = P\left( \inf_{β:||β||_0 = s, ||β||_2 = 1} T(β) \leq x \right) \leq 2P\left( \sup_{β:||β||_2 = 1} ||t(β) - μ(β)||^2 \geq \inf_{β:||β||_0 = s} ||μ(β)||^2 - \sqrt{x} \right).$$

(25)

Now we first prove (i). Fix $s \in \mathbb{N}$ such that $s < ||β^*||_0$ (if $||β^*||_0 = 1$, the proof simplifies and one can consider (26) directly). Then, for all $β \in \mathbb{R}^d$ such that $||β||_0 = s$, we have by Theorem 3 that $\text{Cov}(I - X^Tβ) \neq 0$. Furthermore, using $\beta \mapsto ||\text{Cov}(I - X^Tβ)||_2^2 = ||π - Πβ||_2^2$ is a quadratic form, there exists $c > 0$ such that for all $β \in \mathbb{R}^d$ with $||β||_0 = s$ it holds that $||π - Πβ||_2 > c$. As $n$ tends to infinity, it holds that $\inf_{β:||β||_0 = s} ||μ(β)||_2$ diverges to infinity in probability, since

$$\inf_{β:||β||_0 = s} ||μ(β)||_2 \geq \inf_{β:||β||_0 = s} \sqrt{\frac{n-m}{m}}\|\Sigma(β)\|^{-1/2}c.$$ 

Therefore, by (25) it holds that

$$\lim_{n \to \infty} P(\varphi_s = 1) = \lim_{n \to \infty} P\left( \inf_{β:||β||_0 = s} T(β) > F_{n-m,m}(1-α) \right) \geq 1 - \lim_{n \to \infty} 2P\left( \sup_{β:||β||_2 = 1} ||t(β) - μ(β)||^2 \geq \inf_{β:||β||_0 = s} ||μ(β)||^2 - \sqrt{F_{n-m,m}(1-α)} \right) = 1.$$ 

Since this holds for any $s \in \mathbb{N}$ such that $s < ||β^*||_0$, we have

$$\lim_{n \to \infty} P(||\hat{β}||_0 \leq s_{\max} ||0 = ||β^*||_0) = \lim_{n \to \infty} P\left( \min_{s < ||β^*||_0} \text{φ}_s = 1, \text{φ}_{||β^*||_0} = 0 \right) \leq \lim_{n \to \infty} P(\text{φ}_{||β^*||_0} = 0) \geq 1 - α,$$ 

(26)

where the last statement follows from the fact that $φ_s$ has valid level.

Statement (ii) follows with the same argument noting that for all $ε > 0$ there exists a $c > 0$ such that for all $β \in \mathbb{R}^d$ satisfying $||β||_0 < ||β^*||_0$ or $||β||_0 = ||β^*||_0$ and $||β - β^*||_2 \geq ε$, we have $||π - Πβ||_2 > c > 0$, again, using Theorem 3. This concludes the proof of Theorem 6.

D PROOF OF PROPOSITION 7

Proof. To prove the first statement, we note that

$$\left\{ \bigcap_{S \subseteq \{PA[Y]\} \text{and} \, S \subseteq \{PA[Y]\} \text{accepted} \right\} \supseteq \{H_0(\{PA[Y]\}) \text{ accepted} \}.$$ 

But because

$$T(β^*) \geq T(\hat{β}_{\text{LIML}}(PA[Y])),$$ 

we have

$$P(H_0(\{PA[Y]\}) \text{ accepted } \geq 1 - α.$$ 

Accepted (UPDATED VERSION) for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).
To prove the second statement, observe that by the definition of $M$ it holds that
\[ \{ M \geq \| \beta^* \|_0 \} \supseteq \left\{ \min_{s < \| \beta^* \|_0} \phi_s = 1 \right\} \]
and therefore
\[ \left\{ \bigcap_{S \mid |S| = M} S \subseteq PA[Y] \right\} \supseteq \left\{ \min_{s < \| \beta^* \|_0} \phi_s = 1 \right\} \cap \{ T(\hat{\beta}_{LIML}(PA[Y])) \leq F_{n-m,m}^{-1}(1 - \alpha) \} \]

It follows from the first part of Theorem 3, that for all $\beta \in \mathbb{R}^d$ such that $\| \beta \|_0 < \| \beta^* \|_0$, we have $\text{Cov}(I, Y - X^T \beta) \neq 0$. We can therefore apply the same arguments as in Theorem 6 to argue that for all $s < \| \beta^* \|_0$, we have
\[ \lim_{n \to \infty} P(\phi_s = 1) = 1. \]
The statement then follows from $T(\beta^*) \geq T(\hat{\beta}_{LIML}(PA[Y]))$ and the fact that the Anderson-Rubin test holds level. This completes the proof of Proposition 7.

E  EXAMPLE 1 CONTINUED

Figure 7 discusses the example graph mentioned in Example 1.

F  EXAMPLE VIOLATING ASSUMPTION (A2)

Example 10. Consider an SCM of the following form
\[
\begin{bmatrix}
X^1 \\
X^2 \\
X^3
\end{bmatrix} :=
\begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
1 & 2 & 0
\end{bmatrix}
\begin{bmatrix}
X^1 \\
X^2 \\
X^3
\end{bmatrix} +
\begin{bmatrix}
4 & 0 \\
0 & 3 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
I^1 \\
I^2
\end{bmatrix} + h(H, \varepsilon^X)
\]
\[ Y := (X^1 \ X^2 \ X^3) \begin{bmatrix}
\frac{1}{2} \\
0
\end{bmatrix} + g(H, \varepsilon^Y), \]  
where $I^1$, $I^2$, $H$, $\varepsilon^Y$, $\varepsilon^X$ are jointly independent. Figure 8 shows the corresponding graphical representation. In this case, it holds that
\[ C = \begin{bmatrix}
1 & 0 & 1 \\
0 & 1 & 1
\end{bmatrix}. \]
Hence, the set $S = \{3\}$ violates Assumption (A2). In particular, the coefficient $\beta = (0, 0, 1)^T \in \mathcal{B}$ yields a sparser solution than the causal coefficient $(1, 1, 0)^T$. Therefore, the result of Theorem 3 cannot be valid. Assumption (A2) is violated in this example because the coefficients can be matched exactly. If the coefficients are chosen randomly with a distribution that is absolutely continuous with respect to Lebesgue measure, this happens with probability zero, see Proposition 9.

G  ADDITIONAL SIMULATION RESULTS
Figure 9: Same experiment as in Figure 4 but with TSLS estimator instead of LIML. Results for all random models that satisfy (A1)-(A3) (in total 1871 out of 2000 models). The median RMSE of the spaceIV estimator converges to zero as the sample size increases, which does not hold for OLS-sparse. Note that some of the outliers are cut-off in this plot.

Figure 10: Same experiment as in Figure 5 but with TSLS estimator instead of LIML. Expected fraction of random models for which spaceIV estimated the correct sparsity level. Only random models that satisfy (A1)-(A3) are considered (in total 1871 models). As the sample size increases the estimation of the sparsity level becomes more accurate.

Figure 11: Same experiment as in Figure 6 but with TSLS estimator instead of LIML. Results for all 2000 random models with \( n = 1600 \). We split the models into three cases depending on which of the assumptions (A1) and (A3) are satisfied (the group ‘(A1)’ contains 88 models, the group ‘(A1) & (A3)’ contains 1871 models and the group ‘none’ contains 41 models). If none of the assumptions are satisfied, not even the oracle with known parent set works. If only (A1) is satisfied, multiple sets of size 2 are able to satisfy the moment equation (3) and spaceIV may not estimate the correct set. These findings are in par with Theorem 3.

Accepted (UPDATED VERSION) for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).