Identifiability of Sparse Causal Effects using Instrumental Variables (Supplementary material)

Niklas Pfister[1, *] **Jonas Peters**[1, *]

[1]Department of Mathematical Sciences, University of Copenhagen, Denmark
[*]Authors contributed equally.

A PROOF OF PROPOSITION 2

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

$$\beta = M^\dagger v + w. \tag{20}$$

Therefore, the space $\mathcal{B}$ has a degenerate j-th coordinate if and only if $\mathrm{Null}(M)_j = \{0\}$. Denoting the Moore-Penrose inverse by $M^\dagger$, the null space of M can be expressed as

$$\mathrm{Null}(M) = \{(\mathrm{Id} - M^\dagger M)w \,|\, w \in \mathbb{R}^d\}.$$

Next, (1) and the assumption of joint independence of I, ξ^X, and ξ^Y imply that

$$\begin{aligned} M = \mathrm{Cov}[I, X] &= \mathrm{Cov}\left[I, (\mathrm{Id} - B)^{-1}(AI + \xi^X)\right] \\ &= \mathrm{Cov}[I]A^\top (\mathrm{Id} - B)^{-\top} \\ &= \mathrm{Cov}[I]C. \end{aligned}$$

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

$$M^\dagger M = C^\dagger \,\mathrm{Cov}[I]^{-1}\,\mathrm{Cov}[I]C. \tag{21}$$

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

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*

$$\mathrm{Rank}(B) \le \mathrm{Rank}(A) \quad \textit{and} \quad \mathrm{Im}(A) \neq \mathrm{Im}(B)$$

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

$$\mathbb{P}(AW \in \mathrm{Im}(B)) = 0.$$

Proof. We begin by showing that

$$\mathrm{Im}(B)^\perp \cap \mathrm{Im}(A) \neq \varnothing. \tag{22}$$

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

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

$$\mathrm{span}(b_1, \dots, b_k) = \mathrm{Im}(B)^\perp$$

and

$$\mathrm{span}(b_{k+1}, \dots, b_n) = \mathrm{Im}(B)\,.$$

Then, for every $\ell \in \{1, \dots, m\}$ there exits unique $\alpha_1^\ell, \dots, \alpha_n^\ell \in \mathbb{R}$ such that

$$A_\ell = \sum_{i=1}^n \alpha_i^\ell b_i.$$

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

$$Aw = \sum_{\ell=1}^m w^\ell A_\ell = \sum_{\ell=1}^m \sum_{i=1}^n w^\ell \alpha_i^\ell b_i = \sum_{i=1}^n \left(\sum_{\ell=1}^m w^\ell \alpha_i^\ell\right) b_i.$$

This implies that $Aw \in \mathrm{Im}(B)$ if and only if $\sum_{\ell=1}^m w^\ell \alpha_i^\ell = 0$ for all $i \in \{1, \dots, \ell\}$. Using this we get

$$\begin{aligned} \mathbb{P}(AW \in \mathrm{Im}(B)) &= \mathbb{P}(\forall i \in \{1, \dots, \ell\} : \textstyle\sum_{\ell=1}^m W^\ell \alpha_i^\ell = 0) \\ &\le \mathbb{P}(\textstyle\sum_{\ell \neq \ell^*} W^\ell \alpha_{i^*}^\ell = W^{\ell^*} \alpha_{i^*}^{\ell^*}) \\ &= 0, \end{aligned}$$

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

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. □

C PROOF OF THEOREM 6

Proof. It is known that for $\beta \in \mathbb{R}^d \setminus \mathcal{B}$ with $\beta \neq \beta^*$ the Anderson-Rubin test statistic (given Gaussian noise variables and conditioned on the observations of I and X) satisfies

$$T(\beta) \sim \chi^2\left(1, n\frac{\|\widehat{\mathrm{Cov}}(I,X)(\beta^*-\beta)\|_2^2}{\sigma^2}\right),$$

where $\chi^2(1,\lambda)$ is the non-central χ^2-distribution with one degree of freedom and non-centrality parameter λ, see for example Moreira [2009].

We first prove (i). Fix $s \in \mathbb{N}$ such that $s < \|\beta^*\|_0$ (if $\|\beta^*\|_0 = 1$, the proof simplifies and one can consider (24) directly). Then, for all $\beta \in \mathbb{R}^d$ such that $\|\beta\|_0 = s$, we have by Theorem 3 that $\mathrm{Cov}\left(I, Y - X^\top\beta\right) \neq 0$. Furthermore, there exists $\varepsilon > 0$ such that for all $\beta \in \mathbb{R}^d$ with $\|\beta\|_0 = s$ it holds that $\|\beta - \beta^*\|_2^2 > \varepsilon$. Therefore, since $\beta \mapsto \|\mathrm{Cov}\left(I, Y - X^\top\beta\right)\|_2^2 = \|\mathrm{Cov}\left(I, X^\top(\beta^* - \beta)\right)\|_2^2$ is a quadratic form, there exists $c > 0$ such that $\|\mathrm{Cov}\left(I, X^\top(\beta^* - \beta)\right)\|_2^2 > c$.

Conditioning on the observed data of X and I, we have

$$\begin{aligned} &P\left(\inf_{\beta\,:\,\|\beta\|_0 = s} T(\beta) > c_\alpha \,\middle|\, (X_1, I_1), \ldots, (X_n, I_n)\right) \\ &\quad = 1 - \kappa\left(c_\alpha, n \inf_{\beta\,:\,\|\beta\|_0 = s} \frac{\|\widehat{\mathrm{Cov}}(I,X)(\beta^*-\beta)\|_2^2}{\sigma^2}\right), \end{aligned} \tag{23}$$

where $\kappa(\cdot,\lambda)$ is the $\chi^2(1,\lambda)$-distribution function; here, we have exploited that for all x, $\lambda \mapsto \kappa(x,\lambda)$ is monotonically decreasing.

As n tends to infinity, it holds almost surely that $\|\widehat{\mathrm{Cov}}(I,X)(\beta^*-\beta)\|_2^2 \to \|\mathrm{Cov}(I,X)(\beta^*-\beta)\|_2^2 > c$. Hence, since c does not depend on β, the non-centrality parameter in the χ^2-distribution tends to infinity and (23) converges to 1. Thus,

$$\lim_{n\to\infty} P(\phi_s = 1) = 1.$$

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

$$\begin{aligned} &\lim_{n\to\infty} P(\|\hat{\beta}_{\leq s_{\max}}\|_0 = \|\beta^*\|_0) \\ &\quad = \lim_{n\to\infty} P\left(\min_{s<\|\beta^*\|_0} \phi_s = 1, \phi_{\|\beta^*\|_0} = 0\right) \\ &\quad = \lim_{n\to\infty} P(\phi_{\|\beta^*\|_0} = 0) \\ &\quad = 1 - \alpha, \end{aligned} \tag{24}$$

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 $\varepsilon > 0$ there exists a $c > 0$ such that for all $\beta \in \mathbb{R}^d$ satisfying $\|\beta\|_0 < \|\beta^*\|_0$ or $\|\beta\|_0 = \|\beta^*\|_0$ and $\|\beta - \beta^*\|_2 \geq \varepsilon$, we have $\mathrm{Cov}\left(I, Y - X^\top\beta\right) > c > 0$, 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

$$\begin{aligned} &\left\{\bigcap_{\substack{S:|S|=|\,\mathrm{PA}[Y]|\text{ and }\\ H_0(S)\text{ accepted}}} S \subseteq \mathrm{PA}[Y]\right\} \\ &\qquad \supseteq \{H_0(\mathrm{PA}[Y])\text{ accepted}\}. \end{aligned}$$

But because

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

we have

$$P\left(H_0(\mathrm{PA}[Y])\text{ accepted }\right) \geq 1 - \alpha.$$

To prove the second statement, observe that by the definition of M it holds that

$$\left\{M \geq \|\beta^*\|_0\right\} \supseteq \left\{\min_{s<\|\beta^*\|_0} \phi_s = 1\right\}$$

and therefore

$$\begin{aligned} &\left\{\bigcap_{\substack{S:|S|=M\text{ and }\\ H_0(S)\text{ accepted}}} S \subseteq \mathrm{PA}[Y]\right\} \\ &\supseteq \left\{\left\{\min_{s<\|\beta^*\|_0} \phi_s = 1\right\}\right. \\ &\qquad \left.\cap\, \{T(\hat{\beta}_{\mathrm{LIML}}(\mathrm{PA}[Y])) \leq F^{-1}_{n-m,m}(1-\alpha)\}\right\}. \end{aligned}$$

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 $\mathrm{Cov}\left(I, Y - X^\top\beta\right) \neq 0$. 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 $T(\beta^*) \geq T(\hat{\beta}_{\mathrm{LIML}}(\mathrm{PA}[Y]))$ 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.

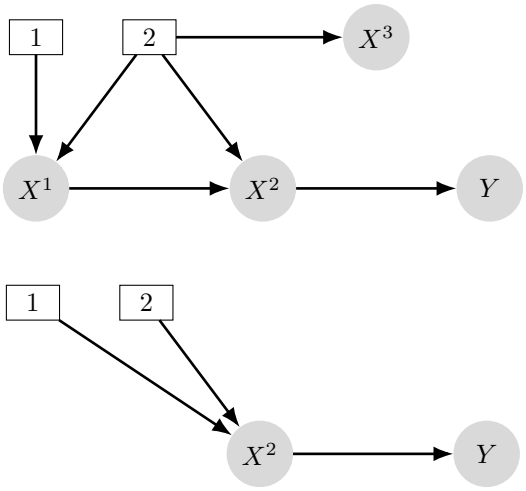

Figure 7: Top: Graph copied from Example 1 and Figure 2. Assumption (B1) holds because of the path $2 \to X^2$, for example. For $S = \{1\}$, (B3) (i) is not satisfied but (B3) (ii) holds: there is no set T of size one, such that all directed paths from I to $\mathrm{PA}(Y)$ go through T. Therefore, if (B2) holds, the effect β^* is identifiable (see Theorem 5). If, however, we were to remove the second instrument node from Example 1, (B3)(i) and (ii) would be violated (for set $S = \{X^1\}$). Bottom: Marginalized graph $\mathcal{G}^{\mathrm{PA}(Y)}$.

F EXAMPLE VIOLATING ASSUMPTION (A2)

Example 10. *Consider an SCM of the following form*

$$\begin{pmatrix} X^1 \\ X^2 \\ X^3 \end{pmatrix} := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} X^1 \\ X^2 \\ X^3 \end{pmatrix} + \begin{pmatrix} 4 & 0 \\ 0 & 3 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} I^1 \\ I^2 \end{pmatrix} + h(H, \varepsilon^X)$$

$$Y := \begin{pmatrix} X^1 & X^2 & X^3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} + g(H, \varepsilon^Y), \tag{25}$$

where I^1, I^2, H, ε^Y, ε^X are jointly independent. Figure 8 shows the corresponding graphical representation. In this case, it holds that

$$C = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Hence, the set $S = \{3\}$ violates Assumption (A2). In particular, the coefficient $\tilde{\beta} = (0, 0, 1)^\top \in \mathcal{B}$ yields a sparser solution than the causal coefficient $(1, 1, 0)^\top$. 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.

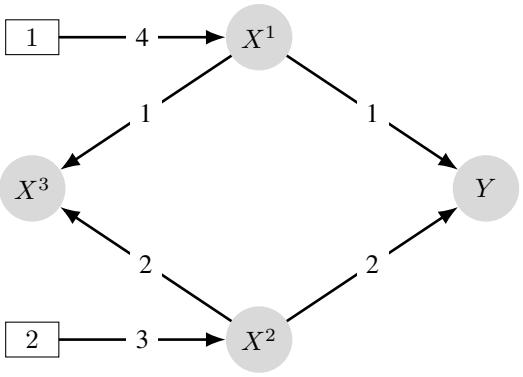

Figure 8: Example graph for which Assumption (A2) can be violated if the edge coefficients are fine-tuned to match each other exactly.

G ADDITIONAL SIMULATION RESULTS

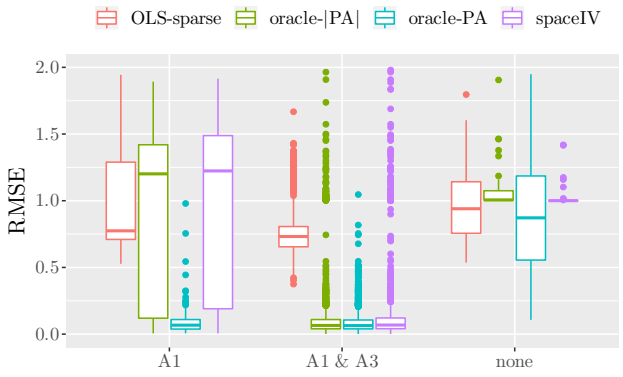

Figure 9: Same experiment as in Figure 6 but with TSLS estimator instead of LIML. Results for all 2000 random models with $n = 1600$. We split the models into three cases depending on which of the assumptions (A1) and (A3) are satisfied (the group '(A1)' contains 88 models, the group '(A1) & (A3)' contains 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.

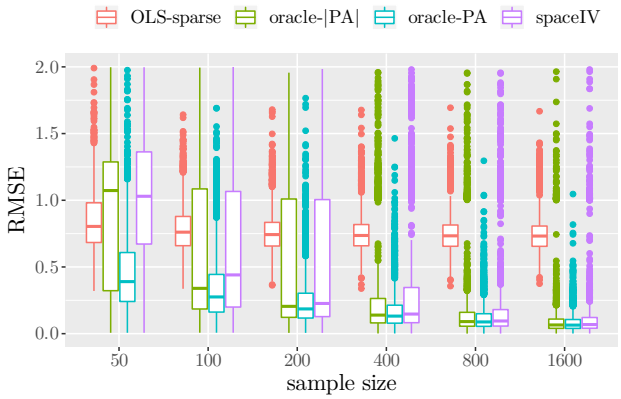

Figure 10: Same experiment as in Figure 4 but with TSLS estimator instead of LIML. Results for all random models that satisfy (A1)-(A3) (in total 1871 out of 2000 models). The median RSME of the `spaceIV` estimator converges to zero as the simple size increases, which does not hold for `OLS-sparse`. Note that some of the outliers are cut-off in this plot.

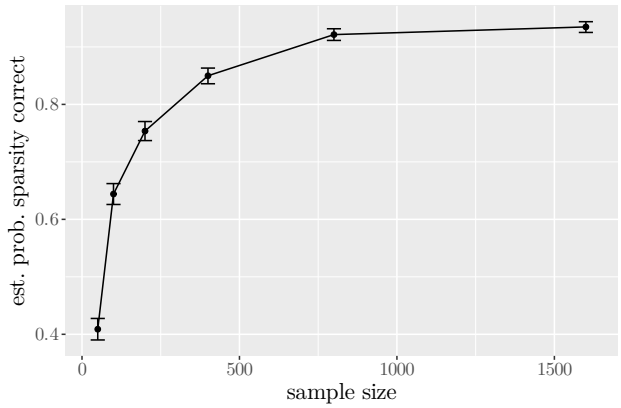

Figure 11: Same experiment as in Figure 5 but with TSLS estimator instead of LIML. Expected fraction of random models for which `spaceIV` estimated the correct sparsity level. Only random models that satisfy (A1)-(A3) are considered (in total 1871 models). As the sample size increases the estimation of the sparsity level becomes more accurate.

References

M. J. Moreira. Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics*, 152(2): 131–140, 2009.