Sequential Underspecified Instrument Selection for Cause-Effect Estimation

Elisabeth Ailer [1 2] **Jason Hartford** [3 4] **Niki Kilbertus** [1 2 5]

Abstract

Instrumental variable (IV) methods are used to estimate causal effects in settings with unobserved confounding, where we cannot directly experiment on the treatment variable. Instruments are variables which only affect the outcome indirectly via the treatment variable(s). Most IV applications focus on low-dimensional treatments and crucially require at least as many instruments as treatments. This assumption is restrictive: in the natural sciences we often seek to infer causal effects of high-dimensional treatments (e.g., the effect of gene expressions or microbiota on health and disease), but can only run few experiments with a limited number of instruments (e.g., drugs or antibiotics). In such underspecified problems, the full treatment effect is not identifiable in a single experiment even in the linear case. We show that one can still reliably recover the projection of the treatment effect onto the instrumented subspace and develop techniques to consistently combine such partial estimates from different sets of instruments. We then leverage our combined estimators in an algorithm that iteratively proposes the most informative instruments at each round of experimentation to maximize the overall information about the full causal effect.

1. Introduction

Motivation. Understanding cause-effect relationships in high-dimensional systems is a common challenge in various scientific areas. For example, how does our gut microbiome (treatment X) causally influence health and disease (outcome Y)? How does the transcriptome of a cell (treatment X) causally influence its function (outcome Y)? Typically, the high-dimensional treatment and the outcome are heavily confounded via unknown mechanisms, rendering the strong assumptions required for cause-effect estimation, i.e., computing $p(y \mid do(x))$, from observational data sampled from $p(x, y)$ untenable. Hence, experimentation is indispensable for the ultimate goal of identifying and estimating these causal relationships. Whenever we can intervene directly on the treatment, we have direct access to $p(y \mid do(x))$ and cause-effect relationships can be captured by mere association in experiments. We are motivated by two crucial realizations: (a) oftentimes practically feasible experiments do not intervene directly on the treatment X but some other variable Z; (b) still, the scientific goal is to estimate the effect of the high-dimensional X on the outcome Y (instead of the effect of the actual intervention on Z).

Regarding (a), administering certain antibiotics has a strong and highly predictable effect on the gut microbiome, but it does not break causal links from potentially unobserved confounders to X; similarly, applying various drug (dosages) to cell cultures influences the transcriptome, but again does not directly intervene on it. Even targeted gene knockouts or CRISPR/Cas gene editing (Zhang et al., 2015; Fu et al., 2013) do not constitute interventions as defined in causal inference (Pearl, 2009). They do not strip the microbiome or transcriptome free of any other causal influences, i.e., they may still be confounded with the outcome of interest. As for (b), ultimately we give a certain antibiotic to learn about *how the microbiome causally influences disease* ($p(y \mid do(x))$) and not merely about what overall effect the antibiotic has on disease ($p(y \mid do(z))$).

In such settings, the variable experimented on (antibiotics, drugs, gene knockout/edits) can at most serve as an *instrument* Z for the treatment: **(A1)** Z (strongly) affects the treatment ($Z \not\perp\!\!\!\perp X$). **(A2)** Z is independent of unobserved confounders U ($Z \perp\!\!\!\perp U$). **(A3)** Z "only affects the outcome via the treatment" ($Y \perp\!\!\!\perp Z \mid \{X, U\}$). Whether we can ascertain these conditions depends on the setting. For example, orally administered sub therapeutic dosages of antibiotics (such that no antibiotics can be detected in the bloodstream) may satisfy this condition (Ailer et al., 2021). Similarly,

[1]HelmholtzAI, Helmholtz Munich, Munich, Germany [2]School of Computation, Information and Technology, Technical University Munich, Munich, Germany [3]MILA - Quebec Artificial Intelligence Institute, Montreal, Quebec, Canada [4]Recursion, Montreal Quebec Canada [5]Munich Center for Machine Learning, Munich, Germany. Correspondence to: Elisabeth Ailer <elisabeth.ailer@helmholtz-munich.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

genetic interventions (or drugs) may indeed only influence cell function via the transcriptome.

As a result, even with access to experimentation, we must resort to instrumental variable (IV) techniques to estimate the treatment effect. A number of challenges arise in this setting even in the fully linear case: **(B1)** At least as many instruments as treatment variables are required for the causal effect to be identified (just- or overspecified setting). This is not feasible in our scenarios with (tens of) thousands of organisms/genes and much fewer antibiotics/drugs to experiment with. **(B2)** Experimentation is typically slow and expensive, putting natural bounds on how many experiments can be run. **(B3)** The cost of an individual experiment depends on the number of randomized instruments and there may be limits to how many instruments can be sensibly combined in one single experiment. For example, too many drugs, gene knockouts, or antibiotics at once may kill or permanently damage the studied organism.

Main goal and contributions. Constrained by (B1)-(B3), we formulate our main goal: *Can we select a bounded number of instruments in a bounded number of sequential experiments and combine the results for an informative estimate of a high-dimensional causal effect?*[1] In answering this question, we make the following contributions:

- We show that in the underspecified linear IV setting, we can estimate the orthogonal projection of the causal effect onto the "instrumented subspace" and construct a $\sqrt{n}$-consistent, asymptotically normal estimator.
- We combine such estimates from experiments with different instruments to consistently recover the estimate had we randomized all IVs simultaneously (without actually having to apply all perturbations at once).
- We develop an algorithm that sequentially proposes subsets of the available instruments to maximally identify the treatment effect from the combined estimate across all experimental rounds. The algorithm trades off the information gained from multiple instruments with the increasing cost of including them in a single experiment based on a pre-specified similarity between instruments.
- We develop techniques to keep track of which components of β have been identified reliably at each round, upper bound the absolute error in unidentified components, and propose a stopping criterion based purely on observational data ($p(x, y)$) that—when reached—guarantees full identification of β under mild assumptions.

Related work. We build on the theory of (linear) instrumental variable estimators, with a specific focus on two-stage methods (Angrist & Pischke, 2008). Instrumental variables have been used since 1928 by Philip G. Wright (Wright, 1928; Stock & Trebbi, 2003) and are an essential part of the econometrics toolkit (Angrist & Pischke, 2008). They have also received renewed interest from the machine learning community recently (Hartford et al., 2017; Zhang et al., 2020; Kilbertus et al., 2020; Singh et al., 2019; Bennett et al., 2019; Padh et al., 2022; Saengkyongam et al., 2022; Muandet et al., 2019). Despite the difficulty of finding valid instruments for a given target effect in practice (Hernán & Robins, 2006), instrumental variable estimation has been applied successfully in genetics via Mendelian randomization (Sanderson et al., 2022; Didelez & Sheehan, 2007) and recently on microbiome data (Sohn & Li, 2019; Wang et al., 2020; Ailer et al., 2021).

Our work is also related to ideas in experiment design for causal structure learning (Hyttinen et al., 2013; Gamella & Heinze-Deml, 2020; Sussex et al., 2021; Tigas et al., 2022). Two key differences are that those works focus on sequential selection of *interventions* (not just instruments) and that they seek to identify causal structure (i.e., the causal graph) instead of a specific causal effect. To the best of our knowledge, there is no literature on adaptively selecting instruments in sequential experiments to identify a causal effect from high-dimensional treatments.

Finally, variants of underspecified instrumental variable settings have been studied by Pfister & Peters (2022). They assume the causal effect from X on Y to be sparse, which allows them to relax standard identifiability assumptions in the linear IV setting. Rothenhäusler et al. (2018) make use of exogenous variables to provide an estimator that interpolates between the ordinary least-squares (OLS) and two stage least-squares (2SLS) estimates, even when IV assumptions are not fully satisfied and point-identification is not guaranteed. Their proposal can be interpreted as "choosing the best performing (in terms of mean squared error) estimate among the compatible ones". Thus, both works pursue goals orthogonal to ours.

2. Background and Problem Setting

We aim to estimate the causal effect of treatments $X \in \mathbb{R}^{d_x}$ on a scalar outcome $Y \in \mathbb{R}$. There may be unobserved confounding between X and Y, but we assume access to valid instruments $Z \in \mathbb{R}^{d_z}$. We focus on the linear setting

$$X = Z\alpha + \epsilon_X \,, \qquad Y = X\beta + \epsilon_Y \,, \tag{1}$$

where $\alpha \in \mathbb{R}^{d_z \times d_x}$ and $\beta \in \mathbb{R}^{d_x}$ represent the linear structural functions. We interpret β flexibly as a row or column vector as needed. Because of unobserved confounding, the noise variables ϵ_X, ϵ_Y are typically not independent, but we assume $\mathbb{E}[\epsilon_X] = \mathbb{E}[\epsilon_Y] = 0$. The standard instrumental variable assumptions (A1)-(A3) become $\alpha \neq 0$, $Z \perp\!\!\!\perp \{\epsilon_X, \epsilon_Y\}$, and $Z \perp\!\!\!\perp Y \mid \{X, \epsilon_Y\}$.

[1] We highlight that while we call the high-dimensional X the "treatment", we experiment (or intervene) on lower-dimensional instruments Z.

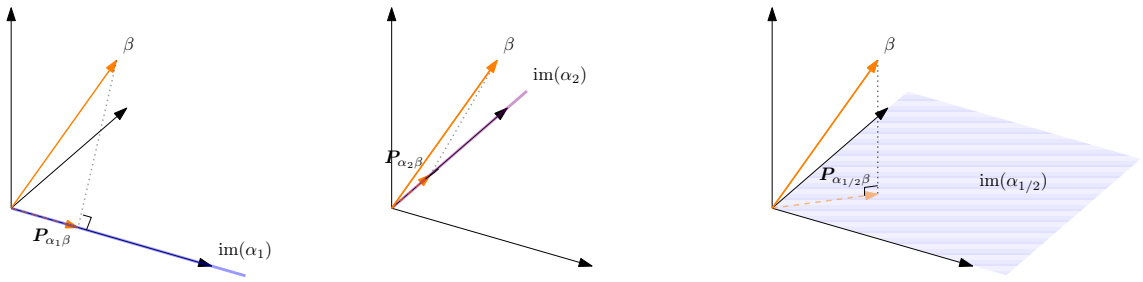

Figure 1. The illustration shows the possible experimental settings, including their estimates in a setting with $d_z = 2, d_x = 3$: (1) The experiments with one of the two instruments respectively, denoted by α_1 and α_2 and (2) - if feasible - with both instruments at once, i.e. $\alpha_{1/2}$. In each figure, the true causal effect β (orange) is projected onto the corresponding instrumented spaces, i.e. $\text{im}(\alpha_1), \text{im}(\alpha_2)$ and $\text{im}(\alpha_{1/2})$ (shaded blue) by $\boldsymbol{P}_{\alpha_1}\beta, \boldsymbol{P}_{\alpha_2}\beta$ and $\boldsymbol{P}_{\alpha_{1/2}}\beta$ (dashed orange).

Therefore, estimating the causal effect simply corresponds to estimating β. The OLS estimator for β will be biased, but if $d_x \leq d_z$ and a rank condition on the covariance of Z and X is satisfied, β is point-identified and can be estimated consistently, for example via the standard 2SLS estimator (Angrist & Pischke, 2008). When collecting n i.i.d. observations in matrices $\boldsymbol{X} \in \mathbb{R}^{n\times d_x}$, $\boldsymbol{Z} \in \mathbb{R}^{n\times d_z}$, and $\boldsymbol{y} \in \mathbb{R}^n$, the 2SLS estimator for β is given by

$$\hat{\beta}_{\text{2SLS}} = (\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y}\,, \tag{2}$$

where $\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$ is a projection matrix. The 2SLS estimator can be viewed as (a) regressing X on Z, and (b) regressing Y on the predicted X values from the first-stage regression. Intuitively, the influence of ϵ_X on X is "regressed out" in the first stage, leaving only the direct causal effect of X on Y in the second stage.

Following traditional application settings, most proposed IV methods assume the *just-identified* setting $d_x = d_z$ (as well as the rank condition on $\text{cov}(X, Z)$, which translates into a full rank condition on α). Typically, these methods can also accommodate the *overidentified* (or *overspecified*) setting $d_z > d_x$. Having many instruments available in the overspecified setting can also be exploited for consistent estimators under relaxed assumptions such as correlated or weak instruments (Hahn & Hausman, 2005; Kang et al., 2016). In contrast, the *underidentified* (or *underspecified*) case $d_z < d_x$, where there are fewer instruments than treatments, has received little attention in the literature. In this case, since $\text{rk}(\boldsymbol{P}_{\boldsymbol{Z}}) \leq d_z$, we have that $\text{rk}(\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}) \leq d_z < d_x$ and consequently cannot take the inverse in Equation (2). More generally, in this situation the causal effect β is not fully identified. However, the available instruments still constrain the possible values of β. We will later estimate and exploit those confines.

In our setting, we assume access to a fixed number of $N_{\text{IV}} \in \mathbb{N}$ instruments (e.g., available antibiotics, or drugs) and—in high-dimensional treatment settings—typically have $N_{\text{IV}} < d_x$. Because we will also consider subsets of instruments, we generically denote by $d_z \in [N_{\text{IV}}] := \{1, \ldots, N_{\text{IV}}\}$ the number of IVs in a given estimation (or experiment). Further, we assume experimental access to the instruments in that we can run experiments in which a chosen set of instruments is randomized (e.g., in mouse studies or on cell cultures). Since the specific distribution of Z does not affect identifiability, we assume without loss of generality that the components of Z all independently follow a Rademacher distribution. This can be interpreted as whether a drug or antibiotic is applied.[2] With this choice, α fully characterizes the effect of Z on X and we consequently also call the rows of $\alpha \in \mathbb{R}^{N_{\text{IV}}\times d_x}$ "the instruments".[3] Since the components of Z are jointly independent, the full rank condition on $\text{cov}(X, Z)$ translates to α having full rank.

To maximally constrain the effect estimate, one would randomize all N_{IV} available instruments simultaneously. Due to (B3), this is typically infeasible in practice. We model this via a cost function, which can also incorporate hard constraints such as limiting the number of instruments per experiment to at most $N_{\text{IV}/\text{exp}} < N_{\text{IV}}$. For (B2) we limit the total number of possible experiments to $T \in \mathbb{N}$.

We proceed as follows. In Section 3, we first construct a consistent, asymptotically normal estimator for the orthogonal projection of β onto the image of α viewed as a linear map $\alpha : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$. We then establish a method to combine multiple such estimates obtained from (different) subsets of IVs to obtain a consistent estimate for the orthogonal projection of β on the linear subspace spanned by all instruments combined. Furthermore, we introduce a method to determine which components of β have been successfully identified and a condition that guarantees full identification

[2] Other illustrative choices are $\frac{1}{2}$-Bernoulli or the uniform distribution on a finite interval representing (standardized) dosage levels. All our results immediately transfer.

[3] Generically, $\alpha \in \mathbb{R}^{d_z\times d_x}$ with $d_z \leq N_{\text{IV}}$ represents some choice of d_z instruments in a given estimation/experiment.

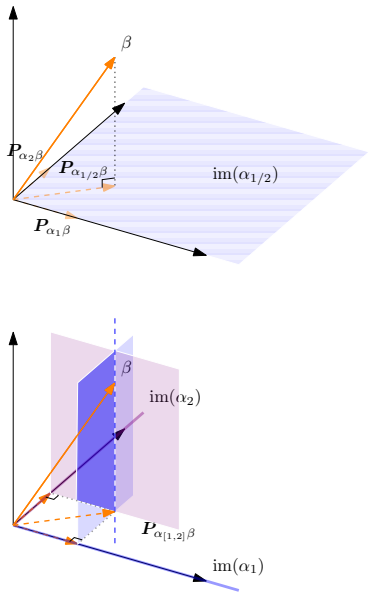

Figure 2. The illustration shows a setting with $d_z = 2, d_x = 3$. It contrasts the single estimation steps with the combination step. Upper panel, single estimation: the true causal effect β (orange) is projected onto the individual instrumented spaces by $\boldsymbol{P}_{\alpha_1}\beta, \boldsymbol{P}_{\alpha_2}\beta$ and $\boldsymbol{P}_{\alpha_{1/2}}\beta$. Lower panel, combination step: The intersection (dashed blue line) of the orthogonal complements of $\{\mathrm{im}(\alpha_1), \mathrm{im}(\alpha_2)\}$ represents all vectors that are compatible with both individual estimates. Equation (4) then selects the minimum norm point on that line. In the illustration it comes to show that the combination of individual estimators $\boldsymbol{P}_{\alpha_{[1,2]}}\beta$ recovers the effect of the single estimation with both instruments $\boldsymbol{P}_{\alpha_{1/2}}\beta$.

of β under mild assumptions. In Section 4, we then leverage all these findings to develop a procedure that proposes subsets of instruments for sequential experimentation to maximally identify β under the given constraints.

Figure 1 and Figure 2 illustrate the estimation resp. the combination step and the role of linearity in this context visually. We suggest to consult these illustrations for intuition when reading the formal statements in the following section.

3. Underidentified IV Estimates

3.1. Estimator for a Single Experiment

Consider a set of instruments $\alpha \in \mathbb{R}^{d_z \times d_x}$ with $d_z < d_x$ and $\mathrm{rk}(\alpha) = d_z$. The predicted treatment values via OLS $\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X} \in \mathbb{R}^{n \times d_x}$ are confined to a proper linear subspace of $\mathbb{R}^{d_x}$, rendering $\boldsymbol{X}^T P_{\boldsymbol{Z}}\boldsymbol{X}$ singular and $\hat{\beta}_{\mathrm{2SLS}}$ in Equation (2) ill-defined in the underspecified setting. With $\hat{\alpha}$ being the first-stage OLS estimate, we have $\hat{\boldsymbol{X}}_i = \boldsymbol{P}_{\hat{\alpha}}\boldsymbol{X}_i \in \mathrm{im}(\hat{\alpha}) := \{\hat{\alpha}z \,|\, z \in \mathbb{R}^{d_z}\} \subset \mathbb{R}^{d_x}$ for all $i \in [n]$, where $\boldsymbol{P}_{\hat{\alpha}} = \hat{\alpha}^T(\hat{\alpha}\hat{\alpha}^T)^{-1}\hat{\alpha} \in \mathbb{R}^{d_x \times d_x}$ denotes the orthogonal projection onto $\mathrm{im}(\hat{\alpha})$. That is, all first-stage predictions lie in the image of $\hat{\alpha}$ and we can consider the second stage as a mapping $\mathbb{R}^{d_x} \supset \mathrm{im}(\hat{\alpha}) \to \mathbb{R}$. We call $\mathrm{im}(\hat{\alpha})$ the *instrumented subspace* (in this experiment). Intuitively, while the low-dimensional Z cannot "shake" or "instrument" the entire treatment space to identify β, it still "instruments" a non-trivial subspace, inducing non-trivial constraints on β. In particular, one may expect to recover "the part of β within the instrumented subspace". The following statement formalizes this intuition.

Proposition 1. *Let $\alpha \in \mathbb{R}^{d_z \times d_x}$ have full rank and assume we have i.i.d. data $\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}$ from an experiment in which all d_z instruments have been randomized. Then*

$$\widehat{\boldsymbol{P}_\alpha\beta} := (\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y} \xrightarrow{d} \mathcal{N}(\boldsymbol{P}_\alpha\beta, \Sigma) \qquad (3)$$

with $\Sigma := \frac{1}{n}\alpha^+\Sigma_Z^{-1}(\alpha^T)^+ \mathrm{Var}[\epsilon_Y] = \frac{1}{n}(\alpha^T\alpha)^+ \mathrm{Var}[\epsilon_Y]$,

where $(\cdot)^+$ denotes the Moore-Penrose pseudoinverse, Σ_Z is the covariance matrix of Z, and the simple form of Σ (in gray) applies when Z are i.i.d. Rademacher variables.

Proof. For $d_z \geq d_x$, the pseudoinverse in Equation (3) can be replaced with a regular matrix inverse and the result follows from the asymptotic normality of $\hat{\beta}_{\mathrm{2SLS}}$ in Equation (2) (Angrist & Pischke, 2008, Sec. 4.2.1). For the underspecified case $d_z < d_x$, we start with the singular value decomposition (SVD) $\hat{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, with which $\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} = \boldsymbol{V}\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{V}^T$. Assuming that the singular values in the rectangular diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{n \times d_x}$ are sorted in non-ascending order, only the first d_z entries are nonzero, because $\mathrm{rk}(\hat{\boldsymbol{X}}) = d_z$. Accordingly, the first d_z rows of $\boldsymbol{V}$ form an orthonormal basis of $\mathrm{im}(\alpha)$. We write $\boldsymbol{V}_{\hat{\alpha}} \in \mathbb{R}^{d_x \times d_z}$ for the first d_z columns of $\boldsymbol{V}$, $\boldsymbol{D}_{\hat{\alpha}} \in \mathbb{R}^{d_z \times d_z}$ for the upper left $d_z \times d_z$ block of $\boldsymbol{D}$, and $\boldsymbol{U}_{\hat{\alpha}} \in \mathbb{R}^{n \times d_z}$ for the first d_z columns of $\boldsymbol{U}$. With $\hat{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$ we compute

$$\begin{aligned}
\widehat{\boldsymbol{P}_\alpha\beta} &= (\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{Z}}^T\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{Z}}^T\boldsymbol{P}_{\boldsymbol{Z}}(\boldsymbol{X}\beta + \epsilon_Y) \\
&= (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^+(\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}}\beta + \hat{\boldsymbol{X}}^T\epsilon_Y) \\
&= \boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D})^+\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{V}^T\beta + \boldsymbol{V}(\boldsymbol{D}^T)^+\boldsymbol{U}^T\epsilon_Y \\
&= \boldsymbol{V}_{\hat{\alpha}}\boldsymbol{V}_{\hat{\alpha}}^T\beta + \boldsymbol{V}_{\hat{\alpha}}\boldsymbol{D}_{\hat{\alpha}}^{-1}\boldsymbol{U}_{\hat{\alpha}}^T\epsilon_Y \,,
\end{aligned}$$

where we have used

$$(\boldsymbol{D}^T\boldsymbol{D})^+\boldsymbol{D}^T\boldsymbol{D} = \begin{pmatrix} \boldsymbol{I}_{d_z \times d_z} & \boldsymbol{0}_{d_z \times (d_x - d_z)} \\ \boldsymbol{0}_{(d_x - d_z) \times d_z} & \boldsymbol{0}_{(d_x - d_z) \times (d_x - d_z)} \end{pmatrix} .$$

From the asymptotic normality of the first-stage OLS estimate $\hat{\alpha}$, it follows that $\boldsymbol{V}_{\hat{\alpha}}$ and $\boldsymbol{D}_{\hat{\alpha}}$ are $\sqrt{n}$-consistent estimators for $\boldsymbol{V}_\alpha$ and $\boldsymbol{D}_\alpha$ respectively (Bura & Pfeiffer, 2008). Therefore, $\boldsymbol{P}_{\hat{\alpha}} = \boldsymbol{V}_{\hat{\alpha}}\boldsymbol{V}_{\hat{\alpha}}^T$ converges in probability to $\boldsymbol{P}_\alpha$ in the large sample limit. With $\mathbb{E}[\epsilon_Y] = 0$ we thus have

$$\mathbb{E}[\widehat{\boldsymbol{P}_\alpha\beta}] = \mathbb{E}[\boldsymbol{V}_{\hat{\alpha}}\boldsymbol{V}_{\hat{\alpha}}^T]\beta \xrightarrow{p} \boldsymbol{P}_\alpha\beta \,.$$

Similarly, with $\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\alpha}$ we compute the covariance

$$\begin{aligned}
\mathrm{Cov}[\widehat{\boldsymbol{P}_\alpha\beta}] &= \mathrm{Cov}[\boldsymbol{V}_{\hat{\alpha}}\boldsymbol{D}_{\hat{\alpha}}^{-1}\boldsymbol{U}_{\hat{\alpha}}^T\epsilon_Y] = \hat{\boldsymbol{X}}^+(\hat{\boldsymbol{X}}^T)^+ \mathrm{Var}[\epsilon_Y] \\
&= \hat{\alpha}^+(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}(\hat{\alpha}^T)^+ \mathrm{Var}[\epsilon_Y] \,.
\end{aligned}$$

Since the matrix inverse is continuous (on invertible matrices), the continuous mapping theorem yields that $\hat{\Sigma}_Z^{-1}$ consistently estimates Σ_Z^{-1}. Again using the asymptotic normality of $\hat{\alpha}$, we thus have (for centered) $\mathrm{Cov}[\widehat{\boldsymbol{P}_\alpha \beta}] \xrightarrow{p} \frac{1}{n}\alpha^+ \Sigma_Z^{-1} (\alpha^T)^+ \mathrm{Var}[\epsilon_Y]$. For i.i.d. Rademacher instruments with covariance matrix $\Sigma_Z = \boldsymbol{I}_{d_z \times d_z}$ we have $\mathrm{Cov}[\widehat{\boldsymbol{P}_\alpha \beta}] \xrightarrow{p} \frac{1}{n}(\alpha^T \alpha)^+$. □

Proposition 1 substantiates the first intuition that in the underidentified setting we can still consistently estimate the orthogonal projection of β onto the instrumented subspace.

We refer to Figure 1 for an illustration of the estimation in a linear setting with $d_z = 2, d_x = 3, T = 2$. For visualization purposes, we choose axis-aligned $\alpha_1 = (1, 0, 0)$ and $\alpha_2 = (0, 1, 0)$. For both instruments, we estimate the projection of β onto the respective instrumented spaces ($P_{\alpha_1}\beta$ and $P_{\alpha_2}\beta$). The illustration on the right hand side also includes the estimation which uses both instruments ($\alpha_{1/2}$) at once. In the following, this is the effect we want to recover solely based on the individual estimates.

3.2. Combined Estimator

In sequential experiments, we select pairwise disjoint instruments $\alpha_1, \alpha_2, \ldots, \alpha_T$, where each α_i is a subset of the N_{IV} available ones. Each corresponding individual estimator $\widehat{\boldsymbol{P}_{\alpha_1}\beta}, \widehat{\boldsymbol{P}_{\alpha_2}\beta}, \ldots, \widehat{\boldsymbol{P}_{\alpha_T}\beta}$ estimates an orthogonal projection of β onto the linear subspace $\mathrm{im}(\hat{\alpha}_i)$. Our goal is to reconstruct the estimate we would have obtained, had we used all instruments in the sequential experiments at once in a single one. Denoting the union (or concatenation) of all available instruments by a single matrix $\alpha_{[T]}$, we aim to reconstruct $\boldsymbol{P}_{\alpha_{[T]}}\beta$ from the individual $\widehat{\boldsymbol{P}_{\alpha_i}\beta}$.[4] In other words, we are looking for the least-square vector compatible with all projections

$$\min_{\gamma \in \mathbb{R}^{d_x}} \|\gamma\|_2^2 \text{ s.t. } \widehat{\boldsymbol{P}_{\alpha_i}\beta} = \boldsymbol{V}_{\hat{\alpha}_i} \boldsymbol{V}_{\hat{\alpha}_i}^T \gamma \text{ for all } i \in [T]\,. \quad (4)$$

Proposition 2. *Let $\alpha \in \mathbb{R}^{N_{\mathrm{IV}} \times d_x}$ have full rank and let $(\boldsymbol{Z}_1, \boldsymbol{X}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{Z}_T, \boldsymbol{X}_T, \boldsymbol{y}_T)$ be i.i.d. datasets from T experiments with disjoint subsets $\alpha_1, \ldots, \alpha_T$ of randomized instruments. Then the solution of Equation (4) is a consistent estimator for $\boldsymbol{P}_{\alpha_{[T]}}\beta$.*

Proof. Let $\boldsymbol{A} := \boldsymbol{V}_{\hat{\alpha}_{[T]}} \boldsymbol{V}_{\hat{\alpha}_{[T]}}^T = (\boldsymbol{V}_{\hat{\alpha}_1} \boldsymbol{V}_{\hat{\alpha}_1}^T | \ldots | \boldsymbol{V}_{\hat{\alpha}_T} \boldsymbol{V}_{\hat{\alpha}_T}^T) \in \mathbb{R}^{(T d_x) \times d_x}$ and $\boldsymbol{b} := (\widehat{\boldsymbol{P}_{\alpha_1}\beta}, \ldots, \widehat{\boldsymbol{P}_{\alpha_T}\beta}) \in \mathbb{R}^{T d_x}$, where "|" denotes concatenation. Then the solution to Equation (4) is simply $\boldsymbol{A}^+ \boldsymbol{b}$, i.e., the least-squares solution to $\boldsymbol{A}\gamma = \boldsymbol{b}$. Hence the optimal γ is the orthogonal projection of $\boldsymbol{b}$ onto $\mathrm{im}(\boldsymbol{A})$, which by construction is just $\mathrm{im}(\hat{\alpha}_{[T]})$. By the consistency of $\hat{\alpha}_i$ (and thus the consistency of $\boldsymbol{A}$ and $\boldsymbol{b}$ for their respective expressions without hats as in the proof of Proposition 1), asymptotically $\boldsymbol{A}^+ \boldsymbol{b} \xrightarrow{p} \boldsymbol{P}_{\alpha_{[T]}}\beta$. □

[4] We use similar notation $\hat{\alpha}_{[t]}$ for the concatenation of the instruments (as matrices) used up until and including round t.

Note that the instrument sets need not be disjoint for the proof of Proposition 2. However, since we do not gain further information by including the same instrument in two distinct experiments, which would instrument the same subspace twice, it is reasonable to keep instrument sets distinct across experiments for efficiency. The least-squares problem with linear equality constraints in Equation (4) can be solved efficiently (in high dimensions and for many constraints) by simply solving a linear system (Nocedal & Wright, 2006, Sec. 16.1). In practice, we keep a running estimate $\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta}$, which is the combined estimate for all experiments up to and including round $t \in [T]$. Being able to combine estimates from individual sets of instruments "as if we had run a single experiment using the union of instruments at once" forms the basis for leveraging sequential experiments with different sets of randomized instruments to optimally constrain the overall causal effect in Section 4.

3.3. Full Identification and Identified Components

Full identification. As an interesting special case of Proposition 1, we note that regardless of d_x, a single instrument $d_z = 1$ in principle suffices to identify $\beta \in \mathbb{R}^{d_x}$, namely when $\alpha \in \mathbb{R}^{d_x \times 1}$ is parallel to β (viewed as vectors in $\mathbb{R}^{d_x}$), which implies $\boldsymbol{P}_\alpha \beta = \beta$. Of course, this is unlikely to "happen accidentally" in practice. However, it highlights that if we had full control over the instruments, we could fully identify β by crafting $\alpha \in \mathbb{R}^{d_x \times 1}$ such as to maximize the 2-norm of the resulting estimated $\widehat{\boldsymbol{P}_\alpha \beta}$. This can be seen from the fact that $\|\boldsymbol{P}_\alpha \beta\|_2 \leq \|\beta\|_2$ for any α with equality implying $\beta \in \mathrm{im}(\alpha)$. Therefore, when running sequential experiments, the 2-norm of the combined estimate is a proxy for whether we are still gaining relevant information. We summarize these results in the following Corollary.

Corollary 3. *In the setting of Proposition 1, the following are equivalent: (i) $\beta \in \mathrm{im}(\alpha)$; (ii) $\|\widehat{\boldsymbol{P}_\alpha \beta}\|_2 \longrightarrow \|\beta\|_2$ as $n \to \infty$; (iii) $\widehat{\boldsymbol{P}_\alpha \beta}$ consistently estimates β (not just $\boldsymbol{P}_\alpha \beta$).*

In our underspecified setting, it is generally impossible to determine whether any (and therefore all) of the statements in Corollary 3 hold true. However, recently Janzing & Schölkopf (2018) have shown that under mild assumptions on the confounding model (i.e., on the joint distribution of ϵ_X, ϵ_Y), one can estimate what they call the *confounding strength*. Their estimator has been further refined by Rendsburg et al. (2022). The confounding strength can then be leveraged to estimate $\|\beta\|_2$ itself from purely observational data (Janzing, 2019). This means that we can consistently estimate $\|\beta\|_2$ from data $(\boldsymbol{X}, \boldsymbol{y})$, where no instruments have been randomized in our setting (Janzing, 2019, step 5 in

Alg. ConCorr).[5] Let us denote this estimator by $\widehat{\|\beta\|}_2$. We outline the assumptions on the confounding model, which are satisfied in our empirical evaluation, in Section 5.

Following Corollary 3 we can use the difference of the 2-norm of our running estimate $\|\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta}\|_2$ and $\widehat{\|\beta\|}_2$ as a heuristic for whether we have already fully identified β, i.e., whether the overall instrumented subspace contains β.[6] In future work, with the asymptotic normality of $\widehat{\boldsymbol{P}_\alpha\beta}$ in Proposition 1 and thus an explicit expression for the asymptotic distribution of $\|\widehat{\boldsymbol{P}_\alpha\beta}\|_2^2$ (Moschopoulos, 1985) as well as the asymptotic behavior of $\widehat{\|\beta\|}_2$ in Rendsburg et al. (2022), one could derive (asymptotic) confidence intervals for whether $||\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta}|| = \widehat{\|\beta\|}_2$. In practice, we propose

$$\left|\widehat{\|\beta\|}_2 - \|\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta}\|_2\right| < \epsilon \tag{5}$$

for a fixed tolerance $\epsilon > 0$ as a stopping criterion of our algorithm. When this condition is reached, we can conclude (for sufficiently large n) that β is near fully identified, even though we may have used fewer than d_x instruments.

Identified components. When the above stopping criterion is not reached within our T rounds of experimentation, learning an orthogonal projection of β may still be informative in itself. However, in many situations one is interested in precise values of individual components β_i for $i \in [d_x]$. Notably, when $\beta \notin \mathrm{im}(\alpha)$, the individual components of $\boldsymbol{P}_\alpha\beta$ are not easily interpretable. For example, neither holds $(\boldsymbol{P}_\alpha\beta)_i = 0 \Rightarrow \beta_i = 0$ nor the other way round. Therefore, we now devise a method to determine when we can trust any given component of our running estimate $\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta}$.

Corollary 4. *Let $(e_i)_{i\in[d_x]}$ be the standard basis of $\mathbb{R}^{d_x}$. In the setting of Proposition 2, when $\boldsymbol{V}_{\alpha_{[t]}}\boldsymbol{V}_{\alpha_{[t]}}^T e_i = e_i$, then $(\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta})_i$ consistently estimates β_i.*

Proof. This follows from Proposition 2 and the linearity of projections when considering $\beta = \sum_{i=1}^{d_x} \beta_i e_i$. □

We note that we already compute $\boldsymbol{V}_{\hat{\alpha}_t}$ at each round as part of $\widehat{\boldsymbol{P}_{\alpha_t}\beta}$. Since $\boldsymbol{V}_{\hat{\alpha}_{[t]}}\boldsymbol{V}_{\hat{\alpha}_{[t]}}^T \xrightarrow{p} \boldsymbol{P}_{\alpha_{[t]}}$, the check for identified components in Corollary 4 can therefore be performed efficiently in practice by checking for (approximate) equality of $\boldsymbol{V}_{\hat{\alpha}_{[t]}}\boldsymbol{V}_{\hat{\alpha}_{[t]}}^T e_i \approx e_i$. In our empirical evaluation, we use the absolute value of the cosine similarity, denoted by `cdist`, between the two vectors as a continuous measure for whether β_i has been identified. As an aggregate metric, we report the percentage of identified components

$$\tfrac{1}{d_x}\left|\{i \in [d_x] \mid \texttt{cdist}(V_{\alpha[t]}V_{\alpha[t]}^T e_i, e_i) < \delta\}\right| \tag{6}$$

for some fixed tolerance $\delta > 0$.

Finally, we can estimate an upper bound on the absolute error in each non-identified component. Let $\beta = \boldsymbol{P}_\alpha\beta + \nu$ be the orthogonal decomposition of β into the instrumented subspace $\mathrm{im}(\alpha)$ and its orthogonal complement. Since $\|\beta\|_2^2 = \|\boldsymbol{P}_\alpha\beta\|_2^2 + \|\nu\|_2^2$ and $|\nu_i| \leq \|\nu\|_2$ for all components $i \in [d_x]$, we have

$$|\nu_i| \leq \sqrt{\|\beta\|_2^2 - \|\boldsymbol{P}_\alpha\beta\|_2^2} \text{ for all } i \in [d_x]\,. \tag{7}$$

With our consistent estimates $\widehat{\|\beta\|}_2$ and $\widehat{\boldsymbol{P}_\alpha\beta}$, we can thus upper bound all remaining unidentified components.

We return now to Figure 2 which illustrates the estimation (upper panel) and combination steps (lower panel) in our linear setting for $d_z = 2, d_x = 3, T = 2$. For both instruments, we estimate the projection of β onto the respective instrumented spaces ($P_{\alpha_1}\beta$ and $P_{\alpha_2}\beta$), including the effect $P_{\alpha_{1/2}}\beta$ which is the one that should be recovered. In the lower panel, the two planes are the orthogonal complements of the instrumented spaces and their intersection corresponds to all vectors that are compatible, i.e., would be projected onto $\mathrm{im}(\alpha_1)$ and $\mathrm{im}(\alpha_2)$ respectively. This corresponds to the constraints in Equation (4). Among those, we then select the vector with the smallest norm as our combined estimate. The figure both illustrates the necessity of linearity for the combination of estimates and the increasing norm of the combined estimate converging to $\|\beta\|$.

4. Sequential Selection of Instruments

At each step $t \in [T]$, we seek to select the most informative subset of instruments α_t out of the pool of N_{IV} available choices. After each round we combine the newly obtained estimate $\widehat{\boldsymbol{P}_{\alpha_t}\beta}$ with all previous ones to obtain $\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta}$. Regarding consideration (B3), we assume a cost function $c : [N_{\mathrm{IV}}] \to \mathbb{R}_{\geq 0}$, where $c(d_z)$ is the cost of running a single experiment with d_z randomized instruments. For instance, this may be the actual monetary and logistic cost of randomly administering the selected drugs to a collection of n cell cultures ($\boldsymbol{Z}$), sequencing the cells ($\boldsymbol{X}$), and measuring the outcome of interest ($\boldsymbol{y}$) for each culture. However, the cost may also incorporate a hard limit $N_{\mathrm{IV}/\mathrm{exp}}$ on the number of instruments that can sensibly be combined in a single experiment (e.g., without killing the organism), by setting $c(d) = \infty$ for $d > N_{\mathrm{IV}/\mathrm{exp}}$.

In light of Equation (5), the ultimate goal is to select instruments that maximize $\|\widehat{\boldsymbol{P}_{\alpha_{[T]}}\beta}\|_2$. However, in round t

[5] Since we assume experimental access, we can also collect purely observational (X, Y) data about the system of interest.

[6] Recall that the 2-norm of an orthogonal projection of β is always smaller or equal to the 2-norm of β. This provides additional motivation for why we choose to combine estimates according to Equation (4) using a minimum 2-norm objective.

we cannot anticipate by how much $\|\widehat{\boldsymbol{P}_{\alpha_{[t-1]}}\beta}\|_2$ is going to increase for a candidate set of instruments α_t without actually performing the experiment. Therefore, we must rely on another signal to sequentially select subsets of instruments. Without any information about the available instruments, we cannot do better than selecting instruments (uniformly) at random at each step. While in practice one may not have precise information about the actual effects of individual instruments, information about the similarity of different antibiotics or drugs is typically still available. For example, certain antibiotics may have related active agents (high similarity) or certain drugs may target similar pathways (high similarity). We assume that pairwise normalized similarities $\mathrm{sim}_{i,j} \in [0,1]$ are provided for all available instruments $i, j \in [N_{\mathrm{IV}}]$. Here, $\mathrm{sim}_{i,j} = \mathrm{sim}_{j,i}$ and $\mathrm{sim}_{i,i} = 1$. Such similarities could also be derived from a set of features known about the instruments. To optimally explore the treatment space, it is then natural to attempt to sequentially select highly dissimilar instruments.

Hence, to evaluate the expected gain of adding a new set of instruments $\mathcal{I} \subset [N_{\mathrm{IV}}]$ to the already used instruments $\mathcal{J} \subset [N_{\mathrm{IV}}]$, we define the following gain function

$$\begin{aligned} \mathrm{gain} : 2^{[N_{\mathrm{IV}}]} \times 2^{[N_{\mathrm{IV}}]} \to \mathbb{R}_{\geq 0}\,, \qquad (8) \\ (\mathcal{I}, \mathcal{J}) \mapsto \frac{1}{|\mathcal{I}| + |\mathcal{J}| - 1} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I} \cup \mathcal{J}} (1 - \mathrm{sim}_{i,j})\,. \end{aligned}$$

This takes into account the similarities of the newly proposed instrument set within itself as well as with respect to the previously used ones. For example, when all instruments are maximally dissimilar ($\mathrm{sim}_{i,j} = \delta_{i,j}$), $\mathrm{gain}(\mathcal{I}, \mathcal{J}) = |\mathcal{I}|$ regardless of $\mathcal{J}$. When all instruments are equal ($\mathrm{sim}_{i,j} = 1$), $\mathrm{gain}(\mathcal{I}, \mathcal{J}) = 0$ for all inputs. Finally, we define the score of the set $\mathcal{I}$ when the set $\mathcal{J}$ has already been used by

$$\mathrm{score}(\mathcal{I}, \mathcal{J}) := \mathrm{gain}(\mathcal{I}, \mathcal{J}) - c(|\mathcal{I}|)\,. \qquad (9)$$

These considerations lead to the following setting. At each round t, we select a subset $\mathcal{I}_t \subset [N_{\mathrm{IV}}] \setminus \mathcal{I}_{[t-1]}$ of still unused instruments that maximize the score function given the already used instruments. We run a randomized experiment with those instruments to collect data, estimate (a projection of) β from this data (Proposition 1), and combine the estimate with the previous ones (Proposition 2). By convention, we have $\mathcal{I}_{[0]} = \mathcal{I}_\emptyset = \emptyset$ and we overload terminology to call both $\alpha_t \in \mathbb{R}^{|\mathcal{I}_t| \times d_x}$ and $\mathcal{I}_t \subset [N_{\mathrm{IV}}]$ "the instruments selected at round t". We similarly use $\alpha_{[t]}$ and $\mathcal{I}_{[t]}$ for the instruments selected up to (and including) round t. When the stopping criterion (Equation (5)) is satisfied, we return our current estimate as an estimate of the full β. Otherwise, we return the estimate after T experiments together with the identified components (Corollary 4). We outline our sequential instrument selection (SIS) procedure in Algorithm 1.

We remark that since $\mathrm{gain}(\mathcal{I}, \mathcal{J}) \leq |\mathcal{I}|$ it makes sense to use a sublinear cost function (until a potential hard limit $N_{\mathrm{IV}/\mathrm{exp}}$). Intuitively, while it becomes more expensive to include multiple randomized variables in a single experiment, it is still cheaper than running an individual randomized experiment for each instrument separately. The precise choice of the cost function is informed by the actual experimental setting and determines the trade-off between randomizing many instruments at once (to increase the dimensionality of the instrumented subspace) and the cost of doing so.

Algorithm 1 Sequential selection of instrument sets

Require: maximum rounds T, pairwise similarities $\mathrm{sim} \in [0,1]^{N_{\mathrm{IV}} \times N_{\mathrm{IV}}}$, cost function c, tolerance $\epsilon > 0$

collect observational data $\boldsymbol{X}, \boldsymbol{y}$
compute $\widehat{\|\beta\|}_2$ from $\boldsymbol{X}, \boldsymbol{y}$ ▷*Janzing (2019, ConCorr)*
$\mathcal{C} \leftarrow \emptyset$ ▷*set of identified components*
for $t \in [T]$ **do** ▷*experimental rounds*
  $\mathcal{I}_t \leftarrow \arg\max_{\mathcal{I} \subset [N_{\mathrm{IV}}] \setminus \mathcal{I}_{[t-1]}} \mathrm{score}(\mathcal{I}, \mathcal{I}_{[t-1]})$
  collect $\boldsymbol{Z}_t, \boldsymbol{X}_t, \boldsymbol{y}_t$ ▷*run experiment with $\mathcal{I}_t$*
  $\widehat{\boldsymbol{P}_{\alpha_t}\beta} \leftarrow (\boldsymbol{X}_t^T \boldsymbol{P}_{\boldsymbol{Z}_t} \boldsymbol{X}_t)^+ \boldsymbol{X}_t^T \boldsymbol{P}_{\boldsymbol{Z}_t} \boldsymbol{y}_t$ ▷*Proposition 1*
  $\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta} \leftarrow \arg\min_{\gamma \in \mathbb{R}^{d_x}} \|\gamma\|_2$ ▷*Proposition 2*
    s.t. $\widehat{\boldsymbol{P}_{\alpha_\tau}\beta} = \boldsymbol{V}_{\hat{\alpha}_\tau} \boldsymbol{V}_{\hat{\alpha}_\tau}^T \gamma$ for all $\tau \in [t]$
  if $|\|\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta} - \widehat{\|\beta\|}_2| < \epsilon$ **then** ▷*fully identified, (5)*
    $\mathcal{C} \leftarrow [d_x]$
    return $\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta}, \mathcal{C}$
  $\mathcal{C} \leftarrow \{i \in [d_x] \mid \boldsymbol{V}_{\hat{\alpha}_{[t]}} \boldsymbol{V}_{\hat{\alpha}_{[t]}}^T e_i \approx e_i\}$ ▷*Corollary 4*
return $\widehat{\boldsymbol{P}_{\alpha_{[T]}}\beta}, \mathcal{C}$ ▷*estimate, identified components*

5. Empirical Evaluation

Setup. Since a real-world evaluation of our approach would require access to sequential randomized experimentation in a complex setting, we are restricted to simulation studies. We first illustrate the properties of our proposed (combined) causal effect estimators in the underspecified IV setting and then evaluate our full sequential instrument selection method. We generate the parameters α, β randomly (Appendix A) in order to avoid parameter selection bias. Next, we fix a mixing matrix $\boldsymbol{M} \in \mathbb{R}^{d_x \times d_x}$ with all entries sampled from independent standard Gaussians as well as a direction $v \in \mathbb{R}^{d_x}$ as a uniform sample from the sphere $\mathbb{S}^{d_x - 1}$. In Equation (1), we then sample instrument components independently from a Rademacher distribution, and set $\epsilon_X = \boldsymbol{M}e$, $\epsilon_Y = v^T e$ with all components of $e \in \mathbb{R}^{d_x}$ sampled independently from standard Gaussians for each sample. This confounding model satisfies the assumptions required to estimate $\widehat{\|\beta\|}_2$ (Janzing & Schölkopf, 2018; Janzing, 2019; Rendsburg et al., 2022). Loosely speaking, the

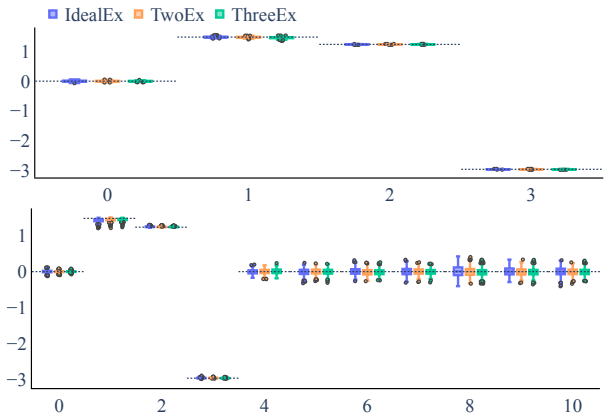

Figure 3. Estimates of β over 500 runs with $d_z = d_x = 3$ (top) and $d_z = 3$ and $d_x = 10$ (bottom). The dotted lines are the true β and boxplots show the median (horizontal line), first and third quartile (box height) and 10/90 percentiles (whiskers).

confounder affects both the treatment and the outcome as (independent) random mixtures of the same independent noise sources. We ensure in our data generation that $\beta \in \mathrm{im}(\alpha)$. Hence, β can be recovered fully in principle. Moreover, we also guarantee that there is a subset of less than $N_{\mathrm{IV}/\mathrm{exp}} \cdot T$ instruments that suffices to fully identify β (see Appendix A for details). Therefore, even if we cannot use all instruments throughout the experiments ($N_{\mathrm{IV}/\mathrm{exp}} \cdot T < N_{\mathrm{IV}}$) a good selection algorithm can in principle fully identify β.

For the similarities between instruments, in our experiments we take the absolute value of the cosine distance $\mathrm{sim}_{i,j} = |\alpha_i^T \alpha_j| / (\|\alpha_i\|_2 \|\alpha_j\|_2)$. Crucially, we thereby do not assume access to α—only the similarities enter the algorithm. While in a given application, the similarities would be provided by practitioners and domain experts, in our simulated study we have to choose *some* similarity measure that is informed by α. We account for uncertainties in the provided similarities by computing them on a noisy version of α, where we add independent standard Gaussian noise to all entries. For the cost function we choose $c(d) = \log(d)$ for $d \leq N_{\mathrm{IV}/\mathrm{exp}}$ and $c(d) = \infty$ otherwise, effectively limiting the maximum number of instruments per round to $N_{\mathrm{IV}/\mathrm{exp}}$. While pairwise similarities can help the sequential selection to converge quickly, we note that our findings from Section 3 are extremely useful for the (often strong) baseline of selecting instruments randomly.

Finite sample properties of our estimators. We first empirically analyze the finite sample behavior of our estimators from Propositions 1 and 2. For illustration, Figure 3 shows a setting with $d_x = d_z = 3$ on the upper panel and $d_x = 10$, $d_z = 3$ on the lower panel. The x-axis shows component indices of $\beta \in \mathbb{R}^{d_x}$ (component 0 is the offset); dotted lines are the ground truth values of β, and boxplots show distribution of estimated values ($\widehat{\boldsymbol{P}_\alpha \beta}$) over 500 random seeds. We compare three different estimators. *IdealEx* randomizes all three available instruments simultaneously and shows the single estimate (from Proposition 1). *TwoEx* combines (via Proposition 2) two individual estimates using only the first two and the last instrument, respectively. *ThreeEx* combines three individuals estimates, obtained by randomizing each of them separately. Figure 3 shows that all methods correctly identify all three components of β (plus the offset) on average with low variance in the low-dimensional $d_x = 3$ setting. The $d_x = 10$ setting (lower panel) in which 3 instruments suffice for full identification, shows similar results for all components. Even though the variance increases compared to the lower-dimensional setting, estimates are still consistent. All details are provided in Appendix A.

Sequential instrument selection. For our instrument selection algorithm we compare the following methods. *IdealEx*: the hypothetical ideal where all instruments are used at once in a single experiment. *Random*: a baseline that selects one of the allowed (smaller or equal $N_{\mathrm{IV}/\mathrm{exp}}$) subsets uniformly at random from the remaining instruments at each round. *Sequential Instrument Selector (SIS)*: our Algorithm 1. We remark that the random baseline, to its advantage, does not respect the cost per experiment, but is allowed to select the maximum number of possible instruments in each round.

We set $d_z = 30, d_{\mathrm{id}} = 15$ and use two different treatment dimensions $d_x = 50$ and $d_x = 150$. Further, we allow $N_{\mathrm{IV}/\mathrm{exp}} = 3$ instruments per round with a total budget of $T = 6$ experimental rounds. Note that $N_{\mathrm{IV}/\mathrm{exp}} \cdot T \leq d_{\mathrm{id}}$, i.e., the algorithm has the budget to fully identify β.

In Figure 4 (*left*) we show boxplots of the mean squared error (MSE) of our estimates for the nonzero components after each optimization round. *SIS* outperforms the random baseline in terms of MSE. Additionally, each boxplot in the lower panel shows the percentage of uncertain components from Equation (6) for $\delta = 0.3$. In Figure 4 (*right*), the norm of our estimator $||\widehat{\boldsymbol{P}_\alpha \beta}||$ converges towards the norm of the actual β (and its estimate $\widehat{\|\beta\|_2}$), i.e., the stopping criterion is reached. Figure 5 compares these estimates for the nonzero components of β after the last round of experiments. Our greedy optimization indeed identifies each of the components just as well as the hypothetical ideal of a single experiment that uses all instruments at once. Empirically, the finite sample properties remain unaffected by our combination procedure. We refer to Appendix B for results on a $d_x = 150$ setting.

6. Conclusion

In this work we made multiple contributions aiming at inferring causal effects of high-dimensional treatments under unobserved confounding by sequential experimentation, where we cannot intervene on the treatments directly, but can only randomize instruments. We proposed consistent, asymptoti-

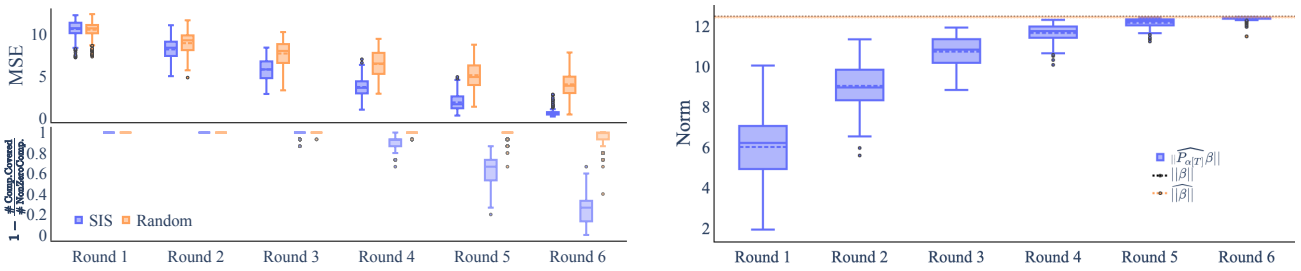

Figure 4. Results for the sequential selection of instruments with $d_z = 30, d_{\text{id}} = 15, N_{\text{IV/exp}} = 3, T = 6$ and $d_x = 50$: each boxplot shows the median and mean (solid resp. dashed line), first and third quartile (box height) and 10/90 percentiles (whiskers) over $n_{\text{runs}} = 250$. **Left:** The upper panel shows the of squared error $\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta} - \beta$ over rounds $t \in \{1, \dots, 6\}$. The lower panel shows the corresponding percentages of unidentified components. **Right:** The $||\widehat{\boldsymbol{P}_{\alpha_{[t]}}\beta}||$ increases for $t \in \{1, \dots, 6\}$ approaching $\widehat{||\beta||}$ (dotted orange line), which perfectly estimates the true $\|\beta\|$ in this case (dotted black line).

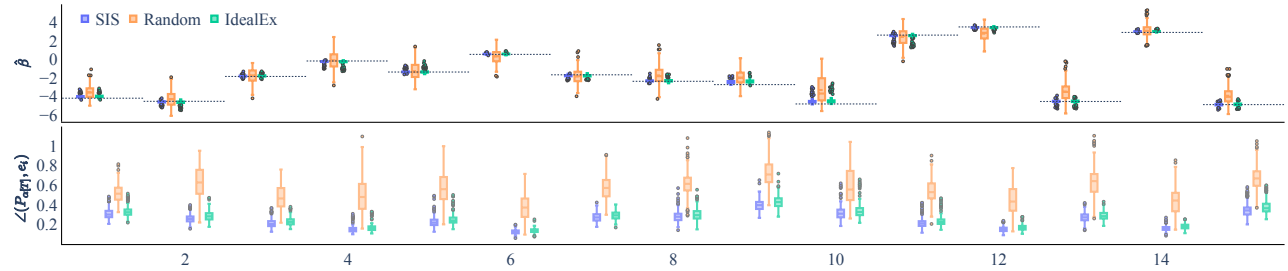

Figure 5. Results for the sequential selection of instruments with $d_z = 30, d_{\text{id}} = 15, N_{\text{IV/exp}} = 3, T = 6$ and $d_x = 50$: each boxplot shows the median and mean (solid resp. dotted line), first and third quartile (box height) and 10/90 percentiles (whiskers) over $n_{\text{runs}} = 250$ after the last round $T = 6$. **Top:** estimates of the nonzero components of β with the black dotted line being ground truth. **Bottom:** distribution of cosine similarities in Equation (6) for each component.

cally normal estimators for the orthogonal projection of a treatment effect onto the instrumented subspace in the linear setting and introduced a method to consistently combine such estimates from separate experiments. Surprisingly, neither the (perhaps intuitive) estimator in Proposition 1, nor the geometric intuition around instrumented subspaces in the underspecified setting can be found in existing literature. These estimators may be of independent interest as a contribution to the largely ignored underspecified IV setting. We then developed an algorithm to sequentially propose subsets of instruments from a given pool that flexibly trades off the expected information gain (informed by provided similarities) with the cost of each experiment. Moreover, we integrated a stopping criterion for when the sequential selection has fully identified the causal effect, a method to keep track of all components that are consistently estimated, and an upper bound on the absolute error of unidentified components: these additions inform the practitioner about whether (and which parts) of the estimate can be trusted.

The linearity assumption may appear restrictive. However, the linear IV setting is still heavily used in econometrics and health, as it reliably captures dominant effects even in noisy settings, and still attracts attention with novel results recently (Pfister & Peters, 2022; Rothenhäusler et al., 2018). The thorough theoretical understanding developed in this work is a challenging and necessary foundation for experiment design via instrument selection. Extensions of our method and of our notion of the instrumented subspace to (certain) non-linear settings is an important direction for future work. A second limitation of our work is inherent to the problem setting: missing real-world experiments due to a lack of access to the required expensive, specialized facilities. Finally, we highlight that independent testing and verification is paramount when using algorithmically obtained insights to inform consequential decisions such as actual clinical treatment decisions.

Code. The implementation as well as experimental details are publicly available on Github: https://github.com/EAiler/underspecified-iv.

Acknowledgments. We thank Sören Becker for useful suggestions and Dominik Janzing for fruitful discussions on estimating $\|\beta\|_2$. EA is supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS". JH is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Recursion Pharmaceuticals. This work has been supported by the Helmholtz Association's Initiative and Networking Fund through CausalCellDynamics (grant # Interlabs-0029).

References

Ailer, E., Müller, C. L., and Kilbertus, N. A causal view on compositional data. *arXiv preprint arXiv:2106.11234*, 2021. 1, 2

Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008. 2, 3, 4

Bennett, A., Kallus, N., and Schnabel, T. Deep generalized method of moments for instrumental variable analysis. 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/15d185eaa7c954e77f5343d941e25fbd-Paper.pdf. 2

Bura, E. and Pfeiffer, R. On the distribution of the left singular vectors of a random matrix and its applications. *Statistics & Probability Letters*, 78(15):2275–2280, 2008. 4

Didelez, V. and Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330, 2007. 2

Fu, Y., Foden, J., Khayter, C., Maeder, M., Reyon, D., Joung, J., and Sander, J. High-frequency off-target mutagenesis induced by crispr-cas nucleases in human cells. *Nature biotechnology*, 31, 06 2013. doi: 10.1038/nbt.2623. 1

Gamella, J. L. and Heinze-Deml, C. Active invariant causal prediction: Experiment selection through stability. *Advances in Neural Information Processing Systems*, 33: 15464–15475, 2020. 2

Hahn, J. and Hausman, J. Estimation with valid and invalid instruments. *Annales d'Économie et de Statistique*, 79/80: 25–57, 2005. ISSN 0769489X, 22726497. URL http://www.jstor.org/stable/20777569. 3

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pp. 1414–1423, 2017. 2

Hernán, M. A. and Robins, J. M. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, pp. 360–372, 2006. 2

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. Experiment selection for causal discovery. *Journal of Machine Learning Research*, 14:3041–3071, 2013. 2

Janzing, D. Causal regularization. *Advances in Neural Information Processing Systems*, 32, 2019. 5, 7

Janzing, D. and Schölkopf, B. Detecting non-causal artifacts in multivariate linear regression models. In *International Conference on Machine Learning*, pp. 2245–2253. PMLR, 2018. 5, 7, 12

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American statistical Association*, 111(513):132–144, 2016. 3

Kilbertus, N., Kusner, M. J., and Silva, R. A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. 2

Moschopoulos, P. G. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, 37(1):541–544, 1985. 6

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. Dual instrumental variable regression. *arXiv preprint arXiv:1910.12358*, 2019. 2

Nocedal, J. and Wright, S. J. *Numerical optimization*. Spinger, 2006. 5

Padh, K., Zeitler, J., Watson, D., Kusner, M., Silva, R., and Kilbertus, N. Stochastic causal programming for bounding treatment effects, 2022. URL https://arxiv.org/abs/2202.10806. 2

Pearl, J. *Causality*. Cambridge university press, 2009. 1

Pfister, N. and Peters, J. Identifiability of sparse causal effects using instrumental variables. *arXiv preprint arXiv:2203.09380*, 2022. 2, 9

Rendsburg, L., Vankadara, L. C., Ghoshdastidar, D., and von Luxburg, U. A consistent estimator for confounding strength. *arXiv preprint arXiv:2211.01903*, 2022. 5, 6, 7

Rothenhäusler, D., Bühlmann, P., Meinshausen, N., and Peters, J. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 01 2018. doi: 10.1111/rssb.12398. 2, 9

Saengkyongam, S., Henckel, L., Pfister, N., and Peters, J. Exploiting independent instruments: Identification and distribution generalization. *arXiv preprint arXiv:2202.01864*, 2022. 2

Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., Palmer, T., Schooling, C. M., Wallace, C., Zhao, Q., and Davey Smith,

G. Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):6, 2022. 2

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pp. 4593–4605, 2019. 2

Sohn, M. B. and Li, H. Compositional mediation analysis for microbiome studies. *Annals of Applied Statistics*, 13(1):661–681, 2019. ISSN 19417330. doi: 10.1214/18-AOAS1210. 2

Stock, J. H. and Trebbi, F. Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194, 2003. 2

Sussex, S., Uhler, C., and Krause, A. Near-optimal multi-perturbation experimental design for causal structure learning. *Advances in Neural Information Processing Systems*, 34:777–788, 2021. 2

Tigas, P., Annadani, Y., Jesson, A., Schölkopf, B., Gal, Y., and Bauer, S. Interventions, where and how? experimental design for causal models at scale. *arXiv preprint arXiv:2203.02016*, 2022. 2

Wang, C., Hu, J., Blaser, M. J., Li, H., and Birol, I. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics*, 2020. ISSN 14602059. doi: 10.1093/bioinformatics/btz565. 2

Wright, P. G. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928. 2

Zhang, R., Imaizumi, M., Schölkopf, B., and Muandet, K. Maximum moment restriction for instrumental variable regression. *arXiv preprint arXiv:2010.07684*, 2020. 2

Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S., and Yang, S.-H. Off-target effects in crispr/cas9-mediated genome engineering. *Molecular Therapy - Nucleic Acids*, 4:e264, 2015. 1

A. Details of Experiments

Data Generation The data generation process follows the model described in (Janzing & Schölkopf, 2018). We adopt their data setup in order to estimate $\widehat{||\beta||}$ consistently:

$$X = Z\alpha + \epsilon_X \,, \qquad Y = X\beta + \epsilon_Y \,, \tag{10}$$

ϵ_X and ϵ_Y are confounded via the common variable e.

$$\epsilon_X = eM \,, \qquad \epsilon_Y = ev \,, \tag{11}$$

with $Z \sim \text{Rademacher}(0.5)$ and $e \sim \mathcal{N}(0, Id_l)$.

Scenario Generation. For each figure, i.e. simulation, we generate random scenarios in order to avoid introducing involuntary bias in the parameter setting.

Each generated scenario is based on a seed and the constants n, d_x, d_z and d_{id}, i.e. the number of instruments it will take to identify the causal effect in full. Further, we assume $d_x = l$, i.e. $M \in \mathbb{R}^{d_x \times d_x}$. Note that this choice is based on the setting in the simulation studies of (Janzing & Schölkopf, 2018).

We sample d_{id} nonzero components of $\alpha_j, j \in \{1, ..., d\}$ and the d_{id} nonzero components of β from a uniformly distributed random variable $\mathcal{U}(-5.0, 5.0)$. The sparsity is introduced by setting the remaining components to 0.0. Our motivation for this choice is to reliably generate a setup for which we are guaranteed to identify the causal effect by a maximum of d_{id} instruments. In addition to d_{id} identifying instruments, we generate $d - d_{\text{id}}$ further instruments by picking two of the necessary instruments on top of which we add a p-dimensional Gaussian noise. Overall, we end up with $d_{\text{id}} - 2$ necessary instruments and two clusters from which we can pick any instrument in order to identify the remaining components of β. For the confounder we sample all entries of $\boldsymbol{M} \in \mathbb{R}^{d_x \times d_x}$ from independent standard Gaussians and the direction $v \in \mathbb{R}^{d_x}$ as a uniform sample from the sphere $\mathbb{S}^{d_x - 1}$.

Scenario Generation for Figure 3. For showcasing the finite sample properties of our estimator, we sampled from the scenario generation above with seed 253 and $d_z = 3$ and $d_x = 3$ resp. $d_x = 10$. This choice of parameters left us with two settings (1) being just-identified $d_z = d_x = 3$ and (2) being underspecified $d_z = 3, d_x = 10$. Moreover, we set β to only have 3 non-zero components in order to be able to identify the causal effect in full. Note that this choice is for illustration purpose and does not affect the method's applicability in a setting where we can only identify parts of the causal effect. However, in those setting where some β_i-components are not part of the instrumented subspace $\text{im}(\alpha)$, full causal recovery is in general impossible.

B. Additional Experiments

In the optimization we compare the baseline which uses all instruments at once (*IdealEx*) to the developed sequential optimization routine (*SIS*) and the random baseline. We include results for the same setting as Figure 4 and Figure 5, except with an increased treatment dimension, i.e. $d_x = 150$ in Figure 6:

$$d_x = 150,\ d_z = 30,\ d_{\text{id}} = 15,\ N_{\text{IV/exp}} = 3,\ T = 6$$

Moreover, it might not always be the case that the stopping criterion in (Janzing & Schölkopf, 2018) works as nicely as in the previous scenarios, see Fig. Figure 7 with parameter setting:

$$d_x = 50,\ d_z = 30,\ d_{\text{id}} = 20,\ N_{\text{IV/exp}} = 4,\ T = 6$$

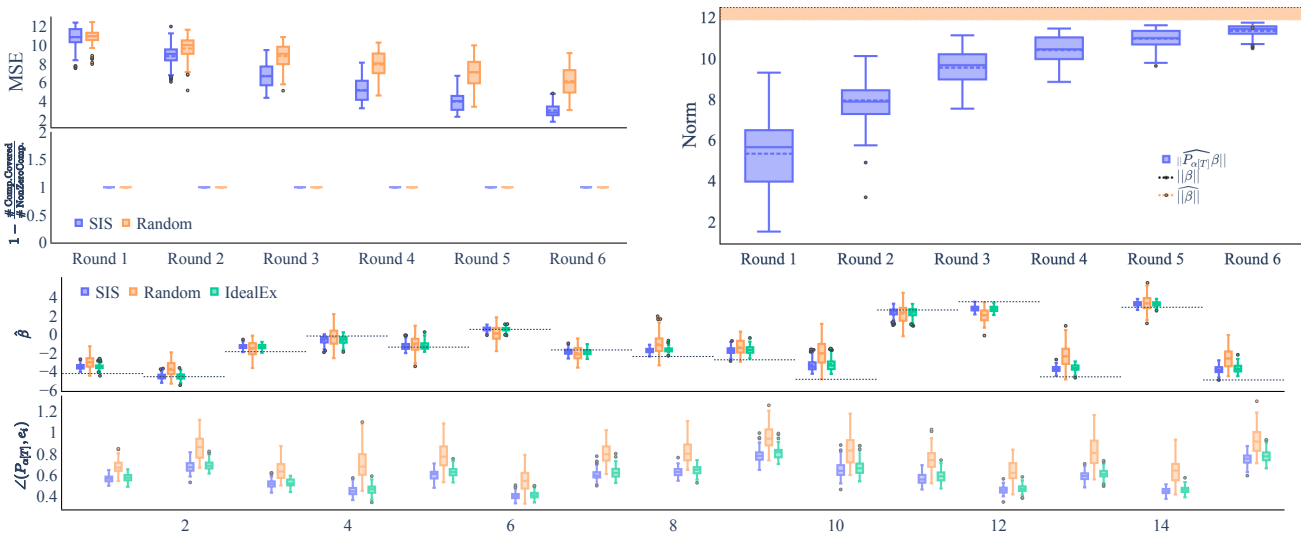

Figure 6. Results for the sequential selection of instruments with $d_z = 30, d_{\text{id}} = 15, N_{\text{IV/exp}} = 3, T = 6$ and $d_x = 150$: each boxplot shows the median and mean (solid resp. dashed line), first and third quartile (box height) and 10/90 percentiles (whiskers) over $n_{\text{runs}} = 250$. **Top Left:** The upper panel shows the of squared error $\widehat{\boldsymbol{P}_{\alpha_{[t]}}}\beta - \beta$ over rounds $t \in \{1, \dots, 6\}$. The lower panel shows the corresponding percentages of unidentified components. As we are in the scenario with $p = 150$, we would need to adjust the threshold $\delta = 0.3$ to a higher value. **Top Right:** The $||\widehat{\boldsymbol{P}_{\alpha[t]}\beta}||$ increases for $t \in \{1, \dots, 6\}$ approaching $\widehat{||\beta||}$ (shaded orange block), which does not perfectly estimate the true $\|\beta\|$ in this case (dotted black line), but performs still reasonably well. **Bottom:** *Upper Panel:* estimates of the nonzero components of β with the black dotted line being ground truth. *Lower Panel:* distribution of cosine similarities in Equation (6) for each component.

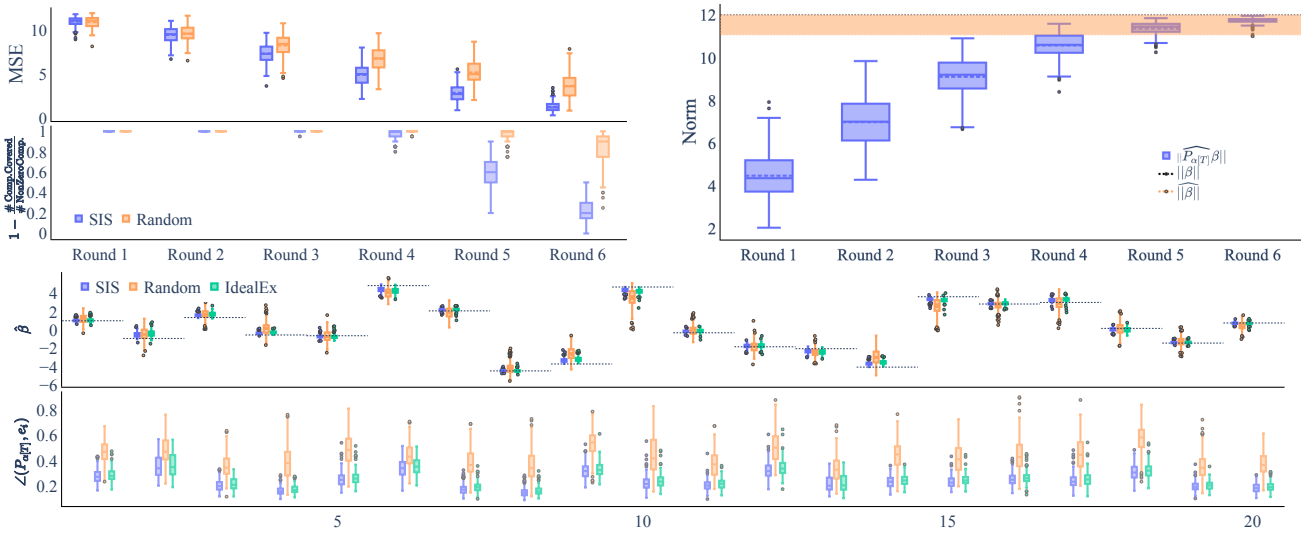

Figure 7. Results for the sequential selection of instruments with $d_z = 30, d_{\text{id}} = 20, N_{\text{IV/exp}} = 4, T = 6$ and $d_x = 50$: each boxplot shows the median and mean (solid resp. dashed line), first and third quartile (box height) and 10/90 percentiles (whiskers) over $n_{\text{runs}} = 250$. **Top Left:** The upper panel shows the of squared error $\widehat{\boldsymbol{P}_{\alpha_{[t]}}}\beta - \beta$ over rounds $t \in \{1, \dots, 6\}$. The lower panel shows the corresponding percentages of unidentified components. **Top Right:** The $||\widehat{\boldsymbol{P}_{\alpha[t]}\beta}||$ increases for $t \in \{1, \dots, 6\}$ approaching $\widehat{||\beta||}$ (shaded orange block), which does not perfectly estimate the true $\|\beta\|$ in this case (dotted black line), but performs still reasonably well. **Bottom:** *Upper Panel:* estimates of the nonzero components of β with the black dotted line being ground truth. *Lower Panel:* distribution of cosine similarities in Equation (6) for each component.