Necessary and Probably Sufficient Test for Finding Valid Instrumental Variables

Amit Sharma * [1,2]

[1]Microsoft Research, New York
[2]amshar@microsoft.com

ABSTRACT

Increasing availability of fine-grained data from diverse fields spanning the social and biomedical sciences presents new opportunities for gaining causal knowledge. While inferring the global causal graph or mechanisms is usually intractable, local causal estimation methods such as instrumental variables (IV) can be used for estimating causal effects. However, IV methods depend on two assumptions—*exclusion* and *as-if-random*—that are largely believed to be untestable from data, thus making it hard to evaluate the validity of an instrumental variable or compare estimates across studies even on the same dataset. In this paper, we show that when all variables are discrete, testing for instrumental variables is possible. We build upon prior work on necessary tests to derive a necessary and probably sufficient test for the validity of an instrumental variable. Given observational data on a pair of cause and effect variables, the proposed test determines whether an instrument is valid within some statistical error. The test works by defining the class of invalid-IV and valid-IV causal models in terms of Bayesian graphical models and comparing their marginal likelihood based on observed data.
Simulations of randomly selected causal models show that the test is most powerful for moderate-to-weak instruments; incidentally, such instruments are most commonly used in observational studies. It is also able to distinguish between invalid and valid IV causal models for an open problem proposed in past work. Applying the test to two seminal studies on instrumental variables and five recent studies from the American Economic Review shows that many of the instruments are flawed, at least when all variables are discretized. The proposed test opens the possibility of algorithmically finding and validating instruments in large datasets and more generally, adopting a data-driven approach to instrumental variable studies.

1 INTRODUCTION

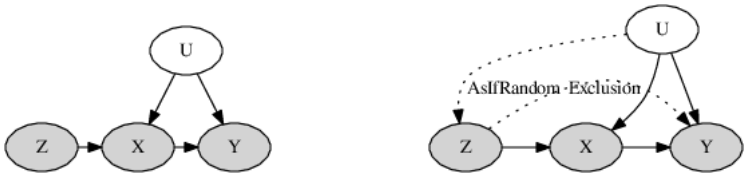
The method of *instrumental variables* is one of the most popular ways to estimate causal effects from observational data in the social and biomedical sciences. The key idea is to find subsets of the data where conditions were nearly the same as they would in a randomized experiment, and use those subsets to estimate causal effect. For example, instrumental variables have been used in economics to study the effect of policies such as military conscription and compulsory schooling on future earnings[1,2], and in epidemiology (under the name *Mendelian* randomization) to study the effect of risk factors on disease outcomes[3].

Figure 1a shows the canonical causal inference problem for the instrumental variables (IV) method. The goal is to estimate the effect of variable X on another variable Y based on observed data. However, there are unobserved (and possibly unknown) common causes for X and Y that contribute to the observed association between X and Y, making the isolation of X's effect on Y a non-trivial problem. Compared to conditioning approaches such as stratification or matching[4], instrumental variables have the advantage that they do not require observing any of the confounders to estimate the causal effect. Rather, identification relies on finding an additional variable Z, that acts as an *instrument* to modify the distribution of X, as shown by the arrow $Z \rightarrow X$ in Figure 1a.

To be a valid instrument, however, Z should satisfy three conditions[2]. First, Z should have a substantial effect on X. That is, Z causes X *(Relevance)*. Second, Z should not cause Y directly (*Exclusion*); the only association between Z and Y should be through X. Third, Z should be independent of all the common causes U of X and Y (*As-if-random*). The latter two conditions are shown in the graphical model in Figure 1b. Taken together, the as-if-random and exclusion conditions are equivalent to $Z \perp\!\!\!\perp U$ and $Z \perp\!\!\!\perp Y || X, U$ respectively.

However, the Achilles heel of any instrumental variable analysis is that these core conditions are never tested systematically. Except for relevance (which can be tested by looking at the observed correlation between Z and X), the other conditions are usually considered as *assumptions*, and are defended with qualitative domain knowledge. This can be problematic, especially because the entire validity of the IV estimate depends on the exclusion and as-if-random conditions. It also makes it hard to compare and evaluate studies that use instrumental variables.

*I acknowledge Jake Hofman and Duncan Watts for their valuable feedback throughout the course of this work.



(a) Valid-IV causal model **(b)** Invalid-IV causal model

Figure 1. Standard instrumental variable causal model and common violations that lead to an invalid-IV model. Exclusion condition is violated when the instrumental variable Z directly affects the outcome Y. As-if-random condition is violated when unobserved confounders U also affect the instrumental variable Z.

As a remedy, there is a line of work using causal graphical models that devises necessary tests that follow from the IV conditions[5,6]. These tests can be used to weed out bad instruments, but are inconclusive whenever an instrument passes the test. Other tests such as the Durbin-Wu-Hausman test[7] are sufficient, but require knowledge of at least one valid instrumental variable; the circular nature of the test makes it impractical for most situations. These shortcomings have led scholars to declare instrumental variables as untestable from data[8,9].

In this paper, we show that whenever X,Y and Z are discrete, it is possible to test for the validity of an instrumental variable using observational data only. By combining ideas from causal graphical models and Bayesian statistics, we present a necessary and *probably* sufficient test for instrumental variables. While not provably sufficient, the proposed test quantifies the likelihood that an instrument is valid given the observed data sample. Specifically, it compares the probability of a valid-IV model to the probability of an invalid-IV model, given the observed data.

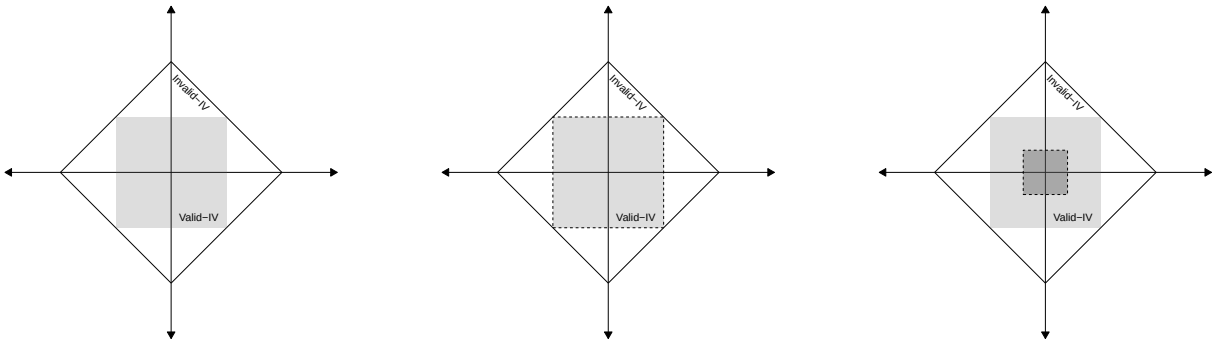
Simply computing the likelihood of generating observed data from these two classes of models—valid-IV and invalid-IV—will not help because the class of invalid-IV models, as shown in Figure 1b also includes the class of valid-IV models by definition. Therefore, any data distribution generated by a valid-IV model can also be generated by an invalid-IV model. To devise an informative statistical test, we need some way of identifying how a data distribution generated by a valid-IV model differs from invalid-IV models. Here we make use of a necessary conditions for instrumental variables proposed by Pearl[5] that restricts the joint probability distributions for $P(X, Y, Z)$ that any valid-IV model can generate. Armed with these conditions, our proposed statistical test proceeds as follows. If the observed data does not satisfy the necessary conditions, then it is declared invalid. If it does, then the relative likelihood of a valid-IV model is computed by estimating the joint likelihood of satisfying all necessary conditions and generating the observed data.

In theory, evaluating the test depends on enumerating and testing all possible causal models—both valid and invalid instrumental variable models—that could have generated the data. However, it is impossible to enumerate all causal models because there can be infinitely many models within the class of valid and invalid-IV models. Therefore, we introduce sampling strategies to approximate the enumeration. For simplicity, we first present a uniform sampling strategy assuming that all causal models are equally likely. However, given observed data, some of the models may be more likely than others. Therefore, we also present a computationally intensive sampling strategy that selects underlying causal models based on their likelihood of generating the observed data.

Finally, any statistical test is only as good as the decisions it helps to support. Using simulations, we show that the proposed NPS test is most effective for validating instruments having a low correlation with the cause of interest. Coincidentally, most of the instruments used in observational studies in the social and biomedical sciences also satisfy this property. To demonstrate its usefulness, we apply the test to datasets from two seminal papers on instrumental variables. In both cases, we find that instrumental assumptions used in the corresponding papers were possibly flawed, at least when variables are discretized. We also apply the NPS test to validate recent instrumental variable studies in the American Economic Review, a premier economics journal. Looking forward, the proposed test makes it possible to compare potential instruments for their validity, allow transparent comparisons between multiple IV studies, and enable more data-driven search for natural experiments[10].

2 BACKGROUND: TESTABILITY OF AN INSTRUMENTAL VARIABLE

Since sufficient conditions for validity of an instrument ($Z \perp\!\!\!\perp U$ and $Z \perp\!\!\!\perp Y|X, U$) depend on an unobserved variable U, the validity of an instrumental variable is largely believed to be untestable from observational data alone[8]. Nevertheless, there are some statistical tests that can be used to examine an instrumental variable. These tests either provide necessary conditions for a



(a) Data distributions generated by Valid-IV and Invalid-IV models

(b) Data distributions that pass Pearl's necessary test

(c) Data distributions that pass proposed NPS test

Figure 2. A 2D schematic of the probability space $P(X, Y|Z)$. Each point on the graph is a conditional probability distribution over X, Y given Z. The two squares in the left panel show polytopes for invalid-IV and valid-IV models in the probability space. Note that data distributions generated by valid-IV are a subset of the data distributions generated by invalid-IV. The right panel shows the test boundary for Pearl's necessary test in dotted lines, which coincides with the Valid-IV boundary for binary X. Still, Pearl's test is not sufficient because data distributions $P(X, Y|Z)$ inside the valid-IV polytope may also be generated by invalid-IV models.

valid IV or depend circularly on the knowledge of at least one valid instrumental variable. They fall into two broad categories: parametric and non-parametric.

2.1 Parametric tests for IV

One of the well-known tests for instrumental variables is the Durbin-Wu-Hausman test[7]. Given a subset of valid instrumental variables, it can identify whether other potential candidates are also valid instrumental variables. However, it provides no guidance on how to find the initial set of valid instrumental variables.

Without having an initial set of valid instrumental variables, one can test for a stronger condition, $Z \perp\!\!\!\perp Y|X$[11]. This will be a sufficient test, barring any incidental equality of parameters that leads to conditional independence between Z and Y. However, it is too restrictive because it can happen only if all the common causes U are constant throughout the data measurement period, or if U is not a common cause for both X and Y. In such a case, using an instrumental variable is redundant: the effect of X on Y is unconfounded and can be estimated simply using the observed probability $P(Y|X)$.

2.2 Graphical model-based tests for IV

More admissive tests can be obtained by considering the restriction on probability distribution for (Z, X, Y) imposed by a valid IV model. Consider the causal IV model from Figure 1a. In structural equations, the model can be equivalently expressed as:

$$y = f(x, u); \quad x = g(z, u) \tag{1}$$

where g and h are arbitrary deterministic functions and U is an unobserved random variable, independent of Z. Using this framework, Pearl derived a necessary test for checking an instrumental variable[5]. As an example, for binary variables Z, X and Y, the Pearl's IV test can be written as the following set of *instrumental inequalities*.

$$\begin{aligned}
&P(Y = 0, X = 0|Z = 0) + P(Y = 1, X = 0|Z = 1) \leq 1 \\
&P(Y = 0, X = 1|Z = 0) + P(Y = 1, X = 1|Z = 1) \leq 1 \\
&P(Y = 1, X = 0|Z = 0) + P(Y = 0, X = 0|Z = 1) \leq 1 \\
&P(Y = 1, X = 1|Z = 0) + P(Y = 0, X = 1|Z = 1) \leq 1
\end{aligned} \tag{2}$$

Typically, researchers make an additional assumption that helps to derive a point estimate for the Local Average Treatment Effect (LATE). This assumption, called monotonicity[12], precludes any *defiers* to treatment in the population[2]. That is, we assume that $g(z_1, u) \geq g(z_2, u)$ whenever $z_1 \geq z_2$. Under these conditions, we obtain tighter inequalities. For binary variables

Z, X and Y, the instrumental inequalities become:

$$P(Y=y, X=1|Z=1) \geq P(Y=y, X=1|Z=0) \qquad \forall y \in \{0,1\} \tag{3}$$
$$P(Y=y, X=0|Z=0) \geq P(Y=y, X=0|Z=1) \qquad \forall y \in \{0,1\} \tag{4}$$

Whenever any of these inequalities are violated, it implies that one or more of the IV assumptions—exclusion, as-if-random or monotonicity—are violated. Based on these instrumental inequalities, different hypothesis tests can been derived to account for sampling variability in observing the true conditional distributions. For example, a null hypothesis tests based on the chi-squared statistic[13] or the Kolmogorov-Smirnov test statistic[14] have been proposed.

When X, Y and Z is binary, this test is not only necessary, it is the strongest necessary test possible[6,14]. In other words, if an observed data distribution satisfies the test, then there does exist at least one valid-IV causal model that could have generated the data; we call this the *existence* property. However, it does not satisfy the existence property when all variables are not binary, allowing probability distributions that cannot be generated by any valid-IV model. To rectify this, Bonet proposed a more general version of the test that ensures the existence property for any discrete-valued X, Y and Z.[6]. We call this test as the Pearl-Bonet necessary and existence test for instrumental variables, or simply the *Pearl-Bonet test*.

While Bonet presented theoretical properties of the test for discrete variables, implementing the test in practice is non-trivial because it involves testing membership of a convex polytope in high-dimensional space. Further, the test does not support the monotonicity assumption, a popular assumption instrumental variable methods. In this paper, therefore, we extend Bonet's work by incorporating monotonicity and present a practical method for testing IVs when variables can have arbitrary number of discrete levels.

2.3 Towards a sufficient test

The above tests can *refute* an invalid-IV model, but are unable to *verify* a valid-IV model[14]. That is, even when an observed data distribution passes the necessary test, it does not exclude the possibility that data was generated by an invalid-IV model. To see why, let us look at Figure 2 that shows the relationship between an observed data distribution and Valid-IV or Invalid-IV models. Each point in Figure2a represents a probability distribution over X, Y and Z.[1] The two squares bound the probability distributions that can be generated by any Valid-IV or Invalid-IV model. As can be seen from the figure, probability distributions generatable from Valid-IV are a strict subset of the distributions generatable by the invalid-IV model. This implies that even if a statistical test can accurately identify the boundary for valid-IV models, as in Figure 2b, we can never be sure whether the probability distribution was actually generated by a valid-IV or invalid-IV model.

Therefore, a harder problem is to establish *sufficiency*: determining whether observed data was generated by a valid IV model. One way to establish this is by comparing the likelihood of different causal models given the data. However, as we argued above, likelihood-based approaches for graphical models[15] are not informative because invalid-IV models (as shown in Figures 1b and 2b) are more general than the valid-IV model and thus are always as likely (or more) to generate the observed data.

In this paper, we show that comparing the Bayesian evidence between valid-IV and invalid-IV models can be used to test for the validity of an instrumental variable. Unlike past statistical tests[13,14] that refute a null hypothesis that observed data was generated from a valid-IV model, we do so by comparing the probability of valid-IV and invalid-IV models given observed data. More precisely, given a probability distribution $P(X,Y,Z)$ from observed data, the test computes the ratio of marginal likelihoods for valid and invalid IV models. Whenever this marginal likelihood ratio is above a pre-determined acceptance threshold, our test concludes that the instrument is valid. To distinguish this statistical notion from deterministic sufficiency—conditions that would determine in absolute whether an instrument is valid or not—we call such a sufficiency as *probable sufficiency*.

Probable Sufficiency for Instrumental Variables: If an observed data distribution passes Pearl's necessary test, how likely is it that it was generated from a valid-IV model compared to an invalid-IV model?

Intuitively, we wish to find out how often does Pearl's necessary test provide a wrong answer. That is, how often does an observed distribution that was generated by an invalid-IV model pass the necessary test? Once we know that, we can compute

[1]The axes represent the space of conditional probabilities $P(X,Y|Z)$. We use the fact that any observed probability distribution $\mathcal{P}$ over X, Y and Z can be specified by a set of conditional probabilities of the form $P(X=x, Y=y|Z=z)$. For example, for binary variables, this would be a set of eight conditional probabilities[6]. The corresponding 8-dimensional real vector would be:

$$F(\mathcal{P}) = \big(P(X=0,Y=0|Z=0), P(X=0,Y=1|Z=0), P(X=1,Y=0|Z=0), P(X=1,Y=1|Z=0),$$
$$P(X=0,Y=0|Z=1), P(X=0,Y=1|Z=1), P(X=1,Y=0|Z=1), P(X=1,Y=1|Z=1)\big) \tag{5}$$

The 2-D squares represented in Figures 2a,b are actually polytopes in this multi-dimensional space. The extreme points for $F(\mathcal{P})$, or equivalently for invalid-IV models are characterized by $P(X=x, Y=y|Z=z) = 1\ \forall z$. The boundary shown for NPS test in Figure 2c is however, an oversimplification. The set of conditional distributions (or equivalently, instruments) that can be validated by the NPS test is unknown and most likely will constitute many regions in the probability space, instead of a single bounded region as shown.

the probability that a given observed distribution was generated by a valid-IV model, based on the result of Pearl's necessary test.

Combined, Pearl's necessary test and our probable sufficiency test provide a framework for testing instrumental variables, which we call the *Necessary and Probably Sufficient (NPS)* test for instrumental variables. Any valid instrument needs to pass Pearl's test. Therefore, NPS test provides necessity: any instrument that fails Pearl's instrumentality inequalities is not a valid instrument. Further, NPS test provides sufficiency: any instrument that satisfies Pearl's instrumental inequalities and passes the probable sufficiency test can be accepted as a valid instrument. That said, NPS test will be inconclusive for some instruments: those that satisfy Pearl's inequalities but the marginal likelihood ratio remains close to 1. Figure 2c shows these possibilities. As shown by the dark grey box in the center, NPS test validates a subset of all possible valid-IV models.

In the next two sections, we describe the NPS test formally. Section 3 presents a general *Validity Ratio* statistic that can be used to compare Valid-IV and Invalid-IV models. We do so by introducing a probabilistic generative meta-model that formalizes the connection between IV assumptions, causal models and the observed data. The key detail for computing the Validiity Ratio is in selecing a suitable sampling strategy for causal models. Section 4 describes one such strategy based on the response variable framework[11].

For the rest of the paper, we assume that X, Y and Z are all discrete. In principle, we can apply the NPS testing framework to both continuous and discrete values for X, Y and Z. However, the test is expected to be most informative when X is discrete. This is because when X is continuous, the region for Valid-IV identified by Pearl's necessary test coincides with the Invalid-IV region and the necessary test becomes uninformative[6].

For ease of exposition, we present the NPS test using binary Z, X and Y. In Section 5, we discuss how the test extends to the case where Z, X and Y can be arbitrary discrete variables. Finally, Sections 6 and 7 demonstrate the practical applicability of the test using simulation data and data from past IV studies.

3 A NECESSARY AND PROBABLY SUFFICIENT (NPS) TEST

The key idea behind the NPS test is that we can compare marginal likelihoods of valid-IV and invalid-IV class of models. To do so, we first describe a Bayesian meta-model that describes how observed data can be generated from different values of the exclusion and as-if-random conditions. We then provide the main result that provides a *Validity Ratio* to compare valid-IV and invalid-IV models, followed by pseudo-code for an algorithm that uses the NPS test to validate an instrumental variable.

3.1 Generating valid-IV and invalid-IV causal models

As mentioned above, our strategy depends on simulating all causal models—both valid-IV and invalid-IV models—that could have generated the observed data. Therefore, we first describe a probabilistic generative *meta-model* of how the observed data is generated from a causal model.

Let us first define the valid-IV and invalid-IV models formally in terms of the IV assumptions. A valid IV model does not contain an edge from $Z \rightarrow Y$ or from $U \rightarrow Z$, as shown in Figure 1a. This implies that both Exclusion and As-if-random conditions hold for a valid-IV model. For our example above, this would mean an encouragement instrument that is not confounded by common causes for peers' activities and encouragement to an individual does not directly lead to an encouragement for her peer, except through social influence. Conversely, a causal model is an invalid IV model when at least one of Exclusion or As-if-random conditions is violated, as shown by the dotted arrows in Figure 1b. Therefore, assuming the causal structure $Z \rightarrow X \rightarrow Y$, there are two classes of causal models that can generate observed data distributions over X, Y and Z:

- Valid IV model: $E = True$ and $R = True$
- Invalid IV model: *Not* ($E = True$ and $R = True$)

where E denotes the exclusion restriction and R denote the as-if-random restriction.

Each of these classes of causal models—valid and invalid IV—in turn contains multiple causal models, based on the specific parameters (θ) describing each edge of the graph. This one-to-many relationship between conditions for IV validity and causal models can be made precise using a generative meta-model, as shown in Figure 3. We show dotted arrows to distinguish this (probabilistic) generative meta-model from the causal models described earlier. The meta-model entails the following generative process: Based on the configuration of the Exclusion and As-if-Random conditions, either of the causal model classes—Valid or Invalid IV—is selected. A specific model (*Causal Model* node) is then generated by parameterizing the selected class of causal models, where we use θ to denote model parameters. The causal model results in a probability distribution over Z, X and Y, from which observed data (*Data* node) is sampled. Finally, we can apply Pearl's necessary test on the observed data, which leads to the binary variable $NecTestResult$.

For a given problem, we observe the data D and result of Pearl's necessary test. All other variables in the meta-model are unobserved.

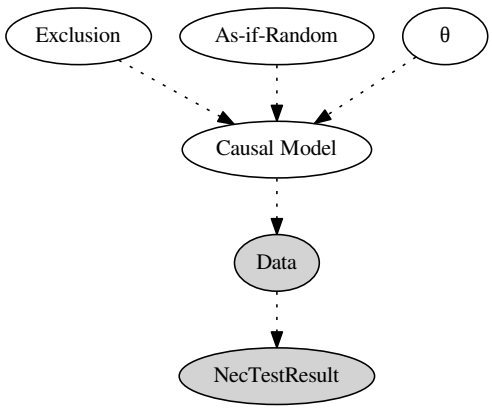


Figure 3. A probabilistic graphical meta-model for describing the connection between IV conditions and specific causal models. Evidence consists of both the results of the necessary test and observed data sample. Therefore, given this evidence, some causal models are expected to become more likely than others. Note that arrows are dotted to distinguish these *probabilistic* diagrams from the causal diagrams in Figure 1.

3.2 Comparing likelihood of valid-IV and invalid-IV models

Let PT denote whether the observed data passed the necessary test. We wish to estimate whether the data was generated from a Valid IV model. We can compare the likelihood of observing PT and D given that both Exclusion and As-if-random conditions are valid, versus when they are not.

Theorem 1. *Given a representative data sample D drawn from $P(X, Y, Z)$ over variables X, Y, Z, result of Pearl's necessary test PT on the data sample, the validity of Z as an instrument for estimating causal effect of X on Y can be decided using the following evidence ratio of valid and invalid classes of models:*

$$\textit{Validity-Ratio} = \frac{P(E,R|PT,D)}{P(\neg(E,R)|PT,D)} = \frac{P(PT,D|E,R) * P(E,R)}{P(PT,D|\neg(E,R)) * P(\neg(E,R))} \tag{6}$$

$$= \frac{P(M1)}{P(M2)} \frac{\int_{M_1:m \text{ is valid}} P(m|E,R)P(D|m)dm}{\int_{M_2:m \text{ is invalid}} I_{PT_m}.P(m|\neg(E,R))P(D|m)dm} \tag{7}$$

where $M1$ and $M2$ denote classes of valid-IV and invalid-IV causal models respectively. $P(D|m)$ represents the likelihood of the data given a causal model m. $P(m|E,R)$ and $P(m|\neg(E,R)$ denote the prior probability of any model m within the class of valid-IV and invalid-IV causal models respectively, and I_{PT_m} is an indicator function which is 1 whenever a causal model m generates a data distribution that passes the Pearl-Bonet necessary test.

This is similar to the Bayes Factor[16], except that we are additionally using the result of the Pearl-Bonet necessary test to compute evidence.

The proof of the theorem follows from the structure of the generative meta-model and properties of Pearl's necessary test.

Proof. Let us first consider the ratio of marginal likelihoods of the two model classes.

$$\textit{ML-Ratio} = \frac{P(PT,D|E,R)}{P(PT,D|\neg(E,R))} \tag{8}$$

Since the Pearl's test is a necessary test, we know that $P(PT|E,R) = 1$. Thus, the numerator becomes:

$$\begin{aligned} &P(PT,D|E,R) \\ &\quad = P(PT|D,E,R)P(D|E,R) \\ &\quad = P(D|E,R) \end{aligned} \tag{9}$$

Further, for any causal model m, we know with certainty whether it follows exclusion and as-if-random restrictions. In particular, $P(m_{invalidIV}|E,R)=0$. Using this observation, we can write the numerator of the *ML-Ratio* as:

$$\begin{aligned}
&P(D|E,R)\\
&\quad=\int_m P(D,m|E,R)dm\\
&\quad=\int_m P(m|E,R)P(D|m)dm\\
&\quad=\int_{M_1:m \text{ is valid}} P(m|E,R)P(D|m)dm
\end{aligned} \tag{10}$$

Similarly, the denominator can be expressed by,

$$\begin{aligned}
&P(PT,D|\neg(E,R))\\
&\quad=\int_m P(PT,D,m|\neg(E,R))dm\\
&\quad=\int_m P(m|\neg(E,R))P(PT,D|m,\neg(E,R))dm\\
&\quad=\int_{M_2:m \text{ is invalid}} P(m|\neg(E,R))P(PT,D|m)dm
\end{aligned} \tag{11}$$

where we use the conditional independencies entailed by the generative meta-model. Now given a model m, the result of Pearl-Bonet necessary test PT is deterministic. Therefore, $P(PT|M)$ is 1 whenever the data distribution $P(x,y,z)$ generated by a causal model m passes the test, and 0 otherwise. Assuming that D is a representative sample from data distribution induced by each m, the denominator then simplifies to:

$$\int_{M2:m \text{ is invalid}} P(m|\neg(E,R))P(PT,D|m)dm=\int_{M2:m \text{ is invalid}} P(m|\neg(E,R))P(D|m)I_{PT_m}dm \tag{12}$$

where I_{PT_m} is an indicator function, evaluating to 1 whenever causal model m passes the Pearl-Bonet test.

Combining Equations 10 and 12, we obtain the ratio of marginal likelihoods:

$$\textit{ML-Ratio}=\frac{P(PT,D|E,R)}{P(PT,D|\neg(E,R))}=\frac{\int_{M1:m \text{ is valid}} P(m|E,R)P(D|m)dm}{\int_{M2:m \text{ is invalid}} I_{PT_m}.P(m|\neg(E,R))P(D|m)dm} \tag{13}$$

Finally, by definition of model classes $M1$ and $M2$, they correspond to valid and invalid classes of causal models. Thus,

$$\frac{P(E,R)}{P(\neg(E,R))}=\frac{P(M1)}{P(M2)} \tag{14}$$

The above two equations lead us to the main statement of the theorem:

$$\textit{Validity-Ratio}=\frac{P(M1)}{P(M2)}\frac{\int_{M1:m \text{ is valid}} P(m|E,R)P(D|m)dm}{\int_{M2:m \text{ is invalid}} I_{PT_m}.P(m|\neg(E,R))P(D|m)dm} \tag{15}$$

□

Since the configuration of Exclusion and As-if-random conditions does not provide any more information apart from restricting the class of causal models, we may assume a uninformative uniform prior on causal models given any configuration of these two assumptions. Using a uniform model prior leads to the following Corollary.

Corollary 1. *Using a uniform model prior $P(M1|E,R)$ for valid-IV models, $P(M2|\neg(E,R))$ for invalid-IV models, the Validity-Ratio from Theorem 1 reduces to:*

$$\textit{Validity-Ratio}=\frac{P(M1)}{P(M2)}\frac{K_2\int_{M1:m \textit{ is valid}} P(D|m)dm}{K_1\int_{M2:m \textit{ is invalid}} I_{PT_m}.P(D|m)dm} \tag{16}$$

where K_1 and K_2 are normalization constants.

3.3 NPS Algorithm for testing IVs

Based on the above theorem, we present the NPS algorithm for testing the validity of an instrumental variable below. Assume that the observational dataset contains values for three discrete variables: cause X, outcome Y and a candidate instrument Z.

1. Estimate $P(Y, X|Z)$ using observational data and run the Pearl-Bonet necessary test. If the necessary test fails, Return *REJECT-IV*.
2. Else, compute the Validity-Ratio from Equation 6 for the one or more of the following types of violations (excluding violations that are known to be impossible):
   - **Exclusion may be violated**: $Z \not\perp\!\!\!\perp Y|X, U$
   - **As-if-random may be violated**: $Z \not\perp\!\!\!\perp U$
   - **Both may be violated**:$Z \not\perp\!\!\!\perp Y|X, U; Z \not\perp\!\!\!\perp U$
3. If all Validity Ratios are above a pre-determined threshold γ, then return *ACCEPT-IV*. Else if any Validity Ratio is less than γ^{-1}, then return *REJECT-IV*. Else, return *INCONCLUSIVE*.

Although the NPS algorithm seems straightforward, in practice, the first two steps involve a number of smaller steps that we discuss in the next two sections. Section 4 describes how to compute the Validity Ratio for the three kinds of violations listed above and Section 5 discusses extensions of the Pearl-Bonet test that enable its empirical application to discrete variables.

4 COMPUTING THE VALIDITY RATIO

The key detail in implementing the NPS test is in evaluating the integrals in Equation 6, since there can be infinitely many valid-IV or invalid-IV causal models. In this section we first present the *response variables* framework from Balke and Pearl[17] that provides a finite representation for any non-parametric causal model with discrete X, Y and Z. We extend this framework to also work with invalid-IV causal models. Armed with this characterization, we describe methods for computing the Validity-Ratio in Section 4.2. Note that neither our finite characterization of causal models nor our methods for evaluating the integrals are unique; any other suitable strategy can be used to implement the NPS test.

4.1 The response variable framework

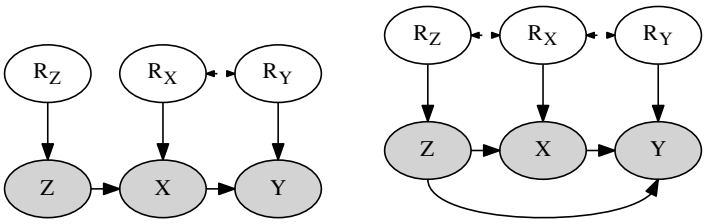
A typical way to characterize causal models is to assume specific functional relationships between observed variables. In most cases, however, the nature of the functional form is not known and thus parameterization in this way arbitrarily restricts the complexity of a causal model. A more general way to make no assumptions on the functional relationships or the unobserved confounders, but rather reason about the space of all possible functions between observed variables. We will use this approach for characterizing valid-IV and invalid-IV causal models.

As an example, suppose we observe the following functional relatioship between Y and X, $y = k(x)$, where the true causal relatioship is $y = f(x, u)$. Conceptually, the variables U are simply additional inputs to this function but it can be hard to reason about them because they are unobserved and may even be unknown. Here we make use of a property of discrete variables that stipulates a finite number of functions between any two discrete variables. When X and Y are discrete, the effect of unobserved confounders can be seen as simply modifying the observed relationship k to another function $k'(x)$. Because the number of possible functions is finite, the combined effect of unobserved confounders U can be characterized by a finite set of parameters. We will call these parameters *response variables*, in line with past work. Note that we make no restriction on U—they can be discrete or continuous—but rather restrict the observed variables to be discrete.

More formally, a response variable acts as a selector on a non-deterministic function and converts it into a a set of deterministic functions, indexed by the response variable. Depending on the value of the response variable, one of the deterministic functions is invoked. Under this transformation, the response variables become the only random variables in the system, and therefore any causal model can be expressed as a probability distribution over the response variables.

4.1.1 Response variables framework for valid-IV models

Let us first construct response variables for a valid-IV model. To do so, we will transform U to a different *response variable* for each observed variable in the causal model. For ease of exposition, we will assume that Z, X and Y are binary variables; however, the analysis follows through for any number of discrete levels.



(a) Valid IV **(b)** Invalid IV

Figure 4. Causal graphical model with response variables denoting the effect of unknown, unobserved U.

For valid-IV causal models, we can write the following structural equations for observed variables X, Y and Z (from Equation 1).

$$\begin{aligned} y &= f(x, u_y) \\ y &= g(z, u_x) \\ z &= h(u_z) \end{aligned} \tag{17}$$

where U_x, U_y and U_z represent error terms. U_x and U_y are correlated. As-if-random condition ($Z \perp\!\!\!\perp U$) stipulates that $U_z \perp\!\!\!\perp \{U_y, U_x\}$. Exclusion condition is satisfied because function f does not depend on z.

Since there are a finite number of functions between discrete variables, we can represent the effect of unknown confounders U as a selection over those functions, indexed by a variable known as a response variable. For example, in Equation 17, Y can be written as a combination of 4 deterministic functions of x, after introducing a response variable, r_y.

$$y = \begin{cases} f_{ry_0}(x) \equiv 0, & \text{if } r_y = 0 \\ f_{ry_1}(x) \equiv x, & \text{if } r_y = 1 \\ f_{ry_2}(x) \equiv \tilde{x}, & \text{if } r_y = 2 \\ f_{ry_3}(x) \equiv 1, & \text{if } r_y = 3 \end{cases} \tag{18}$$

That is, different values of U change the value of Y from what it would have been without U's effect, which we capture through r_y. Intuitively, these r_y refer to different ways in which individuals may respond to the treatment X. Some may have no effect irrespective of treatment($r_y = 0$), some may only have an effect when $X = 1$ ($r_y = 1$), some may only have an effect when X=0 ($r_y = 2$), while others would always have an effect irrespective of X ($r_y = 3$). Such response behavior, as denoted by $r_y = \{0, 1, 2, 3\}$, is analogous to *never-recover*, *helped*, *hurt*, and *always-recover* behavior to treatment in past work[18].

Similarly, we can write a deterministic functional form for x, leading to the transformed causal diagram with response variables in Figure 4.

$$x = \begin{cases} g_{rx_0}(z) \equiv 0, & \text{if } r_x = 0 \\ g_{rx_1}(z) \equiv z, & \text{if } r_x = 1 \\ g_{rx_2}(z) \equiv \tilde{z}, & \text{if } r_x = 2 \\ g_{rx_3}(z) \equiv 1, & \text{if } r_x = 3 \end{cases} \tag{19}$$

Similar to r_y, $r_x = \{0, 1, 2, 3\}$ can be interpreted in terms of a subject's compliance behavior: *never-taker*, *complier*, *defier*, and *always-taker*[2].

Finally, z can be assumed to be generated by its own response variable, r_z.

$$z = \begin{cases} 0, & \text{if } r_z = 0 \\ 1, & \text{if } r_z = 1 \end{cases} \tag{20}$$

Trivially, $Z = R_Z$.

Given this framework, a specific value of the joint probability distribution $P(r_z, r_x, r_y)$ defines a specific, valid causal model for an instrument Z. Exclusion condition is satisfied because the structural equation for y does not depend on Z. For as-if-random condition, we additionally require that $U_z \perp\!\!\!\perp \{U_x, U_y\}$. Since R_X and R_Y represent the effect of U as shown in

Figure 4a, the as-if-random condition translates to $R_Z \perp\!\!\!\perp \{R_X, R_Y\}$, implying that $P(R_Z, R_X, R_Y) = P(R_Z)P(R_X, R_Y)$. Using this joint probability distribution over r_z, r_x, and r_y, any valid-IV causal model for x, y and z can be parameterized. For instance, when all three variables are binary, R_Z, R_X and R_Y will be 2-level, 4-level and 4-level discrete variables respectively. Therefore, each unique causal model can be represented by 2+4x4=18 dimensional probability vector θ where each $\theta_i \in [0, 1]$. In general, for discrete-valued Z, X and Y with levels l, m and n respectively, θ will be a $(l + m^l n^m)$-dimensional vector.

4.1.2 Response variable framework for invalid IVs

While past work only considered Valid-IV models, we now show that the same framework can also be used to represent invalid-IV models. As defined in Section 3.1, a causal model is invalid when either of the IV conditions is violated: Exclusion or As-if-random.

Exclusion is violated

When exclusion is violated, it is no longer true that $Z \perp\!\!\!\perp Y|X, U$. This implies that Z may affect Y directly. To account for this, we redefine the structural equation for Y to depend on both Z and X: $y = h(X, Z)$. This corresponds to adding a direct arrow from Z to Y as shown in Figure 4b. In response variables framework, this translates to:

$$y = \begin{cases} h_{ry_0}(x, z) & \text{if } r_y = 0 \\ h_{ry_1}(x, z) & \text{if } r_y = 1 \\ h_{ry_2}(x, z) & \text{if } r_y = 2 \\ \dots & \\ h_{ry_{15}}(x, z) & \text{if } r_y = 15 \end{cases} \tag{21}$$

where R_Y now has 16 discrete levels, each corresponding to a deterministic function from the tuple (x, z) to y.

As with valid-IV causal models, any invalid-IV causal model can be denoted by a probability vector for $P(R_Z)$ and $P(R_X, R_Y)$. However, the dimensions of the probability vector will increase based on the extent of Exclusion violation. For full exclusion, dimensions will be $l + m^l n^{lm}$.

As-if-random is violated

The violation of as-if-random does not change the structural equations, but it changes the dependence between R_Z and (R_X, R_Y). If as-if-random assumption does not hold, then R_Z is no longer independent of (R_X, R_Y). Therefore, we can no longer decompose $P(R_Z, R_X, R_Y)$ as the product of independent distributions $P(R_z)$ and $P(R_X, R_Y)$ and dimensions of θ will be $lm^l n^m$.

4.2 Computing marginal likelihood for Valid-IV and Invalid-IV models

Based on the response variable framework, choosing an individual causal model is equivalent to sampling a probability vector θ from the joint probability distribution $P(r_x, r_y, r_z)$. The dimensions of this probability vector will vary based on the extent of violations of the instrumental variable conditions, from $l + m^l n^m$ for the valid-IV model to $lm^l n^{lm}$ for invalid-IV model in which both exclusion and as-if-random conditions are violated. Below we describe two ways for computing the validity ratio: first for an average dataset and then for a specific observed dataset.

4.2.1 Assuming uniform likelihood for all causal models

To examine the power of the NPS test in general, we may be interested in how well the test distinguishes valid-IV from invalid-IV models. One way is to simulate the test when all causal models are equally likely, thereby not being dependent on any specific dataset. Another way of interpreting the assumption of uniform likelihood of causal models is that we are integrating over all possible datasets to compute an *Overall-Validity-Ratio*, as shown below:

$$\begin{aligned} \textit{Overall-Validity-Ratio} &= \frac{P(M1)}{P(M2)} \frac{\int_D \int_{M_1:m \text{ is valid}} P(m|E, R)P(D|m)dm.dD}{\int_D \int_{M_2:m \text{ is invalid}} I_{PT_m}.P(m|\neg(E, R))P(D|m)dm.dD} \\ &= \frac{P(M1)}{P(M2)} \frac{\int_{M_1:m \text{ is valid}} P(m|E, R)dm}{\int_{M_2:m \text{ is invalid}} I_{PT_m}.P(m|\neg(E, R))dm} \end{aligned} \tag{22}$$

We can use Monte Carlo simulation with S iterations to approximate the above integrals, using a uniform dirichlet prior on the probability vector θ. The dimensions of the θ vector and the exact sampling strategy depends on the presumed configuration of Exclusion and As-if-random conditions.

Valid-IV: Both conditions are satisfied

When both Exclusion and As-if-random conditions are satisfied, R_Y is a 4-level discrete variable as in Equation 18. Because $R_Z \perp\!\!\!\perp \{R_X, R_Y\}$, we sample θ_{R_Z} independently and separately sample the joint distribution vector of θ_{R_X,R_Y}.

Invalid-IV: At least one of the conditions is violated

When as-if-random assumption is not violated (such as when z is randomized), we sample θ_{R_Z} independently as for a Valid-IV model. However, since Exclusion may be violated, θ_{R_X,R_Y} will be a 4x16-level discrete variable as in Equation 21.

Otherwise, if Exclusion is not violated, then θ_{R_Y} remains a 4-level discrete variable. However, R_Z is no longer independent of (R_X, R_Y) because as-if-random may be violated. Therefore, we will sample a joint probability distribution vector θ_{R_Z,R_X,R_Y} from all possible joint probability vectors. Since our goal is to only generate invalid IV models, we reject any generated probability vector where R_Z turns out to be independent of R_X, R_Y.

When both conditions are violated, R_Y is a 16-level discrete variable and R_Z is not independent of (R_X, R_Y). Thus, we sample a (2x4x16)-dimensional probability vector θ_{R_Z,R_X,R_Y}.

4.2.2 Estimating Validity Ratio for a specific dataset

The above procedure will demonstrate the general power of the test, but will not tell us the likelihood of a specific observed dataset to be generated from a valid-IV model. To compute this, we return to Equation 1.

$$\textit{Validity-Ratio} = \frac{P(M1)}{P(M2)} \frac{K_2 \int_{M1:m \text{ is valid}} P(D|m)dm}{K_1 \int_{M2:m \text{ is invalid}} I_{PT_m}.P(D|m)dm} \tag{23}$$

To compute the integrals in the numerator and the denominator of the above expression, we utilize the fact that there can be a finite number of unique observed data points (Z, X, Y) when all three variables are discrete. For example, for binary Z, X and Y, there can be 2x2x2=8 unique observations. In general, the number of unique data points is lmn. Making the standard assumption of independent data points, we obtain the following likelihood for any causal model m,

$$\begin{aligned} P(D|m) &= \prod_{i=1}^{N} P(Z = z^{(i)}, X = x^{(i)}, Y = y^{(i)}|m) \\ &= (P(Z = 0, X = 0, Y = 0|m))^{R_0} ... P(Z = z_l, X = x_m, Y = y_n|m))^{R_{lmn}} \\ &= \prod_{j=1}^{Q} (P(Z = z_j, X = x_j, Y = y_j|m))^{Q_j} \end{aligned} \tag{24}$$

where N is the number of observed (z, x, y) data points and Q_j the number of times each unique value of (z, x, y) repeats in the dataset. Sincethe model m can be equivalently represented by its probability vector θ_{R_Z,R_X,R_Y}, we can rewrite the above equation as:

$$\begin{aligned} P(D|m) = P(D|\theta) &= \prod_{j=1}^{Q} (P(Z = z_j, X = x_j, Y = y_j|\theta))^{Q_j} \\ &= \prod_{j=1}^{Q} (\sum_{r_{zxy}=000}^{lmn} P(R_{XYZ} = r_{xyz}) P(Z = z_j, X = x_j, Y = y_j|\theta, r_{zxy}))^{Q_j} \end{aligned} \tag{25}$$

As before, we will illustrate closed form expressions for binary variables, but the following derivations follow through for any number of discrete levels.

4.2.3 Calculating the numerator

When both conditions are satisfied, the numerator can be written as:

$$\begin{aligned} P(D|\theta) &= \prod_{j=1}^{Q} (\sum_{r_{zxy}='000'}^{'133'} P(R_Z = r_Z)(P(Z = z_j|\theta, r_z) P(R_{XY} = r_{xy})(P(X = x_j, Y = y_j|\theta, r_{zxy}))^{Q_j} \\ &= \prod_{j=1}^{Q} (\sum_{r_z='0'}^{'1'} P(R_Z = r_Z)(P(Z = z_j|\theta, r_z))^{Q_j} (\sum_{r_{xy}='00'}^{'33'} P(R_{XY} = r_{xy})(P(X = x_j, Y = y_j|\theta, r_{xy}))^{Q_j} \end{aligned} \tag{26}$$

Note that for a fixed value of R_Z, Z can be uniquely determined. Similarly, for a given value of Z and R_{XY}, X and Y can be deterministically evaluated. Thus, the above expression reduces to:

$$
\begin{aligned}
P(D|\theta) &= \prod_{j=1}^{Q}\Big(\sum_{r_z='0'}^{'1'}\theta_{r_z}\Big)^{Q_j}\Big(\sum_{r_{xy}='00'}^{'33'}\theta_{r_{xy}}\Big)^{Q_j} \\
&= \theta_{rz=0}^{Q_0}(\theta_{rxy=00}+\theta_{rxy=20}+\theta_{rxy=02}+\theta_{rxy=22})^{Q_0} \\
&\quad \theta_{rz=0}^{Q_1}(\theta_{rxy=01}+\theta_{rxy=21}+\theta_{rxy=03}+\theta_{rxy=23})^{Q_1} \\
&\quad \theta_{rz=0}^{Q_2}(\theta_{rxy=11}+\theta_{rxy=10}+\theta_{rxy=31}+\theta_{rxy=30})^{Q_2} \\
&\quad \theta_{rz=0}^{Q_3}(\theta_{rxy=12}+\theta_{rxy=13}+\theta_{rxy=32}+\theta_{rxy=33})^{Q_3} \\
&\quad \theta_{rz=1}^{Q_4}(\theta_{rxy=00}+\theta_{rxy=02}+\theta_{rxy=10}+\theta_{rxy=12})^{Q_4} \\
&\quad \theta_{rz=1}^{Q_5}(\theta_{rxy=01}+\theta_{rxy=03}+\theta_{rxy=11}+\theta_{rxy=13})^{Q_5} \\
&\quad \theta_{rz=1}^{Q_6}(\theta_{rxy=20}+\theta_{rxy=30}+\theta_{rxy=21}+\theta_{rxy=31})^{Q_6} \\
&\quad \theta_{rz=1}^{Q_7}(\theta_{rxy=22}+\theta_{rxy=32}+\theta_{rxy=23}+\theta_{rxy=33})^{Q_7}
\end{aligned} \tag{27}
$$

The above equation leads to the following simplification for the numerator of Equation 23.

$$
\begin{aligned}
\int_{M1:m \text{ is valid}} P(D|m)dm &= \iint_{\theta_{rz},\theta_{rxy}} \prod_{j=1}^{Q}\theta_{rz=z}^{Q_j}(\theta_{rxy=a}+\theta_{rxy=b}+\theta_{rxy=c}+\theta_{rxy=d})^{Q_j}d\theta_{rz}d\theta_{rxy} \\
&= \int\prod_{j=1}^{Q}\theta_{rz=z}^{Q_j}d\theta_{rz}\int\prod_{j=1}^{Q}(\theta_{rxy=a}+\theta_{rxy=b}+\theta_{rxy=c}+\theta_{rxy=d})^{Q_j}d\theta_{rxy}
\end{aligned} \tag{28}
$$

The above integral has a form equivalent to the hyperdirichlet integral[19], for which no tractable closed form exists except in a few special cases.[2] We therefore resort to approximate methods for estimating the integral. For binary X, Y and Z, the maximum dimension of the integral will be 16, so we recommend using approximate integral techniques over the unit simplex.[21] For discrete variables, monte carlo methods for estimating marginal likelihood, such as annealed importance sampling[22] or nested sampling[23,24] may be more appropriate.

4.2.4 Calculating the denominator

We can calculate the denominator of Equation 23 in a similar way as the numerator, except that the exact integral expression will vary based on the extent of violation of as-if-random and exclusion restrictions.

When Exclusion is violated

Following Equations 23, the denominator can be expressed similarly to 26. The only difference is that θ_{rxy} will be 4x16=64-dimensional integral.

$$
P(D|\theta) = \prod_{j=1}^{Q}\Big(\sum_{r_z='0'}^{'1'}P(R_Z=r_Z)(P(Z=z_j|\theta,r_z))^{Q_j}\Big(\sum_{r_{xy}='0,0'}^{'3,15'}P(R_{XY}=r_{xy})(P(X=x_j,Y=y_j|\theta,r_{xy}))^{Q_j}
\tag{29}
$$

However, the above formulation corresponds to a full violation of the exclusion condition, assuming all $\theta ry \in \{0,15\} \setminus \{0,3,12,15\}$ are non-zero. In practice, exclusion can be violated even if a single R_Y in that set is non-zero. Therefore, a stronger and realistic way of estimating marginal likelihood under an invalid-IV is to compute the maximum of all marginal likelihoods for causal models where one of the R_Y corresponding to an exclusion violation is non-zero. This would mean computing 12 integrations over 4x5=20 dimensions, one for each for nonzero R_Y that results in an exclusion violation.

[2] We say *tractable* because it is possible to decompose the hyperdirichlet integral into a sum of exponentially many dirichlet integrals[20], but that will not be computationally feasible.

Algorithm 1: NPS Algorithm

Data: Observed tuples (Z, X, Y), Prior-Ratio=$P(M1)/P(M2)$
Result: Validity Ratio for comparing invalid and valid
Select appropriate subclass of invalid-IV models based on domain knowledge about the validity of IV conditions. ;

- **Only Exclusion may be violated**: Assume $y = h(x, z, u)$. Sample $P(r_z)$, $P(r_x, r_y)$ separately. Use Equation 29 to compute marginal likelihood M_{EXCL}.
- **Only As-if-random may be violated**: Assume $y = f(x, u)$. Sample $P(r_z, r_x, r_y)$ as a joint distribution. Use Equation 31 to compute marginal likelihood M_{AIR}.
- **Both conditions may be violated**: Assume $y = h(x, z, u)$. Sample $P(r_z, r_x, r_y)$ as a joint distribution. Use Equation 32 to compute marginal likelihood $M_{AIR,EXCL}$

Compute marginal likelihood of invalid-IV models as $ML_{INVALID} = max(M_{EXCL}, M_{AIR}, M_{AIR,EXCL})$;
Compute marginal likelihood of valid-IV models using Equation 26, assuming $y = f(x, u)$ and sample $P(r_z)$, $P(r_x, r_y)$ separately ;
Compute Validity Ratio as $ML_{VALID}/ML_{INVALID} * \textit{PRIOR-RATIO}$

Figure 5. NPS Algorithm for validating an instrumental variable.

When As-if-random is violated

Fortunately, when as-if-random condition is violated, we can obtain a closed form solution for the integral. Proceeding from Equation 25, we cannot simplify the marginal likelihood as a product of two independent integrals and thus obtain:

$$\int P(D|\theta)d\theta = \int_{M2} \prod_{j=1}^{Q} (\sum_{r_{zxy}=0,0,0}^{1,3,3} P(R_{XYZ} = r_{xyz}) P(Z = z_j, X = x_j, Y = y_j|\theta, r_{zxy}))^{Q_j} \tag{30}$$

This, however, means that each θ_{rzxy} occurs exactly once in the integral, allowing a transformation of the integral to a dirichlet integral. The complete derivation is in the Appendix; the closed form integral is given by,

$$\int P(D|\theta)d\theta = \frac{\prod_{j=1}^{Q} \Gamma(4 + Q_j)}{(\Gamma(4))^Q \Gamma(\sum_{j=1}^{Q} \Gamma(4 + Q_j)))} \tag{31}$$

where $\Gamma(n) = factorial(n - 1)$ is the Gamma function.

When both are violated

When both exclusion and as-if-random conditions are violated, we can again obtain a closed form solution. The integral expression is similar to that for as-if-random violation (Equation 30), except that the number of dimensions of theta increases to 2x4x16=128. The denominator of the Validity Ratio can be evaluated as:

$$\begin{aligned} \int P(D|\theta)d\theta &= \int_{M2} \prod_{j=1}^{Q} (\sum_{r_{zxy}=0,0,0}^{1,3,15} P(R_{XYZ} = r_{xyz}) P(Z = z_j, X = x_j, Y = y_j|\theta, r_{zxy}))^{Q_j} \\ &= \frac{\prod_{j=1}^{Q} \Gamma(16 + Q_j)}{(\Gamma(16))^Q \Gamma(\sum_{j=1}^{Q} \Gamma(16 + Q_j)))} \end{aligned} \tag{32}$$

In the rest of the paper, we use a non-informative uniform prior for $P(M|E, R)$ and $P(M|\neg(E, R))$ using Corollary 1. Based on the above discussion, Algorithm 1 summarizes the algorithm for computing validity ratio for any observed dataset.

5 EXTENSIONS TO THE PEARL-BONET TEST

In this section we describe extensions to the Pearl-Bonet test that are required for empirical application of the test for discrete variables. First, we present an efficient way to evaluate the necessary test for any number of discrete levels. Second, we show how to extend the monotonicity condition to more than two levels. Third, we discuss how to use the test in finite samples, by utilizing an exact test proposed by Wang et al..

5.1 Implementing Pearl-Bonet test for discrete variables

Specifying a closed form for the necessary test becomes complicated when we generalize from binary to discrete variables. Bonet[6] showed that Pearl's instrumental inequalities for binary variables do not satisfy the existence requirement from Section 2 and more inequalities are needed. Further, it is not always feasible to construct analytically all the necessary inequalities for discrete variables.

To derive a practical test for IVs with discrete variables, we employ Bonet's framework[6] that specifies Valid-IV and Invalid-IV class of causal models as convex polytopes in multi-dimensional probability space. In Figure 2, we showed a schematic of Bonet's framework, representing Valid-IV and Invalid-IV classes as polygons on a 2D surface. We now make these notions precise. The axes represent different dimensions of the probability vector $f = P(X, Y|Z)$. Assuming l discrete levels for Z, n for X and m for Y, f will be a lnm dimension vector, given by:

$$\begin{aligned} f =&(P(X = x_1, Y = y_1|Z = z_1),\\ &P(X = x_1, Y = y_2|Z = z_1), ...\\ &P(X = x_1, Y = y_m|Z = z_1),\\ &P(X = x_1, Y = y_1|Z = z_2), ...\\ &P(X = x_n, Y = y_m|Z = z_l)) \end{aligned} \tag{33}$$

U may be either discrete or continuous, we do not impose any restrictions on unobserved variables. Any observed probability distribution over Z, X and Y can be expressed as a point in this lmn-dimension space. Since $\sum_{i,j} P(X = x_i, Y = y_j|Z = z_k) = 1 \forall k \in \{1...l\}$, the extreme points of for valid probability distributions $P(X, Y|Z)$ are given by $P(X = x_i, Y = y_j|Z = z_k) = 1$. We showed a square region as the set of all valid probability distributions in Figure 2, but more generally the region constitutes a lmn-dimensional convex polytope $\mathcal{F}$[6].

Based on the models defined in Figure 1, the set of all valid probability distributions $\mathcal{F}$ can be generated by Invalid-IV class of models. Within that set, we are interested in the probability distributions that can be generated by a valid-IV model. Knowing this subset provides a necessary test for instrumental variables; any observed data distribution that cannot be generated from a valid-IV model fails the test. Bonet showed that such a subset forms another convex polytope $\mathcal{B}$ (the Valid-IV region in Figure 2) whose extreme points can be enumerated analytically. Based on this result, we provide a practical implementation for a necessary test for discrete IVs.

Theorem 2. *Given data on discrete variables Z, X and Y, their observational probability vector $f \in \mathcal{F}$, and extreme points of the polytope B containing distributions generatable from a Valid-IV model, the following linear program serves as a necessary and existence test for instrumental variables:*

$$\sum_{k=1}^{K} \lambda_k . \vec{e_k} = \vec{f}; \qquad \sum_{j=1}^{K} \lambda_j = 1; \qquad \lambda_j \geq 0 \; \forall j \in [1, K] \tag{34}$$

where $e_1, e_2 ... e_K$ represents lmn-dimensional extreme points of B and $\lambda_1, \lambda_2 .. \lambda_K$ are non-negative real numbers. If the linear program has no solution, then the data distribution cannot be generated from a Valid-IV causal model.

Proof. The proof is based on properties of a convex polytope, which is also a convex set. A point lies inside a convex polytope if it can be expressed as a linear combination of the polytope's extreme points. Therefore, an observed data distribution could not have been generated from a Valid IV model if there is no real-valued solution to Equation 34. □

While the test works for any discrete variables, in practice the test becomes computationally prohibitive for variables with large number of discrete levels, because the number of extreme points K grows exponentially with l, m and n. If the number of discrete levels is large, we can an entropy-based approximation instead, as in[25].

5.2 Extending Pearl-Bonet test to include Monotonicity

Monotonicity is a common assumption made in instrumental variables studies, so it will be useful to extend the necessary test for discrete variables when monotonicity holds. No prior necessary test for monotonicity exists for discrete variables with more than two levels, so here we propose a test for monotonicity that can be used in conjunction with Theorem 2.

As defined in Section 2.2, monotonicity implies that:

$$g(z_1, u) \geq g(z_2, u) \textit{ whenever } z_1 \geq z_2 \tag{35}$$

That is, increasing Z can cause X to either increase or stay constant, but never decrease. Note that the above definition is without any loss of generality. In case Z has a negative effect on X, we can do a simple transformation by inverting the discrete levels on Z so that Equation 35 holds.

By requiring this constraint on the structural equation between X and Z, monotonicity restricts the observed data distribution. We use this observation to test for monotonicity.

Theorem 3. *For any data distribution $P(X, Y, Z)$ generated from a valid-IV model that also satisfies monotonicity, the following inequalities hold:*

$$\begin{aligned} P(Y=y, X\geq x|Z=z_0) &\leq P(Y=y, X\geq x|Z=z_1) \quad ... \quad \leq P(Y=y, X\geq x|Z=z_{l-1}) \quad \forall x, y \\ P(Y=y, X\leq x|Z=z_0) &\geq P(Y=y, X\leq x|Z=z_1) \quad ... \quad \geq P(Y=y, X\leq x|Z=z_{l-1}) \quad \forall x, y \end{aligned} \tag{36}$$

where Z, X and Y are ordered discrete variables of levels l, n and m respectively and $z_0 \leq z_1 ... \leq z_{l-1}$.

Proof. Consider the first set of inequalities with $P(Y=y, X\geq x|Z=z_k)$, for some $X=x$ and $Y=y$. Based on the structure of a Valid-IV causal model (Figure 1), we can factorize $P(Y, X|Z)$ as:

$$P(Y=y, X\geq x|Z=z) = P(X\geq x|Z=z)P(Y=y|X=x, Z=z) = P(X\geq x|Z=z)P(Y=y|X\geq x)$$

$P(Y=y|X\geq x)$ is independent of Z. Therefore, as Z varies, $P(Y=y, X\geq x|Z=z)$ only depends on $P(X\geq x, Z=z)$.

Using the structural equations for $x=g(z,u)$ from Equation 19, we obtain for any x and $z\in\{z_0, z_1, ...z_{l-1}\}$:

$$P(X\geq x|Z=z_k) = P(R_X : g(z_k, u)\geq x) \tag{37}$$

By monotonicity, we know that $g(z_{k2}, u)\geq g(z_{k1}, u)$ whenever $z_{k2}\geq z_{k1}$. Thus, we can write:

$$g(z_{k1}, u)\geq x \Rightarrow g(z_{k2}, u)\geq x \qquad \text{if } z_{k2}\geq z_{k1} \tag{38}$$

Combining Equations 37 and 38, for any k, we can argue that the set of response variables r_x that satisfy $g(z_k, u)\geq x$ will always be smaller than the set of response variables that satisfy $g(z_{k+1}, u)\geq x$. Therefore, we obtain the following inequality:

$$P(X\geq x|Z=z_k) = P(R_X : g(z_k, u)\geq x) \leq P(R_X : g(z_{k+1}, u)\geq x) = P(X\geq x|Z=z_{k+1})$$

Iterating over $k\in\{0, 1..., l-1\}$ will provide us the first set of inequalities stated in the Theorem. We can follow a similar reasoning to derive the second set of inequalities with $P(Y=y, X\leq x|Z=z_k)$.

□

For binary variables, Theorem 3 reduces to $P(Y=y, X=1|Z=z_0)\leq P(Y=y, X=1|Z=z_1)$ and $P(Y=y, X=0|Z=z_0)\geq P(Y=y, X=0|Z=z_1)$, same as Equation 3.

5.3 Finite sample testing for Pearl-Bonet test

Finally, Pearl-Bonet test assumes that we can infer conditional probabilities $P(Y, X|Z)$ accurately. However, in any finite observed sample, we will only be able to compute a sample probability estimate. Therefore, we need a statistical test that accounts for the finite sample properties of any observed dataset. There are many tests proposed to deal with finite samples.[13,14,26,27] In this paper we use an exact test proposed by Wang et al.[27], both for its simplicity and because it makes no assumptions about the data-generating causal models. This test converts the inequalities of the necessary test into a version of one-tailed Fisher's exact test. As with all null hypothesis tests, the goal is to refute the null hypothesis. Here the null hypothesis is that the conditional probability distribution satisfies the inequalities of the Pearl-Bonet test. We then quantify the likelihood of seeing the obseved data under this null hypothesis, thus providing us with a p-value for the test. Because we are testing 4 inequalities at once, our analysis can be prone to multiple comparisons. Therefore, Wang et al. recommend a significance level of $\alpha/2$ for each test, where α is the desired significance level.

However, this test does not work under monotonicity assumption. We extend their method for monotonicity, by using the same transformation to convert monotonicity inequalities to the Fisher's exact test. Again, to prevent false positives due to multiple comparisons, it would be ideal to choose a smaller significance level for each inequality, but the results we present are without any correction.

6 SIMULATIONS: HOW POWERFUL IS THE NPS TEST?

We now report simulation results for the NPS test. The goal of this simulation exercise is to determine the power of the NPS test: if the observed data is generated from an invalid-IV model, we want to estimate the probability that NPS test rejects the observed data. To obtain the power of the test in general, we assume a uniform likelihood for all causal models, as described in Section 4.2.1. For any real dataset, the power of the test may be lower or higher than the average case reported here, based

on how easily distinguishable likely invalid-IV and valid-IV models are. We will compute dataset-dependent estimates in Section 7.

As described in Section 4.2.1, we will conduct separate simulations for three kinds of violations of the IV conditions: exclusion only, as-if-random only, and both violated. For these simulations, we assume that Z, X and Y are all binary variables. Further, since monotonicity is a required assumption for obtaining the local average causal effect[12], we also assume monotonicity throughout. To compute the Overall-Validity-Ratio, we use Equation 22 and assume uniform likelihood and prior for causal models. Under these conditions, the Validity-Ratio simplifies to,

$$\textit{Validity-Ratio} = \frac{P(PT, D|E, R)}{P(PT, D|\neg(E, R))} = 1/\sum_{j:M_j \text{ is invalid}} I_{PT_j}.P(M_j|\neg(E, R)) \tag{39}$$

$$= 1/\sum_{j:M_j \text{ is invalid}} I_{PT_j}/N \tag{40}$$

where N is the number of invalid-IV causal models sampled. For each kind of invalid-IV model class, we set $N = 200$, sample N causal models as in Section 4.1.2 and compute the fraction of models that pass the Pearl-Bonet necessary test. The fraction of invalid-IV models that pass the Pearl-Bonet test gives $P(PT, D|\neg(E, R))$ and can be interpreted as the inverse of the Bayesian Validity Ratio.

Because we are directly testing causal models that we sampled, we can compute the probability distributions exactly and thus do not need any finite sample modification for the Pearl-Bonet test. Following guidelines from Kass et al.[16], we declare an instrument as valid if the Validity Ratio is at least 20.[3] Therefore, when the Validity Ratio is at least 20 (or equivalently, when the denominator is at most 0.05), the NPS test is able to declare an instrument valid if it passes the Pearl-Bonet test. If not, the NPS test will be inconclusive.

Although the NPS test can tell us whether an instrument is likely to be valid or not, it does not say anything about the bias in the resulting IV estimate. It could be possible that an instrument is invalid in the strict sense defined above, but still provides causal estimates with low bias. Therefore, besides checking IV validity, we will also estimate the bias incurred when providing causal estimates from an invalid causal model. To compute the causal effect $X \to Y$, we use the Wald estimator[28], which for binary variables, can be written as[11]:

$$\hat{W} = \frac{P(Y = 1|Z = 1) - P(Y = 1|Z = 0)}{P(X = 1|Z = 1) - P(X = 1|Z = 0)}$$

Because we know that the causal effect between binary variables ranges from $[-1, 1]$, we bound the estimate within this interval. We will also compare the power of the NPS test at different values of instrument strength, where the denominator of the Wald estimator can be used as an estimate of the strength of the instrument.

6.1 Exclusion may be violated

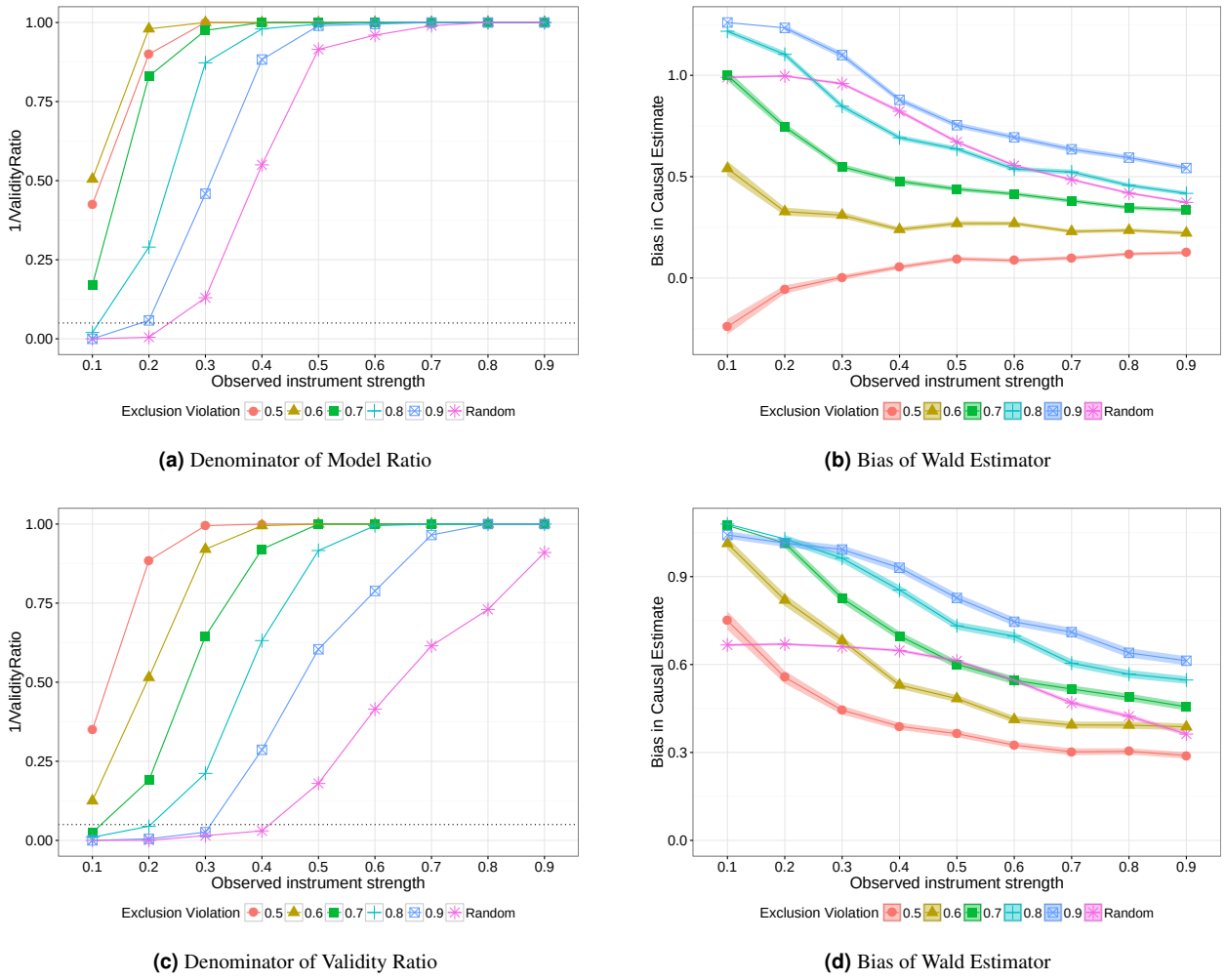
We first test for violation of exclusion only. That is, we assume that as-if-random is satisfied (e.g. through random assigment). As described in Section 4.1.2, violation of the exclusion restriction implies that $y = f(z, x, u)$ and thus there can be 16 different r_y levels. Following the NPS Algorithm, we sample r_x and r_y jointly and sample r_z independently.

The remaining detail is how to sample invalid-IV models that violate exclusion condition. This is non-trival because the degree of violation of the exclusion restriction can vary based on known properties of the underlying causal model. We take one of the weakest properties of the unobserved true causal model—the direction of effect from Z or X to Y—and characterize the power of the NPS test as we vary this property. First we will consider a scenario where either of Z or X has a non-decreasing effect on Y and then weaken this requirement to obtain a more general scenario. For instance, in many empirical studies, we may know Z or X have a non-decreasing effect on Y because there is no plausible mechanism that leads to a decrease in Y with increase in Z or X (common in *encouragement design* studies). However, in many other cases, completely ruling out violations that do not follow the non-decreasing property is justifiably harsh. We therefore relax this restriction and instead stipulate the percentage of data points where this restriction is satisfied. Driving this percentage down to 50% essentially provides the general case, where the direction of the effect is equally likely to be positive or negative.

6.1.1 Instrument assumed to have a non-decreasing effect on Y

We first study the power of the NPS test—the probability of classifying an instrument as valid when the true causal model is also valid-IV—by varying the extent to which Z has a non-decreasing effect on Y. Figure 6a shows the inverse of the

[3] The threshold of 20 can also justified with a frequentist interpretation. Because we are assuming a uniform likelihood for all causal models, the denominator of the Validity Ratio can be considered as equivalent to conducting a null hypothesis test where the null hypothesis is that the true causal model is invalid. Then, the fraction of invalid causal models that pass the test provide a p-value for the effectiveness of the Pearl-Bonet test in identifying a valid instrument. When this p-value is 0.05, the Validity Ratio will be $1/0.05 = 20$.



(a) Denominator of Model Ratio

(b) Bias of Wald Estimator

(c) Denominator of Validity Ratio

(d) Bias of Wald Estimator

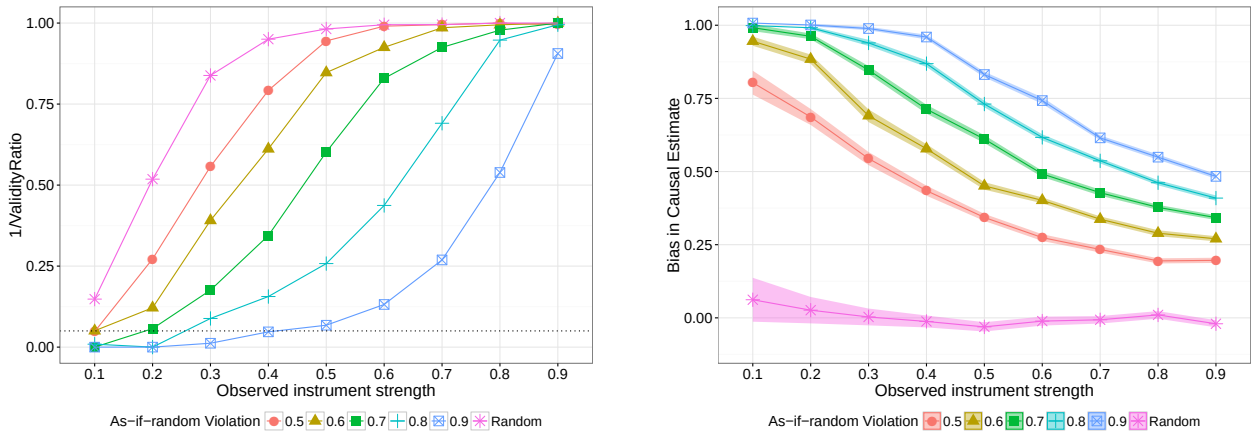
Figure 6. Testing for the exclusion condition. Top panel corresponds to the scenario where Z has a non-decreasing effect on Y, and the bottom panel corresponds to the scenario where both Z and X have a non-decreasing effect on Y. In both panels, the left subfigure shows the inverse of Validity Ratio and the right subfigure shows bias of the Wald IV estimator, as the observed strength of the instrument is varied.

Validity-Ratio as a function of the observed instrument strength. Each of the different lines shows a different fraction of data points that satisfy the non-decreasing effect criterion. When Z can be assumed to have a non-decreasing effect on Y, we can correctly identify a violation of the Exclusion condition with an error rate of less than 5% until an instrument strength of 0.2. An error rate of 5% corresponds to a Validity Ratio of 20, which we consider as the decision threshold for the NPS test. As instrument strength increases, the power of the NPS test decreases. Further, the power of the NPS test decreases as the fraction of data points satisfying the non-decreasing effect $Z \rightarrow Y$ decrease. At a fraction of 0.8, NPS test is only able to identify correctly invalid-IV models with less than 5% error for instruments with strength less than 0.1.

Contrasting these results with estimated bias in the Wald estimate of the causal effect $X \rightarrow Y$ provides more context to the results. Even when the NPS test is unable to detect violation of exclusion, it is also likely that the bias is relatively low (Figure 6b). The magnitude of the bias is larger for weak instruments and for high fractions of data points satisfying the non-decreasing effect, both scenarios where the NPS test has the highest power. When the observed strength of the instrument is high, even clearly invalid-IV models lead to lower bias (less than 0.6).

6.1.2 Both instrument and cause assumed to have a non-decreasing effect on Y

In some IV studies, we may know that both instrument Z and cause X have a non-decreasing effect on the outcome. In such cases, we can strengthen the above assumption by assuming that both Z and X have a non-decreasing effect on Y. Figure 6c shows the fraction of invalid-IV models that pass the Pearl-Bonet test under these conditions. When the non-decreasing



(a) Denominator of Validity Ratio

(b) Bias of Wald Estimator

Figure 7. Testing for the as-if-random condition. Varying the mutual information between R_Z-R_Y. As the severity of as-if-random violation—mutual information—is increased, power of the NPS test increases and so does bias in the resultant IV estimate.

condition is satisfied strictly—that is, there are no data points with an increasing effect of either Z or X on Y—the conventional 5% significance level for false positives is reached up to a maximum instrument strength of 0.4. The power of the NPS test for other scenarios also increases. For thresholds of non-decreasing effect at least 0.7, fraction of false positives lies around 5% at instrument strength of 0.1. Similar to the previous results for bias, Figure 6d shows that bias is highest for weak instruments or when the percentage of data points having a non-decreasing effect is the highest. In both of these situations, the NPS test provides the highest power.

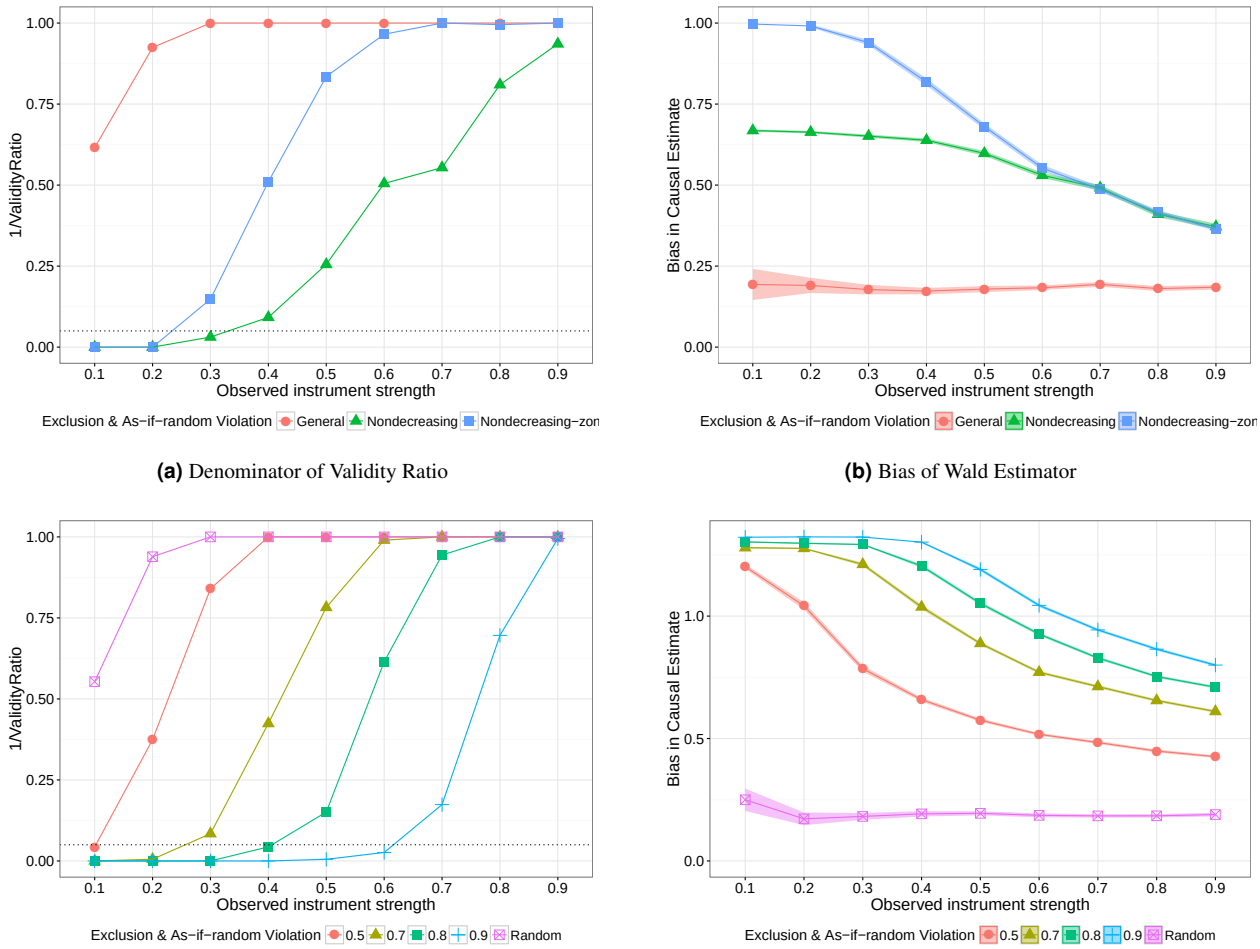
The lack of NPS test's power with a strong instrument Z is not surprising: in the limit, Z could be identical to X (an experiment with full compliance) and then Pearl-Bonet test inequalities (Equation 3) are satisfied trivially because the RHS will be 0. Clearly, these inequalities will be most discriminative when the instrument is weak. As we will see, this pattern will be consistent in the all results we obtain. Similary, we saw that bias in the causal estimate is highest for weak instruments; this trend also repeats across our simulations, consistent with past results that show even small violations in IV conditions can lead to big finite sample biases in the IV estimates, especially when the instrument is weak[29].

6.2 As-if-random may be violated

Violation of the as-if-random condition implies that r_z is not independent of r_x and r_y. Here we assume that Exclusion is not violated. Following Algorithm 1, we generate a joint distribution over r_z, r_x and r_y variables, sampling them uniformly at random. As with the exclusion restriction, there can be a number of ways to define the strength of an as-if-random violation, depending on how we specify the dependence between R_Z, R_X and R_Y. For the results presented, we define the strength of the violation as the mutual information between r_x and (r_x, r_y). When as-if-random is satified, mutual information will be zero. As we increase the mutual information, violation of as-if-random is expected to become more and more severe. Since correlation between R_Z and R_Y is necessary and sufficient for a violation of the as-if-random condition (but not correlation between R_Z and R_X), we modulate the severity of violation by changing the correlation between R_Z and R_Y. To do so, we vary a single conditional probability, $P(r_y = 3|r_z = 1)$ for simplicity; similar results can be obtained by varying other probabilities. We chose $P(r_y = 3|r_z = 1)$ because of the intuitive property that when it is high, Z and Y will also be highly correlated.

Figure 7a shows the results of the NPS test when as-if-random condition is violated. When we sample $P(R_Z, R_Y, R_Y)$ uniformly at random, the NPS test is unable to distinguish effectively between invalid-IV and valid-IV models. Even at low values of instrument strength ($<= 0.2$), nearly half of invalid-IV models pass the Pearl-Bonet test. However, we also see the Wald estimator is reasonably accurate at all levels of instrument strength, even though the as-if-random condition is not satisfied. This indicates that complete uniform sampling of causal models does not introduce a strong enough violation to either be detected by the NPS test or result in a noticeable biased estimate.

When the mutual information is increased beteen R_Z and R_Y, we find that the power of the NPS test increases. When the as-if-random threhold is $>= 0.7$, instruments with strength up to 0.2 have an errror rate of roughly 5%. Thus, the test is more powerful for stronger violations of the as-if-random assumption. That said, the test is unable to capture all violations that lead to noticeable bias. For instance, at a threshold of 0.5, bias in the causal estimate can be as high as 0.5, but the NPS test is able



(a) Denominator of Validity Ratio

(b) Bias of Wald Estimator

(c) Denominator of Validity Ratio

(d) Bias of Wald Estimator

Figure 8. Both violated: Exclusion and as-if-random. Top panel shows the 1/ValidityRatio and bias in the Wald estimator for the general case and when one or more of (Z, X) and Y have a non-decreasing causal relationship. The bottom panel shows the power of the NPS test as the severity of as-if-random violation is increased.

to detect invalid-IV models less than 50% of the time.

6.3 Both exclusion and as-if-random may be violated

Finally, we consider the case when both conditions may be violated. First, let us consider the straightforward case when both exclusion and as-if-random violating models are uniformly sampled. The red line in Figure 8a shows the power of the NPS test for such invalid-IV models is low; even for weak instrument strength, NPS test can correctly identify invalid-IV models less than 40% of the time. Fortunately, the bias is also negligible (Figure 8b), indicating that uniform violation does not lead to a noticeable bias in causal IV estimates.

When we stipulate that Z can only have a non-decreasing effect on Y, as in the above subsection, the power of the NPS improves. The error rate for identifying invalid-IV models is less than 5% for instruments with strength up to 0.2. However, the bias also shoots up. When the observed instrument strength is high (say 0.5), bias in the causal estimate is over 0.6, but the NPS test can correctly identify an invalid-IV model less than 20% accuracy. Assuming that both Z and X have a non-decreasing effect on Y provides slightly better results. Instruments of strength up to 0.4 have error rates nearly 5% and the bias also decreases.

When we modulate the severity of the as-if-random condition (while keeping exclusion violation uniformly at random), the power of the NPS test improves substantially (Figure 8c). For thresholds at least 0.7, the NPS test misses less than 5% of the invalid-IV models at an instrument strength of 0.2. Bias is also high for these thresholds, but we have a higher chance of correctly filtering out invalid-IV models.

Dataset	Log Marginal Likelihood			Validity Ratio
	Exclusion Violated	As-if-random Violated	Valid IV	
$D_0 : Z, X, Y_0$	-3080	-3086	**-3077**	20.1
$D_1 : Z, X, Y_1$	-3168	**-3161**	-3163	0.13
$D_2 : Z, X, Y_2$	**-3366**	-3367	-3397	$3.4x10^{-14}$

Table 1. Validity Ratio estimates for an example open problem proposed for testing binary instrumental variables. The NPS test can distinguish between valid-IV (D_0) and invalid-IV (D_1, D_2) datasets. Bold values denote the maximum marginal likelihood for each dataset.

6.4 Summary of results

Two key patterns emerge. First, the test is more powerful in recognizing violations that also lead to a substantial bias. This is encouraging because the kind of violations that bias the causal estimate are exactly the invalid-IV datasets we want to eliminate. On average, these results suggest that when the NPS test is inconclusive, it is unlikely that applying the Wald Estimator will lead to substantial bias in the causal estimate. Conversely, in cases when the NPS test has high power, eliminating invalid-IV data models will avoid computing Wald Estimates with high bias.

Second, the above results show that detecting the violation of IV conditions is sensitive to the strength of the instrument. This may seem as a big limitation; however, in most observational studies, instruments with high strength are rare. For instance, in economics,[30] recommend an F-value of > 10 to prevent weak instrument bias. At such a low value for F, the instrument is likely to have to low correlation with X. Similarly, in epigenetics, R^2 of 0.1 between Z and X is typical[31]. In these low strength regimes, the NPS test can be effective in testing for validity of an instrumental variable, within some accepted false positive rate.

7 USING NPS TEST TO VALIDATE PAST IV STUDIES

In this section we evaluate the effectiveness of the NPS test in practice. First, we will apply the test to an open problem for binary instrumental variables that was proposed as a limitation of Pearl-Bonet necessary test. Then, we will apply the test to two seminal and highly cited studies on instrumental variables. Finally, we will use the NPS test to validate recent studies from a leading economics journal,*American Economic Review*.

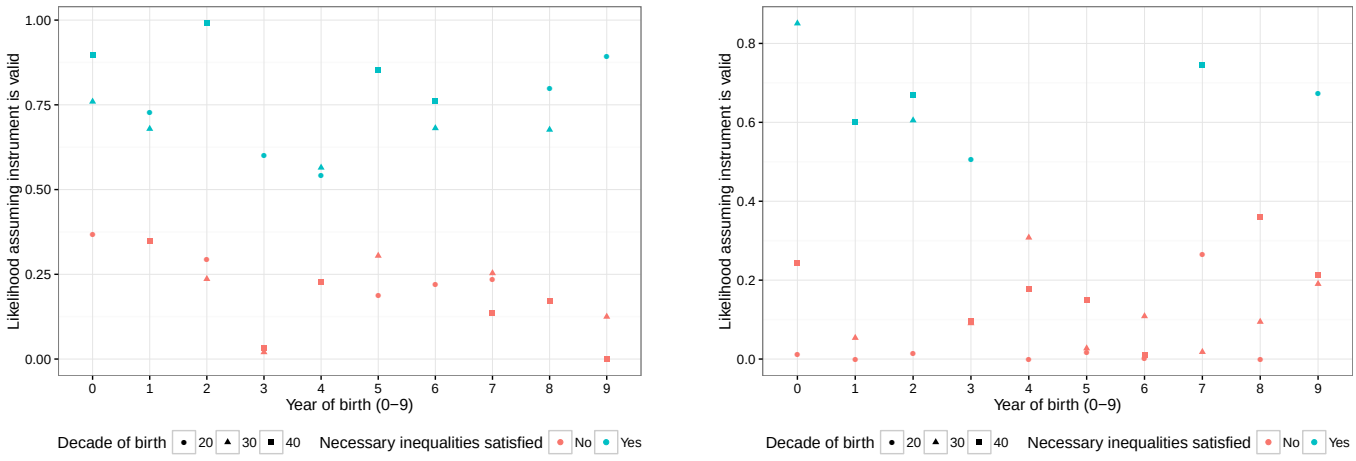
7.1 An example open problem for binary instrumental variables

Palmer et al.[32] proposed the following problem for binary variables where the Pearl-Bonet necessary test fails to detect violation of IV assumptions. Let Z, X, Y be three binary variables generated from the following causal process:

$$
\begin{aligned}
Z \sim & Bern(0.5) \\
U \sim & Bern(0.5) \\
X \sim & Bern(p_X); p_X = 0.05 + 0.1Z + 0.1U \\
Y_0 \sim & Bern(p_0); p_0 = 0.1 + 0.05X + 0.1U \\
Y_1 \sim & Bern(p_1); p_1 = 0.1 + 0.2Z + 0.05X + 0.1U \\
Y_2 \sim & Bern(p_2); p_2 = 0.1 + 0.05Z + 0.05X + 0.1U
\end{aligned} \tag{41}
$$

There can be three possible datasets: $D_0(Z, X, Y_0)$, $D_1(Z, X, Y_1)$,$D_2(Z, X, Y_2)$. Z is a valid instrument only when the outcome is Y_0, not for Y_1 and Y_2. Although Pearl-Bonet test is able to rule out D_1 as an invalid-IV dataset, Palmer et al. find that it is inconclusive for D_0 and D_2. We validate the same datasets using the NPS test by simulating 2000 data points for each dataset. Table 1 shows that comparing Validity-Ratio from the NPS test can be used to idntify the datasets for which Z is a valid instrument. The second and third columns show the log marginal likelihood for invalid-IV models when either of exclusion or as-if-random is violated. In the absence of additional information, we can assume a uniform prior over valid-IV versus invalid-IV models ($P(M_1 = P(M_2)$. This leads to the Validity Ratio computed in the fifth column, as a ratio of marginal likelihood of the Valid-IV model class over marginal likelihood of the Invalid-IV model class. Validity Ratio is the highest (nearly 20) for D_0 and the lowest ($< 10^{-13}$) for D_2, thereby clearly distinguishing between the two datasets. Dataset D_1 has a Validity Ratio less than 1, indicating that it is less likely to be a valid instrument, especially in comparison to dataset D_0.

In practice, a common goal is to select a single instrument. Results from the NPS test indicate that D_0 should be chosen; it has the highest Validity-Ratio among candidate datasets. Further, given that the Validity-Ratio is greater than 1, it is also likely to be a valid instrument. As a guidance, we suggest following standard thresholds for the Bayes Factor in determining when the



(a) Unconditional IV

(b) Conditional IV

Figure 9. Likelihood of being generated from a valid-IV causal model for instruments used in a past IV study, as provided by the output of the necessary test. Left panel shows unconditional instruments, while the right panel instruments conditioned on relevant covariates. Many of the instruments used have a likelihood below 0.05.

Validity-Ratio is high enough[?]. However, we deliberately refrain from providing standard thresholds, because the judgment for validity of an instrument will anyways depend on the prior assumed for Invalid-IV and Valid-IV model classes. We recommend interpreting the ratio of marginal likelihoods as a benchmark for priors: a Validity Ratio of 20 assuming uniform priors indicates that for D_0 to be an invalid-IV dataset, a researcher's prior on finding an invalid-IV dataset should be at least 20 times as strong as the prior for valid-IV dataset.

7.2 Seminal IV studies on the effect of schooling on wages

Returns of schooling on future income was one of the first applications that instrumental variable studies were applied to. We apply the NPS test to two of the seminal instrumental variable studies[1,33]. These studies are both highly cited, yet concerns about their validity continue until today[29,34]. We show how the NPS test can provide evidence and validate the instruments used in these studies.

7.2.1 Effect of compulsory schooling on future wages

The first paper estimates the effect of compulsory schooling on future earnings of students[1]. In the original analysis, Z is the quarter of birth, which was binarized to indicate that the student was either born in the first quarter or the last three quarters. X is years of schooling and Y is the (log) weekly earnings of individuals. Using yearly cohorts, the authors define a separate instrumental variable for each year of birth. The causal effect of interest is the effect of years of schooling on future earnings.[4] Years of schooling and earnings are both reported as continuous numbers, in a bounded range. We follow the simplest discretization by binarizing both these variables at their mean. To check robustness against outliers, we also tried using median as the cutoff and obtained similar results.

As in the original paper, we first use the IV method without conditioning on any covariates. of Applying the NPS test shows that 3 out of yearly instruments do not pass the necessary test at a significance level of $\alpha = 0.05$, as Figure 9a shows. This indicates that one or more of the three assumptions—monotonicity, exclusion and as-if-random—is violated, at least when all variables are binary. Further, interpreting the p-value of the necessary test as the likelihood of the observed data being generated from a valid-IV causal model, Figure 9a shows that nearly half of the instruments have less than 50% likelihood of being generated from a valid-IV causal model. Although we cannot rule out that some of them may be valid, we can use the likelihood values from the necessary test to filter out instruments that more likely to be valid and thus compute more reliable estimates.

In practice, however, unconditional instrumental variables are rare. The original paper proceeds to construct conditional instrumental variables based on related covariates in the dataset. We replicate conditioning on covariates by implementing a partialling out strategy[?]. Based on the Frisch-Waugh-Lovell theorem, using the partialling out technique provides the equivalent effect of conditioning on covariates, provided the underlying causal model is linear (which the authors assume). Figure 9b

[4]Dataset available at http://economics.mit.edu/faculty/angrist/data1/data/angkru1991

Study name	Num. Observations	IV Strength	Pearl-Bonet Test Result	Validity Ratio
Randomized Experiment				
National Job Training Partnership Act (JTPA) Study (2002)[35]				
	5102	0.58	Pass	3.4
Instrumental Variable Studies				
Effect of rural electrification on employment in South Africa (2011)[36]				
–Type 0	1816	0.1	Pass	3.6
–Type 1	1816	0.1	Fail (p=0.26)	0.002
–Type 2	1816	0.16	Fail (p=0.16)	0.0009
–Type 3	1816	0.05	Fail (p=0.11)	0.001
Effect of Chinese import competition on local labor markets (2013)[37]				
–Outcome(population change)	1444	0.59	Pass	0.3
–Outcome(employment)	1444	0.59	Pass	0.3
Effect of credit supply on housing prices (2015)[38]				
–Outcome(nloans)	11107	-0.009	Fail (p=0.003)	0.011
–Outcome(vloans)	11107	-0.003	Fail (p=0.005)	0.006
–Outcome(lir)	11107	-0.01	Fail (p=0.004)	0.0004
Effect of subsidy manipulation on Medicare premiums (2015)[39]				
–Unconditioned	170	0.60	Pass	1.02
–Conditioned	170	0.42	Pass	0.04
Effect of Mexican immigration on crime in United States (2015)[40]				
–Unconditioned	182	0.50	Pass	0.07
–Conditioned	182	0.22	Pass	0.005

Table 2. Split on mean.

shows the results of the necessary test when all instruments are conditioned on covariates. Contrary to intuition, a bigger fraction of conditional instruments are now invalid at the 5% significance level using the necessary test. Combined, our results on unconditional and conditional IVs suggest that many of the instruments used in the original analysis may not be valid, at least when the treatment and outcome is binarized.

7.2.2 Return of college education on future earnings

Another early study in the instrumental variables literature was on the effect of college education on future earnings of students[33]. This study used a person's distance from college as an instrument to estimate the causal effect of college education. We follow a similar protocol as above. Z is distance from college, which was already binarized in the original study. X is years of education and Y is the log weekly wages.[5] After binarizing X and Y, data from Card does not pass the necessary test for IV validity.

7.3 Recent IV studies in the American Economic Review

We now apply the NPS test to validate more recent IV studies. To select recent studies, we searched for papers published in the American Economic Review from 2011-2015 that had 'instrumental variable" or "instrument" mentioned in their title or abstract. From this set, we filtered out studies that did not provide full datasets for replication, leaving us with five studies on the causal effect of diverse economic treatments such as rural electrification[36], credit supply[38], subsidy manipulation[39], foreign import competition[37], and foreign immigration[40]. As a comparison benchmark, we also include an instrumental variable study based on data from a randomized experiment[35], which almost surely should pass the NPS test.

For each of the studies, we use code provided by the authors to construct a dataset of three variables (Z, X, Y), where Z is the instrument. If the authors condition on covariates, then we use the partialling out technique to process the (Z, X, Y) dataset. Finally, we binarize each variable at its mean, unless it is already binarized. Table 2 presents results from the NPS test. First, we find that the randomized experiment passes the Pearl-Bonet test and obtains a Validity Ratio of 3.4, thus providing a weak evidence for a valid instrument. In contrast, 2 out of 5 do not pass Pearl-Bonet test when binarized. They also report significantly low estimates of the Validity Ratio. For the other three studies, the Validity Ratio is less than 1, even though they pass the Pearl-Bonet necessary test. This suggests that the data does not support validity of the instruments used. As an example, the conditional instrument for the study on Mexican immigration obtains a Validity Ratio of 0.005, providing strong

[5]Dataset available at http://davidcard.berkeley.edu/data_sets.html

evidence for the invalidity of the instrument. However, in the other two studies the Validity Ratio is high enough that we cannot reject the validity of the instrument, thereby proving inconclusive about their validity.

8 DISCUSSION AND FUTURE WORK

We presented a probably sufficient test for instrumental variables using necessary tests proposed by past work. Simulation results show that the effectiveness of the test increases as the strength of the instrument decreases. Therefore, while the NPS test cannot always verify whether an instrument is valid, it can do so when the instrument is weak. Fortunately, many observational studies are based on instruments with low Z-X correlation, where NPS can be applied. Future work includes studying the sensitivity of the NPS test to different discretization strategies and different priors on causal models.

More generally, the NPS test is an example of a general Bayesian testing framework for causal models. When necessary existence tests based on specific causal models are available, they can be used to create a *probably* sufficient test. We have shown that the combined evidence from the observed data and a necessary test can be used to distinguish between causal models. Looking forward, the proposed test can be used to compare potential instruments for their validity, allow transparent comparisons between multiple IV studies, and enable a data-driven search for natural experiments.

Acknowledgements

We acknowledge Jake Hofman and Duncan Watts for their valuable feedback throughout the course of this work. We also thank Miro Dudik, Akshay Krishnamurthy, Justin Rao, Vasilis Syrgkanis and Michael Zhao for helpful suggestions.

References

1. Angrist, J. D. & Krueger, A. B. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* **106**, 979–1014 (1991).
2. Angrist, J. D. & Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion* (Princeton university press, 2008).
3. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine* **27**, 1133–1163 (2008).
4. Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**, 1 (2010).
5. Pearl, J. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 435–443 (Morgan Kaufmann Publishers Inc., 1995).
6. Bonet, B. Instrumentality tests revisited. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 48–55 (Morgan Kaufmann Publishers Inc., 2001).
7. Nakamura, A. & Nakamura, M. On the relationships among several specification error tests presented by durbin, wu, and hausman. *Econometrica: journal of the Econometric Society* 1583–1588 (1981).
8. Morgan, S. L. & Winship, C. *Counterfactuals and causal inference* (Cambridge University Press, 2014).
9. Dunning, T. *Natural experiments in the social sciences: a design-based approach* (Cambridge University Press, 2012).
10. Sharma, A., Hofman, J. M. & Watts, D. J. Split-door criterion for causal identification: Automatic search for natural experiments. *arXiv preprint arXiv:1611.09414* (2016).
11. Balke, A. & Pearl, J. Nonparametric bounds on causal effects from partial compliance data. *Technical report, UCLA* (1993).
12. Angrist, J. & Imbens, G. Identification and estimation of local average treatment effects. *Econometrica* (1994).
13. Ramsahai, R. & Lauritzen, S. Likelihood analysis of the binary instrumental variable model. *Biometrika* **98**, 987–994 (2011).
14. Kitagawa, T. A test for instrument validity. *Econometrica* **83**, 2043–2063 (2015).
15. Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT press, 2009).
16. Kass, R. E. & Raftery, A. E. Bayes factors. *Journal of the american statistical association* **90**, 773–795 (1995).
17. Balke, A. & Pearl, J. Probabilistic evaluation of counterfactual queries. *Proc. of AAAI* (1994).
18. Heckerman, D. & Shachter, R. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research* 405–430 (1995).

19. Hankin, R. K. *et al.* A generalization of the dirichlet distribution. *Journal of Statistical Software* **33**, 1–18 (2010).

20. Cooper, G. F. & Herskovits, E. A bayesian method for the induction of probabilistic networks from data. *Machine learning* **9**, 309–347 (1992).

21. Genz, A. & Cools, R. An adaptive numerical cubature algorithm for simplices. *ACM Transactions on Mathematical Software (TOMS)* **29**, 297–308 (2003).

22. Neal, R. M. Annealed importance sampling. *Statistics and Computing* **11**, 125–139 (2001).

23. Skilling, J. *et al.* Nested sampling for general bayesian computation. *Bayesian analysis* **1**, 833–859 (2006).

24. Feroz, F., Hobson, M. & Bridges, M. Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society* **398**, 1601–1614 (2009).

25. Chaves, R. *et al.* Inferring latent structures via information inequalities. *arXiv preprint arXiv:1407.2256* (2014).

26. Huber, M. & Mellace, G. Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics* **97**, 398–411 (2015).

27. Wang, L., Robins, J. M. & Richardson, T. S. On falsification of the binary instrumental variable model. *arXiv preprint arXiv:1605.03677* (2016).

28. Wald, A. The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics* **11**, 284–300 (1940).

29. Bound, J., Jaeger, D. A. & Baker, R. M. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association* **90**, 443–450 (1995).

30. Staiger, D. O. & Stock, J. H. Instrumental variables regression with weak instruments (1994).

31. Pierce, B. L., Ahsan, H. & VanderWeele, T. J. Power and instrument strength requirements for mendelian randomization studies using multiple genetic variants. *International journal of epidemiology* dyq151 (2010).

32. Palmer, T. M., Ramsahai, R. R., Didelez, V., Sheehan, N. A. *et al.* Nonparametric bounds for the causal effect in a binary instrumental-variable model. *Stata Journal* **11**, 345 (2011).

33. Card, D. Using geographic variation in college proximity to estimate the return to schooling. Tech. Rep., National Bureau of Economic Research (1993).

34. Buckles, K. S. & Hungerman, D. M. Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics* **95**, 711–724 (2013).

35. Abadie, A., Angrist, J. & Imbens, G. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* **70**, 91–117 (2002).

36. Dinkelman, T. The effects of rural electrification on employment: New evidence from south africa. *The American Economic Review* **101**, 3078–3108 (2011).

37. David, H., Dorn, D. & Hanson, G. H. The china syndrome: Local labor market effects of import competition in the united states. *The American Economic Review* **103**, 2121–2168 (2013).

38. Favara, G. & Imbs, J. Credit supply and the price of housing. *The American Economic Review* **105**, 958–992 (2015).

39. Decarolis, F. Medicare part d: are insurers gaming the low income subsidy design? *The American Economic Review* **105**, 1547–1580 (2015).

40. Chalfin, A. The long-run effect of mexican immigration on crime in us cities: Evidence from variation in mexican fertility rates. *The American Economic Review* **105**, 220–225 (2015).