Necessary and Probably Sufficient Test for Finding Valid Instrumental Variables

Amit Sharma
Microsoft Research India
Vigyan, Bangalore 560008
amshar@microsoft.com

Abstract

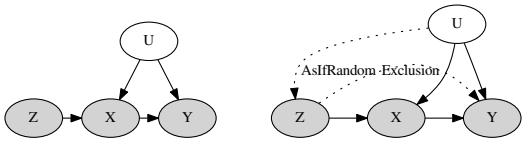
Can instrumental variables be found from data? While instrumental variable (IV) methods are widely used to identify causal effect, testing their validity from observed data remains a challenge. This is because validity of an IV depends on two assumptions, *exclusion* and *as-if-random* treatment assignment, that are largely believed to be untestable from data. In this paper, we show that under certain restrictive conditions, testing for instrumental variables is possible. We build upon prior work on necessary tests to derive a test that characterizes the odds of being a valid instrument, thus yielding the name "necessary and *probably* sufficient". The test works by defining the class of invalid-IV and valid-IV causal models as Bayesian generative models and comparing their marginal likelihood based on observed data. When all variables are discrete, we also provide a method to efficiently compute these marginal likelihoods.

We evaluate the test on multiple simulations for binary data, inspired by an open problem for IV testing. We find that the test is most powerful when the instrument has moderate-to-weak strength; incidentally, such instruments are commonly used in observational studies. Among as-if-random and exclusion, it detects exclusion violations with higher power. That said, the results are sensitive to the choice of prior over causal models. We use a uniform prior; in practice domain knowledge will be useful to select a suitable prior.

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. Increasingly, it is also being used in computing systems to estimate causal effect from log data (Sharma, Hofman, and Watts 2015; Peysakhovich and Eckles 2018). The key idea is to find subsets of the data that resemble a randomized experiment, and use those subsets to estimate causal effect. (Angrist and Pischke 2008). Specifically, consider the canonical causal inference problem shown in Figure 1a. The goal is to estimate the effect of a variable X on another variable Y based on observed data, where X is commonly referred to as the *treatment* and Y as the *outcome*. However, there are unobserved (and possibly unknown) common causes for X and Y that confound observed association between X and Y, making the isolation of X's effect on Y a non-trivial problem. Unlike methods such as stratification or matching that condition on all observed common causes (Morgan and Winship 2014), the instrumental variable method 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. The advantage is that we do not need to assume that all confounding common causes are observed to estimate the causal effect. To be a valid instrument, however, Z should satisfy three conditions (Angrist and Pischke 2008). 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 (*Ignorable* or As-if-random treatment assignment). The latter two conditions are shown in the graphical model in Figure 1b. These conditions can also be expressed as conditional independence constraints: exclusion and ignorable assignment conditions imply $Z \perp\!\!\!\perp Y | X, U$ and $Z \perp\!\!\!\perp U$ 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 measuring observed correlation between Z and X), the other two conditions depend on unobserved variables U and thus are harder to check. Although necessary tests do exist that can weed out bad instruments (Pearl 1995; Bonet 2001), in practice exclusion and as-if-random as considered as *assumptions* and often defended with qualitative domain knowledge. This can be problematic because the entire validity of the IV estimate depends on the exclusion and as-if-random conditions.

In this paper, therefore we propose a test for validating instrumental variables that can be used to find, evaluate and compare potential instruments for their validity. Although instruments are untestable in general (Morgan and Winship 2014), we find that in many cases it is possible to distinguish between invalid and valid instruments. To do so, the proposed test applies the principles of Bayesian model comparison to causal models and estimates marginal likelihood of an valid instrument given the observed data. Comparing this to the corresponding marginal likelihood for an invalid instrument provides a metric for evaluating the validity of an instrument. The intuition is that if the instrument is valid, then causal models with an instrument as in Figure 1a should be able to generate observed data with higher likelihood than



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

Figure 1: Standard instrumental variable causal model and violations of the two assumptions—exclusion and as-if-random—that lead to an invalid-IV model.

all other causal models. Specifically, let *Valid-IV* refer to the class of all causal models that yield a valid instrument and *Invalid-IV* to the class of causal models that yield an invalid instrument. Given an observed data distribution $P(X, Y, Z)$, the proposed method computes the ratio of marginal likelihoods for Valid-IV and Invalid-IV models. Whenever this marginal likelihood ratio is above a pre-determined acceptance threshold, we can conclude that the instrument is likely to be valid. This notion of validity, however, depends on the specific prior used for unidentified distributions and thus the results are sensitive to the choice of prior. We use the uniform prior in this work and leave other variations for future work. To distinguish this probabilistic notion from deterministic sufficiency—conditions that would determine in absolute whether an instrument is valid or not—we call an instrument that passes the marginal likelihood ratio test as *probably sufficient*.

Combining the above approach with necessary tests proposed in past work leads to a *Necessary and Probably Sufficient (NPS)* test for instrumental variables. The NPS test proceeds as follows. If the observed data does not satisfy the necessary conditions, then it is declared invalid. If it does, then we proceed to estimate the marginal likelihood ratio over Valid-IV and Invalid-IV models. When all variables are discrete, we provide a general implementation of this test that makes no assumptions about the nature of functional relationships between the treatment, outcome and instrument.

Finally, any statistical test is only as good as the decisions it helps to support. Among the two IV assumptions, simulations show that the NPS test is more effective at detecting violations of the exclusion restriction. We also find that the proposed NPS test is most effective for validating instruments having low correlation with the treatment X. Incidentally, most of the instruments used in observational studies in the social and biomedical sciences have weak to moderate strength, well-suited to the NPS test. To demonstrate the test's usefulness in practice, we first consider an open problem proposed by (Palmer et al. 2011) for validating an instrumental variable and show that the NPS test can identify valid instruments in that setting. We then evaluate effectiveness of the NPS test on multiple simulated datasets.

Background: Testability of an IV

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 (Morgan and Winship 2014). Pearl (1995), however, discovered that the specific causal graph structure in Figure 1 imposes constraints on the observed probability distribution over Z, X and Y. The causal model from Figure 1a can be written as:

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

where f and g are arbitrary deterministic functions and U represents arbitrary, unobserved random variables that are independent of Z. Using this framework, Pearl derived a necesssary test that any observed data generated from a valid instrumental variable model must satisfy (Pearl 1995). For binary variables Z, X and Y,

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

Typically, researchers make an additional assumption that helps to derive a point estimate for the Local Average Treatment Effect (LATE). This assumption, called monotonicity (Angrist and Imbens 1994), precludes any *defiers* to treatment in the population (Angrist and Pischke 2008). That is, we assume that $g(z_1, u) \geq g(z_2, u)$ whenever $z_1 \geq z_2$. Under these conditions, Pearl showed that we can obtain tighter inequalities.

$$\begin{aligned}
&P(Y=y, X=1|Z=1) \geq P(Y=y, X=1|Z=0) \quad \forall y \in \{0,1\} \\
&P(Y=y, X=0|Z=0) \geq P(Y=y, X=0|Z=1) \quad \forall y \in \{0,1\}
\end{aligned}$$

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.

Moreover, when X, Y and Z are binary, this test is not only necessary, it is the strongest necessary test possible (Bonet 2001; Kitagawa 2015).Bonet (2001) extended this work to create the strongest necessary test when the variables are discrete—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 refer to this test as the *Pearl-Bonet test*.

Such a test weeds out bad instruments, but is inconclusive whenever an instrument passes the test. Sufficient tests exist, but require prohibitive assumptions such as knowing another valid instrumental variable as in the Durbin-Wu-Hausman test (Nakamura and Nakamura 1981), or stipulating that confounders have no effect on the outcome.

Necessary and Probably Sufficient (NPS) test

Without prohibitive assumptions, establishing *sufficiency* for a validity test is non-trivial. In particular, the usual method of comparing the maximum data likelihood of the two classes of IV models, Valid-IV or Invalid-IV, provides us no information. This is because Invalid-IV class of models (as in Figure 1b) is more general than the Valid-IV class and thus is always as likely (or more) to generate the observed data.

Instead of comparing *maximum* likelihoods of model classes, we turn to estimating likelihoods of individual causal models from Invalid-IV and Valid-IV classes. The intuition is that while the Invalid-IV class may always have a causal model that matches likelihood of the Valid-IV class for a valid instrument, there will be many other Invalid-IV

models that provide a lower likelihood for the data. By generating models with different violations of the Exclusion and As-if-random conditions, we can estimate the data likelihood over individual models in the Invalid-IV class. Averaging over all models in the Valid-IV and Invalid-IV classes, we expect marginal likelihood to be higher for the Valid-IV class for data generated from a Valid-IV model. The idea of comparing different models from Valid-IV and Invalid-IV classes is similar to sensitivity analysis, except that we are interested in the likelihood of data rather than resultant causal estimates.

Thus, unlike necessary tests (Ramsahai and Lauritzen 2011; Kitagawa 2015) that refute a null hypothesis that observed data was generated from a valid-IV model, probable sufficiency requires estimating the relative probability of valid-IV and invalid-IV models given observed data. When the relative probability—formally, *marginal likelihood*—is high, the instrument is likely to be valid. Conversely, when it is low, the instrument is likely to be invalid. Based on this, we now provide a definition for probable sufficiency.

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

Combined, the Pearl-Bonet 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. Throughout, we assume that Z, X and Y are all discrete variables.

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—that could have generated the observed data. Therefore, we describe a probabilistic generative *meta-model* of how observed data is generated from a causal model, which in turn, is generated based on the as-if-random and exclusion assumptions.

Let us first define the valid-IV and invalid-IV models formally in terms of the two IV assumptions: exclusion and as-if-random. 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. 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, given 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 assumption and R denotes the as-if-random assumption.

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 2. 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, one 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-Bonet 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 the Pearl-Bonet necessary test. All other variables in the meta-model are unobserved.

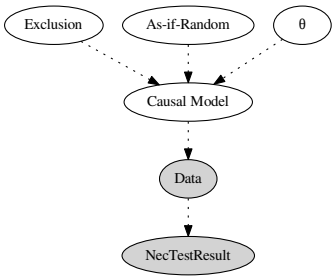


Figure 2: A probabilistic graphical meta-model describing the connection between IV conditions and specific causal models. Note that arrows are dotted to distinguish these *probabilistic* diagrams from the causal diagrams in Fig. 1.

Marginal likelihood of Valid-IV and Invalid-IV

Let PT denote whether the observed data passed the necessary test. 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, and result of the Pearl-Bonet 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:*

$$\begin{aligned} \textit{Validity-Ratio} &= \frac{P(E,R|PT,D)}{P(\neg(E,R)|PT,D)} \\ &= \frac{P(PT,D|E,R) * P(E,R)}{P(PT,D|\neg(E,R)) * P(\neg(E,R))} \\ &= \frac{P(M1)}{P(M2)} \frac{\int_{M_1:m \textit{ is valid}} P(m|E,R)P(D|m)dm}{\int_{M_2:m \textit{ is invalid}} P(m|\neg(E,R))P(D|m)dm} \end{aligned} \quad (3)$$

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.

While we are additionally using the result of the Pearl-Bonet necessary test to compute evidence, the Validity-Ratio

reduces to the Bayes Factor (Kass and Raftery 1995). The proof of the theorem follows from the structure of the generative meta-model and properties of the Pearl-Bonet necessary test.

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

$$\textit{ML-Ratio} = \frac{P(PT=1, D=d|E,R)}{P(PT=1, D=d|\neg(E,R))} \tag{4}$$

Since the Pearl-Bonet test is a necessary test, we know that $P(PT|E,R) = 1$ if the true data distributions are known. However, in practice, we will have a data sample and apply a statistical test. Therefore in some cases the test may return Fail even if E and R are satisfied, leading to the following expression for the numerator:

$$\begin{aligned} &P(PT=1, D=d|E,R) \\ &= P(PT=1|D=d,E,R)P(D=d|E,R) \\ &= P(PT=1|D=d)P(D=d|E,R) \end{aligned} \tag{5}$$

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 $P(D|E,R)$ as:

$$P(D|E,R) = \int_m P(D,m|E,R)dm \tag{6}$$

$$= \int_m P(m|E,R)P(D|m)dm = \int_{M_1:m \text{ is valid}} P(m|E,R)P(D|m)dm$$

Similarly, the denominator can be expressed by,

$$\begin{aligned} &P(PT, D|\neg(E,R)) = P(PT|D,\neg(E,R))P(D|\neg(E,R)) \\ &= P(PT|D)P(D|\neg(E,R)) \\ &= P(PT|D)\int_m P(D,m|\neg(E,R))dm \\ &= P(PT|D)\int_m P(m|\neg(E,R))P(D|m,\neg(E,R))dm \\ &= P(PT|D)\int_{M_2:m \text{ is invalid}} P(m|\neg(E,R))P(D|m)dm \end{aligned} \tag{7}$$

where we use the conditional independencies entailed by the generative meta-model.

Combining Equations 5, 6 and 7, we obtain the ratio of marginal likelihoods:

$$\begin{aligned} \textit{ML-Ratio} &= \frac{P(PT, D|E,R)}{P(PT, D|\neg(E,R))} \\ &= \frac{\int_{M1:m \text{ is valid}} P(m|E,R)P(D|m)dm}{\int_{M2:m \text{ is invalid}} P(m|\neg(E,R))P(D|m)dm} \end{aligned} \tag{8}$$

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

$$\frac{P(E,R)}{P(\neg(E,R))} = \frac{P(M1)}{P(M2)} \tag{9}$$

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

$$\begin{aligned} &\frac{P(PT, D|E,R) * P(E,R)}{P(PT, D|\neg(E,R)) * P(\neg(E,R))} \\ &= \frac{P(M1)}{P(M2)} \frac{\int_{M1:m \text{ is valid}} P(m|E,R)P(D|m)dm}{\int_{M2:m \text{ is invalid}} P(m|\neg(E,R))P(D|m)dm} \end{aligned} \tag{10}$$

□

As with the Bayes Factor, estimation of the Validity-Ratio depends on the prior on causal models because the model is not identified. 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 can assume a uninformative uniform prior on causal models within each of the Valid-IV and Invalid-IV classes. If sufficient data is available, one may use the fractional Bayes Factor (O'Hagan 1995) to split the sample and use the first subsample to find a prior on causal models using data likelihood, and the second to estimate the Validity Ratio. We discuss the effect of using other model priors in the Discussion. Using a uniform model prior leads to the 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*

$$\textit{Validity-Ratio} = \frac{P(M1)}{P(M2)} \frac{K_2 \int_{M1:m \text{ is valid}} P(D|m)dm}{K_1 \int_{M2:m \text{ is invalid}} P(D|m)dm} \tag{11}$$

where K_1 and K_2 are normalization constants.

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.

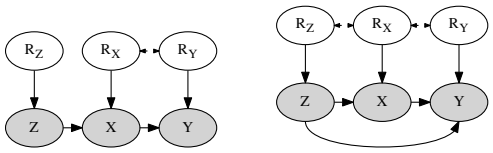
1. Estimate $P(Y, X|Z)$ using observational data and run the Pearl-Bonet necessary test. If the necessary test fails, Return *REJECT-IV*.
2. Else, compute the Validity-Ratio from Equation 3 for the one or more of the following types of violations (can exclude violations that are known *apriori* to be 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*.

Computing the Validity Ratio

The key detail in implementing the NPS test is in evaluating the integrals in Equation 3, since there can be infinitely many valid-IV or invalid-IV causal models. In this section we present a possible approach using the *response variables* framework from (Balke and Pearl 1994). We extend this framework to also work with invalid-IV causal models.

The response variable framework

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,



(a) Valid IV (b) Invalid IV

Figure 3: Causal graphical model with response variables R_X and R_Y denoting the effect of unknown, unobserved U.

and thus any causal model can be expressed as a probability distribution over the response variables. Note that there is no restriction on U—they can be discrete or continuous—but instead restrict the observed variables to be discrete.

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

$$y = f(x, u_y); x = g(z, u_x); \qquad z = h(u_z) \tag{12}$$

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 Eqn 12, Y can be written as a combination of 4 deterministic functions of x, after adding a response variable, r_y.

$$y = \begin{cases} f_{ry_0}(x) \equiv 0, & \text{if } r_y = 0 \\ f_{ry_1}(x) \equiv x, & \text{if } r_y = 1 \\ f_{ry_2}(x) \equiv \tilde{x}, & \text{if } r_y = 2 \\ f_{ry_3}(x) \equiv 1, & \text{if } r_y = 3 \end{cases} \tag{13}$$

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 ($r_y = 3$).

Similarly, we can write a deterministic functional form for X, leading to the transformed causal diagram with response variables in Figure 3a.

$$x = \begin{cases} g_{rx_0}(z) \equiv 0, & \text{if } r_x = 0 \\ g_{rx_1}(z) \equiv z, & \text{if } r_x = 1 \\ g_{rx_2}(z) \equiv \tilde{z}, & \text{if } r_x = 2 \\ g_{rx_3}(z) \equiv 1, & \text{if } r_x = 3 \end{cases} \tag{14}$$

Similar to r_y, $r_x = \{0, 1, 2, 3\}$ can be interpreted in terms of a subject's compliance behavior to an instrument: *never-taker*, *complier*, *defier*, and *always-taker* (Angrist and Pischke 2008). Finally, z can be assumed to be generated by its own response variable, r_z. That is, $Z = R_Z$.

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 3a, 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.

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, which are characterized by violations of Exclusion and/or As-if-random assumptions.

Exclusion is violated Exclusion violation implies that $Z \perp\!\!\!\perp Y|X, U$ does not hold, and thus 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 3b. In response variables framework, this translates to:

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

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 violation, dimensions will be $l + m^l n^{lm}$.

As-if-random is violated 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 $lm^l n^m$.

Both exclusion and as-if-random are violated In this case the structural equation for Y is given by Equation 15 and R_Z is not independent of (R_X, R_Y). Thus the dimensions of θ increase to $lm^l n^{lm}$.

Thus, under the response variable framework, choosing a causal model is equivalent to sampling a probability vector θ from the joint probability distribution $P(r_x, r_y, r_z)$. We use this to compute the integrals in Equation 3 by transforming them to an integral over θ parameters. Details are in the Appendix. When variables are discrete, we also provide extensions to the Pearl-Bonet test to handle monotonicity and practical implementation, described in the Appendix.

Simulations: How powerful is the NPS test?

We first simulate datasets and check whether estimating the Validity Ratio can correctly identify whether they contain a valid instrumental variable or not. Throughout, we assume monotonicity and that Z, X and Y are binary. For statistical significance of the Pearl-Bonet test, we use an exact test by Wang, Robins, and Richardson (2016) which converts the inequalities of the necessary test into a version of one-tailed Fisher's exact test.

An example open problem for binary IV

To start with, we consider the following causal model from (Palmer et al. 2011) where the Pearl-Bonet necessary test fails to detect violation of IV assumptions.

$$\begin{aligned} Z &\sim Bern(0.5); U \sim Bern(0.5) \\ X &\sim Bern(p_X); p_X = 0.05 + 0.1Z + 0.1U \\ Y_0 &\sim Bern(p_0); p_0 = 0.1 + 0.05X + 0.1U \\ Y_1 &\sim Bern(p_1); p_1 = 0.1 + 0.2Z + 0.05X + 0.1U \\ Y_2 &\sim Bern(p_2); p_2 = 0.1 + 0.05Z + 0.05X + 0.1U \end{aligned} \quad (16)$$

where Z, X, Y_i are the instrument, cause and outcome respectively and all variables are binary. There can be three datasets depending on which Y is chosen as the outcome: $D_0(Z, X, Y_0)$, $D_1(Z, X, Y_1)$,$D_2(Z, X, Y_2)$. Z is a valid instrument only when the outcome is Y_0, not for Y_1 and Y_2 because they violate the exclusion restriction. 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 three datasets using the NPS test by simulating 2000 data points from each of their causal models. Table 1 shows that comparing Validity-Ratio can be used to identify the datasets for which Z is a valid instrument. We assume a uniform prior over models within valid-IV and invalid-IV model classes and use the equation from Corollary 1. Further, in the absence of any additional information, we can assume an equal probability of the instrument being valid or invalid ($P(M_1) = P(M_2)$). The second and third columns show the log marginal likelihood for invalid-IV models when either of exclusion or as-if-random is violated. This leads to the Validity Ratio shown 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.

Simulating a broad range of binary datasets

Motivated by the example from (Palmer et al. 2011), we now construct a set of datasets by changing the parameters of the Palmer et al.'s example model presented above. Z and U are generated from a Bernoulli distribution as before, but parameter for effect of Z on X can have five different values: $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Similarly, the effect of X on Y takes values in this set. Each of U's effect on X, U's effect on Y, U's effect on Z, and Z's effect on Y takes on values from the set $\{0, 0.1, 0.3, 0.5\}$. For simplicity, we assume that U's effect on X and Y is the same. Combined, these parameters lead to 5x5x4x4x4=1600 simulations, each of which yields a different causal model. From each causal model, we generate an i.i.d. dataset with size=50000 of $< Z, X, Y >$ tuples.

	Log Marginal Likelihood			
Dataset	Exclusion Violated	As-if-random Violated	Valid IV	Validity Ratio
$D_0 : Z, X, Y_0$	-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	$3.4\text{x}10^{-14}$

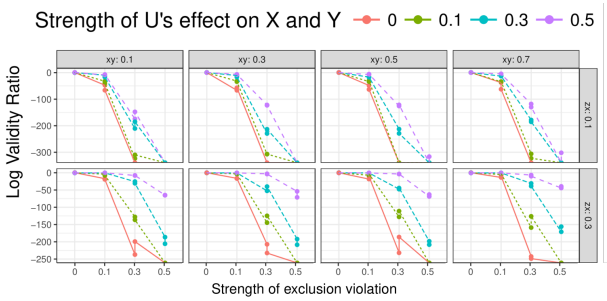
Table 1: Validity Ratio estimates for an example open problem proposed for testing binary instrumental variables. The NPS test can distinguish between valid-IV (D_0) and invalid-IV (D_1, D_2) datasets. Bold values denote the maximum marginal likelihood for each dataset.

These simulated datasets span the range of datasets with a valid or invalid instrument. When the parameters for the effect of U on Z and the effect of Z on Y are zero, the causal model contains a valid instrument. Otherwise, it contains an invalid instrument. We make the same assumptions as before: equal prior probability of an invalid or valid instrument, and a uniform prior over causal models within both Valid-IV and Invalid-IV model classes. On each dataset, we compute the Validity-Ratio using the equation from Corollary 1.

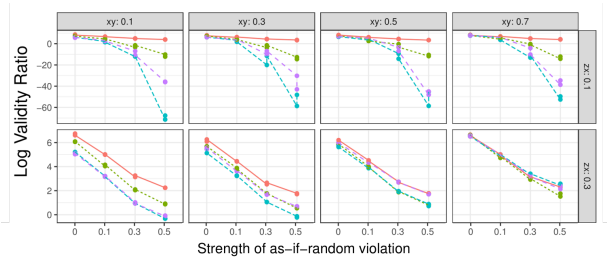
NPS test can detect exclusion violation, except when instrument is strong When only exclusion is violated, Figure 4a shows the log Validity-Ratio as the strength of the exclusion violation is increased. We find that when the parameter for effect of Z on X (*instrument strength*) is below 0.5, Validity Ratio is below 1 consistently even for minor violation of the exclusion restriction. This holds true even as the true causal effect is varied: scanning horizontally through the rows shows a similar trend. Above results indicate that exclusion can be tested using the Validity-Ratio as long as the instrument is not too strong (effect parameters < 0.5).

As-if-random is hard to detect, except when instrument is very weak Next, we look at violation of the as-if-random restriction only (Figure 4b). We find that as-if-random violation is harder to detect than exclusion. When the instrument is very weak (effect parameter for Z on X is 0.1), the Validity-Ratio goes below 1 as the strength of as-if-random violation is increased. This result is consistent even as the direct causal effect from X to Y is varied. However, when instrument strength increases to 0.3, the Validity Ratio stays above 1 and NPS test is unable to detect violation.

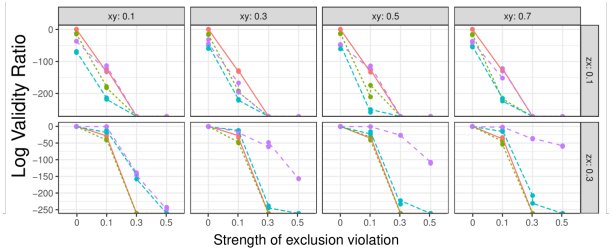
Violations of both assumptions is easier to detect Finally, we look at the case when both exclusion and as-if-random are violated. Figure 4c shows the log Validity-Ratio as the strength of exclusion violation varies, for a fixed as-if-random violation of 0.5 (i.e., the parameter for U's effect on Z is 0.5). When both exclusion and as-if-random are violated, it becomes easier to identify datasets with invalid instruments. Even when the is instrument is moderately strong (effect of Z on X is 0.7), we find that Validity Ratio quickly drops to less than 1 as the strength of exclusion violation increases. This pattern is consistent as the true causal effect of



(a) Only Exclusion Violated



(b) Only As-if-random Violated



(c) Both Violated (strength of as-if-random violation=0.5)

Figure 4: Log Validity-Ratio computed from the NPS test on simulated binary datasets with different violations. *(Rows)* zx denotes the direct effect of Z on X. *(Columns)* xy denotes the direct effect of X on Y.

X on Y is varied across datasets. When the instrument's effect on X is the strongest at 0.9, NPS test can still detect violations of exclusion with a severity higher than 0.3. Detailed results and other simulation configurations are in Appendix.

Validating American Economic Review studies

We also apply the NPS test to validate IVs used by 5 recent studies from a leading economics journal, *American Economic Review*. After binarizing variables, we find that 2 out 5 studies do not pass the Pearl-Bonet test and have low Validity Ratios. Details on the evaluation are in Appendix.

Discussion and Future Work

We presented a probably sufficient test for instrumental variables using necessary tests proposed in past work. Simulations show that the test is more effective for detecting violation of the exclusion assumption, and that effectiveness of the test increases as the strength of the instrument decreases.

Nevertheless, the proposed test has several limitations. First, it relies on the specification of a prior over causal models for both the Valid-IV and Invalid-IV model classes. In this paper we chose a uniform prior. It will be useful to study the sensitivity of the Validity-Ratio to changes in the prior. Second, the proposed implementation of the NPS test works only for discrete variables. Third, even for discrete variables, the test is often inconclusive. If the Validity-Ratio lies close to 1 (from -1 to 0 on the log scale), then we are unable to distinguish between valid and invalid instruments. Based on the simulation results, we conjecture that in such cases the resultant causal estimate will not have high bias even for invalid instruments, but this claim needs more evidence.

More generally, testing is a step towards the final goal of valid causal estimates. We would like to explore connections of our Bayesian testing framework to recent work on estimating bounds for similar problems (Silva and Evans 2016). 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 more data-driven search for natural experiments.

Appendix Details on computing Validity Ratio, extensions to discrete variables, and extensive simulation results are available at: `https://arxiv.org/abs/1812.01412`

References

Angrist, J., and Imbens, G. 1994. Identification and estimation of local average treatment effects. *Econometrica*.

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

Balke, A., and Pearl, J. 1994. Probabilistic evaluation of counterfactual queries. *AAAI*.

Bonet, B. 2001. Instrumentality tests revisited. In *Proc. UAI*, 48–55.

Kass, R. E., and Raftery, A. E. 1995. Bayes factors. *JASA* 90(430):773–795.

Kitagawa, T. 2015. A test for instrument validity. *Econometrica* 83(5):2043–2063.

Morgan, S. L., and Winship, C. 2014. *Counterfactuals and causal inference*. Cambridge University Press.

Nakamura, A., and Nakamura, M. 1981. On the relationships among several specification error tests presented by durbin, wu, and hausman. *Econometrica* 1583–1588.

O'Hagan, A. 1995. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)* 99–138.

Palmer, T. M.; Ramsahai, R. R.; Didelez, V.; and Sheehan, N. A. 2011. Nonparametric bounds for the causal effect in a binary instrumental-variable model. *Stata Journal*.

Pearl, J. 1995. On the testability of causal models with latent and instrumental variables. In *Proc. UAI*, 435–443. Morgan Kaufmann Publishers Inc.

Peysakhovich, A., and Eckles, D. 2018. Learning causal effects from many randomized experiments using regularized instrumental variables. In *Proc. World Wide Web*.

Ramsahai, R., and Lauritzen, S. 2011. Likelihood analysis of the binary instrumental variable model. *Biometrika* 98(4):987–994.

Sharma, A.; Hofman, J. M.; and Watts, D. J. 2015. Estimating the causal impact of recommendation systems from observational data. In *Proc. ACM Economics and Computation*.

Silva, R., and Evans, R. 2016. Causal inference through a witness protection program. *The Journal of Machine Learning Research* 17(1):1949–2001.

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