Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Necessary and Probably Sufficient Test for Finding Valid Instrumental Variables

Amit Sharma AMSHAR@MICROSOFT.COM
Microsoft Research India
Vigyan, Bangalore 560008

Editor: TBD

Abstract

Can instrumental variables be found from observed data? While instrumental variable methods are widely used to identify causal effect, testing their validity remains a big challenge. This is because validity of an IV depends on two assumptions—*exclusion* and *as-if-random*—that are largely believed to be untestable from data. In this paper, we show, however, that testing for instrumental variables is possible under certain conditions. 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 in terms of Bayesian graphical 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 an extensive set of simulations for binary data, inspired by an open problem for IV testing proposed in past work. We find that the test is most powerful when an instrument follows monotonicity—effect on treatment is either non-decreasing or non-increasing—and 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. 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 may be flawed, at least when all variables are discretized. The proposed test opens the possibility of algorithmically finding instruments in large datasets and more generally, adopting a data-driven approach to validating instrumental variable studies.

Keywords: Instrumental variable, Sensitivity analysis, Bayesian model comparison

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 that resemble 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 (Angrist and Krueger, 1991; Angrist and Pischke, 2008), and in epidemiology (under the name *Mendelian* randomization) to study the effect of risk factors on disease outcomes (Lawlor et al., 2008).

Inspite of their popularity, basic questions about the design, analysis and evaluation of instrumental variable (IV) studies remain elusive. In the design phase, it is unclear how to find a suitable instrumental variable. Even with access to bigger and more granular datasets, finding an instrument requires ingenuity and a laborious manual search, thereby restricting most IV studies to instruments derived from a small set of events such as the weather, lotteries or sudden shocks (Dunning, 2012). In the analysis phase, arguments are proposed to justify selection of an instrument, but it is hard to ascertain the extent to, or for which populations, the instrument is likely to be valid. Finally, in the evaluation phase, it is unclear how to evaluate the estimates produced by multiple IV studies (even on the same dataset) due to the absence of any objective criteria of comparing relative validity of instruments.

Specifically, consider the canonical causal inference problem shown in Figure 1a. 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 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 (Stuart, 2010), 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 (*As-if-random*). 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 as-if-random 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, 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; Dunning, 2012), 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 all other causal models. Specifically, let *Valid-IV* refer to the

(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.

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. 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 combined 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 cause, 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. In particular, for certain instrumental variable designs where we restrict the direction of causal effects (such as encouragement designs (**?**)), the NPS test is able to detect invalid instruments with high power. We also find that the proposed NPS test is most effective for validating instruments having low correlation with the cause of interest. 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 is able to identify valid instruments in that setting. Second, we apply the test to datasets from two seminal studies and five recent papers on instrumental variables from the American Economic Review, a premier economics journal. In many cases, we find that instrumental assumptions used in the corresponding papers were possibly flawed,

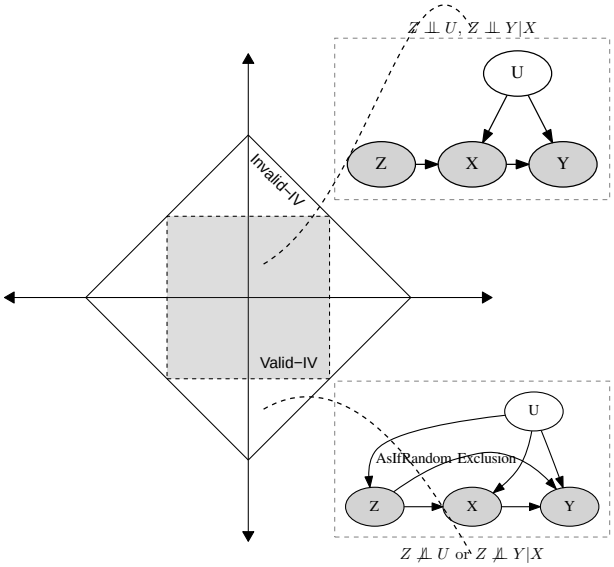


Figure 2: A 2D schematic of the probability space $P(X, Y|Z)$. Each point on the graph is a conditional probability distribution over X, Y given Z. The two squares in the left panel show polytopes for invalid-IV and valid-IV models in the probability space. Note that data distributions generated by valid-IV 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.

at least when variables are binarized. We provide an R package *ivtest* that provides an implementation of the NPS test.

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 (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. These constraints can be used to derive a necessary test for IV conditions (Pearl, 1995; Bonet, 2001). Such a test weeds out bad instruments, but is inconclusive whenever an instrument passes the test. Other tests are sufficient, 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. Due to these shortcomings, the dominant method to evaluate IV estimates involves a sensitivity analysis (Small, 2007), where we

estimate the impact on the causal estimate when assumptions are violated with varying severity. We review prior tests and sensitivity analysis for instrumental variables below.

2.1 Tests for an instrumental variable

A classic test for instrumental variables is the Durbin-Wu-Hausman test (Nakamura and Nakamura, 1981). 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, an intuitive idea is to test whether the instrument is independent of the outcome whenever the treatment is constant ($Z \perp\!\!\!\perp Y|X$) (Balke and Pearl, 1993). This can be seen as an approximation of the exclusion restriction using only observed data. It is sufficient under the assumption that either all common causes U are constant throughout the data measurement period or U is not a common cause for X and Y[1]. However, under any non-trivial common cause U, it will be implausible that Z and Y are independent, because conditioning on X induces a correlation between Z and U. Further, because we ignore the contribution of unobserved confounders, this test is not a necessary condition for a valid instrument.

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:

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

where g and h are arbitrary deterministic functions and U represents unobserved variables that are independent of Z. ϵ_y and ϵ_x are mutually independent error terms. Using this framework, Pearl derived conditions that any observed data generated from a valid instrumental variable model must satisfy (Pearl, 1995). 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 (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. For binary variables Z, X and Y, the instrumental inequalities become:

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

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

1. Here we make the standard assumption of *faithfulness*. That is, observed conditional independence between variables implies causal independence.

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 have been proposed to account for sampling variability in observing the true conditional distributions. For example, a null hypothesis tests based on the chi-squared statistic (Ramsahai and Lauritzen, 2011) or the Kolmogorov-Smirnov test statistic (Kitagawa, 2015) can be applied.

When X, Y and Z are binary, this test is not only necessary, it is the strongest necessary test possible (Bonet, 2001; Kitagawa, 2015). 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. (Bonet, 2001). We refer to this version of the 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 in instrumental variable studies. 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.

2.2 Sensitivity analysis for instrumental variables

Still, the above tests can refute an invalid-IV model, but are unable to *verify* a valid-IV model (Kitagawa, 2015). 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.[2] 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

2. 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 (Bonet, 2001). 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.

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.

As a partial remedy, researchers employ a sensitivity analysis to check the brittleness of a causal estimate when required assumptions are violated. For instance, one may progressively increase the magnitude of violation of the exclusion restriction, and observe when a resultant causal estimate flips in its direction of stated effect (Small, 2007). A Bayesian framework presents an intuitive way to conceptualize sensitivity analysis. The idea is to first estimate a causal effect assuming that the data is generated from a valid-IV causal model. One can then change parameters of the causal model to violate key assumptions and observe how fast the causal estimate changes. As an example, Aart (2010) proposes a Bayesian analysis to check the sensitivity to assumptions for a linear IV model. This analysis shows that even moderate uncertainty in the prior for the exclusion restriction lead to considerable loss of precision in estimating causal effects.

2.3 Developing a sufficient test for instrumental variables

In an ideal world, we would like to reduce uncertainty about assumptions as much as possible and determine precisely whether observed data was generated from a valid-IV model or not. However, as Figure 2 indicates, establishing *sufficiency* for a validity test is non-trivial. In particular, the usual method of comparing the data likelihood of different causal models provides us no information. This is because Invalid-IV class of models (as shown in Figures 1b and 2b) 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. 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.

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 motivation, 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?

Intuitively, we wish to find out how often does the Pearl-Bonet test accept an Invalid-IV model. 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 the probability that a given observed distribution was generated by a valid-IV model, based on the result of Pearl-Bonet test.

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. Any valid instrument needs to pass Pearl-Bonet test. Therefore, NPS test provides necessity: any instrument that fails 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 2 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 (Balke and Pearl, 1993). 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 (Bonet, 2001). 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 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 present 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 our main result that proposes 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, 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 $Z \to Y$ or from $U \to Z$, 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 $Z \to X \to Y$, 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 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, 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.

3.2 Comparing marginal likelihood of Valid-IV and Invalid-IV model classes

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, and result of 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:*

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

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

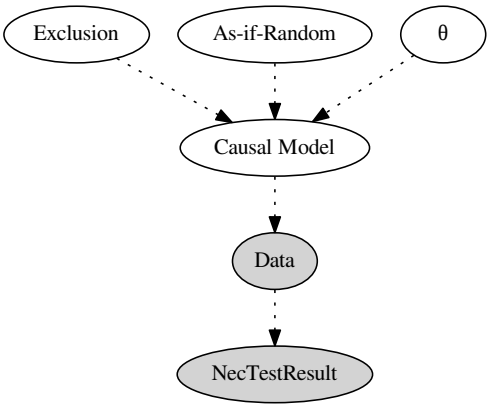


Figure 3: A probabilistic graphical meta-model for describing the connection between IV conditions and specific causal models. Evidence consists of both the results of the necessary test and observed data sample. Therefore, given this evidence, some causal models are expected to become more likely than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

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 Pearl-Bonet necessary test.

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

$$\textit{ML-Ratio} = \frac{P(PT=1, D=d|E,R)}{P(PT=1, D=d|\neg(E,R))} \tag{8}$$

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:

$$\begin{aligned} &P(PT=1, D=1|E,R) \\ &\quad = P(PT=1|D=d,E,R)P(D=d|E,R) \\ &\quad = I_{PT_d}P(D=d|E,R) \end{aligned} \tag{9}$$

where I_{PT_D} is an indicator function that is 1 whenever the test pases for the data sample. 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:

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

Similarly, the denominator can be expressed by,

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

where we use the conditional independencies entailed by the generative meta-model. Now given a data sample d, the result of Pearl-Bonet necessary test PT is deterministic. Therefore, $P(PT|D)$ is 1 whenever the data sample 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:

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

where I_{PT_d} is an indicator function, evaluating to 1 whenever the data sample passes the Pearl-Bonet test.

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

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

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:

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

■

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 given any configuration of these two assumptions. If sufficient data is available, one may use the fractional Bayes Factor () 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 Section 8. Using a uniform model prior leads to the following corollary.

Corollary 2 *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:*

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

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 (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*.

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 (1994) 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 computing the integrals are unique; any other suitable strategy can be used to implement the NPS test.

4.1 The response variable framework

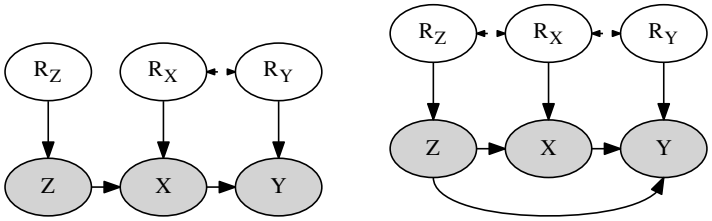
One way to characterize causal models is to assume specific functional relationships between observed variables, such as in the popular linear model for instrumental variables (Imbens and Rubin, 2015). In most cases, however, the nature of the functional form is not known and thus parameterization in 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 relationship between Y and X, $y = k(x)$, where the true causal relationship is $y = f(x, u)$. Conceptually, the variables U are simply additional inputs to this function. Thus the effect of unobserved confounders can be seen as simply transformation the observed relationship k to another function $k'(x) = f(x, u)$. However, it is hard to reason about this transformation because the confounders 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. Because the number of possible functions is finite, the combined effect of unobserved confounders U can be characterized by a finite set of parameters, *without* making any assumptions about U. 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 instead 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 in prior instrumental variable studies (Heckerman and Shachter, 1995).

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 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.

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

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:

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

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 $lm^l n^m$.

Both exclusion and as-if-random are violated

In this case the structural equation for Y is given by Equation 21 and R_Z is not independent of (R_X, R_Y). Thus the dimensions of θ increase to $lm^l n^{lm}$.

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

The response variable framework provides a convenient way to specify an individual causal model: choosing a 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 how to compute the validity ratio for a given observed dataset.

To compute the Validity-Ratio, we return to Equation 2.

$$\textit{Validity-Ratio} = \frac{P(M1)}{P(M2)} \frac{K_2 \int_{M1:m \text{ is valid}} P(D|m)dm}{K_1 \int_{M2:m \text{ is invalid}} P(D|m)dm} \tag{22}$$

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

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 33 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 35 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 36 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 30, assuming $y = f(x, u)$ and sample $P(r_z)$, $P(r_x, r_y)$ separately ;
Compute Validity Ratio as $ML_{VALID}/ML_{INVALID} * PRIOR\text{-}RATIO$

Figure 5: NPS Algorithm for validating an instrumental variable.

the standard assumption of independent data points, we obtain the following likelihood for any causal model m,

$$\begin{aligned} P(D|m) &= \prod_{i=1}^{N} P(Z = z^{(i)}, X = x^{(i)}, Y = y^{(i)}|m) \\ &= (P(Z = 0, X = 0, Y = 0|m))^{R_0} ... P(Z = z_l, X = x_m, Y = y_n|m))^{R_{lmn}} \\ &= \prod_{j=1}^{Q} (P(Z = z_j, X = x_j, Y = y_j|m))^{Q_j} \end{aligned} \tag{23}$$

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. Since the model m can be equivalently represented by its probability vector θ_{R_Z,R_X,R_Y}, we can rewrite the above equation as:

$$\begin{aligned} P(D|m) = P(D|\theta) &= \prod_{j=1}^{Q} (P(Z = z_j, X = x_j, Y = y_j|\theta))^{Q_j} \\ &= \prod_{j=1}^{Q} (\sum_{r_{zxy}=000}^{lmn} P(R_{XYZ} = r_{xyz}) P(Z = z_j, X = x_j, Y = y_j|\theta, r_{zxy}))^{Q_j} \end{aligned} \tag{24}$$

For illustration, we derive the closed form expressions for the numerator and denominator of Equation 22 when all variables are binary, in Appendix A. In general the method works for any number of discrete levels.

Finally, based on the above details, 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. (2016).

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 (2001) 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 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:

$$\begin{aligned} f =& (P(X = x_1, Y = y_1|Z = z_1), \\ & P(X = x_1, Y = y_2|Z = z_1), ... \\ & P(X = x_1, Y = y_m|Z = z_1), \\ & P(X = x_1, Y = y_1|Z = z_2), ... \\ & P(X = x_n, Y = y_m|Z = z_l)) \end{aligned} \tag{25}$$

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 $\mathcal{F}$ Bonet (2001).

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 3 *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:*

$$\sum_{k=1}^{K} \lambda_k . \vec{e_k} = \vec{f}; \qquad \sum_{j=1}^{K} \lambda_j = 1; \qquad \lambda_j \geq 0 \ \forall j \in [1, K] \tag{26}$$

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 26. ■

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 Chaves et al. (2014).

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 3.

As defined in Section 2.1, monotonicity implies that:

$$g(z_1, u) \geq g(z_2, u) \ \textit{whenever} \ z_1 \geq z_2 \tag{27}$$

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 27 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 4 *For any data distribution $P(X, Y, Z)$ generated from a valid-IV model that also satisfies monotonicity, the following inequalities hold:*

$$\begin{aligned} P(Y = y, X \geq x|Z = z_0) &\leq P(Y = y, X \geq x|Z = z_1) \quad ... \quad \leq P(Y = y, X \geq x|Z = z_{l-1}) \quad \forall x, y \\ P(Y = y, X \leq x|Z = z_0) &\geq P(Y = y, X \leq x|Z = z_1) \quad ... \quad \geq P(Y = y, X \leq x|Z = z_{l-1}) \quad \forall x, y \end{aligned} \tag{28}$$

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 of the theorem is in Appendix B. Note that for binary variables, Theorem 4 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, such as by Kitagawa (2015); Huber and Mellace (2015); Wang et al. (2016); Ramsahai and Lauritzen (2011). In this paper we use an exact test proposed by Wang et al. (2016), 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, with the goal of determining the extent to which NPS test can distinguish between a valid and invalid instrumental variable. In general, results of the NPS test are dependent on a given dataset and the associated marginal likelihoods. To obtain a general sense of test characteristics that are independent of a particular dataset, we first ignore the Validity-Ratio test and consider only the necessary test (Pearl-Bonet test). We want to check how likely is it that the instrument is valid, given that it passes the empirical necessary test. This can be interpreted as a simulation where

all causal models are assumed to be equally likely. That is, we assume a uniform likelihood for all causal models (within the respective Valid-IV and Invalid-IV models classes) in generating given data, and so the Validity-Ratio test is uninformative. Such a simulation will provide us a qualitative sense of the efficacy of the necessary test: broadly, under what conditions does it perform well?

Second, to see how the NPS test will perform on individual datasets, we conduct a more realistic simulation by estimating both marginal likelihoods and the result of the necessary test for multiple datasets. Here the results will depend on the choice of datasets. To consider a broad range, we start with an open problem proposed by Palmer et al. (2011) and extend it to construct datasets that span possible relationships between the cause, outcome and instrument.

6.1 Evaluating the necessary test

Our goal is to estimate the probability that an instrument is valid given that it passed the necessary test. Instead of specifying a dataset, we consider simulations over types of datasets that one might encounter in empirical IV studies. That is, we consider different *model configurations* based on key properties of an IV causal model such as the strength of the instrument, monotonicity of relationship between the instrument and outcome, and similarly between the cause and outcome. Under each of these model configurations, we would like to know how likely it is that an instrument is valid, given that it passes the Pearl-Bonet necessary test.

We consider model configurations in each of the three types of violations: exclusion, as-if-random and both violated. For exclusion, we consider two cases—when the instrument has a monotonic effect on the outcome, and when both instrument and cause have a monotonic relationship with outcome—and construct configurations as we vary the strength of these relationships. These model configurations are motivated by encouragement design studies, where it is plausible to assume that a non-decreasing relationship between the instrument and outcome. We also relax the non-decreasing assumption to study more general scenarios. For as-if-random, we consider configurations based on the strength of relationship between confounders and the instrument, or equivalently, between response variables for the instrument on the one hand, and the cause and outcome on the other. Finally, we consider configurations where both exclusion and as-if-random are violated.

For all model configurations, we sample uniformly at random $N = 200$ invalid-IV causal models each. We do so by identifying the specific violation in each configuration and then using the appropriate conditions on response variables from Section 4.1.2. In effect, we sample response variables to sample different causal models.

We assume that the cause, outcome and the instrument are all binary variables. To eliminate errors in estimating the probabilities required by the necessary test (due to sampling variation in simulating a dataset of $< Z, X, Y >$ from each model), we compute exact probabilities from each sampled causal model and use them directly in the necessary test. Given N models for each scenario, we estimate the Validity-Ratio by computing the fraction of invalid-IV causal models that pass the Pearl-Bonet necessary test.

For all simulations below, we assume that Z, X and Y are binary variables. Further, since monotonicity is a required assumption for obtaining the interpretation of a local

average causal effect (Angrist and Imbens, 1994), we also assume monotonicity throughout. We will conduct separate simulations for three kinds of violations of the IV conditions: exclusion only, as-if-random only, and both violated.

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}.

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 estimate the causal effect $X \to Y$, we use the Wald estimator (Wald, 1940), which for binary variables, can be written as (Balke and Pearl, 1993):

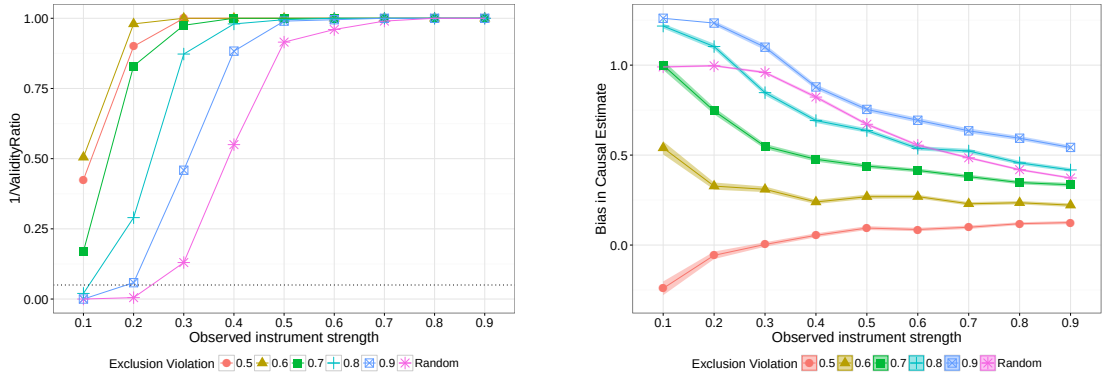
$$\hat{W} = \frac{P(Y=1|Z=1) - P(Y=1|Z=0)}{P(X=1|Z=1) - P(X=1|Z=0)}$$

Because the causal effect between binary variables ranges from $[-1, 1]$, we bound the estimate within this interval. We will compare the effectiveness of the NPS test at different values of *instrument strength*, which is defined as $P(X=1|Z=1) - P(X=1|Z=0)$. Note that the instrument strength also appears in the denominator of the Wald estimator.

6.1.1 Exclusion may be violated

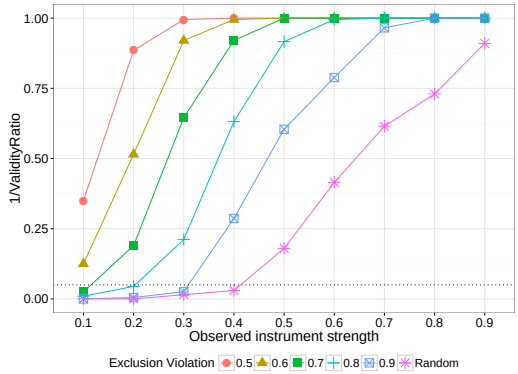
We first test for violation of exclusion only. That is, we assume that as-if-random is satisfied (e.g. through random assignment). As described in Section 4.1.2, violation of the exclusion restriction implies that $y = f(z, x, u)$ and thus there can be 16 different r_y levels. Following the NPS Algorithm,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

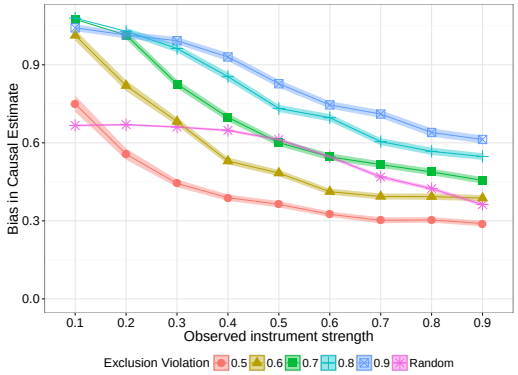


(a) Denominator of Model Ratio

(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.

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 gradually weaken this requirement to obtain a more general scenario. This is motivated by encouragement design IV studies where we may know apriori that 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.

However, in other studies, completely ruling out model configurations where the non-decreasing property does not hold is too strong a condition. 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 from Z to Y is equally likely to be positive or negative.

Extent of non-decreasing effect of instrument on outcome

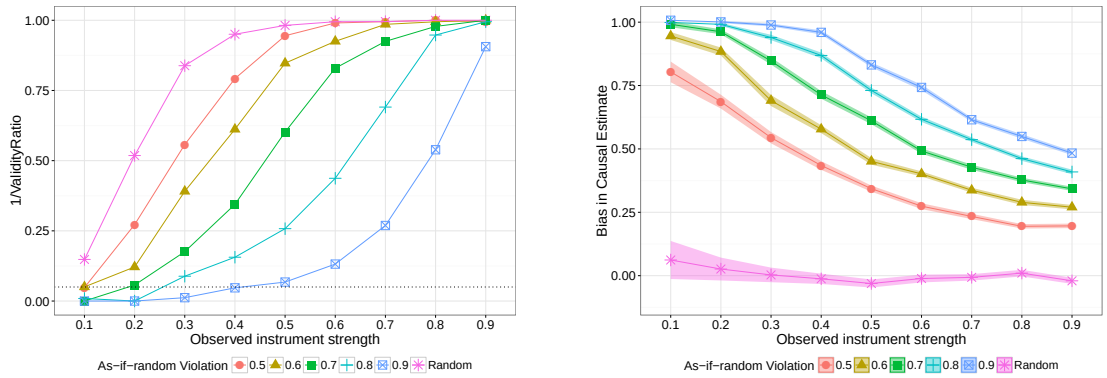
First, we simulate the configurations where exclusion is violated by varying the extent to which Z has a non-decreasing effect on Y. Let us first consider the scenario where the non-decreasing effect is strict: the instrument cannot cause the outcome to decrease. Figure 6a shows the inverse of the Validity-Ratio as we vary observed instrument strength (marked as the 'Random' line). At low instrument strength, less than 5% of invalid-IV models pass the necessary test. Consequently, the Validity-Ratio will be greater than 20. Thus, if we observe empirically that a weak instrument passes the necessary test, it is likely to be a valid instrument (Validity-Ratio$>$ 20) assuming that Z cannot decrease Y and under the uniform likelihood assumption. However, as instrument strength increases, utility of the NPS test decreases.

To relax the non-decreasing assumption, we simulate α-non-decreasing cases such that α fraction of the units have a non-decreasing relationship; the rest do not. When $\alpha = 0.5$, there is an equal chance of Z decreasing or increasing the value of Y. Each of the different lines in Figure 6a shows a different fraction of data points that satisfy the non-decreasing effect criterion. We observe that the power of the NPS test decreases as the fraction of data points satisfying the non-decreasing effect $Z \to Y$ decreases. At a fraction of 0.8, the test is only able to identify correctly invalid-IV models with less than 5% error for instruments with strength less than or equal to 0.1. When non-decreasing relationship does not exist (e.g., $\alpha = 0.5$), the test performs poorly and cannot identify an invalid-IV model reliably. Thus, as we relax the strictness of the non-decreasing effect assumption, more invalid-IV models are passed by the necessary test and thus the NPS test remains inconclusive.

Contrasting these results with estimated bias in the Wald estimate of the causal effect $X \to Y$ 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 is the most discriminative. When the observed strength of the instrument is high, even clearly invalid-IV models lead to comparatively lower bias (less than 0.6).

Extent of non-decreasing effect of both instrument and cause on outcome

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 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 is reached up to a maximum instrument strength of 0.4. The discriminatory power of the NPS test for other scenarios also increases. For thresholds of non-decreasing effect at least 0.7, fraction of incorrectly identified Invalid-IV models lies below 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 differentiating power.



(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, discriminatory power of the NPS test increases and so does bias in the resultant IV estimate.

The lack of NPS test's effectiveness 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. Similarly, 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 (Bound et al., 1995).

6.1.2 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 satisfied, 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 choose $P(r_y = 3|r_z = 1)$ 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 distin-

guish 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 be either 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 discriminatory power of the NPS test increases. When the as-if-random threshold 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 necessary test is unable to detect invalid-IV models more than 50% of the time (and thus the Overall-Validity-Ratio will be deceivingly high).

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

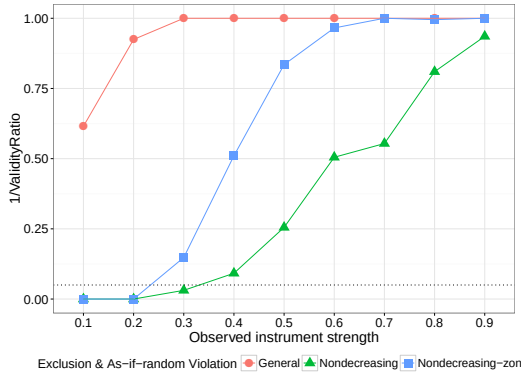
Finally, we consider the case when both conditions may be violated. First, let us consider the straightforward case when both exclusion and as-if-random violating models are uniformly sampled. The red line in Figure 8a shows the power of the NPS test for such invalid-IV models is low; even for weak instrument 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, discriminatory 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 necessary 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.

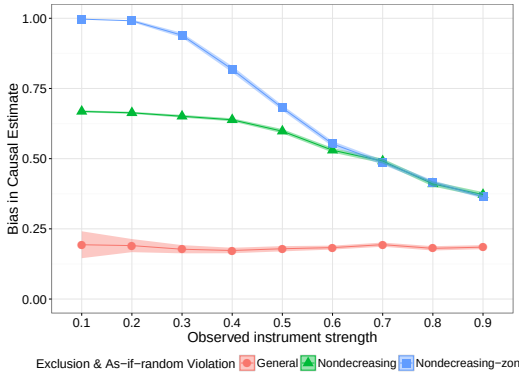
When we modulate the severity of the as-if-random condition (while keeping exclusion violation uniformly at random), the power of the NPS test improves substantially (Figure 8c). For thresholds at least 0.7, the NPS test misses less than 5% of the invalid-IV models at an instrument strength of 0.2. Bias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

6.1.4 Summary of results

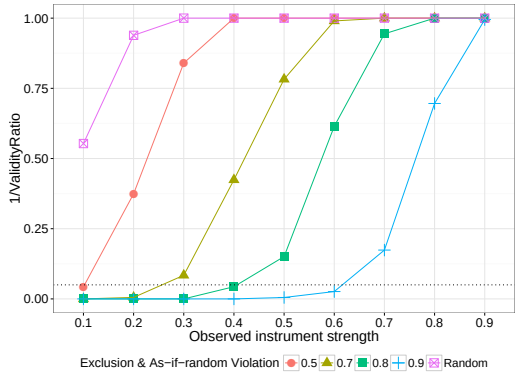
Two key patterns emerge. First, the test is more powerful in recognizing violations that also lead to a substantial bias. This is encouraging because the kind of violations that bias the causal estimate are exactly the invalid-IV datasets we want to eliminate. On average, these results suggest that when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



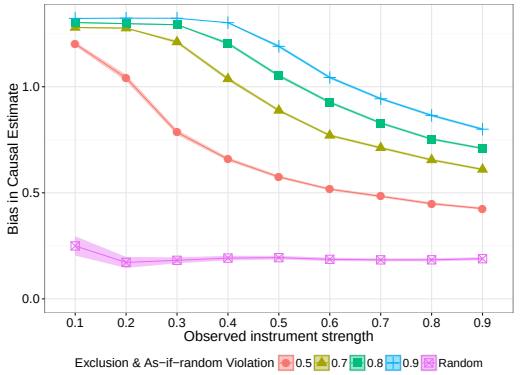
(a) Denominator of Validity Ratio



(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.

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, (Staiger and Stock, 1994) 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 (Pierce et al., 2010). In these low strength regimes, the NPS test can be effective in testing for validity of an instrumental variable.

6.2 An example open problem for binary instrumental variables

While the above results are indicative, they do not provide evidence on how the NPS test will perform on a particular single dataset. We now simulate datasets and check whether the NPS can correctly identify whether they contain a valid instrumental variable or not.

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

$$
\begin{aligned}
Z \sim & Bern(0.5) \\
U \sim & Bern(0.5) \\
X \sim & Bern(p_X); p_X = 0.05 + 0.1Z + 0.1U \\
Y_0 \sim & Bern(p_0); p_0 = 0.1 + 0.05X + 0.1U \\
Y_1 \sim & Bern(p_1); p_1 = 0.1 + 0.2Z + 0.05X + 0.1U \\
Y_2 \sim & Bern(p_2); p_2 = 0.1 + 0.05Z + 0.05X + 0.1U
\end{aligned} \tag{29}
$$

where Z, X, Y_i are the instrument, cause and outcome respectively and all variables are binary. There can be three possible 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 datasets using the NPS test by simulating 2000 data points from each of their causal models. Table 1 shows that comparing Validity-Ratio from the NPS test 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 2. 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 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. In general, to determine validity, one way would be to prespecify standard thresholds for the Validity-Ratio above which an instrument is valid, as suggested by Kass and Raftery (1995). 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. In addition, note that the Validity-Ratio may also be sensitive to the assumption of uniform priors over models within invalid-IV and valid-IV classes. We leave exploration of other priors over models for future work.

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	$3.4\text{x}10^{-14}$

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.

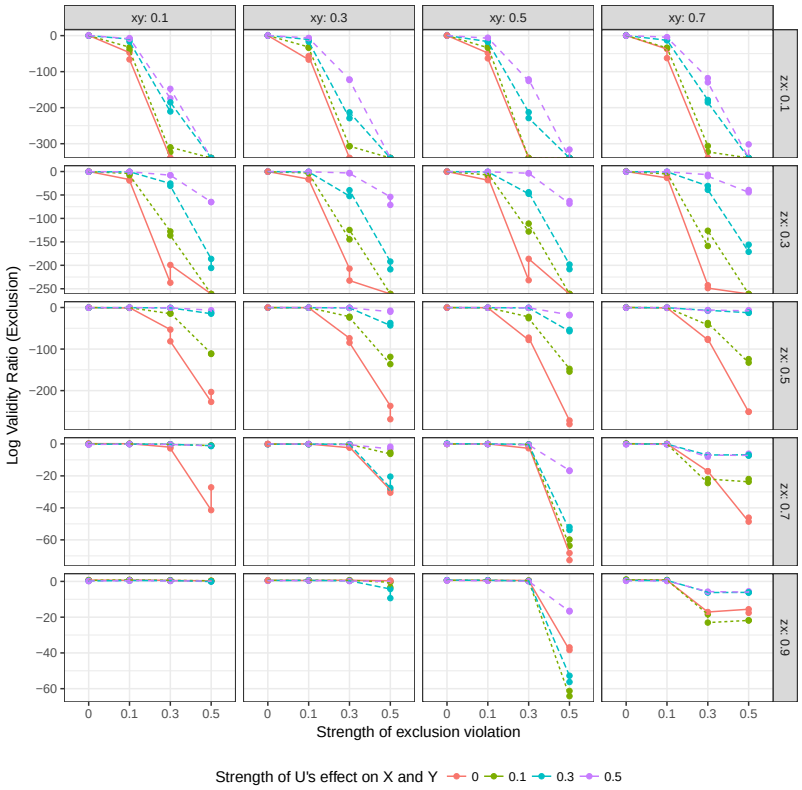


Figure 9: Log Validity-Ratio computed from the NPS test on simulated binary datasets where only exclusion is potentially violated. Strength of exclusion violation increases on the x-axis. *(Rows)* zx denotes the direct effect of Z on X. *(Columns)* xy denotes the direct effect of X on Y.

6.3 Simulating a broad range of binary datasets

Motivated by the example from Palmer et al. (2011), we now construct a set of datasets that cover a broad spectrum of possible datasets with binary variables. We do so 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.

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 apply the NPS algorithm from Figure 5 and compute the Validity-Ratio using the equation from Corollary 2.[3]

6.3.1 NPS test can detect exclusion violation, except when instrument is strong

We first look at violation of the exclusion restriction. For this case, we consider all datasets where as-if-random is satisfied (i.e. effect of U on Z is zero). Figure 9 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 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. In addition, the detectability of exclusion violation depends on the effect of confounders U on X and Y. When confounders do not affect X and Y (red line), Validity Ratio is the most sensitive to violations of exclusion. As the effect of confounders increases, Validity Ratio becomes less sensitive in detecting exclusion violation. In addition to detecting violations, the NPS test also correctly returns a Validity Ratio close to 1 or higher for valid instruments. The mean Validity-Ratio across all valid instruments is 1.8, with a maximum value of 12. There are some valid-IV models for which the Validity-Ratio falls below 1, but in all cases it is higher than 0.3 (or roughly, -1 on the log scale).

Above results indicate that exclusion can be tested by inspecting the Validity-Ratio as long as the instrument is not too strong (effect parameters < 0.5). Further, the NPS test is able to identify violation of the exclusion restriction best when there is little confounding between X and Y. However, this is still a secondary effect and even at a confounding effect of U on X and Y at 0.5, the NPS test can successfully identify instruments that violate exclusion.

6.3.2 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. For this case, we consider all datasets where the exclusion condition is satisfied (i.e., effect of Z on Y is zero). The obtained log Validity-Ratio as the parameter for effect of U on Z is varied is shown in Figure 10. We find that violation of the as-if-random is harder to detect than exclusion.

3. Unlike the NPS algorithm though, we compute the Validity-Ratio for all datasets for comparison, irrespective of whether they pass the Pearl-Bonet necessary test.

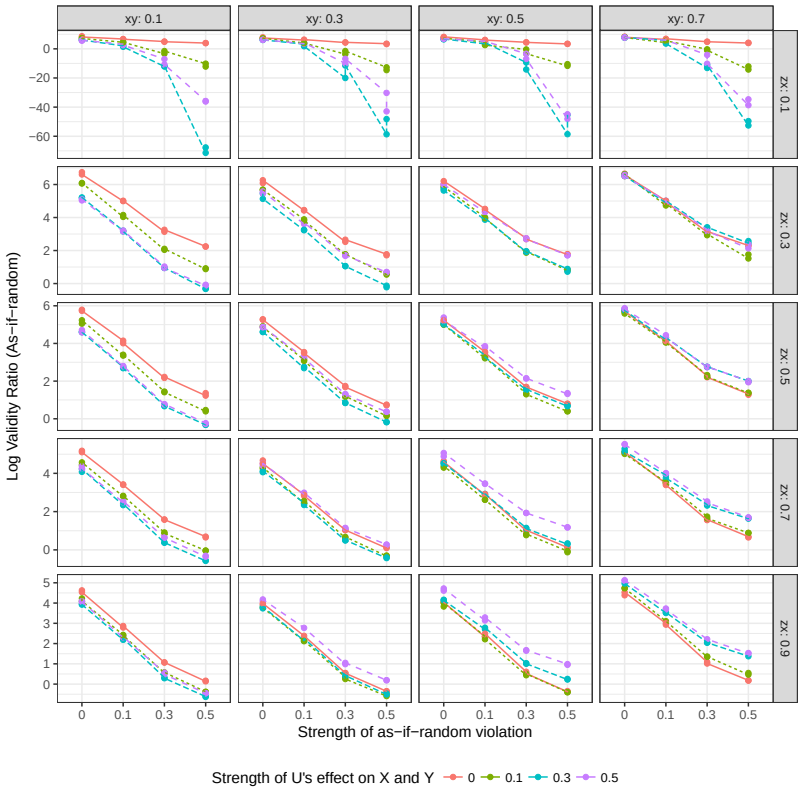


Figure 10: Log Validity-Ratio computed from the NPS test on simulated binary datasets where only as-if-random is potentially violated. Strength of as-if-random violation increases on the x-axis. *(Rows)* zx denotes the direct effect of Z on X. *(Columns)* xy denotes the direct effect of X on Y.

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. Interestingly, the detection of a violation is better when there is a strong confounding effect of U on X and Y, and worse when the confounding effect is near zero. We conjecture that this is because the test is trying to detect an effect of U on Z and it is helpful if U also has a non-trivial on X and Y to gain comparison from. As for the exclusion assumption above, Validity-Ratio for datasets with a valid instrument is close to or higher than 1, indicating that they are probably valid.

However, in the case where instrument strength increases to 0.3 and above, the Validity Ratio stays above 1 even when as-if-random is violated. As we found in Section 6.1.2, these results indicate that the NPS test is unable to detect violations of the as-if-random assumption. As in that section, it is also likely that violations of as-if-random assumption lead to lesser bias in causal estimation than the exclusion assumption.

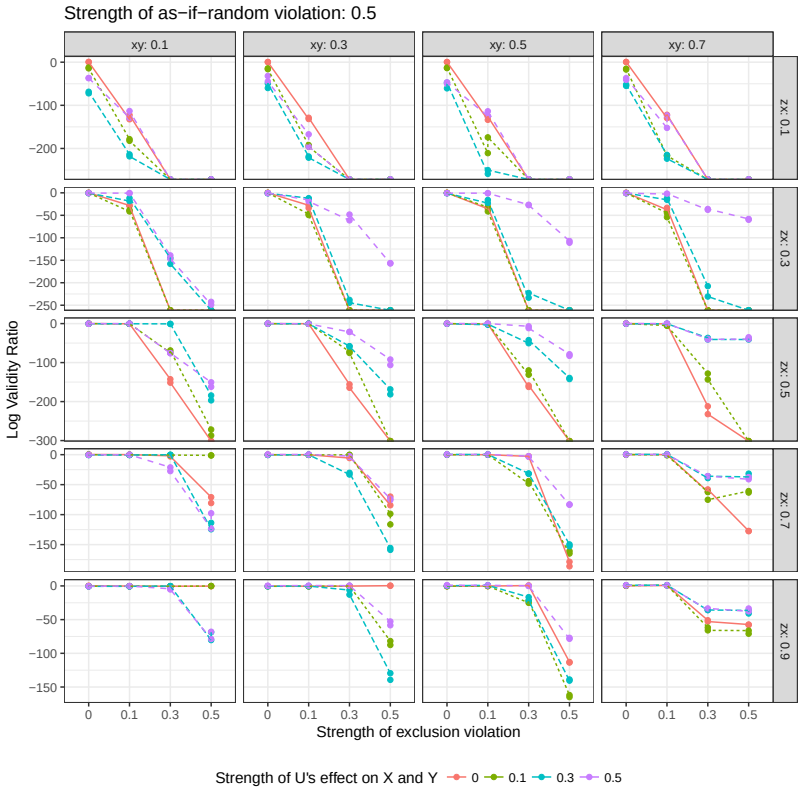
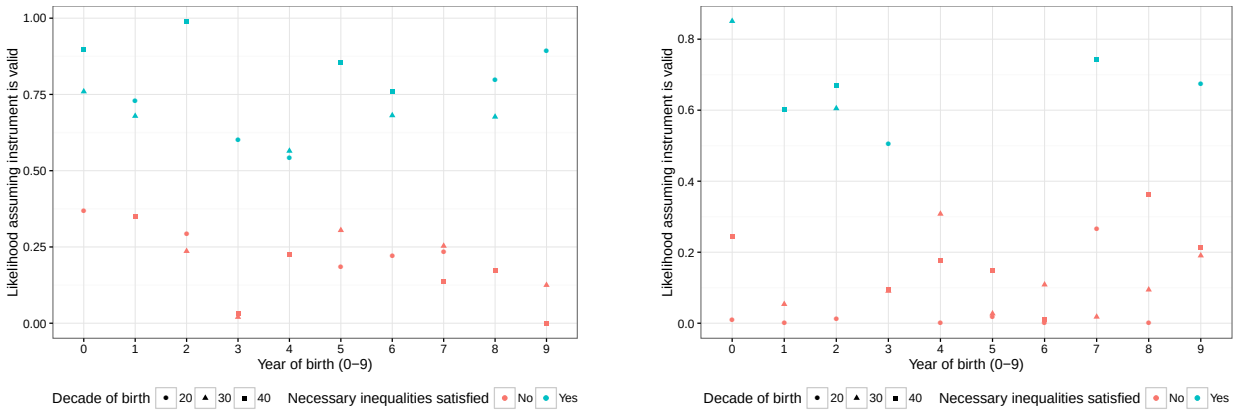


Figure 11: Log Validity-Ratio computed from the NPS test on simulated binary datasets where both exclusion and as-if-random are potentially violated. Strength of as-if-random violation is fixed at 0.5. *(Rows)* zx denotes the direct effect of Z on X. *(Columns)* xy denotes the direct effect of X on Y.

6.3.3 Violations of both assumptions is easier to detect

Finally, we look at the case when both exclusion and as-if-random are violated. Here we considered all simulated datasets. Figure 14 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 dropts to less than 1 as the strength of exclusion violation increases. This pattern is consistent as the true causal effect of 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. Detection of invalid instruments becomes weaker as the strength of as-if-random violation is decreased. We include results for other values of the as-if-random violation (i.e., effect of U on Z) in the Appendix.

Overall, these results show that the NPS test can detect violations of exclusion, but violations of as-if-random are not detected as reliably. Further, if both exclusion and as-if-random are violated, the NPS test can detect them more easily. In all cases, distinguishing



(a) Unconditional IV

(b) Conditional IV

Figure 12: p-values from the Pearl-Bonet necessary test for instruments used in a past IV study. Left panel shows unconditional instruments, while the right panel instruments conditioned on relevant covariates. Assuming a p-value threshold of 0.05, many of the instruments are determined to be invalid by the test.

between an invalid and valid instrument is more reliable when the instrument's effect on the cause is weak, as is the case with many observational studies. Finally, for all datasets that contained a valid instrument, the NPS test correctly returns a Validity Ratio close to or higher than 1, thus indicating the probable validity of the instrument.

7. USING NPS TEST TO VALIDATE PAST IV STUDIES

In this section we use the NPS test to evaluate empirical studies based on instrumental variables. We select two seminal and highly cited studies on instrumental variables and a sample of recent studies from a leading economics journal, *American Economic Review*.

7.1 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 (Angrist and Krueger, 1991; Card, 1993). These studies are both highly cited, yet concerns about their validity continue until today (Bound et al., 1995; Buckles and Hungerman, 2013). We show how the NPS test can provide evidence and help evaluate the instruments used in these studies.

7.1.1 Effect of compulsory schooling on future wages

The first paper estimates the effect of compulsory schooling on future earnings of students Angrist and Krueger (1991). 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. Applying the NPS Algorithm shows that nearly half of the instruments do not satisfy the inequalities of the necessary test. However, some of these may be due to sampling variability. At a significance level of $\alpha = 0.05$, 3 yearly instruments do not pass the necessary test, as Figure 12a 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. The null hypothesis in the necessary test is that an instrument is valid.

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" technique (Baum et al., 2007). 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 12b 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 under the transformation when treatment and outcome are binarized. Note that these results do not necessarily invalidate the analysis in the paper: the instruments in the original data may still be valid when X and Y are continuous, as binarizing the variables can also lead to violation of one of the assumptions.

We also estimate the Validity Ratio for the instruments that pass the necessary test. We find that more than one quarter of the instruments have a Validity-Ratio lower than 0.001, with the minimum and maximum validity ratio being $1.1\text{x}10^{-}7$ and 0.5 respectively. Thus, while some of the instruments may be probably valid, the NPS test indicates that many others are likely to be invalid. In this case, the NPS test can be used to select the top-ranked probably valid instruments for analysis, and filter out the rest.

7.1.2 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 (Card, 1993). 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, we find that data from Card's study does not pass the necessary test for IV validity.

4. Dataset available at `http://economics.mit.edu/faculty/angrist/data1/data/angkru1991`
5. Dataset available at `http://davidcard.berkeley.edu/data_sets.html`

Study name	Num. Observations	IV Strength	Pearl-Bonet Test	Validity Ratio
Randomized Experiment				
National Job Training Partnership Act (JTPA) Study (Abadie et al., 2002)				
	5102	0.58	Pass	3.4
Instrumental Variable Studies				
Effect of rural electrification on employment in South Africa (2011) (Dinkelman, 2011)				
–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) (David et al., 2013)				
–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) (Favara and Imbs, 2015)				
–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) Decarolis (2015)				
–Unconditioned	170	0.60	Pass	1.02
–Conditioned	170	0.42	Pass	0.04
Effect of Mexican immigration on crime in United States (2015) (Chalfin, 2015)				
–Unconditioned	182	0.50	Pass	0.07
–Conditioned	182	0.22	Pass	0.005

Table 2: Results of the NPS test on recent studies from the American Economic Review. Many of the instruments fail the necessary test and have comparatively low Validity-Ratios, indicating that they may be invalid, at least under the binary transformation.

This is corroborated by estimating the Validity-Ratio for the same data. For both mean and median split for binarization, the Validity Ratio is lower than 0.003.

However, in their analysis, Card also considers a conditional instrumental variable that conditions on a number of secondary variables, such as race and geography. When we use the partialling out technique from above to condition on these variables and rerun the NPS test, we find that the dataset passes the necessary test and yields a Validity Ratio of 0.2, indicating a higher likelihood of being valid than the unconditional instrument.

7.2 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 (Dinkelman, 2011), credit supply (Favara and Imbs, 2015), subsidy manipulation (Decarolis,

2015), foreign import competition (David et al., 2013), and foreign immigration (Chalfin, 2015). As a comparison benchmark, we also include an instrumental variable study based on data from a randomized experiment (Abadie et al., 2002), 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 evidence for a probably 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 indicates a measure of their probable validity. As an example, the conditional instrument for the study on Mexican immigration obtains a Validity Ratio of 0.005, providing evidence for the invalidity of the instrument (i.e., the data does not support validity of the instruments used, at least under the transformation of binary variables). In the remaining two studies the Validity Ratio is close to 1 so we cannot reject the validity of the instrument using the NPS test, 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 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. Therefore, while the NPS test cannot always verify whether an instrument is valid, it is more effective when the instrument is weak. Fortunately, many observational studies are based on instruments with low Z-X correlation, where NPS can be applied.

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 but it is possible to choose other priors. If sufficient data is available, a possible method is to split the sample into two and use the first sample to estimate relative likelihoods of different models. This estimated likelihood can then be used as the prior over models in the estimation phase. It will be useful to study the sensitivity of the Validity-Ratio to changes in this prior, which we leave for future work. Second, the proposed implementation of the NPS test works only for discrete variables. Extensions to continuous variables can increase the applicability of this test. 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. In addition, we would also like to study if there is a natural threshold that can be set for identifying valid instruments in such cases.

More generally, the NPS test is an example of a general Bayesian testing framework for 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 (Sharma et al., 2016).

Acknowledgments

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.

Appendix A

Details on computing the Validity Ratio.

Calculating the numerator

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

$$\begin{aligned} P(D|\theta) &= \prod_{j=1}^{Q}\Big(\sum_{r_{zxy}='000'}^{'133'} P(R_Z = r_Z)(P(Z = z_j|\theta, r_z)P(R_{XY} = r_{xy})(P(X = x_j, Y = y_j|\theta, r_{zxy}))^{Q_j} \\ &= \prod_{j=1}^{Q}\Big(\sum_{r_z='0'}^{'1'} P(R_Z = r_Z)(P(Z = z_j|\theta, r_z))^{Q_j}\Big(\sum_{r_{xy}='00'}^{'33'} P(R_{XY} = r_{xy})(P(X = x_j, Y = y_j|\theta, r_{xy}))^{Q_j} \end{aligned} \tag{30}$$

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:

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

The above equation leads to the following simplification for the numerator of Equation 22.

$$\int_{M1:m \text{ is valid}} P(D|m)dm = \iint_{\theta_{rz},\theta_{rxy}} \prod_{j=1}^{Q} \theta_{rz=z}^{Q_j}(\theta_{rxy=a} + \theta_{rxy=b} + \theta_{rxy=c} + \theta_{rxy=d})^{Q_j} d\theta_{rz} d\theta_{rxy}$$
$$= \int \prod_{j=1}^{Q} \theta_{rz=z}^{Q_j} d\theta_{rz} \int \prod_{j=1}^{Q} (\theta_{rxy=a} + \theta_{rxy=b} + \theta_{rxy=c} + \theta_{rxy=d})^{Q_j} d\theta_{rxy} \tag{32}$$

The above integral has a form equivalent to the hyperdirichlet integral Hankin et al. (2010), for which no tractable closed form exists except in a few special cases.[6] 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. (Genz and Cools, 2003) For discrete variables, monte carlo methods for estimating marginal likelihood, such as annealed importance sampling (Neal, 2001) or nested sampling (Skilling et al., 2006; Feroz et al., 2009) may be more appropriate.

Calculating the denominator

We can calculate the denominator of Equation 22 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 22, the denominator can be expressed similarly to 30. The only difference is that θ_{rxy} will be 4x16=64-dimensional integral.

$$P(D|\theta) = \prod_{j=1}^{Q} (\sum_{r_z='1'} P(R_Z = r_Z)(P(Z = z_j|\theta, r_z))^{Q_j} (\sum_{r_{xy}='3,15'} P(R_{XY} = r_{xy})(P(X = x_j, Y = y_j|\theta, r_{xy}))^{Q_j} \tag{33}$$

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.

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 24, we cannot simplify the marginal likelihood

6. We say *tractable* because it is possible to decompose the hyperdirichlet integral into a sum of exponentially many dirichlet integrals Cooper and Herskovits (1992), but that will not be computationally feasible.

as a product of two independent integrals and thus obtain:

$$\int P(D|\theta)d\theta = \int_{M2} \prod_{j=1}^{Q} (\sum_{r_{zxy}=0,0,0}^{1,3,3} P(R_{XYZ} = r_{xyz}) P(Z = z_j, X = x_j, Y = y_j|\theta, r_{zxy}))^{Q_j} \tag{34}$$

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,

$$\int P(D|\theta)d\theta = \frac{\prod_{j=1}^{Q} \Gamma(4 + Q_j)}{(\Gamma(4))^Q \Gamma(\sum_{j=1}^{Q} \Gamma(4 + Q_j)))} \tag{35}$$

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 34), except that the number of dimensions of theta increases to 2x4x16=128. The denominator of the Validity Ratio can be evaluated as:

$$\begin{aligned} \int P(D|\theta)d\theta &= \int_{M2} \prod_{j=1}^{Q} (\sum_{r_{zxy}=0,0,0}^{1,3,15} P(R_{XYZ} = r_{xyz}) P(Z = z_j, X = x_j, Y = y_j|\theta, r_{zxy}))^{Q_j} \\ &= \frac{\prod_{j=1}^{Q} \Gamma(16 + Q_j)}{(\Gamma(16))^Q \Gamma(\sum_{j=1}^{Q} \Gamma(16 + Q_j)))} \end{aligned} \tag{36}$$

Appendix B

Theorem 5 *For any data distribution $P(X, Y, Z)$ generated from a valid-IV model that also satisfies monotonicity, the following inequalities hold:*

$$\begin{aligned} &P(Y = y, X \geq x|Z = z_0) \leq P(Y = y, X \geq x|Z = z_1) \quad ... \quad \leq P(Y = y, X \geq x|Z = z_{l-1}) \quad \forall x, y \\ &P(Y = y, X \leq x|Z = z_0) \geq P(Y = y, X \leq x|Z = z_1) \quad ... \quad \geq P(Y = y, X \leq x|Z = z_{l-1}) \quad \forall x, y \end{aligned} \tag{37}$$

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 $z \in \{z_0, z_1, ...z_{l-1}\}$:

$$P(X \geq x|Z = z_k) = P(R_X : g(z_k, u) \geq x) \tag{38}$$

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

$$g(z_{k1}, u) \geq x \Rightarrow g(z_{k2}, u) \geq x \qquad \text{if } z_{k2} \geq z_{k1} \tag{39}$$

Combining Equations 38 and 39, 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 $P(Y = y, X \leq x|Z = z_k)$.

■

Appendix C

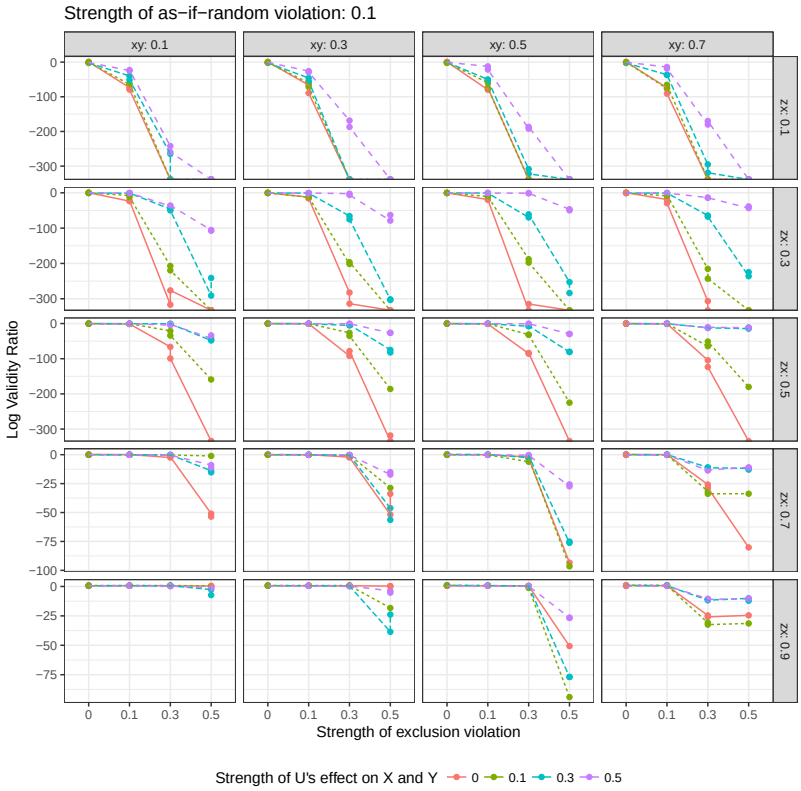


Figure 13: Log Validity-Ratio computed from the NPS test on simulated binary datasets where both exclusion and as-if-random are potentially violated. Strength of as-if-random violation is fixed at 0.1 *(Rows)* zx denotes the direct effect of Z on X. *(Columns)* xy denotes the direct effect of X on Y.

References

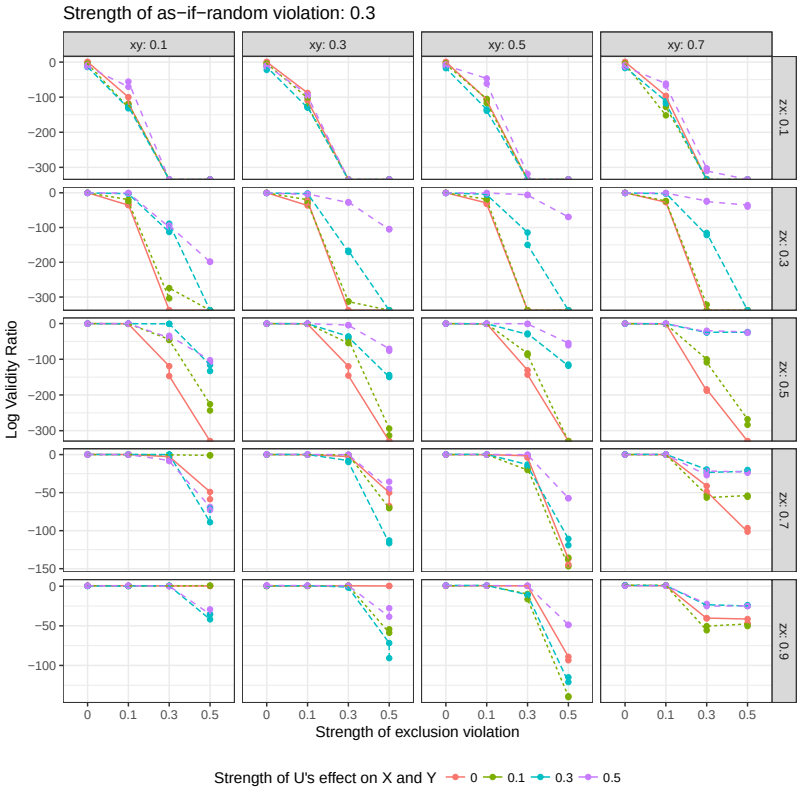
Kraay Aart. Instrumental variables regressions with uncertain exclusion restrictions: a bayesian approach. *Journal of Applied Econometrics*, 27(1):108–128, 2010. doi: 10.1002/jae.1148. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1148`.

Alberto Abadie, Joshua Angrist, and Guido Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70 (1):91–117, 2002.

Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects. *Econometrica*, 1994.

Joshua D Angrist and Alan B Krueger. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014, 1991.

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



]

Figure 14: Log Validity-Ratio computed from the NPS test on simulated binary datasets where both exclusion and as-if-random are potentially violated. Strength of as-if-random violation is fixed at 0.3. *(Rows)* zx denotes the direct effect of Z on X. *(Columns)* xy denotes the direct effect of X on Y.

Alexander Balke and Judea Pearl. Nonparametric bounds on causal effects from partial compliance data. *Technical report, UCLA*, 1993.

Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. *Proc. of AAAI*, 1994.

Christopher F Baum, Mark E Schaffer, Steven Stillman, et al. Enhanced routines for instrumental variables/gmm estimation and testing. *Stata Journal*, 7(4):465–506, 2007.

Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 48–55. Morgan Kaufmann Publishers Inc., 2001.

John Bound, David A Jaeger, and Regina M Baker. 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(430):443–450, 1995.

Kasey S Buckles and Daniel M Hungerman. Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics*, 95(3):711–724, 2013.

David Card. Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research, 1993.

Aaron Chalfin. The long-run effect of mexican immigration on crime in us cities: Evidence from variation in mexican fertility rates. *The American Economic Review*, 105(5):220–225, 2015.

Rafael Chaves, L Luft, TO Maciel, D Gross, D Janzing, and Bernhard Schölkopf. Inferring latent structures via information inequalities. *arXiv preprint arXiv:1407.2256*, 2014.

Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

H David, David Dorn, and Gordon H Hanson. The china syndrome: Local labor market effects of import competition in the united states. *The American Economic Review*, 103 (6):2121–2168, 2013.

Francesco Decarolis. Medicare part d: are insurers gaming the low income subsidy design? *The American Economic Review*, 105(4):1547–1580, 2015.

Taryn Dinkelman. The effects of rural electrification on employment: New evidence from south africa. *The American Economic Review*, 101(7):3078–3108, 2011.

Thad Dunning. *Natural experiments in the social sciences: a design-based approach*. Cambridge University Press, 2012.

Giovanni Favara and Jean Imbs. Credit supply and the price of housing. *The American Economic Review*, 105(3):958–992, 2015.

F Feroz, MP Hobson, and M Bridges. Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614, 2009.

Alan Genz and Ronald Cools. An adaptive numerical cubature algorithm for simplices. *ACM Transactions on Mathematical Software (TOMS)*, 29(3):297–308, 2003.

Robin KS Hankin et al. A generalization of the dirichlet distribution. *Journal of Statistical Software*, 33(11):1–18, 2010.

David Heckerman and Ross Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, pages 405–430, 1995.

Martin Huber and Giovanni Mellace. Testing instrument validity for late identification based on inequality moment constraints. *Review of Economics and Statistics*, 97(2):398–411, 2015.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

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

Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2014.

Alice Nakamura and Masao Nakamura. On the relationships among several specification error tests presented by durbin, wu, and hausman. *Econometrica: journal of the Econometric Society*, pages 1583–1588, 1981.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Tom M Palmer, Roland R Ramsahai, Vanessa Didelez, Nuala A Sheehan, et al. Nonparametric bounds for the causal effect in a binary instrumental-variable model. *Stata Journal*, 11(3):345, 2011.

Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 435–443. Morgan Kaufmann Publishers Inc., 1995.

Brandon L Pierce, Habibul Ahsan, and Tyler J VanderWeele. Power and instrument strength requirements for mendelian randomization studies using multiple genetic variants. *International journal of epidemiology*, page dyq151, 2010.

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

Amit Sharma, Jake M Hofman, and Duncan J Watts. Split-door criterion for causal identification: Automatic search for natural experiments. *arXiv preprint arXiv:1611.09414*, 2016.

John Skilling et al. Nested sampling for general bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.

Dylan S Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007. doi: 10.1198/016214507000000608. URL `https://doi.org/10.1198/016214507000000608`.

Douglas O Staiger and James H Stock. Instrumental variables regression with weak instruments, 1994.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Abraham Wald. The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300, 1940.

Linbo Wang, James M Robins, and Thomas S Richardson. On falsification of the binary instrumental variable model. *arXiv preprint arXiv:1605.03677*, 2016.