Identification and Model Testing in Linear Structural Equation Models using Auxiliary Variables

Bryant Chen [1] **Daniel Kumor** [2] **Elias Bareinboim** [2]

Abstract

We developed a novel approach to identification and model testing in linear structural equation models (SEMs) based on auxiliary variables (AVs), which generalizes a widely-used family of methods known as instrumental variables. The identification problem is concerned with the conditions under which causal parameters can be uniquely estimated from an observational, non-causal covariance matrix. In this paper, we provide an algorithm for the identification of causal parameters in linear structural models that subsumes previous state-of-the-art methods. In other words, our algorithm identifies strictly more coefficients and models than methods previously known in the literature. Our algorithm builds on a graph-theoretic characterization of conditional independence relations between auxiliary and model variables, which is developed in this paper. Further, we leverage this new characterization for allowing identification when limited experimental data or new substantive knowledge about the domain is available. Lastly, we develop a new procedure for model testing using AVs.

1. Introduction

The problem of estimating causal effects is one of the fundamental problems in the data-driven sciences. In order to estimate a causal effect, the desired effect must be *identified* or uniquely expressible in terms of the probability distribution over the available data. Causal effects are identified by design in randomized control trials, but in many applications, such experiments are not possible. When only observational data is available, determining whether a causal effect is identified requires modeling the underlying causal structure, which is generally done using *structural equation models* (SEMs) (also called *structural causal models*) (Pearl, 2009; Bareinboim and Pearl, 2016).

A structural equation model consists of a set of equations that describe the underlying data-generating process for a set of variables. While SEMs, in their most general, non-parametric form do not require any assumptions about the form of these functions, in many fields, including machine learning, psychology, and the social sciences, linear SEMs are used. A linear SEM consists of a set of equations of the form, $X = \Lambda X + U$, where $X = [x_1, ..., x_n]^t$ is a vector containing the model variables, Λ is a matrix containing the *coefficients* of the model, and Λ_{ij} represents the direct effect of x_i on x_j, and $U = [u_1, ..., u_n]^t$ is a vector of normally distributed error terms, which represents omitted or latent variables.[1] The matrix Λ contains zeroes on the diagonal, and $\Lambda_{ij} = 0$ whenever x_i is not a cause of x_j. The covariance matrix of X will be denoted by Σ and the covariance matrix over the error terms, U, by Ω. In this paper, we will restrict our attention to *semi-Markovian* models (Pearl, 2009), models where the rows of Λ can be arranged so that it is lower triangular, and the corresponding graph is acyclic.

When modeling using SEMs, researchers typically specify the model by setting certain entries of Λ and Ω to zero (i.e. exclusion and independence restrictions), while leaving the rest of the entries as free parameters to be estimated from data[2]. Restricting a particular entry Λ_{ij} to zero reflects the assumption that Y_i has no direct effect on Y_j. Similarly, restricting Ω_{ij} to zero reflects the assumption that there are no unobserved common causes of both Y_i and Y_j. Once the

[1]IBM Research, San Jose, California, USA [2]Purdue University, West Lafeyette, Indiana, USA. Correspondence to: Bryant Chen <bryant.chen@ibm.com>, Daniel Kumor <dkumor@purdue.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]Instrumental and auxiliary variables can also be used when normality is not assumed, but to simplify the proofs in the paper, we will, as is commonly done by empirical researchers, assume normality.

[2]There are a number of algorithms for discovering the model structure from data(Spirtes et al., 2000; Shimizu et al., 2006; Pearl, 2009; Zhang and Hyvärinen, 2009; Mooij et al., 2016). However, it is only in very rare instances that these methods are able to uniquely determine the model structure. As a result, model specification generally utilizes knowledge about the domain under study.

parameters are estimated, causal effects (as well as counterfactual quantities) can be computed from the structural coefficients directly (Pearl, 2009; Chen and Pearl, 2014). However, in order to be estimable from data, a parameter must first be identified. In some cases, the modeling assumptions are not strong enough, and there are multiple, often infinite, values for the parameter that are consistent with the observed data. As a result, two fundamental problems in SEMs are to identify and estimate the model parameters and to test the underlying assumptions that enable identification.

The problem of identification has been studied extensively by econometricians and social scientists (Fisher, 1966; Bowden and Turkington, 1984; Bekker et al., 1994; Rigdon, 1995) and more recently by the AI and statistics communities using graphical methods (Spirtes et al., 1998; Tian, 2007; 2009; Brito and Pearl, 2002a;c; 2006; Bareinboim and Pearl, 2016). To our knowledge, the most general, efficient algorithm for model identification is the g-HT algorithm given by Chen (2016) combined with ancestor decomposition (Drton and Weihs, 2016). This method generalizes the half-trek algorithm of Foygel et al. (2012) and utilizes ancestor decomposition, which expands on an idea by Tian (2005) where the model is decomposed into simpler sub-models. Graphical methods have also been applied to the problem of testing the causal assumptions embedded in an SEM. For example, d-separation (Pearl, 2009) and overidentification (Pearl, 2004; Chen et al., 2014) provide the means to discover testable implications of the model, which can be used to test it against data.

Despite decades of attention and work from diverse fields, the identification problem[3] has still not been efficiently solved[4]. There are identifiable parameters and models that none of the above methods are able to identify. Similarly, there are testable implications of SEMs that the above methods are unable to detect. One promising avenue to aid in both tasks are *auxiliary variables* (Chen et al., 2016). Each of the aforementioned methods for identification and model testing only utilizes restrictions on the entries of Λ and Ω to zero. Auxiliary variables can be used to incorporate knowledge of non-zero coefficient values into existing methods for identification and model testing. These coefficient values could be obtained, for example, from a previously conducted randomized experiment, from substantive understanding of the domain, or even from another identification technique. The intuition behind auxiliary variables is simple: if the coefficient from variable w to z, β, is known, then we would like to remove the direct effect of w on z by subtracting it from z. This removal eliminates confounding paths through w and is performed by creating a variable $z^* = z - \beta w$, which is used as a proxy for z. In many cases, z^* allows the identification of parameters or testable implications using existing methods when z could not.

Chen et al. (2016) demonstrated how auxiliary variables could be utilized in simple instrumental sets (instrumental sets that do not utilize conditioning to block spurious paths) (Brito and Pearl, 2002a; van der Zander et al., 2015) and proved that any model identifiable using the g-HT algorithm is also identifiable using auxiliary simple instrumental sets.

Since auxiliary variables allow knowledge of non-zero coefficient values to be incorporated into existing methods for identification, they are also directly applicable to the problem of z-identification (Bareinboim and Pearl, 2012), in which partial experimental data is available. Additionally, the cancellation of paths that results from adding an AV may result in conditional independence constraints between the AV and other variables that can be used to test the model.

In this paper, we generalize the results of Chen et al. (2016) and demonstrate how auxiliary variables can be utilized in generalized instrumental sets, which allow for conditioning to block spurious paths. We prove that, unlike auxiliary simple instrumental sets, this generalization *strictly* subsumes the g-HT algorithm. Additionally, we introduce quasi-instrumental sets, which utilize auxiliary variables to identify coefficients when partial experimental data is available. Quasi-instrumental sets are incorporated into our identification algorithm, allowing it to better address the problem of z-identification. To our knowledge, this algorithm is the first systematic method for tackling z-identification in linear systems. We also demonstrate how auxiliary instrumental sets and quasi-instrumental sets can be used to derive over-identifying constraints, which can be used to test the model specification against data. Moreover, we prove that these overidentifying constraints subsume conditional independence constraints among auxiliary variables. Lastly, we discuss related work, showing how auxiliary IVs are able to unite a variety of disparate methods under a single framework.

2. Preliminaries

The causal graph or path diagram of an SEM is a graph, $G = (V, D, B)$, where V are nodes or vertices, D directed edges, and B bidirected edges. The nodes represent model variables. Directed eges encode the direction of causal-

[3]To be precise, we are referring to the problem of identification almost everywhere (Brito and Pearl, 2002b), also called generic identification (Foygel et al., 2012).

[4]An exhaustive procedure can be obtained using Gröbner bases methods (Foygel et al., 2012). However, these methods are computationally intractable for anything but the smallest of graphs.

ity, and for each coefficient $\Lambda_{ij} \neq 0$, an edge is drawn from x_i to x_j. Each directed edge, therefore, is associated with a coefficient in the SEM, which we will often refer to as its structural coefficient. Additionally, when it is clear from context, we may abuse notation slightly and use coefficients and directed edges interchangeably. The error terms, u_i, are not shown explicitly in the graph. However, a bidirected edge between two nodes indicates that their corresponding error terms may be statistically dependent while the lack of a bidirected edge indicates that the error terms are independent.

We will use standard graph terminology with $Pa(y)$ denoting the parents of y, $Anc(y)$ denoting the ancestors of Y, $De(y)$ denoting the descendants of y, and $Sib(y)$ denoting the siblings of y, the variables that are connected to y via a bidirected edge. $He(E)$ denotes the heads of a set of directed edges, E, while $Ta(E)$ denotes the tails. Additionally, for a node v, the set of edges for which $He(E) = v$ is denoted $Inc(v)$. Lastly, we will utilize d-separation (Pearl, 2009).

We will use $\sigma(x, y|W)$ to denote the partial covariance between two random variables, x and y, given a set of variables, W, and $\sigma(x, y|W)_G$ as the partial covariance between random variables x and y given W implied by the graph G. We will assume without loss of generality that the model variables have been standardized to mean 0 and variance 1.

Definition 1. *For a given unblocked (given the empty set) path, π, from x to y, Left(π) is the set of nodes, if any, that has a directed edge leaving it in the direction of x in addition to x. Right(π) is the set of nodes, if any, that has a directed edge leaving it in the direction of y in addition to y.*

For example, consider the path $\pi = x \leftarrow v_1^L \leftarrow ... \leftarrow v_k^L \leftarrow v^T \rightarrow v_j^R \rightarrow ... \rightarrow v_1^R \rightarrow y$. In this case, Left($\pi$) $= \cup_{i=1}^{k} v_i^L \cup \{x, v^T\}$ and Right(π) $= \cup_{i=1}^{j} v_i^R \cup \{y, v^T\}$. v^T is a member of both Right(π) and Left(π).

Definition 2. *A set of paths, $\pi_1, ..., \pi_n$, has* no sided intersection *if for all $\pi_i, \pi_j \in \{\pi_1, ..., \pi_n\}$ such that $\pi_i \neq \pi_j$, Left(π_i)$\cap$Left(π_j)=Right(π_i)$\cap$Right(π_j) $= \emptyset$.*

Wright's rules (Wright, 1921) allow us to equate the model-implied covariance, $\sigma(x, y)_M$, between any pair of variables, x and y, to the sum of products of parameters along unblocked paths between x and y.[5] Let $\Pi = \{\pi_1, \pi_2, ..., \pi_k\}$ denote the unblocked paths between x and y, and let p_i be the product of structural coefficients along path π_i. Then the covariance between variables x and y is $\sum_i p_i$.

Lastly, we define auxiliary variables and the augmented graph.

Definition 3 (Auxiliary Variable)**.** *Given a linear SEM with graph G and a set of edges E whose coefficient values are known, an* auxiliary variable *is a variable, $z^* = z - \sum_i e_i t_i$, where $\{e_1, ..., e_k\} \subseteq E \cap Inc(z)$ and $t_i = Ta(e_i)$ for all $i \in \{1, ..., k\}$.*

If not otherwise specified, z^* refers to the auxiliary variable, $z - c_1 t_1 - ... - c_l t_l$, where $\{c_1, ..., c_l\}$ are the coefficients of $E \cap Inc(z)$ and E is the set of directed edges whose coefficient values are known. In other words, z^* is the auxiliary variable for z where as many known coefficients are subtracted out as possible. Chen et al. (2016) demonstrated that the covariance between any auxiliary variables and model variables can be computed using Wright's rules on the *augmented graph*, defined below.

Definition 4. *(Chen et al., 2016) Let M be a linear SEM with graph G and a set of directed edges E such that their coefficient values are known. The E-*augmented *model, M^{E+}, includes all variables and structural equations of M in addition to new auxiliary variables, $y_1^*, ...y_k^*$, one for each variable in $He(E) = \{y_1, ..., y_k\}$ such that the structural equation for y_i^* is $y_i^* = y_i - \Lambda_{X_i y_i} T_i^t$, where $X_i = Ta(E) \cap Pa(y_i)$, for all $i \in \{1, ..., k\}$. The corresponding* augmented graph *is denoted G^{E+}.*

For example, consider Figure 1a. If the value of β is known, we can generate an auxiliary variable $x^* = x - \beta t$. The β-augmented graph $G^{\beta+}$ is depicted in Figure 1b. In some cases, x^* allows the identification of coefficients and testable implications using existing methods when x could not, due to the fact that the back-door paths from x to y that go through β cancel with the back-door paths from x^* to y that go through $-\beta$. This can be seen by expressing the covariance of x^* and y in terms of the model parameters using Wright's rules.

3. Auxiliary and Quasi-Instrumental Sets

Two, perhaps the most common, methods for estimating causal effects are OLS regression and two-stage least-squares (2SLS) regression. Both of these methods assume that the underlying causal relationships between variables are linear, in addition to other causal assumptions that guarantee identification. The *single-door criterion* (Pearl, 2009) graphically characterizes when the assumptions sufficient to estimate a causal effect using regression are satisfied in a linear SEM. Similarly, Brito and Pearl (2002a) gave a graphical characterization for when a variable z

[5] Wright's rules characterize the relationship between the covariance matrix and model parameters. Therefore, any question about identification using the covariance matrix can be decided by studying the solutions for this system of equations. However, since these equations are polynomials and not linear, it can be very difficult to analyze identification of models using Wright's rules.

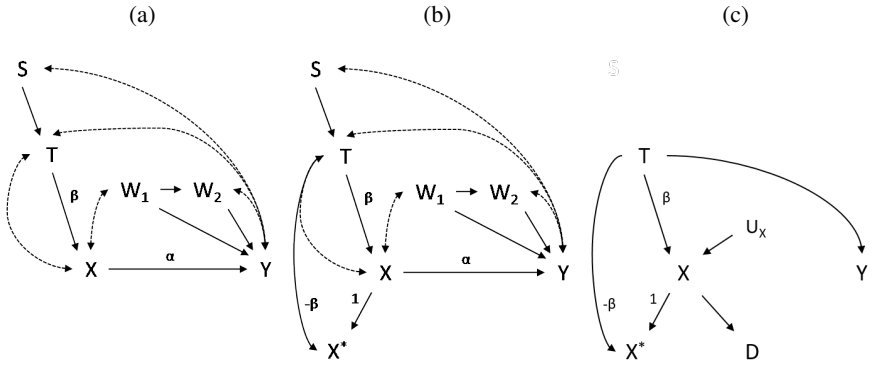

Figure 1. (a) α is not identified using IVs (b) α is identified using x^* as an auxiliary IV given w_1 (c) conditioning on descendants of x induces correlation between x^* and y

qualifies as an IV so that 2SLS regression provides a consistent estimate of the causal effect. In this section, we give a graphical criterion for when AVs can be utilized in generalized instrumental sets, which extends both the single-door criterion and IVs. Additionally, we introduce quasi-instrumental sets, which utilize AVs to better address the problem of z-identification.

First, we give a simple graphical criterion for when an AV would be conditionally independent of another variable, which will allow us to incorporate AVs into instrumental sets, as well as other identification and model testing methods that require the ability to detect conditional independence in the graph.

Theorem 1. *Given a linear SEM with graph G, where $E \subseteq Inc(z)$ is a set of edges whose coefficient values are known, if $W \cup \{y\}$ does not contain descendants of z and G_{E-} represents the graph G with the edges for E removed, then $(z^* \perp\!\!\!\perp y|W)_{G^{E+}}$ if and only if $(z \perp\!\!\!\perp y|W)_{G_{E-}}$.*[6]

Proof. Proofs for all theorems and lemmas can be found in the Appendix (Chen et al., 2017). □

Next, we demonstrate how AVs can be incorporated into generalized instrumental sets, defined below.

[6]The theorem disallows descendants of the generating variable in the conditioning set. At first glance, this may appear to limit the ability to block biasing paths among AVs. However, we conjecture that if z cannot be separated from y in G, then z^* will almost surely not be independent of y given W, if W contains descendants of z. To illustrate, consider the example shown in Figure 1c. $x^* = x - \beta t$ is independent of y, as can be verified using Wright's rules, but x^* is not independent of y given d! An intuitive explanation for this surprising result is that conditioning on d, a descendant of x, in Figure 1c induces correlation between the error term of x and t, since x acts as a "virtual collider". As a result, we have a "virtual path" from x^* to y, $x^* \leftarrow x \leftarrow u_x \leftrightarrow t \rightarrow y$. See Pearl (2009, p. 339) for a detailed discussion of virtual colliders.

Theorem 2. *(Brito and Pearl, 2002a) Given a linear model with graph G, the coefficients for a set of edges $E = \{(x_1, y), ..., (x_k, y)\}$ are identified if there exists triplets $(z_1, W_1, p_1), ..., (z_k, W_k, p_k)$ such that for $i = 1, ..., k$,*

(i) $(z_i \perp\!\!\!\perp y|W_i)_{G_{E-}}$, where W does not contain any descendants of y and G_{E-} is the graph obtained by deleting the edges, E from G,

(ii) p_i is a path between z_i and x_i that is not blocked by W_i, and

(iii) the set of paths, $\{p_1, ..., p_k\}$ has no sided intersection.[7]

If the above conditions are satisfied, we say that Z is a generalized instrumental set *for E or simply an* instrumental set *for E.*[8]

In some cases, a variable z may not satisfy condition (i) above but an auxiliary variable z^* does. For example, in Figure 1a, we cannot identify α using Theorem 2. Blocking the path $x \leftarrow t \leftrightarrow y$ by conditioning on t opens the path, $x \leftrightarrow t \leftrightarrow y$. Moreover, we cannot use t or s in an instrumental set due to the edges $t \leftrightarrow y$ and $s \leftrightarrow y$. However, s is an IV for β, allowing us to generate an AV, $x^* = x - \beta \cdot t_1$, as in Figure 1b. Now, α can be identified using x^* as an auxiliary instrument given w_1.

Theorem 1 tells us when (i) of Theorem 2 can be satisfied using an AV, z_i^*. We simply check whether z_i can be separated from y in $G_{E \cup E_z -}$, where $E_z \subseteq Inc(z_i)$ is the set of z_i's edges whose coefficient values are known. When an instrumental set includes AVs, we call the set an *auxiliary instrumental set* or *auxiliary IV set* for short.

[7]Brito and Pearl (2002a) provided an alternative statement of condition (iii). A proof that the two statement are, in fact, equivalent is given in the Appendix (Chen et al., 2017).

[8]Note that when $k = 1$, z_1 is an IV for (x_1, y). Further, if $z_1 = x_1$, then x_1 satisfies the single-door criterion for (x_1, y).

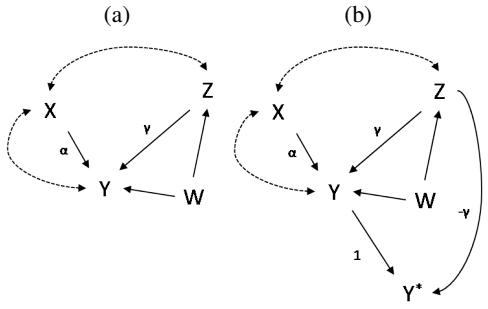

Figure 2. (a) α is not identified using IVs (b) α is identified using Z as a quasi-IV after adding auxiliary variable Y^*

Figure 1a also demonstrates the importance of extending the simple auxiliary instrumental sets introduced by Chen et al. (2016) to allow for conditioning. α can only be identified if we block the paths $x \leftrightarrow w_1 \rightarrow y$ and $x \leftrightarrow w_1 \rightarrow w_2 \rightarrow y$ by conditioning on w_1.

When knowledge of coefficient values are known a priori, it may be helpful to generate an AV from the outcome variable y. For example, in Figure 2a, α cannot be identified. However, suppose that it is possible to run a surrogate experiment and randomize z. This experiment would allow us to estimate γ and generate the AV, $Y^* = Y - \gamma Z$. Now, z is not technically an instrument for α, but it can be shown that $\alpha = \frac{r_{Y*Z.W}}{r_{XZ}}$. Chen et al. (2016) called such variables *quasi-instrumental variables* or *quasi-IVs* for short.

Interestingly, while quasi-IVs are valuable for the problem of z-identification, they do no better than instrumental sets when applied to the standard identification problem, where no external knowledge of coefficient values is available. For example, consider again Figure 2a. In order to use z as a quasi-IV for α, we would first have to identify γ using an IV. If such a variable existed, say z', then we could have simply identified $\{\alpha, \gamma\}$ using the IV set $\{z, z'\}$.

Next, we formally define *quasi-instrumental sets* or *quasi-IV sets* for short. Note that auxiliary IV sets are also quasi-IV sets.

Definition 5. *Given a linear SEM with graph G, a set of edges E_K whose coefficient values are known, and a set of structural coefficients $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_k\}$, the set $Z = \{z_1, ..., z_k\}$ is a quasi-instrumental set if there exist triples $(z_1, W_1, p_1), ..., (z_k, W_j, p_k)$ such that:*

(i) For $i = 1, ..., k$, either:

- *(a) the elements of W_i are non-descendants of y, and $(z_i \perp\!\!\!\perp y | W_i)_{G_{E \cup E_y}}$ where $E_y = E_K \cap Inc(y)$.*
- *(b) the elements of W_i are non-descendants of z_i and y, and $(z_i \perp\!\!\!\perp y | W_i)_{G_{E \cup E_{zy}}}$ where $E_{zy} = E_K \cap (Inc(z) \cup Inc(y))$.*

(ii) for $i = 1, ..., k$, p_i is a path between z_i and x_i that is not blocked by W_i, where $x_i = He(\alpha_i)$, and

(iii) the set of paths $\{p_1, ..., p_k\}$ has no sided intersection

Theorem 3. *If Z^* is a quasi-instrumental set for E, then E is identifiable.*

Lastly, the following corollary provides a simple graphical condition for when a single variable or AV qualifies as a quasi-IV.

Corollary 1. *Given a linear SEM with graph G, z^* is a quasi-IV for α given W if W does not contain any descendants of z, and z is an IV for α given W in $G_{E_z \cup E_y -}$, where $E_z \subseteq Inc(z)$ and $E_y \subseteq Inc(y)$ are sets of edges whose coefficient values are known.*

Auxiliary and quasi-IV sets enable a bootstrapping procedure whereby complex models can be identified by iteratively identifying coefficients and using them to generate new auxiliary variables. For example, consider Figure 3a. First, we are able to identify b and c using IVs, but no other coefficients. Once b is identified, Corollary 1 tells us that e is identified using v_3^* since v_3 is an IV for e when the edge for b is removed (see Figure 3b). Now, the identification of e allows us to identify a and d using v_5^*, since v_5 is an IV for a and d when the edge for e is removed (see Figure 3c). This general strategy is the basis for our identification, z-identification, and model testing algorithm, described next.

4. Identification and z-Identification Algorithm

In this section, we construct an identification algorithm that operationalizes the bootstrapping approach described in Section 3. First, we describe how to algorithmically find a quasi-instrumental set for a set of coefficients E, given a set of known coefficients, IDEdges.

The problem of finding generalized instrumental sets was addressed by van der Zander and Liskiewicz (2016). They provided an algorithm, TestGeneralIVs, that determines whether a given set Z is a generalized instrumental set for a set of edges, E, that runs in polynomial time if we bound the size of the coefficient set to be identified. More specifically, their algorithm has a running time of $O((k!)^2 n^k)$, where n is the number of variables in the graph and $k = |E|$.[9]

Our method, TestQIS, given in the Appendix (Chen et al., 2017), generalizes TestGeneralIVs, for quasi-IV sets.

[9] van der Zander and Liskiewicz (2016) also give an algorithm that tests whether Z is a *simple conditional instrumental sets* in $O(nm)$ time. A simple conditional instrumental set is a generalized instrumental set where $W_1 = W_2 = ... = W_k$

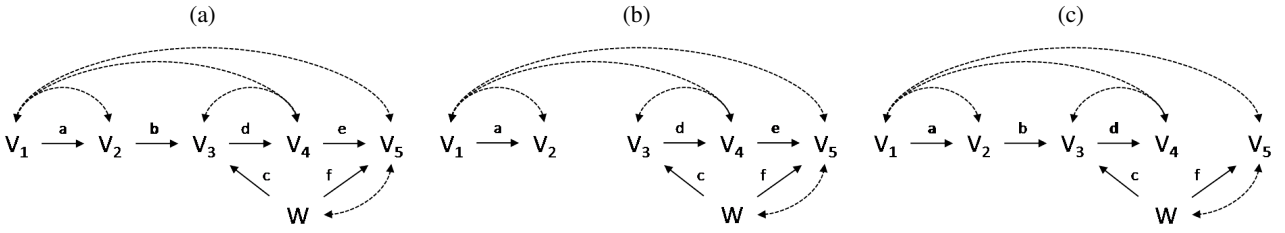

Figure 3. (a) b is identified using either v_2 or v_1 as an instrument and c is identified using w as an instrument (b) e is identified using v_3^* as an auxiliary instrument given (c) a and d are identified using v_5^* as an auxiliary instrument

FindQIS, also given in the Appendix (Chen et al., 2017), searches for a quasi-IV set by checking all subsets of $Z \subseteq (Anc(z_i) \cup Anc(y))$ using TestQIS. It returns a quasi-IV set, as well as its conditioning sets, if one exists.

In some cases an instrumental set may not exist for C, but one exists for $C^{'}$, where $C \subset C^{'}$. Conversely, there may not be an instrumental set for $C^{'}$, but there is one for $C \subset C^{'}$. As a result, we may have to check all possible subsets of a variable's coefficients in order to determine whether a given subset is identifiable using auxiliary instrumental sets. This search can be simplified somewhat by noting that if E is a *connected edge set* (defined below) with no instrumental set, then there is no superset $E^{'}$ with an instrumental set.

Definition 6. *(Chen et al., 2014) For an arbitrary variable, V, let $Pa_1, Pa_2, ..., Pa_k$ be the unique partition of Pa(V) such that any two parents are placed in the same subset, Pa_i, whenever they are connected by an unblocked path. A* connected edge set *with head V is a set of directed edges from Pa_i to V for some $i \in \{1, 2, ..., k\}$.*

The ID algorithm, called qID utilizes FindQIS to identify as many coefficients as possible in a given model with graph G. It iterates through each connected edge set and attempts to identify it using FindQIS. If it is unable to identify the connected edge set, it then attempts to identify subsets of the connected edge set. After the algorithm has attempted to identify each connected edge set, it again attempts to identify each unidentified connected edge set, since each newly identified coefficient may enable the identification of previously unidentifiable coefficients. This process is repeated until all coefficients have been identified or no new coefficients have been identified in the last iteration. The algorithm is polynomial if the degree of each node in the graph is bounded.

Our algorithm identifies the model depicted in Figure 4b in the following way. First, let us assume that the connected edge sets are arbitrarily ordered, $(\{a\}, \{b, c, f\}, \{d\}, \{e\})$. Now, the first edge to be identified would be a using w_1 as an IV. There is no auxiliary IV set for $\{b, c, f\}$, and we would attempt to find one for its subsets. We find that $\{b\}$ is identified using $\{x\}$ as an IV set with conditioning set $\{w_1\}$. Now, $\{d\}$ is identified using $y^* = y - bx$, and e is identified using t_2^*. In the second iteration, we return to $\{b, c, f\}$ and find that it is now identified using the auxiliary IV set, $\{x, w_1, t_3^*\}$.

Algorithm 1 qID$(G, \Sigma, \text{IDEdges})$

Initialize: EdgeSets $\leftarrow$ all connected edge sets in G
repeat
  for all ES in EdgeSets such that $ES \not\subseteq$ IDEdges **do**
    $y \leftarrow He(ES)$
    for all $E \subseteq ES$ such that $E \not\subseteq$ IDEdges **do**
      $(Z, W) \leftarrow \text{FindQIS}(G, ES, \text{IDEdges})$
      if $(Z, W) \neq \perp$ **then**
        Identify ES using Z^* as an auxiliary instrumental set in $G^{(\text{IDEdges} \cap Inc(Z))+}$
        IDEdges $\leftarrow$ IDEdges $\cup$ ES
      end if
    end for
  end for
until All coefficients have been identified or no coefficients have been identified in the last iteration

In contrast, Figure 4b is not identified using simple instrumental sets and auxiliary variables. We cannot identify b without conditioning on w_1, which means that the only coefficients identified using auxiliary simple instrumental sets is a. Since Chen et al. (2016) showed that any coefficient identified using the generalized half-trek criterion (g-HTC) can be identified using auxiliary variables and simple instrumental sets, we know that qID is able to identify coefficients and models that the g-HT algorithm is not. Moreover, qID will identify any coefficients that are identifiable using auxiliary variables and simple instrumental sets, giving us the following theorem.

Theorem 4. *Given an arbitrary linear causal model, if a set of coefficients is identifiable using the g-HT algorithm, then it is identifiable using qID. Additionally, there are models that are not identified using the g-HT algorithm, but identified using qID.*

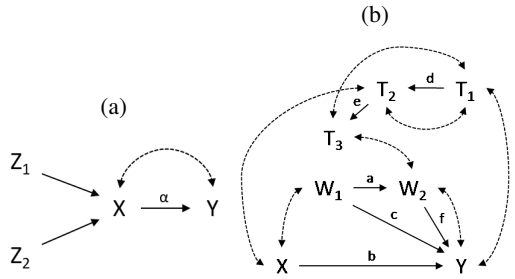

Figure 4. (a) $\sigma(z_2, y^*) = 0$, where $y^* = y - \frac{\sigma(y,z_1)}{\sigma(x,z_1)}$, and, equivalently, α is overidentified using z_1 and z_2 as IVs (b) the model is identified using auxiliary instrumental sets, but not the g-HT algorithm

5. Deriving Testable Implications using AVs

Theorem 1 also enables us to derive new vanishing partial correlation constraints that can be used to test the model. For example, in Figure 4a, α can be identified using z_1 as an instrument. Once α is identified, we can generate the AV $y^* = y - \alpha x = y - \frac{\sigma(y,z_1)}{\sigma(x,z_1)}x$, and Theorem 1 tells us that the correlation of z_2 and y^* should vanish. As a result, we can test the model specification by verifying that this constraint holds in the data.

Theorem 1 also tells us that the correlation between z_1 and y^* should also vanish. However, upon closer inspection, we find that this implication does not actually constrain the covariance matrix:

$$\begin{aligned}\sigma(z_1, y^*) =& \sigma(z_1, y - \alpha x) \\ =& \sigma(z_1, y) - \frac{\sigma(y, z_1)}{\sigma(x, z_1)}\sigma(z_1, x) = 0.\end{aligned}$$

In other words, our "testable implication" that $\sigma(z_1, y^*) = 0$ is equivalent to stating $\sigma(z_1, y) - \sigma(z_1, y) = 0$–a tautology! In contrast,

$$\sigma(z_2, y^*) = \sigma(z_2, y) - \frac{\sigma(z_1, y)}{\sigma(x, z_1)}\sigma(z_2, x) = 0$$

does provide a true testable implication.

Shpitser et al. (2009) noticed a similar phenomenon when deriving dormant independences in non-parametric models, and their explanation applies to conditional independence constraints among AVs as well. The idea is the following: When the model implies that two variables are conditionally independent, it relies on the modeled assumption that there is no edge between those variables. As a result, verifying that the constraint holds in data represents a test that this assumption is valid. However, unlike conditional independence constraints between model variables, conditional independence constraints among AVs rely upon the absence of certain edges in order to identify the coefficients necessary to generate the AV. The key point is that this identification cannot rely on the same lack of edge whose existence we are trying to test!

In the above example, we identified α using z_1 as an IV. $\sigma(z_2, y^*) = 0$ follows from the lack of edge between z_2 and y. However, even if this edge did exist, z^* still equals $z - \frac{\sigma(y,z_1)}{\sigma(x,z_1)}x$. In contrast, $\sigma(z_1, y^*) = 0$ follows from the lack of edge between z_1 and y. The existence of this edge would disallow z_1 as an instrument and $z^* = z - \alpha x \neq z - \frac{\sigma(y,z_1)}{\sigma(x,z_1)}x$.

Another way to derive the constraint $\sigma(z_2, y^*) = 0$ is via overidentification. α can be identified using either z_1 or z_2 and equating the corresponding expressions yields the constraint $\frac{\sigma(y,z_1)}{\sigma(x,z_1)} = \frac{\sigma(y,z_2)}{\sigma(x,z_2)}$, which is clearly equivalent to the previous constraint $\sigma(z_2, y^*) = 0$. In fact, we show (Theorem 6) that whenever a variable z cannot be separated from another variable y, but z^* can be, the resulting AV conditional independence, if it is non-vacuous, is equivalent to an overidentifying constraint that can be derived using quasi-IVs. As a result, all non-vacuous AV conditional independences are captured by overidentifying constraints derived using quasi-IVs!

First, we give a sufficient condition for when a set of edges α is overidentified.

Theorem 5. *Let Z be a quasi-IV set for structural coefficients $\alpha = \{\alpha_1, ..., \alpha_k\}$ and E be a set of known edges. If there exists a node s satisfying the conditions listed below, then α is overidentified and we obtain the constraint .*

(i) $s \notin Z$

(ii) There exists an unblocked path between s and y including an edge in α

(iii) There exists a conditioning set W that does not block the path p, such that either:

- *(a) the elements of W are non-descendants of y, and $(s \perp\!\!\!\perp y|W)_{G_{\alpha \cup E_y -}}$, where $E_y = E \cap Inc(y)$)*
- *(b) the elements of W are non-descendants of s and y, and $(s \perp\!\!\!\perp y|W)_{G_{\alpha \cup E_s \cup E_y -}}$ where $E_s = E \cap Inc(s)$.*

The above theorem can be used to derive an overidentifying constraint for every variable that satisfies (i)-(iii) above. It can also be applied when α is known a priori, yielding a *z-overidentifying constraint*. In this case, $Z = \emptyset$ would be a quasi-IV set that trivially identifies α.

The following theorem states that non-vacuous AV conditional independence constraints are subsumed by quasi-IV overidentifying and z-overidentifying constraints.

Theorem 6. *Let $z^* = z - e_1t_1 - ... - e_kt_k$ and suppose there does not exist W such that $(z \perp\!\!\!\perp y|W)_G$. There exists W such that $W \cap De(z) = \emptyset$ and $(z^* \perp\!\!\!\perp y|W)$ is non-vacuous if and only if y satisfies the conditions of Theorem 5 for $E = \{e_1, ..., e_k\}$.*

The above theorem also applies when y is an AV, called y^*. In this case, we simply replace $(z \perp\!\!\!\perp y|W)_G$ with $(z \perp\!\!\!\perp y^*|W)_{G^{E_y+}}$, where $E_y \subseteq Inc(y)$ is a set of edges whose coefficient values are known.

Algorithm 2 uses quasi-IV sets to output overidentifiying constraints in a graph given an optional set of identified edges. It uses isEIV, which is a slightly modified version of FindQIS that tests whether w fits the conditions of Theorem 6. Details of isEIV can be found in the Appendix (Chen et al., 2017).

Algorithm 2 Finds overidentifying constraints for G

```
function CONSTRAINTFINDER(G,Σ,IDEdges)
    for all ES ∈ Edge Sets of G do
        (Z, W) ← FINDQIS(ES,G,IDEdges)
        if (Z, W) ≠ ⊥ then
            for all w ∈ V \ Z ∪ {He(ES)} do
                if ISEIV(w,ES,G,IDEdges) then
                    Add constraint a_w A^{-1} b = b_w
                end if
            end for
        end if
    end for
end function
```

6. Discussion and Related Work

In this section, we discuss how (single-variable) auxiliary IVs encompass a number of previous identification methods developed in economics (Hausman and Taylor, 1983), computer science (Chan and Kuroki, 2010), and epidemiology (Shardell, 2012).

Hausman and Taylor (1983) showed that if the equation for a given variable, $z = \beta_1 p_1 + ... + \beta_k p_k + u_z$, is identified, then the error term u_z can be estimated and used as an instrument for other coefficients. In this case, the auxiliary variable $z^* = z - \beta_1 p_1 - ... - \beta_k p_k$ is equal to the error term u_z. As a result, whenever the error term is estimable and can be used as an IV, we can also generate an auxiliary instrument. However, there are times when only some of the coefficients in an equation are identifiable, and as a result, the error term cannot be used as an instrument, but we can nevertheless generate an auxiliary instrument. As a result, auxiliary IVs strictly subsume error term IVs.

Chan and Kuroki (2010) gave sufficient conditions for when a descendant of x and a descendant of y could be used in analogous manner to IVs to identify the effect of x on y. In the context of AVs, this method is equivalent to generating an auxiliary instrument from the descendant by subtracting the total effect of x on the descendant or the total effect of y on the descendant (depending on whether the variable is a descendant of x or y). In this paper, we generated AVs by subtracting out direct effects, but clearly the work can be extended to subtracting out total effects. The benefit of AVs over these descendant IVs is that they can be generated from a variety of variables, not just descendants of x and y. Additionally, descendants of x or y can generate AVs from other total or direct effects, not just the effect of x or y on the descendant.

The notion of "subtracting out a direct effect" in order to turn a variable into an instrument was also noted by Shardell (2012) when attemping to identify the total effect of x on y. It was noticed that in certain cases, the violation of the independence restriction of a potential instrument z (i.e. z is not independent of the error term of y) could be remedied by identifying, using ordinary least squares regression, and then subtracting out the necessary direct effects on y. AVs generalize and operationalize this notion so that it can be used on arbitrary sets of known coefficient values and be utilized in conjunction with existing graphical methods for identification and enumeration of testable implications.

Additionally, as we have alluded to earlier, the highly algebraic, state-of-the-art g-HTC can also be understood in terms of auxiliary instruments. Identification using the g-HTC is equivalent to identification using auxiliary simple instrumental sets.

In summary, auxiliary instruments are not only the basis for the most general identification algorithm yet devised, but they also unify disparate identification methods under a single framework. Moreover, AVs are directly applicable to the tasks of z-identification and model testing. Finally, they can, in principle, enhance any method for identification, model testing, or other tasks that relies on graphical separation.

7. Conclusion

In this paper, we graphically characterized conditional independence among AVs, allowing us to demonstrate how they can help generalized instrumental sets in the problem of identification. We provided an algorithm that identifies more models than the g-HT algorithm, subsuming the state-of-the-art for identification in linear models. Additionally, we introduced quasi-IV sets, and constructed an algorithm that utilizes them to attack the problem of z-identification. Finally, we proved that AV conditional independences are subsumed by overidentifying constraints and gave an algorithm for deriving overidentifying constraints.

Acknowledgements

We would like to thank Judea Pearl, Mathias Drton, Thomas Richardson, and Luca Weihs for helpful discussions. This research was supported in parts by grants from NSF #IIS-1302448 and #IIS-1527490 and ONR #N00014-13-1-0153 and #N00014-13-1-0153.

References

BAREINBOIM, E. and PEARL, J. (2012). Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (N. de Freitas and K. Murphy, eds.). AUAI Press, Corvallis, OR.

BAREINBOIM, E. and PEARL, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* **113** 7345–7352.

BEKKER, P., MERCKENS, A. and WANSBEEK, T. (1994). *Identification, Equivalent Models, and Computer Algebra*. Statistical Modeling and Decision Science, Academic Press.

BOWDEN, R. and TURKINGTON, D. (1984). *Instrumental Variables*. Cambridge University Press, Cambridge, England.

BRITO, C. and PEARL, J. (2002a). Generalized instrumental variables. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference* (A. Darwiche and N. Friedman, eds.). Morgan Kaufmann, San Francisco, 85–93.

BRITO, C. and PEARL, J. (2002b). A graphical criterion for the identification of causal effects in linear models. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, 533–538.

BRITO, C. and PEARL, J. (2002c). A new identification condition for recursive models with correlated errors. *Journal Structural Equation Modeling* **9** 459–474.

BRITO, C. and PEARL, J. (2006). Graphical condition for identification in recursive SEM. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Corvallis, OR, 47–54.

CHAN, H. and KUROKI, M. (2010). Using descendants as instrumental variables for the identification of direct causal effects in linear sems. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*.

CHEN, B. (2016). Identification and overidentification of linear structural equation models. In *Advances In Neural Information Processing Systems*. 1579–1587.

CHEN, B., KUMOR, D. and BAREINBOIM, E. (2017). Identification and model testing in linear structural equation models using auxiliary variables. *arXiv preprint arXiv:1612.03451; Technical report R-27, Purdue AI Lab, Dept. of Computer Science, Purdue University.* .

CHEN, B. and PEARL, J. (2014). Graphical tools for linear structural equation modeling. Tech. Rep. R-432, <http://ftp.cs.ucla.edu/pub/stat_ser/r432.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Forthcoming, Psychometrika.

CHEN, B., PEARL, J. and BAREINBOIM, E. (2016). Incorporating knowledge into linear structural equation models using auxiliary variables. In *Proceedings of the Twenty-fifth International Joint Conference on Artificial Intelligence* (S. Kambhampati, ed.).

CHEN, B., TIAN, J. and PEARL, J. (2014). Testable implications of linear structual equation models. In *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence* (C. E. Brodley and P. Stone, eds.). AAAI Press, Palo, CA. <http://ftp.cs.ucla.edu/pub/stat_ser/r428-reprint.pdf>.

DRTON, M. and WEIHS, L. (2016). Generic identifiability of linear structural equation models by ancestor decomposition. *Scandinavian Journal of Statistics* n/a–n/a10.1111/sjos.12227.
URL `http://dx.doi.org/10.1111/sjos.12227`

FISHER, F. (1966). *The Identification Problem in Econometrics*. McGraw-Hill, New York.

FOYGEL, R., DRAISMA, J. and DRTON, M. (2012). Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics* **40** 1682–1713.

HAUSMAN, J. A. and TAYLOR, W. E. (1983). Identification in linear simultaneous equations models with covariance restrictions: an instrumental variables interpretation. *Econometrica: Journal of the Econometric Society* 1527–1549.

MOOIJ, J. M., PETERS, J., JANZING, D., ZSCHEISCHLER, J. and SCHÖLKOPF, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research* **17** 1–102.

PEARL, J. (2004). Robustness of causal claims. In *Proceedings of the Twentieth Conference Uncertainty in Artificial Intelligence* (M. Chickering and J. Halpern, eds.). AUAI Press, Arlington, VA, 446–453.

PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

RIGDON, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research* **30** 359–383.

SHARDELL, M. (2012). Methods to overcome violations of an instrumental variable assumption: Converting a confounder into an instrument. *Computational statistics & data analysis* **56** 2317–2333.

SHIMIZU, S., HOYER, P. O., HYVÄRINEN, A. and KERMINEN, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **7** 2003–2030.

SHPITSER, I., RICHARDSON, T. S. and ROBINS, J. M. (2009). Testing edges by truncations. In *IJCAI*.

SPIRTES, P., GLYMOUR, C. N. and SCHEINES, R. (2000). *Causation, prediction, and search*, vol. 81. MIT press.

SPIRTES, P., RICHARDSON, T., MEEK, C., SCHEINES, R. and GLYMOUR, C. (1998). Using path diagrams as a structural equation modelling tool. *Sociological Methods and Research* **27** 182–225.

TIAN, J. (2005). Identifying direct causal effects in linear models. In *Proceedings of the National Conference on Artificial Intelligence*, vol. 20. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

TIAN, J. (2007). A criterion for parameter identification in structural equation models. In *Proceedings of the Twenty-Third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*. AUAI Press, Corvallis, Oregon.

TIAN, J. (2009). Parameter identification in a class of linear structural equation models. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*.

VAN DER ZANDER, B. and LISKIEWICZ, M. (2016). Searching for generalized instrumental variables. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS-16))*.

VAN DER ZANDER, B., TEXTOR, J. and LISKIEWICZ, M. (2015). Efficiently finding conditional instruments for causal inference. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*.

WRIGHT, S. (1921). Correlation and causation. *Journal of Agricultural Research* **20** 557–585.

ZHANG, K. and HYVÄRINEN, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press.