UvA-DARE (Digital Academic Repository)

Algebraic Equivalence of Linear Structural Equation Models

van Ommen, T.; Mooij, J.M.

[Link to publication](#)

Citation for published version (APA):

Algebraic Equivalence of Linear Structural Equation Models

Thijs van Ommen
Informatics Institute
University of Amsterdam
The Netherlands

Joris M. Mooij
Informatics Institute
University of Amsterdam
The Netherlands

Abstract

Despite their popularity, many questions about the algebraic constraints imposed by linear structural equation models remain open problems. For causal discovery, two of these problems are especially important: the enumeration of the constraints imposed by a model, and deciding whether two graphs define the same statistical model. We show how the half-trek criterion can be used to make progress in both of these problems. We apply our theoretical results to a small-scale model selection problem, and find that taking the additional algebraic constraints into account may lead to significant improvements in model selection accuracy.

1 INTRODUCTION

In a linear structural equation model (SEM), each variable of interest is a linear function of the other variables and a noise term, with possibly correlated noise terms. Linear SEMs are popular in many fields of science, in no small part due to their causal interpretability (Spirtes et al., 2000; Pearl, 2000). However, many questions about these models remain unanswered. For example, it is known that many of these models impose equality constraints on the observational distribution which do not correspond to (conditional) independences (Richardson and Spirtes, 2002). One example of these is the Verma constraint (Robins, 1986; Verma and Pearl, 1991). But no general method exists which enumerates all constraints that hold in a model given its graphical representation. Figure 1 shows another example of a graph that imposes an equality constraint in the linear case; for this type of constraint, no systematic approach exists yet.

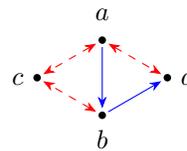Relatedly, given graphical representations of two mod-

Figure 1: A mixed graph imposing the non-independence constraint (3) on the observational distribution.

els, it is often unclear whether the models can be distinguished based on observational data alone. Both of these problems are great impediments to the development of methods that learn the structure of a linear SEM from observational data: constraint-based methods (e.g. PC and FCI (Spirtes et al., 2000)) cannot test for constraints that are not yet well understood and may thus miss signals in the data, while score-based methods (e.g. (Chickering, 2002; Silva and Ghahramani, 2006)) would currently require the scoring of many models that then turn out to be indistinguishable.

The theoretical results of this paper achieve progress in addressing both problems mentioned above. For example, we provide an efficient method to find the constraint imposed by the graph in Figure 1, as well as similar constraints for many other graphs. Our results apply to models with latent confounders (represented by bidirected edges), including confounders between nodes that are also related by a direct causal effect (a structure known as a *bow*), and even to models with directed cycles. We show how these results enable practical improvements on model selection problems.

Suppose we see that our observational data obeys a non-independence equality constraint, such as the one imposed by the graph in Figure 1, and no other constraints. Then we can often draw very specific conclusions about the graph structure. Without knowledge of these constraints, automated methods for causal discovery would
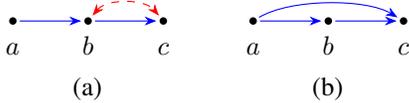
Figure 2: Two graphs whose statistical models are almost, but not entirely identical: (a) the instrumental variable model; (b) a saturated model.

likely select a *saturated* model (one imposing no constraints). This tells us very little about the graph structure, so e.g. it does not allow us to predict the results of interventions. Constraint-based methods generally use only (conditional) independence constraints, so they would not be able to draw any conclusions in the situation described here. For purposes of model selection, we are thus interested in a notion of model equivalence that is more fine-grained than Markov equivalence (which only takes conditional independence constraints into account), yet not so fine-grained as to be impractical

The equivalence concept we propose in this paper is *algebraic equivalence*: Two linear structural equation models are algebraically equivalent if they impose the same algebraic (i.e. equality) constraints on the observational distribution.[1] These constraints take the form of polynomial equations over covariances σ_{vw} of the observed variables. Because the graph in Figure 1 imposes such a constraint, it is not algebraically equivalent to the saturated model on four nodes, so a model selection method based on algebraic equivalence is able to distinguish the two, while a method based on Markov equivalence is not.

For an example of models not distinguished by algebraic equivalence, consider the mixed graph in Figure 2(a), often called the instrumental variable model. This model contains all multivariate Gaussian distributions on the three variables with $\sigma_{ab} \neq 0$, but excludes some with $\sigma_{ab} = 0$. Because it imposes no equality constraints, it is algebraically equivalent to the saturated model on three nodes, represented for example by the graph in Figure 2(b). The difference between the two models is a measure zero subset of their union, so that in a model selection problem, it would rarely be possible to distinguish between these models based on observational data alone. Thus it is appropriate that our proposed equivalence concept treats these models as equivalent.

By considering only equality constraints, we are also treating models as equivalent if they differ only by inequality constraints (of the form $f(\Sigma) \geq 0$). It is known that linear Gaussian models with latent variables may impose inequality constraints. For example, consider the

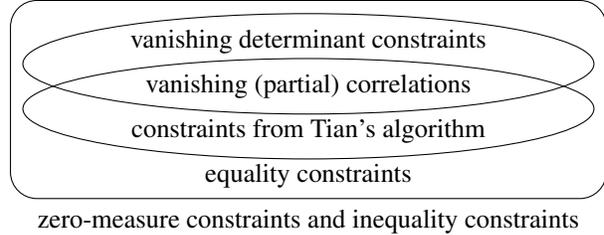zero-measure constraints and inequality constraints

Figure 3: The types of constraints imposed by linear structural equation models.

graph with three observed variables and a latent confounder between every pair of observed variables. The corresponding mixed graph has three nodes and a bidirected edge between each pair of nodes. This corresponds to the saturated model. However, if we consider the entire graph on 6 nodes and then marginalize out the latent variables, the resulting set of distributions obeys a nontrivial inequality constraint (Richardson and Spirtes, 2002). By using mixed graphs instead of including latent variables in our models explicitly, we are already simplifying away such inequality constraints. However, as we will show in Section 3.3, using mixed graphs does not get rid of all inequality constraints. We found that for models imposing an inequality constraint, maximum likelihood estimation can be challenging, so score-based methods for causal discovery would benefit significantly from being able to ignore these models. Algebraic equivalence provides a way to do this, as many models that impose an inequality constraint are algebraically equivalent to a model that imposes no such constraints.[2]

1.1 RELATIONS TO OTHER TYPES OF CONSTRAINTS

When defining linear structural equation models statistically, often the noise terms are chosen to be Gaussian. For Gaussian variables, independence is equivalent to having zero covariance (and in turn to having zero correlation), and conditional independence is equivalent to zero *partial correlation*. Because we do not want to assume that the data are generated by a Gaussian distribution, we need to distinguish between conditional independence and vanishing partial correlation. The equality constraints we consider express vanishing partial correlations, not conditional independences, so we will use that terminology from now on.

A partial correlation $\rho_{vw.S}$ is zero precisely when a certain submatrix of the observed covariance matrix has

---

[1]Using terminology from algebraic geometry, the statistical models have the same Zariski closure (Cox et al., 2015).

[2]In a model selection task, we can of course still check for inequality constraints, but we propose to do this *after* selecting the algebraic equivalence class.

zero determinant. Sullivant et al. (2010) give a graphical characterization for the more general type of constraint where an arbitrary square submatrix has zero determinant; this also includes the vanishing tetrad constraints (Spirtes et al., 2000). Together these are called *vanishing determinant constraints* (see Figure 3).

The questions of constraint enumeration and model equivalence that we study here for the linear case, are studied for the general nonparametric case by Tian and Pearl (2002) and Shpitser et al. (2014). Tian's algorithm gives a sound enumeration of constraints in the general case, and Evans (2015) shows it to be complete in the discrete case. If other parametric assumptions are made, there may be additional constraints. Indeed this happens in the linear Gaussian case: Tian's algorithm returns no constraints for the graph in Figure 1, even though one exists.

Lists of algebraic constraints can also be obtained using algorithms from computer algebra, but these are in general impractically slow, sometimes taking many days even for very small graphs (García-Puente et al., 2010). The methods we propose are based on the graphical criteria proposed by Foygel et al. (2012), which can be checked in polynomial time.

The rest of this paper is structured as follows. Section 2 discusses preliminaries about linear SEMs, the half-trek criterion, algebraic geometry, and our notation for sets of graphs. Our main theoretical contributions are in Section 3, addressing the enumeration of algebraic constraints in Section 3.1, and a sufficient graphical criterion for algebraic equivalence in Section 3.2; further, Section 3.3 gives an example of how inequality constraints may arise in linear SEMs. Experimental results demonstrating the practical usefulness of our results are presented in Section 4. Section 5 concludes the paper. All proofs are in the supplementary material in Appendix A, and a complete description of all algebraic equivalence classes on four nodes (acyclic) can be found in Appendix B.

2 PRELIMINARIES

We follow Foygel et al. (2012) for most of the notation defined in this section.

A *mixed graph* $G = (V, D, B)$ consists of a set of nodes V, a set of directed edges D which are ordered pairs of distinct nodes, and a set of bidirected edges B which are unordered pairs of nodes. In this article, the word 'graph' without qualification refers to mixed graphs. A node $x \in V$ with $(x, v) \in D$ is called a *parent* of v, and the set of parents is denoted $\mathrm{pa}(v)$; similarly, a node x with $\{x, v\} \in B$ is called a *sibling* of v and

the set is denoted $\mathrm{sib}(v)$. G is called *acyclic* if it contains no directed cycle (such a G is also called *acyclic directed mixed graph (ADMG)*). If G is acyclic and contains no bidirected edges, it is called a *directed acyclic graph (DAG)*.

Together with parameter vector λ_0 and parameter matrices Λ and Ω, the graph G describes a distribution on observed variables X via

$$X_v = \lambda_{0v} + \sum_{w \in \mathrm{pa}(v)} \lambda_{wv} X_w + \epsilon_v \qquad \text{for } v \in V,$$

where the noise terms have covariances $\mathrm{Cov}(\epsilon_v, \epsilon_w) = \omega_{vw}$. The parameter space is defined as follows. Let $n = |V|$. $\mathbb{R}^D$ is the set of all $n \times n$ matrices Λ with $\Lambda_{vw} \neq 0$ only if $(v, w) \in D$, and $\mathbb{R}^D_{\mathrm{reg}}$ is the subset of $\mathbb{R}^D$ for which $I - \Lambda$ is invertible (for acyclic G, $\mathbb{R}^D = \mathbb{R}^D_{\mathrm{reg}}$). PD_n is the set of all positive definite $n \times n$ matrices, and $\mathrm{PD}(B)$ is the subset consisting of all Ω with $\Omega_{vw} \neq 0$ only if $v = w$ or $\{v, w\} \in B$. The parameterization map ϕ_G maps parameters $(\Lambda, \Omega) \in \mathbb{R}^D_{\mathrm{reg}} \times \mathrm{PD}(B)$ to covariance matrices $\Sigma \in \mathrm{PD}_n$ on the observed variables X as follows:

$$\phi_G(\Lambda, \Omega) = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}.$$

The *model* defined by a graph G then consists of all covariance matrices Σ that can be obtained for some setting of the parameters:

$$\mathcal{M}(G) := \{\phi_G(\Lambda, \Omega) \mid (\Lambda, \Omega) \in \mathbb{R}^D_{\mathrm{reg}} \times \mathrm{PD}(B)\}. \quad (1)$$

Note that the mean of X can be set arbitrarily by choosing appropriate values for the parameter vector $\lambda_0 \in \mathbb{R}^n$, regardless of the structure of G. Thus these aspects of the model carry no information for model selection, and we will ignore them here.

2.1 THE HALF-TREK CRITERION

A central question about a mixed graph G is that of *(parameter) identifiability*: can the parameters (Λ, Ω) be uniquely recovered from Σ? A graph is called *generically identifiable* (or almost-everywhere identifiable) if this is true of $\phi_G(\Lambda, \Omega)$ for all but a measure zero subset of $\mathbb{R}^D_{\mathrm{reg}} \times \mathrm{PD}(B)$. Similarly, G is called *generically finite-to-one* if for almost all (Λ, Ω), the number of parameter values mapped by ϕ_G to the same Σ is finite, and *generically infinite-to-one* if this number is infinite for almost all parameter values. We will sometimes omit the qualifier 'generically' when talking about (in)finite-to-one models.

Foygel et al. (2012) present two graphical criteria to decide in which of the above categories a graph G belongs. A graph is called *HTC-identifiable* if it meets

the condition for being generically identifiable; *HTC-nonidentifiable* if it meets the condition for being generically infinite-to-one; and *HTC-inconclusive* otherwise. Because neither criterion is necessary, the class of HTC-inconclusive graphs contains generically identifiable, finite-to-one, and infinite-to-one graphs. While the criteria are not complete, they are quite powerful. For example, all bow-free acyclic graphs are HTC-identifiable (thus implying the earlier identification result of Brito and Pearl (2002)), and so are many graphs containing bows or directed cycles.

The proof of HTC-identifiability in (Foygel et al., 2012) is constructive: it gives an algorithm that, given $\Sigma \in \mathcal{M}(G)$, computes parameters such that $\Sigma = \phi_G(\Lambda, \Omega)$ (except for a measure zero subset). To apply our Theorem 1 below, some details of HTC-identifiability and this algorithm are needed; for the rest, we refer to (Foygel et al., 2012, proof of Theorem 1). A *half-trek* from v to w is either a directed path, or a path consisting of one bidirected edge followed by directed edges towards w. We write $\mathrm{htr}(v)$ (*half-trek reachable*) for the set of nodes that are reachable from v by half-treks.[3] HTC-identifiability of a graph G requires that for each node $v \in V$, a set $Y_v \subseteq V \setminus (\{v\} \cup \mathrm{sib}(v))$ exists, consisting of nodes y with $v \in \mathrm{htr}(y)$, and the set as a whole satisfying $|Y_v| = |\mathrm{pa}(v)|$. There are some additional restrictions which we omit here, except to point out that for all $v, w \in V$, at most one of $v \in Y_w$ and $w \in Y_v$ can hold. As an example, the instrumental variable model (Figure 2(a)) is HTC-identifiable with $Y_a = \varnothing$ and $Y_b = Y_c = \{a\}$.

Using these sets Y_v, the algorithm for finding Λ solves a sequence of linear systems, one for each $v \in V$: Let $Y_v = \{y_1, \ldots, y_n\}$ and $\mathrm{pa}(v) = \{p_1, \ldots, p_n\}$, and define $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{b} \in \mathbb{R}^n$ as

$$\mathbf{A}_{ij} = \begin{cases} [(I-\Lambda)^T\Sigma]_{y_i p_j} & \text{if } y_i \in \mathrm{htr}(v), \\ \Sigma_{y_i p_j} & \text{if } y_i \notin \mathrm{htr}(v); \end{cases}$$

$$\mathbf{b}_i = \begin{cases} [(I-\Lambda)^T\Sigma]_{y_i v} & \text{if } y_i \in \mathrm{htr}(v), \\ \Sigma_{y_i v} & \text{if } y_i \notin \mathrm{htr}(v). \end{cases}$$

Then the vector $\Lambda_{\mathrm{pa}(v),v}$ is found by solving $\mathbf{A} \cdot \Lambda_{\mathrm{pa}(v),v} = \mathbf{b}$. After the entire matrix Λ has been found this way, Ω is given by $(I-\Lambda)^T\Sigma(I-\Lambda)$.

2.2 SOME TERMINOLOGY FROM ALGEBRAIC GEOMETRY

This section very briefly highlights the main terms from algebraic geometry; we refer to Cox et al. (2015) for further reading.

---
[3]Here we follow the (more natural) definition of Chen et al. (2014) rather than that of Foygel et al. (2012).
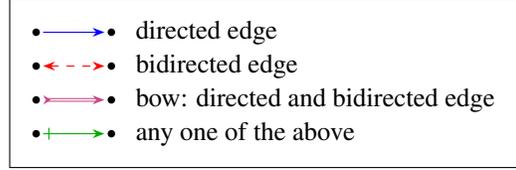

Figure 4: Legend for the edges we use to denote (sets of) graphs.

Algebraic geometry studies sets of points defined by systems of polynomial equations. For points in $\mathbb{R}^n$, such a set is called an *affine variety*. The set of polynomials that are identically zero on some set of points form an *ideal*. The *Zariski closure* of a set of points is obtained by first finding the ideal of the set, then taking all points for which the polynomials in the ideal all vanish.

2.3 NOTATION FOR SETS OF GRAPHS

The notation described here will be used in the remaining figures in this paper, and extensively in Appendix B in the supplementary material. When drawing a mixed graph, we draw directed edges in solid blue and bidirected edges in dashed red (see Figure 4; note that the colours are redundant but may aid visual distinction). Because we often want to show a set of graphs with a common node set V (e.g., an algebraic equivalence class of graphs), we use some new notation to avoid listing all graphs one by one. This notation is based on the *skeletons* of the graphs. Formally, the skeleton $S(G)$ of $G = (V, D, B)$ is the undirected graph on V that has an edge between a pair of nodes if there is at least one edge of any type between them in G. Similarly, our *graph patterns* also have at most one edge between each pair of nodes, with different markings to show what (combinations of) edges may occur between those nodes. For example, in this notation, a bow (a directed and a bidirected edge between two nodes) is represented visually as a double magenta line with 'fletching' at the back.

In a pattern representing a set of graphs, other edges occur in places where the graphs differ. The only such edge that appears in the patterns in the main paper is a green arrow with a plus sign as its tail. In all of these patterns, it can be understood to mean either a directed edge, a bidirected edge, or both. Its meaning changes if another edge is incident at the endpoint with the plus sign; for this case, additional markings are used and defined in Appendix B.

3 THEORETICAL RESULTS

We now present our main theoretical results.

3.1 ALGEBRAIC CONSTRAINTS

The statistical model $\mathcal{M}(G)$ for a graph G is defined by (1) in terms of the parameterization ϕ_G. This makes it hard to decide whether a given Σ belongs to a model, or whether two models differ. For these purposes, a more usable description of $\mathcal{M}(G)$ would be as a set of constraints that must hold for any $\Sigma \in \mathcal{M}(G)$: a set of functions $f_1, \ldots, f_k : \Sigma \to \mathbb{R}$ such that

$$\mathcal{M}(G) \subseteq \{\Sigma \in \mathrm{PD}_{|V|} \mid f_1(\Sigma) = \ldots = f_k(\Sigma) = 0\},$$

with the set difference a measure zero subset of the right-hand set.[4] For a DAG G, this can be done by choosing each f_i to be a partial correlation: $f_i(\Sigma) = \rho_{v_i w_i . S_i}(\Sigma)$ (Richardson and Spirtes, 2002). Put differently, vanishing partial correlation constraints fully describe such models.

In the more general case where G is allowed to contain bidirected edges and directed cycles, vanishing partial correlation constraints are not always expressive enough to describe $\mathcal{M}(G)$ (as we saw in the example of Figure 1). Below we present a method that allows us to write down a list of equality constraints that together describe $\mathcal{M}(G)$, for any *HTC-identifiable graph* G (Foygel et al., 2012).

Any model $\mathcal{M}(G)$ is a *semi-algebraic* subset of $\mathrm{PD}_{|V|}$: it can be described by a set of polynomial (in)equalities (Sullivant et al., 2010). Thus we may use the terms 'equality constraint' and 'algebraic (i.e. polynomial) constraint' interchangeably, and similarly for 'inequality' and 'semi-algebraic'.

The question of finding constraints that describe $\mathcal{M}(G)$ is related to the question of parameter identifiability. Identifiability of a graph G means that for given Σ, *at most* one pair of parameter values (Λ, Ω) exists for which $\phi_G(\Lambda, \Omega) = \Sigma$ (in other words, the *fibre* $\phi_G^{-1}(\Sigma)$ has cardinality at most one). We ask for what Σ (up to measure zero subsets as above) *exactly one* such pair exists. The main tool used here is the algorithm for retrieving the parameters of an HTC-identifiable graph that appears in the proof of Theorem 1 of Foygel et al. (2012), and that we sketched in Section 2.1. For given G and $\mathcal{Y} = (Y_v)_v$ (which we will call *HTC-identifying sets*) satisfying the conditions of the HTC-identifiability theorem, this algorithm defines a rational function that maps Σ to Λ by solving a sequence of linear equation systems. We write this function as $\Lambda_{\mathcal{Y}}(\Sigma)$.[5] Points Σ for which the algorithm encounters a singular matrix are excluded from the domain of $\Lambda_{\mathcal{Y}}$.

---

[4]Put more precisely, the difference is contained in an affine variety of lower dimension.

[5]Note that this function does not depend on the ordering $\prec$ that also appears in the HTC-identifiability theorem.

Theorem 1 below shows that a graph G imposes the following constraints:

$$[(I - \Lambda_{\mathcal{Y}}(\Sigma))^T \Sigma (I - \Lambda_{\mathcal{Y}}(\Sigma))]_{vw} = 0$$
$$\text{for all } \{v, w\} \notin B \text{ with } v \notin Y_w \text{ and } w \notin Y_v. \quad (2)$$

Each left-hand side in these equations is a rational function of Σ: a function of the form $p(\Sigma)/q(\Sigma)$, with p and q polynomials. Instead of these rational constraints, it will often be useful to consider the polynomial constraints $p(\Sigma) = 0$ obtained by multiplying out the denominator. For Σ with $q(\Sigma) \neq 0$, $p(\Sigma) = 0$ iff $p(\Sigma)/q(\Sigma) = 0$, so the two forms of the constraints agree whenever both are defined (see the proof of Theorem 1 for details). The polynomial constraints have the advantage that they are defined everywhere.

As an example, consider again the graph in Figure 1. For $\mathcal{Y}$ with $Y_b = \{a\}$, $Y_d = \{c\}$ and $Y_a = Y_c = \varnothing$, (2) gives us one rational constraint (for $v = b$, $w = d$):

$$\begin{bmatrix} 1 & -\Lambda_{\mathcal{Y}}(\Sigma)_{ab} \end{bmatrix} \begin{bmatrix} \sigma_{bd} & \sigma_{bb} \\ \sigma_{ad} & \sigma_{ab} \end{bmatrix} \begin{bmatrix} 1 \\ -\Lambda_{\mathcal{Y}}(\Sigma)_{bd} \end{bmatrix} = 0,$$

where $\Lambda_{\mathcal{Y}}(\Sigma)_{ab} = \sigma_{ab}/\sigma_{aa}$ and $\Lambda_{\mathcal{Y}}(\Sigma)_{bd} = \sigma_{cd}/\sigma_{bc}$. Multiplying out the denominators, we obtain the polynomial constraint

$$\sigma_{aa}\sigma_{bd}\sigma_{bc} - \sigma_{aa}\sigma_{bb}\sigma_{cd} - \sigma_{ab}\sigma_{ad}\sigma_{bc} + \sigma_{ab}^2\sigma_{cd} = 0. \quad (3)$$

Theorem 1. *For an HTC-identifiable graph G with HTC-identifying sets $\mathcal{Y} = (Y_v)_v$ and generic Σ, $\Sigma \in \mathcal{M}(G)$ iff $\Lambda_{\mathcal{Y}}(\Sigma) \in \mathbb{R}_{reg}^D$ and Σ satisfies the rational constraints (2). A stronger statement holds in one direction: All (not merely generic) $\Sigma \in \mathcal{M}(G)$ satisfy the polynomial constraints described above.*

This means that the constraints (2) define the model $\mathcal{M}(G)$ up to a measure zero set that may satisfy all constraints, but still be missing from $\mathcal{M}(G)$. In particular, it shows that the model of an HTC-identifiable graph imposes no semi-algebraic (i.e. inequality) constraints. This is not true for general graphs, as we will see in Section 3.3.

The HT-overidentifying constraints from Chen et al. (2014) are also based on (a version of) the half-trek criterion, but unlike (2), they do not give a full description of the algebraic constraints imposed by a graph G: no constraint is found for pairs $\{v, w\}$ with $v \notin \mathrm{htr}(w)$ and $w \notin \mathrm{htr}(v)$. An example is given in Appendix A.

There may often be several ways of expressing Λ in terms of Σ, using different HTC-identifying sets $\mathcal{Y}$. As a result, the rational constraints may look very different, though they may become the same when converted to polynomial form. For example, a different choice of $\mathcal{Y}$ for the

graph in Figure 1 would have led to a different rational expression than we found above. Similarly, sets of constraints may be found which are different when compared one by one, even in polynomial form, but which together still define the same model (in the terminology of algebraic geometry, they generate the same ideal). For example, if $\sigma_{ab} = 0$, then $\sigma_{ac} = 0$ iff $\rho_{ac.b} = 0$, so that a model satisfying all these constraints can be described in two ways using two constraints, or redundantly using all three constraints. So while the set of constraints (2) is not unique, Theorem 1 shows that they are a complete description of $\mathcal{M}(G)$: the model imposes no algebraic constraints beyond these.

Theorem 1 only applies to HTC-identifiable graphs. This immediately excludes graphs which are not generically identifiable; we will revisit those in Sections 3.2 and 3.3. However, it also excludes graphs that are generically identifiable but not HTC-identifiable. For the case of acyclic graphs on four variables, HTC-identifiability is complete (in the sense that all generically identifiable graphs are also HTC-identifiable), but this is no longer true when either more nodes or cycles are allowed. Finding more general sufficient graphical criteria for generic identifiability is the topic of ongoing research (Chen et al., 2014; Chen, 2016; Drton and Weihs, 2016; Chen et al., 2017). Because these criteria are extensions of HTC-identifiability, our Theorem 1 might be extended to work with these criteria as well.

3.2 ALGEBRAIC EQUIVALENCE AND INFINITE-TO-ONE GRAPHS

We now turn to the second central problem we address in this paper, namely checking whether two graphs G_1 and G_2 are algebraically equivalent. Theorem 1 from the previous section in principle allows us to do this: we need to check that each equality constraint imposed by G_1 is implied by G_2's equality constraints, and vice versa. However, the general solution for such tasks from algebraic geometry (computing Gröbner bases (Cox et al., 2015)) is computationally extremely expensive. The theorem below gives a sufficient condition for algebraic equivalence that relies only on graphical criteria that can be checked efficiently.

Another limitation of Theorem 1 is that it only applies to HTC-identifiable graphs, so it does not help us in finding equivalences involving graphs that are not HTC-identifiable. This issue is also addressed by the following theorem.

Theorem 2. *If G is generically infinite-to-one, G' generically finite-to-one and obtained by deleting k edges from G, and this k is the smallest number for which such a G' exists, then G and G' are algebraically equivalent. If*

further G' imposes no inequality constraint, $\mathcal{M}(G)$ and $\mathcal{M}(G')$ are equal up to a measure zero subset.

A relation between forms of model equivalence and lack of parameter identifiability has been shown previously by Bekker et al. (1994). The following two corollaries serve to illustrate the power of this theorem.

Corollary 3. *All infinite-to-one graphs are algebraically equivalent to some finite-to-one graph.*

This implies that it is not a limitation that Theorem 1 does not apply to infinite-to-one graphs: for any such graph, these always exists a finite-to-one graph that we can consider instead. In particular, if this finite-to-one graph is HTC-identifiable, then this gives us a complete description of the infinite-to-one graph as a set of equality constraints.

Corollary 4. *If two generically finite-to-one graphs G'_1, G'_2 are each obtained by deleting an edge from a single generically infinite-to-one graph G, then G'_1 and G'_2 are algebraically equivalent to each other (and their models are equal up to a measure zero subset if they impose no inequality constraints).*

Combined with graphical criteria for generic (in)finite-to-oneness, this gives a sufficient *graphical* condition for model equivalence. As an example, take G'_1 and G'_2 to be the two graphs in Figure 2, and G to be their union (thus having three nodes and four edges). G'_1 and G'_2 are generically finite-to-one, while G is generically infinite-to-one (it is HTC-nonidentifiable). Then Corollary 4 states that G'_1 and G'_2 are algebraically equivalent. In fact, both graphs are HTC-identifiable, so they impose no inequality constraints, and their models are thus equal up to measure zero subsets.

By repeatedly applying Corollary 4, equivalence of many more pairs of models may be established. This is illustrated by the following proposition (which is similar to (Nowzohour et al., 2017, Theorem 2), though there the stronger relation of distributional equivalence is shown).

Proposition 5. *If two bow-free acyclic graphs have the same skeleton, and any 2-edge path through three distinct nodes that is a collider in one graph is a collider in both, the graphs are algebraically equivalent.*

3.2.1 Algebraic Equivalence Classes on Four Nodes

Using Theorems 1 and 2 and information on the identifiability of the acyclic graphs on four nodes, we can determine the algebraic equivalence classes of these graphs. An explicit description of these classes is given in Appendix B. Here we describe some details of how these results were derived.

Using HTC-identifiability and HTC-nonidentifiability, almost all acyclic graphs on four nodes can be classified as either generically identifiable or generically infinite-to-one. With the additional information from (Foygel et al., 2012) that the remaining graphs are generically finite-to-one (see Section 3.3), we can apply Theorem 2 to all these graphs. Together with transitivity, this partitions the set of graphs into 419 subsets, which we will call *clusters* here. Because the theorem only gives a sufficient condition for algebraic equivalence, this partition may be finer than the partition into algebraic equivalence classes. We still need to check if algebraic equivalences exist between different clusters.

Among clusters imposing two or more equality constraints, all but three (up to graph isomorphism) can be described by vanishing (partial) correlation constraints. Because graphs in different Markov equivalence classes must also be in different algebraic equivalence classes, we only need to focus on these remaining three clusters. One of these imposes the Verma constraint (Verma and Pearl, 1991); the other two both impose the constraints $\sigma_{cd} = 0$ and $\sigma_{ac}\sigma_{bd} - \sigma_{ad}\sigma_{bc} = 0$ (a vanishing tetrad constraint). These two latter clusters are shown in Figure 5(a) and (b). So those two clusters are algebraically equivalent, while no other algebraic equivalences among these graphs were missed by Theorem 2.

For graphs imposing only one equality constraint, it is much easier to check if two graphs impose the same equality constraint, as we do not need to worry about the possibility of two equality constraints implying a third (see Section 3.1). In algebraic terminology, these models are described by principal ideals; for these, equality can be checked by normalizing the generating polynomials so that their leading coefficients equal one (Cox et al., 2015). This way, we find that among graphs imposing one equality constraint, the three clusters (up to isomorphism) shown in Figure 5(c), (d) and (e) are actually algebraically equivalent to each other, while all others are different. This leaves a total of 389 algebraic equivalence classes.

Interestingly, the graphs imposing a vanishing tetrad constraint and no other equality constraints were all determined to be algebraically equivalent by Theorem 2.

3.2.2 Consequences for Model Selection

The theoretical results we presented above offer enormous benefits to model selection, in particular to score-based methods. Without any knowledge of model equivalence, a score-based method for model selection may in principle need to score all different graphs. Even for the limited case of acyclic graphs on four nodes, there are 34752 distinct graphs, making such an approach clearly
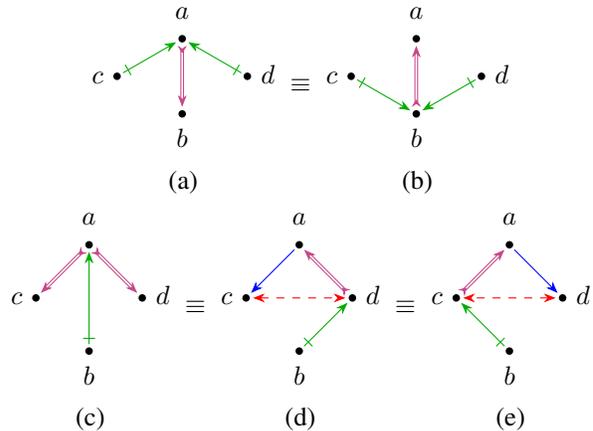

Figure 5: The algebraic equivalences not detected by Theorem 2: clusters (a) and (b) are algebraically equivalent to each other, and the same is true for clusters (c), (d) and (e).

not practical. For this reason, most score-based methods for model selection rely on a coarser concept of equivalence such as Markov equivalence, or even limit themselves to DAGs, ignoring the possibility of latent confounders entirely (Drton and Maathuis, 2017). As demonstrated in Section 3.2.1, our theoretical results can be used to determine the 389 algebraic equivalence classes. Scoring just one representative of each class leads to huge computational savings, making algebraic equivalence class selection feasible. As will be elaborated on in Section 4, the gains become even larger when these representatives are chosen cleverly: for example, the maximum likelihood parameters of a DAG are generally much easier to compute than those of a graph which also includes bidirected edges, so by picking a DAG from each class that contains one, we can avoid many relatively expensive score computations on more complex graphs.

When we are looking for maximum likelihood parameters for an HTC-identifiable graph G but our maximum likelihood fitting procedure has difficulty converging, it may be beneficial to apply the procedure to an 'easier' algebraically equivalent graph G' instead. After maximum likelihood parameters (Λ', Ω') have been found, we can compute $\Sigma = \phi_{G'}(\Lambda', \Omega')$, the point in $\mathcal{M}(G')$ where the likelihood is maximized; this will be the same point for all algebraically equivalent models (up to the nongeneric case where $\Sigma \notin \mathcal{M}(G)$—though then points arbitrarily close to Σ will be included in $\mathcal{M}(G)$). For this Σ, we can compute parameters (Λ, Ω) for which $\phi_G(\Lambda, \Omega) = \Sigma$ using the algorithm in the proof of Theorem 1 of Foygel et al. (2012).

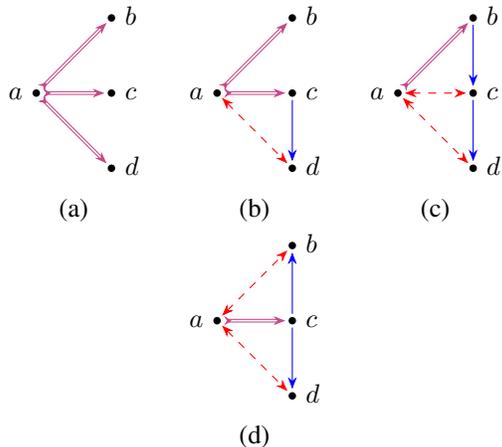Because Theorem 2 only provides a sufficient condition

Figure 6: Four acyclic finite-to-one graphs.

for algebraic equivalence, an automatic procedure for model selection based on this theorem may sometimes fail to recognize that two classes of graphs are equivalent, and consider both separately (unless the results of Theorem 1 are also considered by the algorithm, similar to how we used them above). While this means that some redundant computational work is done, it does not hurt the quality of model selection, and the gains in computation time are still enormous compared to testing all models individually: 34752 acyclic mixed graphs on four nodes are grouped into 419 clusters by Theorem 2, which is very close to the 389 algebraic equivalence classes we would find by also looking at the constraints.

3.3 FINITE-TO-ONE GRAPHS AND INEQUALITY CONSTRAINTS

Among the acyclic graphs on four nodes, four (up to graph isomorphism) are HTC-inconclusive. These graphs are shown in Figure 6. By (Foygel et al., 2012, Table 1), these graphs are generically finite-to-one but not identifiable. We used this in Section 3.2.1 to assign them to algebraic equivalence classes, and found that all are algebraically equivalent to the saturated model.

Because these graphs are not generically identifiable, Theorem 1 does not apply to them, so we cannot rule out the possibility that they impose inequality constraints. Indeed, the proposition below shows that one of the graphs imposes such a constraint, so that its model differs from the saturated model by a subset of nonzero measure. We believe the other graphs listed here impose similar constraints, and expect that the same is true for many other graphs that are generically finite-to-one but not generically identifiable.

Proposition 6. *The model of the graph in Figure 6(a) imposes the inequality constraint $\rho_{bc.a} \cdot \rho_{cd.a} \cdot \rho_{bd.a} \leq 0$.*

4 EXPERIMENTAL RESULTS

In Section 4.2, we will describe the results of a model selection experiment using algebraic equivalence classes, choosing the class with the best BIC score from among those found in Section 3.2.1. We define the BIC score of an equivalence class as the best score among its models; this can be determined by computing the maximum likelihood of just one member model. RICF (Drton et al., 2009) was used to find maximum likelihood parameters. We will first describe some empirical results about its convergence behaviour.

4.1 CONVERGENCE BEHAVIOUR OF RICF

As already mentioned in Section 3.2.2, knowing the algebraic equivalence classes is very useful in a model selection problem. Here we elaborate on this, based on Monte Carlo results.

RICF finds maximum likelihood parameters in its first iteration if the graph is a DAG (Drton et al., 2009). For all algebraic equivalence classes not containing a DAG (so all graphs in the class contain a bidirected edge), we saw evidence of local optima: when RICF was run with the same random data on different graphs in the same class, or with different initialization values, there would be different runs that all reported convergence but achieved different likelihoods. This could be addressed by random restarts. However, Drton and Richardson (2004) observe that for one of these graphs, local optima only present themselves when the model is misspecified. If the same is true for all graphs with bidirected edges, then it would follow that local minima will not change the results of model selection, as they only affect the scores of models that would not have scored well anyway. Either way, it is clearly advantageous to run RICF on a DAG, for every algebraic equivalence class that contains one.

For equivalence classes where all graphs contain a bow, RICF failed to converge relatively often (in fact, Drton et al. (2009) only discuss bow-free graphs). These equivalence classes contain graphs having different skeletons; as an example, the nine graphs represented by the pattern in Figure 5(a) have a different skeleton than those in Figure 5(b), but all are algebraically equivalent to each other. We found that for many data sets, RICF failed to converge on all graphs in such a class having one skeleton, even with random restarts, but did converge easily on graphs with a different skeleton. Knowing the algebraic equivalence classes is a great benefit here, as it would be very difficult to determine scores for all graphs in this situation otherwise. When scoring a model, if we find that RICF fails to converge on a graph of this type (e.g., one from Figure 5(a)), then we run RICF again on an alge-
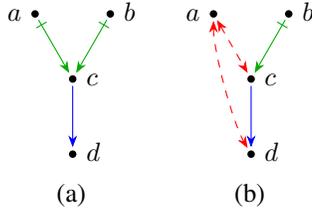
Figure 7: (a) A Y-structure; (b) an extended Y-structure.

braically equivalent graph with a different skeleton (one from Figure 5b).

4.2 USING ALGEBRAIC EQUIVALENCE CLASS SELECTION TO DETECT Y-STRUCTURES

To demonstrate the practical usefulness of algebraic equivalence class selection, we consider the setup of Mooij and Cremers (2015).[6] In a simulated dataset with $p \in \{10, 30, 50\}$ variables (with acyclic ground truth), they look at each ordered 4-tuple of distinct nodes, and use several independence tests to detect whether or not these nodes form a *Y-structure* or an *extended Y-structure* (see Figure 7). These are two of the simplest Markov equivalence classes (coinciding with algebraic equivalence classes) that must contain a directed edge in a fixed place that is not part of a bow, so detecting these structures in observational data allows us to draw conclusions about the results of interventions.

Mooij and Cremers observed that detecting 4-tuples which were in either of these two classes yielded poor precision, especially for larger p. Precision improved when additional tests were added so that only Y-structures were detected.

We applied algebraic equivalence class selection to this problem by taking the 4-tuples that tested positively according to one of these sets of tests, and then filtering out those for which the algebraic equivalence class with the best BIC score was different from the class / pair of classes being tested for. The resulting precisions are shown in Figure 8. The gains of this filtering procedure are significant when testing for both classes together: the precision is close to 0.1 larger for all p. We also show the results of filtering using only Maximal Ancestral Graphs (MAGs) (Richardson and Spirtes, 2002), thus only computing BIC scores of classes which can be described using vanishing partial correlations. This already yields a large improvement, but the gains from considering all algebraic equivalence classes instead are still significant, especially for the larger p. On the other hand, for 4-tuples

---

[6]The code for reproducing these results is available online at `https://github.com/caus-am/aelsem`.
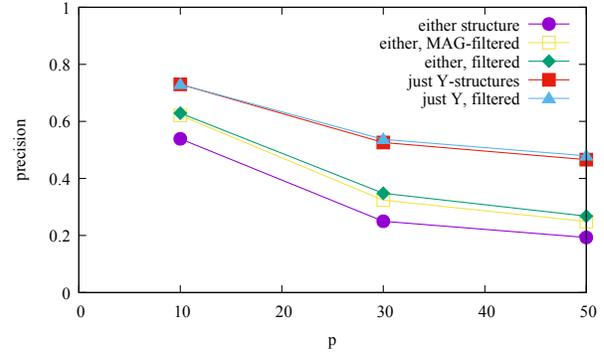

Figure 8: Precision of (extended) Y-structure detection.

that passed the more stringent Y-structure tests, the benefit of filtering is much smaller. Importantly, these gains in precision came at a very small cost in recall: of the true positives detected by a set of independence tests, at least 98% passed through the filter for each p.

5 CONCLUSION

We proposed the notion of algebraic equivalence for use in model selection among linear structural equation models, and showed how graphical criteria can be used to efficiently establish the equivalence of many models. Our experimental results show that the finer granularity of the resulting equivalence classes sometimes allows for improvements in model selection accuracy, compared to methods that only consider vanishing partial correlations.

While the experimental model selection results we show are limited to four nodes, we believe that the theoretical results described here can form the foundation of efficient causal inference algorithms on (much) larger numbers of nodes. For this, more work will need to be put into finding an efficient algorithm, for example by using ideas from Chickering (2002).

References

Paul A. Bekker, Arjen Merckens, and Tom J. Wansbeek. *Identification, Equivalent Models, and Computer Algebra*. Academic Press, Boston, 1994.

Carlos Brito and Judea Pearl. A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9(4):459–474, 2002.

Bryant Chen. Identification and overidentification of linear structural equations models. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 1579–1587, 2016.

Bryant Chen, Jin Tian, and Judea Pearl. Testable implications of linear structural equations models. In C. E. Brodley and P. Stone, editors, *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

Bryant Chen, Daniel Kumor, and Elias Bareinboim. Identification and model testing in linear structural equation models using auxiliary variables. Technical Report R-27-L, Purdue AI Lab, 2017.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

David A. Cox, John Little, and Donal O'Shea. *Ideals, Varieties, and Algorithms*. Springer, New York, fourth edition, 2015.

Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Applications*, 4:365–393, 2017.

Mathias Drton and Thomas S. Richardson. Multimodality of the likelihood in the bivariate seemingly unrelated regression model. *Biometrika*, 91(2):383–392, 2004.

Mathias Drton and Luca Weihs. Generic identifiability of linear structural equation models by ancestor decomposition. *Scandinavian Journal of Statistics*, 43:1035–1045, 2016.

Mathias Drton, Michael Eichler, and Thomas S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10:2329–2348, 2009.

Robin J. Evans. Margins of discrete Bayesian networks. *arXiv preprint arXiv:1501.02103*, 2015.

Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, 40(3):1682–1713, 2012.

Luis D. García-Puente, Sarah Spielvogel, and Seth Sullivant. Identifying causal effects with computer algebra. In P. Grünwald and P. Spirtes, editors, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, 2010.

Joris M. Mooij and Jerome Cremers. An empirical study of one of the simplest causal prediction algorithms. In Ricardo Silva, Ilya Shpitser, Robin J. Evans, Jonas Peters, and Tom Claassen, editors, *UAI 2015 Workshop on Advances in Causal Inference*, number 1504 in CEUR Workshop Proceedings, pages 30–39, Aachen, 2015. URL http://ceur-ws.org/Vol-1504/uai2015aci_paper2.pdf.

Christopher Nowzohour, Marloes H. Maathuis, Robin J. Evans, and Peter Bühlmann. Structure learning for bow-free acyclic path diagrams. *arXiv preprint arXiv:1508.01717v3*, 2017.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000.

Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *The Annals of Statistics*, 7:1393–1512, 1986.

Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

Ricardo Silva and Zoubin Ghahramani. Bayesian inference for Gaussian mixed graph models. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pages 453–460, 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, second edition, 2000.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for Gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI 2002)*, 2002.

T. S. Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI 1991)*, 1991.

SUPPLEMENTARY MATERIAL

A PROOFS

Proof of Theorem 1. First assume $\Sigma \in \mathcal{M}(G)$, and consider the equations

$$[(I - \Lambda)^T \Sigma (I - \Lambda)]_{vw} = 0 \quad \text{for all } \{v, w\} \notin B \quad (4)$$

(this is equation (9.2) in Foygel et al. (2012)). Treating the entries of Λ and Σ as variables λ and σ, define the rational function $\delta := \det(I - \Lambda)^{-1}$. Then the left-hand sides of (4) are polynomials in $\mathbb{R}[\lambda, \sigma, \delta]$ generating an ideal $\mathcal{I}$. This ideal consists of all polynomials that are zero everywhere on the set of all (λ, σ) with δ well-defined and $\Sigma = \phi_G(\Lambda, \Omega)$ for some $\Omega \in \mathrm{PD}(B)$ (Foygel et al., 2012, Section 8 of the supplement). In addition to the left-hand sides of (4), (using that G is rationally identifiable) the ideal $\mathcal{I}$ contains for each directed edge $(u, v) \in D$ an element $a(\sigma)\lambda_{uv} - b(\sigma)$ for some nonzero polynomials $a, b \in \mathbb{R}[\sigma]$. By multiplying a left-hand side of (4) by some polynomial and adding polynomial multiples of the ideal elements mentioned above, we can eliminate λ and find an element in $\mathcal{I} \cap \mathbb{R}[\sigma]$. This represents a polynomial constraint that *all* $\Sigma \in \mathcal{M}(G)$ must satisfy, and we see that the polynomial constraints described in the main text can be obtained in this way. This implies that also the rational constraints (2) are satisfied for any Σ that is additionally in the domain of $\Lambda_{\mathcal{Y}}$.

If $\Sigma \in \mathcal{M}(G)$, then there exist $\Lambda \in \mathbb{R}^D_{\mathrm{reg}}$ and $\Omega \in \mathrm{PD}(B)$ for which $\Sigma = \phi_G(\Lambda, \Omega)$. In the generic case, $\Lambda_{\mathcal{Y}}$ will recover the parameter matrix Λ; in particular, $\Lambda_{\mathcal{Y}}(\Sigma) \in \mathbb{R}^D_{\mathrm{reg}}$.

Conversely, suppose Σ is in the domain of $\Lambda_{\mathcal{Y}}$, $\Lambda_{\mathcal{Y}}(\Sigma) \in \mathbb{R}^D_{\mathrm{reg}}$, and Σ satisfies (2). Let $\Lambda = \Lambda_{\mathcal{Y}}(\Sigma)$ and $\Omega = (I - \Lambda)^T \Sigma (I - \Lambda)$. Σ and Λ again satisfy (4):

For $\{v, w\}$ with $v \notin Y_w$ and $w \notin Y_v$: by (2);

For $\{v, w\}$ with $w = y_i \in Y_v \cap \mathrm{htr}(v)$:

$$[(I - \Lambda)^T \Sigma \Lambda]_{wv} = (\mathbf{A} \cdot \Lambda_{\mathrm{pa}(v),v})_i = \mathbf{b}_i$$
$$= [(I - \Lambda)^T \Sigma]_{wv}$$

(with $\mathbf{A}$ and $\mathbf{b}$ as in Foygel et al. (2012) / in Section 2.1), which implies (4) for $\{v, w\}$;

For $\{v, w\}$ with $w \in Y_v \setminus \mathrm{htr}(v)$: Fix any node v, and let $W = Y_v \setminus \mathrm{htr}(v)$ and $I = \{i \mid y_i \in W\}$. Then

$$[\Sigma \Lambda]_{Wv} = (\mathbf{A} \cdot \Lambda_{\mathrm{pa}(v),v})_I = \mathbf{b}_I = \Sigma_{Wv},$$

from which follows

$$\begin{aligned}
\mathbf{0} &= [\Sigma(I - \Lambda)]_{Wv} \\
&= [(I - \Lambda)^{-T}(I - \Lambda)^T \Sigma (I - \Lambda)]_{Wv} \\
&= [(I - \Lambda)^{-T} \Omega]_{Wv} = \sum_{x \in V} [(I - \Lambda)^{-T}]_{Wx} \Omega_{xv}.
\end{aligned}$$

In the case distinction that follows, we use that for all $x \neq v$ with $x \notin \mathrm{sib}(v)$, we have $x \in Y_v$, $v \in Y_x$, or neither. For x in one of the sets $Y_v \cap \mathrm{htr}(v)$, $\{x \mid v \in Y_x \cap \mathrm{htr}(x)\}$, or $\{x \mid x \neq v, x \notin \mathrm{sib}(v), x \notin Y_v, v \notin Y_x\}$, we know from the previous two cases that $\Omega_{xv} = \Omega_{vx} = 0$. For $x = v$ or $x \in \mathrm{sib}(v)$, $[(I - \Lambda)^{-T}]_{wx} = 0$ for any $w \in W$, because such a directed path from such x to w would form a half-trek from v to w. Similarly, for x such that $v \in Y_x \setminus \mathrm{htr}(x)$, also $[(I - \Lambda)^{-T}]_{wx} = 0$, or there would again be a half-trek from v to w. The above covers all cases except $x \in W$. So all terms in the previous sum with $x \notin W$ are zero, and we get

$$\mathbf{0} = [(I - \Lambda)^{-T}]_{WW} \Omega_{Wv}.$$

The entries of $[(I - \Lambda)^{-T}]_{WW}$ are polynomials in the variables Λ_{ij} and $\delta := \det(I - \Lambda)^{-1}$. The matrix $(I - \Lambda)^{-1}$ equals δ times the adjugate matrix of $(I - \Lambda)$; this adjugate matrix includes the constant term 1 in each diagonal entry, and no constant terms in any other entries. We see that the determinant of $[(I - \Lambda)^{-T}]_{WW}$ contains the term $\delta^{|W|}$ so it is not the zero polynomial. It follows that $[(I - \Lambda)^{-T}]_{WW}$ is generically invertible, so the only solution to the above equation is $\Omega_{wv} = 0$ for all $w \in W$. This again implies (4).

Then $\Omega = (I - \Lambda)^T \Sigma (I - \Lambda)$ is positive definite (using that Σ is positive definite and $I - \Lambda$ is invertible) and obeys $\Omega_{vw} = 0$ for $\{v, w\} \notin B$. For these parameters, $\Sigma = \phi_G(\Lambda, \Omega)$, showing that $\Sigma \in \mathcal{M}(G)$.

For generic $\Sigma \in \mathcal{M}(G)$, the matrices $\mathbf{A}$ that occur in the HTC-identifiability algorithm are invertible by (Foygel et al., 2012, Lemma 2), so the assumption above that Σ is in the domain of $\Lambda_{\mathcal{Y}}$ holds in the generic case. $\square$

Example of a constraint that is not HT-overidentifying. Figure 9 gives an example graph which imposes one equality constraint, namely a vanishing tetrad constraint, which is not reported by Algorithm 2 of Chen et al. (2014). The graph is HTC-identifiable, so Theorem 1 will report the constraint. $\square$

Figure 9: Graph imposing an equality constraint that is not HT-overidentifying.

Proof of Theorem 2. Define the *rank deficiency* of a graph as the difference between the number of columns and the generic rank of the Jacobian in (Foygel et al., 2012, proof of Theorem 2). The rank deficiency of a graph is zero iff that graph is finite-to-one, so the rank deficiency of G' is zero. Removing a directed edge from a graph removes a column from the Jacobian, which will not increase the rank, so this reduces the rank deficiency by at most one; removing a bidirected edge adds a row, which will increase the largest order of a non-vanishing minor (this order equals the rank) by at most one. It follows that the rank deficiency of G is at most k. In fact, it must be equal to k: suppose the rank deficiency of G is $k' < k$. Then by deleting k' columns from the Jacobian, a matrix of full column rank can be obtained. By deleting the corresponding k' directed edges from G, a graph with rank deficiency 0 is obtained; however, this is a contradiction with the conditions of the theorem, which imply that all such graphs are infinite-to-one.

By the statements in the beginning and end of the proof of Lemma 2 in the supplement of (Foygel et al., 2012), the dimensions of the images of ϕ_G and $\phi_{G'}$ (i.e. of $\mathcal{M}(G)$ and $\mathcal{M}(G')$) equal the ranks of the respective Jacobians as considered above plus the respective numbers of (co)variance parameters in Ω. It follows that the models' dimensions are equal. Because the models are defined by rational parameterizations, their Zariski closures are irreducible affine varieties (Cox et al., 2015, Proposition 4.5.6). For two such varieties, $W \subsetneq V$ would imply that the dimension of W is smaller than that of V. Because $\mathcal{M}(G') \subseteq \mathcal{M}(G)$, the models' Zariski closures must be equal; in other words, the models are algebraically equivalent.

If additionally G' imposes no inequality constraints, then $\mathcal{M}(G')$ is equal to its Zariski closure up to a measure zero subset. Using again that $\mathcal{M}(G') \subseteq \mathcal{M}(G)$ and that $\mathcal{M}(G)$ is contained in the models' common Zariski closure, we find that $\mathcal{M}(G')$ and $\mathcal{M}(G)$ are equal up to a measure zero subset. $\square$

Proof of Proposition 5. First consider the special case of two acyclic bow-free graphs G'_1, G'_2 that differ in only one arrowhead which does not form a collider. Then the graph G containing the union of their edges (so having a bow where G'_1 and G'_2 had a different type of edge) is HTC-nonidentifiable, as the directed edge of the bow cannot be identified. An application of Corollary 4 shows that G'_1 and G'_2 are algebraically equivalent.

The algebraic equivalence of two arbitrary graphs G'_1, G'_2 with the same skeleton and colliders can now be shown using a finite sequence of such steps. Let $G_1^\triangle$ be the subgraph consisting of the edges of G'_1 that are not

in G'_2, and let $W \subseteq V$ be the set of nodes where G'_1 and G'_2 have different incoming arrowheads. In $G_1^\triangle$, at each node $w \in W$, there is at most one arrowhead (otherwise there is a collider in either G'_1 or G'_2 that is not in the other) and at most one arrowtail (two tails would become two colliding heads in G'_2), so each node $w \in W$ has degree at most 2 in $G_1^\triangle$.

Now find the finest partition of the edges in $G_1^\triangle$ that groups two edges together if they are incident at a node $w \in W$. The edge sets in this partition form paths that may only meet each other at the endpoints, at nodes not in W. None of these paths can have two directed edges pointing away from each other (with possibly other edges in between), as that would imply G'_2 contains a collider not in G'_1. So each of these paths is a half-trek. At each step, modify the first edge of such a half-trek by adding an arrowhead that appears in G'_2 or removing one that does not. $\square$

Proof of Proposition 6. Assume w.l.o.g. that Σ is normalized to a correlation matrix (with ones on the diagonal). Let $W = \{b, c, d\}$. Then for each pair of distinct nodes $x, y \in W$,

$$
\begin{aligned}
\sigma_{xy} &= \lambda_{ax}\lambda_{ay} + \lambda_{ax}\omega_{ay} + \omega_{ax}\lambda_{ay} \\
&= (\lambda_{ax} + \omega_{ax})(\lambda_{ay} + \omega_{ay}) - \omega_{ax}\omega_{ay} \\
&= \sigma_{ax}\sigma_{ay} - \omega_{ax}\omega_{ay}.
\end{aligned}
$$

It follows that

$$
\omega_{ax}\omega_{ay} = \sigma_{ax}\sigma_{ay} - \sigma_{xy}. \tag{5}
$$

We have three of these equations for the three possible pairs in W. If all three ω_{aw}'s have the same sign, all these expressions will be positive; otherwise, two will be negative and one positive. Other combinations of positive and negative cannot be attained by any choice of Ω. Using that (5) has the same sign as $-\rho_{xy.a}$, we can write this as the inequality constraint $\rho_{bc.a} \cdot \rho_{cd.a} \cdot \rho_{bd.a} \leq 0$.

The observed covariance matrix

$$
\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1/2 & 1/2 \\ 0 & 1/2 & 1 & 1/2 \\ 0 & 1/2 & 1/2 & 1 \end{bmatrix}
$$

is an example of a positive definite matrix (so an element of the saturated model) that does not satisfy this inequality constraint. In fact, it is easy to see that an open ball around Σ has this property, showing that the inequality constraint rules out a part of the saturated model having nonzero measure. $\square$

B ACYCLIC ALGEBRAIC EQUIVALENCE CLASSES ON FOUR NODES

Before presenting in Section B.2 a full description of the algebraic equivalence classes of four-node acyclic graphs, we first introduce the additional graph notation we will use.

B.1 NOTATION FOR SETS OF GRAPHS, CONTINUED

In Section 2.3, we described how our graph patterns represent directed edges, bidirected edges, and bows. The full list of markings we use in Section B.2 is displayed in Figure 10. To describe these markings, we need to extend the well-known concept of colliders (see e.g. Spirtes et al. (2000)) to graph skeletons. A 2-edge path in the skeleton $S(G)$ of G along three distinct nodes (v_1, v_2, v_3) is called a *collider* if for both edges of the path, a corresponding edge exists in the original graph G with an arrowhead at v_2. Such an edge could be either bidirected, or directed towards v_2. We distinguish two cases: if *all* corresponding edges in the original graph G have arrowheads at v_2 (in other words, there are no directed edges in G from v_2 to either v_1 or v_3), we call the path a *full collider*; otherwise we call it a *partial collider*. An example of a partial collider is the path (a, b, c) in the instrumental variable model (Figure 2(a)). When looking at the endpoint at v_2 of an edge in $S(G)$ between v_1 and v_2, we say that it *forms a (full/partial) collider* if a node v_3 exists for which the path along (v_1, v_2, v_3) is a (full/partial) collider. Note that an endpoint can simultaneously form a full collider and a partial collider, for different choices of v_3.

In a graph pattern denoting a set of graphs, if all graphs in the set have the same edge(s) between two given nodes, this edge is displayed as for a single graph, as in Section 2.3. If the graphs in the set have different edges between two nodes, these are represented in the graph pattern by a green edge, which may have different endpoint markings. First of all, if all edges to be represented have an arrowhead at an endpoint, then the graph pattern will also show an arrowhead. Otherwise, the following special markings are used (illustrated in Figure 10): a plus sign if the endpoint never forms a collider in any of the graphs; a bracket if it always forms a partial collider; and a star in all other cases. Finally, if two nodes are adjacent in some graphs in the set but not in others, then an edge is shown in dotted grey; other aspects of the edge's appearance are determined using only the graphs in which the nodes are adjacent.
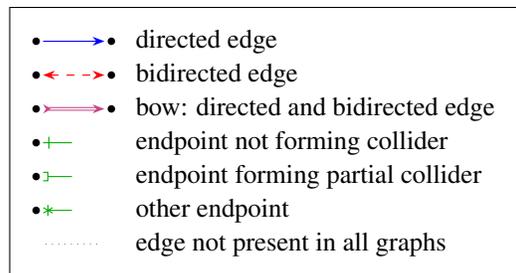

Figure 10: Legend for the edges and endpoints we use to denote (sets of) graphs.

B.2 TABLE OF EQUIVALENCE CLASSES

Each entry in the table below represents one algebraic equivalence class and its isomorphisms (obtained by re-labelling the nodes). The leftmost column shows the set of graphs in this equivalence class, using the notation for graph patterns explained in Sections 2.3 and B.1. For each equivalence class, its algebraic constraints (found by Theorem 1 and converted to polynomial form) are also listed.

For many equivalence classes, the set of graphs that fit the pattern in the leftmost column is a superset of the equivalence class: some acyclic graphs that match the pattern do not actually belong in the class, but our notation is not expressive enough to show the precise inclusions in one figure. For each equivalence class where this is the case, a second column of smaller graph patterns appears. The union of the sets of graphs described by these patterns equals the equivalence class exactly. Because many of the graph patterns in the union are graph isomorphisms of each other, we show only one member of each isomorphism class, and indicate how many isomorphisms there are by writing '$n\times$'. (Each of these isomorphisms is an automorphism of the leftmost graph pattern, and there are always exactly n such isomorphic patterns, so no information is lost by not listing them explicitly.) For example, the first equivalence class listed in Section B.2.3 consists of all acyclic graphs with at least one edge of any type between each pair of nodes in $\{a, b, c\}$, as well as all graphs in which one of those pairs has no edge between it, but the other two pairs make up a partial collider.

The equivalence classes are split out below according to their dimension. This refers to the dimension of the set of *correlation* rather than covariance matrices, thus normalizing out the variance Σ_{vv} associated with each node. This number has the convenient property that it equals the minimum number of edges among all graphs in an equivalence class.

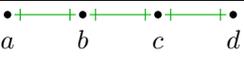### B.2.0 Dimension 0

$\sigma_{ab} = 0, \sigma_{ac} = 0, \sigma_{ad} = 0,$
$\sigma_{bc} = 0, \sigma_{bd} = 0, \sigma_{cd} = 0$

B.2.1 Dimension 1

$\sigma_{ac} = 0, \sigma_{ad} = 0, \sigma_{bc} = 0, \sigma_{bd} = 0, \sigma_{cd} = 0$
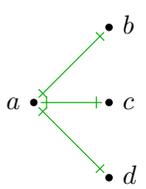
B.2.2 Dimension 2

$\sigma_{ac} = 0, \sigma_{ad} = 0, \sigma_{bc} = 0, \sigma_{bd} = 0$

$\rho_{ac.b} = 0, \sigma_{ad} = 0, \sigma_{bd} = 0, \rho_{cd.b} = 0$

$\sigma_{ac} = 0, \sigma_{ad} = 0, \sigma_{bd} = 0, \sigma_{cd} = 0$

B.2.3 Dimension 3

$1\times$
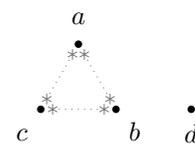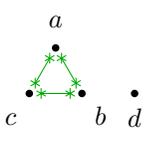
$\sigma_{ad} = 0, \sigma_{bd} = 0,$
$\sigma_{cd} = 0$

$3\times$
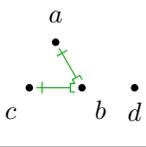
$\rho_{ac.b} = 0, \rho_{ad.b} = 0, \rho_{bd.c} = 0$

$\sigma_{ac} = 0, \sigma_{ad} = 0, \rho_{bd.c} = 0$

$\sigma_{ac} = 0, \sigma_{ad} = 0, \sigma_{bd} = 0$

$\rho_{bc.a} = 0, \rho_{bd.a} = 0, \rho_{cd.a} = 0$
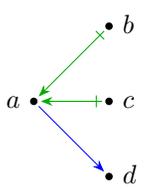
$\sigma_{bc} = 0, \rho_{bd.a} = 0, \rho_{cd.a} = 0$
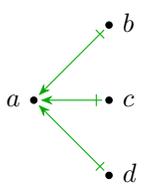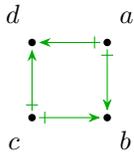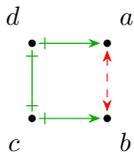

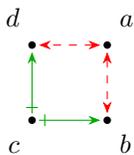$\sigma_{bc} = 0, \sigma_{bd} = 0, \sigma_{cd} = 0$

B.2.4 Dimension 4
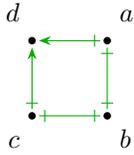
$\rho_{ac.b} = 0$, $\rho_{bd.ac} = 0$
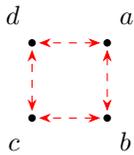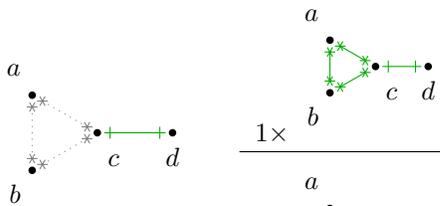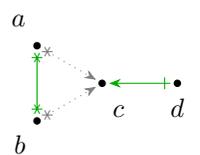

$\sigma_{ac} = 0$, $\rho_{bd.ac} = 0$

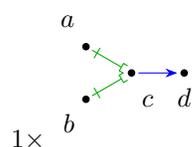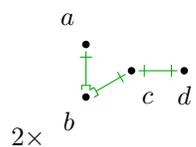$\rho_{ac.d} = 0$, $\rho_{bd.c} = 0$

$\sigma_{ac} = 0$, $\rho_{bd.c} = 0$

$\sigma_{ac} = 0$, $\sigma_{bd} = 0$

$\rho_{ad.c} = 0$, $\rho_{bd.c} = 0$

$1\times$

$2\times$

$1\times$
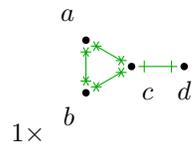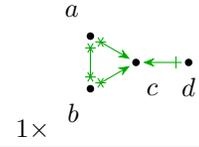
$\sigma_{ad} = 0$, $\sigma_{bd} = 0$

$1\times$
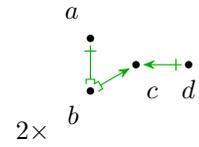
$2\times$
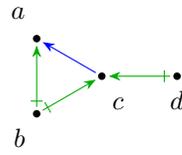
$\rho_{ad.bc} = 0$, $\sigma_{bd} = 0$

$\rho_{bd.c} = 0$,

$\sigma_{cc}\sigma_{ab}\sigma_{bd} - \sigma_{cc}\sigma_{bb}\sigma_{ad} - \sigma_{bc}\sigma_{ac}\sigma_{bd} + \sigma_{bc}^2\sigma_{ad} = 0$

$\sigma_{ad} = 0$, $\rho_{bd.c} = 0$
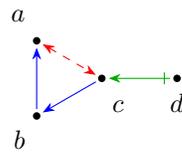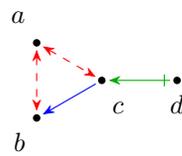

$2\times$

$\sigma_{cd} = 0$, $\sigma_{ac}\sigma_{bd} - \sigma_{ad}\sigma_{bc} = 0$

B.2.5 Dimension 5

$\sigma_{cd} = 0$

$1\times$

$2\times$

$4\times$

$4\times$

$\rho_{cd.ab} = 0$

$2\times$

$2\times$

$4\times$

$\rho_{cd.b} = 0$

$2\times$

$2\times$

$2\times$

$2\times$

$(p_1\sigma_{aa} - p_2\sigma_{ad})(\sigma_{cd}\sigma_{bb} - \sigma_{bc}\sigma_{bd})$
$- (p_2\sigma_{cd} - p_1\sigma_{ac})(\sigma_{ab}\sigma_{bd} - \sigma_{ad}\sigma_{bb}) = 0$

where $p_1 = \sigma_{bb}\sigma_{dd} - \sigma_{bd}^2$ and $p_2 = \sigma_{ad}\sigma_{bb} - \sigma_{bd}\sigma_{ab}$

$p_0^2\sigma_{cd} + p_0\sigma_{cb}p_2 + p_1\sigma_{bd}p_0 + p_1\sigma_{bb}p_2 = 0$

where $p_0 = \sigma_{aa}\sigma_{bb} - \sigma_{ab}^2$,

$p_1 = \sigma_{ab}\sigma_{ac} - \sigma_{aa}\sigma_{bc}$, and $p_2 = \sigma_{ab}\sigma_{ad} - \sigma_{aa}\sigma_{bd}$

$(\sigma_{aa}\sigma_{cc} - \sigma_{ac}^2)(\sigma_{bd}\sigma_{bc} - \sigma_{bb}\sigma_{cd})$
$+ (\sigma_{ac}\sigma_{bc} - \sigma_{cc}\sigma_{ab})(\sigma_{ad}\sigma_{bc} - \sigma_{ab}\sigma_{cd}) = 0$

$\sigma_{aa}\sigma_{cd}\sigma_{bb} - \sigma_{aa}\sigma_{cb}\sigma_{bd}$
$- \sigma_{ac}\sigma_{ad}\sigma_{bb} + \sigma_{ac}\sigma_{ab}\sigma_{bd} = 0$
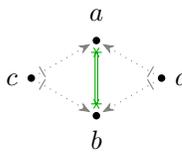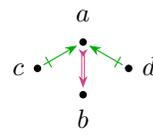
$$\sigma_{aa}\sigma_{bd}\sigma_{bc} - \sigma_{aa}\sigma_{bb}\sigma_{cd} - \sigma_{ab}\sigma_{ad}\sigma_{bc} + \sigma_{ab}^2\sigma_{cd} = 0$$

$1\times$

$$\sigma_{bd}\sigma_{ac}\sigma_{ab} - \sigma_{bd}\sigma_{aa}\sigma_{bc} - \sigma_{ab}^2\sigma_{cd} + \sigma_{ab}\sigma_{ad}\sigma_{bc} = 0$$

$2\times$

$$\sigma_{ab}\sigma_{cd} - \sigma_{ac}\sigma_{bd} = 0$$

$4\times$

$4\times$

B.2.6   Dimension 6

(No algebraic constraints)

$1\times$

$4\times$

$12\times$

$24\times$

$12\times$

$24\times$

$24\times$

$24\times$

$24\times$

$24\times$

$24\times$

$24\times$

$24\times$

$24\times$

$12\times$