Identifiability of binary directed graphical models with hidden variables

Elizabeth S. Allman
Dept. of Mathematics
and Statistics
University of Alaska
Fairbanks, AK 99775

John A. Rhodes
Dept. of Mathematics
and Statistics
University of Alaska
Fairbanks, AK 99775

Elena Stanghellini
Dipartimento di Economia
Finanza e Statistica
Università di Perugia
06100 Perugia, Italy

Marco Valtorta
Dept. of Computer Science
and Engineering
Univ. of South Carolina
Columbia, SC 29208

Abstract

Whether parameters of a DAG model with hidden variables can be identified is a difficult question. Here we give algebraic arguments establishing identifiability for two special DAG models with certain restrictions on the size of the finite state spaces of all variables. These results can be used to shed light on many other models. As an illustration, we address identifiability for all binary DAG models with at most 5 nodes and a single hidden variable parental to all observable ones.

1 Introduction

A parameterized statistical model is said to have *identifiable parameters* if a joint distribution for that model uniquely determines the parameters that produced it. Identifiability of parameters is a basic property that is essential for a model to be useful in most settings. In this work we focus on graphical models specified by directed acyclic graphs (DAGs), *i.e.* Bayesian networks not necessarily of the causal variety, and assume all variables have finite state spaces. Related work for undirected graphical models exists; see, for example, Stanghellini and Vantaggi (2013). If all variables of a DAG model are observable, then under mild assumptions (*e.g.*, positivity of all parameters) parameter identifiability is easy to establish.

Work initiated by Pearl (1995, 2009) investigated the identification of causal effects in causal Bayesian networks when some variables are assumed observable and some others are hidden. In a non-parametric setting, with no assumptions about the state space of variables, there is a complete algorithm for determining which causal effects between variables are identifiable (Huang and Valtorta, 2006; Shpitser and Pearl, 2008; Tian and Pearl, 2002; Pearl, 2012).

As powerful as this theory is, however, it does not address identifiability when one does make assumptions about the nature of the hidden variables. Indeed, by specializing to finite state spaces, causal effects that were non-identifiable according to the above mentioned theory may become identifiable. One fundamental result is due to Kruskal (1977), as developed in Allman et al. (2009).

More generally, with finite state spaces the question of whether parameters are identifiable for DAG models with hidden variables can be cast in an algebraic framework, as the parameterization map for such a model is polynomial. Given a distribution arising from the model, the parameters are identifiable precisely when a certain system of multivariate polynomial equations has exactly one solution (up to label-swapping of states for hidden variables). In principle, then, computational algebra software can be used to investigate parameter identifiability. However, the necessary calculations are usually intractable for even moderately large DAGs and/or state spaces. In addition, one runs into issues of complex versus real roots, and the difficulty of determining when real roots lie within stochastic bounds.

As a step toward a general approach to understanding parameter identifiability for DAG models, in this paper we consider this question for all possible DAG models with at most 5 binary variables, where one variable is hidden and the parent of all observable variables. See Table 1 of the Appendix for these graphs. For each such model, we establish that the parameterization map is generically k-to-one for a specific value of k. Our arguments are fundamentally algebraic, and do not depend on any machine computations. One particular example among these has also been studied in Kuroki and Pearl (2012), with reference to a specific causal effect.

We view the main contribution of this paper not as the determination of parameter identifiability for these specific models, but rather as the development of the techniques by which we show our results. We believe these examples will lead to a more general understanding of identifiability for finite state DAG models. Ultimately, one would like fairly simple graphical rules to determine which parameters are identifiable, and perhaps even to yield formulas for them in terms of the joint distribution. While it is unclear to what extent this is possible, even partial results covering only certain classes of DAGs, or some state spaces, are useful.

2 Discrete DAG models and parameter identifiability

The models we consider are specified in part by DAGs $\mathcal{G} = (V, E)$ in which nodes $v \in V$ represent random variables X_v, and directed edges in E imply certain independence statements for the joint distribution of all variables (Lauritzen, 1996). A bipartition of $V = O \sqcup H$ is given, in which variables associated to nodes in O or H are observable or hidden, respectively. Finally, we fix finite state spaces, of size n_v for each variable X_v.

A DAG $\mathcal{G}$ entails a collection of conditional independence statements on the variables associated to its nodes, via d-separation, or an equivalent separation criterion in terms of the moral graph on ancestral sets. The joint distribution of variables satisfying these statements has a factorization according to $\mathcal{G}$ as

$$P = \prod_{v \in V} P(X_v | X_{\mathrm{pa}(v)}),$$

with pa(v) denoting the set of parents of v in the DAG. We refer to the conditional probabilities $\theta = (P(X_v|X_{\mathrm{pa}(v)}))_{v \in V}$ as the *parameters* of the DAG model, and denote the space of all possible choices of parameters by $\Theta = \Theta_{\mathcal{G}}$. The parameterization map for the joint distribution of all variables, both observable and hidden, is denoted

$$\phi : \Theta \to \Delta^{(\prod_{v \in V} n_v)-1},$$

where Δ^k is the k-dimensional probability simplex of stochastic vectors in $\mathbb{R}^{k+1}$. Thus $\phi(\Theta)$ is precisely the collection of all probability distributions satisfying the conditional independence statements associated to $\mathcal{G}$ (and possibly additional ones).

Since the probability distribution for the model with hidden variables is obtained from that of the fully observable model, its parameterization map is

$$\phi^+ = \sigma \circ \phi : \Theta \to \Delta^{(\prod_{v \in O} n_v)-1},$$

where σ denotes the appropriate map marginalizing over hidden variables. $\phi^+(\Theta)$ is thus the collection of all distributions that may arise from the hidden variable model. This collection depends not only on the DAG and designated state spaces of observable variables, but also on the state spaces of hidden variables, even though the sizes of hidden state spaces are not readily apparent from an observable joint distribution.

Since all variables have finite state spaces, the parameter space Θ can be identified with the closure of an open subset of $[0, 1]^L$, for some L. We refer to L as the dimension of the parameter space. In the case of all binary variables, the dimension of Θ is easily seen to be

$$\dim(\Theta) = \sum_{v \in V} 2^{|pa(v)|} = \sum_{k=0}^{\infty} m_k 2^k, \qquad (1)$$

where m_k is the number of nodes in $\mathcal{G}$ with in-degree k.

If a statement is said to hold for *generic parameters* or *generically* then we mean it holds for all parameters in a set of the form $\Theta \smallsetminus E$, where the exceptional set E is a proper algebraic subset of Θ. (Recall an *algebraic subset* is the zero set of a finite collection of polynomials.) As proper algebraic subsets of $\mathbb{R}^n$ are always of Lebesgue measure zero, a statement that holds generically can fail only on a set of measure zero.

As an example of this language, for any DAG model with all variables finite and observable, generic parameters lead to a distribution faithful to the DAG, in the sense that those conditional independence statements implied by d-separation rules will hold, and no others (Meek, 1995). Equivalently, a generic distribution from such a model is faithful to the DAG.

There are several notions of identifiability of parameters of a model; we refer the reader to Allman et al. (2009). If a model has hidden variables an important issue is *label swapping*, by which one can always permute the names of the states of hidden variables, making appropriate changes to the parameters, without changing the joint distribution of the observable variables. Thus for a model with one binary hidden variable, for any generic $\theta_1 \in \Theta$ there is at least one other point $\theta_2 \in \Theta$ with $\phi^+(\theta_1) = \phi^+(\theta_2)$. Note, however, that there are exceptional parameter points which are fixed by the label swapping, and thus are identifiable in a strict sense.

The strongest useful notion of identifiability for models with hidden variables is that for generic $\theta_1 \in \Theta$, if $\phi^+(\theta_1) = \phi(\theta_2)^+$, then θ_1 and θ_2 differ only up to label swapping for hidden variables. This notion is our primary focus in this paper, which we refer to it as *generic identifiability up to label swapping.* In particular, for models with a single binary hidden variable it is equivalent to the parameterization map being generically 2-to-one.

3 Overview of results

In Table 1 of the Appendix, we list each of the binary DAG models considered in this paper, up to the naming of the observable nodes. We number the graphs as A-Bx where $A = |O| = |V| - 1$ is the number of observed variables, $B = |E| - |O|$ is the number of directed edges between the observed variables, and x is a letter appended to distinguish between several graphs with these same features. As the table presents only the case that all variables are binary, the joint distribution lies in a space of dimension $2^A - 1$.

The primary information in this table is in the column for k, indicating the parameterization map is generically k-to-one. In fact, the existence of such a k is not obvious, and does not follow from the behavior of general polynomial maps in real variables, as we now review.

If a single polynomial $p(x)$ in one variable is given, of degree n, then it is well known that the map from $\mathbb{C}$ to $\mathbb{C}$ that it defines will be generically n-to-one. Indeed the equation $p(x) = a$ will be of degree n for each choice of a, and generically will have n distinct roots. This fact generalizes to polynomial maps from $\mathbb{C}^n$ to $\mathbb{C}^m$; there always exists a $k \in \mathbb{N} \cup \{\infty\}$ such that the map is generically k-to-one.

However if $p(x)$ has real coefficients, and is instead viewed as a map from (a subset of) $\mathbb{R}$ to $\mathbb{R}$, it may not have a generic k-to-one behavior. For instance, it is immediately clear from a typical graph of a cubic that there are some sets of positive measure on which it is 3-to-one, and others on which it is one-to-one, as well as an exceptional set of measure zero on which the cubic is 2-to-one. While this exceptional set arises since a polynomial may have repeated roots, the lack of a generic k-to-one behavior is due to passing from considering a complex domain for the function, to a real one.

The fact that the polynomial parameterizations for the models in the table have a generic k-to-one behavior, then, depends on the particular form of the parameterizations. In later sections we prove this essentially one model at a time, while obtaining the value for k. In the case of finite k, we actually go further and characterize the k elements of $\phi^{-1}(\phi(\theta))$ in terms of a generic θ. Of course when $k = 2$ this is nothing more than label swapping, but for the cases of $k = 4$ more is required. Precise statements appear in later sections. In some cases, we also give descriptions of an exceptional subset of Θ where the generic behavior may not hold. In all cases, the reader can deduce such a set from our arguments.

The models 4-3e and 4-3f, for which the parameterization maps are generically 4-to-one, are particularly interesting cases, as for these models there are non-identifiability issues that arise neither from overparameterization (in the sense of a parameter space of larger dimension than the distribution space) nor from label swapping. While these models are ones that can plausibly be imagined as being used for data analysis, they in fact have a rather surprising failure of identifiability, which is explored more precisely in Section 6.3.

In establishing all these results, we first show that we need only consider DAGs up to Markov equivalence. We then consider two special models, 3-0 and 4-3b, whose identifiability we study through certain matrix factorizations. Importantly, the results we obtain for them are not limited to the binary case that we otherwise focus on in this paper. These two models subsequently play a key role in analyzing many of the others we consider.

4 Markov equivalence and parameter identifiability

Two DAGs on the same sets of observable and hidden nodes are said to be *Markov equivalent* if they entail the same conditional independence statements through d-separation. (Note this notion does not distinguish between observable and hidden variables; all are treated as observable.) Thus for fixed choices of state spaces of the variables, two different but Markov equivalent DAGs, $\mathcal{G}_1 \cong \mathcal{G}_2$, have different parameter spaces Θ_1, Θ_2, and different parameterization maps, yet $\phi_1(\Theta_1) = \phi_2(\Theta_2)$.

The relevance of this notion to parameter identifiability is made clear by the following:

Theorem 1. *With all variables having finite state spaces, consider two Markov equivalent DAGs, $\mathcal{G}_1$ and $\mathcal{G}_2$, possibly with hidden nodes. If the parameterization map ϕ_1^+ is generically k-to-one for some $k \in \mathbb{N}$, then ϕ_2^+ is also generically k-to-one.*

In particular if such a model has parameters that are generically identifiable up to label swapping, so does every Markov equivalent model.

This theorem is a consequence of the following:

Lemma 2. *With all variables having finite state spaces, consider two Markov equivalent DAGs, $\mathcal{G}_1$ and $\mathcal{G}_2$, with parameter spaces Θ_i and parameterization maps ϕ_i for the joint distribution of all variables. Then there are generic subsets $S_i \subseteq \Theta_i$ and a rational homeomorphism $\psi : S_1 \to S_2$, with rational inverse, such* *that for all $\theta \in S_1$*

$$\phi_1(\theta) = \phi_2(\psi(\theta)).$$

Proof. Recall that an edge $i \to j$ of a DAG is said to be covered if $\mathrm{pa}(j) = \mathrm{pa}(i) \cup \{i\}$. By Chickering (1995), Markov equivalent DAGs differ by applying a sequence of reversals of covered edges.

We thus first assume the $\mathcal{G}_i$ differ by the reversal of a single covered edge $i \to j$ of $\mathcal{G}_1$. Let $W = \mathrm{pa}_{\mathcal{G}_1}(i) = \mathrm{pa}_{\mathcal{G}_2}(j)$, so $\mathrm{pa}_{\mathcal{G}_1}(j) = W \cup \{i\}$, $\mathrm{pa}_{\mathcal{G}_2}(i) = W \cup \{j\}$. Now any $\theta \in \Theta_1$ is a collection of conditional probabilities $P(X_v|X_{\mathrm{pa}(v)})$, including $P(X_i|W), P(X_j|X_i, W)$. From these, successively define

$$P(X_i, X_j|W) = P(X_j|X_i, W)P(X_i|W),$$
$$P(X_j|W) = \sum_k P(X_i = k, X_j|W),$$
$$P(X_i|X_j, W) = P(X_i, X_j|W)/P(X_j|W).$$

Using these last two conditional probabilities, along with those specified by θ for all $v \neq i, j$, define parameters $\psi(\theta) \in \Theta_2$. Now ψ is defined and continuous on the set S_1 where $P(X_i|W)$ and $P(X_j|X_i, W)$ are strictly positive.

One easily checks that the same construction applied to the edge $j \to i$ in $\mathcal{G}_2$ gives the inverse map.

If $\mathcal{G}_1, \mathcal{G}_2$ differ by a sequence of edge reversals, one defines the S_i as subsets where all parameters related to the reversed edges are strictly positive, and let ψ be the composition of the maps for the individual reversals. □

Proof of Theorem 1. Suppose that ϕ_1^+ is k-to-one when restricted to $S_1' = \Theta_1 \smallsetminus E_1'$, and $S_1 = \Theta_1 \smallsetminus E_1$, $S_2 = \Theta_2 \smallsetminus E_2$ are the sets of Lemma 2, with E_1', E_1, and E_2 proper algebraic subsets. Since ϕ_i^+ is polynomial, we may replace the E_i with the smallest algebraic sets containing $(\phi_i^+)^{-1}(\phi_i^+(E_i))$, which, since k is finite, are also proper subsets. Then using the map ψ of Lemma 2, let $E_2' \subsetneq \Theta_2$ be the smallest algebraic set containing $\psi(E_1' \cap S_1)$. The identity

$$\phi_2^+(\theta) = \phi_1^+(\psi^{-1}(\theta))$$

shows that ϕ_2^+ is k-to-one when restricted to $S_2 \smallsetminus E_2' = \Theta_2 \smallsetminus (E_2 \cup E_2')$. Thus the set on which we have shown ϕ_2^+ to be k-to-one omits only a proper algebraic subset from Θ_2. □

5 Two special models

In this section, we explain how one may explicitly solve for parameter values in the models 3-0 and 4-3b from a joint distribution of the observable variables. We work in more generality than is necessary for the rest of this paper, allowing certain cases of non-binary variables, as the arguments extend easily to these cases.

The generic identifiability up to label swapping of model 3-0 is an instance of a much more general theorem of Kruskal (1977). See also (Stegeman and Sidiropoulos, 2007; Rhodes, 2010). However, Kruskal's theorem does not yield an explicit procedure for recovering parameters. Nonetheless, a more restricted theorem (the essential idea of which is not original to this work, and has been rediscovered several times) does. We include this argument in Theorem 3 below, since it is still not widely known and provides motivation for the approach to the proof of identifiability for model 4-3b. Our analysis of model 4-3b appears to be entirely novel. For both models, we characterize the exceptional parameters for which these procedures fail, giving a precise characterization of the set containing all non-identifiable parameters.

5.1 Special cases of Kruskal's Theorem with explicit solutions

The model we consider corresponds to DAG 3-0 in Table 1, but we allow more general finite state spaces than binary ones.

Parameters for the model are:

1. $\mathbf{p}_0 = P(X_0) \in \Delta^{n_0-1}$, a stochastic vector giving the distribution for the n_0-state hidden variable X_0.
2. For each of $i = 1, 2, 3$, a $n_0 \times n_i$ stochastic matrix $M_i = P(X_i|X_0)$.

We also use the following terminology.

Definition. The *Kruskal row rank* of a matrix M is the maximal number r such that every set of r rows of M is linearly independent.

Note that the Kruskal row rank of a matrix may be less than its rank, which is the maximal r such that *some* set of r rows is independent.

Our special case of Kruskal's Theorem is the following:

Theorem 3. *Consider the model represented by the DAG of model 3-0, where the variable X_i has $n_i \geq 2$ states, with $n_0 = n_1 = n_2 = n$. Then generic parameters of the model are identifiable up to label swapping, and an algebraic procedure for determination of the parameters from the joint probability distribution $P(X_1, X_2, X_3)$ can be given.*

More specifically, if $\mathbf{p}_0$ has no zero entries, M_1, M_2 have full rank, and M_3 has Kruskal rank at least 2, *then the parameters can be found through determination of the roots of certain n-th degree univariate polynomials and solving linear equations. The coefficients of these polynomials and linear systems are rational expressions in the joint distribution.*

Proof. Let $P = P(X_1, X_2, X_3)$ be a probability distribution of observable variables arising from the model, viewed as a $n \times n \times n_3$ array.

Marginalizing P over X_3 (*i.e.*, summing over the 3rd index), we obtain a matrix which, in terms of the unknown parameters, is the matrix product

$$P_{\cdot\cdot+} = P(X_1, X_2) = M_1^T \operatorname{diag}(\mathbf{p}_0) M_2.$$

Similarly, if $M_3 = (m_{ij})$, then the slices of P with third index fixed at i (*i.e.,* the conditional distributions given $X_i = i$, up to normalization) are

$$\begin{aligned} P_{\cdot\cdot i} &= P(X_1, X_2, X_3 = i) \\ &= M_1^T \operatorname{diag}(\mathbf{p}_0) \operatorname{diag}(M_3(\cdot, i)) M_2, \end{aligned}$$

where $M_3(\cdot, i)$ is the ith column of M_3.

Assuming M_1, M_2 are non-singular, and $\mathbf{p}_0$ has no zero entries, $P_{\cdot\cdot+}$ is invertible and we see

$$P_{\cdot\cdot+}^{-1} P_{\cdot\cdot i} = M_2^{-1} \operatorname{diag}(M_3(\cdot, i)) M_2.$$

Thus the entries of the columns of M_3 can be determined (without order) by finding the eigenvalues of the $P_{\cdot\cdot+}^{-1} P_{\cdot\cdot i}$, and the rows of M_2 can be found by computing the corresponding left eigenvectors, normalizing so the entries add to 1. (If M_3 has repeated entries in the ith column, the eigenvectors may not be uniquely determined. However, since the matrices $P_{\cdot\cdot+}^{-1} P_{\cdot\cdot i}$ for various i commute, and M_3 has Kruskal rank 2 or more, the set of these matrices do uniquely determine a collection of simultaneous 1-dimensional eigenspaces. We leave the details to the reader.) This determines M_2 and M_3, up to the simultaneous ordering of their rows.

A similar calculation with $P_{\cdot\cdot i} P_{\cdot\cdot+}^{-1}$ determines M_1, and M_3, up to the row order. Since the rows of M_3 are distinct (because it has Kruskal rank 2), fixing some ordering of them fixes a consistent order of the rows of all of the M_i.

Finally, one determines $\mathbf{p}_0$ from $M_1^{-T} P_{\cdot\cdot+} M_2^{-1} = \operatorname{diag}(\mathbf{p}_0)$.

The hypotheses on the rank and Kruskal rank of the parameter matrices can be expressed through the nonvanishing of minors, so all assumption on parameters used in this procedure can be phrased as the nonvanishing of certain polynomials. As a result, the exceptional set where it cannot be performed is contained in a proper algebraic subset of the parameter set. Since the computations to perform the procedure involve computing eigenvalues and eigenvectors of matrices whose entries are rational in the joint distribution, the second paragraph of the theorem is justified. □

5.2 Another special model

The model we consider next has the DAG of model 4-3b in Table 1, but we again allow more general finite state spaces than binary ones.

Parameters for the model are:

1. $\mathbf{p}_0 = P(X_0) \in \Delta^{n_0 - 1}$, a stochastic vector giving the distribution for the n_0-state hidden variable X_0.
2. Stochastic matrices $M_1 = P(X_1|X_0)$ of size $n_0 \times n_1$; $M_i = P(X_i|X_0, X_1)$ of size $n_0 n_1 \times n_i$ for $i = 2, 3$; and $M_4 = P(X_4|X_0, X_3)$ of size $n_0 n_3 \times n_4$.

Theorem 4. *Consider the model represented by the DAG of model 4-3b, where the variable X_i has $n_i \geq 2$ states, with $n_0 = n_2 = n_4 = n$. Then generic parameters of the model are identifiable up to label swapping, and an algebraic procedure for determination of the parameters from the joint probability distribution $P(X_1, X_2, X_3, X_4)$ can be given.*

More specifically, suppose $\mathbf{p}_0, M_1, M_3$ have no zero entries, the $n \times n$ matrices

$$\begin{aligned} &M_2^i = P(X_2|X_0, X_1 = i),\ 1 \leq i \leq n_1, \text{ and} \\ &M_4^j = P(X_4|X_0, X_3 = j),\ 1 \leq j \leq n_3 \end{aligned}$$

have full rank, and there exists some i, i' with $1 \leq i < i' \leq n_1$ such that for all $1 \leq j < j' < n_3$, $1 \leq k < k' \leq n_4$ the entries of M_3 satisfy inequality (5) below. Then from the resulting joint distribution unique parameters can be found through determination of the roots of certain n-th degree univariate polynomials and solving linear equations. The coefficients of these polynomials and linear systems are rational expressions in the entries of the joint distribution.

Proof. With $P = P(X_1, X_2, X_3, X_4)$ viewed as an $n_1 \times n \times n_3 \times n$ array, we work with $n \times n$ 'slices' of P,

$$P_{i,j} = P(X_1 = i, X_2, X_3 = j, X_4),$$

(*i.e.*, we essentially condition on X_1, X_3, though omit the normalization).

Note that these slices can be expressed as

$$P_{i,j} = (M_2^i)^T D_{i,j} M_4^j, \tag{2}$$

where $D_{i,j} = \operatorname{diag}(P(X_0, X_1 = i, X_3 = j))$ is the diagonal matrix given in terms of parameters by

$$D_{i,j}(k,k) = \mathbf{p}_0(k) M_1(k,i) M_3((k,i), j),$$

and M_2^i and M_4^j are as in the statement of the Theorem.

Equation (2) implies for $1 \le i, i' \le n_1$ and $1 \le j, j' \le n_3$ that

$$P_{i,j}^{-1} P_{i,j'} P_{i',j'}^{-1} P_{i',j} = \\ (M_4^j)^{-1} D_{i,j}^{-1} D_{i,j'} D_{i',j'}^{-1} D_{i',j} M_4^j, \quad (3)$$

and the hypotheses on the parameters imply the needed invertibility. But this shows the rows of M_4^j are left eigenvectors of this product.

In fact, if $i \ne i'$, $j \ne j'$, then the eigenvalues of this product are distinct, for generic parameters. To see this, note the eigenvalues are

$$M_3((k,i),j')M_3((k,i'),j)/(M_3((k,i),j)M_3((k,i'),j')), \quad (4)$$

for $1 \le k \le n$, so distinctness of eigenvalues means for all $1 \le k < k' \le n$

$$M_3((k,i),j')M_3((k,i'),j)M_3((k',i),j)M_3((k',i'),j') \\ \ne M_3((k,i),j)M_3((k,i'),j')M_3((k',i),j')M_3((k',i'),j), \quad (5)$$

and thus a generic choice of M_3 leads to distinct eigenvalues.

With distinct eigenvalues, the eigenvectors are determined up to scaling. But since each row of M_4^j must sum to 1, the rows of M_4^j are therefore determined by P.

The ordering of the rows of the M_4^j has not yet been determined. To do this, first fix an arbitrary ordering of the rows of M_4^1, say, which imposes an arbitrary labeling of the states for X_0. Then using equation (2), from $P_{i,1}(M_4^1)^{-1}$ we can determine $D_{i,1}$ and M_2^i with their rows ordered consistently with M_4^1. For $j \ge 1$, using equation (2) again, from $(M_2^i)^{-T} P_{i,j}$ we can determine $D_{i,j}$ and M_4^j with a consistent row order. Thus M_2 and M_4 are determined.

To determine the remaining parameters, again appealing to equation (2), we can recover the distribution $P(X_0, X_1, X_2)$ using

$$(M_2^i)^{-T} P_{i,j} (M_4^j)^{-1} = \operatorname{diag}(P(X_0, X_1 = i, X_3 = j)).$$

With X_0 no longer hidden, it is straightforward to determine the remaining parameters. □

Remark. In the case of all binary variables, the expression in (4) is just the conditional odds ratio for the observed variables X_1, X_3, conditioned on X_0. The inequality (5) can thus be interpreted as saying there is a non-zero 3-way interaction between the variables X_0, X_1, X_2, which is the generic situation.

6 Small binary DAG models

We now turn to establishing the remaining results in Table 1. All variables are thus assumed binary.

For many of the models *A-Bx* the dimension of the parameter space computed by equation (1) exceeds the dimension $2^A - 1$ of the probability simplex in which the joint distribution of observed variables lies. In all these cases the following Proposition applies to show the parameterization is generically infinite-to-one. We omit its proof for brevity.

Proposition 5. *Let $f : S \to \mathbb{R}^m$ be any map defined by real polynomials, where S is an open subset of $\mathbb{R}^n$ and $n > m$. Then f is generically infinite-to-one.*

This proposition applies to all models in Table 1 with an infinite-to-one parameterization, with the single exception of 4-2a. For that model, amalgamating X_1 and X_2 together, and likewise X_3 and X_4, we obtain a model with two 4-state observed variables that are conditionally independent given a binary hidden variable X_0. One can show that the probability distributions for this model forms an 11-dimensional object, and then a variant of the above proposition applies.

For models 3-0 and 4-3b (and the Markov equivalent 4-3a), specializing the results of the previous section to binary variables yields the claims in the table.

For the remaining models, the strategy is to first marginalize or condition on an observable variable to reduce the model to one already understood. One then attempts to 'lift' results on the reduced model back to the original one.

We consider in detail only some of the models, indicating how the arguments we give can be adapted to others with minor modifications.

6.1 Model 4-1

Since node 2 is a sink, marginalizing over X_2 gives an instance of model 3-0 with the same parameters, after discarding $P(X_2|X_0, X_1)$. Thus generically all parameters except $P(X_2|X_0, X_1)$ are determined, up to label swapping.

But note that if the (unknown) joint distribution of X_0, X_1, X_2, X_3 is written as an 8×2 matrix U, with

$$U((i,j,k),\ell) = P(X_0 = \ell, X_1 = i, X_2 = j, X_3 = k),$$

and $M_4 = P(X_4|X_0)$, then the matrix product UM_4 has entries

$$(UM_4)((i,j,k),\ell) = P(X_1 = i, X_2 = j, X_3 = k, X_4 = \ell),$$

which form the observable joint distribution. Since generically M_4 is invertible, from the observable distribution and each of the already identified label swapping variants of M_4 we can find U. From U we marginalize to obtain $P(X_0, X_1, X_2)$ and $P(X_0, X_1)$. Under the generic condition that $P(X_0), P(X_1|X_0)$ are strictly positive, $P(X_0, X_1)$ is as well, and so we can compute $P(X_2|X_0, X_1) = P(X_0, X_1, X_2)/P(X_0, X_1)$.

Models 4-0 and 4-2d are handled similarly, by marginalizing over a sink.

An alternative argument for model 4-1 and 4-0 proceeds by amalgamating the observed variables, X_1, X_2, into a single 4-state variable, and applying Theorem 3 directly to that model. We leave the details to the reader.

6.2 Models 4-2b,c

The DAGs for these models are Markov equivalent. Thus by Theorem 1, it is enough to consider model 4-2c.

We condition on $X_1 = j$, $j = 1, 2$ to obtain two related models. Letting $X_i^{(j)}$ denote the conditioned variable at node i, the resulting observable distributions are

$$\begin{aligned}P(X_2^{(j)}, X_3^{(j)}, X_4^{(j)}) &= P(X_2, X_3, X_4 \mid X_1 = j)\\ &= P(X_1 = j)^{-1}P(X_1 = j, X_2, X_3, X_4).\end{aligned}$$

With a hidden variable $X_0^{(j)}$ and observed variables $X_2^{(j)}, X_3^{(j)}, X_4^{(j)}$, these distributions arise from a DAG like that of model 3-0. With parameters for the original model $\mathbf{p}_0 = P(X_0)$, 2×2 matrices $M_i = P(X_i|X_0)$ for $i = 1, 4$, and 2×4 matrices $M_i = P(X_i \mid X_0, X_1)$, $i = 2, 3$ and $\mathbf{e}_j$ the standard basis vector, parameters for the conditioned models are:

1. the vector
$$\begin{aligned}\mathbf{p}_0^{(j)} &= P(X_0^{(j)}) = P(X_0|X_1 = j)\\ &= P(X_1 = j)^{-1}P(X_0, X_1 = j)\\ &= \frac{1}{\mathbf{p}_0^T M_1 \mathbf{e}_j}(\mathrm{diag}(\mathbf{p}_0)M_1\mathbf{e}_j),\end{aligned}$$
2. the 2×2 stochastic matrix $M_4^{(i)} = P(X_4^{(i)}|X_0^{(i)}) = M_4$, and
3. for $i = 2, 3$, the 2×2 stochastic matrix $M_i^{(j)} = P(X_i^{(j)}|X_0^{(j)})$, whose rows are the $(0, j)$ and $(1, j)$ rows of M_i.

Now if $\mathbf{p}_0$ and column j of M_1 have non-zero entries, it follows that $\mathbf{p}_0^{(j)}$ has no zero entries. If additionally $M_2^{(j)}, M_3^{(j)}, M_4$ all have rank 2, by Theorem 3 the parameters of these conditioned models are identifiable, up to the labeling of the states of the hidden variable. As these assumptions are generic conditions on the parameters of the original model, we can generically identify the parameters of the conditioned models.

In particular, M_4 can be identified and is invertible. But let U denote the (unknown) 8×2 matrix with $U((i, j, k), \ell) = P(X_0 = \ell, X_1 = i, X_2 = j, X_3 = k)$. Then $P = UM_4$, has as its entries the observable distribution $P(X_1, X_2, X_3, X_4)$. Thus $U = PM_4^{-1}$ can be determined from P. Since U is the distribution of the induced model on X_0, X_1, X_2, X_3 with no hidden variables, it is then straightforward to identify all remaining parameters of the original model.

Thus all parameters are identifiable generically. More specifically, they are identifiable provided that for either $j = 0$ or 1 the three matrices $M_4, M_2^{(j)}, M_3^{(j)}$ have rank 2, and $\mathbf{p}_0$ and the jth column of M_1 have non-zero entries.

6.3 Models 4-3e,f

Due to Markov equivalence, we consider only 4-3e.

By conditioning on $X_1 = j$, $j = 1, 2$ we obtain two models of the form of 3-0. One checks that the induced parameters for these conditioned models are generic. Indeed, in terms of the original parameters they are $P(X_i \mid X_0, X_1 = j)$, $i = 2, 3, 4$, which are generically non-singular since they are simply submatrices of the $P(X_i \mid X_0, X_1)$, and at the hidden node

$$P(X_0 \mid X_1 = j) = \frac{P(X_1 = j \mid X_0)P(X_0)}{\sum_\ell P(X_1 = j \mid X_0 = \ell)P(X_0 = \ell)}$$

which generically has non-zero entries.

Thus for generic parameters on the original model we can determine $P(X_0 \mid X_1 = j)$ and $P(X_i \mid X_0, X_1 = j)$, $i = 2, 3, 4$ up to label swapping. However, we do not have an ordering of the states of X_0 that is consistent for the recovered parameters for the two models. Thus generically we have 4 choices of parameters for the 2 models taken together. Each of these 4 choices leads to a possible joint distribution $P(X_0, X_1)$ (viewing this joint distribution as a matrix, the 4 versions differ only by independently interchanging the two entries in each column, thus keeping the same marginalization $P(X_1)$), and then different parameters $P(X_0)$ and $P(X_1|X_0)$. The matrices $P(X_i|X_0, X_1)$ $i = 2, 3, 4$ are then obtained using the same rows as in $P(X_i \mid X_0, X_1 = j)$, though the ordering of the rows is dependent on the choice made previously.

Having obtained 4 possible parameter choices, it is straightforward to confirm that they all lead to the same joint distribution. Thus the parameterization map is generically 4-to-one.

7 Conclusion

Paraphrasing Pearl (2012), the problem of identifying causal effects in non-parametric models has been "placed to rest" by the proof of completeness of the *do*-calculus and related graphical criteria. In this paper we show that the introduction of modest (parametric) assumptions on the size of the state spaces of variables allows for identifiability of parameters that otherwise would be non-identifiable. Causal effects can be computed from identified parameters, if desired, but our techniques allow for the recovery of all parameters. In the process of proving parameter identifiability for several small networks, we use techniques inspired by a theorem of Kruskal, and other novel approaches. This framework can be applied to other models as well.

We have at least three reasons to extend the work described in this paper. The first is to develop new techniques and to prove new theoretical results for parameter identifiability; this provides the foundation of our work. A second is to reach the stage at which one can easily determine parameter identifiability for DAG models with hidden variables that are used in statistical modeling; this motivates our work. A third and related focus of future work is to address the scalability of our approach and to automate it. We noted above that some of our proofs do not depend on variables being binary. Also, a strategy that we used successfully to handle larger models is to first marginalize or condition on an observable variable to reduce the model to one already understood, and then to 'lift' results on the reduced model back to the original one. We are working towards turning this strategy into an algorithm.

Acknowledgments

The authors thank the American Institute of Mathematics, where this work was begun during a workshop on Parameter Identification in Graphical Models, and continued through AIM's SQuaRE program, and the anonymous reviewers for their comments.

References

Allman, E., C. Matias, and J. Rhodes. 2009. Identifiability of parameters in latent structure models with many observed variables. Ann. Statist. 37:3099–3132.

Chickering, D. M. 1995. A transformational characterization of equivalent Bayesian network structures. Proc. of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI-95), 87–98.

Huang, Y. and M. Valtorta. 2006. Pearl's calculus of intervention is complete. Proc. of the Twenty-second Conference on Uncertainty in Artificial Intelligence (UAI-06), 217–224.

Kruskal, J. 1977. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. Linear Algebra and Appl. 18:95–138.

Kuroki, M. and J. Pearl. 2012. Measurement bias and effect restoration in causal inference, Technical Report R-366. Cognitive Systems Lab, Dept. of Computer Science, Univ. of California at Los Angeles.

Lauritzen, S. L. 1996. Graphical models vol. 17 of *Oxford Statistical Science Series*. Oxford Univ. Press.

Meek, C. 1995. Strong completeness and faithfulness in Bayesian networks. Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI-95), 411–418.

Pearl, J. 1995. Causal diagrams for empirical research. Biometrika 82:669–710.

Pearl, J. 2009. Causality: Models, Reasoning, and Inference. 2 ed. Cambridge University Press.

Pearl, J. 2012. The do-calculus revisited. Proceedings of the Twenty-eighth Conference on Uncertainty in Artificial Intelligence (UAI-12), 4–11.

Rhodes, J. 2010. A concise proof of Kruskal's theorem on tensor decomposition. Linear Algebra and its Applications 432:1818–1824.

Shpitser, I. and J. Pearl. 2008. Complete identification methods for the causal hierarchy. J. Mach. Learn. Res. 9:1941–1979.

Stanghellini, E. and B. Vantaggi. 2013. On the identification of discrete concentration graph models with one hidden binary variable. To appear in Bernoulli, doi: 10.3150/12-BEJ435.

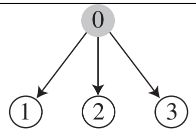
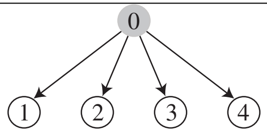
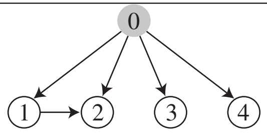
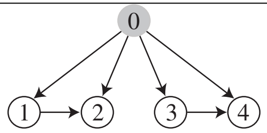
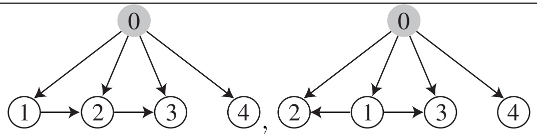
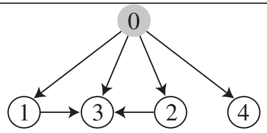
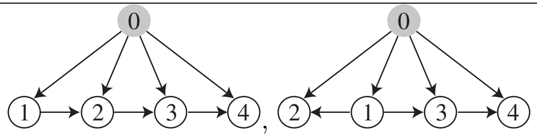
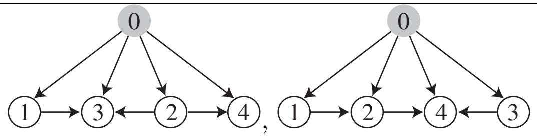
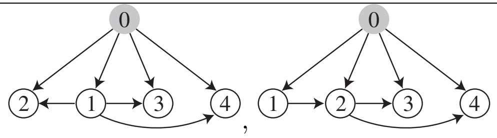
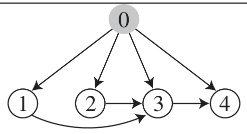
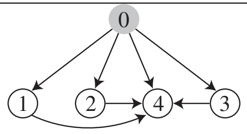
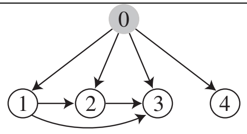
Stegeman, A. and N. D. Sidiropoulos. 2007. On Kruskal's uniqueness condition for the Candecomp/Parafac decomposition. Linear Algebra Appl. 420:540–552.

Tian, J. and J. Pearl. 2002. A general identification condition for causal effects. Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02), 567–573.

Appendix

Table 1 shows all DAGs with 4 or fewer observable nodes and one hidden node that is a parent of all observable ones. See Section 3 for model naming convention. Markov equivalent graphs appear on the same line. The dimension of the parameter space is $\dim(\Theta)$, and $2^A - 1$ is the dimension of the probability simplex in which the joint distribution lies. The parameterization map is generically k-to-one.

Table 1: Small binary DAG models.

Model	Graph	$\dim(\Theta)$	2^A-1	k
2-B, $B \geq 0$		≥ 5	3	∞
3-0		7	7	2
3-Bx, $B \geq 1$		≥ 9	7	∞
4-0		9	15	2
4-1		11	15	2
4-2a		13	15	∞
4-2b,c		13	15	2
4-2d		15	15	2
4-3a,b		15	15	2
4-3c,d		17	15	∞
4-3e,f		15	15	4
4-3g		17	15	∞
4-3h		25	15	∞
4-3i		25	15	∞
4-Bx, $B \geq 4$		≥ 19	15	∞