Recovering from Selection Bias using Marginal Structure in Discrete Models

Robin J. Evans
Department of Statistics
University of Oxford
evans@stats.ox.ac.uk

Vanessa Didelez
School of Mathematics
University of Bristol
Vanessa.Didelez@bristol.ac.uk

Abstract

This paper considers the problem of inferring a discrete joint distribution from a sample subject to selection, such as might arise from a case-control study. Abstractly, we want to identify a distribution $p(x, w)$ from its conditional $p(x \mid w)$. We introduce new assumptions on the marginal model for $p(x)$, under which generic identification is possible. These assumptions are quite general and can easily be tested; they do not require precise background knowledge of $p(x)$ or $p(w)$, such as proportions estimated from previous studies. We particularly consider conditional independence constraints, which often arise from graphical and causal models, although other constraints can also be used. We show that generic identifiability of causal effects is possible in a much wider class of causal models than had previously been known.

1 Introduction

Selection bias occurs when samples are obtained from a population in a manner which depends upon the attributes of the samples themselves. This can happen by accident—such as survey participants self-selecting in a way which is correlated with their responses—or by design, for example in a case-control study.

In this paper we show that under plausible and testable structural hypotheses about the relationships between the variables being measured, we can recover from selection bias, even without having external information about the distribution of variables in the population.

Example 1.1. Consider a case-control study in which participants are selected according to a disease status W with $d_w = 2$ levels. We are interested in the effect of a discrete measured treatment X on W which, assuming no confounding is the same as estimating the conditional distribution $p(w \mid x)$. Since our data are selected according to disease status, what we *observe*, ignoring sampling error, is the conditional distribution $p(x \mid w)$. It is well known that we can use this to obtain the causal odds-ratio for X on W, which is one measure of the causal effect, but that the full conditional distribution $p(w \mid x)$ is generally not identifiable.

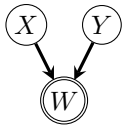


Figure 1: A graph representing a (causal) model implying a marginal independence $X \perp\!\!\!\perp Y$. The variable W has been selected upon, which is represented by the double circle.

However, suppose that we also measure a covariate Y, which is known to be marginally independent of X (see Figure 1). For example if X represents a genetic variant, then Y might be an environmental exposure known to be unrelated to X in the population under study (such assumptions are common in practice, see e.g. [15]). The marginal independence means $p(x, y) = p(x) \cdot p(y)$, so

$$\sum_w p(w)p(x, y \mid w) = \left(\sum_w p(w)p(x \mid w)\right)\left(\sum_w p(w)p(y \mid w)\right) \quad (1)$$

for each value of x, y; letting d_x, d_y be the number of levels of X and Y respectively, this gives at most $(d_x - 1)(d_y - 1)$ non-redundant equations. Since W is binary, each equation is quadratic in the single unknown $p(w = 0)$ and can have at most two solutions. The true marginal distribution of W must be a solution to each equation, so the distribution is identifiable up to at most two solutions[1].

[1]Provided that $Y \not\perp\!\!\!\perp W \mid X$; see the discussion in the

To be concrete, suppose X and Y are binary, and we are given the following conditional distributions $p(x, y \mid w)$:

$W=0$	0	1
0	$\frac{2}{5}$	$\frac{1}{10}$
1	$\frac{1}{10}$	$\frac{2}{5}$

$W=1$	0	1
0	$\frac{1}{5}$	$\frac{3}{10}$
1	$\frac{3}{10}$	$\frac{1}{5}$

Since X and Y are positively correlated given $W = 0$, and negatively correlated given $W = 1$, it is clear that by taking some appropriate convex combination, we should find a table for independence. In fact, choosing $p(w = 0) = \frac{1}{4}$ gives the marginal table:

	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$

Having recovered $p(w = 0)$ we obtain the entire joint distribution $p(x, y, w) = p(w) \cdot p(x, y \mid w)$ and therefore any causal effects identifiable from it, including $p(w \mid x)$.

Note that no information external to the study was required (e.g. background knowledge of the distribution of $p(x)$) other than the marginal independence $X \perp\!\!\!\perp Y$. In addition, we will see that this condition is in principle testable by equality and inequality constraints. In particular (1) may have no solution, so the model can be refuted.

Intuitively the above method works because the independence $X \perp\!\!\!\perp Y$ is 'destroyed' if the marginal distribution of W is specified incorrectly (except in certain degenerate cases). Hence a correct specification of the marginal $p(w)$ results in a distribution in the marginal model for (X, Y), and an incorrect specification usually does not.

1.1 Previous Work and Set-Up

Selection bias has received significant attention in the statistical literature for many decades, dating back at least to [5]. The term is now used to refer to a very wide range of problems: bias can occur due to selection into the sample, drop-out, missingness or due to possibly inadvertently and inappropriately stratifying the statistical analysis. When using DAGs to represent the situation, selection bias can often be illustrated as the result of conditioning on a variable that is a common child of other variables, sometimes called "collider-stratification bias" [12]. In some of these cases, the bias can be avoided e.g. by inverse-probability weighting given a sufficiently informative set of covariates [13]. Further, in the context of a case-control study it is well known that the odds-ratio can be recovered under very mild assumptions, see [11] for a recent application relying on an argument based on conditional independences in DAGs. However, many popular statistical methods aiming to attenuate selection bias rely on specific assumptions about the selection mechanism or external (quantitative) information, see e.g. [6] in the context of meta-analyses. Here, we aim to exploit conditional or marginal independences and the constraints they impose on the model. Our approach is similar in spirit to [20], who use 'distortions' in multivariate Gaussian distributions caused by selection to identify the parameters of the original model.

Recently, a number of papers have considered the specific question of identifying *causal* effects under selection. [8] give conditions for identification of causal odds-ratios as well as the testability of the causal null hypothesis based on assumptions that can be read off DAGs with unobservable variables. [1] extend these results and additionally propose a method of controlling selection bias using instrumental variables to enable recovery of effect measures other than odds-ratios.

Regarding general causal inference under selection, [3, 2] provide results showing the impossibility of full identification of causal effects under selection in a large class of Markovian models. In contrast, we achieve a much wider range of identification results by using the weaker notion of 'generic identifiability' (defined in Section 2). We exploit equality and inequality constraints derived from known conditional or marginal independences. This is closely related to ancestral graph models, which fully characterize the conditional independence constraints implied by selection in DAG models [17]. Similarly, [14] noted such constraints and showed that selection in DAGs could be used to generate arbitrary hierarchical models.

Several of the above authors formalise the problem of selection by including a binary random indicator S of whether or not an individual is selected in the sample. The available data then comes from the conditional distribution $p(x, w|s = 1)$, where X, W may be vectors of random variables. There is little we can do without further assumptions, but suppose that $X \perp\!\!\!\perp S \mid W$, i.e. the selection only depends upon the variables in W; in this case, provided $p(w \mid s = 1) > 0$, we can recover $p(x \mid w) = p(x \mid w, s = 1)$. If the distribution of W in the population is known, say from a previous study, then we can obtain $p(x, w)$ and in effect 'recover' from the selection bias. In this paper we will assume that no such information is available, and ask what assumptions allow $p(x, w)$ to be recovered from $p(x \mid w)$. We

next section.

assume throughout that we observe $p(x \mid w)$ and that $p(x, w) > 0$.

The rest of the paper is structured as follows. After formally defining generic identifiability in Section 2, we address in Section 3 the role of constraints on the marginal model $p(x)$ for recovering $p(x, w)$. In Section 4 we then consider special cases where the marginal constraints arise from various independence assumptions. While we use directed acyclic graphs (DAGs) throughout to depict marginal or conditional independences, Section 5 formally addresses situations where the model is explicitly defined by a DAG, such as is typical for causal models. It turns out that it is an important prerequisite for identification that the model be non-decomposable. Further implications and examples for causal inference are discussed in Section 6, and practical considerations given in Section 7.

2 Identifiability

We distinguish 'strict' (sometimes called 'global') and 'generic' identifiability. The latter allows identifiability to fail on a lower dimensional subset $\mathcal{O}$ of the model $\mathcal{M}$. It can be argued that, under certain assumptions, observations are 'unlikely' to lie exactly in such a subset.

Consider random variables (X, W) taking values in a finite discrete product space $\mathcal{X} \times \mathcal{W}$. Let

$$\Delta = \Delta_{\mathcal{X}\mathcal{W}} \equiv \{\boldsymbol{p} > 0 : \sum_{x,w} p(x, w) = 1\}$$

be the strictly positive probability simplex of distributions over $\mathcal{X} \times \mathcal{W}$. We consider models $\mathcal{M} \subseteq \Delta$ defined by a finite collection of polynomial constraints in the probabilities $p(x, w)$. For example, the model $\mathcal{M}$ under which $X \perp\!\!\!\perp W$ is the set of $\boldsymbol{p} \in \Delta$ such that

$$p(x, w) - \left(\sum_{w'} p(x, w')\right)\left(\sum_{x'} p(x', w)\right) = 0, \quad \forall w, x.$$

Consider the act of conditioning on W, defined by the rational map $\phi : p(x, w) \mapsto p(x, w)/p(w)$. Let $\mathcal{N} = \phi(\mathcal{M})$ be the image of the model under this operation, which is also an irreducible algebraic variety in the invariants $p(x \mid w)$ [7, Proposition 4.5.6].

Define the *fibre* of ϕ at $p \in \mathcal{M}$ as the set

$$F_\phi(p) = \{q \in \mathcal{M} : \phi(q) = \phi(p)\}.$$

An injective map is one for which all fibres have cardinality one.

Definition 2.1. Let $k \in \mathbb{N}$. We say that $\mathcal{M}$ is *generically k-identifiable* if the fibres $F_\phi(p)$ have cardinality at most k for all $p \in \mathcal{M} \setminus \mathcal{O}$, where $\mathcal{O}$ is some proper (i.e. lower dimensional) algebraic subset of $\mathcal{M}$.

In the case $k = 1$ then ϕ is just *generically identifiable*. If a model is not generically k-identified for any $k \in \mathbb{N}$ then it is *unidentifiable*.

In other words, generic k-identifiability requires the map to be at most k-to-one on a set that contains 'almost all' the distributions. Generic identifiability corresponds to the map being injective on $\mathcal{M} \setminus \mathcal{O}$. Throughout this article we use 'generically' to mean 'at all but a strict sub-variety of the parameter values within the model ($\mathcal{M}$)'.

If $\mathcal{O} = \emptyset$ then this is sometimes referred to as *strict (k-)identifiability*. Strict identifiability of $p(x, w)$ is a very stringent condition and essentially never occurs in our framework. If $W \perp\!\!\!\perp X$, for example, it is clear that $p(x \mid w) = p(x)$ does nothing to help us identify $p(w)$, nor therefore $p(x, w)$. We consider the weaker notion of generic identifiability so that identifiability may fail on a small (i.e. measure zero) subset $\mathcal{O}$.

We say a model on X, W is *algebraically testable* if it places non-trivial equality constraints on $p(x \mid w)$, i.e. $\mathcal{N}$ is a strict sub-variety of the set of all conditional distributions $p(x \mid w)$. We will work throughout the paper as though we can observe $p(x \mid w)$ itself; in practice we would instead have a consistent estimator with asymptotically correct standard errors.

3 General Marginal Models

In this section we consider the possibility of using constraints on the marginal model of X to recover $p(x, w)$ from $p(x \mid w)$. The first result gives us a necessary condition for identifiability.

Lemma 3.1. *Suppose that $\mathcal{M}_X$ is a marginal model for X, and let A be the $(d_x \times d_w)$-matrix with entries $a_{x,w} = p(x \mid w)$. Then $p(x, w)$ is k-identifiable from $p(x \mid w)$ for some $k \in \mathbb{N}$ only if A has rank d_w.*

If $W \perp\!\!\!\perp X$ then, as previously noted, the rank condition will not be satisfied. It may fail in a more subtle way if, for example, there is a lower dimensional variable that mediates the relationship between X and W; see Example 6.3.

In general it is difficult to characterize exactly when a model will be generically identifiable, but an important special case comes when W's dependence on X is completely unrestricted: that is, our model makes no restriction on $p(w \mid x)$ no matter what the value of $p(x)$. This variation independence is known as a 'parameter cut', or simply a cut [4].

Theorem 3.2. *Let $\mathcal{M}$ be a model for (X, W) with a*

cut between X and $W|X$, such that $p(w \mid x)$ is unrestricted and the marginal model for $p(x)$ is a variety of dimension $d_x - 1 - l$.

Then given $p(x|w)$, the variety defined by the fibre

$$F(p) \equiv \{q(x,w) \in \mathcal{M} : q(x \mid w) = p(x \mid w)\}$$

generically has dimension $\max(0, d_w - 1 - l)$. *In particular, $p(x,w)$ is generically k-identifiable for some $k \in \mathbb{N}$ if and only if $d_w \leq l + 1$.*

The quantity l is the number of independent constraints on the marginal model for X. We note that Theorem 3.2 implies that *some* constraints on $p(x)$ are needed; when $l = 0$ the inequality in Theorem 3.2 is only satisfied if W is constant, meaning that there is in fact no selection.

Proof. Since $p(x \mid w)$ is known, it is equivalent to consider the set $\alpha(w)$ such that $\alpha(w)p(x \mid w) \in F(p)$. That is, we need $\alpha(w)$ such that $q(x) = \sum_w \alpha(w)p(x \mid w)$ is contained in the marginal model for X; this is precisely those points $q(x)$ which are in both the marginal model for X and the column span of the matrix C with (x,w)th entry $c_{xw} = p(x \mid w)$.

Let $q(x)$ be a point in both the marginal model for X and the column span of C. The tangent space of $\mathcal{M}_X$ at $q(x)$ is generically of dimension $d_x - 1 - l$ by assumption. Proposition A.1 implies that any $l + 1$ columns of C are transverse to $T_q(\mathcal{M}_X)$ (i.e. so that combining them we recover the entire space), so the dimension of the tangent space of the intersection is the number of remaining columns (all of which are linearly independent). Of course, if $d_w < l + 1$ then there are no columns left, so the dimension of the intersection is $\max(0, d_w - l - 1)$.

Now, the fibre of interest is the preimage of this intersection under the linear map C. When $d_w \leq d_x$ the map is generically injective, so the dimension of the preimage under the full rank linear map is also $\max(0, d_w - l - 1)$.

On the other hand, if $d_w > d_x$ then the linear map is generically surjective, so the intersection is just $\mathcal{M}_X$, which has dimension $d_x - l - 1$ and codimension l. The preimage of this set under C also has codimension l, so the dimension of the preimage is $d_w - l - 1$. Note that $d_w < d_x \leq l + 1$, so in this case $d_w - l - 1 > 0$ and we cannot have identifiability. □

It is important to keep in mind that generic identifiability allows for identifiability to fail on interesting sub-models. It may, therefore, be a serious problem if W is (conditionally) independent of some part of X.

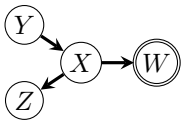


Figure 2: A graphical model in which $p(x, y, z, w)$ is not identifiable under selection on W.

Example 3.3. Consider the graphical model in Figure 2, which implies

$$p(x, y, z, w) = p(y \mid x)p(z \mid x)p(x \mid w)p(w).$$

In this case $Y \perp\!\!\!\perp Z \mid X$ is a marginal constraint, however W does not depend arbitrarily on X, Y, Z so we cannot apply Theorem 3.2. In fact $W \perp\!\!\!\perp Y, Z \mid X$ so

$$p(x, y, z) = p(y \mid x)p(z \mid x) \sum_w p(w)p(x \mid w),$$

which satisfies the required marginal constraint for *any* $p(w)$. We cannot, therefore, use this constraint to identify $p(w)$.

This phenomenon is generalized in the following lemma.

Lemma 3.4. *Suppose distributions in $\mathcal{M}$ are of the form $p(x, y, w) = p(y, w)p(x \mid y, w)$, where $p(x \mid y, w)$ is variation independent of $p(y, w)$. Then $p(x, y, w)$ is identifiable from $p(x, y \mid w)$ if and only if $p(y, w)$ is identifiable from $p(y \mid w)$.*

Proof. We have

$$p(x, y \mid w) = p(x \mid y, w) \cdot p(y \mid w),$$

so the two factors are recoverable from $p(x, y \mid w)$. Variation independence is a graphoid: $p(x \mid y, w)$ is variation independent of $p(y, w)$, so applying the weak union axiom it is also variation independent of $p(w)$ given $p(y \mid w)$ (since these are functions of $p(y, w)$). It follows that no restriction is placed on $p(w)$ by $p(x \mid y, w)$. □

Setting Y to be almost surely constant, the Lemma above reduces to stating that if $p(w)$ and $p(x \mid w)$ are variation independent (i.e. there is a parameter cut between W and $X \mid W$) then $p(x, w)$ is not identifiable from $p(x \mid w)$. In effect no information is shared between $p(w)$ and $p(x \mid w)$, so it is fruitless to try to use one to learn about the other.

In fact our method relies strongly on variation *dependence* between $p(w)$ and $p(x \mid w)$. This is related to the observation of [19] that semi-supervised learning appears to work well in an 'anti-causal' setting (i.e. $W \leftarrow X$) but not in a causal one ($W \rightarrow X$). In

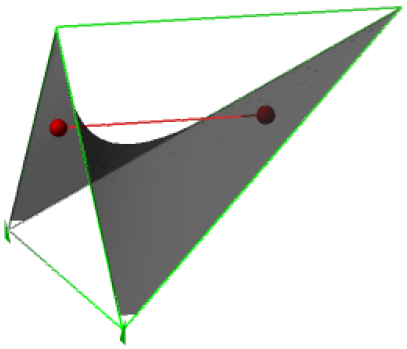


Figure 3: The surface of independence for a 2×2 contingency table in the probability simplex. The plotted line is the set of convex combinations of the two marked points; it intersects the surface twice within the simplex, illustrating the possibility of obtaining two solutions to the all-binary version of equation (1).

semi-supervised learning one effectively uses the margin $p(w)$ to provide information about the conditional $p(x \mid w)$, but if W is a cause of X then we might expect that this 'input' $p(w)$ is unrelated to the causal 'mechanism' $p(x \mid w)$ and hence do not gain anything by doing so.

4 Independences

Marginal and conditional independence constraints often arise in statistical models, commonly from Markov assumptions such as those implied by causal models. In this section we shall focus on when $p(x, w)$ is identifiable in this context.

4.1 Marginal Independence

Consider again Example 1.1 in which $X \perp\!\!\!\perp Y$ but we can only directly observe $p(x, y \mid w)$. Consider each $p(x, y \mid w)$ for $w \in W$ as a single distribution for X, Y; we need to find a convex combination of these distributions which satisfies the independence constraint.

If all the variables are binary we have two 2×2-tables in the 3-dimensional probability simplex, and want to find a point on the line segment between these tables which lies on the surface of marginal independence. This is illustrated in Figure 3. Assuming that neither table lies exactly on the surface, there are three possibilities:

1. the two conditional tables have opposite signed correlations (or log-odds-ratios), in which case there is exactly one convex combination that satisfies marginal independence;

2. the two tables have the same sign, and there is no solution which satisfies the independence;

3. the two tables have the same sign, but there are one or two convex combinations which satisfy the independence.

Situation 2 allows the assumption of $X \perp\!\!\!\perp Y$ to be refuted, and corresponds to an inequality constraint on $p(x, y \mid w)$. This is the only constraint on this model in the binary case.

The possibility of two solutions is due to the quadratic nature of the independence constraint. A line segment which intersects the surface of independence twice is shown in Figure 3. If there are two distinct solutions then there is no way to identify the true distribution uniquely without further assumptions. However, this situation can only occur if the joint distribution exhibits a form of Simpson's paradox: X and Y are positively (or negatively) correlated within each level of W, but are independent after collapsing across the levels of W. It has been argued that this happens relatively rarely in practice [16], and one can see from Figure 4 that it requires the conditional tables to lie in specific corners of the simplex.

The method extends readily to non-binary X and Y; we obtain $(d_x - 1)(d_y - 1)$ separate quadratics in the unknown $p(w = 0)$, and they will generically have distinct solutions; it follows that $p(w = 0)$ is the common solution to these quadratics, and that the model is generically identifiable. If W is non-binary then we obtain quadratic equations in multiple variables, but analogous results are available.

Lemma 4.1. *Consider the model $\mathcal{M}$ on discrete variables X, Y, W such that $X \perp\!\!\!\perp Y$. Then $p(w, x, y)$ is generically k-identifiable from $p(x, y \mid w)$ if and only if $d_w - 1 \leq (d_x - 1)(d_y - 1)$. Identifiability fails if either $X \perp\!\!\!\perp Y, W$ or $Y \perp\!\!\!\perp X, W$.*

If $d_w = 2$ then $p(w, x, y)$ is 2-identifiable from $p(x, w \mid y)$ if $X \not\perp\!\!\!\perp Y \mid W$, and unidentifiable otherwise.

The proof is in the appendix.

Note that, under the causal graphical model in Figure 1, the causal effect of X on W is given by $p(w \mid do(x)) = p(w \mid x)$. This quantity is never strictly identifiable from $p(x \mid w)$, but it is generically identifiable from $p(x, y \mid w)$ under the marginal independence model.

Note that, although we lose identifiability of $p(w)$ if (for example) $X \perp\!\!\!\perp W \mid Y$, we can still test this as a

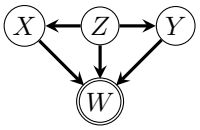


Figure 4: A graph representing a (causal) model implying a conditional independence $X \perp\!\!\!\perp Y \mid Z$.

causal null hypothesis because it is observable directly from $p(x, y \mid w)$.

In the all-binary case, the marginal independence model is not 'algebraically testable' in the sense that it places no equality constraints on $p(x, y \mid w)$: thus for some distributions $p(x, y, w)$ which do not satisfy the marginal independence, we can follow the procedure given above and fallaciously 'recover' some other distribution $q(w)p(x, y \mid w)$ which does satisfy the model. However in the case where $d_x = 3$, there are two independent constraints, so we can use one to recover $p(w)$ and the other to check the model.

4.2 Conditional Independence

Consider the model defined by the graph in Figure 4, which implies that $X \perp\!\!\!\perp Y \mid Z$. Under selection bias on W we can only observe the conditional distribution $p(x, y, z \mid w)$. Following the same approach as in Example 1.1 we obtain

$$p(z) \cdot p(x, y, z) - p(x, z) \cdot p(y, z) = 0, \quad \forall x, y, z;$$

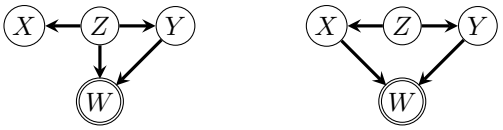
if W is binary we obtain $p(x, y, z) = \alpha p(x, y, z|w = 0) + (1 - \alpha)p(x, y, z|w = 1)$, so each factor is linear in the unknown $\alpha = p(w = 0)$. This gives us quadratic equations in α, which are generically distinct for each level of Z.

Lemma 4.2. *Let $\mathcal{M}$ be the model on X, Y, Z, W defined by $X \perp\!\!\!\perp Y \mid Z$. Then $p(x, y, z, w)$ is generically k-identifiable from $p(x, y, z \mid w)$ if and only if $d_w - 1 \leq (d_x - 1)(d_y - 1)d_z$.*

If $d_w = 2$ the model is unidentifiable if and only if $X \perp\!\!\!\perp Y \mid Z, W$.

The result is really just an application of Lemma 4.1 to separate models for each level of Z. If Z is trivial it reduces to the marginal independence case. Note that the condition $X \perp\!\!\!\perp Y \mid W, Z$ for non-identifiability is testable directly from the observed conditional distribution. The unidentifiability arises because any choice of $p(w)$ will give the required conditional independence.

Particular cases of unidentifiability arise when either $X \perp\!\!\!\perp Y, W \mid Z$ or $Y \perp\!\!\!\perp X, W \mid Z$ (see Figure 5(a), Example 3.3); in fact if W is binary then unidentifiability



(a) (b)

Figure 5: Two graphs representing sub-models of the model in Figure 4. The problem in (a) is not identifiable, but (b) is.

is equivalent to at least one of these constraints holding [9]. In these instances, selecting on W does not destroy the conditional independence we are using to recover the joint distribution: we can observe it directly in the conditional distribution. In a sense, because no structure is lost, we cannot try to recover it by choosing the correct margin $p(w)$.

As the next example illustrates, we can find that extra independences give 'reduced' identifiability by making some constraints redundant, without leading to full unidentifiability. In particular, although the all-binary case of a conditional independence under selection generically gives two non-redundant quadratic equations, in degenerate cases these equations may have the same roots.

Example 4.3. Consider the model in Figure 5(b), in which $Z \perp\!\!\!\perp W \mid X, Y$ as well as $X \perp\!\!\!\perp Y \mid Z$. In this case

$$p(x, y, z \mid w) = p(z \mid x, y) \cdot p(x, y \mid w).$$

In order to have $X \perp\!\!\!\perp Y \mid Z$ we need $p(x, y, z) = p(x, z)p(y \mid z)$. So since $p(x, y, z) = p(z \mid x, y)p(x, y) = p(z \mid x, y) \sum_w p(x, y \mid w)p(w)$, we need to pick $p(w)$ so that the odds-ratio between X and Y exactly cancels out the x, y factor of $p(z \mid x, y)$. This gives a single quadratic equation; and solution to this gives $p(w = 0)$ such that $p(w)p(x, y, z \mid w)$ is inside the model, so we have generic 2-identifiability. Note that although there is only one equation to solve in one unknown, unlike the marginal independence model this one *is* algebraically testable, since $Z \perp\!\!\!\perp W \mid X, Y$ can be checked directly from the observed conditional $p(x, y, z \mid w)$.

The example shows that no three-way interaction is present between X, Y, Z in $p(x, y, z \mid w)$, and that this remains true for any weighted combination $\sum_w \alpha(w)p(x, y, z \mid w)$. If the three-way interaction is zero then the two constraints $X \perp\!\!\!\perp Y \mid Z = 0$ and $X \perp\!\!\!\perp Y \mid Z = 1$ become equivalent, thus we have a partial degeneracy: there is only one non-redundant constraint to fulfil. This issue is explored more generally in Section 5.1.

5 Directed Acyclic Graph Models

In the previous sections, we used directed acyclic graphs (DAGs), also known as Bayesian networks, to represent marginal and conditional independences. Here we give explicit results for situations where the model is defined by a DAG, such as is typical for causal models. A DAG model is defined by a recursive factorization according to the structure of a graph, or equivalently a collection of conditional independence constraints. Let $\mathcal{G}$ be a directed acyclic graph with disjoint sets of vertices $V \dot{\cup} W$ representing vectors of random variables X_V, X_W. We make use of the fairly standard terminology of parents ($\mathrm{pa}_{\mathcal{G}}$), children, ancestors ($\mathrm{an}_{\mathcal{G}}$), etc. See, for example, [17].

We will assume that our distribution obeys the Markov property with respect to a DAG $\mathcal{G}$, so that

$$p(x_V, x_W) = \prod_{v \in V \cup W} p(x_v \,|\, x_{\mathrm{pa}(v)}).$$

No other constraints are imposed.

Our first result notes that for the purposes of recovering from selection on X_W, we can ignore any variables which are not ancestors of W in $\mathcal{G}$.

Lemma 5.1. *Let $p(x_V \,|\, x_W)$ be a conditional distribution from a DAG. Then $p(x_V, x_W)$ is identifiable from $p(x_V \,|\, x_W)$ if and only if $p(x_{\mathrm{an}(W)\setminus W}, x_W)$ is identifiable from $p(x_{\mathrm{an}(W)\setminus W} \,|\, x_W)$.*

Proof. If $V \neq \mathrm{an}_{\mathcal{G}}(W) \setminus W$ then there exists some childless $v \in V$ in $\mathcal{G}$. Hence there is a parameter cut between $W \cup V \setminus \{v\}$ and $\{v\} | W \cup V \setminus \{v\}$, so $p(x_V, x_W) = p(x_v \,|\, x_{V\setminus v}, x_W) \cdot p(x_{V\setminus v}, x_W)$ where $p(v \,|\, x_{V\setminus v}, x_W) = p(x_v \,|\, x_{\mathrm{pa}(v)})$ is variation independent of $p(x_{V\setminus v}, x_W)$. The result then follows from Lemma 3.4 and repeating over all $v \notin \mathrm{an}_{\mathcal{G}}(W)$. □

In light of this result, we henceforth assume that all variables are ancestors of the selection variables W.

5.1 Hierarchical Models

A *hierarchical* model over X_V is the set of distributions $p(x_V)$ which factorize as

$$p(x_V) = \prod_{C \in \mathcal{C}} \phi_C(x_C)$$

for some collection of inclusion maximal sets $\mathcal{C}$; undirected graphical models and decomposable models are special cases. Note that if $p(x_V, x_W)$ is hierarchical with maximal sets $\mathcal{C}$, then $p(x_V \,|\, x_W) = p(x_V, x_W)/p(x_W)$ also factorizes into functions over the maximal sets $\mathcal{C} \cup \{W\}$. Any DAG model is contained in the hierarchical model with maximal sets $\{v\} \cup \mathrm{pa}_{\mathcal{G}}(v)$, which has consequences for our conditional distributions.

Lemma 5.2. *Let $p(x_V, x_W)$ obey the global Markov property for a DAG $\mathcal{G}$ on $V \cup W$. Then $p(x_V \,|\, x_W)$ is of the form*

$$p(x_V \,|\, x_W) = \phi_W(x_W) \prod_{v \in V \cup W} \phi_v(x_v, x_{\mathrm{pa}(v)}). \qquad (2)$$

That is, a hierarchical function with maximal sets W and $\{v\} \cup \mathrm{pa}_{\mathcal{G}}(v), v \in V \cup W$.

This follows directly from writing out the factorization for DAGs.

Note that (2) yields algebraically testable constraints on $p(x_V \,|\, x_W)$, but these constraints do not give any way to identify $p(x_W)$. Clearly the true $p(x_V, x_W)$ must satisfy the same hierarchical model, but so will *any* distribution of the form $q(x_W)p(x_V \,|\, x_W)$.

Proposition 5.3. *Let $\mathcal{M}$ be the model implied by a DAG $\mathcal{G}$, and $\overline{\mathcal{M}}$ the hierarchical model with maximal sets*

$$\mathcal{C} = \left\{ \{v\} \cup \mathrm{pa}_{\mathcal{G}}(v) : v \in V \cup W \right\}.$$

Then $p(x_V, x_W)$ is generically k-identifiable from $p(x_V \,|\, x_W)$ only if $d_w - 1 \leq d(\overline{\mathcal{M}}) - d(\mathcal{M})$.

In particular $p(x_V, x_W)$ is generically k-identifiable from $p(x_V \,|\, x_W)$ only if $\mathcal{G}$ is not *decomposable.*

The proof is given in the appendix. Proposition 5.3 is illustrated by Example 3.3 where the model is indeed decomposable, while in Example 4.3 it is not decomposable. We conjecture that the converse of Proposition 5.3 also holds, so that if $d_w - 1 \leq d(\overline{\mathcal{M}}) - d(\mathcal{M})$ then generic k-identifiability follows. We can obtain the following weaker result from Theorem 3.2, however.

Proposition 5.4. *Let $\mathcal{G}$ be a DAG with vertices $V \cup \{w\}$ such that $\mathrm{pa}_{\mathcal{G}}(w) = V$, and such that $\mathcal{G}$ imposes l independent constraints on X_V. Then $p(x_V, x_w)$ is identifiable from $p(x_V \,|\, x_w)$ if and only if $d_w \leq l + 1$.*

6 Causal Models and Examples

In the context of causal inference, our results will be useful when it is known *a priori* that the causal effect of interest is identified from $p(x, w)$, $p(x)$, or $p(w|x)$. Before quantifying a causal effect, one may first want to test the null hypothesis of no causal effect. As noticed before, when the null translates into a conditional independence (or absence of a directed edge in a DAG), this may actually hamper identification, but we conjecture that in these cases the null can always be checked directly from a factorization of $p(x|w)$ itself; the result from [8] is an example of this.

Example 6.1. It may be plausible that the causal effect of X on W, $p(w|do(x))$, is identified given a sufficient set of covariates C, e.g. by the back-door formula. In the context of a case control study, however, one would then often just estimate the (conditional) causal odds-ratio between X and W given C as the most straightforward causal quantity. Our results would be useful if data on additional covariates Z is available, where it is known that, for example, $X \perp\!\!\!\perp Z \,|\, C$ because subject matter tells us that X is only determined by C. The conditional independence then imposes the sort of constraints which may enable identification of $p(x, c, z, w)$. Hence we can obtain other causal effect measures, such as risk differences or risk ratios. In particular we can obtain the intervention distribution $p(y|do(x))$ by integrating out C, so that unconditional causal parameters can be reported which are more easily compared with results from randomized controlled trials. However, problems may occur: if the association between Z and W is weak, identification becomes unstable, as with a weak instrument in instrumental variable methods. Additionally, if W represents a very rare disease, then the marginal distribution of X is approximately the same as that for the controls, $p(x) \approx p(x|w = 0)$, so recovering $p(x)$ exactly will typically not lead to interesting new insights. However, our method could still be used for sensitivity analysis in such situations.

Example 6.2. Consider the causal model represented by the graph in Figure 6 under selection on X_4. This model arises in the context of dynamic treatment regimes, where X_1, X_3 are treatments, X_2, X_4 outcomes, and X_0 covariates. A typical quantity of interest is $p(x_4 \,|\, do(x_1, x_3))$, which can be expressed as (for example) either of

$$\begin{aligned} p(x_4 \,|\, do(x_1, x_3)) &= \sum_{x_0} p(x_4 \,|\, x_1, x_3, x_0) p(x_0) \\ &= \sum_{x_2} p(x_4 \,|\, x_1, x_2, x_3) p(x_2 \,|\, x_1). \end{aligned}$$

However neither of these quantities is a function of $p(x_{0123} \,|\, x_4)$, so we need to recover the distribution $p(x_4)$ before proceeding. The graph exhibits the conditional independence constraints $X_0 \perp\!\!\!\perp X_1$ and $X_0, X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ and these independences are not visible after selection on X_4, so they provide a method for recovering $p(x_4)$.

The conditional distribution has the form of a hierarchical model

$$p(x_{0123} \,|\, x_4) = \phi(x_0, x_1, x_2)\phi(x_2, x_3)\phi(x_0, x_1, x_3, x_4),$$

and so $q(x_4)p(x_{0123} \,|\, x_4)$ factorizes in the same way for any margin $q(x_4)$. Notice that the constraints which are lost are precisely that $X_0 \perp\!\!\!\perp X_1$ and $X_0 \perp\!\!\!\perp X_3 \,|\, X_2$.

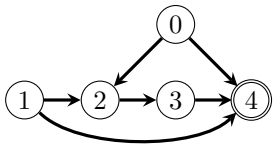


Figure 6: Directed acyclic graph in which certain conditional distributions are identifiable after selecting on X_4.

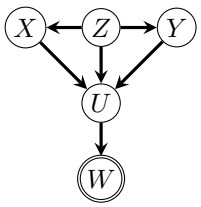


Figure 7: A graph with a 'choke point' for identification.

These constitute $(d_0-1)(d_1-1)$ and $(d_0-1)(d_3-1)d_2$ constraints respectively, but the three-way interaction X_0, X_2, X_3 is not present in the hierarchical model, so in fact there are only $(d_0 - 1)(d_3 - 1)$ *additional* constraints from the conditional independence (this is essentially the same as what happens in Example 4.3). Thus we can generically identify the model only if

$$(d_0 - 1)(d_1 + d_3 - 2) \geq d_4 - 1.$$

In the all-binary case this corresponds to two constraints to resolve one parameter.

Example 6.3. The ability to identify distributions in DAG models may suffer from 'choke points' caused by lower dimensional variables. Consider the model in Figure 7 with selection on W. Then $p(x, y, z, u \,|\, w) = p(x, y, z \,|\, u)p(u \,|\, w)$, so that provided $d_u - 1 \leq (d_x - 1)(d_y - 1)d_z$ we can generically identify $p(u)$ from the first factor. However, knowing $p(u)$ and $p(u \,|\, w)$ will not allow us to generically identify $p(w)$ unless $d_w \leq d_u$. Such a lack of identification is because $p(w \,|\, x, y, z)$ is not unrestricted if $d_w > d_u$, leading to a rank deficiency of the matrix with entries $p(x, y, z \,|\, w)$ and violating the conditions of Lemma 3.1. Note, however, that this rank problem is detectable even if U is unobserved.

7 Discussion

We envisage that the practical application of our results will work as follows. In a given data situation, we would need to know that sampling was subject to selection and which variable(s) this affected. Then, a set of marginal or conditional independences the true

distribution should obey needs to be postulated. Identifiability can then be checked using our results and if it holds the full model can be fitted with appropriate numerical procedures. Note that if these postulated constraints are just enough to enable identifiability they cannot empirically be tested and must be based on subject matter knowledge which will typically be informed by causal assumptions.

In principle it is a relatively straightforward matter to fit these models; to perform maximum likelihood estimation we need to maximize the conditional log-likelihood

$$\begin{aligned} l_{X|W}(p) &\equiv \sum_{x,w} n_{xw} \log p(x \,|\, w) \\ &= \sum_{x,w} n_{xw} \log p(x, w) - \sum_{x,w} n_{xw} \log p(w) \\ &= l_{XW}(p) - l_W(p). \end{aligned}$$

The TM algorithm of [10] deals with this computation by linearizing the marginal log-likelihood $l_W(p)$ at each iteration; this may be useful if the marginal likelihood is hard to calculate but finding its gradient at a single point is not. Naïve numerical methods can also work.

The likelihood behaves much like that of a latent variable model, with some parameter values leading to poor identifiability in finite samples. The stronger the dependence between W and the variable(s) in the marginal model, the better. If identification fails, this will be manifested as slow convergence and a flat log-likelihood. The likelihood may be multi-modal, so trying different starting points would be advisable.

7.1 Testability and Constraints

The all-binary conditional independence model from Section 4.2 requires the existence of a common root to two separate quadratic equations. This corresponds to a single polynomial constraint on the conditional probabilities $p(x, y, z \,|\, w)$, which can be used to test the model. Using the computational algebra software `Singular` we can compute this polynomial, and determine that it is homogeneous of degree 8, but it does not obviously admit a simple interpretation.

Such constraints may be considered analogous to the *Verma constraints* of [18, 21], which arise from marginal distributions of models defined by conditional independence constraints.

7.2 Extensions

All the variables in our causal models were assumed to be observed, but in principle an extension could be made to conditional independence or other constraints which arise in the presence of latent variables. We conjecture that distributions arising from the graph in Figure 6 are recoverable even if X_0 is unobserved, for example (this is the *Verma model* [18, 21]).

Although this paper only considers discrete variables, some of the results here could be extended to the case where the marginal model involves continuous variables, provided the selection variable W is still discrete. In the marginal independence case (Example 1.1), for example, the problem then becomes one of mixing a finite number of densities $f_w(x, y)$ with weights $\alpha(w)$ such that $\sum_w \alpha(w) f_w(x, y)$ factorizes over x and y. This will not be possible for general f_w so, as in the discrete case, we obtain a testable constraint on the model.

Acknowledgements

We thank Stijn Vansteelandt, Ilya Shpitser and others for useful conversations, and an anonymous reviewer for their helpful comments. Most of this work was done while RJE visited Bristol with support from an LMS 'Research in Pairs' grant.

References

[1] E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *Proceedings of AAAI-12*, pages 100–108, 2012.

[2] E. Bareinboim and J. Tian. Recovering causal effects from selection bias. In *Proceedings of AAAI-15*, 2015.

[3] E. Bareinboim, J. Tian, and J. Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of AAAI-14*, 2014.

[4] O. Barndorff-Nielsen. *Information and exponential families in statistical theory.* John Wiley & Sons, 2014.

[5] J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, pages 47–53, 1946.

[6] J. Copas. What works?: selectivity models and meta-analysis. *J. Roy. Statist. Soc. Ser. A*, 162(1):95–109, 1999.

[7] D. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra.* Springer, 2007. Third Edition.

[8] V. Didelez, S. Kreiner, and N. Keiding. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, pages 368–387, 2010.

[9] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics.* Springer Science & Business Media, 2008.

[10] D. Edwards and S. L. Lauritzen. The TM algorithm for maximising a conditional likelihood function. *Biometrika*, 88(4):961–972, 2001.

[11] S. Geneletti, S. Richardson, and N. Best. Adjusting for selection bias in retrospective, case–control studies. *Biostatistics*, 10(1):17–31, 2009.

[12] S. Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, 2003.

[13] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

[14] S. L. Lauritzen. Generating mixed hierarchical interaction models by selection. Unpublished tech report., 1999.

[15] B. Moerkerke, S. Vansteelandt, and C. Lange. A doubly robust test for gene–environment interaction in family-based studies of affected offspring. *Biostatistics*, pages 213–225, 2010.

[16] M. G. Pavlides and M. D. Perlman. How likely is Simpson's paradox? *The American Statistician*, 63(3), 2009.

[17] T. S. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.

[18] J. M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.

[19] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. Semi-supervised learning in causal and anticausal settings. In *Empirical Inference*, pages 129–141. Springer, 2013.

[20] E. Stanghellini and N. Wermuth. On the identification of path analysis models with one hidden variable. *Biometrika*, 92(2):337–350, 2005.

[21] T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence (UAI-90)*, pages 255–270, 1990.

A Proofs

Proof of Lemma 4.1. The first part follows from Theorem 3.2 with $l = (d_x - 1)(d_y - 1)$, and the failure under the additional independence from Lemma 3.4.

Suppose $d_w = 2$, write $\alpha = p(w = 0)$. Marginal independence entails that $p(x, y) - p(x)p(y) = 0$ for each x, y, so using

$$p(x, y) = \alpha p(x, y \,|\, w) + (1 - \alpha)p(x, y \,|\, \bar{w})$$

(here w is used as an abbreviation for $w = 0$, and $\bar{w}$ for $w = 1$) leads to a quadratic equation $a_{xy}\alpha^2 + b_{xy}\alpha + c_{xy} = 0$ with coefficients

$$\begin{aligned} a_{xy} &\equiv (p(x \,|\, w) - p(x \,|\, \bar{w}))(p(y \,|\, w) - p(y \,|\, \bar{w})) \\ b_{xy} &\equiv -a_{xy} - c_{xy} - p(x, y|w) + p(x|w)p(y|w) \\ c_{xy} &\equiv p(x, y|\bar{w}) - p(x|\bar{w})p(y|\bar{w}). \end{aligned}$$

We have complete unidentifiability if and only if $a_{xy} = b_{xy} = c_{xy} = 0$ for all x, y. From the forms of c_{xy} and b_{xy} it is clear that this occurs only if $X \perp\!\!\!\perp Y \,|\, W$. Conversely, if $X \perp\!\!\!\perp Y \,|\, W$ and $X \perp\!\!\!\perp Y$ then for binary W this implies that either $X \perp\!\!\!\perp W, Y$ or $Y \perp\!\!\!\perp W, X$ [9], so identifiability fails if and only if $X \perp\!\!\!\perp Y \,|\, W$. □

Proof of Proposition 5.3. We prove the case $W = \{w\}$, from which the main result is a fairly easy extension.

The map $\phi : p(x_V, x_w) \mapsto p(x_V \,|\, x_w)$ maps $\mathcal{M}$ into the $d(\overline{\mathcal{M}}) - d_w + 1$ dimensional space of hierarchical models after conditioning on X_w; therefore the image of $\mathcal{M} \subset \overline{\mathcal{M}}$ also lies in a $d(\overline{\mathcal{M}}) - d_w + 1$ dimensional space. It follows that generic fibres of ϕ when applied only to $\mathcal{M}$ must have dimension at least $d(\mathcal{M}) - d(\overline{\mathcal{M}}) + d_w - 1$, and in particular have positive dimension if $d(\overline{\mathcal{M}}) - d(\mathcal{M}) < d_w - 1$; it follows that $d(\overline{\mathcal{M}}) - d(\mathcal{M}) \geq d_w - 1$ is necessary for generic identifiability.

In particular, if $\mathcal{G}$ is decomposable then $\mathcal{M} = \overline{\mathcal{M}}$, so there is no identifiability for any $d_w \geq 2$. □

A.1 Generic Identifiability

Proof of Lemma 3.1. Suppose we find $p(w)$ such that $p(x) = \sum_w p(w) \cdot p(x \,|\, w)$ is contained in $\mathcal{M}_X$. This is simply a sum over the columns of A, so if it is not of full column rank then we can clearly add any vector in the kernel of A to $p(w)$ and obtain the same distribution. Hence $p(w)$ is not identifiable. □

Let V be a vector space, and A, B two subspaces of V. We say A and B are *transverse* if $A + B = V$.

Proposition A.1. *Let $\mathcal{M}$ be a model for (X, W) in which there is a parameter cut between X and $W|X$, and such that $p(w \,|\, x)$ is unrestricted. The matrix C with entries $c_{xw} = p(x \,|\, w)$ generically has full rank $r = \min(d_x, d_w)$.*

In addition, let C_r be any $r \leq d_w$ columns of C, and E any $d_x - r$ dimensional linear space. Generically, C_r and E are transverse.

Proof. Fix $p(x) \in \mathcal{M}_X$. Since $p(w \,|\, x)$ is unrestricted, the matrix A with entries $a_{xw} = p(w \,|\, x)$ satisfies this conditions asserted for C.

Now let C' be the matrix with entries $c'_{xw} = a_{xw} \cdot p(x) = c_{xw} \cdot p(w)$. This clearly has the same column span as C. Since C' is obtained is just DA, where D is the non-singular diagonal matrix with entries $p(x)$. Any restriction of (any subset of) the columns of C' to a subspace would imply a similar restriction on A, so by contradiction no such restriction exists. □