Graphical Models and Structure Learning

Karhan Kaan Kayan

September 15, 2022

Abstract

This note gives an overview of the basic concepts behind Bayesian network structure learning, with a focus on identifiability of graphical structures from observational data. The objective is to provide the reader with a rapid introduction to identifiability, graphical models, and structure learning. We review the Markov properties and their relationships, minimal I-maps, and faithfulness. From here, we discuss the relationship between existing results on identifiability and minimality, and illustrate how these results can be viewed as identifying minimal I-maps under different modeling assumptions.

Note. This is an early draft that has not been publicly shared. Please do not distribute it or share it with anyone else. Please report any typos or corrections to `bryon@chicagobooth.edu`.

Notes for students

There are a great number of textbooks on graphical models and causality. Here are the ones I have found the most useful and find myself going back to most often:

(TB1) Koller and Friedman (2009): Foundations and basics

(TB2) Lauritzen (1996): General theory of graphical models

(TB3) Studený (2006): Abstract theory of conditional independence

(TB4) Pearl (2009); Spirtes et al. (2000): Causal interpretation of DAGs

(TB5) Berzuini et al. (2012): Different approaches to causal inference

Roughly speaking, there are two tracks:

- Structure learning: (TB1) ⟶ (TB2) ⟶ (TB3)
- Causality: (TB1) ⟶ (TB4) ⟶ (TB5)

The purpose of this note is to get the reader up to speed as quickly as possible, with the expectation that additional self-study from one of the above tracks will follow.

Prerequisites This note assumes familiarity with some basic concepts from probability and graph theory. To get up to speed, I recommend the following resources:

- Graph theory: Koller and Friedman (2009); Lauritzen (1996) (note that most traditional mathematical textbooks are not appropriate for our purposes)
- Probability theory: Jacod and Protter (2012); Pollard (2002)

Of particular importance is the concept of conditional independence; see Koller and Friedman (2009); Lauritzen (1996) for a refresher.

> *Remark* 1. If you don't know what a "measure" is, don't worry! Just replace the word "measure" with "distribution" everywhere and nothing changes. Our use of measure-theoretic terminology is simply to emphasize the generality of these ideas.

Source material Our convention in this note is to cite the most general source (known to the authors) for a given result, in order to avoid presenting special cases and making unnecessary assumptions. For more instructive resources, each section is marked by one of the textbooks above for self-study.

1 Introduction

In structure learning, we seek to learn a graphical model G given samples from a distribution P. In this note, we consider the special case where G is a directed acyclic graph (DAG), also called a Bayesian network (BN) or structural causal model (SCM). Formally, the structure learning problem is defined as follows:

(SL) Given n i.i.d. samples $X^{(1)}, \dots, X^{(n)} \overset{\text{iid}}{\sim} P$, learn the DAG G that generates P.

Intuitively, we think of G representing some generative or *causal* process that generates the data, and we wish to recover the structure of this process given i.i.d. samples from P. In order for (SL) to make sense, there must be a way to assign a unique DAG G to P. This is known as *identifiability*, and begs two questions:

(Q1) How to assign *any* (i.e. not necessarily unique) DAG G to a general distribution P?

(Q2) When is G uniquely defined from P?

Our goal in this note is to distill a minimal set of results that are necessary to make sense of (SL) at the most basic level.

Remark 2. This note summarizes well-known facts from the graphical modeling literature. The reader interested in tracing the history of these ideas is referred to one of the many textbooks on the subject, such as Pearl (1988); Lauritzen (1996); Studený (2006); Koller and Friedman (2009); Maathuis et al. (2018). For a causal perspective, see Pearl (2009); Spirtes et al. (2000); Berzuini et al. (2012); Peters et al. (2017). Wherever possible, we try to include references for certain results, but make no claim to be exhaustive in doing so.

1.1 Motivation

Our goal is to provide a distribution-free, model-agnostic, and nonparametric approach to structure learning. To accomplish this goal, we will review a single conceptual framework that subsumes now standard models such as nonlinear additive noise models, equal variance models, and linear models (including both LinGAM and gaussian models as special cases).

Unfortunately, the model-free approach complicates matters. To see this, consider arguably the simplest parametric model, the linear gaussian model, which can be written as follows:

$$X = W^T X + z, \quad W = (w_{ij}) \in \mathbb{R}^{d \times d}, \quad z \sim \mathcal{N}(0, \Omega). \tag{1}$$

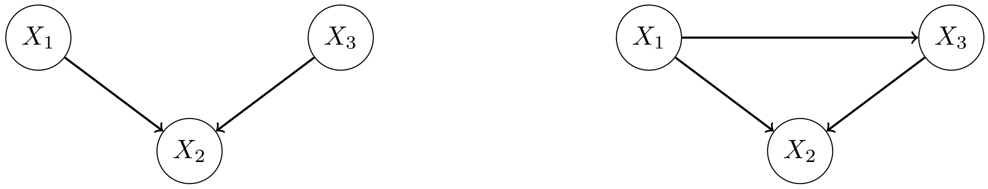
We often assume that Ω is a diagonal matrix, although this is not strictly necessary. Then we can associate to X a graph G by reading off the nonzero entries of the matrix W: Simply define $G = (V, E)$ by

$$V = X, \quad E = \{(i, j) : w_{ij} \neq 0\}. \tag{2}$$

It turns out that as long as G is acyclic, then G is a Bayesian network (i.e. a DAG) for X. In other words, the matrix W explicitly encodes a Bayesian network for X.[1] Evidently, the parametrization (1) makes it easy to find G. But what if we do not have a parametrization such as (1)? How do we define G for general, nonparametric P? This is the content of (Q1) and (Q2), which are prerequisites to answering the main question (SL). We will review some foundational concepts to answer these questions before showing how existing approaches fit under this umbrella.

Remark 3. The assumption that G is acyclic is not necessary, and several extensions of these ideas to *cyclic* graphs are available. In the interest of brevity, we do not go into these very interesting extensions here.

[1] Although we have not yet formally defined Bayesian networks, I-maps, etc., the claims in this discussion should simply be taken for granted at this stage. The reader is suggested to re-visit this discussion after the appropriate definitions are made in Sections 2-3.



(a) A v-structure with an unshielded collider.

(b) A general v-structure.

1.2 Preliminaries

We begin by recalling some conventional notation: We write $[d] = \{1, \dots, d\}$ and $G = (V, E)$ for a graph over the vertices V. Typically, there will be a d-dimensional random vector $X = (X_1, \dots, X_d)$, and the nodes of G correspond to these random variables, i.e. $V = X$. We will frequently abuse notation by using indices j interchangeably with the corresponding random variable X_j and by writing $V = X = [d]$.

The general setting will be the following: We have an abstract probability space $(\mathbb{P}, \Omega, \mathcal{A})$, a random vector $X = (X_1, \dots, X_d)$ whose joint probability measure is denoted by $P = \mathbb{P}(X_1, \dots, X_d)$, and a DAG $G = (V, E)$ whose vertex set V is identified with the random vector X. We will also abuse notation by conflating P with $\mathbb{P}(X)$ and X, as long as no confusion may arise.

In order to avoid measure-theoretic technicalities with conditional measures, we make the following assumption throughout:

(A1) P has a density with respect to some base measure.

When we write expression such as $P(X_j \mid \mathrm{pa}(j))$, it is to be understood that such references are to conditional densities (see Remark 4 below). This abuse of notation should cause no confusion. The following stronger assumption will be used from time to time, but *do not assume from the outset*:

(A2) P has a strictly positive density with respect to some base measure.

The choice of base measure in (A1) and (A2) is typically unimportant and will be ignored.

> *Remark* 4. The purpose of (A1) is to ensure that conditional densities are well-defined. More generally, these can be replaced by regular conditional probabilities. In general, (A1) can be replaced by any regularity condition that ensures the existence of conditional measures, for example, when X is a Polish space and we use regular conditional probabilities. See Rao (2005) for (gory) details.

1.3 Graph notation

A directed *graph* is a pair $G = (V, E)$ where V is a set of vertices (nodes) and $E \subseteq \{(x, y) : (x, y) \in V^2, x \neq y\}$ is a set of edges between the vertices. We sometimes denote the existence of an edge $(x, y) \in E$ by $x \to y$. A *subgraph* of G is formed by taking a subset $A \subset V$, ignoring the vertices outside of A, and removing any edges whose vertices are not in A; i.e. $G[A] = (A, E[A])$ where $E[A] = \{(x, y) \in E : x, y \in A\}$. Given a directed graph G, we call the undirected graph that is obtained by replacing each directed edge with an undirected edge the *skeleton* of G. A *topological ordering* of a (directed) graph G is an ordering of nodes $X_1, \dots, X_d$ such that $X_i \to X_j$ holds only when $i < j$. We call a sequence of nodes $X_1, \dots, X_k$ a *directed path* if $X_i \to X_{i+1}$ for $i = 1, \dots, k-1$. Without qualification, a *path* is a sequence of nodes $X_1, \dots, X_k$ such that $X_i \to X_{i+1}$ or $X_i \leftarrow X_{i+1}$ for $i = 1, \dots, k-1$; i.e. orientation along the path is ignored. A directed path $X_1, \dots, X_k$ is called a *cycle* if $X_1 = X_k$. A directed graph is a *directed acyclic graph* (DAG) if it does not contain any cycles.

If $X \to Y$ is an edge, then Y is called the *child* of X; the node X is called the *parent* of Y, and X and Y are said to be *adjacent* to each other. The set of parents of a node X is denoted by $\mathrm{pa}(X)$. We say that Y is a *descendant* of X, and X is an *ancestor* of Y if there exists a directed path from X to Y. If Y is a not a descendant of X (i.e. there is no directed path from X to Y), we call Y a *non-descendant* of X and denote the set of non-descendants of X by $\mathrm{nd}(X)$. A node X is called a *source* if $\mathrm{pa}(X) = \emptyset$ and a *sink* if $\mathrm{ch}(X) = \emptyset$. The set of all source nodes in a graph G is denoted by $\mathrm{src}(G)$. A graph is *complete* if for any pair of nodes X, Y, we have either $X \to Y$ or $Y \to X$. A subgraph is called a *v-structure* if it is of the form $X \to Y \leftarrow Z$, and we call the child node Y a *collider* (see Figure 1b). We say that a collier is *unshielded* if the parents of the collider do not share an edge (see Figure 1a).

1.4 Conditional independence

We assume the reader is familiar with the concept of conditional independence; see Koller and Friedman (2009); Lauritzen (1996) for a refresher. Given disjoint sets of variables $A, B, C \subset V$, we write $X_A \perp\!\!\!\perp X_B \mid X_C$ to indicate that X_A is conditionally independent (CI) of X_B given X_C. This will often be shortened simply to $A \perp\!\!\!\perp B \mid C$, which is consistent with our convention to abuse notation by conflating indices with random variables. The relation $A \perp\!\!\!\perp B \mid C$ is a ternary relation on subsets of V called a *CI relation*. An *elementary CI relation* is any CI relation such that A and B are singletons.

1.5 The punchline

To give a peek at what's to come, we preview the punchline here. When we say that G is uniquely determined by P, we mean the following:

> *There exists exactly one DAG G that (a) is a minimal I-map of P and (b) satisfies a given additional property (i.e. based on the application).*

See Section 4.1 for more details. The property in (b) could be equal variances, additive noise, post-nonlinearity, a linear non-gaussian model, etc. This is discussed in detail in Section 4, after developing the necessary background to define a minimal I-map in Sections 2-3.

2 Bayesian networks

Self-study: (TB1) §3

In this section and the next, we seek to address (Q1) by formally defining a Bayesian network. There are several ways to define a Bayesian network, not all of which are equivalent in general. We choose a definition that holds in general, i.e. without needing to make additional assumptions on P beyond (A1), and which does require additional definitions up front:

Definition 1 (Bayesian network)

A DAG G is called a *Bayesian network* for P if

$$P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i \mid \mathrm{pa}(X_i)). \tag{BN}$$

When (BN) holds for the pair (G, P), we will say that (G, P) is a Bayesian network, or BN for short.

The property (BN) is also known as the *recursive factorization* of P.

Example 1 (Linear models)**.** Recall the linear model (1) and let P denote as usual the joint distribution of the X_j. For simplicity, let us assume that $\Omega = \mathrm{cov}(z)$ is diagonal and that $z_j \perp\!\!\!\perp \mathrm{pa}(X_j)$ for each j. As long as the associated graph G (cf. (2)) is acyclic, one can check that G satisfies (BN) for P.

A key consequence of Definition 1—of particular importance in the context of structure learning—is that a BN *need not be uniquely defined*. More formally, we would say that G is not *identifiable* from P. Why should we care about this? Recalling the main problem (SL), we need to be careful: (SL) asks for *the* BN G that generates P, but according to (BN), this graph is not even well-defined!

First, let us illustrate this lack of uniqueness in a concrete way:

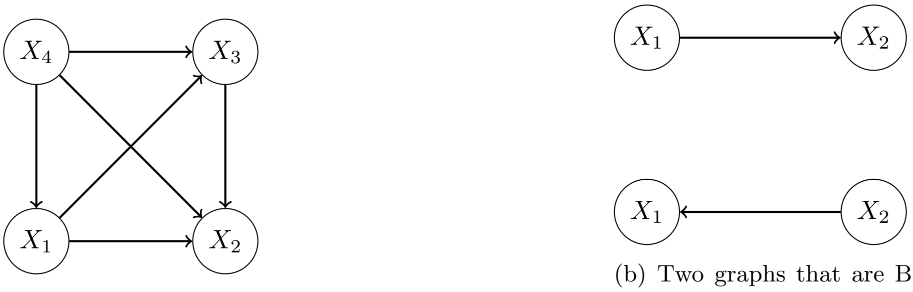
Example 2 (Edge flipping)**.** We consider two variables with distribution $P(X_1, X_2)$. We have

$$P(X_1, X_2) = P(X_1)P(X_2 \mid X_1).$$

Similarly, by the definition of $P(X_1 \mid X_2)$, we also have

$$P(X_1, X_2) = P(X_2)P(X_1 \mid X_2).$$

These to equations show that P factorizes with respect to both the graph $X_1 \to X_2$ and the graph $X_2 \to X_1$ shown in figure 2b. Therefore, both graphs are Bayesian networks for $P(X_1, X_2)$. Thus, G is not identifiable from P.



(a) A complete DAG on 4 nodes

(b) Two graphs that are Bayesian networks for $P(X_1, X_2)$.

Example 3 (Complete graph)**.** Consider a complete DAG G on d variables. We claim that G is a Bayesian network for any distribution $P(X_1, \dots, X_d)$. To see this, recall the chain product rule:

$$P(X_1, \dots, X_d) = P(X_1)P(X_2 \,|\, X_1)P(X_3 \,|\, X_2, X_1) \dots P(X_d \,|\, X_{d-1} \dots X_1).$$

Thus if we let G be the DAG such that $X_k \to X_i$ for $k < i$ and $X_k \leftarrow X_i$ for $k > i$, P factorizes with respect to G. Furthermore, there is nothing special about this choice of G: We can re-order the variables however we like, orient edges from ancestors to children in this ordering, and the resulting complete DAG will always factor P. Thus, there are at least $d!$ BNs for any distribution P (one for each order of the variables).

Evidently, before we can even discuss how to solve (SL), we need to formulate the problem in such a way that it is well-defined—namely, such that we can assign to P a *unique* DAG G. We undertake this program in several steps: A crucial concept is the Markov property, which is introduced in Section 3, along with the concept of a minimal I-map. Minimality is a crucial property for any reasonable solution to (SL) to have. Unfortunately, minimal I-maps are also not unique, so we complete this program in Section 4 by discussing various approaches to identifying minimal I-maps.

3 Markov properties and minimal I-maps

Self-study: (TB1) §3.2.3 $\longrightarrow$ (TB2) §3.2.2

One drawback of the definition (BN) is that it allows for too much redundancy. For example in Example 3, we might expect that we could eliminate some edges in G while still satisfying (BN). In general, this is possible, but when this is *not* possible, we call G a *minimal I-map*. The purpose of this section is to explain how this idea is intimately related to the conditional independence structure of P through what are known as the *Markov properties*.

The basic idea behind the Markov properties is the following: Both distributions P and graphs G naturally define their own ternary relations:

- In a distribution P, we can talk about conditional independence $A \perp\!\!\!\perp B \,|\, C$ for subsets $A, B, C \subset V$;
- In a graph G, we can talk about separation: A subset $A \subset V$ is separated from another subset $B \subset V$ by a third subset $C \subset V$ if every directed path between A and B "contains" a vertex in C.

We put "contains" in quotes to emphasize that different versions of separation for different graphs and assumptions exist in the literature, and each lead to their own ternary relation. The Markov properties simply assert that conditional independence and graphical separation are related, even though each relation makes perfect sense in the absence of the other. In other words, the distribution "respects" the graph in some sense and vice versa.

3.1 Markov properties

Every distribution P defines a collection of conditional independence (CI) relations, which we denote by $\mathcal{I}(P)$. Formally, we define

$$\mathcal{I}(P) = \{\langle A, B, C\rangle : A \perp\!\!\!\perp B \,|\, C \text{ in } P\}. \tag{3}$$

It is implicit in the above that A, B, C are disjoint subsets of X.

The usefulness of a graphical model is that, as we shall see, the graph G encodes (possibly a subset of) these CI relations in a compact way. Let us see one example of this:

Example 4 (Independent Bernoulli trials)**.** Consider the joint distribution of d independent Bernoulli trials $X_1, \ldots, X_d$. Then these variables encode $\binom{d}{2}2^{d-2}$ total elementary CI relations. However, all of these relations can be represented with a graph on d nodes that has no edges. Furthermore, note that we would naively parametrize this distribution using 2^{d-1} parameters (one for each outcome excluding the last one). Yet, using (BN), we would have the compact factorization

$$P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i),$$

which needs only d parameters (one for each probability of success).

3.1.1 Local Markov property

One way a graph can encode CI relations is the local Markov property:

Definition 2 (Local Markov property)

A distribution P satisfies the *local Markov property* with respect to a DAG G if

$$X_i \perp\!\!\!\perp \mathrm{nd}(X_i) \mid \mathrm{pa}(X_i) \tag{DL}$$

holds for each $i \in [d]$.

Theorem 1 (Theorem 1, Lauritzen et al., 1990)

Under (A1), (BN) $\iff$ (DL).

Any directed graph $G = (V, E)$ encodes a ternary relation by

$$\mathcal{I}_{\mathrm{loc}}(G) = \{\langle X_i, \mathrm{nd}(X_i), \mathrm{pa}(X_i)\rangle \mid X_i \in V\}, \tag{4}$$

hence the local Markov property may be equivalently written as

$$\mathcal{I}_{\mathrm{loc}}(G) \subset \mathcal{I}(P). \tag{5}$$

By Theorem 1, we deduce the following: If (G, P) is a BN, then (5) holds. This partially motivates our interest in Bayesian networks and structure learning in particular: If we can learn a BN for P from data, then we immediately infer multiple conditional (in)dependence statements about P.

> *Remark* 5. If our interest is in CI relations, one might wonder why we do not simply run a CI test to test a relation such as $A \perp\!\!\!\perp B \mid C$. One reason is that we may not be interested in testing a *particular* CI relation, but instead learning *some* (or all) of the CI relations present in P. Although we could in principle test all $O(4^d)$ possible CI relations, in practice it is desirable to find more efficient algorithms for this.

3.1.2 Global Markov property

The set $\mathcal{I}_{\mathrm{loc}}(G)$ constitutes only a small subset of all the relations one could read off the DAG G. To get a more comprehensive list, we use the concept of d-separation, which is a directed version of separation that is useful in directed acyclic graphs. In an undirected graph, separation is easy to define: A and B are separated by C if $G[(A \cup B) - C]$ is disconnected, i.e. removing the vertices in C eliminates all paths from A to B. This defines a ternary relation on any undirected graph. Think of d-separation as a generalization of separation in undirected graphs where edge orientation matters. We postpone further details for the time being (see Appendix A). For now, let us simply note that d-separation allows us to define a *global* Markov property as follows:

Definition 3 (Global Markov property)

A distribution P satisfies the *global Markov property* with respect to a DAG G if

$$A \perp\!\!\!\perp B \mid C \tag{DG}$$

holds whenever A and B are d-separated by C.

Write

$$\text{d-sep}_G(AB \mid C) \iff A \text{ and } B \text{ are } d\text{-separated by } C, \tag{6}$$

so a graph G encodes a set of global relations by

$$\mathcal{I}(G) = \{\langle A, B, C\rangle \mid \text{d-sep}_G(AB \mid C)\}. \tag{7}$$

In analogy with (5), we can rewrite the global Markov property as:

$$\mathcal{I}(G) \subseteq \mathcal{I}(P). \tag{8}$$

Furthermore, Theorem 1 can be extended to include the global Markov property (indeed, this is how this result is typically presented):

Theorem 2 (Factorization Theorem; Theorem 1, Lauritzen et al., 1990)

Under (A1), (BN) $\iff$ (DL) $\iff$ (DG).

The factorization theorem establishes a deep connection between the definition of a Bayesian network G and its (global and local) separation properties and shows that they are, in fact, equivalent. This means that we can define a Bayesian network by any of those assumptions, and the others will follow.

Remark 6. The assumption (A1) is really only needed to make sense of the conditional probabilities that appear in the factorization (BN). In fact, the equivalence (DL) $\iff$ (DG) holds without (A1), and there is an appropriate generalization of (BN) to distributions without densities; see §6 of Lauritzen et al. (1990) for details.

We give DAGs that satisfy (DG), or equivalently (8), a special name:

Definition 4 (I-map)

If $\mathcal{I}(G) \subseteq \mathcal{I}(P)$ then G is called an *I-map* (short for *independence map*) of P.

In other words, an I-map is a graph that does not "lie" about the independence structure of a distribution. But it may not tell us everything.

3.1.3 Perfect maps and faithfulness

Although an I-map does not lie about the independencies in the distribution it represents, it is not guaranteed to capture all of the independencies. In the case that it does, we call it a *perfect map*.

Definition 5 (Perfect map)

If $\mathcal{I}(G) = \mathcal{I}(P)$ then G is called a *perfect map* of P.

Note that G is a perfect map if (DG) is satisfied and furthermore $\mathcal{I}(P) \subseteq \mathcal{I}(G)$ holds. This second assumption is called *faithfulness*:

Definition 6 (Faithfulness)

If $\mathcal{I}(P) \subseteq \mathcal{I}(G)$ then P is called *faithful* to the graph G.

The distinction between faithfulness and perfectness is entirely technical: In practice, we *always* assume G is an I-map, in which case perfectness is equivalent to faithfulness. This assumption will be discussed further in Section 4.2.

3.2 Minimal I-maps

Adding edges to a graph G will remove local independencies and removing edges from G will introduce new local independencies. In other words, as we remove edges from G, the set $\mathcal{I}(G)$ gets larger. Eventually, unless $\mathcal{I}(P)$ contains all possible CI relations, an I-map G of a distribution P will stop being an I-map if we remove enough edges. This motivates the notion of a minimal I-map:

Definition 7 (Minimal I-map)

An I-map G for P is minimal if no subgraph of G is also an I-map of P.

In other words, an I-map is minimal if the removal of any edge in the graph introduces independencies that are not in $\mathcal{I}(P)$. We also observe that looking for maximal I-maps does not make much sense: For instance, the complete DAG is a trivial I-map for every distribution since its set of independencies is empty.

Minimality is a natural requirement to impose on G; without it we are accepting added complexity (i.e. extra edges) with less information (i.e. fewer independencies). Luckily, we can convert any distribution into a minimal I-map: Simply remove edges until it is no longer possible to remove any edges while remaining an I-map. More precisely, starting from a complete DAG $G = (V, E)$, look for a redundant edge $e = (i, j) \in E$ such that $G - e$ is an I-map of P. Remove this edge and repeat this until no such edge exists.

Theorem 3

There exists a minimal I-map G for any distribution P.

3.3 Constructing minimal I-maps

There is a more methodical way of describing this procedure as follows. First we need the following definitions:

Definition 8 (Markov blanket and Markov boundary)

A *Markov blanket* of a node X_i in a subset $S \subseteq X$ is a set $M \subseteq X$ such that

$$X_i \perp\!\!\!\perp (S \setminus M) \,|\, M.$$

A *Markov boundary* m of a node is a minimal Markov blanket: No proper subset of m is also a Markov blanket.

We use $\mathrm{MB}(X_i; S)$ to denote the Markov boundary of X_i in S. Note that our definitions differ slightly from the definitions given in Koller and Friedman (2009) (see Definitions 4.11-12).

Using the concept of a Markov boundary, we can construct minimal I-maps.

Theorem 4

The following procedure constructs a minimal I-map for any distribution P:

1. Choose an ordering $\prec$ of the random variables $X_1, \dots, X_d$.
2. For each X_i, draw an edge from every node in $\mathrm{MB}(X_i; \{X_1, \dots, X_{i-1}\})$ to X_i, i.e. set $\mathrm{pa}(X_i) = \mathrm{MB}(X_i; \{X_1, \dots, X_{i-1}\})$.

Conversely, any minimal I-map arises in this way.

To deduce the converse, take a minimal I-map, find any topological sort, and run the above procedure. Thus, if we write $G(\prec)$ for the minimal I-map given by the ordering $\prec$, then the set of *all* minimal I-maps can be written as

$$\{G(\prec) : \prec \text{ is an ordering of the variables}\}.$$

Remark 7. Since this procedure depends on the ordering chosen in the first step, and different orderings can lead to different graphs, this also shows that P will in general have several different, equally valid minimal I-maps.

A complication arises when some Markov boundary is not unique (i.e. there is more than one minimal subset m satisfying Definition 8). Unfortunately, this is possible in general (see Statnikov et al., 2013, for examples). In this case, the output of the above procedure is not well-defined, although the output is always guaranteed to be a minimal I-map.

Fortunately, there are reasonable conditions under which uniqueness is guaranteed. We say that a distribution P satisfies the *intersection property* if the following holds for any subset of random variables $A, B, C, D \subseteq X$:

$$(A \perp\!\!\!\perp B \,|\, C, D) \text{ and } (A \perp\!\!\!\perp C \,|\, B, D) \implies (A \perp\!\!\!\perp B, C) \,|\, D. \tag{9}$$

Theorem 5 (Uniqueness of Markov boundaries; Theorem 4, Pearl, 1988)

If P satisfies the intersection property (9), then for all $X_i \in X$ and $S \subseteq X$, the Markov boundary $\mathrm{MB}(X_i; S)$ exists and is unique.

It can be proven that as long as (A2) holds, P will satisfy the intersection property. Thus, we have the following useful corollary:

Corollary 1

If P satisfies the intersection property (9), then in order to identify a minimal I-map G from P, it is enough to identify a topological ordering of G.

This fact can be used to reduce (SL) to learning a topological ordering of G: Once this ordering is known, Corollary 1 guarantees that G is identifiable. This reduction can be used to design algorithms for learning DAGs by first learning an ordering and then using variable selection to learn Markov boundaries.

Remark 8. (A2) is sufficient but not necessary for (9); see Dawid (1980); San Martin et al. (2005); Peters (2015).

4 Identifiability of I-maps

Having resolved (Q1) in the previous two sections, we now turn our attention to (Q2). Recall that answering these questions is necessary in order to make sense of the main problem (SL). Based on Sections 2-3, we can immediately see there is trouble: Not only does every distribution have more than one I-map (Example 3), in general, even minimal I-maps are not unique (Remark 7), and may not even be well-defined.

4.1 Uniqueness

What does it mean for a graph G to be uniquely determined by P? Since we have already established that *any* P has several valid minimal I-maps, such a statement requires some explanation.

We can make an (imperfect) analogy with the modes of a distribution as follows:[2] Let $\text{mode}(P)$ denote the modes of the distribution P. In general, without more context or additional assumptions, there is no reason to prefer one mode to another. We can, however, single out a particular mode as "special" depending on the application: For example, $m^* \in \text{mode}(P)$ could be the mode closest to the origin, or the mode such that the density of P evaluated at m^* is the largest. Note that both cases, this still may not be enough to uniquely specify m^*.

A similar situation arises with the I-maps of P: In general, there are many I-maps, all equally valid, however, based on context or application, there may be a particular I-map of interest. When we talk about uniqueness of I-maps, what we really mean is that there is at most one minimal I-map of P satisfying some set of assumptions that have been specified in advance. Examples from the literature include:

- Linear non-gaussian models
- Additive noise models
- Post-nonlinear models
- Equal variances

A popular assumption that is notably absent from this list is *faithfulness*. Given its ubiquity and special status, we begin by discussing faithfulness, and most importantly, why it does not identify a DAG.

4.2 Faithfulness

Self-study: (TB1) §3.3.2

Recall the definition of faithfulness in Definition 6. It turns out that a distribution P can have more than one faithful DAG G, i.e. faithfulness does not identify a DAG. Although faithfulness is not enough to identify a minimal I-map in general, it is enough to identify minimal I-maps up to *Markov equivalence*. We begin with a definition:

Definition 9

Two DAGs G and G' are called *Markov equivalent* if $\mathcal{I}(G) = \mathcal{I}(G')$.

In other words, G and G' are Markov equivalent if and only if they encode the exact same separation statements. Under faithfulness and the Markov assumption, we have $\mathcal{I}(G) = \mathcal{I}(G') = \mathcal{I}(P)$, so this can be interpreted as encoding the exact same CI relations (via d-separation) as well. Markov equivalence has a convenient graphical characterization:

Theorem 6

Two DAGs G and G' are Markov equivalent if and only if they share the same skeleton and unshielded colliders.

Putting everything together, we can define the *Markov equivalence class* of P as the collection of DAGs G such that $\mathcal{I}(P) = \mathcal{I}(G)$, i.e. all DAGs to which P is faithful. By Theorem 6, this equivalence is characterized by the skeleton and unshielded colliders of any Markov equivalent DAG. Evidently, by reversing edges that do not participate in a v-structure, we can construct distinct Markov equivalent DAGs.

The importance of faithfulness is that it forces G to encode *all* of the CI relations in P. In other words, assuming faithfulness is tantamount to declaring the CI relations in P as your primary target of inference. If this is the case, the lack of uniqueness of G is not an issue. In this sense, instead of identifying a DAG, its instructive to think of faithfulness as identifying a *set of CI relations*. These CI relations are fundamental in applications such as causal inference.

[2] Imperfect because some distributions have a unique mode, which is rarely the case for I-maps.

4.3 Linear models

Self-study: Kagan et al. (1973) §2-3, §10

Consider the linear model

$$X = W^T X + z, \quad W = (w_{ij}) \in \mathbb{R}^{d\times d}, \quad z_j \perp\!\!\!\perp \mathrm{pa}(X_j). \tag{10}$$

Unlike (1), we no longer assume necessarily gaussianity. Recall the graph G associated to X via (2), and assume that G is acyclic so that it is a BN for X (Example 1). Then we have the following:

(L1) *Non-gaussian errors.* If each z_j is non-gaussian, then there is exactly one minimal I-map such that the model (10) holds.

(L2) *Gaussian errors.* If each z_j is gaussian, then *every* minimal I-map satisfies (10) with independent gaussian noise.

In fact, in general, the linear model (10) can be decomposed into two parts: An essentially unique non-gaussian part and a non-unique gaussian part (§10.3, Kagan et al., 1973). This extreme dichotomy is a consequence of the Marcinkiewicz theorem (Marcinkiewicz, 1939) that says that *the only distribution whose log-characteristic function is a polynomial is the gaussian.* See Kagan et al. (1973) for details.

4.3.1 Linear gaussian models

Under the model (10), we can write the covariance matrix $\Sigma = \mathrm{cov}(X)$ as

$$\Sigma = (I-W)^{-1}\Omega(I-W)^{-T} = LL^T, \quad L = L(W,\Omega) := (I-W)^{-1}\Omega^{1/2}.$$

Evidently, if we re-order the variables (as in Section 3.2), then we obtain a different (but unique) Cholesky factor $\widetilde{L} = (I-\widetilde{W})^{-1}\widetilde{\Omega}^{1/2}$, which can easily be solved for the parameters $(\widetilde{W},\widetilde{\Omega})$, and it is not hard to check that $(\widetilde{W},\widetilde{\Omega}) \neq (W,L)$ in general.

This simple fact explains both (L1) and (L2): If the errors are gaussian as in (1), then $X - X\widetilde{W}$ will always be gaussian with independent marginals. In other words, the errors are independent. This is a special property of gaussians: If the errors are non-gaussian, the errors are no longer guaranteed to be independent. This is why independence of errors is so important in linear non-gaussian models.

> *Remark* 9. Thinking back to Section 4.1, the idea is that although there are many linear non-gaussian models that generate the same model as (10), at most *one* of them has independent errors.

4.4 Additive noise models

A natural nonlinear generalization of (4.3) is

$$X_j = g_j(\mathrm{pa}(X_j)) + z_j, \quad g_j : \mathbb{R}^{|\,\mathrm{pa}(X_j)|} \to \mathbb{R}, \quad z_j \perp\!\!\!\perp \mathrm{pa}(X_j). \tag{11}$$

Under some additional regularity conditions on the g_j to rule out exceptional cases, this model is also identifiable in the sense that at most one minimal I-map satisfies (11). This can be generalized to so-called *post-nonlinear* models with an additional nonlinearity $h_j(g_j(\mathrm{pa}(X_j)) + z_j)$.

4.5 Equal variances

Consider once again the linear model (10). If all noise terms have equal variance, that is $\mathrm{var}(z_j) = \sigma^2$ for all j, we say that this model satisfies the *equal variance* condition. It has been shown that assuming equal variances, the matrix W and the underlying graph G are identifiable.

The equal variance assumption is easily generalized to *nonlinear* models and *dependent* errors. Let (P,G) be a BN. We say that it satisfies the equal variance assumption if

$$\mathbb{E}\,\mathrm{var}(X_j \mid \mathrm{pa}(X_j)) \text{ does not depend on } j. \tag{12}$$

In the case of linear models, this reduces to the original equal variance assumption. The condition (12) is far from necessary, and can be relaxed substantially. There exist polynomial-time algorithms to learn the true DAG under (12).

5 Reduction to order search

A standard trick is to reduce (SL) to *order search*, i.e. finding a topological sort of G. In order for this work, the distribution P must have unique Markov boundaries. Therefore, in the remainder of this section we assume that P satisfies the intersection axiom (cf. (9)).

5.1 Order identification

Suppose G is a minimal I-map of P, and that G is identifiable from P (in the sense of Section 4). Now suppose further that we know any ordering of G, let's call this $\prec$. Then by running the algorithm in Theorem 4, we can reconstruct G from $\prec$ alone. Thus, we have the following:

Theorem 7

In order to identify G from P, it suffices to identify any ordering $\prec$ of G.

This is a convenient theoretical device: To prove that G is identifiable from P, one need only show that some ordering of G can be identified from P. This is how identifiability is often proved.

Remark 10. G may have multiple topological orders; it is only necessary to identify a single such order from P.

Practical considerations This is purely a theoretical device: Actually recovering G from $\prec$ in practice is a difficult problem. In fact, it is essentially equivalent to sparse (nonlinear) regression, and more generally, Markov boundary search. These problems are well-studied, and many practical methods are available to solve them, however, one should not be misled into thinking this is a "solved" or even "easy" problem.

5.2 Layer decomposition

There is a further simplification that can be used. Given a DAG G, define a collection of sets as follows:

$$L_0 := \emptyset, \quad A_j = \cup_{m=0}^{j} L_m, \quad L_j := \operatorname{src}(V - A_{j-1}) - A_{j-1},$$

i.e. for $j > 0$, L_j is the set of all source nodes in the subgraph $G[V - A_{j-1}]$ formed by removing the nodes in A_{j-1}. So, e.g., L_1 is the set of source nodes in G and $A_1 = L_1$. This decomposes G into layers, where each layer L_j consists of nodes that are sources in the subgraph $G[V - A_{j-1}]$, and A_j is an ancestral set for each j. Let r denote the number of "layers" in G, $L(G) := (L_1, \dots, L_r)$ be the corresponding layers. The quantity r effectively measure the depth of a DAG.

Identifying G is equivalent to identifying the sets $L_1, \dots, L_r$, since any topological order $\prec$ of G can be determined from $L(G)$, and from any order $\prec$, the graph G can be determined as described in Section 5.2. Unlike a topological order of G, which may not be unique, the layer decomposition $L(G)$ is always unique.

6 Interventions

Self-study: (TB4) §3.2-3.4 of Pearl (2009)

Up to this point we have interpreted Bayesian networks as representing dependencies between random variables. But, we can also interpret them as representing causal models. From this point of view, the edges represent direct causal relations rather than probabilistic dependencies. The causal interpretation of Bayesian networks allows us to answer *interventional queries*. Intuitively, intervention on a variable amounts to bypassing the causal mechanism affecting that variable and directly setting its value by outside intervention. Interventions are denoted by the *do* operator, and concurrently the joint distribution resulting from interventions by

$$P(X \mid \operatorname{do}(S = s)).$$

An important assumption we have is that interventions are *modular*, which means that post-intervention conditional probability distributions corresponding to each node are only effected if the node is intervened. And for those nodes, we have

$$P(X_i = s_i \mid \mathrm{pa}(X_i)) = 1$$

where s_i is the value set by the intervention.

Theorem 8 (Truncated Factorization)

We have

$$P(X \mid \mathrm{do}(S = s)) = \begin{cases} \prod_{X_i \notin S} P(X_i \mid \mathrm{pa}(X_i)), & \text{for } X \text{ consistent} \\ 0, & \text{otherwise} \end{cases}$$

where consistent means that X agrees with the intervention values on S.

7 Structural Causal Models

A very useful perspective on BNs is offered by structural causal models, defined below:

Definition 10 (SCM)

A structural causal model (SCM) is a tuple $M = (X, \mathcal{E}, \mathcal{F})$, where

1. $X = (X_1, \ldots, X_d)$ is a set of endogenous variables,
2. $\mathcal{E} = (z_1, \ldots, z_d)$ is a set of mutually independent exogenous ("noise") variables,
3. $\mathcal{F} = (f_1, \ldots, f_d)$ is a set of functions $f_j : \mathbb{R}^{d+1} \to \mathbb{R}$,

and each endogenous variable $X_j \in X$ satisfies

$$X_j = f_j(X, z_j).$$

The intuition behind an SCM is that each endogeneous variable X_j is determined by the other variables in X and the independent noise z_j.

Example 5. Additive noise models (Section 4.4) are SCMs of the form

$$X_j = g_j(\mathrm{pa}(X_j)) + z_j \implies f_j(v, z) = g_j(v) + z,\ V_j = \mathrm{pa}(X_j).$$

If g_j are linear, we have a linear ANM.

> *Remark* 11. We assume the domain of f_j is $\mathbb{R}^{d+1}$ in order to write $X_j = f_j(X, z_j)$, however, this does not imply that each X_j in fact depends on every variable in X. Clearly, if f_j is constant with respect to any of its first d arguments, then f_j does not depend on the corresponding arguments. Intuitively, we should think of the arguments over which f_j varies as the "direct causes" of X_j. An alternative definition of SCM that makes this subset explicit is that $f_j : \mathbb{R}^{k_j+1} \to \mathbb{R}$, where $k_j \in \mathbb{Z}_+$ and $X_j = f_j(V_j, z_j)$ for some endogenous subset $V_j \subseteq X$. We choose the former in order to simplify the notation and statements.

Interventional queries have a particularly nice interpretation for SCMs. Suppose we have an SCM M and we would like to determine the joint distribution after an intervention $\mathrm{do}(S = s)$. Let $M_{S=s}$ be the following modified SCM:

$$X_i = \begin{cases} f_i(V_i, z_j), & \text{if } X_i \notin S \\ s_i, & \text{if } X_i \in S \end{cases}$$

where $X_i = s_i$ is the value set by the intervention. In other words, we keep M the same except replace the intervened f_is with the constants set by the intervention. Then, $M_{S=s}d$ gives the post intervention distribution for M. In other words,

$$P_M(Y \mid \mathrm{do}(S = s)) = P_{M_{S=s}}(Y).$$

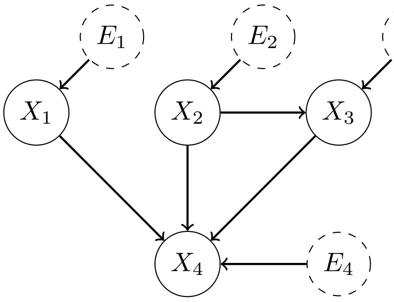


Figure 3(a): An example of an SCM graph M

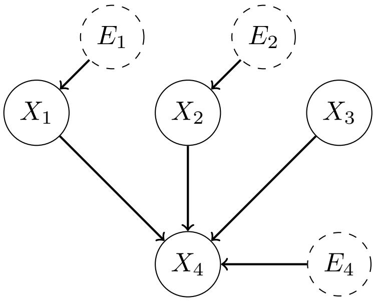


Figure 3(b): Mutilated M with $\mathrm{do}(X_3)$

7.1 Relation to graphs

It is very natural to associate a graph to an SCM M. We construct a graph $G = G(M)$ such that there is a node for each endogenous variable X_i. We put a directed edge $X_i \to X_j$ if and only if f_j is nonconstant with respect to X_i (i.e. its ith argument). Formally, let

$$V_j = \{X_i \in V : f_j \text{ is nonconstant with respect to } X_i\}$$

denote this subset and declare

$$X_i \to X_j \in G \iff X_i \in V_j.$$

For instance the graph in Figure 3a, would have an SCM with equations

$$\begin{aligned} X_1 &= f_1(z_1) \\ X_2 &= f_2(z_2) \\ X_3 &= f_3(X_2, z_3) \\ X_4 &= f_4(X_2, X_3, z_4). \end{aligned}$$

When this graph is acyclic, we call the SCM M acyclic as well.

7.2 Relation to Bayesian networks

It turns out that structural causal models and Bayesian networks are equivalent under mild conditions. A Bayesian network (P, G) defines an SCM with joint distribution P on G and vice versa. To obtain a Bayesian network from an SCM, note that the variables X form a joint distribution given an SCM. Then the following result can be proven:

Theorem 9 (SCM $\iff$ BN)

Let M be an acyclic SCM defined, P be the probability distribution induced by M, and G the DAG induced by M. Then P factorizes over G. Conversely, given any P satisfying (A1), there is an acyclic SCM that induces P and whose DAG G is an I-map of P.

This result implies that the distribution P induced by the SCM forms a Bayesian network over the same graph.

Constructing an SCM from a Bayesian network involves a different approach. Let (P, G) be a Bayesian network. This gives us local conditional probability distributions $P(X_j \mid \mathrm{pa}(X_j))$ on each node. We assume that these distributions have a density function. To obtain the functions f_j and noise terms z_j, we use inverse transform sampling. Let $F_{X_j \mid \mathrm{pa}(X_j)}$ be the CDF of the conditional distribution $P(X_j \mid \mathrm{pa}(X_j))$, $z_j \sim \mathrm{Unif}[0, 1]$, and define

$$X_j := f_j(\mathrm{pa}(X_j), z_j) := F^{-1}_{X_j \mid \mathrm{pa}(X_j)}(z_j)$$

Then, X_j is distributed with the CDF $F_{X_j \mid \mathrm{pa}(X_j)}$, and thus has the same local density as in the original Bayesian network. Note that the noises are also mutually independent, and so the resulting model is a valid SCM. Furthermore, both define the same joint distribution due to the factorization property.

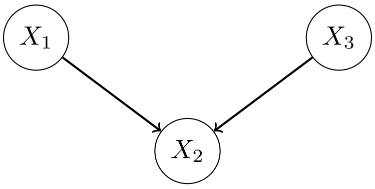


Figure 4: A v-structure where X_2 is a collider.

A d-separation and global independencies

As we mentioned in Section 2, the local Markov property encodes a set of local independencies in Bayesian networks. Although this set is purely defined in terms of the graph, under (DL) this set also encodes real conditional independencies in the distribution P. But, what if we also wanted to know about nonlocal independecies? More precisely, an independence relation such as

$$A \perp\!\!\!\perp B \mid C$$

where C is not the set of parents of A or B, is not encoded in $\mathcal{I}_{\text{loc}}(G)$. So, if we want to relate an arbitrary conditional independence to the graph structure, we need to come up with a notion of *global independence* in the graph. This brings us to *d-separation*, which is the way to define this set of global independencies.

Intuitively, if a pair of nodes have a path connecting them, we expect them to be dependent since there is a flow of information from one to another. However, if we condition on any node on this path, the flow of information would be blocked. The only exceptions to this are the so-called *colliders*, which have the opposite effect of enabling the flow of information when they are conditioned on. More precisely, we call a graph structure a *v-structure* if it is in the form

$$X \to Y \leftarrow Z,$$

and we call the child node Y a collider. The effect of enabling flow of information by conditioning on a collider is known as *Berkson's Paradox* or *collider bias*. Note that conditioning on a descendant of a collider also enables information through the collider since it gives us information about the collider.

For the following definitions, recall that a path (no qualification) means a sequence of nodes connected by edges without regard to orientation (see Section **??**). For instance, a structure such as $X_1 \leftarrow X_2 \to X_3 \to X_4 \leftarrow X_5$ would be an undirected path.

Definition 11 (Active path)

Let G be a DAG. A path $\pi = (X_1, \dots, X_n)$ in G is active given Z if the following hold:

1. For every collider X_i in π, either X_i or one of its descendants is in Z.
2. No other node in π is in Z.

If a path is not active, we call it *blocked*.

Definition 12 (d-separation)

We say that X_i and X_j are d-separated given Z if there are no active paths from X_i to X_j given Z (i.e. all paths are blocked). We denote d-separation by $\text{d-sep}_G(X_i, X_j \mid Z)$.

As mentioned in Section 2, d-separation defines a ternary relation on subsets of V which we have denoted by $\mathcal{I}(G)$ (cf. (7)).

Furthermore, as we would expect, local independencies are a part of global independencies. So, it can be shown that

$$\mathcal{I}_{\text{loc}}(G) \subseteq I(G).$$

Therefore, the local Markov property, global Markov property, and factorization are all equivalent.

References

C. Berzuini, P. Dawid, and L. Bernardinell. *Causality: Statistical perspectives and applications.* John Wiley & Sons, 2012.

A. P. Dawid. Conditional independence for statistical operations. *Annals of Statistics*, pages 598–617, 1980.

J. Jacod and P. Protter. *Probability essentials.* Springer Science & Business Media, 2012.

A. M. Kagan, C. R. Rao, and Y. V. Linnik. *Characterization problems in mathematical statistics.* Wiley, 1973.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

S. L. Lauritzen. *Graphical models.* Oxford University Press, 1996.

S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990.

M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright. *Handbook of graphical models.* CRC Press, 2018.

J. Marcinkiewicz. Sur une propriété de la loi de Gauß. *Mathematische Zeitschrift*, 44(1):612–618, 1939.

J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann, 1988.

J. Pearl. *Causality.* Cambridge university press, 2009.

J. Peters. On the intersection property of conditional independence and its application to causal discovery. *Journal of Causal Inference*, 3(1):97–108, 2015.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms.* MIT press, 2017.

D. Pollard. *A user's guide to measure theoretic probability.* Number 8. Cambridge University Press, 2002.

M. M. Rao. *Conditional measures and applications.* Chapman and Hall/CRC, 2005.

E. San Martin, M. Mouchart, and J.-M. Rolin. Ignorable common information, null sets and basu's first theorem. *Sankhyā: The Indian Journal of Statistics*, pages 674–698, 2005.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*, volume 81. The MIT Press, 2000.

A. Statnikov, N. I. Lytkin, J. Lemeire, and C. F. Aliferis. Algorithms for discovery of multiple markov boundaries. *Journal of Machine Learning Research*, 14(Feb):499–566, 2013.

M. Studený. *Probabilistic conditional independence structures.* Springer Science & Business Media, 2006.