Probabilistic Graphical Models

Chad M. Schafer

Department of Statistics & Data Science

Carnegie Mellon University

May 22, 2025

Introduction

Probabilistic Graphical Models (PGMs) provide a framework for visualizing the dependencies between random variables, parameters, and observations.

This framework can then be used as part of inference procedures to estimate the parameters of these models.

A large proportion of the usage of PGMs focuses on *Bayesian Networks*, and that is what we will cover here.

For a broader introduction see, for instance, *Probabilistic Graphical Models* by Sucar (2021). Bayesian Networks are covered in Chapters 7 and 8 of that text.
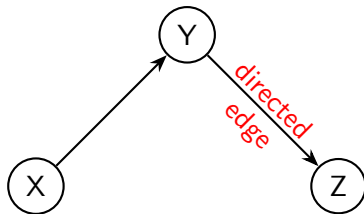
Basic Graph Theory

Some basic concepts from **graph theory** are needed for PGMs.

Graphs consist of **nodes (or vertices)** joined by **edges**. Edges (graphs) can be **undirected** or **directed**. A graph is **mixed** if it has both undirected and directed edges.
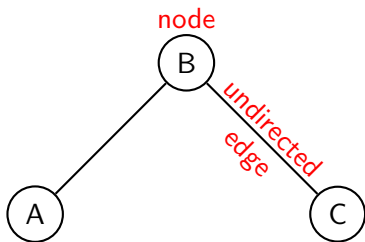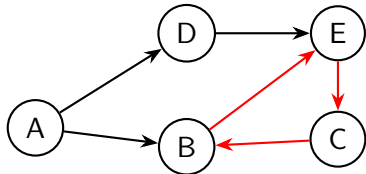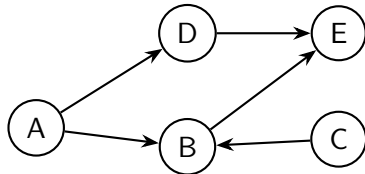

Undirected Graph   Directed Graph

Directed Acyclic Graphs (DAGs)

An **acyclic graph** is one such that there are no closed loops (cycles).

Directed Cyclic Graph   Directed Acyclic Graph (DAG)

Note: In a directed graph, the cycle must follow the direction of the arrows.

The **ancestors** of a node X are all of the nodes from which there is a (directed) path to X. In the DAG above, every other node is an ancestor of E, but C has no ancestors.

PGMs as DAGs

Key Result: Every Bayesian Network can be represented as a DAG.

Open circles represent random variables, which model both parameters and observable quantities (data).

The edges (and their directions) provide a description of the **dependence** between those random variables.
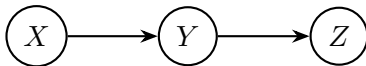
Quick Review

The **joint distribution** between X, Y, and Z can be decomposed as

$$p(X, Y, Z) = p(Z|X, Y)p(X, Y) = p(Z|X, Y)p(Y|X)p(X).$$

A pair of random variables X and Y are **independent** if

$$p(X, Y) = p(X)p(Y), \text{ and it follows that } p(X|Y) = p(X).$$

Random variables X and Z are **independent conditional on** Y if
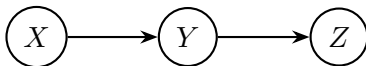
$$p(X, Z|Y) = p(X|Y)p(Z|Y), \text{ which implies that } p(X|Y, Z) = p(X|Y).$$

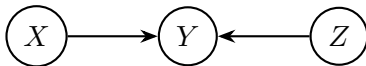These results can be extended to any number of random variables.

Describing the Joint Distribution
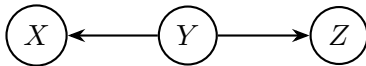
There are three key types of structures in DAGs.

Chain:

$$X \longrightarrow Y \longrightarrow Z$$

Collider:

$$X \longrightarrow Y \longleftarrow Z$$

Fork:

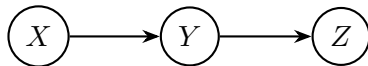$$X \longleftarrow Y \longrightarrow Z$$

Describing the Joint Distribution

If there is a (directed) path (**chain**) from X to Z, then X and Z are dependent.
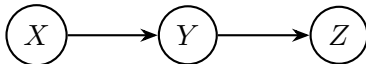

Here,

$$p(Z|X) \neq p(Z) \text{ and } p(X|Z) \neq p(X).$$

Describing the Joint Distribution

But, if every (directed) path from X to Z passes through Y, then X and Z are **independent conditional on** Y.

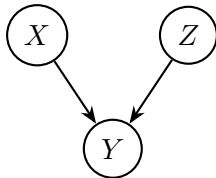Here, X and Z are independent conditional on Y, i.e.,

$$p(X, Z|Y) = p(X|Y)p(Z|Y) \text{ and } p(Z|X, Y) = p(Z|Y).$$

So, we can decompose the joint as

$$p(X, Y, Z) = p(Z|X, Y)p(X, Y) = p(Z|Y)p(Y|X)p(X).$$

Describing the Joint Distribution

Random variables which have no shared ancestors, or are not each other's ancestor, are marginally independent. This is shown as a **collider** below.

Here, without any conditioning X and Z are independent, so $p(X, Z) = p(X)p(Z)$ and

$$p(X, Y, Z) = p(Y | X, Z)p(X, Z) = p(Y | X, Z)p(X)p(Z).$$

Note, however, that X and Z are **dependent** conditional on Y.

Describing the Joint Distribution

The structure below (a **fork**) implies that X and Z are independent conditional on Y.
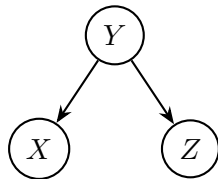
Hence,

$$p(X, Y, Z) = p(X, Y \mid Z)p(Z) = p(X \mid Z)p(Y \mid Z)p(Z)$$

Describing the Joint Distribution

These ideas can be extended to additional random variables to find general conditional independence structure. An informal statement is "Each variable is conditionally independent of its non-descendants, given its parents."
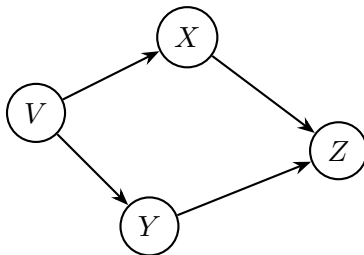
Here, Z and V are independent conditional on both X and Y, i.e.,

$$p(V, Z \mid X, Y) = p(V \mid X, Y)p(Z \mid X, Y)$$

Describing the Joint Distribution

Technical Comment: The formal way of identifying all of the conditional independence structure is somewhat more complicated, and relies on the notion of **D-separation**.

One way to characterize it as follows:

A **trail** (a path along edges that ignores directions) between two nodes is **blocked** by a set of nodes $\mathcal{Z}$ if at least one of the following is true:

- The trail contains a chain or fork structure and its middle node is in $\mathcal{Z}$.
- The trail contains a collider and neither its middle node nor the descendants of that middle node are in $\mathcal{Z}$.

If all trails between two nodes are blocked by $\mathcal{Z}$, then the nodes are D-separated by $\mathcal{Z}$, meaning they are conditionally independent given $\mathcal{Z}$.

Exercise

Reconsider the DAG below. Simplify the joint distribution of the four random variables.
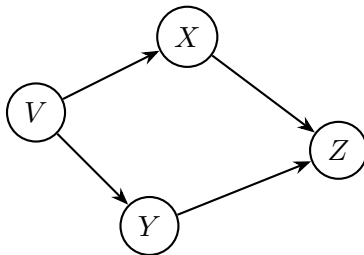
Exercise

Solution:

$$p(V, X, Y, Z) = p(Z|V, X, Y)p(X, Y|V)p(V)$$
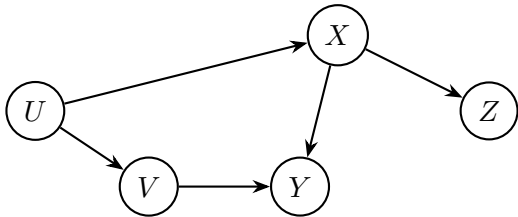$$\text{(this is always true)}$$

$$= p(Z|X, Y)p(X, Y|V)p(V)$$
$$\text{(since } Z \text{ and } V \text{ independent conditional on } X \text{ and } Y)$$

$$= p(Z|X, Y)p(X|V)p(Y|V)p(V)$$
$$\text{(since } X \text{ and } Y \text{ independent conditional on } V)$$

Describing the Joint Distribution

We notice something: **To decompose the joint distribution, you only need to condition each node on its parents.**
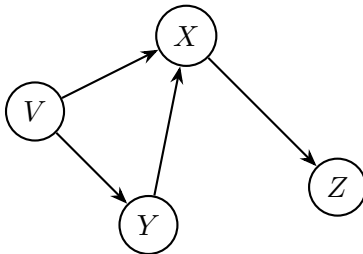

The above DAG represents the joint distribution

$$p(U, V, X, Y, Z) = p(U)p(V|U)p(X|U)p(Y|V, X)p(Z|X).$$

To see this, for each random variable ask "Who are its parents?"

Exercise

Decompose the joint distribution implied by this DAG.

List any pairs of random variables which are independent.

The Bayesian Approach

In the Bayesian framework, **parameters are random variables**.

We can talk about the distribution of a parameter, e.g. the **prior** and the **posterior**.

Recall **Bayes Theorem**:

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|t)\pi(t)}.$$

Likewise, we can characterize the joint distribution of the parameters and the observable random variables.

This makes DAGs a natural tool for describing a Bayesian model.

The Bayesian Approach

In a graph, nodes can represent both parameters **and** observable random variables.

It is common to see observable random variables represented as shaded nodes.

For example, suppose that the redshift z of a galaxy is modeled as a random variable drawn from some distribution parameterized by θ and τ:
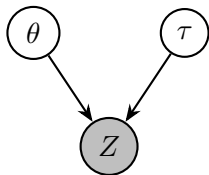
Plates

Of course, we want to model the redshift of many galaxies. To simplify notation, **plates** are used to represent repeated quantities.
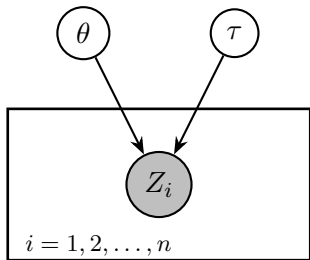
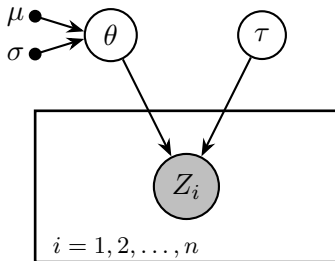This example models the redshift of n galaxies, whose (shared) distribution is described by parameters θ and τ.

Fixed Nodes

You will sometimes see closed, small markers used to represent quantities not considered random variables. These are called **fixed nodes**.

In the example below, it is emphasized that there are two parameters (μ and σ) for the prior on θ.

Hierarchical Bayes

It is common to see **hierarchical Bayes** models described by their DAG.

Here, we have now modeled μ and σ as random variables as well, so that the prior for θ has parameters which have their own priors.

$$i = 1, 2, \ldots, n$$
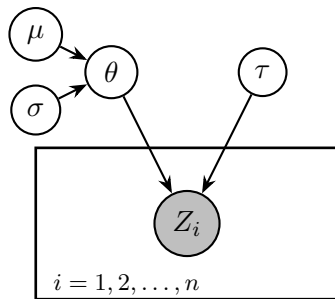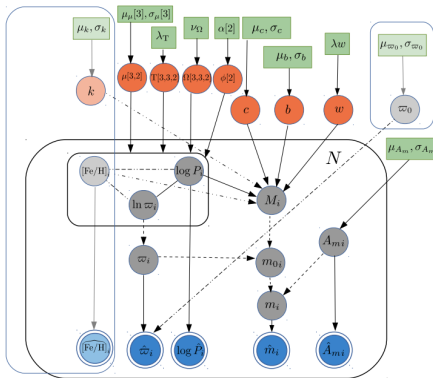
Example

Fig. 1: Directed Acyclic Graph that represents the forward model used to infer the *PLZ* relation coefficients when the prior of true metallicities, logarithm of true periods and (natural) logarithm of true parallaxes is assumed to be a 3D Gaussian mixture distribution.

Delgado, et al. (2019) "A hierarchical Bayesian model to infer PL(Z) relations using Gaia parallaxes."
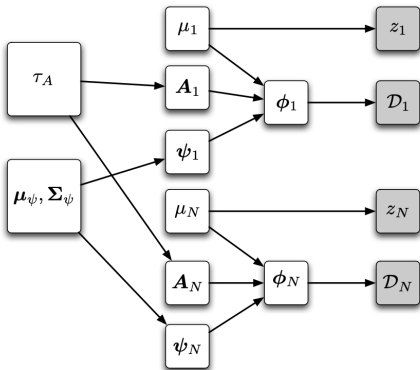
Example

FIG. 1.— Directed acyclic graph for hierarchical Bayesian inference from a training set of Type Ia SN light curves. This is a graphical representation of the joint distribution of unknown parameters and observations for a training set of N SN Ia. Each parameter is represented by a node, and the links between node indicate relationships of conditional probability. The variables in the far left column are the hyperparameters which describe the population probability distribution of supernova characteristics, and the population distribution of extinction values. The variables in the middle left column describe the distances, extinctions, and absolute light curves of individual supernovae. The variables in the middle right column are the observable parameters that describe the apparent light curves of individual SN Ia. The final column contains the observations of the redshifts and multi-band light curves of individual SN Ia. The open nodes describe unknown and hidden parameters, whereas the shaded nodes describe observed values that are conditioned upon in the posterior density.

Mandel, et al. (2009) "Type IA Supernova Light Curve Inference: Hierarchical Bayesian Analysis in the Near Infrared."

Example

In this example:

- τ_A characterizes the exponential distribution from which extinctions $\mathbf{A}_i$ are drawn.
- $\boldsymbol{\mu}_\psi$, $\boldsymbol{\Sigma}_\psi$ describe the multivariate normal distribution from which light curve parameters ψ_i are drawn.
- μ_i are the distance modulus for each SN.
- ϕ_i are the parameters of the apparent light curve.
- z_i is the measured redshift.
- $\mathcal{D}_N$ is the measured light curve.

(Note that this example does not use a plate.)

DAGs as Generative Models

A DAG efficiently defines a **generative model**.

This factorization simplifies inference and learning in probabilistic models, including guiding the design of efficient MCMC algorithms.

In Gibbs sampling, the DAG tells us which variables to condition on:

$$P(X \mid \text{rest}) = P(X \mid \text{Markov blanket of } X)$$

The **Markov blanket** of X consists of a X's parents, children, and co-parents (other nodes that point to a child of X). This reduces computational complexity by avoiding unnecessary conditioning.

DAGs also help identify independent substructures for parallel sampling.

DAFT

DAFT is a Python package designed specifically for creating graphs to illustrate DAGS.

https://docs.daft-pgm.org/en/latest/

Install using
```
$ pip install daft-pgm
```

See the accompanying notebook for examples and exercises