DGMs: Local Factors
10708, Fall 2020
Pradeep Ravikumar

1 DGMs: Conditional Probability based Local Factors

Recall that DGMs have joint distributions that factor into a product of node-conditional distributions:

$$P(X) = \prod_{j=1}^{p} P(X_i|\mathrm{PA}_i).$$

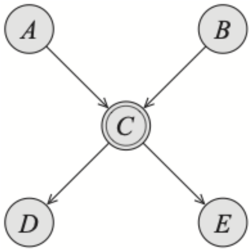
Let us briefly consider popular approaches to specify these local conditional distributions.

2 Conditional Probability Tables

The most common uses of DGMs are when all the variables are discrete. In that case, the most general form of a conditional distribution $P(Y|X)$ is a conditional probability table: with one row for each configuration x of X, and K columns corresponding to the K values $\{P(Y = y|X = x)\}_{y\in[K]}$.

3 Deterministic Nodes

In many cases, Y will be a deterministic function of X. For instance, we could have that $Y = X_1$ OR X_2 for boolean variables. In such cases, additional independencies beyond that dictated by d-separation could arise.



Example 1 *Consider Figure 1, where C is a deterministic function of A and B. In such a case, conditioning on A, B would be equivalent to also conditioning on C. We would thus conclude that $D \perp\!\!\!\perp E \,|\, A, B$ which does not follow from just d-separation in the DAG.*

Given a DAG G, and a set of nodes $\mathbf{Z}$, we define its augmentation $\mathbf{Z}^+$ as the smallest set such that:

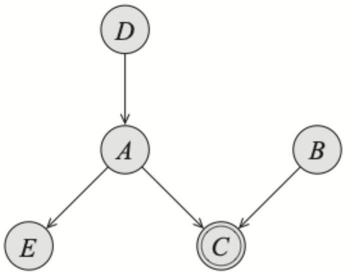
- $\mathbf{Z} \subseteq \mathbf{Z}^+$
- all nodes X such that X is deterministically specified given nodes in $\mathbf{Z}^+$ also lie in $\mathbf{Z}^+$

In particular, if there are deterministic nodes X such that $\mathrm{PA}_X \in \mathbf{Z}^+$, then they are deterministically specified given $\mathbf{Z}^+$, and hence they should also lie in $\mathbf{Z}^+$.

In the example above, we would thus have that $\{A, B\}^+ = \{A, B, C\}$, so that we do get that $D \perp\!\!\!\perp E \,|\, \{A, B\}^+$.

Definition 2 (det-d-separation) *For a DAG G with some deterministic nodes, we say that X is det-d-separated from Y given $\mathbf{Z}$ if X is d-separated from Y given $\mathbf{Z}^+$.*

But it does not suffice to only include additional deterministic nodes in the augmented set.



Example 3 *Consider the DAG in Figure 3. If $C = A$ XOR B, then, given B, C, we also know the value of A. We would thus have that: $D \perp\!\!\!\perp E \,|\, C, B$. We can see that $\{C, B\}^+ = \{C, B, A\}$, so that we do get that $D \perp\!\!\!\perp E \,|\, \{C, B\}^+$.*

4 Context Specific Independence

We have just seen that deterministic nodes can allow for additional dependencies due to implicit conditioning on a larger set of nodes. But in certain cases, they might incur such independencies only for certain values of some of the variables.

Suppose in Figure 1, we have that $C = A$ OR B. Then if $A = 1$, then this specifies the value of C: $C = 1$. So we would have that $D \perp\!\!\!\perp B \,|\, (A = 1)$. But it does not follow that $D \perp\!\!\!\perp B \,|\, (A = 0)$, since just given that $A = 0$ does not specify the value of C.

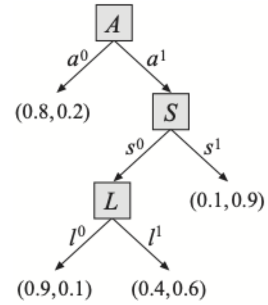
We can thus constrain the notion of conditional independence as follows.

Definition 4 (Context-Specific Independence) *Given a DAG $G = (V, E)$, let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be disjoint subsets of V. Let C be some other set (that could overlap with $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$) and let $\mathbf{c} \in Val(\mathbf{C})$. We say that $\mathbf{X}$ is contextually independent of $\mathbf{Y}$ given $\mathbf{Z}$ and the context $\mathbf{c}$, denoted by $\mathbf{X} \perp\!\!\!\perp_{\mathbf{c}} \mathbf{Y} \,|\, \mathbf{Z}$ if:*

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}, (\mathbf{C} = \mathbf{c}).$$

While the example above might suggest that this is perhaps restricted to specific deterministic functions, contextual conditional independencies can arise even with purely stochastic conditional distributions. Let us consider a few popular classes of such conditional distributions.

4.1 Tree CPDs



Example 5 *Here $P(Y|X)$ is specified by a decision density tree. See for instance Figure 5.*

4.2 Gated CPDs

Suppose $P(Y|A = a, X_1, \ldots, X_k) = I[Y = X_a]$. In this case, the variable A acts as a gate that specifies which of the parents of Y will be set as the value of Y. This is a very popular choice when we are not sure which of a set of choices to pick, and can delegate that to a gate variable A.

4.3 d-separation variant

One caveat of general context-specific independencies is that they cannot be read directly from the DAG G, as we have seen in the examples above. However, we could use the specifics of the conditional probabilities to specify a reduced graph given a context $\mathbf{c}$. Given a DGM with DAG G, and context $\mathbf{c}$, we then say that $G^-_{\mathbf{c}}$ is the contextually reduced DAG if we remove from G all edges $Y \to X$ such that $X \perp\!\!\!\perp_{\mathbf{c}} Y \,|\, \mathrm{PA}_X - \{Y\}$.

We then say that $\mathbf{X}$ is contextually d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ and context $\mathbf{c}$ if $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}, \mathbf{C}$ in DAG $G^-_{\mathbf{c}}$.

5 Independent Parents

One of the caveats with DGM factors is that the local factor $P(X|\mathrm{PA}_X)$ while local could still require large storage complexity if the size of PA_X is large (exponential in this size in the worst case). There are however many popular parametric classes of conditional distributions that have much lower representational complexity. Loosely, these can be motivated by limiting interactions among the parent “causes”.

5.1 Noisy OR Model

One of the most popular instances of such a conditional distribution is a noisy OR model, for the case where all variables are binary. Consider the simple OR function: $Y = X_1$ OR $\ldots$ OR X_k. This can be expressed compactly as:

$$1 - Y = \prod_{i=1}^{k}(1 - X_i),$$

so that $Y = 1 - \prod_{i=1}^{k}(1 - X_i)$.

A natural stochastic variant of this is given as:

$$P(Y = 0) = \alpha_0 \prod_{i=1}^{k} \alpha_i^{X_i}.$$

Here $\alpha_0 = 1 - \epsilon_0$ for some very small ϵ_0, and $\alpha_i \in (0, 1)$. The more variables that are equal to one, the lower the probability that Y is equal to zero. If none of the variables are equal to one, then $P(Y = 0) = \alpha_0$. The purpose of α_0 is to allow for a small probability for $Y = 1$ even when none of the variables are equal to one.

This can be seen to be a very natural conditional distribution when there is weak or no dependence among the parents. For instance, in a medical diagnosis context, where are symptom has multiple parents, each of which is a relatively independent disease, one could argue that one would see the symptom if at least one of the diseases are active. The noisy OR model allows for a stochastic version of this simple OR function to accomodate for any uncertainty.

5.2 Generalized Linear Models

Another popular class of conditional distributions that do not parameterize dependencies among the parents are the broad class of generalized linear models (GLMs). We can consider the following simplified instances of such GLMs:

$$P(Y|X) = \exp(\theta^T XY + C(Y) - B(\theta, X)),$$

where $C(Y)$ is some fixed function of the child, and $B(\theta, X)$ is the log-normalization constant:

$$B(\theta, X) = \log \sum_y \exp(\theta^T Xy + C(y)).$$

Logistic Regression Model. Here $Y \in \{0, 1\}$ (though this can also be generalized to the case where Y takes a larger number of finitely many values, in which case it is called a multiclass logistic regression model). We then have:

$$P(Y = 1|X) = \text{sigmoid}(\theta^T X + \theta_0),$$

which can be seen to be an instance of the general GLM form above with $C(Y) = \theta_0 Y$.

Linear Regression Model. Here $Y \in \mathbb{R}$, and $Y \sim \mathcal{N}(\theta^T X + \theta_0, \sigma^2)$. Here $C(Y) = -1/2Y^2 + \theta_0 Y$.

6 Structural Equation Models

One could use DGMs for causal reasoning by explicitly specifying the conditional distribution in equation form (also called a *strucural equation*):

$$Y = f_Y(X, N_Y),$$

where f_Y is a deterministic function, and N_Y is a noise random vector independent of all other variables in the DGM. An important instance is where $N_Y =$, and Y is a deterministic function of X. Some specific parametric instances are popular for reasons of identifiability (distinct SEMs give rise to the same observational distribution), as we will see later when we consider learning such models from data.

6.1 Linear Gaussian Models

The simplest instance of such an SEM is where:

$$Y = \theta^T X + N_Y,$$

where $N_Y \sim \mathcal{N}(0, \sigma^2)$. These are the most well-studied class of SEMs, but as we will see this class suffers from a lack of identifiability without additional restrictions.

6.2 Linear Additive Non-Gaussian Noise Models

Here we again have the linear additive noise model:

$$Y = \theta^T X + N_Y,$$

but where N_Y is now non-Gaussian. It turns out that one can consider the ill-specified setting where we only impose non-Gaussianity, but can still recover the linear model coefficients.

6.3 Non-linear Additive Noise Models

Here we assume that:

$$Y = f_\theta(X) + N_Y,$$

for some non-linear function $f_\theta(\cdot)$.

6.4 Post-nonlinear Models

Here:

$$Y = g_\beta(f_\theta(X) + N_Y),$$

so that there is a non-linear transformation of an non-linear additive noise model.