12 Graphical Models

Imagine how hard physics would be if electrons could think.
—Murray Gell-Mann

12.1 Graphical Models

12.1.1 Probabilistic Graphical Models

In this section we discuss how to represent conditional dependencies graphically. Specifically, we work with a *directed graphical model (DGM)*, in which the nodes correspond to random variables and the edges have direction which encodes conditional independence relations among the nodes. In this section, for simplicity, we assume that all directed graphical models are actually *directed acyclic graphs* (DAGs), that is, they have no cycles; see Figure 12.1.

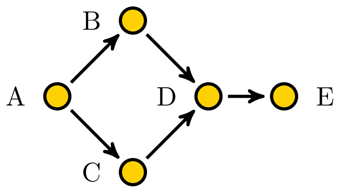


Figure 12.1: An example of a directed acyclic graph (DAG). Because a DAG has no cycles, all paths terminate in a finite number of steps.

Nota Bene 12.1.1. Not all DGMs are DAGs, however it is more difficult to prove probabilistic and causal relationships on graphs that are not DAGs. Since this section is only a brief overview of the huge subject of causal inference, we restrict our study to the simpler case of causal diagrams on DAGs.

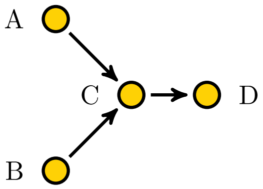


Figure 12.2: A DGM indicating some conditional independence relations, as described in Example 12.1.4.

Definition 12.1.2. *In a DAG a node Y is a* parent *of X if there is a directed edge from Y to X. A node Z is a* descendant *of X if there is a path from X to Z. If Z is a descendant of X, then X is called an* ancestor *of Z.*

The key property of directed graphical model is that all the information about all predecessor nodes is contained in the parent nodes. For example, the DGM in Figure 12.1 indicates that $P(E \,|\, ABCD) = P(E \,|\, D)$ since the only parent to E is D. Similarly, $P(D \,|\, ABC) = P(D \,|\, BC)$. Since A is not a parent of D, knowledge of A does not add any information about D beyond what was passed to B and C. This informal statement about how all the information is contained in the parent nodes can be expressed more precisely in the language of conditional independence. Recall that random variables X and Y are conditionally independent given Z, denoted $X \perp\!\!\!\perp Y \,|\, Z$, if $P(X, Y \,|\, Z) = P(X \,|\, Z)\, P(Y \,|\, Z)$. (See Definition 1.4.14).

DGMs are useful because they allow us to encode a large number of conditional independence relations efficiently. Indeed, fewer arrows in the DGM means more conditional independence relations among the variables.

Definition 12.1.3. *A* directed graphical model (DGM) *is a directed graph G whose nodes correspond to random variables, such that for any node X in G, conditioned on the parents of X, the node X is conditionally independent of every other non-parent node in G that is not a parent or a descendant of X. That is, if $W_1, \ldots, W_k$ are the parents of X, then for any other node $Y \in G$ that is not a descendant or a parent of X, we have $X \perp\!\!\!\perp Y \,\big|\, W_1, \ldots, W_k$. If P is the joint distribution for the random variables corresponding to the nodes of a DGM, then the graph G is said to* represent *P.*

Example 12.1.4. Consider the DGM G in Figure 12.2. Since A and B have no parents (and hence, also, neither is a descendant of the other), they are independent (unconditionally). Moreover, since the only parent of D is C, we have $D \perp\!\!\!\perp A \,|\, C$ and $D \perp\!\!\!\perp B \,|\, C$.

The product rule of probability gives

$$P(A, B, C, D) = P(D \,|\, A, B, C)\, P(C \,|\, A, B)\, P(B \,|\, A) P(A).$$

The conditional independence properties allow us to simplify this expression further:

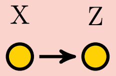
$$\begin{aligned} P(A,B,C,D) &= P(D \,|\, A,B,C)\, P(C \,|\, A,B) P(B \,|\, A) P(A) \\ &= P(D \,|\, C) P(C \,|\, A,B) P(B) P(A). \end{aligned}$$

Example 12.1.5. . Condidtional independence rules can greatly simlify the computation of joint probabilities. For example, if $X \perp\!\!\!\perp Y \,|\, Z$ and $X \perp\!\!\!\perp W \,|\, Z$ and $X \perp\!\!\!\perp V \,|\, Z$ and $Z \perp\!\!\!\perp W \,|\, V$, then

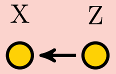
$$\begin{aligned} P(X,Y,Z,W,V) &= P(V)P(X,Y,Z,W \,|\, V) \qquad \text{product rule} \\ &= P(V)P(Z,W \,|\, V)P(X,Y \,|\, Z,W,V) \qquad \text{product rule} \\ &= P(V)P(Z \,|\, V)P(W \,|\, V)P(X \,|\, Z,W,V)P(Y \,|\, Z,W,V) \end{aligned}$$

In Exercise 12.1, you will be asked to find a DGM that produces the same rules.

Nota Bene 12.1.6. This section uses graphical models to indicate conditional dependence and independence relations among random variables. In this setting it is tempting, but incorrect, to assume that an arrow indicates causality, because there are multiple ways to draw a graphical model that incorporates the same probabilistic information but which would suggest different causal information. For example the probabilitistic graphical model

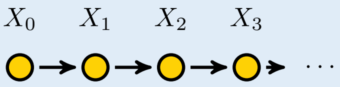


indicates only that X and Z are not necessarily independent. In the case of this simple model, that same probabilistic information could also be conveyed with the arrow reversed.



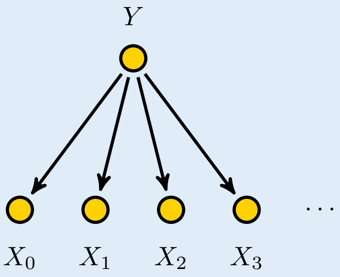
Nevertheless, there is a more enriched setting in which causal inference is possible and in which the arrows do suggest causality. We discuss that later in this chapter.

Example 12.1.7. A Markov chain corresponds to the following graphical model.



The lack of arrow from X_0 to X_2 indicates that X_2 is conditionally independent from X_0, given X_1. And similarly, the lack of any arrow from X_t to X_{t+k} for $k > 1$ indicates that X_{t+k} is conditionally independent from X_t, given X_{t+k-1}.

Example 12.1.8. The naïve Bayes classifier corresponds to a probability model where each feature X_i is conditionally independent of every other X_j given Y. A graphical model for this is given below.



Remark 12.1.9. A DGM is sometimes called a *Bayesian network* or *Bayes net*. It is also sometimes called a *belief network*. If the arrows in the model actually represent causality, then these are called *causal diagrams*.

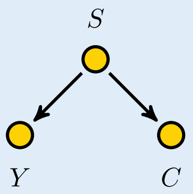
There are at least three benefits of using DGMs:

(i) Transparent representation of an otherwise-complicated collection of relations and independence.

(ii) Efficient inference about a situation modeled by the DGM, that is, efficiently answering questions about the probability of some variables, given what we know about others.

(iii) Effective learning of a model from data.

Example 12.1.10. Here is a possible DGM for the relation between smoking S, cancer C, and yellow fingernails Y



The lack of arrow from Y to C does not mean that there is no relation between yellow nails and cancer. It means that Y and C are conditionally independent after conditioning on the parent node S. If I know the smoking habits of a patient, then looking at his nails will not give me more information about his cancer risk. But if I do not know about his smoking habits, then seeing yellow nails gives me information about his smoking, which gives me information about his cancer risk.

12.1.2 Factorization of the Joint Distribution

The conditional independence relations in a DGM give a way to factor the joint distribution into something potentially much simpler. And conversely, if the joint distribution factors in the right way, then that implies that conditional independence relations must hold.

Theorem 12.1.11. *Let G be a finite DAG with nodes $X_1, \dots, X_n$ and P be a probability distribution on the nodes. The graph G represents P if and only if P factors as*

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i \,|\, \mathrm{par}(X_i)), \tag{12.1}$$

where $\mathrm{par}(X)$ *denotes the set of parents of X in the graph G.*

The proof relies on the following lemma.

Lemma 12.1.12. *The vertices of a finite DAG G can be sorted in such a way that if (u, v) is a directed edge in the DAG, then u precedes v (this is called a* topological sorting of G*).*

Proof. The topological sorting is constructed explicitly with an algorithm called *Kahn's algorithm*. Begin with an empty list L which will eventually be the desired sorted list of nodes. Since the G has no cycles, there must be at least one node without any incoming edges (a *source* node). Let S be the set of all source nodes. For each $s \in S$, do the following:

(i) Remove s from S.

(ii) Add s to the sorted list L.

(iii) Remove all edges out of s from the graph G.

(iv) The previous step may create new source nodes in the new graph; if so, add all the new source nodes to S.

Once all the nodes in S are processed, return L. This algorithm must terminate because there are only a finite number of vertices in G. Moreover, the result L must be a topological sorting of the nodes of G because for any edge (u, v), the node v cannot be added to S, and hence cannot be added to L, until u has been processed and added to L. Hence $u < v$ in L. □

We are now ready for the proof of Theorem 12.1.11.

Proof. (of Theorem 12.1.11)

If (12.1) holds, then reindex so that the nodes have a topological ordering. For any node $A = X_k$, and any fixed values $x_1, \ldots, x_k$, and letting $x_{k+1}, \ldots, x_n$ range over all possible values, we have

$$\begin{aligned} P(x_1, \ldots, x_k) &= \sum_{x_{k+1},\ldots,x_n} P(x_1, \ldots, x_{k-1}, x_k, \ldots x_n) \\ &= \sum_{x_{k+1},\ldots,x_n} \prod_i P(x_i \,|\, \mathrm{par}(x_i)) \\ &= \prod_{i\le k} P(x_i \,|\, \mathrm{par}(x_i)) \sum_{x_{k+1},\ldots,x_n} \prod_{j>k} P(x_j \,|\, \mathrm{par}(x_j)) \\ &= \prod_{i\le k} P(x_i \,|\, \mathrm{par}(x_i)), \end{aligned}$$

where the penultimate equality follows from the topological ordering, which implies that the parents of x_i lie in the set $x_1, \ldots, x_{i-1}$.

Moreover, we have

$$\begin{aligned} P(x_k \,|\, x_1, \ldots, x_{k-1}) &= \frac{P(x_1, \ldots, x_{k-1}, x_k)}{P(x_1, \ldots, x_{k-1})} \\ &= \frac{\prod_{i\le k} P(x_i \,|\, \mathrm{par}(x_i))}{\prod_{i\le k-1} P(x_i \,|\, \mathrm{par}(x_i))} \\ &= P(x_k \,|\, \mathrm{par}(x_k)) \end{aligned}$$

Therefore, for any $j < k$ if $B = X_j$ is not a parent of $A = X_k$, then $A \perp\!\!\!\perp B \,|\, \mathrm{par}(A)$.

Furthermore, we claim that if $j > k$ and node $B = X_j$ is not a descendant of node $A = X_k$, then there is a topological ordering of the graph so that, after reordering, $B = X_\ell$ and $A = X_m$ with $\ell < m$. To see that the claim holds, start at B and follow arrows backward to find an ancestor C that has no incoming edges. To create the desired topological ordering, use Kahn's algorithm, but start by choosing the node C from S, and then each time that the next $s \in S$ must be chosen, choose the next node on the path from C to B until B is chosen. Once B has been chosen, finish Kahn's algorithm as usual, choosing $s \in S$ at random. Because A does not lie on the path from C to B, this procedure will give B a smaller index than A, and the claim is proved.

Applying the earlier independence argument to the new topological ordering gives $A \perp\!\!\!\perp B \,|\, \mathrm{par}(A)$. Therefore G represents P, as required.

The proof of the converse is Exercise 12.3. □

12.2 Inference, Junctions and d-Separation

In this section we address two different questions. First, given a DGM with $P(X \mid \text{par}(X))$ for each node X, can we find other probabilities involving the nodes of the model? This is the question of *inference*. The second question is the following: For two nodes A and B in a DGM and another collection $\mathbf{Z}$ of nodes that we condition on, we'd like to know whether A is conditionally independent of B, conditioned on $\mathbf{Z}$; that is, we want to know if $A \perp\!\!\!\perp B \mid \mathbf{Z}$. To study this problem, we first consider *junctions* in a DGM, which are simple three-node subgraphs, and then we show that all questions of independence can be simplified to questions about sequences of junctions.

12.2.1 Inference in a DGM

Any question you might ask about probabilities of random variables $X_1, \ldots, X_n$ can be answered by the full joint distribution $P(X_1, \ldots, X_n)$. If these random variables are nodes in a DGM, then the joint distribution factors as in (12.1), which usually makes the joint distribution much easier to store, understand, and compute with.

For n random variables, each of which takes d possible values, a table storing the full joint distribution requires d^n entries. But the conditional independence assumptions of the DGM can greatly reduce this number. For example, consider the DGM in Figure 12.3.

The joint distribution factors as

$$P(A, V, L, C) = P(A)P(V)P(L \mid A, V)P(C \mid L),$$

which means we need only specify the four factors: $P(A)$ and $P(V)$ each has d values, while $P(L \mid A, C)$ has d^3 values, and $P(C \mid L)$ has d^2 values, for a total of $2d + d^3 + d^2$ entries in the various tables, instead of d^4. As the number of nodes increases, and especially when the number of edges is relatively small, the overall spatial savings can be significant.

Example 12.2.1. Let DGM in Figure 12.3 model the situation of respiratory patients who could get irritated lungs (node L) either from air pollution (node A) or from a virus (node V). Let's say that each of these is a Bernoulli random variable, and that exposure to air pollution occurs with probability 0.8, infection by a virus occurs with probability 0.4, and the conditional probabilities of lung irritation are given in the following table:

A	V	$P(L = T \mid A, V)$
T	T	0.99
T	F	0.3
F	T	0.8
F	F	0.1

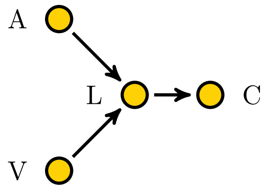


Figure 12.3: A DGM for the situation of respiratory patients whose lungs may be irritated (L) by air pollution (A) or by a viral infection (V). When the lungs are irritated, that can cause a cough (C).

If the patients lungs are irritated they might also develop a cough (node C). The conditional probabilities for the cough are $P(C = T \,|\, L = T) = 0.75$ and $P(C = T \,|\, L = F) = 0.05$.

Any probability involving only these random variables can be computed from the joint distribution $P(A, V, L, C)$. The fact that the joint distribution factors as

$$P(A, V, C, L) = P(A)P(V)P(L \,|\, A, V)P(C \,|\, L)$$

makes many of these computations easier. For example, suppose we want to know the probability that the patient has a viral infection, given that they have a cough and there is air pollution present. We have

$$\begin{aligned}
&P(V = T \,|\, A = T, C = T) \\
&\qquad = \frac{P(A = T, C = T, V = T)}{P(A = T, C = T)} \\
&\qquad = \frac{\sum_\ell P(A = T, C = T, V = T, L = \ell)}{\sum_\ell \sum_v P(A = T, C = T, V = v, L = \ell)} \\
&\qquad = \frac{\sum_\ell \cancel{P(A = T)} P(V = T)P(L = \ell \,|\, A = T, V = T)P(C = T \,|\, L = \ell)}{\sum_\ell \sum_v \cancel{P(A = T)} P(V = v)P(L = \ell \,|\, A = T, V = v)P(C = T \,|\, L = \ell)} \\
&\qquad = \frac{P(V = T) \sum_\ell P(L = \ell \,|\, A = T, V = T)P(C = T \,|\, L = \ell)}{\sum_v P(V = v) \sum_\ell P(L = \ell \,|\, A = T, V = v)P(C = T \,|\, L = \ell)},
\end{aligned}$$

where the sums run over $\ell, v \in \{T, F\}$. This gives

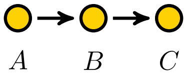
$$\begin{aligned}
&P(V = T \,|\, A = T, C = T) \\
&\qquad = \frac{0.4\,[(0.99)(0.75) + (0.01)(0.05)]}{0.4\,[(0.99)(0.75) + (0.01)(0.05)] + 0.6\,[(0.3)(0.75) + (0.7)(0.05)]}
\end{aligned}$$

12.2.2 Junctions

Three-node subgraphs with exactly two arrows in a DGM are called *junctions*, and they form the basic building blocks of all DGMs.

There are three types of junctions:

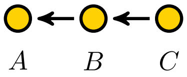
(i) *Chain:*



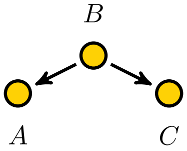
In a chain the middle term B is called a *mediator* because B mediates the action of A on C, and $A \perp\!\!\!\perp C \,\big|\, B$. But if we don't condition on B, then A and C need not be independent.

An example is Fire (F) → Smoke (S) → Alarm (A). A smoke alarm is only triggered by smoke—not fire. Smoke is the mediator. If there is a fire but no smoke (S^c)—for example, if the smoke is all sucked away with a fume hood—then the alarm will not ring. And similarly, if there is smoke, then the alarm will ring whether there is a fire or not. Thus, we have $F \perp\!\!\!\perp A \,\big|\, S$.

Of course, chains need not run from left to right, they can also run in the other direction, like this:



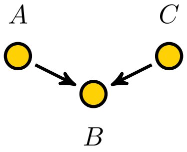
(ii) *Fork:*. The second sort of junction is a fork, which has the following form.



In a fork the middle term B is called a *common cause* or *confounder*. In this case, A and C are conditionally independent, given B; but if we don't condition on B, then A and C need not be independent.

In Example 12.1.10 smoking is a confounder for the relation between yellow nails and cancer. If we know the smoking habits of the patient, then looking at his nails will give us no more information about his risk of cancer. But if we do not know about his smoking habits, then seeing yellow nails gives information about those habits, and thus gives information about his cancer risk.

(iii) *Collider:* The final type of junction is a collider, which has the following form.



The node B in the center of a collider junction is called a *common effect* of A and C or a *collider*. One commonly used example of a collider is the dating situation for a woman (Alice) who will only date men if the sum of their niceness and handsomeness is greater than some fixed threshold. The diagram for her dating is a collider

$$\text{Nice} \to \text{Date} \leftarrow \text{Handsome}$$

A collider works differently than a chain or fork, where conditioning on the middle term B mades the outer nodes independent. Instead, A and C are unconditionally independent, but conditioning on the common effect B in a collider can make the otherwise-independent A and C correlated. Niceness and handsomeness are not statistically correlated in the general population, but if we know that Alice is willing to date someone, then niceness and handsomeness become negatively correlated because if Alice dates a rude man, he must be very handsome, and if she dates an ugly man, he must be very nice. This correlation arising from conditioning on a collider is called *collider bias*.

12.2.3 Junctions Account for all Dependence

The conditional independence relations encoded in a DGM imply other conditional independence relations as well. As mentioned in the introduction to this section, we'd like to describe all such relations. For two nodes A and B in a DGM and another collection $\mathbf{Z}$ of nodes that we condition on, we'd like to know whether A is conditionally independent of B, conditioned on $\mathbf{Z}$; that is, we want to know if $A \perp\!\!\!\perp B \,\big|\, \mathbf{Z}$.

Surprisingly, all true correlation (failure to be independent) can be accounted for by collider bias (conditioning on a common effect) and chains and forks. If we condition on a set $\mathbf{Z}$ of nodes in a DGM, every pair of nodes in the DGM that are not conditionally independent given $\mathbf{Z}$ must be connected by a sequence (a *trail*) of such junctions. Every pair that is not connected by such a trail is conditionally independent, given $\mathbf{Z}$.

Definition 12.2.2. *A* trail *in a directed graph Γ from A to C is sequence of nodes $A = B_0, B_1, B_2, \dots, B_{k-1}, B_k = C$ with a directed edge connecting each successive pair, but the orientation of that edge can be arbitrary; that is, we have $B_i \to B_{i+1}$ or $B_i \leftarrow B_{i+1}$ for each $i \in \{0, \dots, k-1\}$. We denote such a trail by*

$$A = B_0 \rightleftharpoons B_1 \rightleftharpoons \cdots \rightleftharpoons B_{k-1} \rightleftharpoons B_k = C. \tag{12.2}$$

Given a set $\mathbf{Z}$ of nodes in a directed acyclic graph Γ, a pair of nodes A and C in Γ but not in $\mathbf{Z}$ are d-connected given (conditioned on) $\mathbf{Z}$*, if there exists a trail* (12.2) *such that for each $i \in \{1, \dots, k-1\}$ the following hold:*

(*i*) *If the junction in the trail at i is a chain or a fork, then neither B_i nor any of its descendants lie in $\mathbf{Z}$*

(*ii*) *If the junction in the trail at i is a collider, then $B_i \in \mathbf{Z}$ or a descendant of B_i lies in $\mathbf{Z}$.*

The pair A, C is d-separated by $\mathbf{Z}$ *in Γ if A and C are not d-connected given $\mathbf{Z}$.*

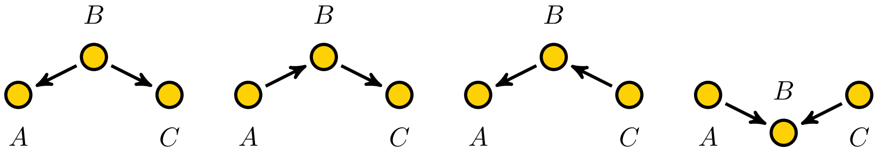


Figure 12.4: The fundamental junctions. The leftmost DAG is a fork (common cause) the middle two are chains and the rightmost has a collider (common effect) at B

Example 12.2.3. Consider the junctions in Figure 12.4.

(i) When B is not a collider, A and C are d-connected without conditioning, but they are d-separated by (given) B.

(ii) If A and C collide at B, then A and C are d-separated without conditioning, but they are d- connected given B .

(iii) Conditioning on the descendant of a collider has the same effect as conditioning on the collider. Thus in Figure 12.5, A and C are unconditionally d-separated, but conditioning on D changes that: they are d-connected given D.

The following theorem shows the essential relation between of d-separation and conditional independence. The proof would take us beyond the scope of this text. The interested reader can find it in [KF09, Sections 3.3.2 & 4.5.1.1].

Theorem 12.2.4. *In a DGM* Γ*, if* A *and* C *are d-separated by* $\mathbf{Z}$ *in* Γ *then* $A \perp\!\!\!\perp C \,\big|\, \mathbf{Z}$*. Conversely, if* A *and* C *are not d-separated by* $\mathbf{Z}$ *in a directed graph* Γ*, then there exists a probability distribution on the nodes of* Γ *such that* Γ *is a DGM and* A *and* C *are not independent given* $\mathbf{Z}$*. Moreover,*

In fact a much stronger version of the second part holds: if A and C are not d-separated, then, in the space of all probability distributions represented by Γ, the subset of distributions for which A and C are independent has measure zero (see [KF09, Theorem 3.5]).

Nota Bene 12.2.5. There are two situations where a correlation might not be the result of a d-connected trail in a causal diagram. These are first, when there is no correlation, just the appearance of one; and second, when the diagram is not correct.

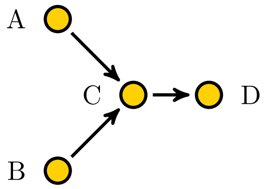


Figure 12.5: A collider with a descendent.

First a correlation may appear to exist due to random chance, even when the events are actually independent. A simple example is the case of flipping two coins, once each. If they both come up the same, we might erroneously conclude that they are positively correlated. And, if they come up different, we might erroneously conclude they are negatively correlated. Sampling more times shows that both of these apparent correlations are false. This may seem like a silly example, but the principle underlies many cases where people find and believe a false correlation.

Second, we may have an incorrect model, where the graph does not correspond to a DGM for the true distribution. For example when a common cause has been omitted from the diagram, then there really is a d-connected causal path in the true model, but we don't see it because we are using an incorrect model.

12.3 Causal Reasoning

One of the most common logical fallacies is when someone falsely assumes a causal relationship when only a correlation or association is present. In 1947, B. F. Skinner performed a series of experiments that demonstrated birds forming causal beliefs. In the experiment he presented food at 20-second intervals to hungry pigeons with no reference whatsoever to their current behavior. Soon the birds started to display ritualistic behavior such as turning around several times about the cage, bobbing their head up and down, and flapping their wings. As Skinner noticed, the birds happened to be maneuvering similarly when the food first appeared and they tended to repeat these movements, presumably thinking their action *caused* the food to appear.

Superstitions aren't just for the birds. It is widely observed in human endeavors as well—particularly in situations of high stress. Prominent stock market traders on Wall Street have been known to have lucky ties. Other traders have various beliefs about trading before a full moon or on Friday the 13th.

An industry well known for superstition is professional baseball. There are numerous rituals that can be observed before, during, and after games. Some players will hit the bat on their feet with a certain succession or tug on various articles of clothing between pitches. One famous player was observed kissing a religious ornament on a necklace after each home run. Another famous pitcher was seen touching a statue of Babe Ruth before every home game.

As Skinner noted, the pigeons in his study had been reduced to 75% of their previous body weight before the experiment commenced—they were hungry. In the case of stock market traders and major league baseball players, slumps are common and and the threat of being fired or demoted to the minors is very real; when a payer or trader enjoys the relief of intermittent success following hard times, they will try not to jinx their new winning streak by changing behavior. Their objectivity seems to wane in the presence of stress.

In everyday life we see superstitions all the time. Many buildings, especially hotels, do not have a 13th floor—and seldom will an airline have a row 13 in passenger seating. It's hard to have a Friday the 13th and not hear about it, and a long series of horror films have even been made commemorating that fact. Some people carry around ornaments for good luck. The number 666 always evokes comment. The idiom that bad luck comes in threes. The list goes on. People are innately superstitious, even when the stakes are low. Perhaps it is when they are stressed that it becomes more prominent.

In this section we examine the difference between association and cause-and-effect relationships. In medicine, causal reasoning is essential to understanding whether a drug or procedure is actually improving a patient's condition and is not just a placebo. Hence the gold standard for medical research is the double blind randomized control study, where neither the patient nor the clinician knows whether the patient is in the exposed group or the control group. But this gold standard only works when you have the time and budget to carry out such a study. There are many situations where it's infeasible or even impossible. For example, in most economics applications it would be very difficult to do a double blind randomized control study to test a market.

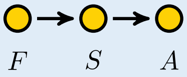
12.3.1 Causal Diagrams

Probabilistic models, on their own, cannot give information about causality, hence the common warning in statistics classes that *correlation is not causation* (see also Figure 7.7). An expression like $P(X \mid Y)$ gives the probability of observing X if we have already observed Y, but it does not give the probability of observing X if we *do* Y. For example, the probability of seeing lung cancer X in a patient, given that he has yellow fingernails Y, may be much higher than the probability of seeing lung cancer in a patient with normal fingernails, but it does not tell us the probability of seeing lung cancer in a patient if we dye his fingernails yellow.

A fundamental tool for studying causality is a *causal diagram*.

Definition 12.3.1. *A* causal diagram *is a DGM in which any arrow $Y \to X$ indicates the* possibility *that Y is a direct cause or influence on X. The lack of an arrow between two random variables R and S indicates that, conditioned on the parents of R, there is no impact by S on R.*[42]

Example 12.3.2. Consider the situation of a smoke alarm (A), which is only triggered by smoke (S)—not by fire (F). The corresponding causal diagram has an arrow from F to S because fire is a direct cause of smoke, and an arrow from S to A because smoke is a direct cause cause of the alarm's sounding. But there is no direct arrow from F to A because although fire causes the alarm to sound, it does so only through the mediator of smoke. If there is a fire but no smoke—for example, if the smoke is all sucked away with a fume hood—then the alarm will not ring. And similarly, if there is smoke, then the alarm will ring regardless of whether there is a fire or not.



If this were only a DGM, with no causal meaning, the graph would only imply that A is independent of F, conditioned on the parents of A, namely S. That is to say, $A \perp\!\!\!\perp F \mid S$. There are other DGMs that convey exactly the same probability information but which are incompatible with the causal structure of this causal model; see Exercise 12.8

A causal diagram alone does not completely explain the causal relationship among variables. A more complete model would also describe the value of each variable as a function of the incoming variables and some random (noise) process. Such a model is called a *functional causal model (FCM)*. The subject of causal diagrams and FCMs is extensive and important. In this section we only give a brief overview of a few of the ideas in the subject of causal inference. For more complete coverage of these ideas, see the references at the end of the section.

[42] A causal diagram is sometimes called a *causal Bayesian network*, despite the fact that it has no special connection to Bayesian statistics.

Nota Bene 12.3.3. Just as with a DGM, most of the information in a causal diagram is in the missing arrows, rather than in the arrows that are present. The presence of an arrow from X to Y does not mean X causes Y, but rather only that X might cause Y. But the absence of an arrow from Z to W means that Z does not directly cause W (it might indirectly cause W if Z causes an ancestor of W).

Vista 12.3.4. Although we focus on directed causal diagrams (on DAGs) in this section, there is also a more general theory that includes causal diagrams with cycles and bidirected edges.

12.3.2 Constructing and Testing Causal Diagrams

When constructing a causal diagram one must carefully consider not only the conditional independence of the variables, but also which other variables might affect a variable and which others cannot affect it. Data alone cannot always distinguish between two causal graphs if their corresponding DGMs encode the same independence relations. To construct a causal diagram we must consider not only the data themselves but also the processes by which the data were generated. This is an important enough principle that it deserves its own red box.

Nota Bene 12.3.5. In order to construct a causal diagram or to draw conclusions about causality, one must consider not just the data themselves, but also how the data are generated.

Considering the data-generation process and constructing a good causal diagram is more work than constructing a DGM, but the payoff is that the causal diagram can answer more questions. A DGM can only tell us probabilities, but a causal diagram can answer questions about interventions (what happens if I do X) and counterfactuals (what would have been if...).

Despite the need to consider the additional causal meaning when constructing a causal diagram, most of the time when we construct DGMs we use causal factors to guide our thinking about independence. Moreover, there seems to be some evidence from psychology that people usually tend to ignore probabilistic information entirely and only think in terms of causality [Pea09, p.22]. Thus, in some sense causal diagrams may be more fundamental to human reasoning than DGMs.

Although constructing a causal diagram usually needs information about the data-generating process and not just the data alone, we can still use data to test some of the predictions of a causal diagram. Because a causal diagram Γ is also a DGM, it imposes some requirements about which random variables should be independent. The property of d-separation tells us which variables in Γ must be independent conditioned on certain other variables, and conditional independence is something we can test for using a data set. Suppose we list the d-separation conditions in Γ, and note that variables A and B must be independent conditioned on $\mathbf{Z}$. Then, suppose we estimate the probabilities based on the data, and discover that the data suggests that A and B are not independent conditional on $\mathbf{Z}$. We can then reject Γ as a possible causal model for the data.

12.3.3 Interventions

Setting X to a particular value x without regard to other factors that might otherwise have influenced X gives a different probability distribution than just observing a case that $X = x$. For example, observing that a person's fingernails (N) have turned yellow implies a higher probability that they are a heavy smoker (S) and hence have a higher probability of lung cancer (C), but dyeing a person's fingernails yellow has no impact on their probability of lung cancer. We denote the process of setting a variable N to a value *yellow* by $\mathbf{do}(N = \text{yellow})$. The fact that $\mathbf{do}(N = \text{yellow})$ has a different impact on cancer risk than just observing $N = \text{yellow}$ can be written symbolically as

$$P(C \,|\, N = \text{yellow}) \neq P(C \,|\, \mathbf{do}(N = \text{yellow})).$$

Throughout this section we use the common notational convenience of writing lower case roman letters to denote that the random variable with the corresponding uppercase name is equal to the lowercase value, so $P(x_i)$ means $P(X_i = x_i)$ and $P(y \,|\, x)$ means $P(Y = y \,|\, X = x)$. This also applies to a set of variables like the parents $\mathrm{PAR}(X_i)$ of X_i; so in a graph G with nodes X_i, if values $x_1, \ldots x_n$ are given, then we write $P(x_j \,|\, \mathrm{par}(x_i))$ to mean $P(X_j = x_j \,|\, \mathrm{PAR}(X_i) = \mathrm{par}(x_i))$, where $\mathrm{par}(x_i)$ are the (lowercase) values of parent nodes $\mathrm{PAR}(X_i)$ of X_i. We also extend this notation to the case of the $\mathbf{do}$ operator, so that $P(y \,|\, \mathbf{do}(x))$ means $P(Y = y \,|\, \mathbf{do}(X = x))$. The following theorem is fundamental to computing the result of an intervention in a causal diagram because it gives the probability $P(x_1, \ldots, x_n \,|\, \mathbf{do}(x_i))$ in terms of the original probabilities in the model (not involving any $\mathbf{do}$ statements).

Theorem 12.3.6 ([Pea09]). *In a causal diagram Γ with nodes $X_1, \ldots, X_n$ and joint distribution $P(X_1, \ldots, X_n)$, the result of doing $X_i = x_i$ on the joint distribution is*

$$P(x_1, \ldots, x_n \,|\, \boldsymbol{do}(x_i)) = \frac{P(x_1, \ldots, x_n)}{P(x_i \,|\, \mathrm{par}(x_i))} = \prod_{j \neq i} P(x_j \,|\, \mathrm{par}(x_j)) \tag{12.3}$$

Graphically, the theorem says that the effect of $\mathbf{do}(X_i = x_i)$ amounts to mutilating the graph Γ by removing all the edges from the parents of X_i into X_i and setting $P(X_i = x_i) = 1$. Note, however, that the probabilities on the right side of (12.3) are all *preintervention*; that is, they use the original probabilities from the original model before doing $X_i = x_i$.

Corollary 12.3.7. *If X and Y are random variables in a causal diagram Γ and $\mathrm{PAR}(X)$ are the parents of X, then*

$$P(y \,|\, \boldsymbol{do}(x)) = \sum_{\mathrm{par}} \frac{P(x, y, \mathrm{par})}{P(x \,|\, \mathrm{par})}, \tag{12.4}$$

where the sum runs over all values par *that the variables* $\mathrm{PAR}(X)$ *could take. If X has no parents, then*

$$P(y \,|\, \boldsymbol{do}(x)) = \frac{P(x, y)}{P(x)} = P(y \,|\, x). \tag{12.5}$$

Proof. Let $\mathbf{Z}$ denote all the variables in Γ except X, Y, and $\mathrm{PAR}(X)$. For any choice of values x, y, $\mathbf{z}$, and par of X, Y, $\mathbf{Z}$, and $\mathrm{PAR}(X)$, respectively, (12.3) gives

$$P(x, y, \mathbf{z}, \mathrm{par} \,|\, \mathbf{do}(x)) = \frac{P(x, y, \mathbf{z}, \mathrm{par})}{P(x \,|\, \mathrm{par})}.$$

Marginalizing to find $P(y \,|\, \mathbf{do}(x))$ gives

$$\begin{aligned} P(y, \,|\, \mathbf{do}(x)) &= \sum_{\mathbf{z}, \mathrm{par}} P(x, y, \mathbf{z}, \mathrm{par} \,|\, \mathbf{do}(x)) \\ &= \sum_{\mathrm{par}} \sum_{\mathbf{z}} \frac{P(x, y, \mathbf{z}, \mathrm{par})}{P(x \,|\, \mathrm{par})} \\ &= \sum_{\mathrm{par}} \frac{P(x, y, \mathrm{par})}{P(x \,|\, \mathrm{par})}. \end{aligned}$$

If X has no parents, then

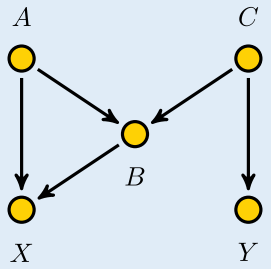
$$P(x, y, \mathbf{z} \,|\, \mathbf{do}(x)) = \frac{P(x, y, \mathbf{z})}{P(x)} = P(y, \mathbf{z} \,|\, x),$$

and

$$\begin{aligned} P(y \,|\, \mathbf{do}(x)) &= \sum_{\mathbf{z}} P(x, y, \mathbf{z} \,|\, \mathbf{do}(x)) \\ &= \sum_{\mathbf{z}} \frac{P(x, y, \mathbf{z})}{P(x)} \\ &= \frac{P(x, y)}{P(x)} = P(y \,|\, x). \qquad \square \end{aligned}$$

When using (12.5) we say that we *control for* the possible values of the parents $\mathrm{PAR}(X)$ by summing over all possible values par of $\mathrm{PAR}(X)$ and using $P(x \,|\, \mathrm{par})$ in the denominator.

Example 12.3.8. Consider the following causal diagram.



In this diagram B has parents A and C, so for any values x and b Corollary 12.3.7 gives

$$\begin{aligned} P(X=x \,|\, \mathbf{do}(B=b)) &= \sum_a \sum_c \frac{P(X=x, A=a, B=b, C=c)}{P(B=b \,|\, A=a, C=c)} \\ &= \sum_a \sum_c P(X=x \,|\, A=a, B=b) P(C=c) \\ &= \sum_a P(X=x \,|\, A=a, B=b). \end{aligned}$$

where the first two sums run over all possible values of a and c, and the last sum runs over all possible values of a. The second equality follows from the standard factorization (12.1) of the joint distribution in the DGM.

In contrast to that example, consider the node A, which has no parents. For any values a and b Corollary 12.3.7 now gives

$$P(B=b \,|\, \mathbf{do}(A=a)) = P(B=b \,|\, A=a).$$

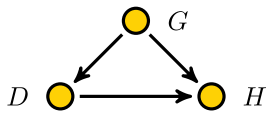
12.3.4 Example: Simpson's Paradox Part 1

The following example is due to [PM18]] Consider the following fictitious data about a drug intended to reduce heart attacks.

	Control (no drug)		Treatment (took drug)	
	Heart attack	No heart attack	Heart attack	No heart attack
Female	1	19	3	37
Male	12	28	8	12
Total	13	47	11	49

Notice that in the control group $\frac{13}{60} \approx 22\%$ had heart attacks and in the treatment group $\frac{11}{60} \approx 18\%$ had heart attacks. That makes it seem like taking the drug might help prevent heart attacks. But breaking it down by gender shows that in the female control group $\frac{1}{20} = 5\%$ had heart attacks and in the female treatment group $\frac{3}{40} = 7.5\%$ had heart attacks. So taking the drug seems to have made things worse for women. Similarly, for the male control group $\frac{12}{40} = 30\%$ had heart attacks, but in the male treatment group $\frac{8}{20} = 40\%$ had heart attacks. So this drug appears to be bad for men, bad for women, and good for people overall. Of course that makes no sense. This is an example of *Simpson's paradox*, where an effect is seen in a group as a whole but the opposite effect is seen for every subgroup.

In this example the solution is to consider a causal diagram, because we want to understand is the effect of taking the drug ($\mathbf{do}(D = 1)$) on the probability of a heart attack (H); that is, we want to compute $P(H \mid \mathbf{do}(D))$. We expect the drug to have a direct causal effect on heart attacks (H), so there should be an arrow from D to H. Gender (G) of the subject is also important. Female subjects have many fewer heart attacks than male subjects do, and so gender appears to be a direct cause of heart attack, corresponding to an arrow from G to H. But we also need to consider the possibility of arrows between D and G. To do that we must consider the data generation process. This study was observational, not randomized, so patients and their physicians decided whether the patient should take the drug, presumably based on things like the patients' risk of heart attack and the patients' own willingness to risk a heart attack. That suggests that gender probably played a role in whether the patients chose to take the drug. That is, we should have an arrow from G to D, giving the following causal diagram.



To compute $P(H = 1 \mid \mathbf{do}(D))$ we use (12.5). The fact that the only parent of D is G means we must sum over the possible values g of G and use $P(D = 1 \mid G = g)$ in the denominator.

$$
\begin{aligned}
P(H = 1 \mid \mathbf{do}(D = 1)) &= \sum_g \frac{P(H = 1, G = g, D = 1)}{P(D = 1 \mid G = g)} \\
&= \frac{P(H = 1, G = f, D = 1)}{P(D = 1 \mid G = f)} + \frac{P(H = 1, G = m, D = 1)}{P(D = 1 \mid G = m)} \\
&= \frac{\frac{3}{120}}{\frac{40}{60}} + \frac{\frac{8}{120}}{\frac{20}{60}} = 0.2475
\end{aligned}
$$

and

$$
\begin{aligned}
P(H=1 \,\big|\, \mathbf{do}(D=0)) &= \sum_g \frac{P(H=1,G=g,D=0)}{P(D=0 \,\big|\, G=g)} \\
&= \frac{P(H=1,G=f,D=0)}{P(D=0 \,\big|\, G=f)} + \frac{P(H=1,G=m,D=0)}{P(D=0 \,\big|\, G=m)} \\
&= \frac{\frac{1}{120}}{\frac{20}{60}} + \frac{\frac{12}{120}}{\frac{40}{60}} = 0.175
\end{aligned}
$$

This shows that taking the drug causes heart attacks.

12.3.5 Example: Simpson's Paradox Part 2

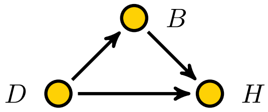
This example is also due to [PM18].

Consider a situation of a drug intended to reduce heart attacks by reducing blood pressure. The drug was only given to patients with high blood pressure, but in some it reduced blood pressure and in others it did not. We'll use the same data as in the example of Section 12.3.4, but where gender is replaced by the subjects' blood pressure status after taking the drug.

	Control (no drug)		Treatment (took drug)	
	Heart attack	No heart attack	Heart attack	No heart attack
Low BP	1	19	3	37
High BP	12	28	8	12
Total	13	47	11	49

Since we are using the same data, we have the same paradox, where the drug appears to be good for patients overall but bad for patients in the low blood pressure group and bad for patients in the high blood pressure group.

Again, we use a causal diagram to resolve the paradox, but the diagram is different. The drug might have a direct causal effect on heart attacks, so we have an arrow from D to H, but because we believe it works by reducing blood pressure B, we also expect an arrow from D to B and an arrow from B to H. A subject's blood pressure status after taking the drug did not influence the choice of whether a patient took the drug or not, so there is no arrow from B to D.



There are no parents of D, so we need not control for any variables. Equation (12.5) gives

$$
\begin{aligned}
P(H=1 \,\big|\, \mathbf{do}(D=1)) &= \frac{P(H=1,D=1)}{P(D=1)} \\
&= P(H=1 \,\big|\, D=1) \\
&= \frac{11}{120} = 0.0917
\end{aligned}
$$

and

$$\begin{aligned} P(H=1 \,\big|\, \mathbf{do}(D=0)) &= P(H=1 \,\big|\, D=0) \\ &= \frac{13}{120} = 0.1083 \end{aligned}$$

This shows that taking the drug causes a reduction in heart attacks. Comparing this to the example in Section 12.3.4 shows that the same data generated in different ways can give different causal diagrams, which can result in very different conclusions about $P(H=1 \,\big|\, \mathbf{do}(D))$.

Until the development of causal diagrams it was common in statistics to control for everything that might have something to do with the final effect being studied. In the situation of the example in Section 12.3.4, that turns out to be the right thing to do. But in the situation of the example in Section 12.3.5 it is wrong to control for blood pressure.

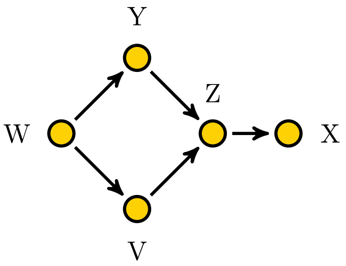
Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

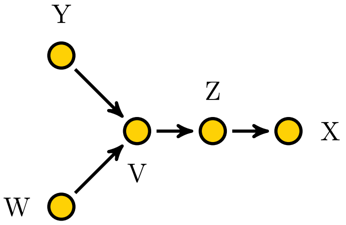
You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

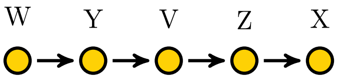
12.1. Which of the following DAGs represent the conditional independence rules given in Example 12.1.5? Justify your answer. Hint: There may be more than one because the list of rules in Example 12.1.5 is not guaranteed to be complete, so each DAG may also imply more conditional independence rules than those listed.



(a.)

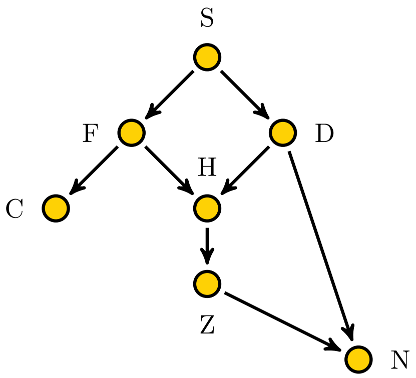


(b.)



(c.)

12.2. Consider the following DGM describing relationships among the following variables: Season (S), Flu (F), Dehydration (D), Chills (C), Headache (H), Nausea (N), Dizziness (Z):



Indicate which of the following statements is guaranteed true and which could possibly be false according to this model. Provide a brief justification for each answer.

(i) $S \perp\!\!\!\perp C$

(ii) $S \perp\!\!\!\perp C \,\big|\, F$

(iii) $S \perp\!\!\!\perp H \,\big|\, F$

(iv) $S \perp\!\!\!\perp H \,\big|\, F, D$

(v) $S \perp\!\!\!\perp N \,\big|\, D$

(vi) $S \perp\!\!\!\perp N \,\big|\, D, Z$

(vii) $F \perp\!\!\!\perp D$

(viii) $F \perp\!\!\!\perp D \,\big|\, S$

(ix) $C \perp\!\!\!\perp N$

(x) $C \perp\!\!\!\perp N \,\big|\, H$

12.3. Finish the proof of Theorem 12.1.11 by showing that if a graph G represents P, then P factors as (12.1). Hint: It may help to put a topological ordering on the nodes of the graph.

12.4. Assume your house has an alarm system against burglary. You live in a seismically active area and the alarm system can occasionally be set off by an earthquake. You have two neighbors, Mary and John, who do not know each other and do not interact. If they hear the alarm they call you, but they might not hear the alarm. Let A be the binary random variable indicating whether the alarm rings, E the binary random variable of an earthquake occurring or not, B the binary random variable of a burglary occurring, and J and M the binary random variables of whether John or Mary, respectively, calls you. Assume the conditional probabilities for this model are as follows:

$$\begin{aligned}
&P(B=1)=0.002\\
&P(E=1)=0.001\\
&P(A=1 \mid B=1, E=1)=0.95\\
&P(A=1 \mid B=1, E=0)=0.94\\
&P(A=1 \mid B=0, E=1)=0.29\\
&P(A=1 \mid B=0, E=0)=0.001\\
&P(J=1 \mid A=1)=0.9\\
&P(M=1 \mid A=1)=0.7\\
&P(J=1 \mid A=0)=0.05\\
&P(M=1 \mid A=0)=0.01
\end{aligned}$$

(i) Draw a DGM for this situation.

(ii) Compute the following probabilities:

(a) $P(J=1)$.

(b) $P(J=1 \mid B=1)$.

(c) $P(B=1) \mid J=1, M=0)$

(d) $P(E=1 \mid A=1)$

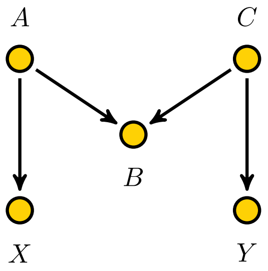
(e) $P(B=1 \mid A=1)$

(f) $P(E=1, B=1 \mid A=1)$

(iii) What do the last three probability computations say about conditional independence of E and B given A?

(iv) Why doesn't that require an edge between B and E in the DGM?

12.5. For the given graph, which of the following sets $\mathbf{Z}$ will d-separate X from Y given $\mathbf{Z}$? Mark all that apply.



(i) $\mathbf{Z} = \emptyset$

(ii) $\mathbf{Z} = \{A\}$

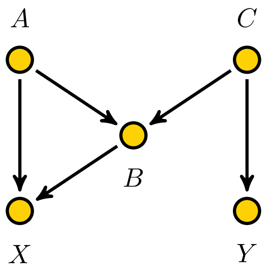
(iii) $\mathbf{Z} = \{B\}$

(iv) $\mathbf{Z} = \{C\}$

(v) $\mathbf{Z} = \{A, B\}$

(vi) $\mathbf{Z} = \{B, C\}$

12.6. For the given graph, identify every possible subset $\mathbf{Z}$ of $\{A, B, C\}$ that would d-separate X from Y.



12.7. Collider bias: Consider a bivariate random variable (X, Y) distributed as $\mathscr{N}(\mathbf{0}, I)$ with mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance $\Sigma = I$.

(i) Show that the variables X and Y are independent, by factoring the joint p.d.f. of (X, Y) as a product of the univariate p.d.f.s for X and Y.

(ii) Let Z be the random variable $Z = X + Y$. We have $Y = Z - X$, so for a given value of $Z = z$, we have $Y = z - X$, so X and Y not independent. We are interested in the case of $Z \geq 0$. Show that the conditional p.d.f. satisfies $f_{X,Y|Z\geq 0}(x, y) \propto f_{X,Y}(x, y)\mathbb{1}_{x+y\geq 0}(x, y)$, which does not factor into a product of univariate p.d.f.s.

(iii) Draw 10^5 times from $\mathscr{N}(\mathbf{0}, I)$ and scatterplot the results.

(iv) Discard any samples for which $Z = X + Y$ is less than 0 and scatterplot the results. This is a draw from $(X, Y \mid Z \geq 0)$.

(v) Compute the sample covariance between X and Y for the remaining points.

(vi) Estimate $\mathbb{E}[X \mid Z \geq 0]$ and $\mathbb{E}[Y \mid Z \geq 0]$ using sample means.

If the points represent men, with X corresponding to niceness and Y to handsomeness, then $Z \geq 0$ could represent Alice's willingness to date. Conditioning on $Z \geq 0$ causes X and Y to become negatively correlated. Note that despite the negative correlation, the average niceness in Alice's dating pool is higher than the average among all men, and the average handsomeness in her dating pool is also higher than average.

12.8. Construct a DGM for the fire alarm example (12.3.2) that encodes the same probabilistic information as the DGM given in that example but which is incompatible with the causal assumption that alarms don't cause smoke. What causal assumptions does your model encode?

12.9. When studying public health data, it is common to sort babies by birthweight because those that have low birthweight tend to have much higher mortality than those with normal birthweight. Surprisingly, babies of smoking mothers who have low birthweight have better survival than those babies of low birthweight with nonsmoking mothers, and similarly, babies of normal birthweight with smoking mothers also tend to have better survival than those of normal birthweight and nonsmoking mothers. Despite all this, babies of smoking mothers tend to have higher mortality than babies of nonsmoking mothers. To summarize, we have

$$\begin{aligned}&P(D|L,S) < P(D|L,S^c) \text{ and } P(D|L^c,S) < P(D|L^c,S^c)\\ &\text{but } P(D|S) > P(D|S^c). \end{aligned} \tag{12.6}$$

(i) Use the law of total probability to write $P(D|S)$ in terms of $P(D|L,S)$ and $P(D|L^c,S)$, and similarly write $P(D|S^c)$ in terms of $P(D|L,S^c)$ and $P(D|L^c,S^c)$. Use this to explain mathematically what is happing in this paradox.

(ii) Low birthweight may also be caused by some other abnormality, independent of smoking. Writing S for the event that the mother is a smoker, A for the event of another abnormality, and L for the event that a baby has low birthweight, and D for the event of the baby's death, draw a causal diagram to model this situation.

(iii) According to your causal diagram and (12.5), write $P(D \mid \mathbf{do}(S))$ and $P(D \mid \mathbf{do}(S^c))$ in terms of the preintervention probabilities (12.6) in part (i) of this exercise.

12.10. The data in the following table show the number of patients observed with bone disease B or respiratory disease R and their recent hospitalization status H, where $H = 1$ denotes a hospitalization in the last six months and $H = 0$ denotes no recent hospitalization, $R = 1$ denotes the presence of respiratory disease, and $B = 1$ denotes the presence of bone disease.

	$H=0$		$H=1$	
	$B=1$	$B=0$	$B=1$	$B=0$
$R=1$	17	207	5	15
$R=0$	184	2,376	18	219

(i) Compute observed probabilities $P(B = 1 \mid R = 1, H = 0)$ and $P(B = 1 \mid R = 0, H = 0)$. What does this suggest about the correlation of bone and respiratory disease in the general population?

(ii) Compute observed probabilities $P(B = 1 \mid R = 1, H = 1)$ and $P(B = 1 \mid R = 0, H = 1)$. What does this suggest about the correlation of bone and respiratory disease in the recently hospitalized population?

(iii) Of course, most people do not believe that being hospitalized makes respiratory patients more likely to develop a bone disease. Draw a causal diagram for R, B and H modeling our common beliefs about this situation.

(iv) Compute $P(B = 1)$ and $P(R = 1)$ and $P(B = 1, R = 1)$. Are these results compatible with your diagram?

(v) Compute the probability $P(B = 1 \mid \mathbf{do}(H = 1))$ and $P(R = 1 \mid \mathbf{do}(H = 1))$ and $P(B = 1, R = 1 \mid \mathbf{do}(H = 1))$. Are these results compatible with your diagram?

12.11. One important way to deal with unknown causes or with causes on which it is hard to condition, is the randomized controlled trial (RCT). As an example, consider the situation where we have a treatment T that we believe will improve the subject's health H, but there are many possible confounders that may affect whether the subject chooses to have the treatment applied or not. For example, if they don't believe the treatment is useful they may not choose to participate. If the treatment is expensive and they can't afford it, they may choose not to participate.

(i) Belief state may also be correlated to the subject's health via the placebo effect. And financial considerations may have an impact on the subject's health because they can't afford to take time off work to exercise or for other reasons. Denoting health by H, belief state by B, financial considerations by F, and the application or not of treatment by T, draw a causal diagram modeling this situation.

(ii) To compute the effect of $\mathbf{do}(T = 1)$ on H (12.5) requires that we condition on the parents of T, but it may not be easy to do that, for example when people don't openly admit they can't afford treatment. An RCT randomly assigns the treatment to patients and thereby breaks the causal arrows from B into T and F into T and instead creates a new, unique parent of T, namely the random variable R that is truly random. Draw a causal diagram for this situation with an RCT.

(iii) Write the formula for $P(H \mid \mathbf{do}(T))$ in the case of the RCT and explain why this is always easier to compute than $P(H \mid \mathbf{do}(T))$ without the RCT.

Notes

13 Latent Variable Models

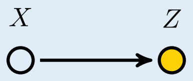
13.1 Latent Variable Models

A latent variable model for a random variable Z arises when the probability $P(Z)$ is actually a marginal distribution of $P(Z, X)$, where X is an unobserved random variable.

Example 13.1.1. Imagine we have two coins that appear identical, with probabilities θ_A and θ_B, respectively, of coming up heads. Randomly (with probability $\frac{1}{2}$) choose one of the coins. Let the random variable X be the choice of coin A or B. After choosing the coin, flip it ten times, and let Z be the total number of heads observed. For convenience, we use $\boldsymbol{\theta}$ to denote the pair (θ_A, θ_B). When $\boldsymbol{\theta}$ and X are known, the probability $P(Z|X, \boldsymbol{\theta})$ is easy to compute:

$$\begin{gathered} P(Z = z \,\big|\, X = A, \boldsymbol{\theta}) = \binom{10}{z} \theta_A^z (1 - \theta_A)^{10-z} \\ \text{and} \\ P(Z = z \,\big|\, X = B, \boldsymbol{\theta}) = \binom{10}{z} \theta_B^z (1 - \theta_B)^{10-z}. \end{gathered} \tag{13.1}$$

The DGM for this model is



We leave the node X unfilled to indicate that it is unobserved.

Assuming that the hidden variable X takes only discrete values, and that X and Z depend on parameter θ, we may write

$$\begin{aligned} P(Z \,|\, \theta) &= \sum_x P(Z, X = x \,|\, \theta) \\ &= \sum_x P(Z \,|\, X = x, \theta) P(X = x \,|\, \theta). \end{aligned}$$

Example 13.1.2. Continuing with the same setup as Example 13.1.1, the probability $P(Z = z \,|\, \boldsymbol{\theta})$ is the sum

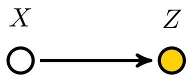
$$\begin{aligned} P(Z = z \,|\, \boldsymbol{\theta}) &= P(Z = z \,|\, X = A, \boldsymbol{\theta}) P(X = A \,|\, \boldsymbol{\theta}) \\ &\quad + P(Z = z \,|\, X = B, \boldsymbol{\theta}) P(X = B \,|\, \boldsymbol{\theta}) \\ &= \frac{1}{2}(P(Z = z \,|\, X = A, \boldsymbol{\theta}) + P(Z = z \,|\, X = B, \boldsymbol{\theta})) \end{aligned}$$

The probability $P(X \,|\, Z, \boldsymbol{\theta})$ can be computed by Bayes' rule

$$P(X \,|\, Z = z, \boldsymbol{\theta}) = \frac{P(Z = z \,|\, X, \boldsymbol{\theta}) P(X \,|\, \boldsymbol{\theta})}{P(Z = z \,|\, \boldsymbol{\theta})}.$$

13.1.1 Mixture Models

The two-coin situation of Examples 13.1.1 and 13.1.2 is an example of a *mixture model*, which is a latent variable model with DGM



where the latent variable X has a categorical distribution $X \sim \mathrm{Cat}(\mathbf{w})$, with $\mathbf{w} = (w_1, \ldots, w_K)$ and $\sum_{k=1}^K w_k = 1$. The conditional distributions $p(Z = \mathbf{z} \,|\, X = k, \boldsymbol{\theta})$, which we write as $p_k(Z = \mathbf{z} \mid \boldsymbol{\theta})$ can be arbitrary. The overall model is called a mixture model because it is a convex combination (mixture) of the distributions p_k:

$$P(Z = \mathbf{z} \,|\, \boldsymbol{\theta}) = \sum_{k=1}^K w_k p_k(Z = \mathbf{z} \,|\, \boldsymbol{\theta}). \tag{13.2}$$

To draw from a mixture model, one first draws $X = k$ from $\mathrm{Cat}(\mathbf{w})$ and then draws z from p_k.

Mixture models greatly expand our universe of parametrized distributions, allowing us to construct models that are considerably more intricate than the standard list of named distributions.

Binomial Mixture Models

The two-coin example is a mixture of binomials model with $w_A = w_B = \frac{1}{2}$, where p_A and p_B are binomial, as given in (13.1).

Gaussian Mixture Models

A mixture of Gaussians is called a *Gaussian mixture model (GMM)*. In this case each $p_k(\mathbf{z} \,|\, \boldsymbol{\theta})$ is Gaussian

$$p_k(\mathbf{z} \,|\, \boldsymbol{\theta}) = \mathscr{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_k, \Sigma_k) = |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma_k^{-1}(\mathbf{z}-\boldsymbol{\mu}_k)}$$

and

$$p(\mathbf{z} \,|\, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \mathscr{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_k, \Sigma_k).$$

As in the binomial mixture, to draw from this GMM one first draws $X = k$ from $\text{Cat}(\mathbf{w})$ and then draws z from $\mathscr{N}(\boldsymbol{\mu}_k, \Sigma_k)$.

13.1.2 Mixture of Experts

OLS is a great tool, but sometimes the basic Gaussian linear regression model of the form

$$P(y \,|\, \mathbf{x}) = \mathscr{N}(y \,|\, \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}, \sigma^2)$$

only describes the random variable y for certain values of $\mathbf{x}$, but for other values of $\mathbf{x}$ a different model should apply. For example there might be a phase change (such as water turning to ice or steam), and one regression model works well in one phase but a different model is needed to describe a different phase. This can be modeled with what is called a *mixture of experts*, where there is a discrete latent variable Z (the phase, for example), which is partially determined by $\mathbf{x}$, say by

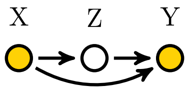
$$P(z \,|\, \mathbf{x}, \boldsymbol{\theta}) = \text{Cat}(\mathscr{S}(V^{\mathsf{T}}\mathbf{x})),$$

where $\mathscr{S}$ is the softmax function. For each value z of Z there is a different choice of parameters $(\boldsymbol{\beta}_z, \sigma_z^2)$ giving

$$P(y \,|\, \mathbf{x}, z) = \mathscr{N}(y \,|\, \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_z, \sigma_z^2).$$

The idea is that each submodel $(\boldsymbol{\beta}_z, \sigma_z^2)$ is the expert in the region of the input space where $P(z \,|\, \mathbf{x}, \boldsymbol{\theta})$ is maximal; see Figure 13.1.

The directed graphical model for the mixture of experts is



Again, Z is unfilled to indicate that it is unobserved. The arrow from X to Z denotes that X and Z are not independent, and the two arrows into Y denote that Y is not independant from either X or Z.

As with a GMM, a mixture of experts can be trained using expectation maximization.

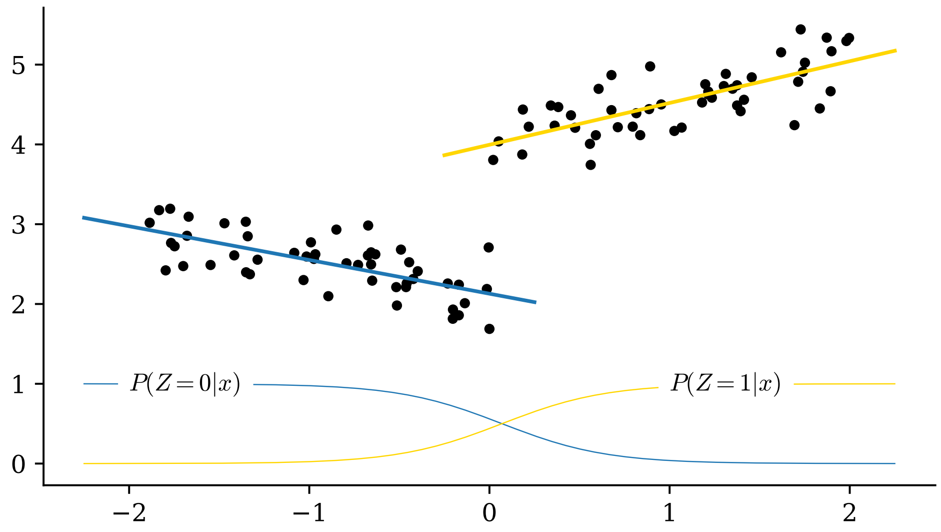
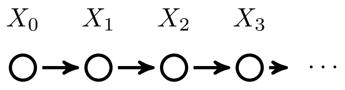


Figure 13.1: An example where a good linear fit to the data in one region is not a good fit in another region. Here the data (black) are modeled well by the heavy blue line when $x < 0$ and are modeled well by the heavy yellow line when $x > 0$. A mixture of experts model involves a latent variable Z (which line to use) influenced by X and influencing Y. The mixture of experts model plotted here uses x (horizontal axis) to choose $z \in \{0, 1\}$ according to $P(z \mid x)$, and then predicts y (vertical axis) on the blue line if $Z = 0$ or predicts y on the yellow line if $Z = 1$. The probabilities $P(Z = 0 \mid x)$ and $P(Z = 1 \mid x)$ are plotted as functions of x along the bottom of this figure as the thin blue and yellow curves, respectively.

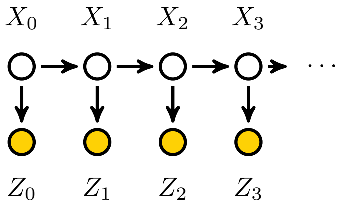
13.1.3 Hidden Markov Models

Consider the following example, due to Mark Stamp [Sta17, Chapter 2]. Imagine you are trying to deduce historical weather information from the rings in an old tree. In years when the weather was wetter, the tree tended to grow more, and when the weather was drier, the tree grew less. We cannot observe the weather directly—we can only observe the ring size (say small, medium, or large). Assuming that the weather is Markov (each year's weather is influenced by the previous year's weather, but is conditionally independent of earlier years' weather), this situation can be modeled with an important latent variable model called a *hidden Markov model (HMM).*

In an HMM the latent variables form a Markov chain (in our example, this is the weather—dry or wet), denoted graphically as



The model also has observables $Z_0, Z_1, \ldots$ (in our example, these are the size of the tree rings) where each Z_i depends on X_i, but is conditionally independent from all other variables, given X_i. We denote this as



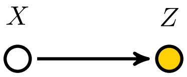
We discuss HMMs in more detail in Section 13.4. As with the GMM and mixture of experts, one of the main tools for working with HMMs is expectation maximization.

13.2 Expectation Maximization

Recall that the maximum likelihood estimation (MLE) problem consists of finding the parameter $\boldsymbol{\theta}$ that maximizes the likelihood $L(\boldsymbol{\theta}) = P(\mathbf{D} \,|\, \boldsymbol{\theta})$, where $\mathbf{D} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$ represents the data. This is equivalent to maximizing the log-likelihood

$$\ell(\boldsymbol{\theta}) = \log P(\mathbf{D} \,|\, \boldsymbol{\theta}),$$

If the data $\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_N$ are i.i.d., then $P(\mathbf{D} \,|\, \boldsymbol{\theta})$ factors as $P(\mathbf{D} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{N} P(Z = \mathbf{z}_i \,|\, \boldsymbol{\theta})$, and if this is a latent variable model with DGM



(13.3)

then

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(\mathbf{z}_i \,|\, \boldsymbol{\theta}) = \sum_{i=1}^{N} \log \sum_{x} P(\mathbf{z}_i, x \,|\, \boldsymbol{\theta}),$$

where x runs over all possible values of the latent variable X. This last sum is usually difficult to compute using standard optimization methods. This difficulty is due, in part, to the fact that the log of the sum makes differentiation messy, at best.

Example 13.2.1. Consider again the two-coin setup of Examples 13.1.1 and 13.1.2. For each coin A or B the number of heads that comes up is binomially distributed with probability θ_A or θ_B, respectively. Imagine that the experiment (that is, a coin has been drawn and flipped ten times, and then replaced) has been repeated five times, giving a sequence of observations $\mathbf{D} = (z_1, \ldots, z_5) = (5, 9, 8, 4, 7)$, but the parameters θ_A, θ_B are unknown.

For each experiment i there is a latent random variable $X_i \in \{A, B\}$. To estimate the values of θ_A and θ_B using maximum likelihood, we must find

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{5} \log P(z_i \,|\, \boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{5} \log \left(\frac{1}{2} (P(Z = z_i \,|\, X_i = A, \boldsymbol{\theta}) + P(Z = z_i \,|\, X_i = B, \boldsymbol{\theta})) \right) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{5} \log(\theta_A^{z_i}(1-\theta_A)^{10-z_i} + (\theta_B^{z_i}(1-\theta_B)^{10-z_i})). \qquad (13.4)\end{aligned}$$

But the obvious approach to maximization, by differentiating (13.4), is messy because of the sum inside the logarithm, even in this relatively simple case with only two hidden states.

In this section we describe an iterative algorithm called *expectation maximization (EM)* for approximating the MLE $\widehat{\boldsymbol{\theta}}$ in latent variable models.

13.2.1 The Expectation Maximization Algorithm

Definition 13.2.2. *The* expectation maximization (EM) *algorithm is an iterative method for finding the maximum likelihood estimate* $\operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ *of the parameter* $\boldsymbol{\theta}$ *in a latent variable model* (13.3)*. Given an estimate* $\boldsymbol{\theta}^t$ *of* $\widehat{\boldsymbol{\theta}}_{MLE}$*, it constructs a new estimate* $\boldsymbol{\theta}^{t+1}$ *by completing the following two steps. Performed repeatedly, it gives a sequence* $(\boldsymbol{\theta}^k)_{k=0}^{\infty}$*.*

Expectation: *Compute*

$$q_i^t(x) = P(X_i = x \,|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t) \qquad (13.5)$$

for all possible latent states x *and all* $i \in \{1, \dots, n\}$*. Note that the superscript* t *here is an index, not an exponent. Use the computed values of* $q_i^t(x)$ *to construct*

$$Q^t(\boldsymbol{\theta}) = \sum_{i=1}^{n} Q_i^t(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{x} q_i^t(x) \log P(X_i = x, Z_i = \mathbf{z}_i \,|\, \boldsymbol{\theta}). \qquad (13.6)$$

Maximization: *Compute*

$$\boldsymbol{\theta}^{t+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q^t(\boldsymbol{\theta}). \qquad (13.7)$$

In the next section we prove the following theorem, which suggests, but does not guarantee, that the sequence constructed by the EM algorithm should converge to a critical point of ℓ.

Theorem 13.2.3. *The EM algorithm gives a sequence $(\boldsymbol{\theta}^t)_{t=0}^{\infty}$ with the property that $(\ell(\boldsymbol{\theta}^t))_{t=0}^{\infty}$ is nondecreasing and bounded above, hence it must converge to some finite value. If, moreover, $Q^t(\boldsymbol{\theta})$ is strictly concave, then $\ell(\boldsymbol{\theta}^{t+1}) > \ell(\boldsymbol{\theta}^t)$, unless $\boldsymbol{\theta}^t$ is a critical point of ℓ.*

Nota Bene 13.2.4. Although Theorem 13.2.3 guarantees that the sequence $(\ell(\boldsymbol{\theta}^t))_{t=0}^{\infty}$ converges, it does *not* guarantee that the sequence of parameters $(\theta^t)_{t=0}^{\infty}$ converges. Moreover, although the sequence $(\ell(\boldsymbol{\theta}^t))_{t=0}^{\infty}$ is nondecreasing and in the strictly concave case can only stop increasing if it reaches a stationary point (where the derivative of ℓ vanishes), it does not guarantee that the limit point $\lim_{t\to\infty} \ell(\theta^t)$ is a stationary point, and thus does not guarantee that the limit point is a local maximum of ℓ.

Example 13.2.5. Returning to the two-coin example (see Examples 13.1.1–13.2.1), we wish to find $\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, where

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{5} \log P(z_i \,|\, \boldsymbol{\theta}) = \sum_{i=1}^{5} \log(\theta_A^{z_i}(1-\theta_A)^{10-z_i} + (\theta_B^{z_i}(1-\theta_B)^{10-z_i}))$$

and $\mathbf{D} = (z_1, \ldots, z_5) = (5, 9, 8, 4, 7)$.

To approximate $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_A, \widehat{\theta}_B)$ using the EM algorithm, assume that $\boldsymbol{\theta}^t = (\theta_A^t, \theta_B^t)$ is given (computed or guessed). The expectation step is to compute $q_i^t(x)$ for each $x \in \{A, B\}$ and each $i \in \{1, \ldots, 5\}$. This can be done with Bayes' rule, using the fact that $P(X_i = x \,|\, \boldsymbol{\theta}^t) = \frac{1}{2}$ and

$$P(Z_i = z_i \,|\, X_i = x, \boldsymbol{\theta}^t) = \binom{10}{z_i} (\theta_x^t)^{z_i} (1 - \theta_x^t)^{10-z_i},$$

For example, in the case of $x = A$, we have

$$\begin{aligned} q_i^t(A) &= P(X_i = A \,|\, Z_i = z_i, \boldsymbol{\theta}^t) \\ &= \frac{P(Z_i = z_i \,|\, X_i = A, \boldsymbol{\theta}^t) P(X_i = A \,|\, \boldsymbol{\theta}^t)}{P(Z_i = z_i \,|\, \boldsymbol{\theta}^t)} \\ &= \frac{\binom{10}{z_i}(\theta_A^t)^{z_i}(1-\theta_A^t)^{10-z_i}(\frac{1}{2})}{\binom{10}{z_i}(\theta_A^t)^{z_i}(1-\theta_A^t)^{10-z_i}(\frac{1}{2}) + \binom{10}{z_i}(\theta_B^t)^{z_i}(1-\theta_B^t)^{10-z_i}(\frac{1}{2})} \\ &= \frac{(\theta_A^t)^{z_i}(1-\theta_A^t)^{10-z_i}}{(\theta_A^t)^{z_i}(1-\theta_A^t)^{10-z_i} + (\theta_B^t)^{z_i}(1-\theta_B^t)^{10-z_i}}. \end{aligned}$$

Computing $q_i^t(B)$ is similar.

The maximization step is to compute

$$\begin{aligned}\boldsymbol{\theta}^{t+1} &= \operatorname*{argmax}_{\boldsymbol{\theta}} Q^t(\boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} Q_i^t(\boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{x} q_i^t(x) \log P(X_i = x, Z_i = z_i \,\big|\, \boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{x} q_i^t(x) \log(\binom{10}{z_i} \theta_x^{z_i}(1-\theta_x)^{10-z_i})\end{aligned}$$

Setting the derivatives $\frac{\partial Q}{\partial \theta_A}$ and $\frac{\partial Q}{\partial \theta_B}$ to zero and solving these for θ_A and θ_B gives (see Exercise 13.5)

$$\theta_A^{t+1} = \frac{\sum_{i=1}^n q_i^t(A) z_i}{\sum_{i=1}^n 10 q_i^t(A)} \qquad \text{and} \qquad \theta_B^{t+1} = \frac{\sum_{i=1}^n q_i^t(B) z_i}{\sum_{i=1}^n 10 q_i^t(B)}. \tag{13.8}$$

Repeat the process until the sequence seems to converge.

Remark 13.2.6. As with most iterative optimization routines, a bad initial guess can get you bad results. It's generally best, if possible, to choose an initial value that is relatively close to the true optimizer. It's also wise to avoid initial values with a lot of symmetry (for example, choosing $\theta_A = \theta_B$ in Example 13.2.5), including special values like those lying on the axes or at the origin (choosing θ_A or θ_B in $\{0, 1\}$ in Example 13.2.5 would generally be worse than choosing random values in $(0, 1)$).

Example 13.2.7. Consider a variant of Example 13.2.5 where the probabilities $\pi_A = P(A)$ and $\pi_B = P(B)$ of choosing the coins are unknown. In this case we wish to find $\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_A, \pi_A, \theta_B, \pi_B)$.

Assume that $\boldsymbol{\theta}^t = (\theta_A^t, \pi_A^t, \theta_B^t, \pi_B^t)$ is given (computed or guessed). As before, we compute the expectation step $q_i^t(x)$ using Bayes rule, but now $P(X_i = x \,\big|\, \boldsymbol{\theta}^t) = \pi_x$. For example, in the case of $x = A$, we have

$$\begin{aligned} q_i^t(A) &= P(X_i = A \,\big|\, Z_i = z_i, \boldsymbol{\theta}^t) \\ &= \frac{P(Z_i = z_i \,\big|\, X_i = A, \boldsymbol{\theta}^t) P(X_i = A \,\big|\, \boldsymbol{\theta}^t)}{P(Z_i = z_i \,\big|\, \boldsymbol{\theta}^t)} \\ &= \frac{\binom{10}{z_i}(\theta_A^t)^{z_i}(1-\theta_A^t)^{10-z_i}\pi_A}{\binom{10}{z_i}(\theta_A^t)^{z_i}(1-\theta_A^t)^{10-z_i}\pi_A + \binom{10}{z_i}(\theta_B^t)^{z_i}(1-\theta_B^t)^{10-z_i}\pi_B}. \end{aligned}$$

Computing $q_i^t(B)$ is similar.

The maximization step is to compute

$$\begin{aligned}\boldsymbol{\theta}^{t+1} &= \operatorname*{argmax}_{\boldsymbol{\theta}} Q^t(\boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n Q_i^t(\boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n \sum_x q_i^t(x) \log P(X_i = x, Z_i = z_i \,|\, \boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n \sum_x q_i^t(x) \log(\pi_x \binom{10}{z_i} \theta_x^{z_i} (1-\theta_x)^{10-z_i})\end{aligned}$$

This is now a constrained optimization problem because $\pi_A + \pi_B = 1$. A straightforward computation with Lagrange multipliers shows that the optimizers θ^{t+1} are given by the same formula (13.8) as before, and

$$\pi_A^{t+1} = \frac{\sum_{i=1}^n q_i^t(A)}{\sum_{i=1}^n (q_i^t(A) + q_i^t(B))} \qquad \text{and} \qquad \pi_B^{t+1} = \frac{\sum_{i=1}^n q_i^t(B)}{\sum_{i=1}^n (q_i^t(A) + q_i^t(B))}. \tag{13.9}$$

Remark 13.2.8. In the coin flip examples, the choice of which coin is called coin A and which is called coin B is arbitrary. Therefore, for any choice of parameters $\boldsymbol{\theta} = (\theta_A, \pi_A, \theta_B, \pi_B)$ there is another choice $\boldsymbol{\theta}' = (\theta_B, \pi_B, \theta_A, \pi_A)$ with $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}')$. Except in the very special case when the maximizer $\widehat{\boldsymbol{\theta}}_{MLE}$ is symmetric, with $\boldsymbol{\theta} = \boldsymbol{\theta}'$, this means that ℓ does not have a unique maximizer, and the choice of which one the EM algorithm will find depends on things like the initial guess of $\boldsymbol{\theta}^0$. That is, the EM algorithm and any other form of maximum likelihood estimation will essentially only tell you that one of the coins has parameters θ_A, π_A and the other has parameters θ_B, π_B, but it cannot tell you which coin is which. A similar symmetry and indeterminacy exists in the the parameters for a GMM, which we treat in the next section.

13.2.2 The Idea of the EM Algorithm

In this section we describe the main idea behind the EM algorithm. In some ways it resembles Newton's method for optimization, but instead of finding a quadratic approximation and optimizing that, the EM algorithm finds a different function that approximates ℓ and always lies below the graph of ℓ. Given an estimate $\boldsymbol{\theta}^t$, the idea is to construct a function $\widetilde{Q}^t(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$, satisfying the following properties:

(i) $\widetilde{Q}^t(\boldsymbol{\theta}) \leq \ell(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

(ii) $\widetilde{Q}^t(\boldsymbol{\theta}^t) = \ell(\boldsymbol{\theta}^t)$.

(iii) It is relatively easy to maximize $\widetilde{Q}^t(\boldsymbol{\theta})$.

Given such a $\widetilde{Q}^t$, compute a new estimate $\boldsymbol{\theta}^{t+1}$ as

$$\boldsymbol{\theta}^{t+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} \widetilde{Q}^t(\boldsymbol{\theta}).$$

See Figure 13.2 for an illustration of one step of the EM algorithm.

Exercise 13.3 shows that

$$\ell(\boldsymbol{\theta}^{t+1}) \geq \ell(\boldsymbol{\theta}^t), \tag{13.10}$$

and the sequence $(\ell(\boldsymbol{\theta}^t))_{t=0}^{\infty}$ is bounded above and hence converges to a finite limit. Exercise 13.4 shows that if $\widetilde{Q}^t(\boldsymbol{\theta})$ is strictly concave, then $\ell(\boldsymbol{\theta}^{t+1}) > \ell(\boldsymbol{\theta}^t)$, unless $\boldsymbol{\theta}^t$ is a critical point of ℓ.

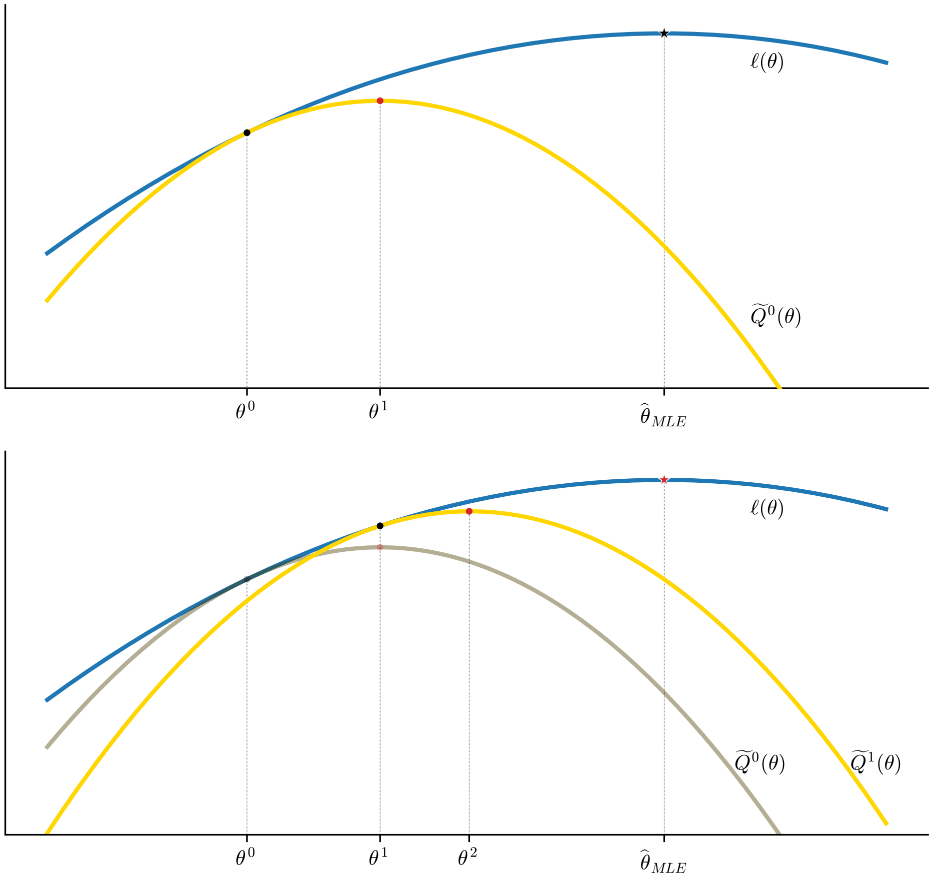


Figure 13.2: An illustration of the first step (top panel) of the EM algorithm for maximizing likelihood. For a given $\boldsymbol{\theta}^0$ find $\widetilde{Q}^0(\boldsymbol{\theta})$ (yellow), which always lies below the log likelihood function (blue), and which agrees (black dot) with the log likelihood function at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. The maximizer of $\widetilde{Q}^0(\boldsymbol{\theta})$ is the new estimate $\boldsymbol{\theta}^1$. The bottom panel shows the next step, with $\widetilde{Q}^1(\boldsymbol{\theta})$ (yellow), which always lies below the log likelihood function (blue), and which agrees (black dot) with the log likelihood function at $\boldsymbol{\theta} = \boldsymbol{\theta}^1$. The maximizer of $\widetilde{Q}^1(\boldsymbol{\theta})$ is the new estimate $\boldsymbol{\theta}^2$ Repeating the process gives a sequence that must climb upward on the log likelihood function. It often converges to the MLE estimate $\widehat{\boldsymbol{\theta}}_{MLE}$, but that is not always guaranteed.

13.2.3 Derivation of the EM Algorithm

In this section we complete the proof of Theorem 13.2.3 by doing the following:

(i) Construct a function $\widetilde{Q}^t(\boldsymbol{\theta})$ satisfying the two conditions (i) and (ii) listed at the beginning of Section 13.2.2.

(ii) Show that maximizing the function $Q^t(\boldsymbol{\theta})$ defined in (13.6) is equivalent to maximizing $\widetilde{Q}^t(\boldsymbol{\theta})$.

Given a $\boldsymbol{\theta}^t$, we construct $\widetilde{Q}^t$ as a sum:

$$\widetilde{Q}^t(\boldsymbol{\theta}) = \sum_{i=1}^{n} \widetilde{Q}^t_i(\boldsymbol{\theta}).$$

To construct the function $\widetilde{Q}^t_i(\boldsymbol{\theta})$ first let

$$q_i^t(x) = P(X_i = x \,\big|\, \mathbf{D}, \boldsymbol{\theta}^t) = P(X_i = x \,\big|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t),$$

where the last equality follows from the fact that the samples X_i and Z_i are independent of all the other data X_j and Z_j for all $j \neq i$. The function q_i^t is a distribution (a p.d.f. or p.m.f.) for the possible states of the latent variable X_i. Note that q_i^t is independent of $\boldsymbol{\theta}$ but depends on $\boldsymbol{\theta}^t$.

Now set

$$\widetilde{Q}^t_i(\boldsymbol{\theta}) = \mathbb{E}_{X_i|\boldsymbol{\theta}^t}\left[\log\left(\frac{P(Z_i = \mathbf{z}_i, X_i = x \,|\, \boldsymbol{\theta})}{q_i^t(x)}\right)\right] \tag{13.11}$$

$$= \sum_x q_i^t(x) \log\left(\frac{P(Z_i = \mathbf{z}_i, X_i = x \,|\, \boldsymbol{\theta})}{q_i^t(x)}\right), \tag{13.12}$$

where the expectation is taken with respect to the distribution q_i^t. The computation of $\widetilde{Q}^t_i(\boldsymbol{\theta})$ is called the *expectation (E)* step of the algorithm, because $\widetilde{Q}^t$ is an expected value; whereas, the computation of $\boldsymbol{\theta}^{t+1} = \operatorname{argmax}_{\boldsymbol{\theta}} \widetilde{Q}^t_i(\boldsymbol{\theta})$ is the *maximization (M)* step. The EM algorithm alternates between these two steps.

Lemma 13.2.9. *The function $\widetilde{Q}^t(\boldsymbol{\theta})$, as defined in* (13.11) *satisfies conditions* (*i*) *and* (*ii*) *described above.*

Proof. Condition (i), that $\widetilde{Q}^t(\boldsymbol{\theta}) \leq \ell(\boldsymbol{\theta})$, follows from Jensen's inequality, since $\sum_x q_i^t(x) = 1$ and log is concave:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\sum_x P(Z_i = \mathbf{z}_i, X_i = x \,|\, \boldsymbol{\theta})\right) = \sum_{i=1}^{n} \log\left(\sum_x q_i^t(x) \frac{P(Z_i = \mathbf{z}_i, X_i = x \,|\, \boldsymbol{\theta})}{q_i^t(x)}\right)$$
$$\geq \sum_{i=1}^{n} \sum_x q_i^t(x) \log\left(\frac{P(Z_i = \mathbf{z}_i, X_i = x \,|\, \boldsymbol{\theta})}{q_i^t(x)}\right) = \widetilde{Q}^t(\boldsymbol{\theta}).$$

To prove Condition (ii), that $\widetilde{Q}^t(\boldsymbol{\theta}^t) = \ell(\boldsymbol{\theta}^t)$, observe that

$$\begin{aligned}
\widetilde{Q}^t(\boldsymbol{\theta}^t) &= \sum_{i=1}^{n} \sum_{x} q_i^t(x) \log \left(\frac{P(Z_i = \mathbf{z}_i, X_i = x \,\big|\, \boldsymbol{\theta}^t)}{P(X_i = x \,\big|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t)} \right) \\
&= \sum_{i=1}^{n} \sum_{x} q_i^t(x) \log \left(\frac{P(X_i = x \,\big|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t) P(Z_i = \mathbf{z}_i \,\big|\, \boldsymbol{\theta}^t)}{P(X_i = x \,\big|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t)} \right) \\
&= \sum_{i=1}^{n} \log \left(P(Z_i = \mathbf{z}_i \,\big|\, \boldsymbol{\theta}^t) \right) \sum_{x} P(X_i = x \,\big|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t) \\
&= \sum_{i=1}^{n} \log P(Z_i = \mathbf{z}_i \,\big|\, \boldsymbol{\theta}^t) \\
&= \ell(\boldsymbol{\theta}^t).
\end{aligned}$$

Here the third equality follows from the fact that, after cancelling the common factor of $P(X_i = x \,\big|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t)$ in the fraction, the expression inside the logarithm is independent of x, and thus the expression with the logarithm can be pulled outside the sum. The fourth equality follows from the fact that $\sum_x P(X_i = x \,\big|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t) = 1$. □

Remark 13.2.10. Note that the argument for Condition (i) only used the fact that $\sum_x q_i^t(x) = 1$, so $\ell(\boldsymbol{\theta}) \geq \widetilde{Q}^t(\boldsymbol{\theta})$ regardless of the choice of distribution q, but it can be shown that the optimal choice for q is $q_i^t(x)$.

In the maximization step finding argmax $\widetilde{Q}^t(\boldsymbol{\theta}))$ is equivalent to solving a simpler problem:

$$\begin{aligned}
\boldsymbol{\theta}^{t+1} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \widetilde{Q}^t(\boldsymbol{\theta}) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{x} q_i^t(x) \log \left(\frac{P(x, Z_i = \mathbf{z}_i \,\big|\, \boldsymbol{\theta})}{q_i^t(x)} \right) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{x} q_i^t(x) \log P(x, Z_i = \mathbf{z}_i \,\big|\, \boldsymbol{\theta}) - \sum_{x} q_i^t(x) \log(q_i^t(x)) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{x} q_i^t(x) \log P(x, Z_i = \mathbf{z}_i \,\big|\, \boldsymbol{\theta})
\end{aligned}$$

The last equality follows from the fact that $q_i^t(x)$ is independent of $\boldsymbol{\theta}$. This means that we can replace $\widetilde{Q}^t(\boldsymbol{\theta})$ by $Q^t(\boldsymbol{\theta}) = \sum_{i=1}^n Q_i^t(\boldsymbol{\theta})$, where

$$Q_i^t(\boldsymbol{\theta}) = \sum_{x} q_i^t(x) \log P(x, Z_i = \mathbf{z}_i \,\big|\, \boldsymbol{\theta}).$$

The previous argument shows that the two steps listed in Definition 13.2.2 are equivalent to items (i)–(iii) above, and hence produce a sequence $(\boldsymbol{\theta}^t)_{t=0}^{\infty}$ with the property that $(\ell(\boldsymbol{\theta}^t))_{t=0}^{\infty}$ is nondecreasing. This completes the proof of Theorem 13.2.3.

Remark 13.2.11. C.F.J. Wu proved convergence results in 1983.

13.3 EM for Gaussian Mixture Models

Recall that Gaussian mixture model (GMM) is a mixture model of the form

$$P(Z = \mathbf{z} \,|\, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k p_k(Z = \mathbf{z} \,|\, \boldsymbol{\theta}),$$

where

$$\begin{aligned} p_k(\mathbf{z} \,|\, \boldsymbol{\theta}) &= P(Z = \mathbf{z} \,|\, X = k, \boldsymbol{\theta}) \\ &= \mathscr{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_k, \Sigma_k) = |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma_k^{-1}(\mathbf{z}-\boldsymbol{\mu}_k)} \end{aligned}$$

and

$$P(\mathbf{z} \,|\, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \mathscr{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_k, \Sigma_k).$$

This section describes how to use the EM algorithm to estimate the parameters in a GMM. One of the main applications is clustering. We also discuss how the K-means algorithm can be thought of as a simplified version of EM for GMMs.

13.3.1 GMM for Clustering

One use of Gaussian mixture models is clustering, that is, the unsupervised learning task of identifying clusters in unlabeled data. Assume the data set is of the form $\mathbf{D} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, where each $\mathbf{z}_i$ belongs to some (unobserved) cluster $X = k$, and the distribution associated to the kth cluster is Gaussian, then the data are distributed according to a GMM with p.d.f. equal to

$$p(\mathbf{z}) = \sum_{k=1}^{K} w_k \mathscr{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_k, \Sigma_k).$$

If $\boldsymbol{\theta} = (w_1, \boldsymbol{\mu}_1, \Sigma_1, \ldots, w_k, \boldsymbol{\mu}_K, \Sigma_K)$ is known, then for each $\mathbf{z}$ we can compute the probability that $\mathbf{z}$ belongs to cluster k using Bayes' rule

$$P(X = k \,|\, \mathbf{z}, \boldsymbol{\theta}) = \frac{P(\mathbf{z} \,|\, X = k, \boldsymbol{\theta}) p(X = k \,|\, \boldsymbol{\theta})}{\sum_{k'} p(\mathbf{z} \,|\, X = k', \boldsymbol{\theta}) p(X = k' \,|\, \boldsymbol{\theta})} = \frac{\mathscr{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_k, \Sigma_k)) w_k}{\sum_{k'} \mathscr{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_{k'}, \Sigma_{k'})) w_{k'}}.$$

Assigning a point $\mathbf{z}$ to the cluster

$$\operatorname*{argmax}_k p(X = k \,|\, \mathbf{z}, \boldsymbol{\theta})$$

gives a clustering of the data. This assignment is identical to the GDA generative classifier described in Section 7.8.1. The difference is in how the model is trained. In GDA, the values of X are given, which makes training relatively easy. In the case of clustering, the labels/clusters are unobserved (latent), but we can still estimate them using EM.

The construction described above is sometimes called *hard clustering*, since every point $\mathbf{z}$ is assigned a unique label/cluster. In contrast, *soft clustering* amounts to assigning to $\mathbf{z}$ the vector of probabilities $(r_1(\mathbf{z}), \ldots, r_K(\mathbf{z}))$, where $r_k(\mathbf{z}) = p(X = k \mid \mathbf{z}, \boldsymbol{\theta})$. This allows us a more nuanced view of which clusters are most likely associated to each $\mathbf{z}$.

13.3.2 EM for GMM

Applying the EM algorithm to the case of a GMM with p.d.f. equal to $P(\mathbf{z}) = \sum_{k=1}^{K} w_k f_k(\mathbf{z})$ with $f_k(\mathbf{z}) = \mathscr{N}(\mathbf{z} \,|\, \boldsymbol{\mu}_k, \Sigma_k)$ for each k, we have

$$\begin{aligned} q_i^t(k) &= P(X_i = k \,|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t) \\ &= \frac{P(Z_i = \mathbf{z}_i \,|\, X_i = k, \boldsymbol{\theta}^t) P(X_i = k \,|\, \boldsymbol{\theta}^t)}{\sum_{k'=1}^{K} P(Z_i = \mathbf{z}_i \,|\, X_i = k', \boldsymbol{\theta}^t) P(X_i = k' \,|\, \boldsymbol{\theta}^t)} \\ &= \frac{|\Sigma_k^t|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{z}_i - \boldsymbol{\mu}_k^t)^\intercal (\Sigma_k^t)^{-1} (\mathbf{z}_i - \boldsymbol{\mu}_k^t)) w_k^t}{\sum_{k'=1}^{K} |\Sigma_{k'}^t|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{z}_i - \boldsymbol{\mu}_{k'}^t)^\intercal (\Sigma_{k'}^t)^{-1} (\mathbf{z}_i - \boldsymbol{\mu}_{k'}^t)) w_{k'}^t} \end{aligned} \tag{13.13}$$

And the function to maximize is

$$\begin{aligned} Q^t(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{k \in \mathscr{X}} q_i^t(k) \log P(X_i = k, Z_i = \mathbf{z}_i \,|\, \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \sum_{k \in \mathscr{X}} q_i^t(k) [\log(w_k) - \frac{1}{2} \log(|2\pi\Sigma_k|) - \frac{1}{2} (\mathbf{z}_i - \boldsymbol{\mu}_k)^\intercal (\Sigma_k)^{-1} (\mathbf{z}_i - \boldsymbol{\mu}_k)] \end{aligned} \tag{13.14}$$

A straightforward KKT argument (see Volume 2, Section 14.4) to account for the constraints $w_k \geq 0$ and $\sum_k w_k = 1$ shows that the maximizing w_k is

$$w_k^{t+1} = \frac{1}{n} \sum_{i=1}^{n} q_i^t(k). \tag{13.15}$$

Essentially the same argument that gives the usual MLE for a normal distribution shows that in this case we have

$$\begin{aligned} \boldsymbol{\mu}_k^{t+1} &= \frac{\sum_{i=1}^{n} q_i^t(k) \mathbf{z}_i}{\sum_{i=1}^{n} q_i^t(k)} \\ \Sigma_k^{t+1} &= \frac{\sum_{i=1}^{n} q_i^t(k) (\mathbf{z}_i - \boldsymbol{\mu}_k^{t+1})(\mathbf{z}_i - \boldsymbol{\mu}_k^{t+1})^\intercal}{\sum_{i=1}^{n} q_i^t(k)}. \end{aligned} \tag{13.16}$$

13.3.3 K-means as a Special Case of EM for GMM

The K-means algorithm is a simplified variant of EM for GMM. It assumes that $\Sigma_k = \sigma^2 I$ is the same for all k, and that $w_k = \frac{1}{K}$ for all k. Thus the only parameters that need estimating are the means $\boldsymbol{\mu}_k$.

Given estimates $\boldsymbol{\mu}_k^t$, the algorithm computes the quantity

$$\kappa_i^t = \operatorname*{argmax}_k P(X_i = k \,|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t), \tag{13.17}$$

which is just the value of k for which $\boldsymbol{\mu}_k^t$ is nearest to $\mathbf{z}_i$, so it is much easier to compute than $P(X_i = k \,|\, Z_i = \mathbf{z}_i, \boldsymbol{\theta}^t)$. The algorithm then substitutes the following rough approximation of (13.13)

$$q_i^t(k) \approx \begin{cases} 1 & \text{if } k = \kappa_i^t \\ 0 & \text{otherwise.} \end{cases} \tag{13.18}$$

Using the approximation (13.18) in (13.16) to compute $\boldsymbol{\mu}_k^{t+1}$ reduces to

$$\boldsymbol{\mu}_k^{t+1} = \frac{1}{N_k} \sum_{i:\kappa_i^t = k} \mathbf{z}_i, \tag{13.19}$$

where N_k is the number of $\mathbf{z}_i$ with $\kappa_i^t = k$.

Thus the algorithm reduces to the following:
Given K and estimated means $\boldsymbol{\mu}_k^t$

(i) For each i compute κ_i^t to be the k such that $\boldsymbol{\mu}_k^t$ is nearest to $\mathbf{z}_i$.

(ii) For each k compute $\boldsymbol{\mu}_k^{t+1}$, using (13.19).

(iii) Repeat until the means converge or the number of iterations reaches some predefined termination point.

K-means is faster to compute than the full EM for GMM algorithm, but if there is reason to believe that the probabilities w_k are not all equal or the covariances Σ_k are not all equal and diagonal, then you may get better clustering results using the full GMM for clustering.

13.1. Consider a mixture-of-experts model with $\boldsymbol{\theta} = (V, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1, \sigma_2)$, where $V = \begin{bmatrix} -20 & 0 \\ 10 & 0 \end{bmatrix}$, $\boldsymbol{\beta}_1 = (1, 3)$, $\boldsymbol{\beta}_2 = (4, -1)$, $\sigma_1 = 0.1$, $\sigma_2 = 0.2$. Code up a way to sample from this model, given an input x. Your code should take as input $x \in \mathbb{R}$, and draw $z \sim \operatorname{Cat}(\mathscr{S}(V^{\mathsf{T}}\mathbf{x}))$, where $\mathbf{x} = (1, x)$. It should then return a draw from $y \sim \mathscr{N}(\mathbf{x}^{\mathsf{T}}\beta_z, \sigma_z^2)$.

(i) Draw 10^3 times from $x \sim \operatorname{Uniform}(0, 4)$ and feed the results to your mixture of experts to produce corresponding y values.

(ii) Scatterplot the results (x_i, y_i), and on the same graph plot the line segment $y = \mathbf{x}^{\mathsf{T}}\beta_1$ for $x \in [0, 2]$ and the line segment $y = \mathbf{x}^{\mathsf{T}}\beta_2$ for $x \in [2, 4]$.

13.2. Consider the weather-and-tree-ring HMM where the Markov chain of the unobserved variables X_i has state space $\{W, D\}$ with transition matrix $A = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix}$ and initial probability $\boldsymbol{\pi} = (0.1, 0.9)$. Assume, further, that the observations Z_i take values in $\{S, M, L\}$ and the probabilities of observing Z_i given X_i are given by the matrix $B = \begin{bmatrix} 0.1 & 0.7 \\ 0.4 & 0.2 \\ 0.5 & 0.1 \end{bmatrix}$ (so, for example, $P(Z = M \,|\, X = D) = 0.2$). We denote all these parameters by $\lambda = (A, B, \boldsymbol{\pi})$. Assume that we have observed the sequence $z_0, z_1, z_2 = M, S, L$.

(i) Compute the probability $P((M, S, L) \,|\, (W, W, W), \lambda) = P(Z_0 = M, Z_1 = S, Z_2 = L \,|\, \lambda, X_0 = X_1 = X_2 = W)$ by hand.

(ii) For every possible sequence of hidden states x_0, x_1, x_2, compute the probability $P((M, S, L) \,|\, \lambda, (x_0, x_1, x_2))$. Hint: Code it.

(iii) The values of X_i are hidden (unobserved), but we can calculate $P((M, S, L) \,|\, \lambda)$ by marginalizing (summing over all possible values of the hidden variables). Use the results of the previous step to compute $P((M, S, L) \,|\, \lambda) = \sum_{x_0, x_1, x_2} P((M, S, L) \,|\, \lambda, (x_0, x_1, x_2)) P((x_0, \dots, x_{n-1}))$.

(iv) For an observed sequence $(z_0, \dots, z_{n-1})$ of length n, how many individual terms will need to be computed to find the sum $P((z_0, \dots, z_{n-1}) \,|\, \lambda) = \sum_{x_0, \dots, x_{n-1}} P((z_0, \dots, z_{n-1}) \,|\, \lambda, (x_0, \dots, x_{n-1})) P((x_0, \dots, x_{n-1}))$? Later in the chapter we give an alternative algorithm to compute $P((z_0, \dots, z_{n-1}) \,|\, \lambda)$ much more efficiently.

13.3. Using only the definition of ℓ and the two conditions (i) and (ii) listed at the beginning of Section 13.2.2, show the following properties of the sequence $(\ell(\theta^t))_{t=0}^{\infty}$ constructed by the EM algorithm:

(i) The sequence $(\ell(\theta^t))_{t=0}^{\infty}$ is nondecreasing.

(ii) The log likelihood $\ell(\theta)$ is bounded above.

(iii) The sequence $(\ell(\theta^t))_{t=0}^{\infty}$ converges to some finite value (Hint: see Lemma 5.4.28 in Volume 1).

13.4. Assume that $\widetilde{Q}^t(\theta)$ and ℓ are both continuously differentiable in θ and that $\widetilde{Q}^t(\theta)$ is strictly concave and satisfies the conditions (i) and (ii) described at the beginning of Section 13.2.2. Prove that if $\ell(\theta^{t+1}) = \ell(\theta^t)$ then $D\ell(\theta^t) = \mathbf{0}$ as follows:

(i) Show that $\widetilde{Q}^t(\theta^t) = \widetilde{Q}^t(\theta^{t+1})$.

(ii) Show that $\theta^t = \theta^{t+1}$.

(iii) Show that $D\widetilde{Q}^t(\theta^t) = \mathbf{0}$

(iv) Show (by way of contradiction) that if $D\ell(\theta^t) \neq \mathbf{0}$, then there exists $\mathbf{v}$ such that the directional derivative $D_{\mathbf{v}}\ell(\theta^t)$ satisfies $D_{\mathbf{v}}\ell(\theta^t) < 0$.

(v) Use the limit definition of the directional derivative and the two properties (i) and (ii) from the beginning of Section 13.2.2 to show that $D_{\mathbf{v}}\widetilde{Q}^t(\theta^t) \leq D_{\mathbf{v}}\ell(\theta^t)$

(vi) Show that the previous step contradicts something that you have already proved, and hence the assumption in (iv) above is false; that is, $D\ell(\theta^t)$ vanishes, as required.

13.5. Prove that the update equations (13.8) and (13.9) for EM applied to the coin flips in Example 13.2.5 and 13.2.7 are correct; that is, show that these do indeed give the correct values of $\operatorname{argmax}_{\boldsymbol{\theta}} Q^t(\boldsymbol{\theta})$.

13.6. Code up the EM algorithm for the Binomial mixture model of Example 13.2.5. Your code should accept an array $\mathbf{z}$ (the sample $z_1, \ldots, z_n$), and, optionally, an initial guess of the parameter $\boldsymbol{\theta}^0 = (\theta_A^0, \theta_B^0)$. If no initial $\boldsymbol{\theta}^0$ is provided, construct one randomly (What's a good way to do this?). With each iteration and existing estimate $\boldsymbol{\theta}^t$, your code should compute $q_i^t(x)$ for each $i \in \{1, \ldots, n\}$ and each $x \in \{A, B\}$ and the updates θ_A^{t+1} and θ_B^{t+1}. Include a stopping criterion (a maximum number of iterations and some measure of how well the estimates are converging). Return the final value of $\boldsymbol{\theta}^t$.

(i) Apply your code to the data from Example 13.2.5. What are its predictions for the values of θ_A and θ_B?

(ii) Let $\theta_A = 0.3$ and $\theta_B = 0.8$, and take a draw of length 20 from a Bernoulli($\frac{1}{2}$) random variable to get values $x_1, \ldots, x_{20}$ for the latent variable $X \in \{A, B\}$. For each of the twenty values of x_i, draw a corresponding z_i from a Binomial$(10, \theta_{x_i})$. The sequence $\mathbf{D} = \{z_1, \ldots, z_{20}\}$ is your data set.

(iii) Apply your code to the data $\mathbf{D}$ from the previous step. What are its predictions for the values of θ_A and θ_B?

13.7. Prove that the (13.15) gives the value of $\mathbf{w}$ that maximizes (13.14), and that (13.16) gives the maximizing $\boldsymbol{\mu}$ and Σ.
Hint: For the computation of Σ you may want to differentiate with respect to Σ^{-1}. Also feel free to assume the following properties of matrices and derivatives:

(i) $D_A \log(\det(A)) = (A^{-1})^{\mathsf{T}}$ (If you don't want to assume this fact, you can verify it directly with a tedious computation, first by expressing both $\det(A)$ and A^{-1} in terms of their cofactor expansions and then taking the partial derivatives $\frac{\partial}{\partial A_{ij}}$ of the expression $\log(\det(A))$ for each i, j).

(ii) $\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA)$.

(iii) $(\mathbf{z} - \boldsymbol{\mu})^{\mathsf{T}} A(\mathbf{z} - \boldsymbol{\mu}) = \operatorname{tr}((\mathbf{z} - \boldsymbol{\mu})^{\mathsf{T}} A(\mathbf{z} - \boldsymbol{\mu}))$

(iv) $D_A \operatorname{tr}(AB) = B^{\mathsf{T}}$

13.8. In your own words explain what the EM algorithm is for and why one might want to use it instead of something like Newton's method or gradient descent.

13.9. K-means vs. EM for GMMs:

(i) In your own words, explain how the k-means algorithm could be thought of as a simplified variant of EM for GMMs.

(ii) Explain the relative benefits and disadvantages of doing clustering using EM for GMMs as compared to the k-means algorithm.

(iii) Draw or plot a dataset where clustering using k-means would perform poorly compared to clustering with a GMM learned using EM.