Exercise: Hidden variables in DGMs

Consider the DGMs in Figure 1 which both define $p(X_{1:6})$, where we number empty nodes left to right, top to bottom. The graph on the left defines the joint as

$$p(X_{1:6}) = \sum_h p(X_1)p(X_2)p(X_3)p(H=h|X_{1:3})p(X_4|H=h)p(X_5|H=h)p(X_6|H=h) \tag{1}$$

where we have marginalized over the hidden variable H. The graph on the right defines the joint as

$$p(X_{1:6}) = p(X_1)p(X_2)p(X_3)p(X_4|X_{1:3})p(X_5|X_{1:4})p(X_6|X_{1:5}) \tag{2}$$

1. Assuming all nodes (including H) are binary and all CPDs are tabular, prove that the model on the left has 17 free parameters.
2. Assuming all nodes are binary and all CPDs are tabular, prove that the model on the right has 59 free parameters.
3. Suppose we have a data set $\mathcal{D} = X_{1:6}^n$ for $n = 1 : N$, where we observe the Xs but not H, and we want to estimate the parameters of the CPDs using maximum likelihood. For which model is this easier? Explain your answer.

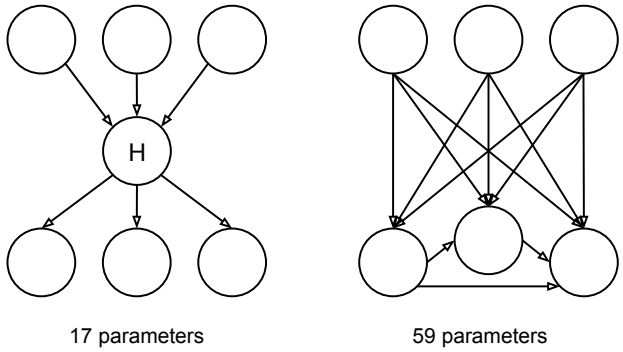

Figure 1: A DGM with and without hidden variables. For example, the leaves might represent medical symptoms, the root nodes primary causes (such as smoking, diet and exercise), and the hidden variable can represent mediating factors, such as heart disease. Marginalizing out the hidden variable induces a clique.