Lecture topics:

- Learning Bayesian networks from data
  - maximum likelihood, BIC
  - Bayesian, marginal likelihood

Learning Bayesian networks

There are two problems we have to solve in order to estimate Bayesian networks from available data. We have to estimate the parameters given a specific structure, and we have to search over possible structures (model selection).

Suppose now that we have d discrete variables, $x_1, \ldots, x_d$, where $x_i \in \{1, \ldots, r_i\}$, and n *complete* observations $D = \{(x_1^t, \ldots, x_d^t), t = 1, \ldots, n\}$. In other words, each observation contains a value assignment to all the variables in the model. This is a simplification and models in practice (e.g., HMMs) have to be estimated from incomplete data. We will also assume that the conditional probabilities in the models are fully parameterized. This means, e.g., that in $P(x_1|x_2)$ we can select the probability distribution over x_1 separately and without constraints for each possible value of the parent x_2. Models used in practice often do have parametric constraints.

Maximum likelihood parameter estimation

Given an acyclic graph G over d variables, we know from previous lecture that we can write down the associated joint distribution as

$$P(x_1, \ldots, x_d) = \prod_{i=1}^{d} P(x_i|x_{pa_i}) \tag{1}$$

The parameters we have to learn are therefore the conditional distributions $P(x_i|x_{pa_i})$ in the product. For later utility we will use $P(x_i|x_{pa_i}) = \theta_{x_i|x_{pa_i}}$ to specify the parameters.

Given the complete data D, the log-likelihood function is

$$\begin{aligned} l(D;\theta,G) &= \log P(D|\theta) && (2)\\ &= \sum_{t=1}^{n} \log P(x_1^t,\ldots,x_d^t|\theta) && (3)\\ &= \sum_{t=1}^{n}\sum_{i=1}^{n} \log \theta_{x_i^t|x_{pa_i}^t} && (4)\\ &= \sum_{i=1}^{n}\sum_{x_i,x_{pa_i}} n(x_i,x_{pa_i}) \log \theta_{x_i|x_{pa_i}} && (5) \end{aligned}$$

where we have again collapsed the available data into counts $n(x_i, x_{pa_i})$, the number of observed instances with a particular setting of the variable and its parents. These are the *sufficient statistics* we need from the data in order to estimate the parameters. This will be true in the Bayesian setting as well (discussed below). Note that the statistics we need depend on the model structure or graph G. The parameters $\hat{\theta}_{x_i|x_{pa_i}}$ that maximize the log-likelihood have simple closed form expressions in terms of empirical fractions:

$$\hat{\theta}_{x_i|x_{pa_i}} = \frac{n(x_i,x_{pa_i})}{\sum_{x_i'=1}^{r_i} n(x_i',x_{pa_i})} \tag{6}$$

This simplicity is due to our assumption that $\theta_{x_i|x_{pa_i}}$ can be chosen freely for each setting of the parents x_{pa_i}. The parameter estimates are likely not going to be particularly good when the number of parents increases. For example, just to provide one observation per a configuration of parent variables would require $\prod_{j\in pa_i} r_j$ instances. Introducing some regularization is clearly important, at least in the fully parameterized case. We will provide a Bayesian treatment of the parameter estimation problem shortly.

BIC and structure estimation

Given the ML parameter estimates $\hat{\theta}_{x_i|x_{pa_i}}$ we can evaluate the resulting maximum value of the log-likelihood $l(D;\hat{\theta},G)$ as well as the corresponding BIC score:

$$BIC(G) = l(D;\hat{\theta},G) - \frac{dim(G)}{2}\log(n) \tag{7}$$

where $dim(G)$ specifies the number of (independent) parameters in the model. In our case this is given by

$$dim(G) = \sum_{i=1}^{d}(r_i-1)\prod_{j\in pa_i} r_j \tag{8}$$

where each term in the sum corresponds to the size of the probability table $P(x_i|x_{pa_i})$ minus the associated normalization constraints $\sum_{x_i'} P(x_i'|x_{pa_i}) = 1$.

BIC and likelihood equivalence

Suppose we have two different graphs G and G' that nevertheless make exactly the same independence assumptions about the variables involved. For example, neither graph in

makes any independence assumptions and are therefore equivalent in this sense. The resulting BIC scores for such graphs are also identical. The principle that equivalent graphs should receive the same score is known as *likelihood equivalence*. How can we determine if two graphs are equivalent? In principle this can be done by deriving all the possible independence statements from the graphs and comparing the resulting lists but there are easier ways. Two graphs are equivalent if they differ only in the direction of arcs and possess the same *v-structures*, i.e., they have the same set of converging arcs (two or more arcs pointing to a single node). This criterion captures most equivalences. Figure 1 provides a list of all equivalence classes of graphs over three variables. Only one representative of each class is shown and the number next to the graph indicates how many graphs there are that are equivalent to the representative.

Equivalence of graphs and the associated scores highlight why we should not interpret the arcs in Bayesian networks as indicating the direction of causal influence. While models are often drawn based on one's causal understanding, when learning them from the available data we can only distinguish between models that make different probabilistic assumptions about the variables involved (different independence properties), not based on which way the arcs are pointing. It is nevertheless possible to estimate causal Bayesian networks, models where we can interpret the arcs as causal influences. The difficulty is that we need *interventional* data to do so, i.e., data that correspond to explicitly setting some of the variables to specific values (controlled experiments) rather than simply observing the values they take.

Bayesian estimation

The idea in Bayesian estimation is to avoid reducing our knowledge about the parameters into point estimates (e.g., ML estimates) but instead retain all the information in a form of a distribution over the possible parameter values. This is advantageous when the available data are limited and the number of parameters is large (e.g., only a few data points per

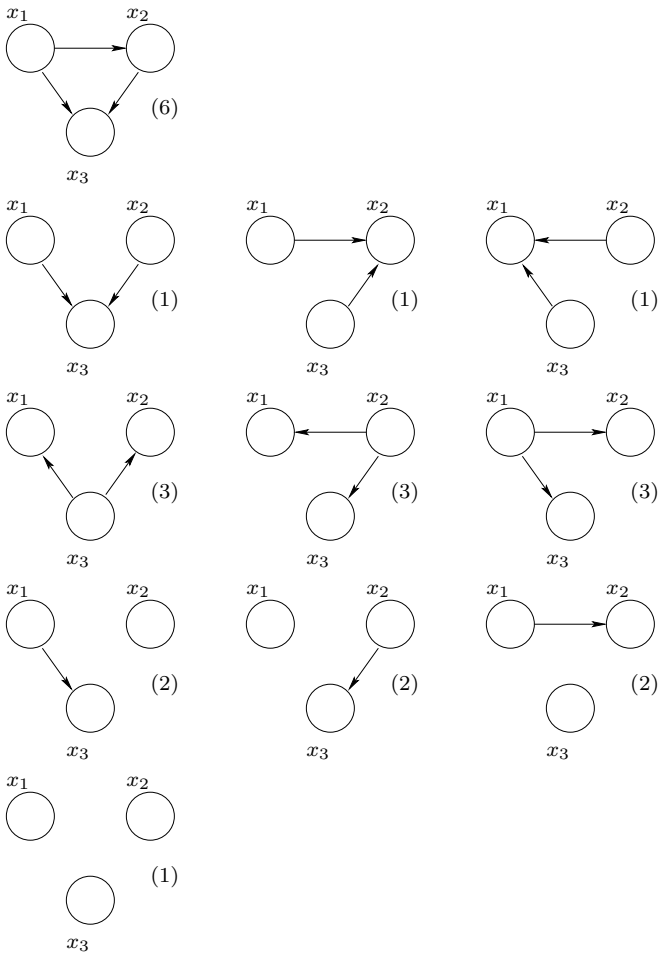
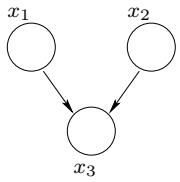


Figure 1: Equivalence classes of graphs over three variables.

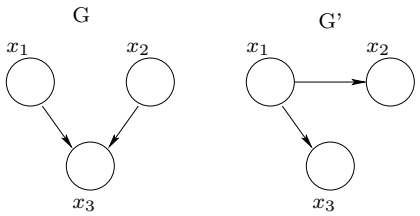
parameter to estimate). The Bayesian framework requires us to also articulate our knowledge about the parameters prior to seeing any data in a form of a distribution, the prior distribution. Consider the following simple graph with three variables



The parameters we have to estimate are $\{\theta_{x_1}\}$, $\{\theta_{x_2}\}$, and $\{\theta_{x_3|x_1,x_2}\}$. We will assume that the parameters are a priori independent for each variable and across different configurations of parents (parameter independence assumption):

$$P(\theta) \;=\; P(\{\theta_{x_1}\}_{x_1=1,\ldots r_1})\, P(\{\theta_{x_2}\}_{x_2=1,\ldots,r_2}) \prod_{x_1,x_2} P(\{\theta_{x_3|x_1,x_2}\}_{x_3=1,\ldots,r_3}) \tag{9}$$

We will also assume that we will use the same prior distribution over the same parameters should they appear in different graphs (parameter modularity). For example, since x_1 has no parents in G and G' given by



we need the same parameter $\{\theta_{x_1}\}$ in both models. The parameter modularity assumption corresponds to using the same prior distribution $P(\{\theta_{x_1}\}_{x_1=1,\ldots,r_1})$ for both models (other parameters would have different prior distributions since, e.g., $\theta_{x_3|x_1,x_2}$ does not appear in graph G'). Finally, we would like the marginal likelihood score to satisfy likelihood equivalence similarly to BIC. In other words, if G and G' are equivalent, then we would like $P(D|G) = P(D|G')$ where, e.g.,

$$P(D|G) = \int P(D|\theta, G)P(\theta)d\theta \tag{10}$$

If we agree to these three assumptions (parameter independence, modularity, and likelihood equivalence), then we can only choose one type of prior distribution over the parameters,

the Dirichlet distribution. To specify and use this prior distribution, it will be helpful to change the notation slightly. We will denote the parameters by θ_{ijk} where i specifies the variable, j the parent configuration (see below), and k the value of the variable x_i. Clearly, $\sum_{k=1}^{r_i} \theta_{ijk} = 1$ for all i and j. The parent configurations are simply indexed from $j = 1, \ldots, q_i$ as in

j	x_1	x_2
1	1	1
2	2	1
$\ldots$	$\ldots$	$\ldots$
q	r_1	r_2

(11)

where $q = r_1 r_2$. When x_i has no parents we say there is only one "parent configuration" so that $P(x_i = k) = \theta_{i1k}$. Note that writing parameters as θ_{ijk} is graph specific; the parents of each variable, and therefore also parent configurations, vary from one graph to another. We will define $\theta_{ij} = \{\theta_{ijk}\}_{k=1,\ldots,r_i}$ so we can talk about all the parameters for x_i given a fixed parent configuration.

Now, the prior distribution of each θ_{ij} has to be a Dirichlet:

$$P(\theta_{ij}) = \frac{\Gamma(\sum_k \alpha_{ijk})}{\prod_k \Gamma(\alpha_{ijk})} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1} = \text{Dirichlet}(\theta_{ij}; \alpha_{ij1}, \ldots, \alpha_{ijr_i}) \tag{12}$$

where, for integers, $\Gamma(z+1) = z!$. The mean of this distribution is

$$\int P(\theta_{ij})\, \theta_{ijk}\, d\theta_{ij} = \frac{\alpha_{ijk}}{\sum_{k'} \alpha_{ijk'}} \tag{13}$$

and it is more concentrated around the mean the larger the value of $\sum_{k'} \alpha_{ijk'}$. We can further write the hyper-parameters $\alpha_{ijk} > 0$ in the form $\alpha_{ijk} = n' p'_{ijk}$ where n' is the equivalent sample size specifying how many observations we need to balance the effect of the data on the estimates in comparison to the prior. There are two subtleties here. First, the number of available observations for estimating θ_{ij} varies with j, i.e., depends on how many times the parent configurations appear in the data. To keep n' as an equivalent sample size across all the parameters, we will have to account for this variation. The parameters p'_{ijk} are therefore not normalized to one across the values of variable x_i but across its values and the parent configurations: $\sum_{j=1}^{q_i} \sum_{k=1}^{r_i} p'_{ijk} = 1$ so we can interpret p'_{ijk} as a distribution over (x_i, x_{pa_j}). In other words, they include the expectation of how many times we would see a particular parent configuration in n' observations.

The second subtlety further constrains the values p'_{ijk}, in addition to the normalization. In order for the likelihood equivalence to hold, p'_{ijk} should be possible to interpret as marginals $P'(x_i, x_{pa_i})$ of some *common distribution* over all the variables $P'(x_1, \ldots, x_d)$ (common to all the graphs we consider). For example, simply normalizing p'_{ijk} across the parent configurations and the values of the variables does not ensure that they can be viewed as marginals from some common joint distribution P'. This subtlety does not often arise in practice. It is typical and easy to set them based on a uniform distribution so that

$$p'_{ijk} = \frac{1}{r_i \prod_{l \in pa_i} r_l} = \frac{1}{r_i q_i} \quad \text{or} \quad \alpha_{ijk} = \frac{n'}{r_i q_i} \tag{14}$$

This leaves us with only one hyper-parameter to set: n', the equivalent sample size.

We can now combine the data and the prior to obtain posterior estimates for the parameters. The prior factors across the variables and across parent configurations. Moreover, we assume that each observation is complete, containing a value assignment for all the variables in the model. As a result, we can evaluate the posterior probability over each θ_{ij} separately from others. Specifically, for each $\theta_{ij} = \{\theta_{ijk}\}_{k=1,\ldots,r_i}$, where i and j are fixed, we get

$$P(\theta_{ij}|D, G) \propto \left[\prod_{t:\, x^t_{pa_i} \to j} P(x^t_i | x^t_{pa_i}, \theta_{ij}) \right] P(\theta_{ij}) \tag{15}$$

$$= \left[\prod_{k=1}^{r_i} \theta_{ijk}^{n_{ijk}} \right] P(\theta_{ij}) \tag{16}$$

$$\propto \left[\prod_{k=1}^{r_i} \theta_{ijk}^{n_{ijk}} \right] \left[\prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1} \right] \tag{17}$$

$$= \prod_{k=1}^{r_i} \theta_{ijk}^{n_{ijk}+\alpha_{ijk}-1} \tag{18}$$

where the product in the first line picks out only observations where the parent configuration maps to j (otherwise the case would fall under the domain of another parameter vector). n_{ijk} specifies the number of observations where x_i had value k and its parents x_{pa_i} were in configuration j. Clearly, $\sum_{j=1}^{q_i} \sum_{k=1}^{r_i} n_{ijk} = n$. The posterior has the same form as the prior[1] and is therefore also Dirichlet, just with updated hyper-parameters:

$$P(\theta_{ij}|D, G) = \text{Dirichlet}(\theta_{ij}; \alpha_{ij1} + n_{ij1}, \ldots, \alpha_{ijr_i} + n_{ijr_i}) \tag{19}$$

[1] Dirichlet is a *conjugate prior* for the multi-nomial distribution.

The normalization constant for the posterior in Eq.(15) is given by

$$\int \left[\prod_{t:\, x^t_{pa_i} \to j} P(x^t_i | x^t_{pa_i}, \theta_{ij}) \right] P(\theta_{ij}) d\theta_{ij} = \frac{\Gamma(\sum_k \alpha_{ijk})}{\Gamma(\sum_k \alpha_{ijk} + \sum_k n_{ijk})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \quad (20)$$

This is also the marginal likelihood of data pertaining to x_i when x_{pa_i} are in configuration j. Since the observations are complete, and the prior is independent for each set of parameters, the marginal likelihood of all the data is simply a product of these local normalization terms. The product is taken across variables and across different parent configurations:

$$P(D|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_k \alpha_{ijk})}{\Gamma(\sum_k \alpha_{ijk} + \sum_k n_{ijk})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \quad (21)$$

We would now find a graph G that maximizes $P(D|G)$. Note that Eq.(21) is easy to evaluate for any particular graph by recomputing some of the counts n_{ijk}. We can further penalize graphs that involve a large number of parameters (or edges) by assigning a prior probability $P(G)$ over the graphs, and maximizing instead $P(D|G)P(G)$. For example, the prior could be some function of the number of parameters in the model or $\sum_{i=1}^{n}(r_i - 1)q_i$ such as

$$P(G) \propto \frac{1}{\sum_{i=1}^{n}(r_i - 1)q_i} \quad (22)$$