23 Learning Structure in Directed Graphs

Last time, we looked at parameter estimation for a directed graphical model given its graphical model structure. Today, we'll instead focus on learning graphical model structure for DAG's. As with the previous lecture, we provide both frequentist and Bayesian perspectives.

We'll stick to a setup where we have N random variables $x_1, x_2, \ldots, x_N$ that are nodes in a graphical model, where each x_i has alphabet size M. The number of possible edges is thus $\binom{N}{2} = N(N-1)/2$, and since an edge is either present or not in a graph, there are $2^{N(N-1)/2}$ possible graphs over the N nodes. Viewing each graph as a different model, we treat learning graphical model structure as a *model selection problem*. Of course, we'd be on a road to nowhere without any data, so we'll assume that we have at our disposal K i.i.d. observations of $\mathbf{x} = (x_1, \ldots, x_N)$. In particular, each observation has N values and no random variable is hidden. The two main questions that arise are:

- How do we score any particular model (i.e., graph) given data?
- How do we find a model with the highest score?

We present the frequentist perspective first before looking at the Bayesian perspective.

23.1 Frequentist Perspective: Likelihood Score

As a reminder, the frequentist perspective has the fundamental interpretation that the parameters of interest, which in our case are parameters $\theta_{\mathcal{G}}$ of the graphical model $\mathcal{G}$, are *deterministic but unknown*. With this interpretation, it does not make sense to talk about parameter $\theta_{\mathcal{G}}$ having a distribution, since it's deterministic!

Moving right along, note that for a fixed model (i.e., graph), evaluating the log likelihood for the model at the ML estimate for the model gives what we'll refer to as the *likelihood score*, which we could use to score a graph! In particular, it seems that the likelihood score should be higher for more plausible graphs. Let's formalize this idea. Recycling notation from the previous lecture, denote $\ell((\mathcal{G}, \theta_{\mathcal{G}}); D)$ to be the log likelihood of graphical model $(\mathcal{G}, \theta_{\mathcal{G}})$ evaluated on our observations D. Then, a best graph is one that is a solution to the following optimization problem:

$$\max_{\mathcal{G},\, \theta_{\mathcal{G}}} \ell((\mathcal{G}, \theta_{\mathcal{G}}); D) = \max_{\mathcal{G}} \underbrace{\max_{\theta_{\mathcal{G}}} \ell((\mathcal{G}, \theta_{\mathcal{G}}); D)}_{\substack{\text{involves ML estimation} \\ \text{given graph structure } \mathcal{G}}} = \max_{\mathcal{G}} \hat{\ell}(\mathcal{G}; D),$$

where $\hat{\ell}(\mathcal{G}; D) \triangleq \ell((\mathcal{G}, \hat{\theta}^{\mathrm{ML}}_{\mathcal{G}}); D)$ is the likelihood score.

We'll now look at what maximizing the likelihood score means for DAG's, for which we will relate the likelihood score to mutual information and entropy. First, we recall some definitions before forging on. The mutual information of two random variables u and v is given by

$$I(\mathsf{u};\mathsf{v}) \triangleq \sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \log \frac{p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{u}}(u)p_{\mathsf{v}}(v)},$$

which is symmetric (i.e., $I(\mathsf{u};\mathsf{v}) = I(\mathsf{v};\mathsf{u})$) and non-negative. The entropy of random variable u is given by

$$H(\mathsf{u}) \triangleq -\sum_{u} p_{\mathsf{u}}(u) \log p_{\mathsf{u}}(u) \geq 0.$$

We'll use the following formula:

$$\begin{aligned}
H(\mathsf{u}|\mathsf{v}) &= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \log p_{\mathsf{u}|\mathsf{v}}(u|v) \\
&= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \log \frac{p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{v}}(v)} \\
&= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \log \frac{p_{\mathsf{u}}(u)p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{u}}(u)p_{\mathsf{v}}(v)} \\
&= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \left[\log \frac{p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{u}}(u)p_{\mathsf{v}}(v)} + \log p_{\mathsf{u}}(u)\right] \\
&= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \log \frac{p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{u}}(u)p_{\mathsf{v}}(v)} - \sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \log p_{\mathsf{u}}(u) \\
&= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \log \frac{p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{u}}(u)p_{\mathsf{v}}(v)} - \sum_{u} p_{\mathsf{u}}(u) \log p_{\mathsf{u}}(u) \\
&= -I(\mathsf{u};\mathsf{v}) + H(\mathsf{u}). \qquad (1)
\end{aligned}$$

Finally, whenever we are looking at mutual information or entropy that comes from empirical distributions, we'll put on hats: $\hat{I}$ and $\hat{H}$.

For a DAG $\mathcal{G}$ with parameters $\theta_{\mathcal{G}}$, we have

$$p_{\mathsf{x}_1,\ldots,\mathsf{x}_N}(x_1,\ldots,x_N;\theta_{\mathcal{G}}) = \prod_{i=1}^{N} p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(x_i|x_{\pi_i};\theta^i_{\mathcal{G}}),$$

where π_i denotes the set of parents of node i and $\theta^i_{\mathcal{G}}$ denote parameters (i.e., table entries) of conditional probability table $p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}$; note that which nodes are parents of node i depends on the graph $\mathcal{G}$. We can more explicitly write out how conditional probability table $p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}$ relates to its parameters $\theta^i_{\mathcal{G}}$:

$$p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(x_i|x_{\pi_i};\theta^i_{\mathcal{G}}) = [\theta^i_{\mathcal{G}}]_{x_i,x_{\pi_i}}.$$

Assuming that parameters from different tables are independent, then we can build the likelihood score from likelihood scores of individual conditional probability tables:

$$\hat{\ell}(\mathcal{G}; D) = \sum_{i=1}^{N} \hat{\ell}_i(\mathcal{G}; D_i) \tag{2}$$

where D_i here refers to the data corresponding to random variables $\{\mathsf{x}_i, \mathsf{x}_{\pi_i}\}$. Recall that the ML estimate $\hat{\theta}^i_{\mathcal{G}}$ for the $\theta^i_{\mathcal{G}}$ is the empirical distribution:

$$[\hat{\theta}^i_{\mathcal{G}}]_{x_i, x_{\pi_i}} = \hat{p}_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(x_i|x_{\pi_i}) = \frac{\hat{p}_{\mathsf{x}_i,\mathsf{x}_{\pi_i}}(x_i, x_{\pi_i})}{\hat{p}_{\mathsf{x}_{\pi_i}}(x_{\pi_i})}.$$

Furthermore,

$$\hat{\ell}_i(\mathcal{G}; D_i) = \sum_{a,b} \hat{p}_{\mathsf{x}_i,\mathsf{x}_{\pi_i}}(a, b) \log \hat{p}_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(a|b) = -\hat{H}(\mathsf{x}_i|\mathsf{x}_{\pi_i}) = \hat{I}(\mathsf{x}_i; \mathsf{x}_{\pi_i}) - \hat{H}(\mathsf{x}_i), \tag{3}$$

where the last step uses equation (1). Putting together (2) and (3), we arrive at

$$\hat{\ell}(\mathcal{G}; D) = \sum_{i=1}^{N} \hat{\ell}_i(\mathcal{G}; D_i) = \sum_{i=1}^{N} \left\{ \hat{I}(\mathsf{x}_i; \mathsf{x}_{\pi_i}) - \hat{H}(\mathsf{x}_i) \right\} = \sum_{i=1}^{N} \hat{I}(\mathsf{x}_i; \mathsf{x}_{\pi_i}) - \underbrace{\sum_{i=1}^{N} \hat{H}(\mathsf{x}_i)}_{\text{independent of } \mathcal{G}}.$$

The main message here is that given a topological ordering (i.e., a node ordering where parents occur before children in the ordering), we can evaluate the likelihood score by summing up empirical mutual information terms while subtracting off the node entropies. Note that all graphs will have the same node entropies getting subtracted off! Hence, for comparing graphs, we only need to compare the sum of empirical mutual information terms per graph.

Example 1. (Chow-Liu 1968) Suppose we restrict ourselves to optimizing over trees across all N nodes. In this scenario, all models are equally complex in the sense that they each have $N - 1$ edges. Our earlier analysis suggests that all we need to do to find a tree with the highest likelihood score is to find one where the sum of the empirical mutual information terms is largest. This is just a maximum spanning tree problem!

In particular, we can just compute the empirical mutual information between every pair of nodes and then run Kruskal's algorithm: We sort the empirical mutual information values across all pairs. Then, starting from an empty (no edges) graph on the N nodes, we iterate through the edges from largest mutual information value to smallest, adding an edge if it does not introduce a cycle. We repeat this until we have $N - 1$ edges.

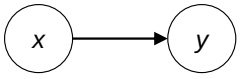
Care must be taken for assigning edge orientations to get a directed tree, since we do not want any node having more than one parent, which would have required looking at an empirical mutual information that involves more than two nodes.

Example 2. Our second example will reveal a critical pitfall of using the likelihood score to evaluate the quality of a model. In particular, what we'll see is that the likelihood score favors more complicated models!

Suppose we're choosing between graphs $\mathcal{G}_0$ and $\mathcal{G}_1$ shown below.

$\mathcal{G}_0$: 

$\hat{\ell}(\mathcal{G}_0; D) = -\hat{H}(x) - \hat{H}(y)$

$\mathcal{G}_1$: 

$\hat{\ell}(\mathcal{G}_1; D) = \hat{I}(x; y) - \hat{H}(x) - \hat{H}(y)$

Note that we have

$$\hat{\ell}(\mathcal{G}_1; D) - \hat{\ell}(\mathcal{G}_0; D) = \hat{I}(x; y) \geq 0,$$

which means that the more complicated model $\mathcal{G}_1$ will always be at least as good as $\mathcal{G}_0$. Thus, it is safe to always prefer $\mathcal{G}_1$. This phenomenon extends to larger models as well, where the likelihood score will favor more complex models. We'll address this by placing priors on parameters, bringing us into Bayesian statistics.

23.2 Bayesian Perspective: Bayesian Score

Unlike frequentists, Bayesians say that since we don't know parameters $\theta_{\mathcal{G}}$ of graph $\mathcal{G}$, then we might as well treat these parameters as random. In particular, we place a prior distribution $p(\mathcal{G})$ on graph $\mathcal{G}$ and a prior $p(\theta_{\mathcal{G}}|\mathcal{G})$ on parameters $\theta_{\mathcal{G}}$ given graph $\mathcal{G}$. Then the quality of a graph can be quantified by its posterior distribution evaluated on our data D. Applying Bayes' rule, the posterior is given by

$$p(\mathcal{G}|D) = \frac{p(D|\mathcal{G})p(\mathcal{G})}{p(D)}.$$

Since data D is observed and fixed, maximizing the above across all graphs is the same as just maximizing the numerator; in particular, we'll maximize the log of the numerator, which we'll call the *Bayesian score* ℓ_{B}:

$$\ell_{\mathrm{B}}(\mathcal{G}; D) = \log p(D|\mathcal{G}) + \log p(\mathcal{G}), \tag{4}$$

where

$$p(D|\mathcal{G}) = \int p(D|\mathcal{G}, \theta_{\mathcal{G}}) p(\theta_{\mathcal{G}}|\mathcal{G}) d\theta_{\mathcal{G}} \tag{5}$$

is called the *marginal likelihood*. Note, importantly, that we marginalize out random parameters $\theta_{\mathcal{G}}$.

It is possible to approximate the marginal likelihood using a *Laplace approximation*, which is beyond the scope of this course and involves approximating the integrand of (5) as a Gaussian. The end result is that we have

$$p(D|\mathcal{G}) \approx p(D|\hat{\theta}_{\mathcal{G}}^{\mathrm{ML}}, \mathcal{G}) \underbrace{p(\hat{\theta}_{\mathcal{G}}^{\mathrm{ML}}|\mathcal{G})\sigma_{\theta|D}}_{\text{Occam factor}},$$

where $\sigma_{\theta|D}$ is a width associated with the Gaussian and the Occam factor turns out to favor simpler models, alluding to Occam's razor.

23.2.1 Marginal Likelihoods with Unknown Distributions

We'll now study a graph with a single node, which may not seem particularly exciting, but in fact looking at the marginal likelihood for this single-node graph and comparing it with the likelihood score in an asymptotic regime will give us a good sense of how the Bayesian score differs from the likelihood score.

Suppose random variable x takes on values in $\{1, 2, \ldots, M\}$. Then x is a single-trial multinomial random variable (this single-trial case is also called a *categorical random variable*) with parameters $\theta = (\theta_1, \ldots, \theta_M)$. We choose prior $p_\theta(\theta; \alpha)$ to be the conjugate prior, i.e. θ is Dirichlet with parameters $\alpha = (\alpha_1, \ldots, \alpha_M)$. Then if we have K i.i.d. observations $x_1, \ldots, x_K$ of x, where we let $K(m)$ denote the number of times alphabet symbol m occurs, then

$$\begin{aligned}
p(D; \alpha) &= p_{\mathsf{x}_1, \ldots, \mathsf{x}_K}(x_1, \ldots, x_K) \\
&= \int p_\theta(\theta; \alpha) \prod_{k=1}^{K} p_{\mathsf{x}_k|\theta}(x_k|\theta) d\theta \\
&= \int p_\theta(\theta; \alpha) \prod_{k=1}^{K} \prod_{m=1}^{M} \theta_m^{\mathbf{1}\{x_k = m\}} d\theta \\
&= \int p_\theta(\theta; \alpha) \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_m^{\mathbf{1}\{x_k = m\}} d\theta \\
&= \int p_\theta(\theta; \alpha) \left(\prod_{m=1}^{M} \theta_m^{K(m)} \right) d\theta \\
&= \int \left(\frac{\Gamma(\sum_{i=1}^{M} \alpha_i)}{\prod_{i=1}^{M} \Gamma(\alpha_i)} \prod_{i=1}^{M} \theta_i^{\alpha_i - 1} \right) \left(\prod_{m=1}^{M} \theta_m^{K(m)} \right) d\theta \\
&= \frac{\Gamma(\sum_{i=1}^{M} \alpha_i)}{\prod_{i=1}^{M} \Gamma(\alpha_i)} \int \prod_{m=1}^{M} \theta_m^{\alpha_m + K(m) - 1} d\theta, \qquad (6)
\end{aligned}$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ is the gamma function. Note that $\Gamma(n) = (n-1)!$ for positive integer n, so it is a continuous extension of the factorial function. Next, observe that the integral in (6) just evaluates to be the partition function of a Dirichlet distribution with parameters $(\alpha_1 + K(1), \alpha_2 + K(2), \ldots, \alpha_M + K(M))$. Hence,

$$\int \prod_{m=1}^{M} \theta_m^{\alpha_m + K(m) - 1} d\theta = \frac{\prod_{i=1}^{M} \Gamma(\alpha_i + K(i))}{\Gamma(\sum_{i=1}^{M} (\alpha_i + K(i)))} = \frac{\prod_{i=1}^{M} \Gamma(\alpha_i + K(i))}{\Gamma(\sum_{i=1}^{M} \alpha_i + K)}. \qquad (7)$$

Stitching together (6) and (7), we get

$$p(D;\alpha) = \frac{\Gamma(\sum_{i=1}^M \alpha_i)}{\prod_{i=1}^M \Gamma(\alpha_i)} \frac{\prod_{i=1}^M \Gamma(\alpha_i + K(i))}{\Gamma(\sum_{i=1}^M \alpha_i + K)} = \frac{\Gamma(\sum_{i=1}^M \alpha_i)}{\Gamma(\sum_{i=1}^M \alpha_i + K)} \prod_{i=1}^M \frac{\Gamma(\alpha_i + K(i))}{\Gamma(\alpha_i)}. \quad (8)$$

We'll now see what happens with asymptotics. For simplicity, we shall assume $M = 2$ and $\alpha_1 = \alpha_2 = 1$. Effectively this means that x is Bernoulli except that we'll let it take on values in $\{1, 2\}$ instead of $\{0, 1\}$. Then using equation (8),

$$\begin{aligned}
p(D;\alpha) &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha + \alpha_2 + K)} \frac{\Gamma(\alpha_1 + K(1))}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_2 + K(2))}{\Gamma(\alpha_2)} \\
&= \frac{\Gamma(2)}{\Gamma(2 + K)} \frac{\Gamma(1 + K(1))}{\Gamma(1)} \frac{\Gamma(1 + K(2))}{\Gamma(1)} \\
&= \frac{K(1)!K(2)!}{(K+1)!} \\
&= \frac{K(1)!K(2)!}{(K+1)K!} \\
&= \frac{1}{K+1} \left[\frac{K!}{K(1)!K(2)!} \right]^{-1} \\
&= \frac{1}{K+1} \left[\binom{K}{K(1)} \right]^{-1}. \quad (9)
\end{aligned}$$

We shall use Stirling's approximation of the binomial coefficient:

$$\binom{n}{m} \approx e^{nH(\mathrm{Ber}(m/n))} \sqrt{\frac{n}{2\pi m(n-m)}}, \quad (10)$$

where $H(\mathrm{Ber}(m/n))$ denotes the entropy of a Bernoulli random variable with parameter m/n. Combining (9) and (10) results in

$$\begin{aligned}
p(D;\alpha) &= \frac{1}{K+1} \left[\binom{K}{K(1)} \right]^{-1} \\
&\approx \frac{1}{K+1} \left[e^{KH(\mathrm{Ber}(K(1)/K))} \sqrt{\frac{K}{2\pi K(1)K(2)}} \right]^{-1} \\
&= \frac{1}{K+1} e^{-KH(\mathrm{Ber}(K(1)/K))} \sqrt{\frac{2\pi K(1)K(2)}{K}} \\
&= \frac{1}{K+1} e^{-K\hat{H}(\mathsf{x})} \sqrt{\frac{2\pi K(1)K(2)}{K}}.
\end{aligned}$$

Taking the log of both sides yields the approximate Bayesian score where we note that there is no $p(\mathcal{G})$ term since there's only one graph possible:

$$\ell_{\mathrm{B}}(D) = \log p(D;\alpha) \approx -K\hat{H}(\mathsf{x}) - \log(K+1) + \frac{1}{2}\log(2\pi K(1)K(2)) - \frac{1}{2}\log K. \quad (11)$$

???

In contrast, the likelihood score is:

$$\begin{aligned}
\hat{\ell}(D) &= \log p(D; \hat{\theta}^{\mathrm{ML}}) \\
&= \log\{(\hat{\theta}_1^{\mathrm{ML}})^{K(1)}(\hat{\theta}_2^{\mathrm{ML}})^{K(2)}\} \\
&= \log\left\{\left(\frac{K(1)}{K}\right)^{K(1)}\left(\frac{K(2)}{K}\right)^{K(2)}\right\} \\
&= K(1)\log\frac{K(1)}{K} + K(2)\log\frac{K(2)}{K} \\
&= K\left(\frac{K(1)}{K}\log\frac{K(1)}{K} + \frac{K(2)}{K}\log\frac{K(2)}{K}\right) \\
&= -K\hat{H}(\mathsf{x}). \qquad (12)
\end{aligned}$$

Comparing (11) and (12), we see that the Bayesian score incurs an extra $-O(\log K)$ factor, which essentially penalizes a model when there is insufficient data. In fact, a more general result holds, which we'll mention at the end of lecture that shows that the Bayesian score will asymptotically penalize model complexity when we have a large number of samples K.

23.2.2 Marginal Likelihoods with DAG's

Let's return to Example 2 considered in the likelihood score setup and see how the Bayesian score differs. Assuming that $p(\mathcal{G}_0) = p(\mathcal{G}_1)$, then comparing the Bayesian score (4) just involves comparing the marginal likelihoods $p(D|\mathcal{G})$. Note that $\mathcal{G}_0$ has parameters $\theta_{\mathsf{x}}, \theta_{\mathsf{y}} \in [0,1]^M$, whereas $\mathcal{G}_1$ has parameters $\theta_{\mathsf{x}} \in [0,1]^M$ and $\theta_{\mathsf{y}} = (\theta_{\mathsf{y}|1}, \theta_{\mathsf{y}|2}, \ldots, \theta_{\mathsf{y}|M})$ with each $\theta_{\mathsf{y}|m} \in [0,1]^M$. Of course, each length-M parameter vector describing a distribution must sum to 1.

The marginal likelihood for $\mathcal{G}_0$ is given by

$$\begin{aligned}
p(D|\mathcal{G}_0) &= \int\int p(\theta_{\mathsf{x}}, \theta_{\mathsf{y}}|\mathcal{G}_0)p(D|\theta_{\mathsf{x}}, \theta_{\mathsf{y}}, \mathcal{G}_0)d\theta_{\mathsf{x}}d\theta_{\mathsf{y}} \\
&= \int\int p(\theta_{\mathsf{x}}|\mathcal{G}_0)p(\theta_{\mathsf{y}}|\mathcal{G}_0)\prod_{k=1}^{K} p(x_k|\theta_{\mathsf{x}}, \mathcal{G}_0)\prod_{k=1}^{K} p(y_k|\theta_{\mathsf{y}}, \mathcal{G}_0)d\theta_{\mathsf{x}}d\theta_{\mathsf{y}} \\
&= \int p(\theta_{\mathsf{x}}|\mathcal{G}_0)\prod_{k=1}^{K} p(x_k|\theta_{\mathsf{x}}, \mathcal{G}_0)d\theta_{\mathsf{x}} \int p(\theta_{\mathsf{y}}|\mathcal{G}_0)\prod_{k=1}^{K} p(y_k|\theta_{\mathsf{y}}, \mathcal{G}_0)d\theta_{\mathsf{y}}.
\end{aligned}$$

Meanwhile, the marginal likelihood for $\mathcal{G}_1$ is given by

$$
\begin{aligned}
&p(D|\mathcal{G}_1) \\
&= \int\int p(\theta_x, \theta_y|\mathcal{G}_1) p(D|\theta_x, \theta_y, \mathcal{G}_1) d\theta_x d\theta_y \\
&= \int\int p(\theta_x|\mathcal{G}_1) \left(\prod_{m=1}^{M} p(\theta_{y|m}|\mathcal{G}_1) \right) \prod_{k=1}^{K} \left\{ p(x_k|\theta_x, \mathcal{G}_1) p(y_k|x_k, \theta_{y|x_k}, \mathcal{G}_1) \right\} d\theta_x d\theta_y \\
&= \int\int \left(p(\theta_x|\mathcal{G}_1) \prod_{k=1}^{K} p(x_k|\theta_x, \mathcal{G}_1) \right) \left(\prod_{m=1}^{M} p(\theta_{y|m}|\mathcal{G}_1) \right) \prod_{k=1}^{K} p(y_k|x_k, \theta_{y|x_k}, \mathcal{G}_1) d\theta_x d\theta_y \\
&= \int\int \left(p(\theta_x|\mathcal{G}_1) \prod_{k=1}^{K} p(x_k|\theta_x, \mathcal{G}_1) \right) \left(\prod_{m=1}^{M} p(\theta_{y|m}|\mathcal{G}_1) \right) \\
&\qquad \times \prod_{m=1}^{M} \prod_{k=1}^{K} p(y_k|x_k, \theta_{y|x_k}, \mathcal{G}_1)^{\mathbf{1}\{x_k=m\}} d\theta_x d\theta_y \\
&= \int \left(p(\theta_x|\mathcal{G}_1) \prod_{k=1}^{K} p(x_k|\theta_x, \mathcal{G}_1) \right) d\theta_x \\
&\qquad \times \prod_{m=1}^{M} \int p(\theta_{y|m}|\mathcal{G}_1) \prod_{k=1}^{K} p(y_k|x_k, \theta_{y|x_k}, \mathcal{G}_1)^{\mathbf{1}\{x_k=m\}} d\theta_{y|m}.
\end{aligned}
$$

If we choose hyperparameters (which are the Dirichlet parameters α) of θ_x to be the same across the two models $\mathcal{G}_0$ and $\mathcal{G}_1$, then the first factors involving an integral over θ_x in marginal likelihoods $p(D|\mathcal{G}_0)$ and $p(D|\mathcal{G}_1)$ are the same, so we just need to compare the second factors; whichever is larger dictates the model favored by the Bayesian score:

$$
\int p(\theta_y|\mathcal{G}_0) \prod_{k=1}^{K} p(y_k|\theta_y, \mathcal{G}_0) d\theta_y \overset{?}{\lessgtr} \prod_{m=1}^{M} \int p(\theta_{y|m}|\mathcal{G}_1) \prod_{k=1}^{K} p(y_k|x_k, \theta_{y|x_k}, \mathcal{G}_1)^{\mathbf{1}\{x_k=m\}} d\theta_{y|m}.
$$

What will happen is that when we have only a little data, we will tend to prefer $\mathcal{G}_0$, and when we have more data, we will prefer $\mathcal{G}_1$ provided that x and y actually are correlated. The transition point at which we switch between favoring $\mathcal{G}_0$ and $\mathcal{G}_1$ depends on the strength of the correlation between x and y.

23.2.3 Approximations for large K

For large K, there is a general result that the Bayesian score satisfies

$$
\ell_{\mathrm{B}}(\mathcal{G}; D) \approx \underbrace{\hat{\ell}(\mathcal{G}; D) - \frac{\log K}{2} \dim \mathcal{G}}_{\text{Bayesian Information Criterion (BIC)}} + O(1),
$$

where $\dim \mathcal{G}$ is the number of independent parameters in model $\mathcal{G}$ and the trailing $O(1)$ constant does not depend on K. Note that $\hat{\ell}(\mathcal{G}; D)$ is the likelihood score and scales linearly with K (a specific example that showed this is in equation (12)). Importantly, the second term, which comes from the Occam factor, penalizes complex models. However, as the number of observations $K \to \infty$, since the likelihood score grows linearly in K, it will dominate the second term, pushing us toward the frequentist solution and finding the correct model provided that the correct model has nonzero probability under our prior $p(\mathcal{G})$.

6.438 Algorithms for Inference

Fall 2014