Bayesian_network_structure

December 9, 2022

1 Estimation of edge directionality from data

The following notes represent an expansion of an example from MacKay "*Information Theory, Inference, and Learning Algorithms*" concerning the Bayesian estimation of the edge direction between two binary nodes in a Bayesian network.

1.0.1 Background

Consider two variables A and B, both of which are represented by distinct nodes (also labelled A and B, respectively) in a Bayesian network. The general difficulty with estimating edge direction comes from the fact that the joint distribution may equally be factored in either of two ways, namely:

$$p(A, B) \quad = \quad p(A)\, p(B \mid A) \quad = \quad p(B)\, p(A \mid B)\,, \tag{1}$$

where the middle term represents the network $A \to B$, and the last term represents the network $B \to A$. Note that there is also a third possible network where A and B are independent and thus not connected by an edge, namely

$$p(A, B) \quad = \quad p(A)\, p(B)\,; \tag{2}$$

we examine independence in a later section.

Despite this apparent difficulty, MacKay briefly discusses the fact that a proper Bayesian treatment can indeed distinguish between the two hypotheses of edge directionality, as demonstrated by a simple example.

1.0.2 Categorical data

For convenience, suppose that A and B are categorical variables. Furthermore, let A have possible states $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$, and let B have states $\mathcal{B} = \{b_1, b_2, \ldots, b_m\}$. Consequently, we assume an empirically observed dataset of joint values, $D = [(a_{i_k}, b_{j_k})]_{k=1}^{N}$, where each $i_k \in \{1, 2, \ldots, n\}$ and $j_k \in \{1, 2, \ldots, m\}$. These data may be summarised by a table $\mathbf{C} \equiv \mathbf{C}_{A,B}$ of joint counts of the form:

A

B

$\mathbf{b_1}$

$\mathbf{b_2}$

$\ldots$

$\mathbf{b_m}$

$\mathbf{a_1}$

c_{11}

c_{12}

$\ldots$

c_{1m}

$\mathbf{a_2}$

c_{21}

c_{22}

$\ldots$

c_{2m}

$\vdots$

$\ldots$

$\mathbf{a_n}$

c_{n1}

c_{n2}

$\ldots$

c_{nm}

where c_{ij} denotes the number of times the pair of categories (a_i, b_j) appear in data D.

1.0.3 Ordering of variables

For a network with just the two nodes A and B, we may define the directed edge as being either $A \to B$ or $B \to A$. However, for a general network of multiple nodes, we may more loosely specify an ordering of the nodes, such that if node A appears earlier in the ordering than node B, denoted by the predicate $A \prec B$, then A is potentially an ancestor of B.

One possible benefit of this scheme is that the node ordering might be able to be estimated from the data D by considering all pairs of variables. Conceptually, given a collection of local node orderings, e.g. either $A \prec B$ or $B \prec A$, the overall ordering of all nodes could be established by topological sorting, provided that the local orderings preserve transitivity and do not induce cycles.

For our example two-node network, we consider the two distinct hypotheses that either $A \prec B$ or $B \prec A$.

1.0.4 Multinomial likelihoods

Under the hypothesis $A \prec B$, we suppose the existence of a generative model

$$p(A, B \mid A \prec B, \boldsymbol{\theta}_A, \boldsymbol{\Phi}_{B|A}) \quad = \quad p(A \mid A \prec B, \boldsymbol{\theta}_A)\, p(B \mid A, A \prec B, \boldsymbol{\Phi}_{B|A})\,, \tag{3}$$

such that pairs of values (A, B) are generated by first sampling A and then by sampling B conditionally on A.

Each value of variable A, say $A = a_i$, is sampled from the categorical distribution

$$p(A = a_i \mid A \prec B, \boldsymbol{\theta}_A) \quad = \quad \theta_{Ai}\,, \tag{4}$$

where $\boldsymbol{\theta}_A = (\theta_{A1}, \theta_{A2}, \ldots, \theta_{An})$ is the vector parameter of category probabilities, such that $\theta_{Ai} \geq 0$ and $|\boldsymbol{\theta}_A| \doteq \sum_{i=1}^{n} \theta_{Ai} \doteq \theta_{A\cdot} \equiv 1$.

The total number of occurrences of each category a_i in data D is then given by $c_{i\cdot} \doteq \sum_{j=1}^{m} c_{ij}$, using the table of counts defined in a previous section. We let the vector of marginal counts be denoted by $\mathbf{c}_A \doteq (c_{1\cdot}, c_{2\cdot}, \ldots, c_{n\cdot})$, which follows the multinomial distribution

$$p(\mathbf{c}_A \mid A \prec B, \boldsymbol{\theta}_A) \quad = \quad \frac{\Gamma(c_{\cdot\cdot} + 1)}{\prod_{i=1}^{n} \Gamma(c_{i\cdot} + 1)} \prod_{i=1}^{n} \theta_{Ai}^{c_{i\cdot}}\,, \tag{5}$$

where $c_{\cdot\cdot} \doteq \sum_{i=1}^{n} c_{i\cdot} \doteq |\mathbf{c}_A|$.

Likewise, for each sampled value of A, say $A = a_i$, a corresponding value of B, say $B = b_j$, is sampled from the conditional categorical distribution

$$p(B = b_j \mid A = a_i, A \prec B, \boldsymbol{\Phi}_{B|A}) \quad = \quad \phi_{(B|a_i)j}\,, \tag{6}$$

where $\boldsymbol{\Phi}_{B|A} = [\boldsymbol{\phi}_{B|a_i}^T]_{i=1}^{n}$ is a row-stochastic matrix, such that $\boldsymbol{\phi}_{B|a_i} \geq \mathbf{0}$ and $|\boldsymbol{\phi}_{B|a_i}| \doteq \sum_{j=1}^{m} \phi_{(B|a_i)j} \doteq \phi_{(B|a_i)\cdot} \equiv 1$.

Subsequently, given $A = a_i$, the associated vector of variable B category counts in data D is given by $\mathbf{c}_{a_i,B} \doteq (c_{i1}, c_{i2}, \ldots, c_{im})$, which has the distribution

$$p(\mathbf{c}_{a_i,B} \mid A = a_i, A \prec B, \boldsymbol{\Phi}_{B|A}) \quad = \quad \frac{\Gamma(c_{i\cdot} + 1)}{\prod_{j=1}^{m} \Gamma(c_{ij} + 1)} \prod_{j=1}^{m} \phi_{(B|a_i)j}^{c_{ij}}\,. \tag{7}$$

Note that since the matrix of counts, $\mathbf{C} \equiv \mathbf{C}_{A,B} \doteq [\mathbf{c}_{a_i,B}^T]_{i=1}^{n}$, completely summarises the observed data D, we take the data likelihood to be

$$p(D \mid A \prec B, \boldsymbol{\theta}_A, \boldsymbol{\Phi}_{B|A}) \quad = \quad p(\mathbf{c}_A \mid A \prec B, \boldsymbol{\theta}_A) \prod_{i=1}^{n} p(\mathbf{c}_{a_i,B} \mid A = a_i, A \prec B, \boldsymbol{\Phi}_{B|A})\,. \tag{8}$$

Conversely, under the alternative hypothesis $B \prec A$, we suppose the existence of a different generative model

$$p(A, B \mid B \prec A, \boldsymbol{\theta}_B, \boldsymbol{\Phi}_{A|B}) \quad = \quad p(B \mid B \prec A, \boldsymbol{\theta}_B)\, p(A \mid B, B \prec A, \boldsymbol{\Phi}_{A|B})\,, \tag{9}$$

where $\boldsymbol{\theta}_B = (\theta_{B1}, \theta_{B2}, \ldots, \theta_{Bm})$ is a stochastic vector of category probabilities satisfying $\boldsymbol{\theta}_B \geq \mathbf{0}$ and $|\boldsymbol{\theta}_B| \equiv 1$, and $\boldsymbol{\Phi}_{A|B} = [\boldsymbol{\phi}_{A|b_j}]_{j=1}^{m}$ is a column-stochastic matrix satisfying $\boldsymbol{\phi}_{A|b_j} \geq \mathbf{0}$ and $|\boldsymbol{\phi}_{A|b_j}| \equiv 1$.

This alternative generative process is now such that pair (A, B) is obtained first by sampling B, say $B = b_j$, from

$$p(B = b_j \mid B \prec A, \boldsymbol{\theta}_B) \quad = \quad \theta_{Bj}\,, \tag{10}$$

and then sampling A, say $A = a_i$, from

$$p(A = a_i \mid B = b_j, B \prec A, \boldsymbol{\Phi}_{A|B}) \quad = \quad \phi_{(A|b_j)i}\,. \tag{11}$$

The associated data counts are given by $\mathbf{c}_B \doteq (c_{\cdot 1}, c_{\cdot 2}, \ldots, c_{\cdot m})$, which follows the multinomial distribution

$$p(\mathbf{c}_B \mid B \prec A, \boldsymbol{\theta}_B) \quad = \quad \frac{\Gamma(c_{\cdot\cdot} + 1)}{\prod_{j=1}^{m} \Gamma(c_{\cdot j} + 1)} \prod_{j=1}^{m} \theta_{Bj}^{c_{\cdot j}}\,, \tag{12}$$

and $\mathbf{c}_{A,b_j} \doteq (c_{1j}, c_{2j}, \ldots, c_{nj})$, which follows the conditional multinomial distribution

$$p(\mathbf{c}_{A,b_j} \mid B = b_j, B \prec A, \boldsymbol{\Phi}_{A|B}) \quad = \quad \frac{\Gamma(c_{\cdot j} + 1)}{\prod_{i=1}^{n} \Gamma(c_{ij} + 1)} \prod_{i=1}^{n} \phi_{(A|b_j)i}^{c_{ij}}\,. \tag{13}$$

The alternative data likelihood is then

$$p(D \mid B \prec A, \boldsymbol{\theta}_B, \boldsymbol{\Phi}_{A|B}) \quad = \quad p(\mathbf{c}_B \mid B \prec A, \boldsymbol{\theta}_B) \prod_{j=1}^{m} p(\mathbf{c}_{A,b_j} \mid B = b_j, B \prec A, \boldsymbol{\Phi}_{A|B})\,. \tag{14}$$

1.0.5 Dirichlet priors

Under the hypothesis $A \prec B$, we suppose that prior to the sequence of A values being sampled, first the parameter $\boldsymbol{\theta}_A$ is sampled from a Dirichlet distribution, namely:

$$p(\boldsymbol{\theta}_A \mid A \prec B, \boldsymbol{\alpha}_A) \quad = \quad \frac{\Gamma(\alpha_{A\cdot})}{\prod_{i=1}^{n} \Gamma(\alpha_{Ai})} \prod_{i=1}^{n} \theta_{Ai}^{\alpha_{Ai}-1}\,, \tag{15}$$

where $\boldsymbol{\alpha}_A = (\alpha_{A1}, \alpha_{A2}, \ldots, \alpha_{An})$ and $\alpha_{A\cdot} \doteq \sum_{i=1}^{n} \alpha_{Ai} \doteq |\boldsymbol{\alpha}_A|$.

Coupled with the corresponding multinomial likelihood, it then follows from the Dirichlet-multinomial distribution that

$$p(\mathbf{c}_A \mid A \prec B, \boldsymbol{\alpha}_A) \quad = \quad \frac{\Gamma(|\boldsymbol{\alpha}_A|)\, \Gamma(|\mathbf{c}_A|)}{\Gamma(|\boldsymbol{\alpha}_A| + |\mathbf{c}_A|)} \prod_{i=1}^{n} \frac{\Gamma(\alpha_{Ai} + c_{Ai})}{\Gamma(\alpha_{Ai})\, \Gamma(c_{Ai})} \tag{16}$$

$$\quad = \quad \frac{\Gamma(\alpha_{A\cdot})\, \Gamma(c_{\cdot\cdot})}{\Gamma(\alpha_{A\cdot} + c_{\cdot\cdot})} \prod_{i=1}^{n} \frac{\Gamma(\alpha_{Ai} + c_{i\cdot})}{\Gamma(\alpha_{Ai})\, \Gamma(c_{i\cdot})}\,. \tag{17}$$

Similarly, prior to the B values being sampled, each parameter $\boldsymbol{\phi}_{B|a_i}$ is first sampled from the Dirichlet distribution

$$p(\boldsymbol{\phi}_{B|a_i} \mid A = a_i, A \prec B, \boldsymbol{\mathcal{A}}_{B|A}) = \frac{\Gamma(\alpha_{(B|A)i\cdot})}{\prod_{j=1}^{m}\Gamma(\alpha_{(B|A)ij})}\prod_{j=1}^{m}\phi_{(B|a_i)j}^{\alpha_{(B|A)ij}-1}, \tag{18}$$

where $\boldsymbol{\mathcal{A}}_{B|A} \doteq \left[\boldsymbol{\alpha}_{(B|A)i}^{T}\right]_{i=1}^{n}$ is a matrix of pseudo-counts, with $\boldsymbol{\alpha}_{(B|A)i} \doteq (\alpha_{(B|A)i1}, \alpha_{(B|A)i2}, \ldots, \alpha_{(B|A)im})$. The corresponding Dirichlet-multinomial distribution is therefore

$$p(\mathbf{c}_{a_i,B} \mid A = a_i, A \prec B, \boldsymbol{\mathcal{A}}_{B|A}) = \frac{\Gamma(\alpha_{(B|A)i\cdot})\,\Gamma(c_{i\cdot})}{\Gamma(\alpha_{(B|A)i\cdot}+c_{i\cdot})}\prod_{j=1}^{m}\frac{\Gamma(\alpha_{(B|A)ij}+c_{ij})}{\Gamma(\alpha_{(B|A)ij})\,\Gamma(c_{ij})}. \tag{19}$$

Putting these together, we obtain

$$\begin{aligned}
p(D \mid A \prec B, \boldsymbol{\alpha}_A, \boldsymbol{\mathcal{A}}_{B|A}) &= p(\mathbf{c}_A \mid A \prec B, \boldsymbol{\alpha}_A)\prod_{i=1}^{n} p(\mathbf{c}_{a_i,B} \mid A = a_i, A \prec B, \boldsymbol{\mathcal{A}}_{B|A}) && (20)\\
&= \frac{\Gamma(\alpha_{A\cdot})\,\Gamma(c_{\cdot\cdot})}{\Gamma(\alpha_{A\cdot}+c_{\cdot\cdot})}\prod_{i=1}^{n}\frac{\Gamma(\alpha_{Ai}+c_{i\cdot})}{\Gamma(\alpha_{Ai})\,\Gamma(c_{i\cdot})}\prod_{i=1}^{n}\left\{\frac{\Gamma(\alpha_{(B|A)i\cdot})\,\Gamma(c_{i\cdot})}{\Gamma(\alpha_{(B|A)i\cdot}+c_{i\cdot})}\prod_{j=1}^{m}\frac{\Gamma(\alpha_{(B|A)ij}+c_{ij})}{\Gamma(\alpha_{(B|A)ij})\,\Gamma(c_{ij})}\right\} && (21)\\
&= \frac{\Gamma(\alpha_{A\cdot})\,\Gamma(c_{\cdot\cdot})}{\Gamma(\alpha_{A\cdot}+c_{\cdot\cdot})}\prod_{i=1}^{n}\left\{\frac{\Gamma(\alpha_{Ai}+c_{i\cdot})\,\Gamma(\alpha_{(B|A)i\cdot})}{\Gamma(\alpha_{Ai})\,\Gamma(\alpha_{(B|A)i\cdot}+c_{i\cdot})}\prod_{j=1}^{m}\frac{\Gamma(\alpha_{(B|A)ij}+c_{ij})}{\Gamma(\alpha_{(B|A)ij})\,\Gamma(c_{ij})}\right\}. && (22)
\end{aligned}$$

Likewise, under the alternative hypothesis $B \prec A$, we follow similar reasoning to obtain

$$\begin{aligned}
p(D \mid B \prec A, \boldsymbol{\alpha}_B, \boldsymbol{\mathcal{A}}_{A|B}) &= p(\mathbf{c}_B \mid B \prec A, \boldsymbol{\alpha}_B)\prod_{j=1}^{m} p(\mathbf{c}_{A,b_j} \mid B = b_j, B \prec A, \boldsymbol{\mathcal{A}}_{A|B}) && (23)\\
&= \frac{\Gamma(\alpha_{B\cdot})\,\Gamma(c_{\cdot\cdot})}{\Gamma(\alpha_{B\cdot}+c_{\cdot\cdot})}\prod_{j=1}^{m}\left\{\frac{\Gamma(\alpha_{Bj}+c_{\cdot j})\,\Gamma(\alpha_{(A|B)\cdot j})}{\Gamma(\alpha_{Bj})\,\Gamma(\alpha_{(A|B)\cdot j}+c_{\cdot j})}\prod_{i=1}^{n}\frac{\Gamma(\alpha_{(A|B)ij}+c_{ij})}{\Gamma(\alpha_{(A|B)ij})\,\Gamma(c_{ij})}\right\} && (24)
\end{aligned}$$

1.0.6 Uniform priors

How do we simplify these data likelihoods under each hypothesis?

Firstly, we note that there are N data-points in D, so that $c_{\cdot\cdot} = N$. Secondly, if we have no good reasons for preferring one category over another, then we could set all of the pseudo-counts to a constant value. If we choose a value of unity, then the Dirichlet priors reduce to uniform priors.

Under these conditions, we now obtain

$$\begin{aligned}
p(D \mid A \prec B) &= \frac{\Gamma(n)\,\Gamma(N)}{\Gamma(n+N)}\prod_{i=1}^{n}\left\{\frac{\Gamma(1+c_{i\cdot})\,\Gamma(m)}{\Gamma(1)\,\Gamma(m+c_{i\cdot})}\prod_{j=1}^{m}\frac{\Gamma(1+c_{ij})}{\Gamma(1)\,\Gamma(c_{ij})}\right\} && (25)\\
&= \frac{\Gamma(N)\,\Gamma(n)\,\Gamma(m)^n}{\Gamma(N+n)}\prod_{i=1}^{n}\frac{\Gamma(c_{i\cdot}+1)}{\Gamma(c_{i\cdot}+m)}\prod_{i=1}^{n}\prod_{j=1}^{m}\frac{\Gamma(c_{ij}+1)}{\Gamma(c_{ij})}, && (26)
\end{aligned}$$

and

$$p(D \mid B \prec A) = \frac{\Gamma(m)\,\Gamma(N)}{\Gamma(m+N)} \prod_{j=1}^{m} \left\{ \frac{\Gamma(1+c_{\cdot j})\,\Gamma(n)}{\Gamma(1)\,\Gamma(n+c_{\cdot j})} \prod_{i=1}^{n} \frac{\Gamma(1+c_{ij})}{\Gamma(1)\,\Gamma(c_{ij})} \right\} \tag{27}$$

$$= \frac{\Gamma(N)\,\Gamma(m)\,\Gamma(n)^m}{\Gamma(N+m)} \prod_{j=1}^{m} \frac{\Gamma(c_{\cdot j}+1)}{\Gamma(c_{\cdot j}+n)} \prod_{j=1}^{m} \prod_{i=1}^{n} \frac{\Gamma(c_{ij}+1)}{\Gamma(c_{ij})}\,. \tag{28}$$

1.0.7 Hypothesis testing

We are now in a position to compare the relaive evidence for each hypothesis. Firstly, observe that

$$\frac{p(A \prec B \mid D)}{p(B \prec A \mid D)} = \frac{p(D \mid A \prec B)\,p(A \prec B)\,/\,p(D)}{p(D \mid B \prec A)\,p(B \prec A)\,/\,p(D)} = \frac{p(D \mid A \prec B)\,p(A \prec B)}{p(D \mid B \prec A)\,p(B \prec A)}\,. \tag{29}$$

Secondly, if we have no prior reason to prefer one hypothesis to the other, then we may choose $p(A \prec B) = p(B \prec A)$, such that

$$\frac{p(A \prec B \mid D)}{p(B \prec A \mid D)} \doteq \frac{p(D \mid A \prec B)}{p(D \mid B \prec A)}\,. \tag{30}$$

Hence, substituting in the respective likelihoods for uniform priors, we obtain

$$\frac{p(A \prec B \mid D)}{p(B \prec A \mid D)} \doteq \frac{\Gamma(N+m)\,\Gamma(m)^{n-1}}{\Gamma(N+n)\,\Gamma(n)^{m-1}} \prod_{i=1}^{n} \frac{\Gamma(c_{i\cdot}+1)}{\Gamma(c_{i\cdot}+m)} \Bigg/ \prod_{j=1}^{m} \frac{\Gamma(c_{\cdot j}+1)}{\Gamma(c_{\cdot j}+n)}\,. \tag{31}$$

Note that in the binary example of MacKay, we have $n = m = 2$, whereupon the ratio reduces to

$$\frac{p(A \prec B \mid D)}{p(B \prec A \mid D)} \doteq \frac{\prod_{j=1}^{2}(c_{\cdot j}+1)}{\prod_{i=1}^{2}(c_{i\cdot}+1)}\,, \tag{32}$$

which agrees with the solution provided by MacKay.

1.0.8 Independence of variables

As mentioned briefly in the introduction, there exists the possibility of a special case where two variables are actually independent of each other, and therefore have no edge between them in the Bayesian network.

In practice, exact independence, even if it exists, is rarely demonstrated in data due to sampling fluctuations. In general, we would need to use an approximate test such as using χ^2 or covariances to detect independence.

How does the above ratio test perform on independent data? For simplicitly, let us invent data such that variable A is either 0 or 1 with probabilities 0.1 and 0.9, respectively. Independently, variable B is also either 0 or 1 with probabilities 0.4 and 0.6, respectively. For $N = 1000$ observations, the counts are given by the table:

A	B		
	0.4	*0.6*	**sum**
0.1	40	60	**100**
0.9	360	540	**900**
sum	**400**	**600**	**1000**

On these data, the ratio test gives

$$\frac{p(A \prec B \mid D)}{p(B \prec A \mid D)} \doteq \frac{(400+1)\,(600+1)}{(100+1)\,(900+1)} \approx \frac{8}{3}\,. \tag{33}$$

In other words, even though the counts are exactly independent, the test suggests that hypothesis $A \prec B$ is two to three times as likely as hypothesis $B \prec A$.

In practical terms, does it matter if an edge is created between two supposedly independent variables? Surely if the parent-child dependence is weak, then the model should also demonstrate weak dependence, e.g. by inferring similar model parameters for different categories. Conversely, if the dependence is strong, then the model parameters should be quite different for different categories.

However, the problem still remains of attempting to interpret the presence and direction of edges as implying some sort of causality, especially if an edge is spurious.