Contents

0.1 Inferring a three variable Bayesian network

I reparametrise the joint probability of the graph by replacing the conditional probability $P(child|parent)$ to be the quotient of two joint distirbution $\frac{P(child,parent)}{P(parent)}$. Because $P(child|parent)$ enters the marginalised likelihood as a Beta-Binomial probability, $P(parent)$ enters the term as a Drichlet-Multinomial probability:

$$P(x_k) = \frac{(n!)\,\Gamma\left(\sum \alpha_k\right)}{\Gamma\left(n + \sum \alpha_k\right)} \prod_{k=1}^{K} \frac{\Gamma(x_k + \alpha_k)}{(x_k!)\,\Gamma(\alpha_k)}$$

where $\sum x_k = N$ is the partition of sample into k categories, α_k is the imaginary sample size for each category (also known as prior concentration). This formulation has the advantage of easier coding.

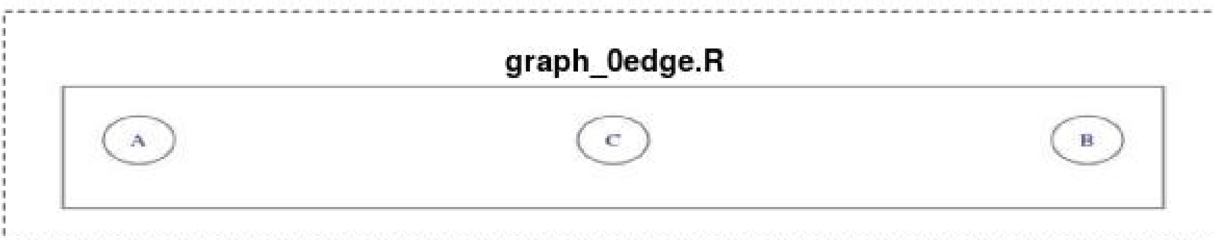
To be consistent with bnlearn and deal, I did discared the multinomial terms in the calculation, leading to

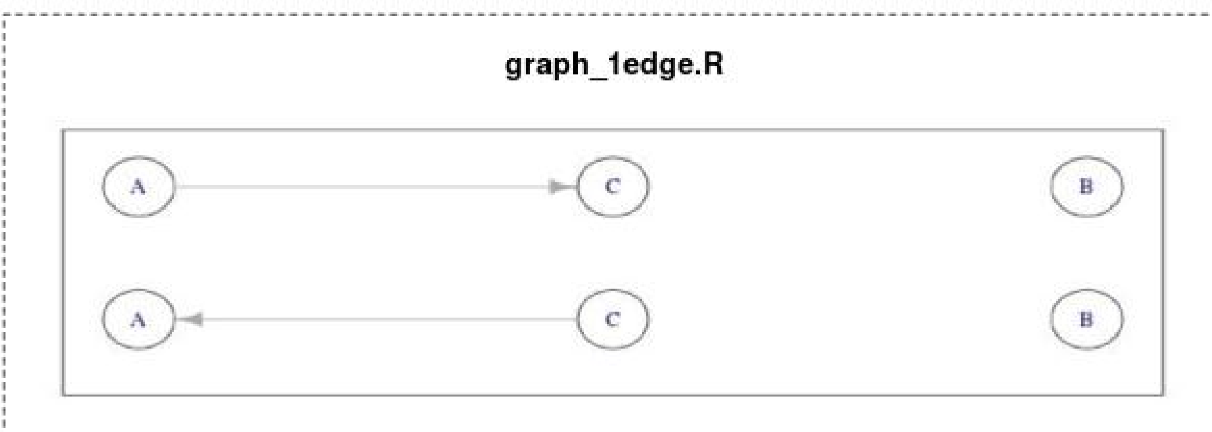
$$P(x_k) = \frac{\Gamma\left(\sum \alpha_k\right)}{\Gamma\left(n + \sum \alpha_k\right)} \prod_{k=1}^{K} \frac{\Gamma(x_k + \alpha_k)}{\Gamma(\alpha_k)}$$

0.2 Number of Bayesian networks:

A V-variable network has $V(V-1)/2$ bivariate interaction (edges), each interaction can have 3 possible status (A->B, A<-B, A B). Hence altogether there are $n(V) = 3^{V(V-1)/2}$ possible networks. For $V = 3$,$n(3) = 27$

However, for this exerecise, the serach space is restircted to the graph set $G = \{$no-edge, A-C only, A-C and B-C$\}$.





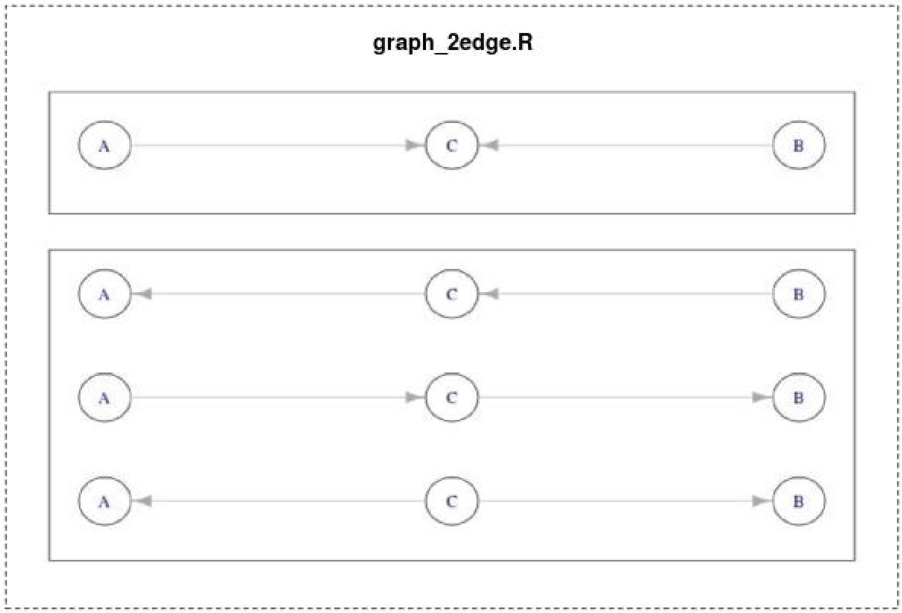


Figure 1: Graph with two edges but not A-B

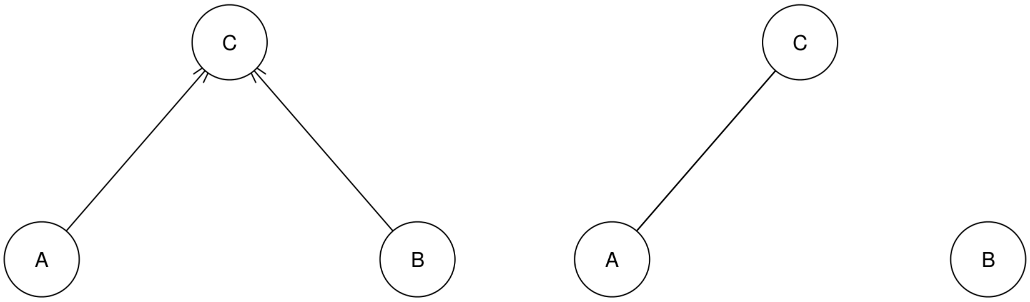


Figure 2: Best networks inferred using "bnlearn::pc.stable". Left: Dataset1. Right: Dataset2

0.2.1 Comment on the likelihood-equivalent prior

The likelihood-equivalent prior is set so that the imaginary sample size decreases as data is stratified by more variables. For example, if $P(A = 1) \sim Beta(\eta(A_0), \eta(A_1))$, then the imaginary sample size for (A=1) is $\eta(A_1) = 2$. Hence if we then ask for $P(B = 1 \mid A = 1) \sim Beta(\eta(B_0 A_1), \eta(B_1 A_1))$, the imaginary η's must add up to the imageinary sample size of the condition $\mid A = 1$, (aka $\eta(B_0 A_1) + \eta(B_1 A_1) = \eta(A_1) = 2$). Assuming two events are equally probable gives $\eta(B_0 A_1) = \eta(B_1 A_1) = 1$. For 3 variable, we can deduce $8\eta(ABC) = 4\eta(AB) = 2\eta(A) = \eta(0)$, setting $\eta(ABC) = 1$ gives $\eta(ABC) = 1, \eta(AB) = 2, \eta(A) = 4, \eta(0) = 8$, corresponding to different levels of stratification.

If a likelihood-preserving prior is used, then it is only the correlation structure that determines the relative feasibility of different graphs. Consider the 1-edge and 0-edge examples, the 0-edge example asserts $P(A \mid C = 0) = P(A \mid C = 1) = P(A)$, whereas the 1-edge example implies $P(A \mid C = 0) \neq P(A \mid C = 1)$, allowing an additional degree of freedom. The striking fact is that this additional DOF does not necessarily leads to a better model, in constrast to conventional mixture models where additional components always reduce likelihood. One of the reason is that the partiaion of $(A_0) = (A_0 C_0) + (A_0 C_1)$ is not arbitratry, but the general case is still confusing. A possible intution is that the additional DOF project the paramteric space to a higher dimension where the likelihood function overlaps less with the prior distribution.

0.2.2 Drawbacks of binary bayesian networks

If there are hidden latent variables in the bayes net, for example where the common parent of A and B (which is C) is conceived from the observers, then one will have to consider a graph with hidden variable in order to explain the data. In other words, a graphical prior needs to accommodate additional nodes to explain such data. Even though this is the case, it will be hard to express the case where n(A__0)=n(B__0)

0.2.3 Effect of imaginary sample size

Here we consider two imaginary sample sizes $\eta(0) = 8$ and $\eta(0) = 1$. A higher η indicates a sharper distribution of binomial probability θ (Setting $\eta(0) = 1$ implies $\eta(ABC) = 0.125, \eta(AB) = 0.25, \eta(A) = 0.5, \eta(0) = 1$)

The corresponding likelihood are calculated for both datasets (dat1 and dat2, see table 1).

- For dat1, the chain network (A-B-C) is the best at ISS=1, the A-B..C network is the best at ISS=8.
- For dat2, the chain network (A-B-C) is the best for ISS=1 and ISS=8

The prediction made by pc.stable is somewhat different (figure 2)

0.2.4 Plot posteiror for P(C|A) in different models

Here I visulise the posterior distribution using dataset 1 only. In order to show how different graphical models lead to different likelihood, I chose to contrast P(C|A) between [A][B][C](0-edge model) and [A][B][C|A] (1-edge model)

In 0-edge model, $P(C|A) = P(C)$ and the distribution is indifferent for $A = 0$ and $A = 1$ (figure 3). The term enters likelihood function as a beta-binomial.

In contrast, the 1-edge model prescribes that $P(A, C) \neq P(A)P(C)$, and two separate distribution must be considered for $P(C|A)$ (figure 4). The $\prod_{C,A} P(C|A)$ term factors out to be $\prod_{C|A=0} P(C|A = 0) \prod_{C|A=0} P(C|A = 1)$, as the product of two beta-binomial with independent probability but the same prior. It would be interesting to explore the precise condition under which the factored likelihood exceeds the original single beta-binomial. Clearly, the 0-edge model fails to capture the difference between $P(C|A = 0)$ and $P(C|A = 1)$ (loglik=-194.3, compared to 1-edge loglik=-175.5), but the underlying mathematics remains to be dissected.

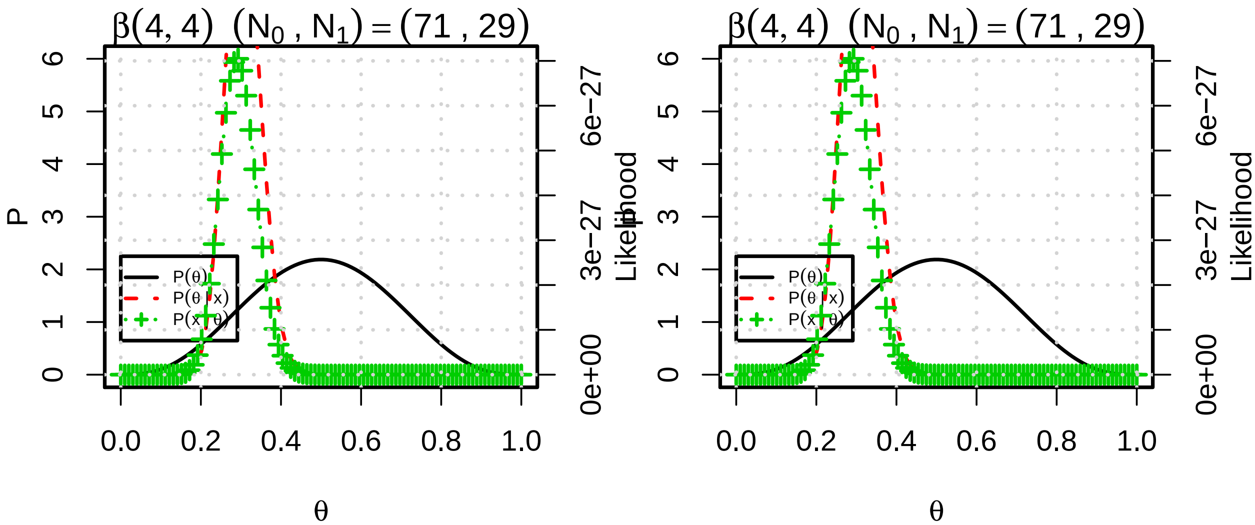


Figure 3: P(C|A)=P(C) according to the 0-edge graph, Left: P(C|A=0). Right: P(C|A=1)

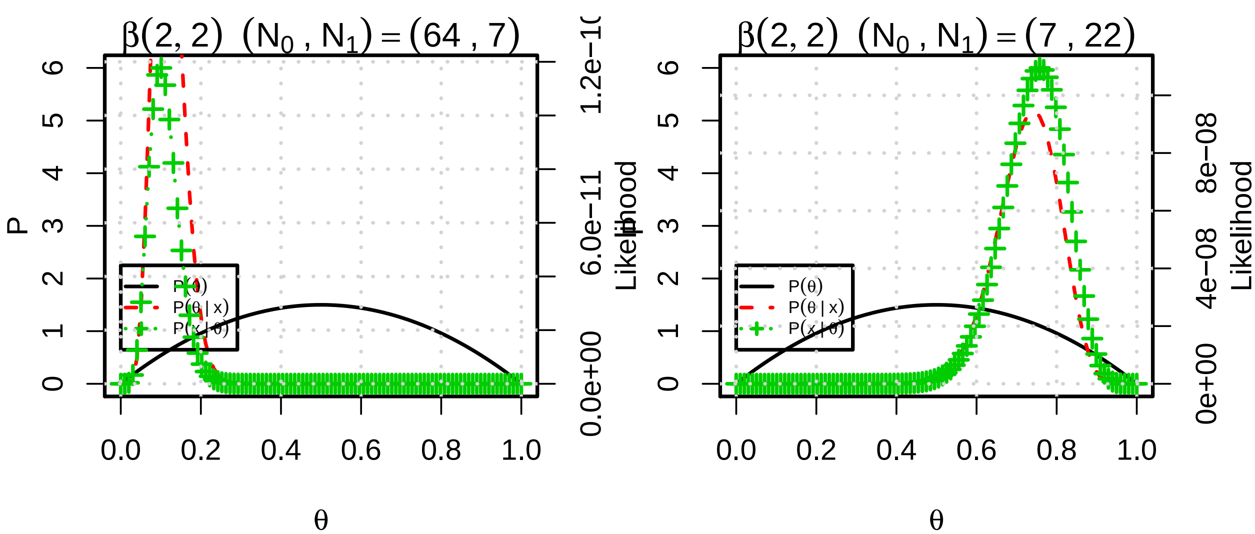


Figure 4: P(C|A) needs to be stratified according to the 1-edge graph, Left: P(C|A=0). Right: P(C|A=1)

Table 1: Marginalised likelihood of different network topology

model	myalgo.bde.iss	bnlearn.bde.iss	bnlearn.bic	iss	dat
[A][B][C]	-194.323	-194.323	-195.931	8	dat1
[A][B][C|A]	-175.536	-175.536	-176.906	8	dat1
[B][C][A|C]	-175.536	-175.536	-176.906	8	dat1
[B][C|B][A|C]	-168.817	-168.817	-170.591	8	dat1
[A][C|A][B|C]	-168.817	-168.817	-170.591	8	dat1
[C][A|C][B|C]	-168.817	-168.817	-170.591	8	dat1
[A][B][C|A:B]	-168.819	-168.819	-170.894	8	dat1
[A][B][C]	-196.617	-196.617	-195.931	1	dat1
[A][B][C|A]	-177.715	-177.715	-176.906	1	dat1
[B][C][A|C]	-177.715	-177.715	-176.906	1	dat1
[B][C|B][A|C]	-171.742	-171.742	-170.591	1	dat1
[A][C|A][B|C]	-171.742	-171.742	-170.591	1	dat1
[C][A|C][B|C]	-171.742	-171.742	-170.591	1	dat1
[A][B][C|A:B]	-170.894	-170.894	-170.894	1	dat1
[A][B][C]	-195.699	-195.699	-197.408	8	dat2
[A][B][C|A]	-180.318	-180.318	-181.025	8	dat2
[B][C][A|C]	-180.318	-180.318	-181.025	8	dat2
[B][C|B][A|C]	-178.579	-178.579	-179.994	8	dat2
[A][C|A][B|C]	-178.579	-178.579	-179.994	8	dat2
[C][A|C][B|C]	-178.579	-178.579	-179.994	8	dat2
[A][B][C|A:B]	-180.568	-180.568	-183.104	8	dat2
[A][B][C]	-198.093	-198.093	-197.408	1	dat2
[A][B][C|A]	-181.803	-181.803	-181.025	1	dat2
[B][C][A|C]	-181.803	-181.803	-181.025	1	dat2
[B][C|B][A|C]	-181.165	-181.165	-179.994	1	dat2
[A][C|A][B|C]	-181.165	-181.165	-179.994	1	dat2
[C][A|C][B|C]	-181.165	-181.165	-179.994	1	dat2
[A][B][C|A:B]	-183.684	-183.684	-183.104	1	dat2