

Appendix: confirmation measures and the difficulty about conjunction

Marcello Di Bello and Rafal Urbaniak

Contents

Bayesian networks and probabilistic independence	1
Bayes factor: claims and proofs	4
Bayes factor: simulations	6
Likelihood ratio: claims and proofs	7
Likelihood ratio: simulations	9

Bayesian networks and probabilistic independence

One assumption often made in the formulation of the conjunction paradox is that claims A and B are probabilistically independent. This is not always the case—we have seen that the paradox does subside even if the two claims are dependent. However, two fairly natural set-ups for conjunctive hypotheses and evidence supporting them (Figure 1) indeed do have some independencies built in, and so it is also natural what can be said about the conjunction problem given these independencies. Moreover, in some context, we will be freely using the independence assumptions, as a counterexample to a general claim still remains one even if it satisfies an additional requirement, that is, independence conditions.

We will be considering the conjunction of two hypotheses, A and B , their respective pieces of evidence a and b , and their conjunction AB , in two set-ups, illustrated by the Bayesian networks shown in Figure 1. The key difference here is that we allow direct dependence between the hypotheses in the second network. In both Bayesian networks, the CPT for the conjunction trivially mirrors the one for conjunction, as in Table 1.

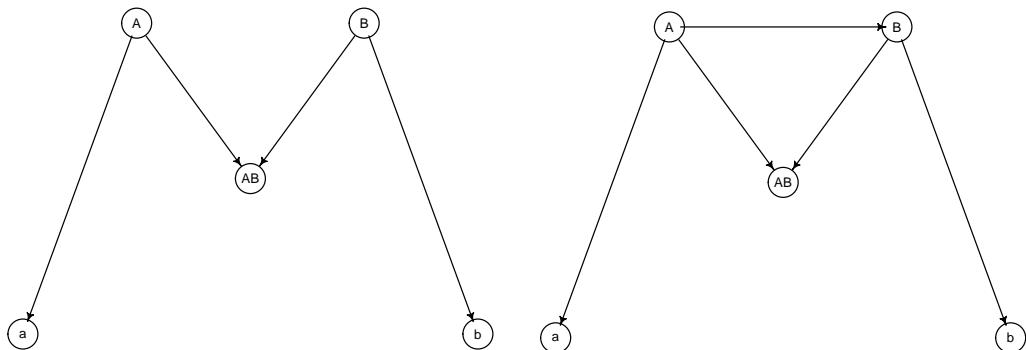


fig:conjunctionBNs}

(a) DAG1

(b) DAG2

Figure 1: Two DAGs for the conjunction problem.

AB	A	B	Pr
1	1	1	1
0	1	1	0
1	0	1	0
0	0	1	1
1	1	0	0
0	1	0	1
1	0	0	0
0	0	0	1

ab:CPTconjunction2}

Table 1: Conditional probability table for the conjunction node.

Directed Acyclic Graphs (DAGs) are useful for representing graphically these relationships of independence. The edges, intuitively, are meant to capture direct influence between the nodes. The role that such direct influence plays is that in a Bayesian network built over a DAG any node is conditionally independent of its nondescendants (including ancestors), given its parents. If this is the case for a given probabilistic measure $P()$ and a given DAG, we say that $P()$ is compatible with G , and they can be put together to constitute a Bayesian network.

The graphical counterpart of probabilistic independence is the so-called **d-separation**, $\perp\!\!\!\perp_d$. We say that two nodes, X and Y , are d-separated given a set of nodes Z — $X \perp\!\!\!\perp_d Y | Z$ — iff for every undirected path from X to Y there is a node Z' on the path such that either:

- $Z' \in Z$ and there is a *serial* connection, $\rightarrow Z' \rightarrow$, on the path,
- $Z' \in Z$ and there is a diverging connection, $\leftarrow Z' \rightarrow$, on the path,
- There is a connection $\rightarrow Z' \leftarrow$ on the path, and neither Z' nor its descendants are in Z .

Finally, two sets of nodes, X and Y , are d-separated given Z if every node in X is d-separated from every node in Y given Z . With serial connection, for instance, if:

Node	Proposition
G	The suspect is guilty.
B	The blood stain comes from the suspect.
M	The crime scene stain and the suspect's blood share their DNA profile.

We naturally would like to have the connection $G \rightarrow B \rightarrow M$. If we don't know whether B holds, G seems to have an indirect impact on the probability of M . Yet, once we find out that B is true, we expect the profile match, and whether G holds has no further impact on the probability of M .

Take an example of a diverging connections. Say you have two coins, one fair, one biased. Conditional on which coin you have chosen, the results of subsequent tosses are independent. But if you don't know which coin you have chosen, the result of previous tosses give you some information about which coin it is, and this has impact on your estimate of the probability of heads in the next toss. Whether a coin is fair, F or not has an impact on the result of the first toss, H_1 , and on the result of the second toss, H_2 . So $H_1 \leftarrow F \rightarrow H_2$ seems to be appropriate. Now, on one hand, as long as we don't know whether F , H_1 increases the probability of H_2 . On the other, once we know that F , though, H_1 and H_2 become independent, and so conditioning on the parent in a fork makes its children independent (provided there is no other open path between them in the graph).

For converging connections, let G and B be as above, and let:

Node	Proposition
O	The crime scene stain comes from the offender.

Both G and O influence B . If he's guilty, it's more likely that the blood stain comes from him, and if the blood crime stain comes from the offender it is more likely to come from the suspect (for instance, more so than if it comes from the victim). Moreover, G and O seem independent – whether the suspect is guilty doesn't have any bearing on whether the stain comes from the offender. Thus, a converging connection $G \rightarrow B \leftarrow O$ seem appropriate. However, if you do find out that B is true, that the stain comes from the suspect, whether the crime stain comes from the offender becomes relevant for whether the suspect is guilty.

One important reason why d-separation matters is that it can be proven that if two sets of nodes are d-separated given a third one, then they are independent given the third one, for any probabilistic measures compatible with a given DAG. Interestingly, lack of d-separation doesn't entail dependence for any probabilistic measure compatible with a given DAG. Rather, it only allows for it: if nodes are d-separated, there is at least one probabilistic measure fitting the DAG according to which they are independent. So, at least, no false independence can be inferred from the DAG, and all the dependencies are built into it.

Now, getting back to the conjunction problem, the d-separations entailed by these networks differ (examples can be found in Table 2)—in fact, DAG1 entails 31 d-separations, while DAG2 entails 22 of them. Attention should be paid to the notation. In the above, variables represent nodes, and so each d-separation entails a probabilistic statement about all combination of the node states involved. For instance, assuming each node is binary with two possible states, 1 and 0, $B \perp\!\!\!\perp_d a$ entails that for any $B_i, a_i \in \{0, 1\}$ we have $P(B = B_i) = P(B = B_i | a = a_i)$.

Bayesian network 1	Bayesian network 2
$A \perp\!\!\!\perp_d B$	$A \perp\!\!\!\perp_d b B$
$A \perp\!\!\!\perp_d b$	$AB \perp\!\!\!\perp_d a A$
$AB \perp\!\!\!\perp_d a A$	$AB \perp\!\!\!\perp_d b B$
$AB \perp\!\!\!\perp_d b B$	$B \perp\!\!\!\perp_d a A$
$B \perp\!\!\!\perp_d a$	$a \perp\!\!\!\perp_d b B$
$a \perp\!\!\!\perp_d b$	$a \perp\!\!\!\perp_d b A$

{tab:indepBNS}

Table 2: Some of d-separations entailed by DAG1 and DAG2 in the conjunction problem. One minimal testable implication (with the smallest possible conditioning set) is returned per missing edge of the graph.

In what follows, however, we will sometimes use a finer level of granularity, being very explicit on what independence assumptions are used in the derivations. In such contexts, we will be talking about states rather than nodes, and so when we present the derivation, $b \perp\!\!\!\perp A \wedge a \mid \neg B$ is a claim about events (or propositions) means the same as $P(b = 1 | B = 0) = P(b = 1 | A = 1, a = 1, B = 0)$. This distinction matters, as, first, independence conditional on $B = 0$ doesn't entail independence given $B = 1$ (for instance, your final grade might depend on how hard you work if the teacher is fair, but this might fail if the teacher is not fair), and, second, sometimes only some of the independencies entailed by a Bayesian network will be actually required for a given claim to hold, and we want to be explicit about such cases. We hope this slight ambiguity in notation will cause no confusion, as whether we talk about nodes or events will be clear from the context. So, moving to events, here is a list of independence claims used in the arguments that follow. We also marked whether they are entailed by the DAG under consideration.

{eq:indAB}	$A \perp\!\!\!\perp B$	DAG1	(1)
{eq:indab}	$b \perp\!\!\!\perp a$	DAG1	(2)
{eq:I1}	$A \perp\!\!\!\perp b a$	DAG1	(3)
{eq:I2}	$B \perp\!\!\!\perp a \wedge A b$	DAG1 , DAG2	(4)
{eq:I3}	$a \perp\!\!\!\perp b A \wedge B$	DAG1 , DAG2	(5)
{eq:I3a}	$a \perp\!\!\!\perp b A$	DAG1 , DAG2	(6)
{eq:I4}	$a \perp\!\!\!\perp B A$	DAG1 , DAG2	(7)
{eq:I4a}	$a \perp\!\!\!\perp B \neg A$	DAG1 , DAG2	(8)
{eq:I4b}	$a \perp\!\!\!\perp \neg B A$	DAG1 , DAG2	(9)
{eq:I4c}	$a \perp\!\!\!\perp \neg B \neg A$	DAG1 , DAG2	(10)
{eq:I5}	$b \perp\!\!\!\perp A \wedge a B$	DAG1 , DAG2	(11)
{eq:I5a}	$b \perp\!\!\!\perp \neg A \wedge a B$	DAG1 , DAG2	(12)
{eq:I5b}	$b \perp\!\!\!\perp A \wedge a \neg B$	DAG1 , DAG2	(13)
{eq:I5c}	$b \perp\!\!\!\perp \neg A \wedge a \neg B$	DAG1 , DAG2	(14)
{eq:I6}	$b \perp\!\!\!\perp a B$	DAG1 , DAG2	(15)

In what follows, we will provide theorems where we managed to obtain them. However, in some cases calculations are somewhat unmanageable, and so we decided to inspect such issues by means of a simulation. Moreover, even for cases in which we obtained the relevant proofs, the results of a simulation provide further insights, for instance, in terms of relative frequencies of various facts. For each DAG under consideration, we generated 100k random Bayesian networks which share that DAG. For each of these networks, we calculated all the relevant probabilities, Bayes factors, and likelihood ratios for each of them. With the output of such calculations, various questions can be asked and the answers visualized.

Alicja: R: Double check if we in fact use I3, or if there are any unused independence assumptions listed above.

Bayes factor: claims and proofs

We will start with the Bayes factor $P(E|H)/P(E)$ and its relation to conjunction. Let's abbreviate:

$$BF_A = \frac{P(a|A)}{P(a)},$$

$$BF_B = \frac{P(b|B)}{P(b)}$$

{fac:BFindep} **Fact 1.** If the independence assumptions (1), (2), (7) and (11) hold, then $BF_{AB} = BF_A \times BF_B$.

Proof.

$$\begin{aligned} \frac{P(a \wedge b|A \wedge B)}{P(a \wedge b)} &= \frac{P(A \wedge B \wedge a \wedge b)}{P(A \wedge B)} / P(a \wedge b) && \text{(conditional probability)} \\ &= \frac{P(A) \times P(B|A) \times P(a|A \wedge B) \times P(b|A \wedge B \wedge a)}{P(A \wedge B)} / P(a \wedge b) && \text{(chain rule)} \end{aligned}$$

Now, let's deploy the respective independence assumptions, as follows:

$$\begin{aligned}
&= \frac{\underbrace{\mathbb{P}(B)}_{\mathbb{P}(A) \times \mathbb{P}(B) \text{ by (1)}} \times \underbrace{\mathbb{P}(a|A)}_{\mathbb{P}(A \wedge B)} \times \underbrace{\mathbb{P}(b|B)}_{\mathbb{P}(A \wedge B \wedge a)} \times \underbrace{\mathbb{P}(a \wedge b)}_{\mathbb{P}(a) \times \mathbb{P}(b) \text{ by (2)}}}{\mathbb{P}(A \wedge B)} \\
&= \frac{\mathbb{P}(A) \times \mathbb{P}(B) \times \mathbb{P}(a|A) \times \mathbb{P}(b|B)}{\mathbb{P}(A) \times \mathbb{P}(B)} \times \mathbb{P}(a) \times \mathbb{P}(b) \\
&= \frac{\mathbb{P}(a|A)}{\mathbb{P}(a)} \times \frac{\mathbb{P}(b|B)}{\mathbb{P}(b)} \\
&= BF_A \times BF_B
\end{aligned}$$

□

This fact has the following straightforward consequence which simply follows from the simple fact that if $a = b \times c$ and $b, c > 1$, then $a > \max(b, c)$.

Corollary 1.1. *If the independence assumptions (1), (2), (7) and (11) hold, and BF_A and BF_B are both greater than 1, then BF_{AB} is greater than one. In fact, BF_{AB} is then greater than both BF_A and BF_B .*

Note that the claim has its negative counterpart.

Corollary 1.2. *If the independence assumptions (1), (2), (7) and (11) hold, and BF_A and BF_B are both strictly less than 1, then BF_{AB} is less than $\min(BF_A, BF_B)$.*

Now, does a similar claim if we drop the independence assumption specific to DAG1? A claim somewhat weaker than Fact 1 can be proven without employing independencies entailed by DAG1, relying only on some independencies entailed by DAG2. Let's abbreviate:

$$\begin{aligned}
BF'_B &= \frac{\mathbb{P}(b|B)}{\mathbb{P}(b|a)} \\
BF'_A &= \frac{\mathbb{P}(a|A)}{\mathbb{P}(a|b)}
\end{aligned}$$

Fact 2. *If (7) and (11) hold, then $BF_{AB} = BF_A \times BF'_B = BF_B \times BF'_A$.*

Proof. We start with the definition of conditional probability and the chain rule, as in the proof of Fact 1, but now we use fewer independencies (all of them entailed by DAG2), as we use the chain rule instead in a few cases.

$$\begin{aligned}
\frac{\mathbb{P}(a \wedge b|A \wedge B)}{\mathbb{P}(a \wedge b)} &= \frac{\mathbb{P}(A) \times \mathbb{P}(B|A) \times \underbrace{\mathbb{P}(a|A \wedge B)}_{\mathbb{P}(A \wedge B)} \times \underbrace{\mathbb{P}(b|B \wedge a)}_{\mathbb{P}(a) \times \mathbb{P}(b|a) \text{ by the chain rule}}}{\mathbb{P}(A \wedge B)} \times \underbrace{\mathbb{P}(a \wedge b)}_{\mathbb{P}(a) \times \mathbb{P}(b) \text{ by the chain rule}} \\
&= \frac{\mathbb{P}(a|A)}{\mathbb{P}(a)} \times \frac{\mathbb{P}(b|B)}{\mathbb{P}(b|a)} \\
&= BF_A \times BF'_B
\end{aligned}$$

If instead of obtaining $\mathbb{P}(a)\mathbb{P}(b|a)$ in the denominator we deploy the chain rule differently, resulting in $\mathbb{P}(b)\mathbb{P}(a|b)$, we end up with:

$$\begin{aligned}
&= \frac{\mathbb{P}(a|A)}{\mathbb{P}(a|b)} \times \frac{\mathbb{P}(b|B)}{\mathbb{P}(b)} \\
&= BF'_A \times BF_B
\end{aligned}$$

□

Now, to obtain a corollary analogous to Corollary 1.1, we need an additional assumption: that either $P(b|a) \leq P(b)$ or $P(a|b) \leq P(b)$.

Corollary 2.1. Suppose (7) and (11) hold and $BF_{BA}, BF_B > 1$. Then if either $P(b|a) \leq P(b)$ or $P(a|b) \leq P(b)$, we have $BF_{AB} > BF_A, BF_B$.

Proof. Notice that by Fact 2 we have that on the assumptions of the corollary:

$$\begin{aligned} BF_{AB} &= BF_A \times BF'_B \\ &= BF_B \times BF'_A \end{aligned}$$

If we now can show that either $BF'_B > 1$, or $BF'_A > 1$, the reasoning we used before applies: the product of two numbers greater than one is greater than either of them. Consider BF_B , which we assumed to be greater than 1. If we can show $BF'_B \geq BF_B$, we're done. But this holds on one of the disjuncts assumed in the corollary, $P(b|a) \leq P(b)$, as then we have:

$$\frac{P(b|B)}{P(b|a)} \geq \frac{P(b|B)}{P(b)}$$

Similarly, if the other disjunct holds, we run an analogous argument, this time focusing on BF'_A . \square

Bayes factor: simulations

add negative simulation result for DAG2, analogous to the ones for DAG1

First, let's look at simulations based on DAG1. We can illustrate the distribution of Bayes factors in the two separate scenarios used as assumptions of the above corollaries.

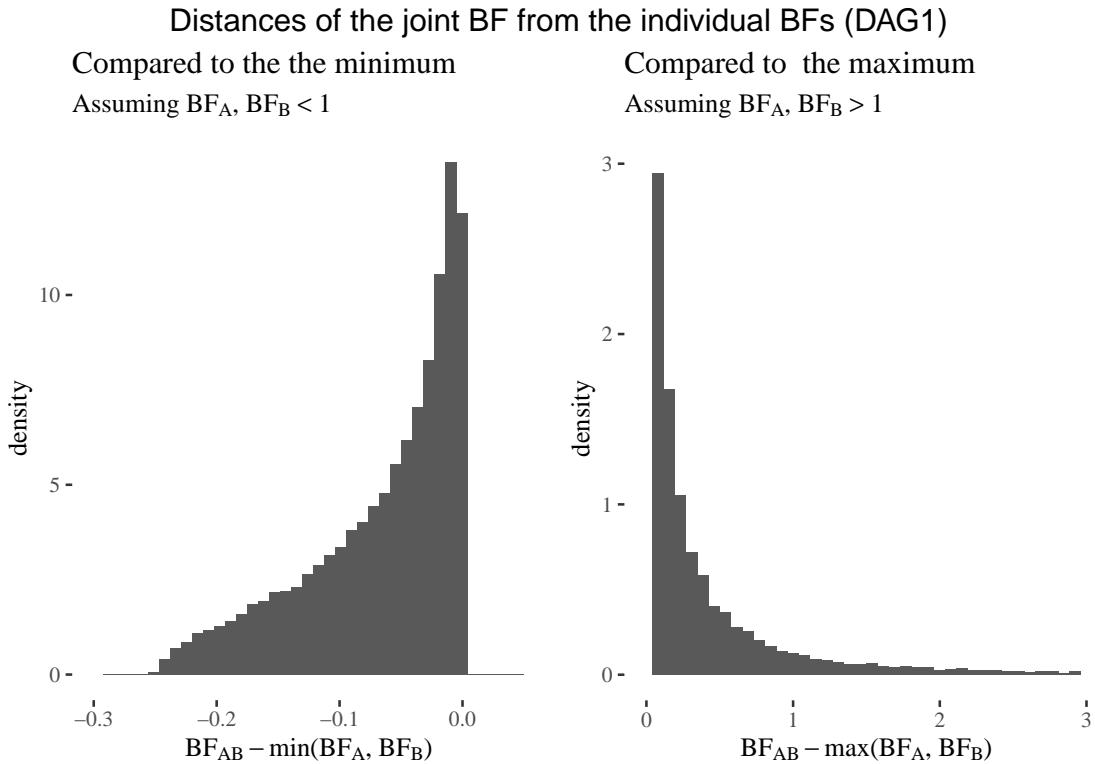
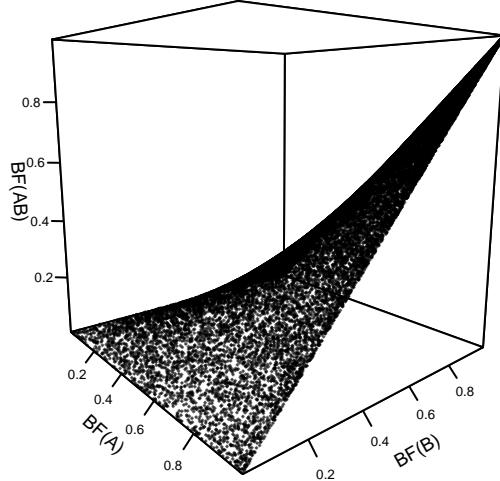


Figure 2: Distances of the joint Bayes factor from maxima and minima of individual Bayes factors, depending on whether the individual support levels are both positive or both negative. Simulation based on 100k Bayesian networks build over the DAG of DAG1.

For the DAG corresponding to DAG1, the simulated frequency of cases in which $BF_{AB} < BF_A, BF_B$ is 25% (which is twice higher than for the likelihood ratio), and the structure of such cases is visualized in Figure 3.

Cases in which $BF(AB) < BF(A), BF(B)$ (frequency=.25)



{fig:BFfails}

Figure 3: Ca. 25k cases (out of simulated 100k) in which the joint BF is below each of the individual BFs.

The picture doesn't change when we move to DAG2 (Figure ??). What does change, however, is that the joint BF is no longer the result of multiplying the individual BFs. Nevertheless, the values are fairly close (Figure 5).

Likelihood ratio: claims and proofs

Now, let's turn to the likelihood understood as:

$$\frac{P(E|H)}{P(E|\neg H)} = \frac{\text{sensitivity}}{1 - \text{specificity}}$$

Let's introduce the following abbreviations:

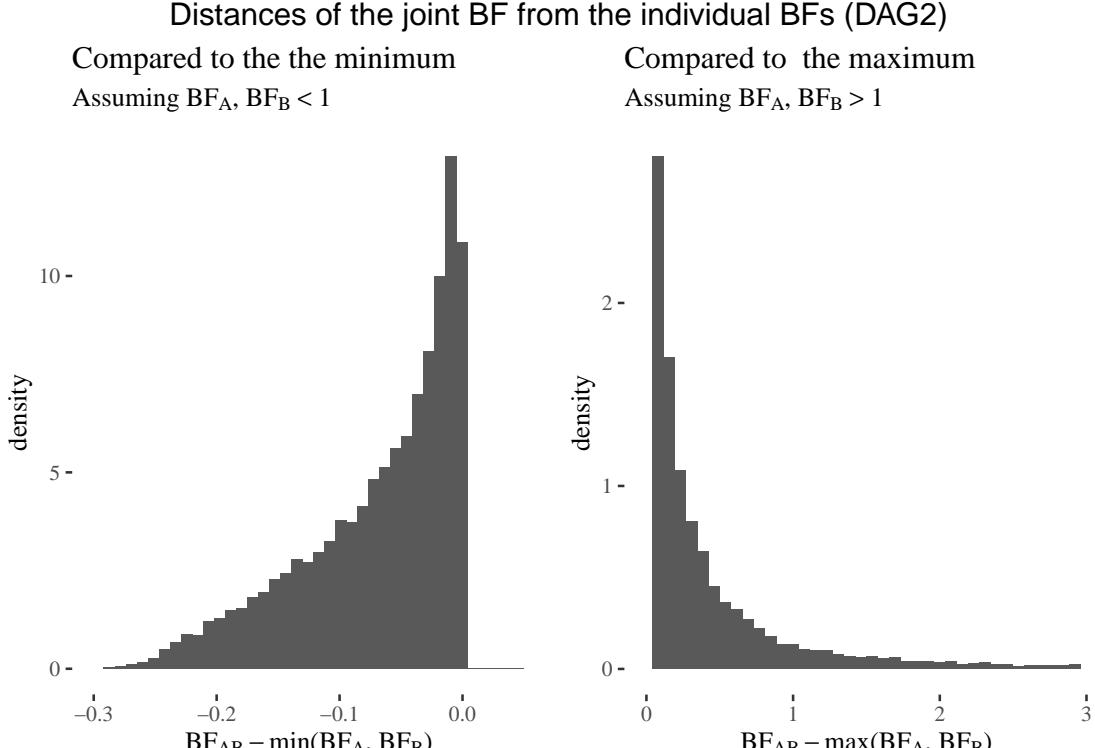
$$\begin{aligned} LR_{AB} &= \frac{P(a \wedge b | a \wedge B)}{P(a \wedge b | \neg(a \wedge B))} \\ LR_A &= \frac{P(a|A)}{P(a|\neg A)} \\ LR_B &= \frac{P(b|B)}{P(b|\neg B)}. \end{aligned}$$

Fact 3. If independence conditions (7), (8), (9), (10), (11), (12), (13), and (14) hold, then:

$$LR_{AB} = \frac{P(a|A) \times P(b|B)}{P(\neg A)P(\neg B|\neg A)P(a|\neg A)P(b|\neg B) + P(A)P(\neg B|A)P(a|A)P(b|\neg B) + P(\neg A)P(\neg B|\neg A)P(a|\neg A)P(b|\neg B)}$$

Note that these independence assumptions are entailed not only in DAG1, but also in DAG2.

Proof. Let's first compute the numerator of LR_{AB} :



labelfig:BFind2

Figure 4: Distances of the joint Bayes factor from maxima and minima of individual Bayes factors, depending on whether the individual support levels are both positive or both negative. Simulation based on 100k Bayesian networks build over the DAG of DAG2.

$$\begin{aligned}
 P(a \wedge b | A \wedge B) &= \frac{P(A \wedge B \wedge a \wedge b)}{P(A \wedge B)} && \text{(conditional probability)} \\
 &= \frac{P(A) \times P(B|A) \times P(a|A \wedge B) \times P(b|A \wedge B \wedge a)}{P(A) \times P(B|A)} && \text{(chain rule)}
 \end{aligned}$$

We deploy the relevant independencies as follows:

$$\begin{aligned}
 &\frac{P(a|A) \text{ by (7)} \quad P(b|B) \text{ by (11)}}{P(A) \times P(B|A)} \\
 &= \frac{P(A) \times P(B|A) \times \overbrace{P(a|A \wedge B)}^{\text{by (7)}} \times \overbrace{P(b|A \wedge B \wedge a)}^{\text{by (11)}}}{P(A) \times P(B|A)} \\
 &= P(a|A) \times P(b|B) && \text{(algebraic manipulation)}
 \end{aligned}$$

The denominator of LR_{AB} is more complicated, mostly because of the conditioning on $\neg(A \wedge B)$.

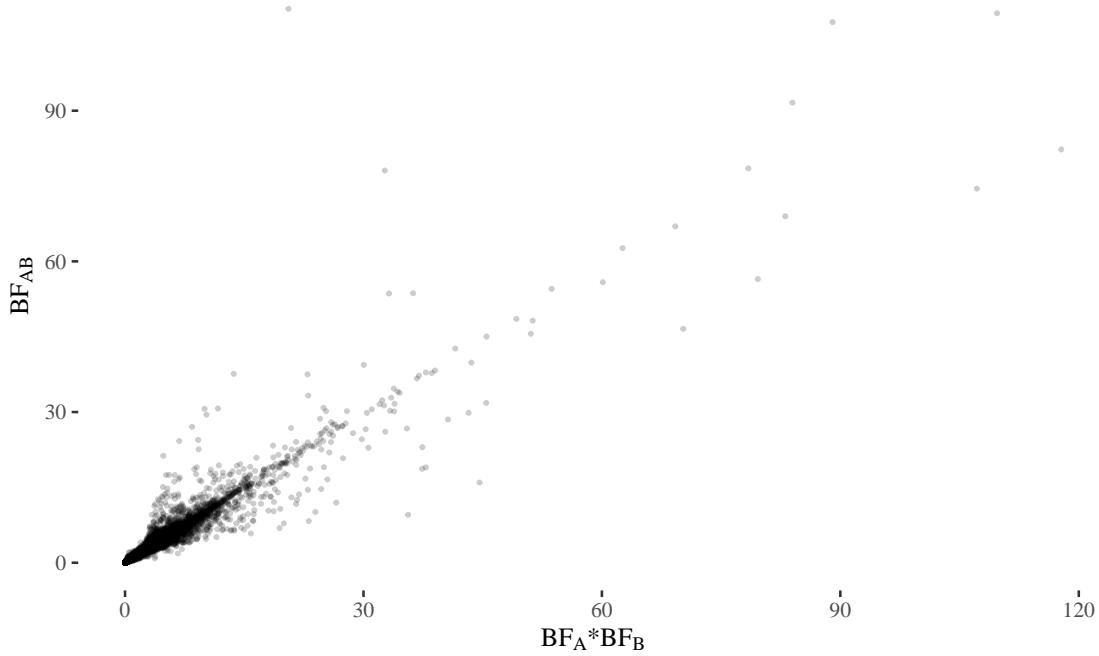
$$\begin{aligned}
 P(a \wedge b | \neg(A \wedge B)) &= \frac{P(a \wedge b \wedge \neg(A \wedge B))}{P(\neg(A \wedge B))} && \text{(conditional probability)} \\
 &= \frac{P(a \wedge b \wedge \neg A \wedge B) + P(a \wedge b \wedge A \wedge \neg B) + P(a \wedge b \wedge \neg A \wedge \neg B)}{P(\neg A \wedge B) + P(A \wedge \neg B) + P(\neg A \wedge \neg B)} && \text{(logic \& additivity)}
 \end{aligned}$$

Now consider the first summand from the numerator:

$$\begin{aligned}
 P(a \wedge b \wedge \neg A \wedge B) &= P(\neg A)P(B|\neg A)P(a|\neg A \wedge B)P(b|a \wedge \neg A \wedge B) && \text{(chain rule)} \\
 &= P(\neg A)P(B|\neg A)P(a|\neg A)P(b|B) && \text{(independencies (8) and (12))}
 \end{aligned}$$

Multiplicative claim fails for DAG2

Pearson's correlation coefficient = .95



{fig:BFmulti}

Figure 5: In DAG2, the result of multiplying individual BFs does not equal the joining BF, but often is a good approximation thereof.

The simplification of the other two summands is analogous (albeit with slightly different independence assumptions—(9) and (13) for the second one and (10) and (14) for the third. Once we plug these into the denominator formula we get:

$$\begin{aligned} P(a \wedge b | \neg(A \wedge B)) &= \frac{P(\neg A)P(B|\neg A)P(a|\neg A)P(b|B) + P(A)P(\neg B|A)P(a|A)P(b|\neg B) + P(\neg A)P(\neg B|\neg A)P(a|\neg A)P(b|\neg B)}{P(\neg A)P(B|\neg A) + P(A)P(\neg B|A) + P(\neg A)P(\neg B|\neg A)} \\ &= \frac{P(\neg A)P(B|\neg A)P(a|\neg A)P(b|B) + P(A)P(\neg B|A)P(a|A)P(b|\neg B) + P(\neg A)P(\neg B|\neg A)P(a|\neg A)P(b|\neg B)}{P(\neg A)P(B|\neg A) + P(A)P(\neg B|A) + P(\neg A)P(\neg B|\neg A)} \end{aligned}$$

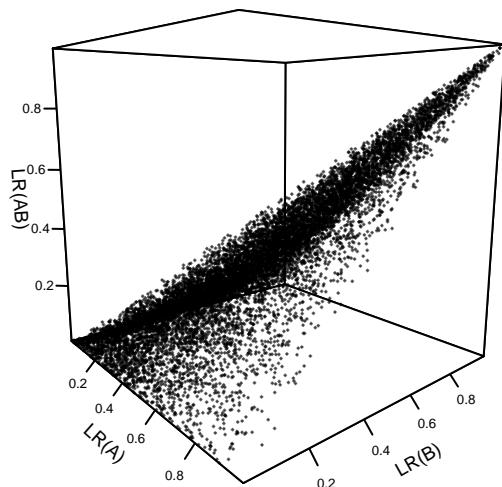
□

Likelihood ratio: simulations

While the analytic approach turns out to be cumbersome, let's inspect the problem using simulations. First of all, there are cases in which the joint likelihood ratio are lower than each of the individual likelihood ratios. Their frequency is twice lower than the corresponding frequency for the Bayes factor (recall Figure 3). For the DAG corresponding to both DAGs, the simulated frequency of cases in which $LR_{AB} < LR_A, LR_B$ is 12-12.5% and the distribution of such cases, somewhat of a different shape, is visualized in Figure 6 (the picture for DAG2 is very similar).

Interestingly, even if the individual likelihood ratios are < 1 , the joint likelihood ratio can be higher than their minimum, but is never higher than their maximum (Figure 7).

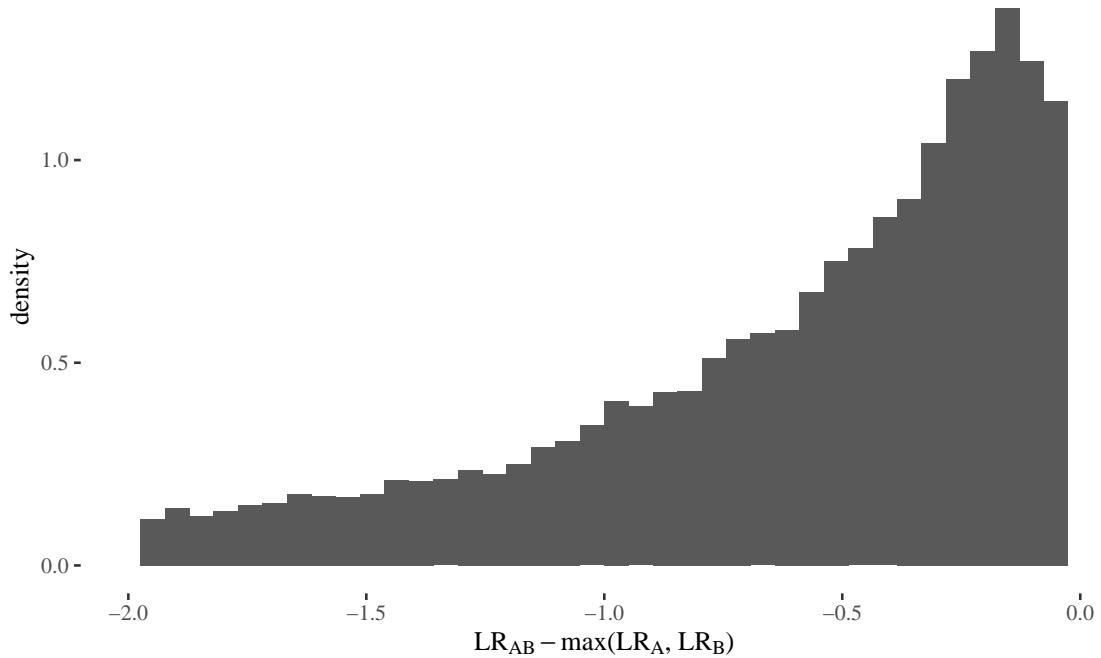
Cases in which $LR(AB) < LR(A), LR(B)$ (frequency=.125 (DAG1))



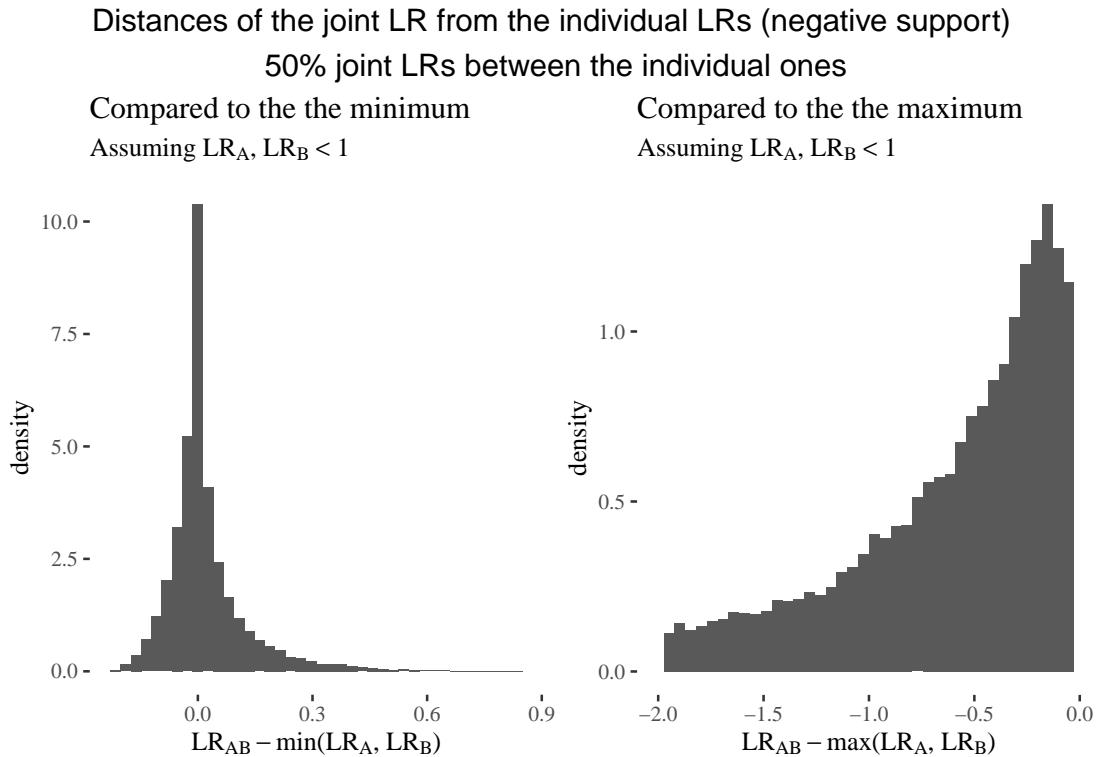
{fig:LRfails}

Figure 6: Ca. 25k cases (out of simulated 100k) in which the joint BF is below each of the individual BFs.

Compared to the the maximum
Assuming $LR_A, LR_B < 1$

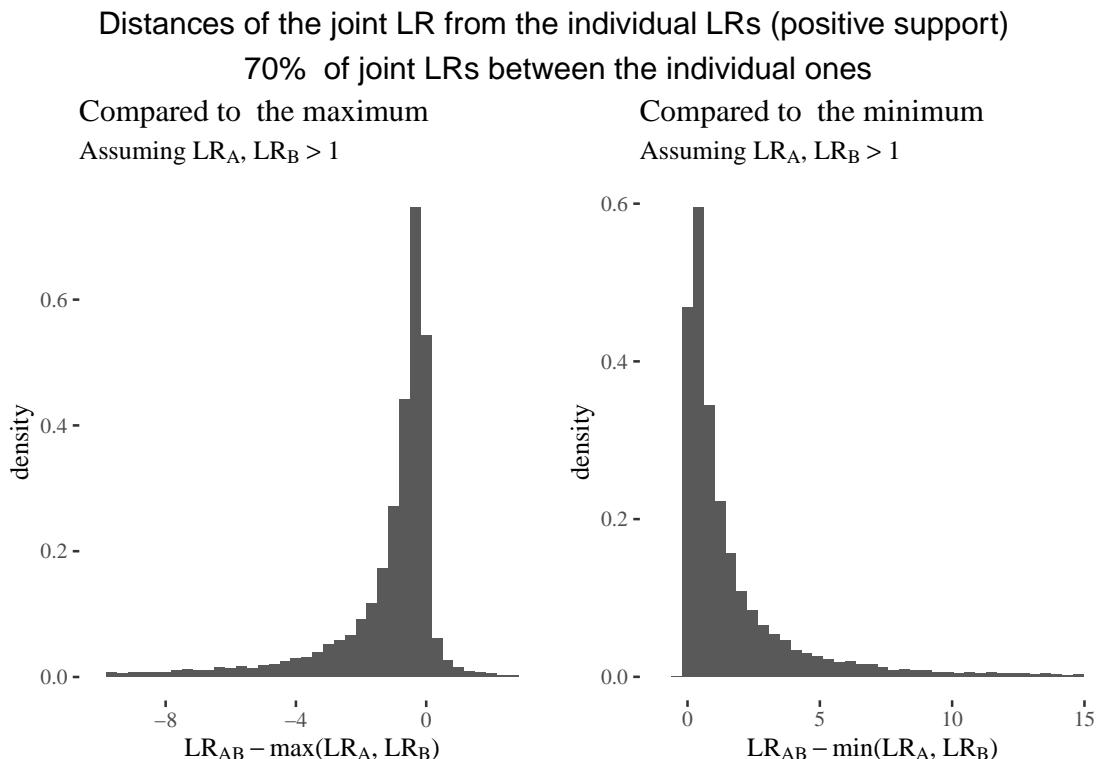


Once the individual likelihood ratios are above 1, the joint likelihood ratio can be lower than the maximum, but is not lower than the minimum of the individual likelihood ratios (Figure 8).



{fig:LRlowerPlot}

Figure 7: Distances of joint likelihood ratios for the minima and the maxima of the individual likelihood ratios if the individual likelihood ratios are below 1, DAG used in DAG1.



{fig:LRabovePlot}

Figure 8: Distances of joint likelihood ratios for the minima and the maxima of the individual likelihood ratios if the individual likelihood ratios are above 1, DAG used in DAG1.