Appendix: Bayes factor, likelihood ratio, and the difficulty about conjunction

Marcello Di Bello and Rafal Urbaniak

This appendix takes a closer look at the behavior of the Bayes factor and the likelihood ratio in the conjunction problem. We do so by building Direct Acyclic Graphs (DAGs) to study to what extent these measures of evidential support comply with the conjunction principle. We rely on analytic proofs, but when calculations become unmanageable, we rely instead on computer simulations. These simulations are conducted as follows. For each DAG of interest, 100,000 random Bayesian networks were generated whose probability tables had values sampled from the Uniform(0,1) distribution. For each of these networks, we calculated the relevant probabilities, Bayes factors, and likelihood ratios. Even for cases in which we derived analytic proofs, the results of a computer simulation sometimes provided additional insights.

Now, for the conceptual development. First, we describe two DAGs at play: one of them represents a conjunction when the hypotheses are independent, and the other one drops this independence assumptions. We then discuss the relation between DAGs and independence, and introduce the independencies in the proofs used later on.

Then we turn to the Bayes factor. First we prove that for if the stronger independence assumptions hold, the joint Bayes factor is just the result of multiplying the individual Bayes factors. It follows that aggregation is satisfied in such cases, if individual Bayes factors are greater than one. Once the hypotheses are not independent, a weaker result can be obtained, which entails that the aggregation is satisfied for the Bayes factor, if a certain additional constraint is satisfied.

We investigate simulations based on the first DAG: in general, aggregation fails 25% of the time if the individual Bayes factors are not constrained to be greater than one, but holds once this constraint is added. Switching to the second DAG does not change the picture, and so the question of whether the additional constraint needed for the weaker theorem holds in all Bayesian networks based on this DAG. Inspired by this observation, we show that in fact the additional constraint needed for aggregation to hold falls out of a pair of other independencies entailed by the second DAG. What the simulation reveal is that there is a large class of cases in which individual Bayes factors are above one, aggregation is satisfied, but distribution fails.

Next, we turn to the likelihood ratio. Since the analytic approach is less feasible here, we approach Bayesian networks based on the simpler DAG analytically, but rely on simulations for cases in which the hypotheses are not assumed to be independent. Without any constraint on individual likelihood ratios, aggregation fails in 12.5% cases (twice less often than aggregation for the Bayes factor). Another difference was that while the joint Bayes factor in cases with positive individual Bayes factor was always not less than the larger of the individual factors, now around 70% of joint likelihood ratios falls between the individual ratios, and is no lower than the smaller of these (if the individual likelihood ratios are assumed to be greater than one). Thus, aggregation is satisfied if individual likelihood ratios equal at least one, but it no longer holds that the joint support is greater than any of the individual support levels. Still, distribution fails in a large class of cases.

Finally, we identify cases in which aggregation can fail even if the individual BFs or LRs are at least one: this can happen if there is a direct dependence between the pieces of evidence.

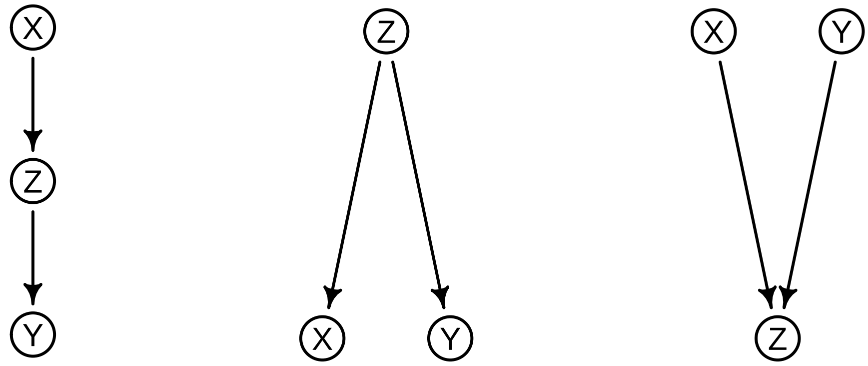
The **R** code we used in the simulations, calculations and visualizations is made available on the book website [LINK TO DOCUMENTATION TO BE ADDED LATER].

Bayesian networks and probabilistic independence

We begin with a refresher of the basic notions. A Bayesian network consists of a graphical part—a directed acyclic graph (DAG)—and a probability measure defined over the nodes (variables) in the graph. Bayesian networks satisfy the Markov condition. That is, any node is conditionally independent of its nondescentants (including ancestors), given its parents. If a probabilistic measure $\mathsf{P}()$ that is defined over the nodes (variables) in a graph G respects the Markov condition, $\mathsf{P}()$ is said to be compatible with G. Graph G and measure $\mathsf{P}()$ can then be combined to form a Bayesian network.

The graphical counterpart of probabilistic independence is **d-separation**, $\perp\!\!\!\perp_d$. Two nodes, X and Y, are d-separated given a set of nodes Z—$X \perp\!\!\!\perp_d Y|\mathsf{Z}$ — iff for every undirected path from X to Y there is a node Z' on the path such that either (see Figure 1):

- $Z' \in \mathsf{Z}$ and there is a **serial** connection, $\rightarrow Z' \rightarrow$, on the path (**pipe**),
- $Z' \in \mathsf{Z}$ and there is a **diverging** connection, $\leftarrow Z' \rightarrow$, on the path (**fork**),
- There is a **converging** connection $\rightarrow Z' \leftarrow$ on the path (in which case Z' is a **collider**), and neither Z' nor its descendants are in Z.



(a) Serial: Pipe (b) Diverging: Fork (c) Converging:Collider

Figure 1: Three basic types of connections.

Serial, converging and diverging connections represent common scenarios. Consider the nodes:

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 the same DNA profile.

This scenarios is naturally represented by the serial connection $G \rightarrow B \rightarrow M$. If we don't know whether B holds, G has 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.

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

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

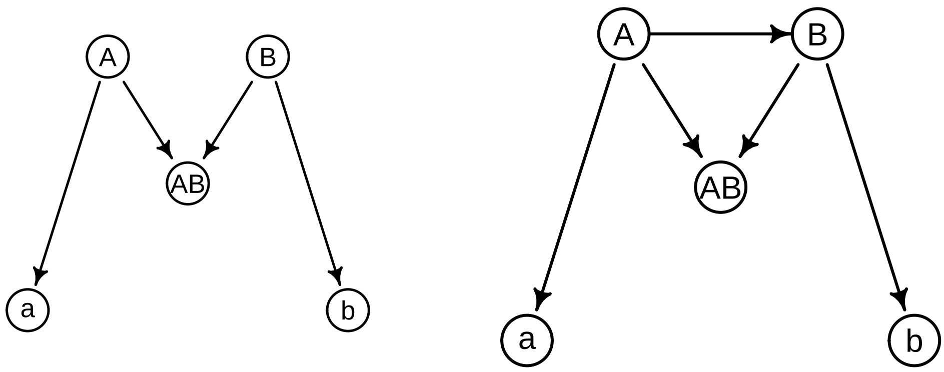
Both G and O influence B. If suspect guilty, it is more likely that the blood stain comes from him, and if the blood crime stain comes from the offender it is more likely to come from the suspect (for instance, more so than if it comes from the victim). Moreover, G and O seem independent. Whether the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

Take an example of a diverging connection. 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 an 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, $H1$, and on the result of the second toss, $H2$. So $H1 \leftarrow F \rightarrow H2$ seems to be appropriate. Now, on one hand, so long as you do not know whether F, the truth of $H1$ increases the probability of $H2$. On the other hand, once you know that F is true, $H1$ and $H2$ become independent, and so conditioning on the parent in a fork makes its childern independent (provided there is no other open path between them in the graph).

As a final piece of terminology, 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. Interestingly, it can be proven that if two sets of nodes are d-separated given a third one, they are independent given the third one, for any probabilistic measures compatible with a given DAG. However, lack of d-separation does not necessarily entail dependence for any probabilistic measure compatible with a given DAG. 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 dependent. So, at least, no false independencies can be inferred from the DAG, and all the dependencies are built into it.

Independencies in the conjunction problem

Back to the difficulty with conjunction. One assumption often made in the formulation of the problem is that hypotheses A and B are probabilistically independent. We will endorse this assumption, but also relax it to see whether the problem subsides (it does not). In particular, we will consider the two Bayesian networks shown in Figure 2. They represent two hypotheses, A and B, their supporting pieces of evidence, a and b, and their conjunction AB.[1] The difference is that a direct dependence between the hypotheses exists in the second network.



(a) DAG1

(b) DAG2

Figure 2: Two DAGs for the conjunction problem.

[1] In the Bayesian networks in this appendix, the conditional probability table for the conjunction node AB mirrors the truth table for the conjunction in propositional logic, as in Table 1.

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

Table 1: Conditional probability table for the conjunction node.

Unsurprisingly, the relations of d-separation entailed by the two networks differ. Examples can be found in Table 2. In fact, DAG1 entails 31 d-separations, while DAG2 entails 22 of them. One warning about the notation. Nodes represent variables, and so each d-separation entails a probabilistic statement about all combination of the node states (values of variables) 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 $\mathsf{P}(B = B_i) = \mathsf{P}(B = B_i | a = a_i)$.

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

Table 2: Some of d-separations entailed by DAG1 and DAG2.

Turning from nodes to states (or events, propositions), Figure 3 lists the independencies between propositions.[2] It also shows which independencies are entailed by either of the two DAGs.

$A \perp\!\!\!\perp B$	DAG1	(1)	$a \perp\!\!\!\perp B \vert A$	DAG1 , DAG2	(10)
$b \perp\!\!\!\perp a$	DAG1	(2)	$a \perp\!\!\!\perp B \vert \neg A$	DAG1 , DAG2	(11)
$A \perp\!\!\!\perp b \vert a$	DAG1	(3)	$a \perp\!\!\!\perp \neg B \vert A$	DAG1 , DAG2	(12)
$B \perp\!\!\!\perp a \wedge A \vert b$	DAG1	(4)	$a \perp\!\!\!\perp \neg B \vert \neg A$	DAG1 , DAG2	(13)
$a \perp\!\!\!\perp b \vert A \wedge B$	DAG1 , DAG2	(5)	$b \perp\!\!\!\perp A \wedge a \vert B$	DAG1 , DAG2	(14)
$a \perp\!\!\!\perp b \vert A$	DAG1 , DAG2	(6)	$b \perp\!\!\!\perp \neg A \wedge a \vert B$	DAG1 , DAG2	(15)
$a \perp\!\!\!\perp b \vert \neg A$	DAG1 , DAG2	(7)	$b \perp\!\!\!\perp A \wedge a \vert \neg B$	DAG1 , DAG2	(16)
$a \perp\!\!\!\perp b \vert B$	DAG1 , DAG2	(8)	$b \perp\!\!\!\perp \neg A \wedge a \vert \neg B$	DAG1 , DAG2	(17)
$a \perp\!\!\!\perp b \vert \neg B$	DAG1 , DAG2	(9)	$b \perp\!\!\!\perp a \vert B$	DAG1 , DAG2	(18)

Figure 3: Independencies among propositions according to DAG1 and DAG2.

An ambiguity in our notation is worth mentioning. Table 2 lists independencies between *nodes*. But

[2]Some caveats. In (3) the conditioning on a is not essential, because it's not on the path between the nodes: the key reason why the independence remains upon this conditioning is that there is an unconditioned collider on the path. Still, we need this independence in the proof later on. In (5) what we are conditioning on is A and B jointly. Technically, independence conditional on the conjunction node AB does not fall out of the d-separations present in the network—it follows given that AB and A, B are connected deterministically: fixing AB to true fixes both A and B to true.

Figure 3 is about *states* rather than nodes. Each particular instantiation of a node (a state) corresponds to a proposition, for example, $A = 1$ means that the proposition corresponding to A holds, while $A = 0$ means that the negation of A holds. Crucially, an expression such as $b \perp\!\!\!\perp A \wedge a | \neg B$ should be understood as a claim about states (events, propositions), which means the same as $\mathsf{P}(b = 1|B = 0) = \mathsf{P}(b = 1|A = 1, a = 1, B = 0)$. The distinction betwen nodes and their states (or variables and their values) matters because independence conditional on $B = 0$ doesn't entail independence given $B = 1$. For instance, one's final grade might depend on hard work if the teacher is fair, but this might fail if the teacher is not fair. We hope this ambiguity in notation will cause no confusion. Whether we talk about nodes or states (or events, propositions) should be clear from the context.

Posterior probabilities

We first examine how posterior probabilities behave in the conjunction problem. The joint posterior $\mathsf{P}(A \wedge B|a \wedge b)$ is often lower than the individual posterior $\mathsf{P}(A|a)$ and $\mathsf{P}(B|b)$, whether the hypotheses A and b are independent or not. We establish this fact via a computer simulation. We simulated 10,000 random Bayesian networks based on DAG1 (independent hypotheses) and DAG2 (dependent hypotheses). If any such network has an equal probability of occurring, the joint posterior is lower than both individual posteriors 68% of the time for DAG1, and around 60% for DAG2. Figure 4 displays the distributions of the distances of the joint posterior from the lowest of the individual posteriors.

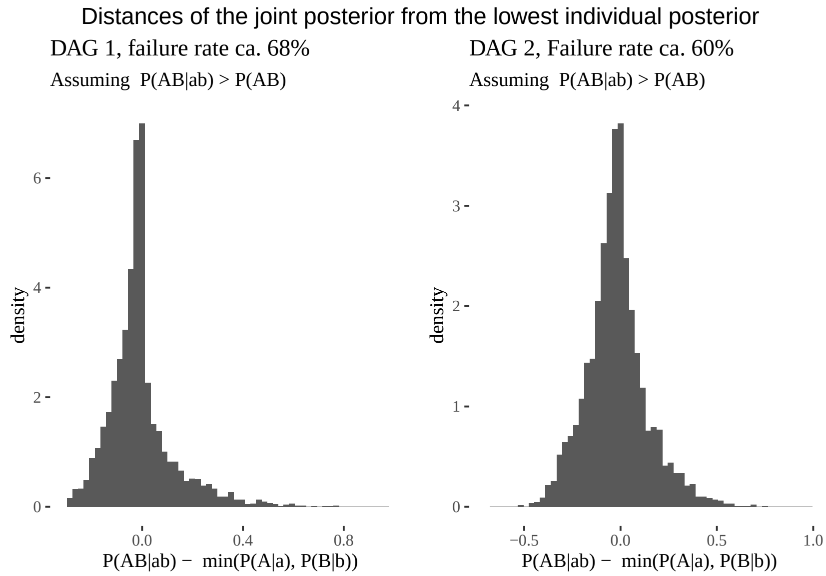


Figure 4: Even assuming the joint support is positive, the joint posterior often is lower than individual posteriors.

Bayes factor: proofs

Next, we turn to the Bayes factor as our measure of evidential support. For ease of reference, we use the following abbreviations:

$$BF_A = \frac{\mathsf{P}(a|A)}{\mathsf{P}(a)},$$
$$BF_B = \frac{\mathsf{P}(b|B)}{\mathsf{P}(b)},$$
$$BF_{AB} = \frac{\mathsf{P}(a \wedge b|A \wedge B)}{\mathsf{P}(a \wedge b)}$$

The objective here is to study how the combined support BF_{AB} compares to the individual supports BF_A and BF_B. We prove the following general theorem:

Theorem 1. *Given a measure* $\mathsf{P}()$ *compatible with* DAG1*, if both* BF_A *and* BF_B *are grater than one, then* $BF_{AB} \geq \max(BF_A, BF_B)$*. The same holds for a measure compatible with* DAG2*.*

In other words, the combined support BF_{AB} is never below the individual supports BF_A and BF_B, whether claims A and B are independent (DAG1) or not (DAG2).

Proof. For DAG1, the theorem holds by Fact 1 (and corollary). For DAG2, the theorem holds by Fact 2 (and corollary), Lemma 1, and Fact 3. □

Fact 1. *If the independence assumptions* (1)*,* (2)*,* (10) *and* (14) *hold (all of which are entailed by* DAG1*), then* $BF_{AB} = BF_A \times BF_B$*.*

Proof.

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

Next, we apply the relevant independence assumptions:

$$\begin{aligned}
&= \frac{\mathsf{P}(A) \times \overbrace{\mathsf{P}(B|A)}^{\mathsf{P}(B) \text{ by } (1)} \times \overbrace{\mathsf{P}(a|A \wedge B)}^{\mathsf{P}(a|A) \text{ by } (10)} \times \overbrace{\mathsf{P}(b|A \wedge B \wedge a)}^{\mathsf{P}(b|B) \text{ by } (14)}}{\underbrace{\mathsf{P}(A \wedge B)}_{\mathsf{P}(A) \times \mathsf{P}(B) \text{ by } (1)}} \Big/ \underbrace{\mathsf{P}(a \wedge b)}_{\mathsf{P}(a) \times \mathsf{P}(b) \text{ by } (2)}\\
&= \frac{\mathsf{P}(A) \times \mathsf{P}(B) \times \mathsf{P}(a|A) \times \mathsf{P}(b|B)}{\mathsf{P}(A) \times \mathsf{P}(B)} \Big/ \mathsf{P}(a) \times \mathsf{P}(b)\\
&= \frac{\mathsf{P}(a|A)}{\mathsf{P}(a)} \times \frac{\mathsf{P}(b|B)}{\mathsf{P}(b)}\\
&= BF_A \times BF_B
\end{aligned}$$

□

This fact has the following straightforward consequences. They hold because, if $a = b \times c$ and $b, c > 1$, then $a > \max(b, c)$, and if $b, c < 1$, then $a < \min(b, c)$.

Corollary 1.1. *If the independence assumptions* (1)*,* (2)*,* (10) *and* (14) *hold, and* BF_A *and* BF_B *are both greater than 1, then* BF_{AB} *is greater than one. In fact,* BF_{AB} *is greater than* $\max(BF_A, BF_B)$*.*

Corollary 1.2. *If the independence assumptions* (1)*,* (2)*,* (10) *and* (14) *hold, and* BF_A *and* BF_B *are both strictly less than 1, then* BF_{AB} *is less than* $\min(BF_A, BF_B)$*.*

This establishes Theorem 1 for DAG 1. After dropping the independence assumptions specific to DAG1 and shifting to DAG 2, the combined BF_{AB} can no longer be obtained by multiplying the individual ones, although multiplication still provides a decent approximation (see Figure 5).

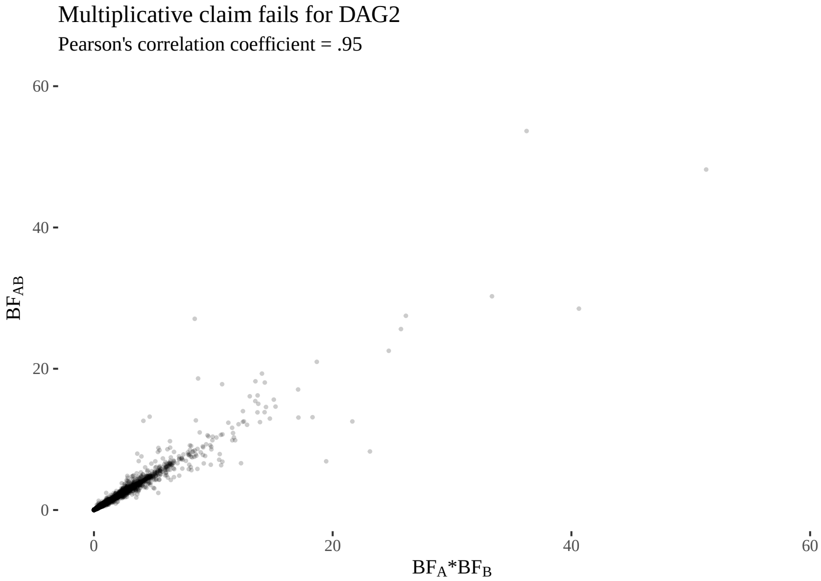


Figure 5: In DAG2, the result of multiplying individual BFs does not equal the joint BF, but often is a good approximation thereof. Axes restricted to $0,60$ (one extreme outlier lying close to the diagonal dropped).

However, if the probabilistic measure fits DAG 2, Theorem 1 is still satisfied. First, we abbreviate:

$$BF_B^{'} = \frac{\mathsf{P}(b|B)}{\mathsf{P}(b|a)}$$
$$BF_A^{'} = \frac{\mathsf{P}(a|A)}{\mathsf{P}(a|b)}$$

A claim weaker than Fact 1 can be proven by relying only on the independencies entailed by DAG2.

Fact 2. *If* (10) *and* (14) *hold (and they do in BNs based on DAG 2), 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).

$$\frac{\mathsf{P}(a \wedge b|A \wedge B)}{\mathsf{P}(a \wedge b)} = \frac{\mathsf{P}(A) \times \mathsf{P}(B|A) \times \overbrace{\mathsf{P}(a|A \wedge B)}^{\mathsf{P}(a|A) \text{ by } (10)} \times \overbrace{\mathsf{P}(b|A \wedge B \wedge a)}^{\mathsf{P}(b|B) \text{ by } (14)}}{\underbrace{\mathsf{P}(A \wedge B)}_{\mathsf{P}(A) \times \mathsf{P}(B|A) \text{ by the chain rule}}} \Big/ \underbrace{\mathsf{P}(a \wedge b)}_{\mathsf{P}(a) \times \mathsf{P}(b|a) \text{ by the chain rule}}$$
$$= \frac{\mathsf{P}(a|A)}{\mathsf{P}(a)} \times \frac{\mathsf{P}(b|B)}{\mathsf{P}(b|a)}$$
$$= BF_A \times BF_B^{'}$$

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

$$= \frac{\mathsf{P}(a|A)}{\mathsf{P}(a|b)} \times \frac{\mathsf{P}(b|B)}{\mathsf{P}(b)}$$
$$= BF_A^{'} \times BF_B$$

□

Corollary 2.1. *Suppose* (10) *and* (14) *hold (they are entailed by DAG 2), and* $BF_{BA}, BF_B > 1$. *Then if both* $\mathsf{P}(a|b) \leq \mathsf{P}(a|A)$ *and* $\mathsf{P}(b|a) \leq \mathsf{P}(b|B)$, *we have* $BF_{AB} \geq BF_A, BF_B$.

Proof. Assume the first conjunct holds. Then $\frac{\mathsf{P}(a|A)}{\mathsf{P}(a|b)} \geq 1$ and so:

$$BF_{AB} = BF_A^{'} \times BF_B \geq BF_B$$

The argument for the other comparison is analogous. □

The proof of Theorem 1 for DAG 2 is not complete yet. This corollary relies on the additional assumptions $\mathsf{P}(a|b) \leq \mathsf{P}(a|A)$ and $\mathsf{P}(b|a) \leq \mathsf{P}(b|B)$. They seem plausible. If, say, a is used as evidence for A, we often expect A and a to be fairly strongly connected, that is, we expect $\mathsf{P}(a|A)$ to be rather high, while the connection between different pieces of evidence for different hypotheses, intuitively, is not expected to be as strong. We provide a proof of these assumptions below.

We start with the following lemma.

Lemma 1. *For any probabilistic measure* P, *if* $BF_A > 1$, *then* $LR_A > 1$.

Proof. We start with our assumption.

$$\begin{aligned}
1 &\leq \frac{\mathsf{P}(a|A)}{\mathsf{P}(a)} && (BF_A \geq 1)\\
\mathsf{P}(A) &\leq \frac{\mathsf{P}(a|A)}{\mathsf{P}(a)}\mathsf{P}(A) && \text{(algebraic manipulation)}\\
\mathsf{P}(A) &\leq \mathsf{P}(A|a) && \text{(Bayes' theorem)}\\
-\mathsf{P}(A) &\geq -\mathsf{P}(A|a) && \text{(algebraic manipulation)}\\
1-\mathsf{P}(A) &\geq 1-\mathsf{P}(A|a) && \text{(algebraic manipulation)}\\
1-\mathsf{P}(A) &\geq \mathsf{P}(\neg A|a) && \text{(algebraic manipulation)}\\
\mathsf{P}(a)\,(1-\mathsf{P}(A)) &\geq \mathsf{P}(a)\mathsf{P}(\neg A|a) && \text{(algebraic manipulation)}\\
\mathsf{P}(a) &\geq \frac{\mathsf{P}(a)\mathsf{P}(\neg A|a)}{\mathsf{P}(\neg A)} && \text{(algebraic manipulation, negation)}\\
\mathsf{P}(a) &\geq \mathsf{P}(a|\neg A) && \text{(conditional probability)}
\end{aligned}$$

From this and our assumption that $\mathsf{P}(a|A) \geq \mathsf{P}(a)$ it follows that $\mathsf{P}(a|A) \geq \mathsf{P}(a|\neg A)$, that is, that $LR_A \geq 1$. □

Now the main claim.

Fact 3. *For any probabilistic measure* P *appropriate for DAG 2, if* $BF_A > 1$, *then* $\mathsf{P}(a|A) \geq \mathsf{P}(a|b)$ *and* $\mathsf{P}(b|B) \geq \mathsf{P}(b|a)$.

Proof. Let's focus on the first conjunct. First, we have:

$$\begin{aligned}
\mathsf{P}(a|b) &= \mathsf{P}(a \wedge A|b) + \mathsf{P}(a \wedge \neg A|b) && \text{(total probability)}\\
&= \underbrace{\mathsf{P}(a|b \wedge A)}_{\mathsf{P}(a|A) \text{ by (6)}}\mathsf{P}(A|b) + \underbrace{\mathsf{P}(a|b \wedge \neg A)}_{\mathsf{P}(a|\neg A) \text{ by (7)}}\mathsf{P}(\neg A|b) && \text{(chain rule)}
\end{aligned}$$

Now let's introduce some abbreviations:

$$= \underbrace{\mathsf{P}(a|A)}_{k}\underbrace{\mathsf{P}(A|b)}_{x} + \underbrace{\mathsf{P}(a|\neg A)}_{t}\underbrace{\mathsf{P}(\neg A|b)}_{(1-x)}$$

Note that the assumption that $BF_A \geq 1$ entails, by Lemma 1, that $k \geq t$, and so $k - t \geq 0$. Also, since x is a probability, we know $0 \leq x \leq 1$. This allows us to reason algebraically as follows:

$$
\begin{aligned}
k &\geq k \\
k &\geq t + (k-t) \\
k &\geq t + (k-t)x \\
k &\geq kx + t - tx \\
\mathsf{P}(a|A) = k &\geq kx + t(1-x) = \mathsf{P}(a|b)
\end{aligned}
$$

For the second conjunct, notice that we have a similar reasoning, albeit it relies on a different pair of independencies (which nevertheless holds in DAG1 and DAG 2).

$$
\begin{aligned}
\mathsf{P}(b|a) &= \mathsf{P}(b \wedge B|a) + \mathsf{P}(b \wedge \neg B|a) && \text{(total probability)} \\
&= \underbrace{\mathsf{P}(b|a \wedge B)}_{\mathsf{P}(b|B) \text{ by (8)}} \mathsf{P}(B|a) + \underbrace{\mathsf{P}(b|a \wedge \neg B)}_{\mathsf{P}(b|\neg B) \text{ by (9)}} \mathsf{P}(\neg B|a) && \text{(chain rule)}
\end{aligned}
$$

The rest of the reasoning for this case is algebraically the same as the one used above. □

Bayes factor: simulations

Computer simulations provide additional insights. The joint BF_AB may be lower than the individual BF_A and BF_B. Simulated cases in which $BF_{AB} < BF_A, BF_B$ are about 25% of the total (which is twice higher than for the likelihood ratio; more o this later). The structure of such cases is visualized in Figure 6.

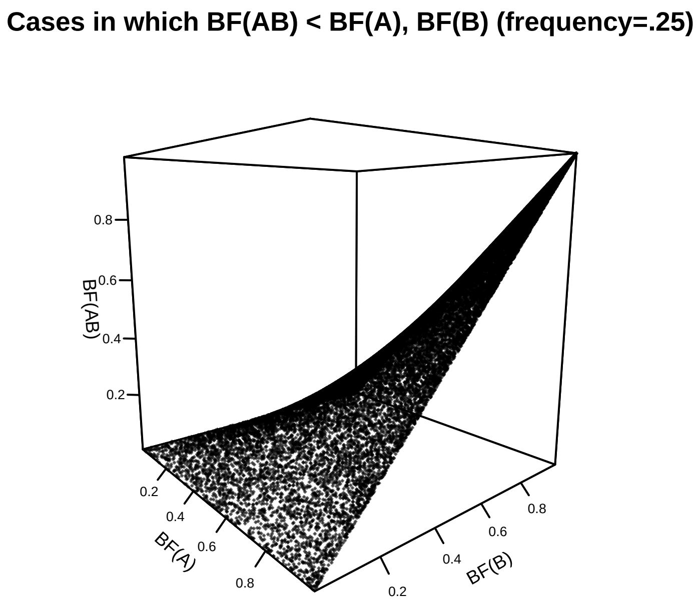


Figure 6: 25k cases (out of simulated 100k) in which the joint BF is below each of the individual BFs.

This result should not be surprising. It happens when BF_A and BF_B are lower than one. When they are greater than one, the joint BF_{AB} exceeds the individual ones. The distribution of Bayes factors based on DAG1 are displayed in Figure 7. Interestingly, the distribution is unchanged under DAG2 (Figure 8).

These simulations and the earlier theorem demonstrate that, whenever individual BF_A and BF_B are above a fixed threshold, so is the combined BF_{AB}. This fact justifies the principle of aggregation, as explained in the main text of the chapter. However, the converse does not hold. Whenever BF_{AB} is above a threshold, BF_A or BF_B may be below the threshold. Such cases for DAG1 and DAG2 are displayed in Figure 9. Hence, the converse of aggregation, the principle of distribution, fail in some cases.

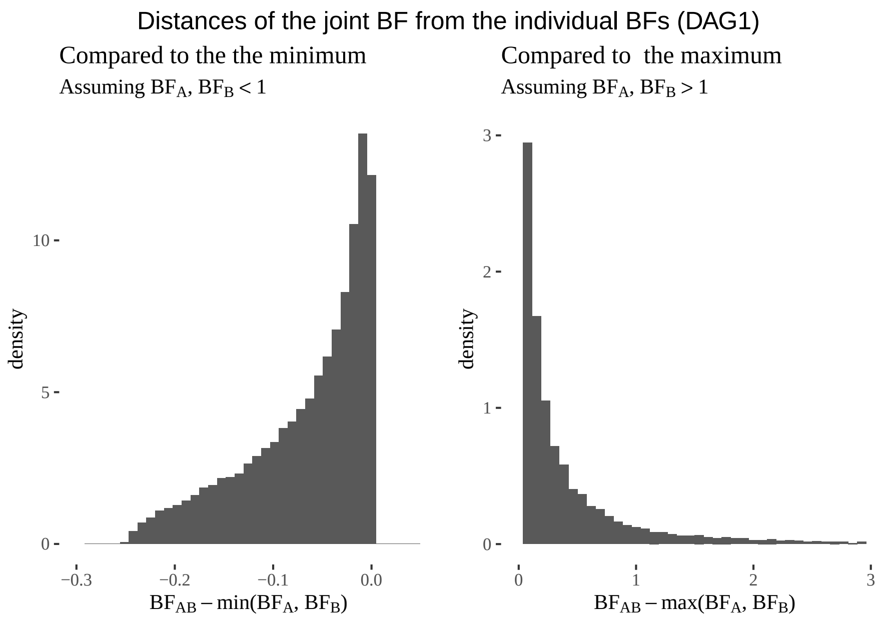


Figure 7: Distances of the joint Bayes factor from maxima and minima of individual Bayes factors, depending on whether the individual support levels are both positive or both negative. Simulation based on 100k Bayesian networks build over the DAG of DAG1.

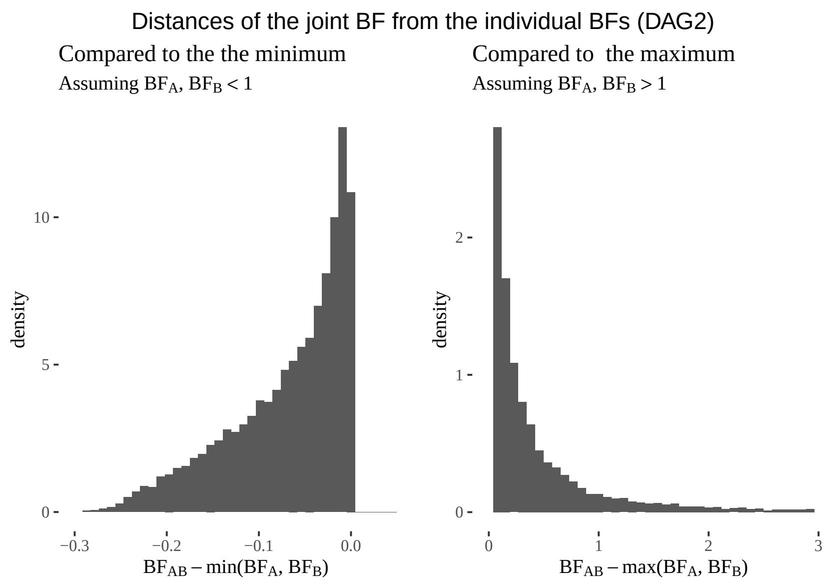


Figure 8: Distances of the joint Bayes factor from maxima and minima of individual Bayes factors, depending on whether the individual support levels are both positive or both negative. Simulation based on 100k Bayesian networks build over DAG2.

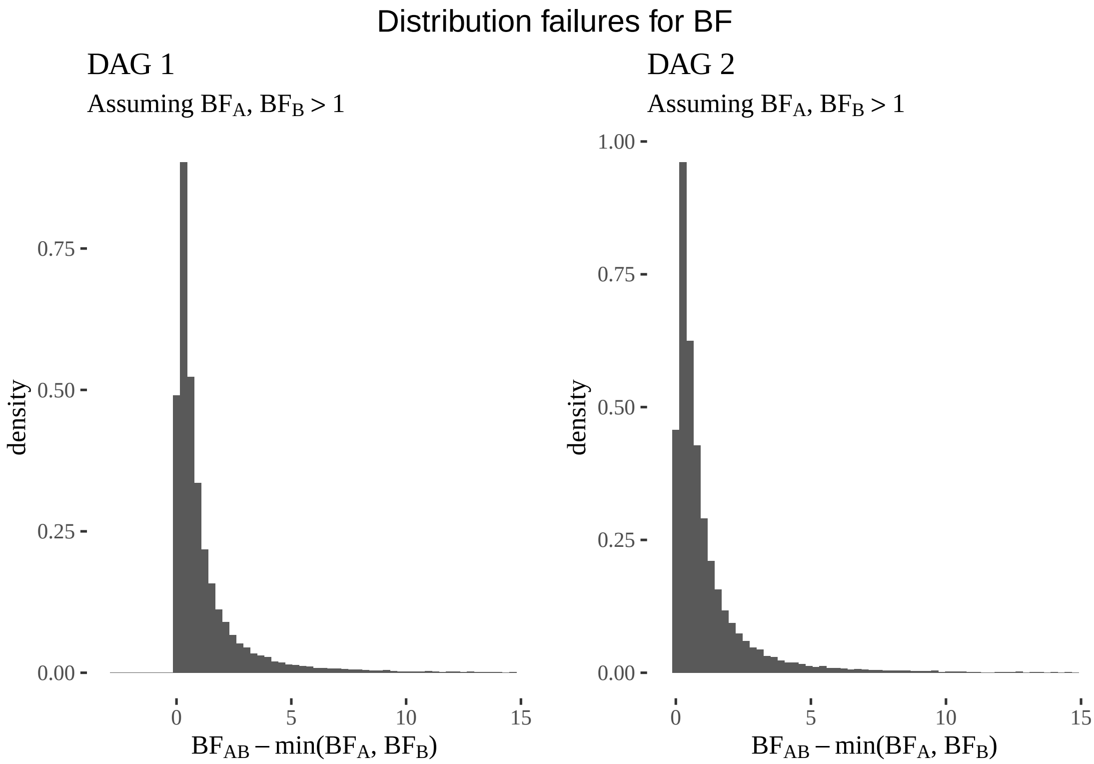


Figure 9: Distribution failure for the Bayes factor, DAG 1. The x axis restricted to $(-3, 15)$ for visibility.

Likelihood ratio: proofs

We now turn to the likelihood ratio. For ease of reference, we use the following abbreviations:

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

Fact 4. *If independence conditions* (10), (11), (12), (13), (14), (15), (16), *and* (17) *hold, then:*

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

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}:

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

We deploy the relevant independencies as follows:

$$= \frac{\mathsf{P}(A) \times \mathsf{P}(B|A) \times \overbrace{\mathsf{P}(a|A \wedge B)}^{\mathsf{P}(a|A) \text{ by (10)}} \times \overbrace{\mathsf{P}(b|A \wedge B \wedge a)}^{\mathsf{P}(b|B) \text{ by (14)}}}{\mathsf{P}(A) \times \mathsf{P}(B|A)}$$

$$= \mathsf{P}(a|A) \times \mathsf{P}(b|B) \qquad \text{(algebraic manipulation)}$$

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

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

Now consider the first summand from the numerator:

$$\begin{aligned}
\mathsf{P}(a \wedge b \wedge \neg A \wedge B) &= \mathsf{P}(\neg A)\mathsf{P}(B|\neg A)\mathsf{P}(a|\neg A \wedge B)\mathsf{P}(b|a \wedge \neg A \wedge B) && \text{(chain rule)}\\
&= \mathsf{P}(\neg A)\mathsf{P}(B|\neg A)\mathsf{P}(a|\neg A)\mathsf{P}(b|B) && \text{(independencies (11) and (15))}
\end{aligned}$$

The simplification of the other two summanda is analogous (albeit with slightly different independence assumptions—(12) and (16) for the second one and (13) and (17) for the third. Once we plug these into the denominator formula we get:

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

$\square$

Likelihod ratio: simulations

In the case of the likelihood ratio, the analytic approach is cumbersome. Instead, we most rely on computer simulations. First of all, the joint LR_{AB} can be lower than any of the individual LR_A and LR_B. Based on DAG1 and DAG2, the frequency of such cases in which $LR_{AB} < LR_A, LR_B$ is about 12.5%, half the frequency for the Bayes factor. The distribution of these cases is displayed in Figure 10.

Consider now cases in which both individual likelihood ratios are above one. Interestingly, one of the individual likelihood ratios, LR_A or LR_B, may be greater than the joint $LRAB$ (see example in Figure 11). This does not happen with the Bayes factor. But even though the joint likelihood ratio can be lower than the maximum, it is never lower than the minimum of the individual likelihood ratios (Figures 12 and 13). Conversely, if both both individual likelihood ratios are below one, the joint likelihood ratio can be higher than their minimum, but is never higher than their maximum (Figures 14 and 15).

0.1 Keeping evidence fixed

Dependent evidence

Both DAG 1 and DAG 2 ensures that the items of evidence are conditionally independent on their respective hypothesis (specifically, that a is independent both conditional on A and conditional on $\neg A$, and the same for b and B.).[3] The results established so far, then, rests of this assumption. What happens if this independence is dropped? To investigate this question, we run a simulation based on DAG 3, illustrated in Figure 16.

[3]In fact, DAG 1 ensure that they are also unconditionally independent.

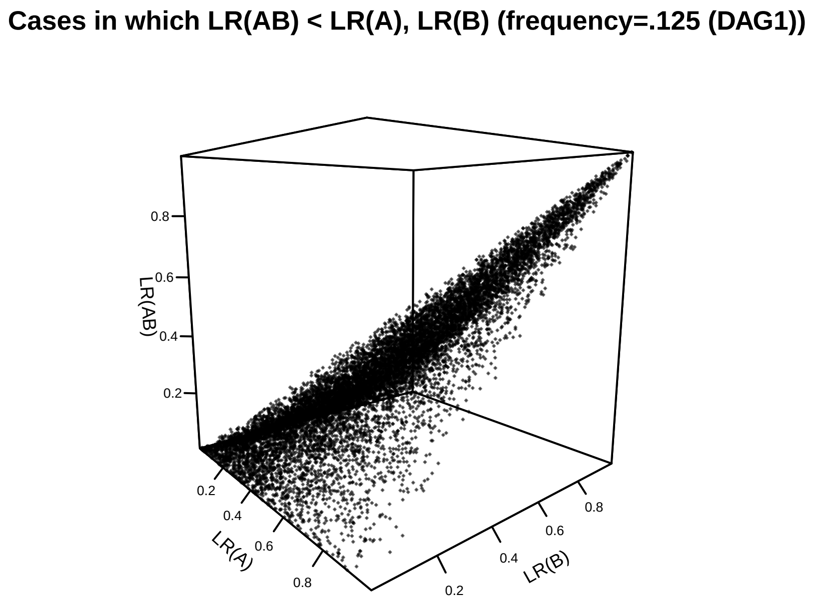


Figure 10: 12.5k cases (out of simulated 100k) in which the joint BF is below each of the individual BFs. The picture for DAG2 is very similar.

A	Pr
1	0.892
0	0.108

B	A = 1	A = 0
1	0.551	0.457
0	0.449	0.543

a	A = 1	A = 0
1	0.957	0.453
0	0.043	0.547

b	B = 1	B = 0
1	0.678	0.573
0	0.322	0.427

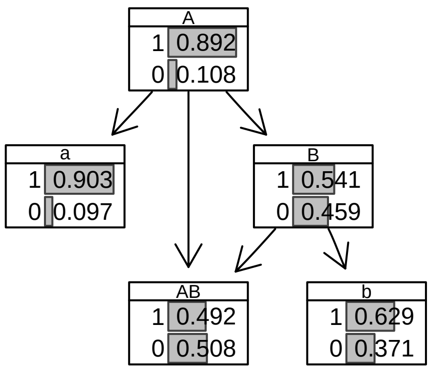


Figure 11: $LR_A \approx 2.11$, $LR_B \approx 1.183$, $LR_{AB} \approx 1.319$.
$BF_A \approx 1.06, BF_B \approx 1.076, BF_{AB} \approx 1.14$.

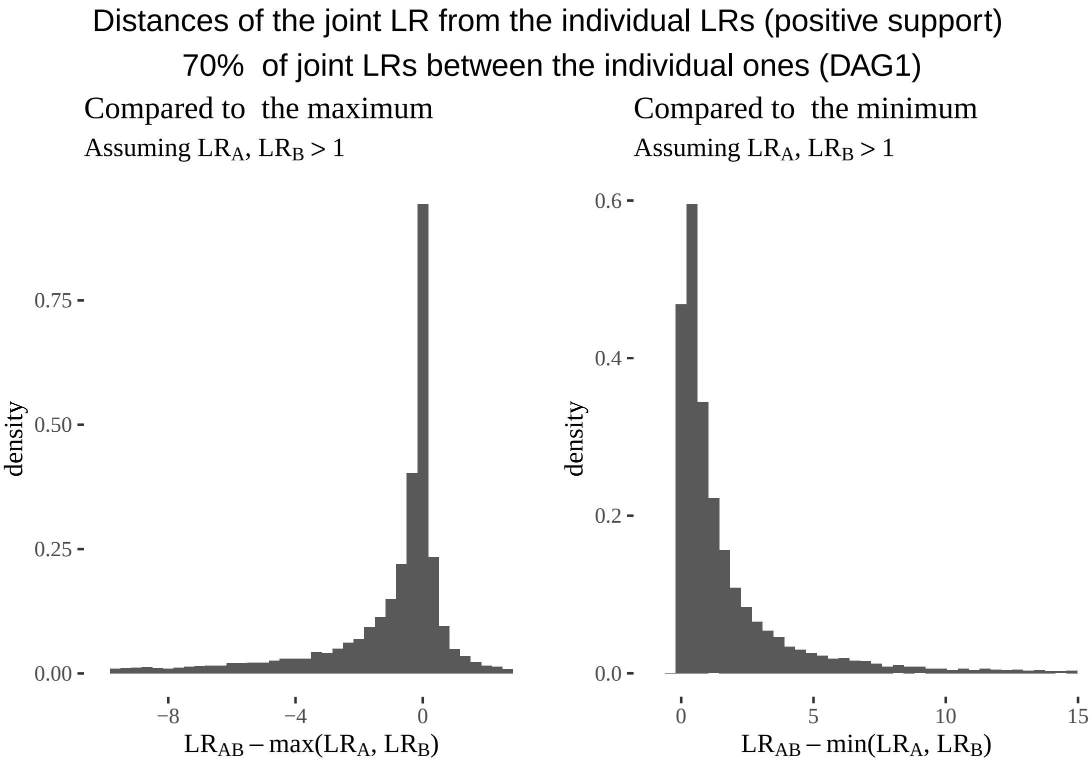


Figure 12: Distances of joint likelihood ratios for the minima and the maxima of the individual likelihood ratios if the individual likelihood ratios are above 1, DAG used in DAG1.

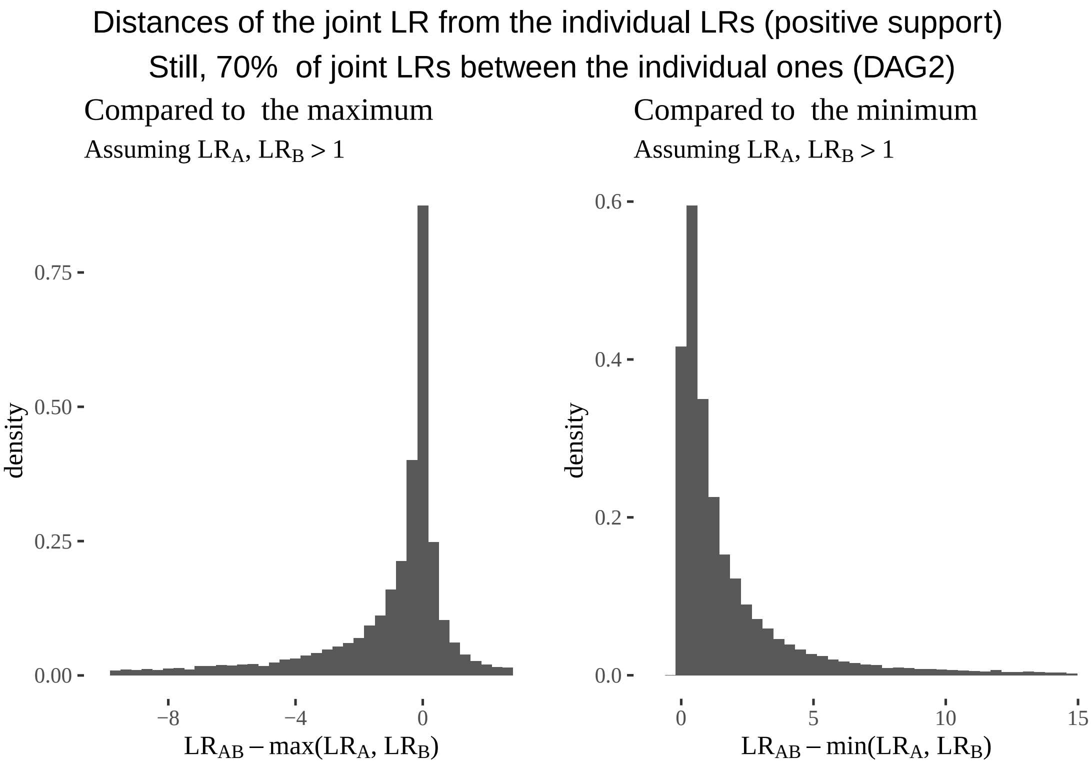


Figure 13: Distances of joint likelihood ratios for the minima and the maxima of the individual likelihood ratios if the individual likelihood ratios are above 1, DAG used in DAG2.

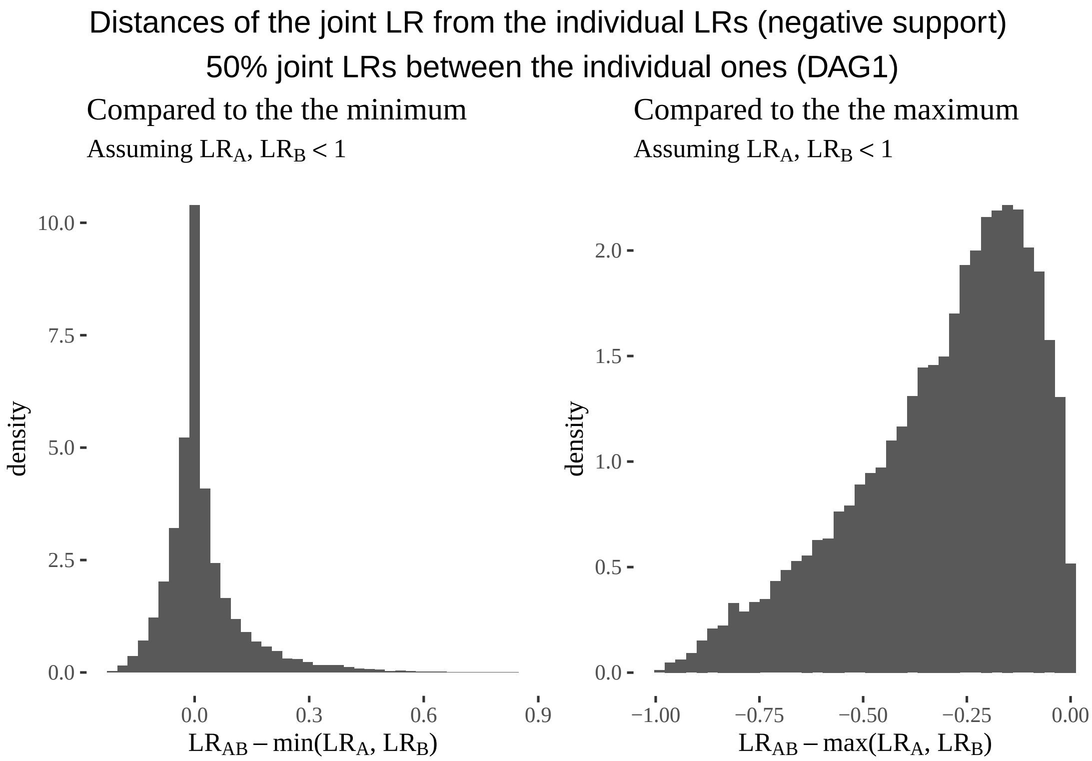


Figure 14: Distances of joint likelihood ratios for the minima and the maxima of the individual likelihood ratios if the individual likelihood ratios are below 1, DAG used in DAG1.

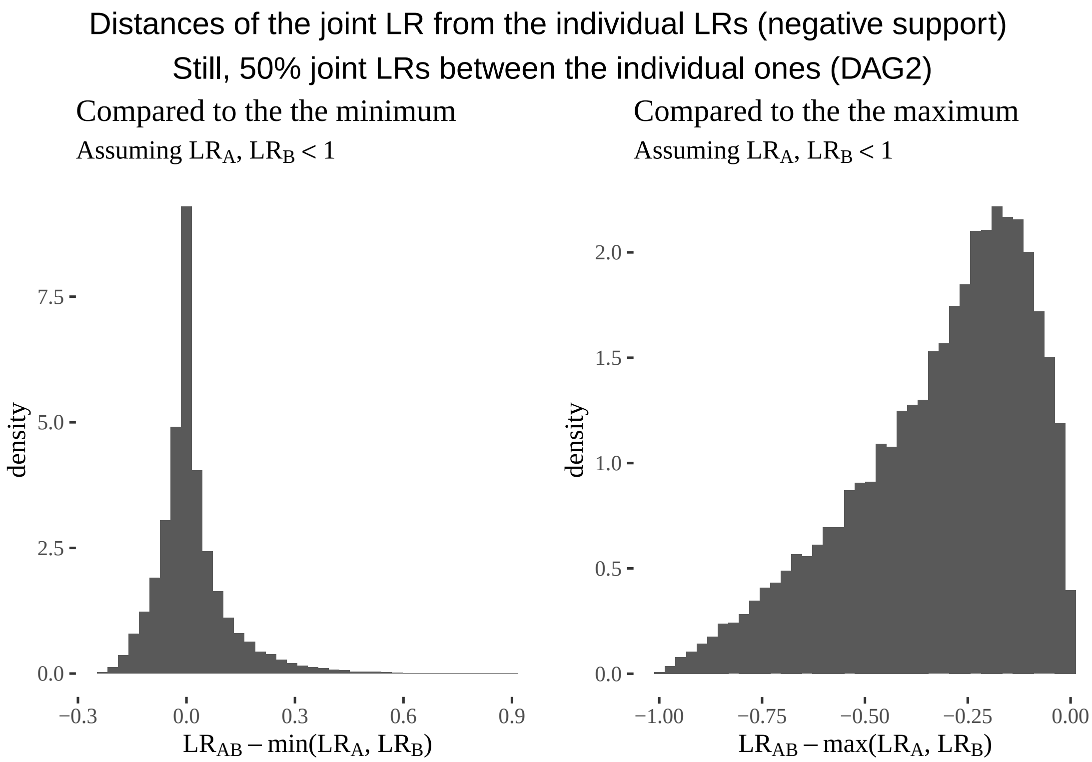


Figure 15: Distances of joint likelihood ratios for the minima and the maxima of the individual likelihood ratios if the individual likelihood ratios are below 1, DAG used in DAG2.

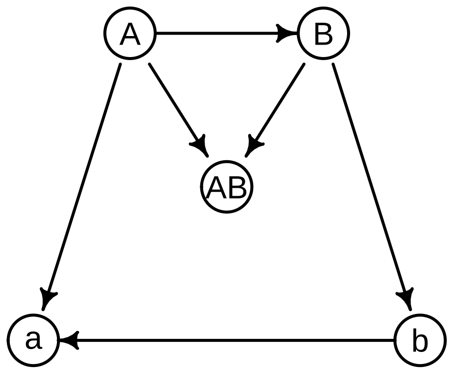


Figure 16: DAG 3 with direct dependence between the pieces of evidence.

In fact, it turns out that simultaneously the joint likelihood ratio is lower than both individual likelihood ratios and the joint Bayes factor is lower than each individual Bayes factor in 14% cases in which the individual Bayes factor (and therefore also the individual likelihood ratios) are greater than one. One particular counterexample is illustrated in Figure 17.

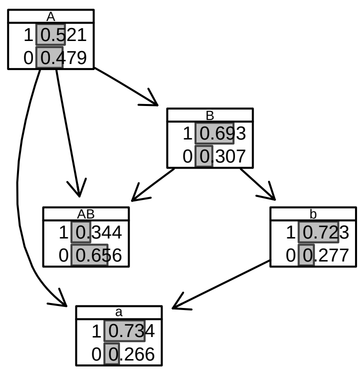
A	Pr
1	0.521
0	0.479

B	A	
	1	0
1	0.66	0.729
0	0.34	0.271

	A	b	
a			Pr
1	1	1	0.3989398
0	1	1	0.6010602
1	0	1	0.9673984
0	0	1	0.0326016
1	1	0	0.9693564
0	1	0	0.0306436
1	0	0	0.7267025
0	0	0	0.2732975

b	B	
	1	0
1	0.995	0.108
0	0.005	0.892

(a) Conditional probabilities for the counterexample (the one for AB does not change).



(b) Marginal probabilities.

Figure 17: A counterexample based on DAG 3, with independence between the items of evidence dropped. $LR_A \approx 1.063$, $LR_B \approx 1.159742$, $LR_{AB} \approx 0.651$. $BF_A \approx 1.022, BF_B \approx 1.079, BF_{AB} \approx 0.699$.

R: Two new passages on fixed joint evidence, check.

Finally, you might recall that in the chapter we distinguished two variants of the distribution requirement, (DIS1) and (DIS2). The former has been in the focus so far, and the latter differs in keeping the evidence fixed at $a \wedge b$, even when calculating individual LRs. So, sticking to LR, distribution in this variant pertains to the joint LR and the following two ratios: $\mathsf{P}(a\wedge b|A)/\mathsf{P}(a\wedge b|\neg A)$ and $\mathsf{P}(a\wedge b|B)/\mathsf{P}(a\wedge b|\neg B)$. To make sure that switching to (DIS2) doesn't make things easier for legal probabilists, we ran 30k simulations (10k for each BN type), and inspected the status of aggregation and distribution for these ratios.

Here are the results. If no assumption about the direction of support is made, around 12.7% of the time (around twice less often than if the usual individual LRs are used), the individual LRs with fixed

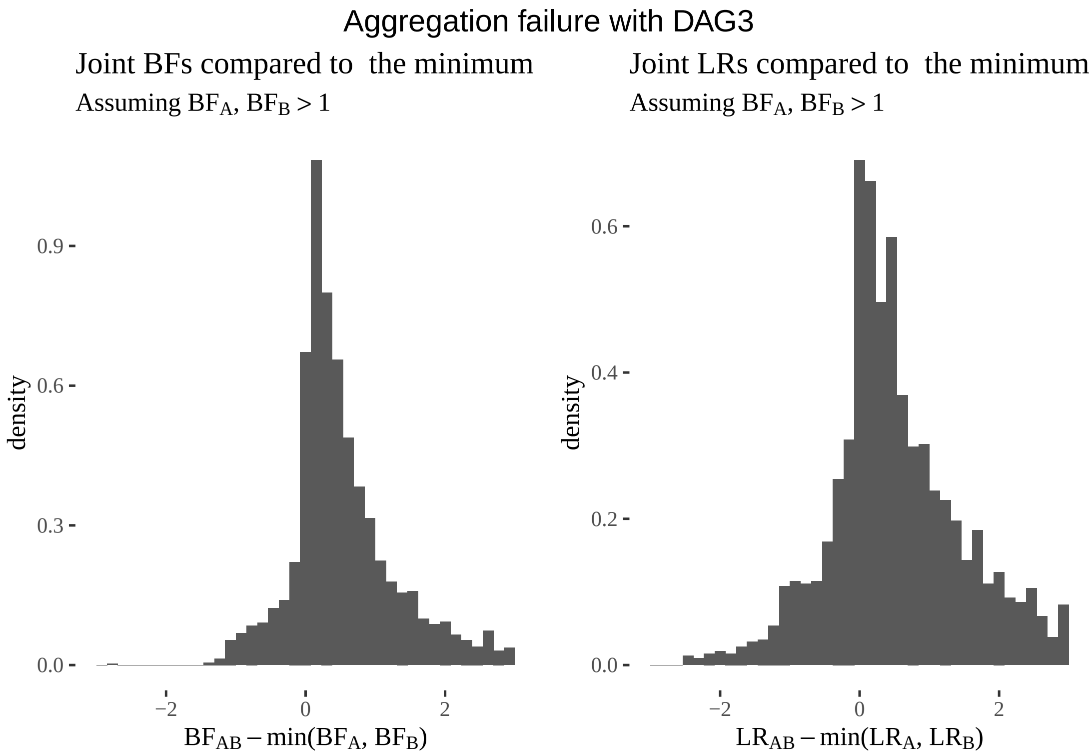


Figure 18: Aggregation failure.

evidence are both greater than the joint LR—this is for DAG1, the frequency goes slightly up to around 13% if we switch to DAG2 and is around the same value if additionally we allow for the dependency between the items of evidence (DAG3). Assuming individual LRs are above one, around 70% of joint likelihoods (75% for BN2, 72% for BN3) are between the individual ones, no joint LR is below the minimum of the individual ratios for DAG1, but is so around 2% times for both DAG2 and DAG3. This is one important difference: even aggregation can fail if dependencies are present, if we keep joint evidence fixed in all the ratios. As for distribution, there are no major surprises: around 30% of the time, the joint LR is strictly greater than both of the individual LRs with evidence fixed for DAG1, 22% for DAG2, and 26% for DAG3. In short, keeping the joint evidence fixed across all ratios makes things even harder, when it comes to preserving aggregation and distribution.

.