JMLR: Workshop and Conference Proceedings vol 52, 426-437, 2016 PGM 2016

# Evidence Evaluation: a Study of Likelihoods and Independence

**Silja Renooij** S.RENOOIJ@UU.NL
*Department of Information and Computing Sciences*
*Utrecht University, The Netherlands*

## Abstract

In the context of evidence evaluation, where the probability of evidence given a certain hypothesis is considered, different pieces of evidence are often combined in a naive way by assuming conditional independence. In this paper we present a number of results that can be used to assess both the importance of a reliable likelihood-ratio estimate and the impact of neglecting dependencies among pieces of evidence for the purpose of evidence evaluation. We analytically study the effect of changes in dependencies between pieces of evidence on the likelihood ratio, and provide both theoretical and empirical bounds on the error in likelihood occasioned by assuming independences that do not hold in practice. In addition, a simple measure of influence strength between pieces of evidence is proposed.

**Keywords:** Evidence evaluation; independence violations; error in overall likelihood; influence measures.

## 1. Introduction

The evaluation of evidence is the investigation of how evidence changes the relation between competing hypotheses. In various domains where ultimately a choice between competing hypotheses is required, the evaluation of evidence is not only an intermediate step of the process, but conveys important information in itself. In evidence-based medicine, for example, likelihood ratios $\Pr(t \mid d)/\Pr(t \mid \overline{d})$ of the sensitivity and $(1-)$ the specificity of a diagnostic test for a disease are used for assessing the value of performing that diagnostic test (Thornbury et al., 1975) and thus for evaluating potential evidence. In forensic science, findings such as for example DNA matches in criminal cases are reported in terms of likelihoods or likelihood ratios (Aitken and Taroni, 2004); incorporating hypothesis priors to arrive at conclusions concerning for example whether or not the evidence shows that the suspect is guilty is left to a judge or jury.

In determining the overall likelihood $\Pr(\mathbf{e} \mid h)$ of an hypothesis $h$, often likelihoods for multiple separate pieces of evidence need to be combined. In practice this is (still) often done in a naive way by simply multiplying the individual likelihoods $\Pr(e_i \mid h)$ and neglecting any possible conditional dependencies among the pieces of evidence. That is, for computing the overall likelihood the same independence assumptions are made as in a naive Bayesian classifier. The performance of a naive Bayesian classifier is known to suffer very little from the unrealistic independence assumptions it makes, in the sense that classification accuracy tends to be quite high. This is repeatedly demonstrated in practical applications (see Hand and Yu (2001) for an overview) and can generally be explained by the fact that a naive Bayesian classifier is optimal as long as the true hypothesis for a set of evidence has the highest posterior in the model (Domingos and Pazzani, 1997); an overview of explanations for special cases can be found in Kuncheva and Hoare (2008).

In this paper we are not interested in classification accuracy, but rather in the effects of neglecting dependencies on likelihoods $\Pr(\mathbf{e} \mid h)$, $\Pr(\mathbf{e} \mid \overline{h})$ and their ratio. Moreover we are interested in

the extent to which the posterior probability can be affected by the likelihood ratio. To address these questions we assume that the true probability distribution over the hypothesis and evidence variables is represented in a Bayesian network (BN). We can then study the effects of changes in probability parameters specified in the network on our outcomes of interest. Moreover, the effects of neglecting dependencies can be studied both theoretically and empirically by comparing the true Bayesian network to a naive Bayesian network approximation. We will propose a new measure to quantify the dependencies neglected by the naive approach and derive bounds on the error in overall likelihood caused by neglecting these dependencies.

This paper is organised as follows. In Section 2 we present the necessary technical background for this paper. In Section 3 we study the impact of the likelihood ratio on the relation between prior and posterior probabilities, followed by an analysis of the effects of parameter changes on the likelihood ratio in Section 4. The errors in overall likelihood as a result of neglecting dependencies is studied both theoretically and empirically in Sections 5 and 6, respectively. We conclude the paper in Section 7.

## 2. Preliminaries

In this paper we consider a joint probability distribution $\Pr(\mathbf{V})$ over a finite set of discrete stochastic variables $\mathbf{V}$; the cardinality of $\mathbf{V}$ is denoted $\#\mathbf{V}$. We assume all variables $V \in \mathbf{V}$ to be binary-valued, with $v$ and $\overline{v} \in \mathbf{cf}(V)$ indicating the possible configurations of $V$. Boldfaced letters are used to indicate sets of variables (upper case) or value-assignments to such sets (lower case).

This paper concerns likelihoods, and likelihood ratios, that basically link prior distributions over an hypothesis variable $H$ to posterior distributions. More specifically, the *posterior odds* for hypotheses $h$ and $\overline{h}$, given evidence $\mathbf{e}$ for a set of evidence variables $\mathbf{E}$, equals the *likelihood ratio* (LR) times the *prior odds* for the hypotheses:

$$\frac{\Pr(h \mid \mathbf{e})}{\Pr(\overline{h} \mid \mathbf{e})} = \frac{\Pr(\mathbf{e} \mid h) \cdot \Pr(h) / \Pr(\mathbf{e})}{\Pr(\mathbf{e} \mid \overline{h}) \cdot \Pr(\overline{h}) / \Pr(\mathbf{e})} = \frac{\Pr(\mathbf{e} \mid h)}{\Pr(\mathbf{e} \mid \overline{h})} \cdot \frac{\Pr(h)}{\Pr(\overline{h})}$$

We will assume that the likelihood ratio, as well as $\Pr(\mathbf{e})$ and any other probability used in the denominator of a fraction, is and remains strictly positive.

The joint probability distribution $\Pr(\mathbf{V})$ under consideration can be captured in a Bayesian network $\mathcal{B} = (G, \Pr)$, where $G = (\mathbf{V}, \mathbf{A})$ is a directed acyclic graph representing the independence relation among the variables by means of the well-known concept of d-separation (Jensen and Nielsen, 2007). The nodes in the graph have a one-to-one correspondence with the stochastic variables $\mathbf{V}$. For a given node $V \in \mathbf{V}$, $\pi_V$ indicates its set of parents in $G$. A joint probability for $\mathbf{v} \in \mathbf{cf}(\mathbf{V})$ is now uniquely defined as the product of the local probabilities associated with each node in the graph and compatible with $\mathbf{v}$:

$$\Pr(\mathbf{v}) = \prod_{V \in \mathbf{V}} \Pr(v \mid \boldsymbol{\pi}_V)$$

The local distributions are typically specified as conditional probability tables (CPTs). The probabilities specified in these tables are termed the network's *parameters* and are bound to be inaccurate.

The robustness of outcomes of a Bayesian network (BN) to parameter inaccuracies can be studied using a sensitivity analysis. Such an analysis investigates the effects of changing one or more network parameters on an output probability of interest. One approach to performing a sensitivity

analysis is to compute *sensitivity functions* (Kjærulff and van der Gaag, 2000). One-way sensitivity functions $f_{\Pr(h|\mathbf{e})}(x)$ that describe a posterior probability $\Pr(h \mid \mathbf{e})$ as a function of a single network parameter $x$ are fractions of two linear expressions in $x$, taking the form of a rectangular hyperbola (van der Gaag and Renooij, 2001).

The effects of arc removal on an output probability can be studied using sensitivity functions that capture the effect of multiple simultaneous parameter changes (Renooij, 2010). Parameter changes that result in simulating the removal of an arc $V_1 \rightarrow V_2$ are those that effectuate a zero *qualitative influence* $S^0(V_1, V_2)$ between the two variables (Wellman, 1990); such a zero influence is defined by $\Pr(v_2 \mid v_1\, \mathbf{z}) - \Pr(v_2 \mid \overline{v}_1\, \mathbf{z}) = 0$ for all $\mathbf{z} \in \mathbf{cf}(\mathbf{Z})$ where $\mathbf{Z} = \pi_{V_2} \setminus \{V_1\}$.

## 3. Likelihood-ratio Impact on Relation between Prior and Posterior Probability

In this section we investigate how much impact the likelihood ratio ($\mathsf{LR}$) can have on the relation between a prior probability $\Pr(h)$ and the posterior $\Pr(h \mid \mathbf{e})$. To this end we construct a function that relates the posterior to the prior for a given value of $\mathsf{LR}$ and study some of its properties. Since the hypothesis variable $H$ is assumed to be binary-valued, this function allows for establishing conclusions concerning prior and posterior odds for $H$ as well.

**Proposition 1** *Consider probability* $\Pr(h \mid \mathbf{e})$ *and likelihood ratio* $\mathsf{LR} = \Pr(\mathbf{e} \mid h)/\Pr(\mathbf{e} \mid \overline{h})$. *Then the function* $f_{\Pr(h|\mathbf{e})}(x)$ *relating* $\Pr(h \mid \mathbf{e})$ *to* $x = \Pr(h)$ *is given by*

$$f_{\Pr(h|\mathbf{e})}(x) = \frac{\mathsf{LR} \cdot x}{(\mathsf{LR} - 1) \cdot x + 1} \tag{1}$$

*and has the following properties:*

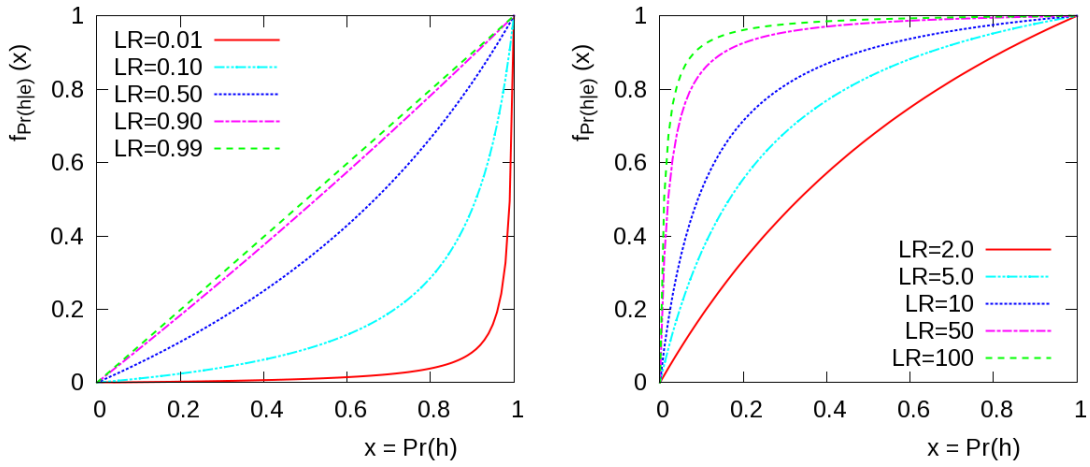
- $f(0) = 0$ *and* $f(1) = 1$, *regardless of the value of* $\mathsf{LR}$, *and* $f(x) = x$ *if* $\mathsf{LR} = 1$;
- *it is an increasing function in* $x$, *convex for* $\mathsf{LR} \in \langle 0, 1\rangle$, *and concave for* $\mathsf{LR} \in \langle 1, \infty\rangle$.

**Proof** Let $y$ denote $\Pr(h \mid \mathbf{e})$ then

$$\begin{aligned}
\frac{\Pr(h \mid \mathbf{e})}{\Pr(\overline{h} \mid \mathbf{e})} = \frac{\Pr(\mathbf{e} \mid h)}{\Pr(\mathbf{e} \mid \overline{h})} \cdot \frac{\Pr(h)}{\Pr(\overline{h})} \quad &\Longleftrightarrow \quad \frac{y}{1-y} = \mathsf{LR} \cdot \frac{x}{1-x} \\
&\Longleftrightarrow \quad y = \mathsf{LR} \cdot \frac{x}{1-x} - y \cdot \mathsf{LR} \cdot \frac{x}{1-x} \\
&\Longleftrightarrow \quad y = \frac{\mathsf{LR} \cdot x/(1-x)}{(1 - x + \mathsf{LR} \cdot x)/(1-x)}
\end{aligned}$$

The first property[1] follows immediately from this equation; as a result the function is obviously increasing for any $\mathsf{LR}$. Like one-way sensitivity functions in BNs, the function takes the form of a rectangular hyperbola. More specifically, for $x \in [0, 1]$ the function is part of one of the hyperbola's branches, whose shape is determined by its asymptotes. The vertical asymptote $x = s$ of the function is found at $s = \frac{-1}{\mathsf{LR}-1}$ and the horizontal asymptote $y = t$ at $t = \frac{\mathsf{LR}}{\mathsf{LR}-1}$. Now for $\mathsf{LR} \in \langle 0, 1\rangle$ we have that $s > 1$, with increasing values for increasing $\mathsf{LR}$, and $t < 0$ with decreasing values for increasing $\mathsf{LR}$. As a result, the function is convex, with an upper bound of $f(x) = x$. For $\mathsf{LR} \in \langle 1, \infty\rangle$

1. We note that the odds version used in the derivation requires that $x, y \neq 1$; the final function $f$, however, is no longer based on these odds and can be safely used for probabilities $x, y = 1$.




(a) (b)

Figure 1: Functions $f_{\Pr(h|\mathbf{e})}(x)$, $x = \Pr(h)$, for different values of (a) $\mathsf{LR} \in \langle 0, 1\rangle$ and (b) $\mathsf{LR} \in \langle 1, \infty\rangle$.

we have that $s < 0$, with increasing values for increasing LR, and $t > 1$ with decreasing values for increasing LR. As a result, the function is concave, with $f(x) = x$ as lower bound. ■

We note that the function in Equation (1) corresponds to a BN sensitivity function if in the BN under consideration $H$ has no parents, and therefore $x$ is an actual network parameter.

Figure 1 illustrates the function $f_{\Pr(h|\mathbf{e})}(x)$ for different values of the likelihood ratio LR. We observe that the closer LR becomes to zero, the smaller the impact of the prior is on the posterior probability, for all but large values of $x$. In fact, the posterior odds will now typically be below 1. Similarly, the larger LR becomes, the smaller the impact of the prior is on the posterior probability, for all but small values of $x$. In this case, the posterior odds will typically be above 1. For a given value or order-of-magnitude for LR the function in Equation (1) thus provides insight in how important an accurate prior is for drawing reliable conclusions about the posterior.

## 4. Likelihood-ratio Sensitivity to Evidence Parameters

In the *naive* likelihood-ratio approach typically individual likelihood ratios for different pieces of evidence are multiplied to establish an overall likelihood ratio. That is, if $\mathbf{e}$ captures $n$ pieces of evidence $e_i$, $i = 1, \ldots, n$, then the overall likelihood ratio is computed from $n$ individual likelihood ratios as follows:

$$\frac{\Pr(\mathbf{e} \mid h)}{\Pr(\mathbf{e} \mid \overline{h})} = \frac{\Pr(e_n \mid h)}{\Pr(e_n \mid \overline{h})} \cdot \ldots \cdot \frac{\Pr(e_1 \mid h)}{\Pr(e_1 \mid \overline{h})} \tag{2}$$

The above computation is only correct if all pieces of evidence are mutually independent given hypothesis $h$. Now, suppose $e_1$ and $e_2$ are in fact not conditionally independent given $h$, then we should include $\Pr(e_2 \mid e_1\, h)$ rather than $\Pr(e_2 \mid h)$ in the above factorisation. The effect of this

error can be studied by interpreting $\Pr(e_2 \mid h)$ as a parameter $x$ of a naive Bayesian network and changing $x$ to a value corresponding with $\Pr(e_2 \mid e_1\, h)$.[2]

**Proposition 2** *Consider a BN $\mathcal{B}$ and likelihood ratio of interest $\mathsf{LR} = \Pr(\mathbf{e} \mid h)/\Pr(\mathbf{e} \mid \overline{h})$. Let $\mathbf{E}^{\mathsf{R}} = \{E_k \in \mathbf{E} \mid H \in \pi_{E_k} \subset \mathbf{E} \cup \{H\}\}$. Let $x = \Pr(e_i \mid \boldsymbol{\pi}_i\, h)$ and $y = \Pr(e_j \mid \boldsymbol{\pi}_j\, \overline{h})$ be parameters for $E_i, E_j \in \mathbf{E}^{\mathsf{R}}$, where $e_k$ and $\boldsymbol{\pi}_k \in \mathbf{cf}(\pi_{E_k} \setminus \{H\})$, $k = i, j$ are compatible with $\mathbf{e}$. Then the function $f_{\mathsf{LR}}(x, y)$ relating $\mathsf{LR}$ to $x$ and $y$ has the following properties:*

- $\dfrac{\partial}{\partial x} f_{\mathsf{LR}}(x, y) > 0 \quad$ *and* $\quad \dfrac{\partial}{\partial y} f_{\mathsf{LR}}(x, y) < 0$;
- $\dfrac{\partial}{\partial y} f_{\mathsf{LR}}(x, y) = -\dfrac{x}{y} \cdot \dfrac{\partial}{\partial x} f_{\mathsf{LR}}(x, y)$.

**Proof** Note that changes in $x$ only affect the numerator of $\mathsf{LR}$; likewise $y$ only affects the denominator. The relation between $\Pr(\mathbf{e} \mid h)$ and $x$, where $x_0$ denotes the value of $x$ as specified in $\mathcal{B}$, is now given by:

$$f_{\Pr(\mathbf{e}|h)}(x) = \frac{f_{\Pr(\mathbf{e}\,h)}(x)}{f_{\Pr(h)}(x)} = \frac{x \cdot \frac{\Pr(\mathbf{e}\,h)}{x_0}}{\Pr(h)} = c_1 \cdot x$$

for constant $c_1 > 0$. We similarly find $f_{\Pr(\mathbf{e}|\overline{h})}(y) = c_2 \cdot y$ for constant $c_2 > 0$. As a result we find:

$$\frac{\partial}{\partial x} f_{\mathsf{LR}}(x, y) = \frac{c_1}{c_2 \cdot y} \quad \text{and} \quad \frac{\partial}{\partial y} f_{\mathsf{LR}}(x, y) = \frac{-c_1 \cdot x}{c_2 \cdot y^2}$$

which proves the proposition. ■

We can use the above proposition, for example, to determine for specific values $x_0$ and $y_0$ what the direction and amount of change in $\mathsf{LR}$ will be for different parameter changes. For a change in direction $(x, y) = (1, 1)$, for example, the directional derivative for $f_{\mathsf{LR}}$ in $(x_0, y_0)$ equals $c_1 \cdot (y_0 - x_0)/c_2 \cdot y_0^2 \cdot \sqrt{2}$, which is strictly positive iff $y_0 > x_0$. As such, we can investigate the effects of compensating for neglecting the dependencies between both $e_i$ and $\boldsymbol{\pi}_i$, and $e_j$ and $\boldsymbol{\pi}_j$. Since the numerator and denominator of $f_{\mathsf{LR}}$ are both linear expressions in any of the evidence parameters, the above proposition and analysis generalises to multiple $x$s and $y$s.

The proposition also shows us that the error in $\mathsf{LR}$ due to neglecting one dependency can be compensated for by neglecting another one. Using the naive approach to computing an overall likelihood ratio therefore doesn't necessarily result in large errors in the overall likelihood ratio, even if independence assumptions are clearly violated in practice. A similar observation was done with respect to the optimality of naive Bayes classifiers: dependencies between evidence variables may cancel each other out without affecting the classification (Zhang, 2004).

We note that the above proposition applies to parameters of evidence variables that have $H$ as direct parent and all remaining parents, if any, among $\mathbf{E}$. Such constrained topology is typical for various Bayesian network classifiers, such as naive Bayes and TAN (Friedman et al., 1997).

2. We note that a change of $\Pr(e_2 \mid h)$ will require a change in $\Pr(\overline{e}_2 \mid h)$ as well; the latter change, however, will not affect the computation of the likelihood ratio under consideration and is therefore disregarded.

## 5. Combining Individual Likelihoods: Theoretical Error

Since the numerator and denominator of the likelihood ratio represent probabilities from different conditional distributions, the error as a result of neglecting dependencies among evidence variables in the numerator is in essence independent of the error in the denominator. Information about the error in computing both $\Pr(\mathbf{e} \mid h)$ and $\Pr(\mathbf{e} \mid \overline{h})$ provides insight in the error in LR, as well as in the posteriors $\Pr(h \mid \mathbf{e})$ and $\Pr(\overline{h} \mid \mathbf{e})$ and their ratio.

In the following two sections our analyses will focus on the error in $\Pr(\mathbf{e} \mid h)$; results for $\Pr(\mathbf{e} \mid \overline{h})$ will be equivalent if $h$ is replaced by $\overline{h}$. The error $\mathsf{Err}(\mathbf{e}|h)$ under consideration is now defined as

$$\mathsf{Err}(\mathbf{e}|h) = \Pr(\mathbf{e} \mid h) - \prod_{i=1}^{n} \Pr(e_i \mid h) \tag{3}$$

where $\#\mathbf{E} = n$ and each $e_i$ is compatible with $\mathbf{e}$. We first study the error in overall likelihood caused by neglecting the dependency between exactly *two* pieces of evidence.

**Proposition 3** *Consider likelihood* $\Pr(\mathbf{e} \mid h) > 0$ *for hypothesis* $h$ *and evidence* $\mathbf{e} \in \mathbf{cf}(\mathbf{E})$, $\#\mathbf{E} = n \geq 2$. *Let* $\mathbf{e}$ *include* exactly two *dependent pieces of evidence, for binary-valued variables* $E_i$ *and* $E_j$. *Then* $|\mathsf{Err}(\mathbf{e}|h)| \leq c \cdot \frac{1}{4}$ *with* $0 < c \leq 1$, *where* $c = 1$ *if* $n = 2$.

**Proof** Without loss of generality, take $i = 1$ and $j = 2$. Then, taking into account the (single!) dependency, we have for $e_i$, $i = 1, \ldots, n$, compatible with $\mathbf{e}$ that

$$\Pr(\mathbf{e} \mid h) = \Pr(e_1\, e_2 \mid h) \cdot \Pr(e_3 \ldots e_n \mid h) = \Pr(e_2 \mid e_1\, h) \cdot \Pr(e_1 \mid h) \cdot \prod_{i=3}^{n} \Pr(e_i \mid h)$$

Moreover, we have that $\Pr(e_2 \mid h) = \Pr(e_2 \mid e_1\, h) \cdot \Pr(e_1 \mid h) + \Pr(e_2 \mid \overline{e}_1\, h) \cdot \Pr(\overline{e}_1 \mid h)$. Now let's introduce the following short-hand notation:

$$\alpha = \Pr(e_1 \mid h),\ \beta_1 = \Pr(e_2 \mid e_1\, h),\ \beta_2 = \Pr(e_2 \mid \overline{e}_1\, h),\ \gamma = \prod_{i=3}^{n} \Pr(e_i \mid h)$$

$$\begin{aligned} \text{Then } \mathsf{Err}(\mathbf{e}|h) \ &= \ \big(\Pr(e_2 \mid e_1\, h) - \Pr(e_2 \mid h)\big) \cdot \Pr(e_1 \mid h) \cdot \prod_{i=3}^{n} \Pr(e_i \mid h) \\ &= \ \big(\beta_1 - (\beta_1 \cdot \alpha + \beta_2 \cdot (1 - \alpha))\big) \cdot \alpha \cdot \gamma \ = \ \gamma \cdot \alpha \cdot (1 - \alpha) \cdot (\beta_1 - \beta_2) \end{aligned}$$

We thus find that $\mathsf{Err}(\mathbf{e}|h) = 0$ whenever $\Pr(e_1 \mid h) = 1$ or $\Pr(e_2 \mid e_1\, h) = \Pr(e_2 \mid \overline{e}_1\, h)$, where the latter indeed indicates independence of the two pieces of evidence given $h$. $\alpha \cdot (1-\alpha)$ is maximal for $\alpha = 0.5$, therefore the error is bounded by $\mathsf{Err}(\mathbf{e}|h) \geq -\frac{1}{4} \cdot \gamma$ when $\beta_1 = 0$ and $\beta_2 = 1$, and $\mathsf{Err}(\mathbf{e}|h) \leq \frac{1}{4} \cdot \gamma$ when $\beta_1 = 1$ and $\beta_2 = 0$. Note that $\gamma$ is absent for $n = 2$; otherwise, taking $c = \gamma$ completes the proof. ■

Figure 2 shows the error in the likelihood $\Pr(e_1\, e_2 \mid h)$ as a result of disregarding the dependency between $e_1$ and $e_2$ for different values of $\Pr(e_1 \mid h)$ and different values of $\Pr(e_2 \mid e_1\, h) - \Pr(e_2 \mid \overline{e}_1\, h)$. Note that the latter difference can be interpreted as a measure of the strength of dependence

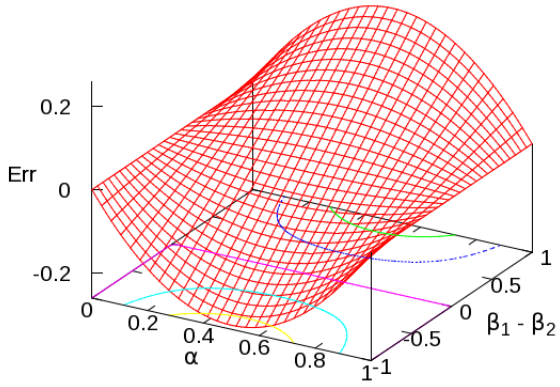



Figure 2: Error Err in likelihood $\Pr(e_1\, e_2 \mid h)$ as a function of $\alpha = \Pr(e_1 \mid h)$ and $\beta_1 - \beta_2 = \Pr(e_2 \mid e_1\, h) - \Pr(e_2 \mid \bar{e}_1\, h)$.

between $E_1$ and $E_2$ in the context of $h$, since $\Pr(e_2 \mid e_1\, h) = \Pr(e_2 \mid \bar{e}_1\, h)$ implies *in*dependence of $E_1$ and $E_2$ in the context $h$.

We note from the proof of the above proposition that the error in overall likelihood is reduced upon including more evidence for which the independence assumption is *not* violated. The question now is what happens to the error in overall likelihood when dependencies exist between more than two pieces of evidence. A theoretical analysis of the general case is complicated by the dimensionality of the problem and the fact that it is not clear-cut how exactly to quantify dependencies. Therefore, we take an experimental approach.

## 6. Combining Individual Likelihoods: Empirical Error

In this section we present an empirical analysis of the error $\mathsf{Err}(\mathbf{e}|h)$ defined in Equation (3).

### 6.1 Experimental set-up

We performed experiments with BNs of restricted topology with 2, 3 and 4 binary evidence variables and varying densities. In addition to computing $\mathsf{Err}(\mathbf{e}|h)$, we analysed the relation between the error and different measures of dependency, or *influence*.

#### 6.1.1 NETWORKS

Let $\mathcal{B}_n^c$ denote a Bayesian network with graph $G = (\{H\} \cup \mathbf{E}, \mathbf{A})$, where

- $\mathbf{E} = \{E_i \mid i = 1, \ldots, n\}$
- $\mathbf{A} = \begin{cases} \text{for } c = \text{`}\mathsf{F}\text{'}: & \mathbf{A}^{\mathsf{F}} = \{E_i \to E_j \mid j > i, 1 \leq i, j \leq n\} \cup \{H \to E_i \mid 1 \leq i \leq n\} \\ \text{for } c = \text{`} - C\text{'}: & \mathbf{A}^{\mathsf{F}} \setminus \{E_i \to E_j \mid ij \in C\} \end{cases}$

then we used the following network types defined by 10 different graph structures:

$$\mathcal{B}_2^{\mathsf{F}}, \mathcal{B}_3^{\mathsf{F}}, \mathcal{B}_4^{\mathsf{F}}, \mathcal{B}_3^{-\{13\}}, \mathcal{B}_3^{-\{23\}}, \mathcal{B}_4^{-\{14\}}, \mathcal{B}_4^{-\{14,24\}}, \mathcal{B}_4^{-\{14,24,13\}}, \mathcal{B}_4^{-\{14,24,23\}}, \mathcal{B}_4^{-\{14,24,13,23\}}$$

For example, $\mathcal{B}_4^{\mathsf{F}}$ has four evidence variables in a fully connected subgraph; in $\mathcal{B}_4^{-\{14,24\}}$ the two arcs $E_1 \rightarrow E_4$ and $E_2 \rightarrow E_4$ are removed from this subgraph.

Furthermore for each of the networks $\mathcal{B}_2^{\mathsf{F}}$, $\mathcal{B}_3^{\mathsf{F}}$, and $\mathcal{B}_4^{\mathsf{F}}$, distributions $\Pr(\mathbf{E}' \mid h)$, $\mathbf{E}' \subseteq \mathbf{E}$, were defined to be the same for the variables shared by the networks. That is, $\Pr(E_1\, E_2 \mid h)$ was the same in all three networks, and $\Pr(E_1\, E_2\, E_3 \mid h)$ was the same in $\mathcal{B}_3^{\mathsf{F}}$ and $\mathcal{B}_4^{\mathsf{F}}$. This was accomplished by generating random numbers for the following probabilities:

$$\Pr(e_j \mid e^*_{j-1} \ldots e^*_1\, h) \text{ for all } e^*_{j-1} \ldots e^*_1 \in \mathbf{cf}(E_j \ldots E_1), \text{ for } j = 2, 3, 4.$$

For each of these probabilities we generated 1000 random numbers, thus obtaining 1000 different distributions $\Pr(E_4\, E_3\, E_2\, E_1 \mid h)$. None of the generated probabilities equaled zero or one. Moreover, no two pieces of evidence were conditionally independent in the generated distributions.

The removal of an arc $E_i \rightarrow E_j$ was subsequently implemented by enforcing a zero qualitative influence in the CPT of $E_j$ in the fully connected network, that is, by setting $\Pr(e_j \mid \overline{e}_i\, \mathbf{z}\, h^*)$ equal to $\Pr(e_j \mid e_i\, \mathbf{z}\, h^*)$, for each $h^* \in \mathbf{cf}(H)$ and each $\mathbf{z} \in \mathbf{cf}(\mathbf{Z})$, $\mathbf{Z} = \pi_{E_j} \setminus \{E_j, H\}$. We thus obtained a total of $1000 \times 10 = 10,000$ models for which we computed $\mathsf{Err}(\mathbf{e}|h)$ where $\mathbf{e}$ was compatible with $E_i = e_i$ for each of the evidence variables under consideration. Note that probabilities conditioned on $\overline{h}$ are irrelevant for these computations and hence are not further discussed.

### 6.1.2 MEASURES OF INFLUENCE STRENGTH

In studying the accuracy of naive Bayes classifiers, previous studies have employed mutual information $I$, and Yule's $Q$ statistic as measures of dependency or influence between evidence variables (Domingos and Pazzani, 1997; Rish et al., 2001; Kuncheva and Hoare, 2008). In that context, all available hypotheses are taken into account by computing expected values of the measures over all possible values of $H$. In this paper we employ the same measures. However, since we focus on the error in overall likelihood for a specific hypothesis $h$, rather than on classification error, no expected values over $H$ need to be considered, so the measures are simply defined by:

- $I_{ij} = I(E_i, E_j|h) = \displaystyle\sum_{e^*_i \in \mathbf{cf}(E_i)} \sum_{e^*_j \in \mathbf{cf}(E_j)} \Pr(e^*_i\, e^*_j \mid h) \cdot \log \frac{\Pr(e^*_i\, e^*_j \mid h)}{\Pr(e^*_i \mid h) \cdot \Pr(e^*_j \mid h)}$
- $Q_{ij} = \dfrac{\Pr(e_i\, e_j \mid h) \cdot \Pr(\overline{e}_i\, \overline{e}_j \mid h) - \Pr(\overline{e}_i\, e_j \mid h) \cdot \Pr(e_i\, \overline{e}_j \mid h)}{\Pr(e_i\, e_j \mid h) \cdot \Pr(\overline{e}_i\, \overline{e}_j \mid h) + \Pr(\overline{e}_i\, e_j \mid h) \cdot \Pr(e_i\, \overline{e}_j \mid h)}$

We note that $I_{ij}$ is non-negative (see Cover and Thomas (2006)); $Q_{ij}$ can be positive or negative (see Yule and Kendall (1940)).

In Section 5 we showed that for two evidence variables $\mathsf{Err}(e_1\, e_2|h)$ can be expressed as a function of $\Pr(e_2 \mid e_1\, h) - \Pr(e_2 \mid \overline{e}_1\, h)$, where this difference can be interpreted as a measure of dependency, or influence, between $E_1$ and $E_2$ given $h$. Exploiting the similarity with the definition of positive (negative) qualitative influences $S^+$ ($S^-$), we will now define another measure of influence strength based upon such differences. Since evidence variables can have multiple other variables as parents, we require our measure to aggregate these contexts of 'other parents'. To this end an aggregation operator is used, represented by placeholder $\odot$ in the following definition.

**Definition 4** *Consider a BN and probability of interest* $\Pr(h \mid \mathbf{e})$. *Then the* influence strength $R^{\odot}_{ij}(\mathbf{e}|h)$ *associated with arc* $E_i \rightarrow E_j$ *is defined by*

$$R^{\odot}_{ij}(\mathbf{e}|h) = \odot_k(\Pr(e_j \mid e_i\, \mathbf{z}_k\, h) - \Pr(e_j \mid \overline{e}_i\, \mathbf{z}_k\, h))$$

| | | | correlation with Err (for $R$, $Q$) or \|Err\| (for $I$) | | | |
|---|---|---|---|---|---|---|
| $\mathcal{B}_n^c$ | $\max \lvert\mathsf{Err}\rvert$ | avg$\lvert\mathsf{Err}\rvert$ | $R_{\mathsf{tot}}^{\mathrm{avg}}$ | $R_{\mathsf{tot}}^{\mathrm{sum}}$ | $Q_{\mathsf{tot}}$ | $I_{\mathsf{tot}}$ |
| $\mathcal{B}_2^{\mathsf{F}}$ | 0.234 | 0.057 | 0.917 | 0.917 | 0.878 | 0.932 |
| $\mathcal{B}_3^{\mathsf{F}}$ | 0.261 | 0.056 | 0.726 | 0.712 | 0.759 | 0.471 |
| $\mathcal{B}_4^{\mathsf{F}}$ | 0.263 | 0.037 | 0.580 | 0.510 | 0.400 | 0.180 |
| $\mathcal{B}_3^{-\{13\}}$ | 0.256 | 0.048 | 0.816 | 0.796 | 0.813 | 0.554 |
| $\mathcal{B}_3^{-\{23\}}$ | 0.273 | 0.043 | 0.826 | 0.789 | 0.816 | 0.643 |
| $\mathcal{B}_4^{-\{14\}}$ | 0.273 | 0.036 | 0.625 | 0.557 | 0.429 | 0.232 |
| $\mathcal{B}_4^{-\{14,24\}}$ | 0.282 | 0.033 | 0.675 | 0.614 | 0.151 | 0.275 |
| $\mathcal{B}_4^{-\{14,24,13\}}$ | 0.300 | 0.030 | 0.726 | 0.659 | 0.517 | 0.320 |
| $\mathcal{B}_4^{-\{14,24,23\}}$ | 0.288 | 0.028 | 0.738 | 0.658 | 0.575 | 0.454 |
| $\mathcal{B}_4^{-\{14,24,13,23\}}$ | 0.185 | 0.022 | 0.709 | 0.609 | 0.678 | 0.469 |

Table 1: For 1000 distributions for each of the 10 different network structures we show: maximum and average absolute overall error $\lvert\mathsf{Err}\rvert$ in the likelihood, and correlations between Err and strengths of dependency among the variables according to 4 influence measures.

*where $e_i$ and $e_j$ are compatible with* $\mathbf{e}$*,* $\mathbf{z}_k \in \mathbf{cf}(\mathbf{Z})$*, where* $\mathbf{Z} = \pi_{E_j} \setminus \{E_i, H\}$*, and* $\odot$ *is an* $n$*-ary operator,* $n = \#\mathbf{Z}$*.*

We note that our definition of influence strength is specifically tailored to our problem at hand:

- it fixes the value of $H$ to the one under consideration;
- it is explicitly defined for an arc $E_i \rightarrow E_j$ and depends on its direction by including all parents of $E_j$, and is therefore asymmetric;
- its evidence variables are assumed to be binary-valued, and their specific values used in the minuend and subtrahend of the subtraction are determined by their values in $\mathbf{e}$.[3]

To capture the total influence strength among the evidence variables, we sum over all the arcs $E_i \rightarrow E_j$ among the variables $\mathbf{E}$:

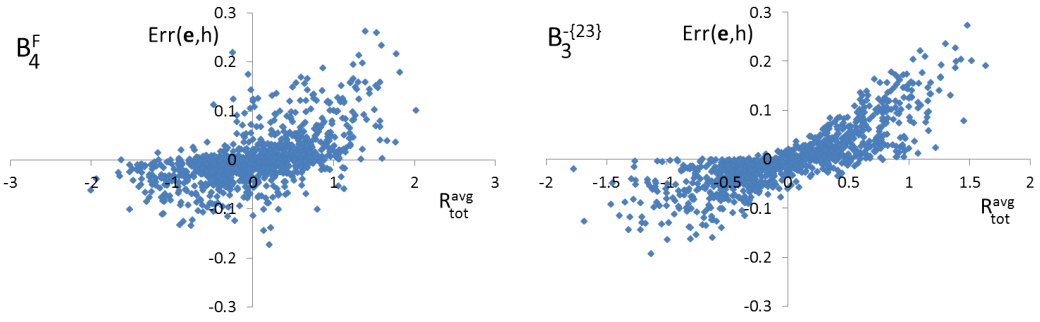
$$R_{\mathsf{tot}}^{\odot} = \sum_{ij} R_{ij}^{\odot}, \quad I_{\mathsf{tot}} = \sum_{ij} I_{ij}, \quad Q_{\mathsf{tot}} = \sum_{ij} Q_{ij}$$

For each of the 10 types of network we then computed the correlation between $\mathsf{Err}(\mathbf{e}\,|\,h)$ and these measures of total influence strength. Note that due to the restricted topology of the used networks, measure $R_{ij}^{\odot}$ requires only probabilities available in the CPT of $E_j$.

### 6.2 Results

We will first consider the differences in $\mathsf{Err}(\mathbf{e}\,|\,h)$ for the different types of network, and then discuss the correlations with the influence measures.

3. That is, for $\mathbf{e} = \overline{e}_1\, e_2$ we use $\Pr(e_2 \mid \overline{e}_1\, h) - \Pr(e_2 \mid e_1\, h)$ and for $\mathbf{e} = e_1\, \overline{e}_2$ we take $\Pr(\overline{e}_2 \mid e_1\, h) - \Pr(\overline{e}_2 \mid \overline{e}_1\, h)$.




(a) (b)

Figure 3: Error $\mathsf{Err}(\mathbf{e}|h)$ as a function of total mean influence $R^{\text{avg}}_{\text{tot}}$ for 1000 distributions for (a) $\mathcal{B}^{\mathsf{F}}_4$ and (b) $\mathcal{B}^{-\{23\}}_3$.

### 6.2.1 ERRORS: DEPENDENT VERSUS INDEPENDENT EVIDENCE

Since the error in overall likelihood can be positive or negative we report the maximum and arithmetic mean for the *absolute* values of the error. Results for the 10 types of network are presented in Table 1. Indeed for the networks with only two evidence variables, the maximum absolute error $|\mathsf{Err}|$ is below the theoretical bound of $0.25$. For the networks including more than two evidence variables and varying dependencies, the maximum absolute error has an empirical bound of $0.30$ (we note that the maximum of $0.300$ reported for networks $\mathcal{B}^{-\{14,24,13\}}_4$ was in fact rounded up). The average absolute error is rather small, indicating that among the 1000 distributions enough had such small errors that they can compensate for those with larger errors. Introducing more evidence variables and more independencies seems to have a positive effect on the average error, but no true effect on the maximum error. For the networks with 3 evidence variables, the average error decreases upon introducing conditional independencies by arc removal, but the maximum error in fact increases upon removing $E_2 \rightarrow E_3$. For the networks with 4 evidence variables we see similar behaviour. Only upon removing 4 out of 6 arcs from $\mathcal{B}^{\mathsf{F}}_4$ do we see a clear decrease in maximum error: in that case only $E_1 \rightarrow E_2$ and $E_3 \rightarrow E_4$ remain.

### 6.2.2 ERROR VERSUS INFLUENCE STRENGTH

In the previous subsection we have compared the maximum and average absolute errors for different network structures. For a given structure, however, the strength of the dependencies captured by the arcs among the evidence variables vary over the 1000 distributions considered. In this section we investigate if we can relate error to influence strength.

For our influence measure $R^{\odot}_{\text{tot}}(\mathbf{e}|h)$ we used two different operators $\odot$ to incorporate the context provided by 'other' evidence parents: summation ($\mathrm{sum}$) and arithmetic mean ($\mathrm{avg}$). Figure 3 shows the overall error in likelihood as a function of the total average influence strength $R^{\text{avg}}_{\text{tot}}(\mathbf{e}|h)$ for the 1000 different networks of type $\mathcal{B}^{\mathsf{F}}_4$ (3(a)) and of type $\mathcal{B}^{-\{23\}}_3$ (3(b)). Moreover, Table 1 shows the correlations between the overall error in likelihood and the different measures of influence strength. More precisely, it shows the correlations between $\mathsf{Err}(\mathbf{e}|h)$ and our total mean

influence strength, our total summed influence strength ($R^{\text{sum}}_{\mathsf{tot}}(\mathbf{e}|h)$), and $Q_{\mathsf{tot}}$; for $I_{\mathsf{tot}}$ it gives the correlations with the absolute values $|\mathsf{Err}(\mathbf{e}|h)|$ of the error, since $I$ is always positive.

The plot in Figure 3(a) displays $\mathsf{Err}(\mathbf{e}|h)$ for the fully connected networks with four evidence variables. The correlation with the used influence measure $R^{\text{avg}}_{\mathsf{tot}}(\mathbf{e}|h)$ is $0.580$, which is the lowest correlation among all correlations with this same influence measure (see Table 1). Note that this network type has the lowest correlation for all influence measures; nonetheless the correlation is higher with $R^{\text{avg}}_{\mathsf{tot}}(\mathbf{e}|h)$ than with the other measures. The plot in Figure 3(b) displays $\mathsf{Err}(\mathbf{e}|h)$ for the networks of type $\mathcal{B}_3^{-\{23\}}$ with three evidence variables and two arcs among them. Here, the correlation with the used influence measure $R^{\text{avg}}_{\mathsf{tot}}(\mathbf{e}|h)$ is $0.826$, which is quite high. From Table 1 we have that all influence measures show similar patterns with correlations decreasing with the addition of more evidence variables, and increasing with the introduction of independences by arc removal. Overall, measure $R^{\text{avg}}_{\mathsf{tot}}(\mathbf{e}|h)$ results in the highest correlations with $\mathsf{Err}(\mathbf{e}|h)$ compared to the other measures, with two exceptions: for $\mathcal{B}_2^{\mathsf{F}}$, the measure $I_{\mathsf{tot}}$ has a higher correlation and for $\mathcal{B}_3^{\mathsf{F}}$, measure $Q_{\mathsf{tot}}$ results in higher correlation. It seems that our tailored measure of influence strength is better at capturing the influences that affect the error than the other influence measures. This observation can partly be attributed to the fact that our measure incorporates 'other parents' of the relations $E_i \rightarrow E_j$ under consideration.

## 7. Conclusions and Further Research

In this paper we have presented a number of results that can be used to assess both the importance of a reliable LR estimate and the impact on overall likelihood of neglecting dependencies among pieces of evidence. Since likelihood is an ingredient used in naive Bayesian network classification, our results also serve to further study optimality conditions for naive Bayes. The plots in Figure 1 show the relation between prior and posterior probabilities of an hypothesis for different values of the LR and give insight in the importance of accuracy of both the LR and the prior for drawing robust conclusions concerning the hypothesis. The properties concerning the sensitivity of the LR to evidence parameter changes show that dependencies between pieces of evidence can cancel each other out in the overall likelihood; the exact effects can be analysed.

We have proven that the error in overall likelihood due to neglecting dependencies has a theoretical bound of $0.25$ when exactly two pieces of evidence are dependent. We established a preliminary empirical bound of $0.30$ for networks with more than two dependent pieces of evidence, where our experiments show a decrease in average absolute error upon introducing more evidence variables and/or more independences. Our experiments moreover show that our proposed influence measure, tailored to the problem at hand, can provide an indication for the error in overall likelihood that can be expected from neglecting independences. Even without exact values for the numbers involved, the measure is simple enough to obtain an indication of its value from rough estimates.

We have experimented with 10 types of network, with at most four evidence variables. We would therefore like to further experiment with larger networks to see if our empirical bound is preserved. Moreover, we would like to compare our bound to a bound on the error that results from replacing an extreme joint distribution by the product of marginals (Rish et al., 2001). Our random distributions did not include extreme distributions, nor do they adhere to the typical monotonicity properties often found in practice. Therefore we would like to run experiments with real networks as well. Finally, we would like to further fine-tune our influence measure and investigate if a theoretical error bound can be formulated in terms of such a measure.

## References

C. Aitken and F. Taroni. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, Ltd, 2004.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29: 131–163, 1997.

D. Hand and K. Yu. Idiot's Bayes – not so stupid after all. *International Statistical Review*, 69: 385–398, 2001.

F. Jensen and T. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Verlag, 2007.

U. Kjærulff and L. van der Gaag. Making sensitivity analysis computationally efficient. In C. Boutilier and M. Goldszmidt, editors, *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 317–325, San Francisco, 2000. Morgan Kaufmann.

L. Kuncheva and Z. Hoare. Error-dependency relationships for the naive Bayes classifier with binary features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:735–740, 2008.

S. Renooij. Bayesian network sensitivity to arc-removal. In P. Myllymäki, T. Roos, and T. Jaakkola, editors, *Proceedings of the 5th European Workshop on Probabilistic Graphical Models*, pages 233–240, Helsinki, Finland, 2010. HIIT Publications.

I. Rish, J. Hellerstein, and J. Thathachar. An analysis of data characteristics that affect naive Bayes performance. Technical report, IBM TJ Watson Research Center, 2001.

J. Thornbury, D. Fryback, and W. Edwards. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. *Radiology*, 114:561–565, 1975.

L. van der Gaag and S. Renooij. Analysing sensitivity data from probabilistic networks. In J. Breese and D. Koller, editors, *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 530–537, San Francisco, 2001. Morgan Kaufmann.

M. Wellman. Graphical inference in qualitative probabilistic networks. *Networks*, 20:687–701, 1990.

G. Yule and M. Kendall. *An Introduction of the Theory of Statistics*. Griffin Co. Ltd., 1940.

H. Zhang. The optimality of naive Bayes. In *Proceedings of the 17th International FLAIRS Conference*, Florida, USA, 2004. AAAI Press.