Bachelor thesis

Artificial Intelligence

Radboud University

Test-selection strategies to increase the number of MAP-independent variables

Author:
Simon Janssen
s1035310

First supervisor:
dr. J.H.P. Kwisthout
Artificial Intelligence
johan.kwisthout@donders.ru.nl

Second reader:
prof. dr. I.J.E.I. van Rooij
Artificial Intelligence
i.vanrooij@donders.ru.nl

June 20, 2022

Abstract

In practical applications, decision support systems that motivate and justify decisions are essential for a user to understand and accept the systems. Recently, the concept of MAP-independence was introduced which aimed to provide such justification for Bayesian networks by explicating the relevant intermediate variables for decisions in MAP problems. Whereas the concept of MAP-independence may work well to justify a decision when the number of relevant variables is small, in this paper we argue that for scenarios in which the number of relevant variables is large, these justifications would be inadequate and a user can only gain limited insight into the decision. Therefore, we focus on finding test-selection strategies that can decrease the number relevant variables in the network by deciding which variable to gather evidence for in order to obtain a decision that is better to justify and motivate. Specific test-selection strategies that we investigate include the in-degree, out-degree and total degree of a variable, the distance of a variable to the explanation variable, the expected utility, the expected Gini index and mutual information with the explanation variable. After running systematic simulations on the ALARM network and analysing the results based on rank- and value-approximation, we conclude that the distance to the explanation variable, expected utility, expected Gini index and mutual information could serve as good test-selection strategies to decrease the number of relevant variables. However, more research is needed to be able to generalize these findings to a larger population of Bayesian networks.

Contents

Chapter 1

Introduction

Bayesian networks have become increasingly popular and relevant over the past years (Gallego, 2005). A Bayesian network, viz. a factored representation of a joint probability distribution, can compute any probability of interest over its variables, which is useful for belief updating, finding a Maximum a Posteriori (MAP) assignment or Most Probable Explanation (MPE) amongst other things (Yuan, Lim, & Lu, 2011). Particularly as an underlying statistical model of decision support systems, Bayesian networks find their application in a variety of public domains. Examples include the field of medicine where Bayesian networks are used to diagnose a patient with a medical condition, given a set of symptoms and health complaints (Kyrimi, McLachlan, Dube, Neves, Fahmi, & Fenton, 2021), weather prediction where Bayesian networks can be used to locally forecast whether or not it will rain, given meteorological variables such as temperature (Cofiño, Cano, Sordo, & Gutiérrez, 2002) or law where Bayesian networks can help a judge or jury in making a verdict based on evidence (Vlek, Prakken, Renooij, & Verheij, 2016).

However, in these practical applications, Bayesian networks underlying decision support systems that merely give the best explanation for the evidence are not enough. A decision support system should be able to support its users, as the name already suggests. This means that the support system, and in particular the Bayesian network as the underlying statistical model of the system, should be able to justify the decisions that it makes (Barredo Arrieta et al., 2020; Lacave & Díez, 2002). The subfield of *explainable AI* which recently emerged has focused on providing motivations and justifications to the users of artificial intelligence (AI) systems, including Bayesian networks. Explainable AI creates a sense of mutual understanding, which is necessary for users to trust, understand and accept AI systems (Gunning, Stefik, Choi, Miller, Stumpf, & Yang, 2019).

In the context of Bayesian networks, one popular approach towards making the decision process more understandable to users is by explicating the

relevant information for the decision. For example, several studies have focused on identifying the relevant variables in the *evidence set* (Kyrimi & Marsh, 2016; Meekes, Renooij, & Van der Gaag, 2015; Suermondt & Cooper, 1993). This allows a user to see which of the evidence variables are influential for the final explanation and why (Suermondt & Cooper, 1993). Other studies have focused on identifying relevant variables in the *explanation set* (Yuan et al., 2011; Gallego, 2005). When the explanation set is large, this is especially useful because it is typically undesirable to burden the user with a large explanation set containing only a few relevant variables. Instead, the aim should be to provide precise and concise explanations with only the most relevant variables (Yuan et al., 2011).

Recently, Kwisthout (2021) proposed a new method to explain the reasoning process of a decision support system by assessing the contribution of *intermediate* variables towards the decision, based on early work by Pearl and Paz (1987). This method focuses on the justification in MAP problems, in which the most probable joint value assignment to the explanation set needs to be computed, given the evidence. In the process of computing the most probable joint value assignment, intermediate variables have to be marginalized out. Kwisthout (2021) argues that this makes the decision more opaque, as information about which intermediate variables have a big impact on the final decision is lost in the marginalization process. Therefore, the concept of MAP-independence is introduced, with which the intermediate variables that are relevant and irrelevant for the best explanation can be identified.

An intermediate variable is considered MAP-independent from the best explanation when the best explanation would be the same, irrespective of which value the intermediate variable takes on and irrespective of whether the variable is observed or not. That is, observing the intermediate variable does not impact the decision. In addition to single intermediate variables, MAP-independence can also be determined for a set of intermediate variables. For this, two variants exist. When the explanation is MAP-independent from each variable in the set independently, the set is called weakly MAP-independent. When interactions between variables of the set are also MAP-independent from the explanation, the set is called strongly MAP-independent (Kwisthout, 2021).

To illustrate the difference (Kwisthout, 2021, p. 8), two variables B and E can be weakly MAP-independent from the best explanation for a variable A. That is, observing a particular value for B alone does not change the best explanation for A and similarly, observing a particular value for E alone does not change the best explanation for A. However, it might be that the combination of observing a particular value for B and E does change the best explanation for A. In that case, the set $\{B, E\}$ is not strongly MAP-independent from A.

Whereas MAP-independence could provide a means to justify the deci-

sions of a decision support system, there are scenarios in which this justification may be inadequate and a user can only gain a limited understanding of the decision. For instance, when the number of relevant variables for a MAP problem according to MAP-independence is large, simply identifying these variables does not help the user much in gaining more insight into the decision. After all, it would be very difficult for a user to comprehend a decision when nearly all unobserved intermediate variables could influence the decision. Therefore, it could be very useful to find a way to decrease the number of relevant variables in order to obtain a decision that is better to motivate.

This may be achieved by gathering additional evidence, which relates to the field of selective evidence gathering. In this field, the focus is on finding test-selection strategies to select the 'best' next variable to observe at a certain stage in the diagnostic process, according to some measure of 'best'. For the problem described above, this means finding a test-selection strategy to increase the number of MAP-independent variables, thereby decreasing the number of relevant variables that can influence the decision. The difficulty with finding the 'best' variable to observe (when solved exactly) is that it is highly intractable in general (Van der Gaag & Bodlaender, 2011), which means that this will be infeasible for larger networks. Hence, previous studies have been looking for heuristics rather than exact procedures. Here, the 'best' variable to gather evidence for was defined as the variable that is expected to provide the most valuable information in the diagnostic process, i.e. the variable that is most informative (Sent & Van der Gaag, 2006; Madigan & Almond, 1996).

In this regard, most previous studies have focused on finding test-selection strategies based on a utility function that can describe for a variable the usefulness of 'knowing' the value of that variable (Van der Gaag & Wessels, 1993). These functions can be based on probabilistic information only. For example, Sent and Van der Gaag (2007) examined the usefulness of the Shannon entropy, Gini index and misclassification error for test-selection purposes. Furthermore, studies have focused on the expected weight of the evidence (Madigan & Almond, 1996) and the expected difference in the posterior probability of the explanation variable for when a variable is observed versus not observed (Van der Gaag & Wessels, 1993). However, a utility function can also incorporate non-probabilistic information. For example, a utility function might incorporate the cost of applying a certain test, as was investigated by Glasziou and Hilden (1989).

Whereas the aim of previous studies has always been to find test-selection strategies that make the best explanation more robust and stable (Van der Gaag & Wessels, 1993) by means of finding the most informative variable, there has not yet been an investigation into a test-selection strategy explicitly aimed at reducing the number of relevant variables in a Bayesian network to the best of our knowledge. This includes test-selection strategies reduc-

ing the number of relevant variables in the context of MAP-independence. While it could be that some of the proposed heuristics in previous work already increase the number of MAP-independent variables implicitly (thereby reducing the number of relevant intermediate variables), more research on this is needed. This will be the focus of the current study. Specifically, in this study we will investigate which test-selection strategy could be used as a heuristic to increase the number of MAP-independent variables. Here, we will focus on weak MAP-independence, given the unfavorable computational complexity of establishing strong MAP-independence (Kwisthout, 2021). Both terms will be described later in the preliminaries.

In order to find a test-selection strategy that can be used as a heuristic, systematic simulations will be run on the benchmarking ALARM network. In each simulation, the 'best' variable to gather evidence for can be identified (where 'best' is defined as resulting in the largest number of MAP-independent variables). After running many simulations, properties of the 'best' variable will be investigated for structural parameters. These include the utility function as described by van der Gaag and Wessels (1993), but also other parameters such as in- or out-degree, the distance between variables in the graphical representation of the Bayesian network, the expected Gini index and mutual information. Insights into which of these measures could be used as a test-selection strategy could help improve the justifications of MAP explanations for users when the number of relevant variables is initially large.

To already relieve some of the computational burdens of computing the 'best' variable to select, two simplifying assumptions will be made in this study. First of all, a myopic approach to evidence gathering is taken, in which only one variable to gather evidence for is selected at a time (Ben-Bassat, 1978). Secondly, only MAP problems with one explanation variable for which the most probable value needs to be computed will be considered. While these assumptions might be an oversimplification of the problem-solving strategy in real-world applications, where experts can order multiple tests at a time (Sent, Van der Gaag, Witteman, Aleman, & Taal, 2005) and might be interested in multiple explanation variables at the same time (Van der Gaag, 1996), relaxing these assumptions poses some serious computational problems (Van der Gaag & Wessels, 1993). Therefore, we will not consider the problem when one or both of these assumptions are relaxed.

The rest of the paper is structured as follows. Section 2 will describe some of the preliminary background on Bayesian networks and MAP-independence, including notational conventions that will be used throughout the paper. In section 3 we will formally introduce the problem that we are addressing in the current paper, we will introduce the test-selection strategies that are explored and we will elaborate on how these test-selection strategies will be tested using simulations. In section 4, we will discuss and analyse the results of the simulations on the benchmarking ALARM network. In section

5, we will conclude the paper and discuss some possible directions for future research.

Chapter 2

Preliminaries

In this section, we will introduce some necessary preliminaries and notational conventions that will be used throughout the paper. We will use the same notational conventions as Kwisthout (2021) in his paper on MAP-independence.

A Bayesian network $\mathcal{B} = (\mathbf{G}_\mathcal{B}, \Pr)$ is a probabilistic graphical model representing a joint probability distribution $\Pr(\mathbf{V}) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$ over a set of discrete random variables $\mathbf{V}$. The network $\mathcal{B}$ includes a directed acyclic graph (DAG) $\mathbf{G}_\mathcal{B} = (\mathbf{V}, \mathbf{A})$, where $\mathbf{V}$ denotes the set of stochastic variables and $\mathbf{A}$ captures the conditional (in)dependencies among them, and a set of parameter probabilities Pr represented by conditional probability tables (CPTs). $\pi(V_i)$ signifies the set of parents of a variable V_i in $\mathbf{G}_\mathcal{B}$. Similar to Kwisthout, we will use upper case to denote individual variables from $\mathbf{V}$ and lower case to denote a specific value of a variable. We will use bold-faced upper case to denote a set of variables and bold-faced lower case to denote a joint value assignment to such a set. $\Omega(V_i)$ indicates the set of possible value assignments to variable V_i, with $\Omega(\mathbf{V_a})$ indicating the set of joint value assignments to the set $\mathbf{V_a}$ (Kwisthout, 2021).

One of the computational problems in Bayesian networks is the problem of finding the most probable explanation for a set of selected variables constituting the explanation set, given a set of observations. In other words, the problem is to find a joint value assignment with maximum posterior probability over the explanation set in the network, given the observations (Kwisthout, 2021). If the network also includes so-called intermediate variables which are neither part of the explanation set nor part of the observations, this problem is referred to as MAP. In this paper, we will use the following formal definition of MAP (Kwisthout, 2021, p. 5):

> MAP
> **Instance:** A Bayesian network $\mathcal{B} = (\mathbf{G}_\mathcal{B}, \Pr)$, where $\mathbf{V}(\mathbf{G}_\mathcal{B})$ is partitioned into a set of evidence nodes $\mathbf{E}$ with a joint value assignment $\mathbf{e}$, a set of intermediate nodes $\mathbf{I}$, and an explanation set

H.
Output: A joint value assignment $\mathbf{h}^*$ to $\mathbf{H}$ such that for all joint value assignments $\mathbf{h}'$ to $\mathbf{H}$, $\Pr(\mathbf{h}^* \mid \mathbf{e}) \geq \Pr(\mathbf{h}' \mid \mathbf{e})$.

A computational problem that can be of interest for justifying and explaining the most probable explanation is to identify the set of relevant intermediate variables that contribute to establishing the best explanation, given the available evidence. Although there may be different opinions on what relevance means for intermediate variables, in this paper we will consider an intermediate variable or set of intermediate variables relevant when the MAP explanation could change, had the variable or variables been observed. The problem of identifying relevant intermediate variables in this context can then be referred to as MAP-independence. When we want to decide MAP-independence for a given subset of intermediate variables $\mathbf{R} \subseteq \mathbf{I}$, we seek to answer the question whether $\mathbf{h}^*$ is MAP-independent from $\mathbf{R}$ given $\mathbf{e}$ (Kwisthout, 2021). This can be formally defined in the following way (Kwisthout, 2021, p. 9):

MAP-INDEPENDENCE
Instance: A Bayesian network $\mathcal{B} = (\mathbf{G}_\mathcal{B}, \Pr)$, where $\mathbf{V}(\mathbf{G})$ is partitioned into a set of evidence nodes $\mathbf{E}$ with a joint value assignment $\mathbf{e}$, a non-empty explanation set $\mathbf{H}$ with a joint value assignment $\mathbf{h}^*$, a non-empty set of nodes $\mathbf{R}$ for which we want to decide MAP-independence relative to $\mathbf{H}$, and a set of intermediate nodes $\mathbf{I}$.
Question: Is $\forall_{\mathbf{r} \in \Omega(\mathbf{R})} \text{argmax}_{\mathbf{H}} \Pr(\mathbf{H}, \mathbf{R} = \mathbf{r} \mid \mathbf{e}) = \mathbf{h}^*$?

The complement problem MAP-DEPENDENCE is defined similarly, where the *yes*- and *no*-answers are reversed. Whereas in the formal definition of MAP-INDEPENDENCE, MAP-independence is established for a set $\mathbf{R}$ in which the variables can interact with each other, a similar problem exists in which MAP-independence is determined for each singleton variable on its own in a set $\mathbf{R}$ (Kwisthout, 2021). This is referred to as WEAK MAP-INDEPENDENCE, which can be defined as follows (Kwisthout, 2021, p. 10):

WEAK MAP-INDEPENDENCE
Instance: As in MAP-INDEPENDENCE.
Question: Is $\forall_{R \in \mathbf{R}} \forall_{r \in \Omega(R)} \text{argmax}_{\mathbf{H}} \Pr(\mathbf{H}, R = r \mid \mathbf{e}) = \mathbf{h}^*$?

It can be observed that the computational complexity of establishing MAP-independence is much higher than the computational complexity of establishing weak MAP-independence. Namely, when establishing weak MAP-independence, at most $\mathcal{O}(c \mid \mathbf{R} \mid)$ assignments have to be tested, where $c = \max_{W \in V(G)} \Omega(W)$. In contrast, establishing strong MAP-independence requires $\mid \Omega(\mathbf{R}) \mid = \mathcal{O}(c^{|\mathbf{R}|})$ assignment tests (Kwisthout, 2021). Given

the unfavourable computational complexity of establishing strong MAP-independence compared to weak MAP-independence, we will in the current study consider only the latter computational problem.

Instead of determining weak MAP-independence for specific sets of intermediate variables, the problem that will be of interest in the current study is to determine the maximal set of weak MAP-independent variables from the explanation set, given the evidence. That is, instead of determining whether a given set of variables $\mathbf{R}$ is weakly MAP-independent from the best explanation, we want to determine the maximal set $\mathbf{R}$ for which every variable in the set is weakly MAP-independent from the best explanation. This problem, which we will refer to as Maximum Weak MAP-independence, can be formally defined in the following way:

> Maximum Weak MAP-independence
> **Instance:** A Bayesian network $\mathcal{B} = (\mathbf{G}_{\mathcal{B}}, \Pr)$, where $\mathbf{V}(\mathbf{G})$ is partitioned into a set of evidence nodes $\mathbf{E}$ with a joint value assignment $\mathbf{e}$, a non-empty explanation set $\mathbf{H}$ with a joint value assignment $\mathbf{h}^*$, and a set of intermediate nodes $\mathbf{I}$.
> **Output:** A subset $\mathbf{R} \subseteq \mathbf{I}$ such that $\forall_{R \in \mathbf{R}} \forall_{r \in \Omega(R)} \arg\max_{\mathbf{H}} \Pr(\mathbf{H}, R = r \mid \mathbf{e}) = \mathbf{h}^*$ and there is no other subset $\mathbf{R}' \subseteq \mathbf{I}$ such that $\mid \mathbf{R} \mid < \mid \mathbf{R}' \mid$ and $\forall_{R' \in \mathbf{R}'} \forall_{r' \in \Omega(R')} \text{argmax}_{\mathbf{H}} \Pr(\mathbf{H}, R' = r' \mid \mathbf{e}) = \mathbf{h}^*$.

Chapter 3

Method

In public domains that make use of decision support systems, explanations for the decisions are important for users to understand and accept the decisions. Therefore, a decision support system, and in particular the Bayesian network as the underlying statistical model of the system, should motivate the decisions that it makes. Whereas the concept of MAP-independence as introduced by Kwisthout (2021) may help to better motivate the decisions of a Bayesian network by identifying relevant intermediate variables for the decision, there could be scenarios in which there are too many relevant variables for a user to easily comprehend the decision. In order to better motivate the decision in these situations and reduce the number of relevant (MAP-dependent) intermediate variables in the network, additional evidence can be gathered.

To help decide what additional evidence to observe (i.e. the 'best' variable to observe), the concept of maximum weak MAP-independence as formally defined in the previous section can be used. However, deciding upon the best variable to observe is intractable in general (Van der Gaag & Bodlaender, 2011), which means that for larger networks this will be infeasible. Therefore, in this study it will be investigated which test-selection strategies could be used as a heuristic to increase the number of weak MAP-independent variables.

In this chapter, the first section will describe the manner in which we will use maximum weak MAP-independence to determine which variable is best to observe. In the second section, we will formally describe the problem that we are addressing. In the third section, the test-selection strategies that are used in this study will be explained. In the last section, we will elaborate on how these test-selection strategies are evaluated using simulations. Note that in the rest of the paper, for the sake of readability, we will use MAP-independence and weak MAP-independence interchangeably, which will both refer to weak MAP-independence.

3.1 Defining 'best' variable to observe

For the current study, we are interested in increasing the number of MAP-independent variables such that the user has more benefit from the concept of MAP-independence to help motivate the decisions of a Bayesian network. When gathering additional evidence, the main aim should thus be to gather evidence for a variable, whose observation results in a Bayesian network that has the largest number of MAP-independent variables according to maximum weak MAP-independence. However, we note that this variable to observe is always associated with a specific value for that variable. That is, if the variable to observe is a binary variable V, this variable will have the largest number of MAP-independent variables when it is observed with a specific value assignment v. The problem with this is that actually observing $V = v$, given the already available evidence $\mathbf{e}$, could have a rather low probability, i.e. $\Pr(v \mid \mathbf{e}) = 0.15$. In this case, there might be situations in which V is not considered the best variable to observe, and a trade-off between impact and prior probability of an observation needs to be made.

For example, imagine we would have another binary variable X with value assignment x, which is the second-best variable to observe. Observing $X = x$ results in a set of MAP-independent variables that contains one less variable than V with v (for example, observing $V = v$ results in a set of 40 MAP-independent variables and observing $X = x$ results in a set of 39 MAP-independent variables). However, actually observing $X = x$, given the available evidence, is much more likely than observing $V = v$, i.e. $\Pr(x \mid \mathbf{e}) = 0.95$. Furthermore, imagine that observing $V = \neg v$ and observing $X = \neg x$ would both result in a set of eight MAP-independent variables, so observing those values for V and X would actually be much worse than observing $V = v$ and $X = x$. Also note that observing $V = \neg v$ would be much worse than observing $X = x$. In this case, we might prefer to observe variable X rather than variable V, since the probability of observing $X = x$ is much higher than observing $V = v$ and the numbers of MAP-independent variables are almost equal.

It thus seems that the probability of actually observing the value for the variable which results in the largest number of MAP-independent variables is also important. Yet, this importance might change when the difference between the numbers of MAP-independent variables changes. For instance, imagine the same setting as before. Only this time around, the set of MAP-independent variables when observing $X = x$ has a size of 12 instead of 39, which is much smaller than the number of MAP-independent variables when observing $V = v$. In this case, we may still prefer to observe V, even considering the low probability of actually observing v. On the other hand, observing variable X with value x would result in a considerably lower number of MAP-independent variables and it is not even certain that value x for variable X is actually observed. Furthermore, observing $V = \neg v$ would

not be much worse than observing $X = x$. Therefore, observing variable X may still be too unattractive compared to observing variable V.

In this example, we can clearly see the multi-objective criterion for the best variable. There is a clear trade-off between the number of MAP-independent variables when observing a variable with a particular value and the probability of actually observing that value for the variable. Generally, we want to select a variable V that has a high probability of taking on value v and, when taken on value v, results in a Bayesian network with the largest number of MAP-independent variables. However, it may not always be possible to satisfy both objectives. Hence, we have to find a formal means to decide between two possible variables.

In this study, we will compute the *expected number* of MAP-independent variables to decide between two or more variables. That is, for a variable V in combination with values v_i, the resulting numbers of MAP-independent variables are weighted by the probability of actually observing the values v_i for variable V, given the already available evidence. The variable to observe will be the variable with the largest expected number of MAP-independent variables, in the rest of the paper referred to as the best variable to observe. For the example above, this means that we can compute the expected number of MAP-independent variables for variable V as follows: $40 \cdot 0.15 + 8 \cdot 0.85 = 12.8$. Similarly, for variable X in the first example, the expected number of MAP-independent variables will be $39 \cdot 0.95 + 8 \cdot 0.05 = 37.45$. As can be seen, variable X is clearly preferred over variable V in this case. For the second example, the expected number of MAP-independent variables for X will be $12 \cdot 0.95 + 8 \cdot 0.05 = 11.8$, which means that X is not preferred over variable V. It can be seen that both cases match our intuition on what would be the best variable to observe.

3.2 Formal problem definition

Now that we have the formal means to decide which variable would be best to observe in a certain situation, the problem which we set out to solve can be formally defined. The computational problem of interest is to decide upon the next variable to observe, such that the number of weak MAP-independent intermediate variables with respect to the explanation set $\mathbf{H}$ is expected to be the largest in the resulting Bayesian network. In other words, we want to know which variable $I \in \mathbf{I}$, when observed, is expected to lead to the largest number of weak MAP-independent variables $\mathbf{R}$ with respect to $\mathbf{H}$ in the resulting Bayesian network. This problem can be formalized as below:

MAP-SELECTION
Instance: A Bayesian network $\mathcal{B} = (\mathbf{G}_\mathcal{B}, \Pr)$, where $\mathbf{V}(\mathbf{G}_\mathcal{B})$ is partitioned into a set of evidence variables $\mathbf{E}$ with a joint value

assignment $\mathbf{e}$, a non-empty explanation set $\mathbf{H}$ with a joint value assignment $\mathbf{h}^*$, and a set of intermediate variables $\mathbf{I}$. Furthermore, the maximum set of weak MAP-independent intermediate variables $\mathbf{I^P} \subseteq \mathbf{I}$ according to MAXIMUM WEAK MAP-INDEPENDENCE($\mathcal{B}, \mathbf{E}, \mathbf{e}, \mathbf{H}, \mathbf{h}^*, \mathbf{I}$).
Output: A variable $I \in \mathbf{I} \setminus \mathbf{I^P}$ such that $exp\_size(I) = \sum_{i \in \Omega(I)} \Pr(i \mid \mathbf{e}) \cdot \mid \mathbf{R} \mid$ is maximized, where $\mathbf{R}$ = MAXIMUM WEAK MAP-INDEPENDENCE($\mathcal{B}, \mathbf{E}_{new}, \mathbf{e}_{new}, \mathbf{H}_{new}, \mathbf{h}^*_{new}, \mathbf{I}_{new}$) and $\mathbf{E}_{new} = \mathbf{E} \cup \{I\}$, $\mathbf{e}_{new} = \mathbf{e} \cup \{i\}$, $\mathbf{H}_{new} = \mathbf{H}$, $\mathbf{h}^*_{new} = \mathbf{h}^*$ and $\mathbf{I}_{new} = \mathbf{I} \setminus \{I\}$. That is, there is no variable $I' \in \mathbf{I} \setminus \mathbf{I^P}$ such that $exp\_size(I') > exp\_size(I)$.

In this study, we will make the simplifying assumption that the explanation set consists of only one explanation variable H with value assignment h^*, which poses fewer computational difficulties.

3.3 Test-selection strategies

In order to address the MAP-SELECTION problem, we have to find a test-selection strategy that can be used as a heuristic to determine which variable, when observed, is expected to generate the largest number of MAP-independent variables in the resulting Bayesian network. That is, the test-selection strategy should help indicate the usefulness of observing a variable, thereby providing a measure to select the variable to observe in each stage of the diagnostic process. In this section, we will provide an overview of the test-selection strategies that we will explore in the current study. All strategies will take a myopic approach to evidence gathering in order to relieve some of the computational burdens.

3.3.1 In-degree, out-degree and total degree

In the qualitative part of a Bayesian network, which is a graphical representation that is encoded in a DAG and models (in)dependencies between the variables in the Bayesian network (Van der Gaag & Wessels, 1993), variables have a number of incoming and outgoing arcs. The incoming arcs from the parents of a variable, generally referred to as the in-degree of the variable, represent the dependencies on other variables in the network while the outgoing arcs to the children of a variable, referred to as the out-degree of a variable, represent the dependencies of other variables on the current variable. It could be argued that variables with many dependencies (either in-degree, out-degree, or both, which is the total degree of a variable) are important variables in the network, as these may directly impact many other variables when observed. Observing one of these 'important' variables could lead to many other intermediate variables becoming MAP-independent, as

the other intermediate variables might be impacted in such a way that the best explanation does not change anymore upon observing any value for these variables. Therefore, the in-degree, out-degree and total degree of the intermediate variables might provide good test-selection strategies, which values should be maximized.

3.3.2 Distance to the explanation variable

With a second test-selection strategy, also focusing on the qualitative part of the Bayesian network, it will be explored whether or not the distance between two variables in the graphical representation of the network could provide a good test-selection strategy for selecting a variable to observe. Here, the distance between two variables is defined as the length of the shortest path between the variables, not taking into account the direction of the arcs. That is, if there is an arc directly connecting two variables, the distance between the variables is one. If there is no such arc, but the variables both have an arc (either incoming or outgoing) connecting to the same other variable, the distance between the variables is two, and so on. The intuition behind using this measure as a test-selection strategy is that an intermediate variable is more likely to contain much information about the explanation variable and block the influence of other intermediate variables on the explanation variable when the distance between the intermediate variable and explanation variable is small. This means that the other intermediate variables are less likely to impact the explanation variable, making them more likely to become MAP-independent from the explanation variable.

3.3.3 Expected utility

In addition to test-selection strategies focusing on the qualitative part of the Bayesian network, strategies focusing on the quantitative part of the network will be explored. This part of the Bayesian network encodes the strengths of the dependencies between variables with a set of conditional probabilities (Van der Gaag & Wessels, 1993). One of the test-selection strategies that is explored in the current study is the expected utility, based on the *linear-value utility function*, which is defined by Van der Gaag and Wessels (1993). In their study, Van der Gaag and Wessels introduced the expected utility and linear-value utility function to make the best explanation more robust and stable (Van der Gaag & Wessels, 1993). While this aim is different from the aim in the current study, the expected utility might still be usable as a test-selection strategy to address the MAP-SELECTION problem. After all, using a test-selection strategy that makes the best explanation more robust and stable also means that the best explanation should remain robust and stable with additional evidence for intermediate variables. Hence, it could be that the expected number of MAP-independent variables implicitly increases

when using this test-selection strategy.

Suppose we want to know the posterior probability over the explanation variable H, given the evidence $\mathbf{e}$ and suppose that h is the value assignment of H. Now, for an unobserved variable I with observed value $i \in \Omega(I)$, the linear-value utility function can be defined as

$$u(i) = \mid \Pr(h \mid \mathbf{e}) - \Pr(h \mid \mathbf{e} \cup \{i\}) \mid \tag{3.1}$$

Subsequently, the expected utility can be computed using the linear-value utility function. Here, the utilities for all possible value assignments of variable I are weighted by the probability of observing that value assignment given the evidence. Whereas this utility function is tailored to binary variables (Van der Gaag & Wessels, 1993), this can be extended to also incorporate non-binary variables, resulting in the following definition:

$$\hat{u}(I) = \sum_{i \in \Omega(I)} \Pr(i \mid \mathbf{e}) \cdot u(i) \tag{3.2}$$

The best intermediate variable to select will be the variable that is expected to cause the largest difference in posterior probability of the explanation variable for when the variable is observed versus not observed. That is, the best intermediate variable to select will be the variable that maximizes $\hat{u}(I)$.

3.3.4 Expected Gini index

A second test-selection strategy focusing on the quantitative part of the Bayesian network is the expected *Gini index*. The expected Gini index can be used to represent the level of uncertainty in the explanation variable and will be applied to see whether observing an intermediate variable can reduce the level of uncertainty in the explanation variable. Similar to the expected utility, the expected Gini index may be a good test-selection strategy to increase the number of MAP-independent variables, because the best explanation should become more stable and hence less likely to be influenced by other variables when the level of uncertainty in the explanation variable decreases. To compute the expected Gini index, first a definition of the Gini index is needed. The Gini index over the probability distribution of a variable H is defined as

$$G(\Pr(H)) = 1 - \sum_{h_i \in \Omega(H)} \Pr(h_i)^2 \tag{3.3}$$

The expected Gini index can be defined for observing a variable I, where all possible probability distributions over the explanation variable H with

$I = i$ as additional evidence are weighted by the probability of observing $I = i$, given the evidence $\mathbf{e}$ (Sent & Van der Gaag, 2007):

$$G(\Pr(H \mid I)) = \sum_{i \in \Omega(I)} \Pr(i \mid \mathbf{e}) \cdot G(\Pr(H \mid \mathbf{e} \cup \{i\})) \tag{3.4}$$

The best intermediate variable to observe will be the variable that is expected to result in the largest decrease of uncertainty of the explanation variable. That is, the variable that maximizes $\tilde{G}(I) = G(\Pr(H)) - G(\Pr(H \mid I))$ for an intermediate variable I (Sent & Van der Gaag, 2007).

3.3.5 Mutual information

The last test-selection strategy we will focus on in the current study is *mutual information*. The mutual information between two variables describes the amount of information that the first variable contains about the second variable and can be used to see how much uncertainty can be reduced in the second variable by observing the first one. As such, this measure can be used to see how much uncertainty can be reduced in the explanation variable by observing an intermediate variable. Following the same reasoning as with the linear-value utility function and expected Gini index, mutual information could provide a good test-selection strategy in this way. Mutual information has a relationship with entropy and conditional entropy. The entropy of a variable H can be defined as

$$H(H) = -\sum_{h_i \in \Omega(H)} \Pr(h_i) \log_2(\Pr(h_i)) \tag{3.5}$$

The conditional entropy between a variable H and I, given the already available evidence $\mathbf{e}$, can be defined as

$$H(H \mid I) = \sum_{i \in \Omega(I)} \Pr(i \mid \mathbf{e}) \cdot H(H \mid \mathbf{e} \cup \{i\}) \tag{3.6}$$

Based on the entropy and conditional entropy, the mutual information between an intermediate variable I and explanation variable H can be defined as

$$I(I; H) = H(H) - H(H \mid I) \tag{3.7}$$

The best intermediate variable to observe will be the variable that contains most information about the explanation variable and can thus reduce most uncertainty in the explanation variable, thereby making the explanation variable more robust. That is, the variable that maximizes $I(I; H)$.

3.4 Simulations

We test the useability of the different test-selection strategies using simulations on the benchmarking ALARM Bayesian network (Beinlich, Suermondt, Chavez, & Cooper, 1989). This network simulates an anesthesia monitoring system, in which text messages about diagnoses are provided for possible problems. The network consists of 37 variables, eight of which constitute the set of possible explanation variables. Sixteen variables can be observed or measured (evidence variables). The other variables are intermediate variables that can neither be observed nor belong to the set of variables to be explained (Beinlich et al., 1989).

In every simulation run, the ALARM network is first partitioned into a set of randomly selected evidence variables $\mathbf{E}$ with a random joint value assignment $\mathbf{e}$, an explanation set consisting of one randomly selected explanation variable H with value assignment h^* and a set of intermediate variables $\mathbf{I}$, which consists of all other variables in the network. The number of evidence variables in the evidence set is varied between zero and four in the simulation runs because for larger evidence sets, there is a higher probability that the number of MAP-independent variables is already large at the start. This makes finding simulation runs in which the test-selection strategies can (significantly) increase the number of MAP-independent variables much harder. Note that in the simulation runs, we do not care about the original partitioning of the network into the evidence, explanation and intermediate variables. Preliminary results suggest that the data to analyse would be fairly limited when taking this into account. Therefore, we make use of the full network to generate queries.

For the partitioned Bayesian network, the maximum subset of MAP-independent variables $\mathbf{R} \subseteq \mathbf{I}$ is computed according to Maximum Weak MAP-independence (Algorithm 1). Furthermore, the quantitative test-selection strategies expected utility, expected Gini index and mutual information are computed for the intermediate variables, based on the formulas described in the previous section. For efficiency reasons, this is done together with computing Maximum Weak MAP-independence[1]. The qualitative test-selection strategies do not have to be computed, as they can be directly read off from the graphical representation of the network.

[1]Both the computation of maximum weak MAP-independence as well as the computations for the test-selection strategies need the posterior probabilities $\Pr(H \mid \mathbf{e})$ and $\Pr(H \mid \mathbf{e} \cup \{i\})$ for all $i \in \Omega(I)$, where H is the hypothesis variable, $\mathbf{e}$ is the available evidence and variable I is an unobserved intermediate variable that is not MAP-independent. When computing maximum MAP-independence and the test-selection strategy values together, these values only need to be computed once, which saves execution time.

Algorithm 1 MAXIMUM WEAK MAP-INDEPENDENCE

Input: Bayesian network $\mathcal{B}$ partitioned in $\mathbf{E} = \mathbf{e}$, $H = h^*$ and $\mathbf{I}$.
Output: A set $\mathbf{R} \subseteq \mathbf{I}$ which contains all intermediate variables that are weakly MAP-independent from H given $\mathbf{e}$.

```
R = new set
for all I ∈ I do
    independent = true
    for all i ∈ Ω(I) do
        if argmax_H Pr(H, I = i | e) ≠ h* then
            independent = false
        end if
    end for
    if independent then
        R.add(I)
    end if
end for
return R
```

Subsequently, iteratively one of the intermediate variables $I \in \mathbf{I}$ is added to the evidence set $\mathbf{E}$ with a particular value assignment i and removed from the set of intermediate variables, provided that this variable is not MAP-independent (not in set $\mathbf{R}$). For the resulting Bayesian network, again the maximum set of MAP-independent variables is computed according to MAXIMUM WEAK MAP-INDEPENDENCE. Furthermore, the probability of observing value i, given the already available evidence $\mathbf{e}$ is computed ($\Pr(i \mid \mathbf{e})$). This probability is then multiplied by the size of the maximum set of MAP-independent variables. When doing this for all values $i \in \Omega(I)$ and adding the results together, the expected number of MAP-independent variables when observing variable I can be computed. After computing the expected number of MAP-independent variables for every variable that is not MAP-independent, the best variable(s) to observe can be identified. The entire algorithm to determine the best variables to observe is given in the pseudo-code below (Algorithm 2).

Algorithm 2 MAP-SELECTION

Input: Bayesian network $\mathcal{B}$ partitioned in $\mathbf{E} = \mathbf{e}$, $H = h^*$ and $\mathbf{I}$; a set $\mathbf{R}$, consisting of all weak MAP-independent variables in the network;

Output: A set of variables $\mathbf{V} \subseteq \mathbf{R}$ which have the largest expected number of weak MAP-independent variables when observed, or alternatively an empty set if no variable can increase set $\mathbf{R}$.

$bestVariables$ = new set
$maximalSize =\mid \mathbf{R} \mid$
for all $R \in \mathbf{R}$ **do**
  $expectedSize = 0$
  for all $r \in \Omega(R)$ **do**
    $\mathbf{E}_{new} = \mathbf{E} \cup \{R\}$
    $\mathbf{e}_{new} = \mathbf{e} \cup \{r\}$
    $H_{new} = H$
    $h^*_{new} = h^*$
    $\mathbf{I}_{new} = \mathbf{I} \setminus \{R\}$
    $\mathbf{R}' =$MAXIMUM WEAK MAP-INDEPENDENCE($\mathcal{B}$, $\mathbf{E}_{new}$, $\mathbf{e}_{new}$, H_{new}, h^*_{new}, $\mathbf{I}_{new}$)
    $expectedSize = expectedSize + (\Pr(r \mid \mathbf{e}) \cdot \mid \mathbf{R}' \mid)$
  end for
  if $expectedSize > maximalSize$ **then**
    $maximalSize = expectedSize$
    $bestVariables =$new set(R)
  else if $expectedSize == maximalSize$ **then**
    $bestVariables$.add(R)
  end if
end for
return $bestVariables$

Chapter 4

Results

We analysed the usability of the test-selection strategy using 164 simulations on the ALARM network, varying the size of the evidence set. Figure 4.1 shows the distribution of the number of MAP-dependent variables over the simulation runs. The number of MAP-dependent variables in a given simulation run indicates how many variables can potentially become (also) MAP-independent in that simulation run. As can be seen, the number of relevant MAP-dependent variables is fairly limited (between zero and six) in most simulation runs, showing that the number of MAP-independent variables is already quite high in those runs. However, we also see that some simulation runs contain many MAP-dependent variables, with as most prominent example the simulation run in which there are initially 26 MAP-dependent variables which would change the best explanation of the hypothesis variable. Especially in those scenarios, using a test-selection strategy to select the best variable to observe could potentially help to significantly decrease the number of MAP-dependent variables, thereby increasing the number of MAP-independent variables.

Table 4.1 furthermore confirms our beliefs that with a larger evidence set, the number of MAP-independent variables is more often already large at the start of a simulation. Namely, the table summarizes the mean number of MAP-dependent variables in a simulation run for the different sizes of the set evidence variables. In general, it can be seen that the mean number of MAP-dependent variables in simulation runs tends to decrease when increasing the size of the evidence set, which means that the number of MAP-independent variables tends to become larger when more variables are observed. This seems to justify the decision to keep the number of relevant variables limited in the current study.

To assess how well the test-selection strategies as described in the previous section can be used to solve the MAP-SELECTION problem and increase the number of MAP-independent variables, we used two different measures: (1) rank-approximation of the best variable, indicating whether or not a

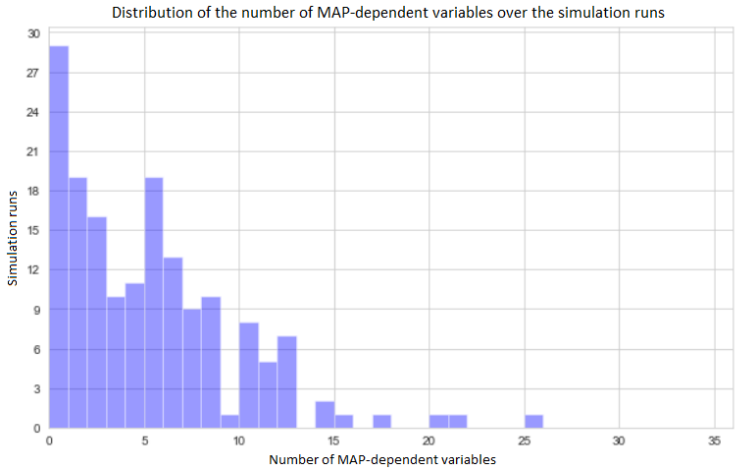

Figure 4.1
The distribution of the number of relevant MAP-dependent variables over the simulation runs, indicating how many variables can potentially become (also) MAP-independent. It can be seen the number of MAP-dependent variables is rather small (between zero and six) in most simulations. However, the distribution has a long right tail, which indicates that there are also scenarios in which the number of MAP-independent variables can be significantly increased.

Table 4.1
Overview of the mean number of MAP-dependent variables over the simulation runs for different sizes of the evidence set.

Size of evidence set	**Nr of sim runs**	**Mean nr of MAP-dependent vars**
0	19	5.684 (min:0, max:26)
1	47	6.468 (min:0, max:21)
2	51	4.039 (min:0, max:12)
3	24	4.542 (min:0, max:14)
4	23	3.000 (min:0, max:10)

test-selection strategy can systematically identify which variable in a simulation run is expected to yield the largest number of MAP-independent variables, (2) value-approximation of the expected number of MAP-independent variables, indicating whether or not the value of a test-selection strategy correlates with the expected number of MAP-independent variables. The intuition behind these measures will be explicated in the respective subsections.

For the analyses, simulation runs in which all variables or all variables except one are MAP-independent will not be considered, which rules out 48 simulation runs. Namely, when all variables are MAP-independent, there will be no point in observing an additional variable, because the best explanation of the explanation variable would not change. Similarly, when all variables are MAP-independent except for one, it is obvious which variable to observe next, because only one variable changes the best explanation of the explanation variable. Therefore, these simulations do not provide us with any insight into which test-selection strategy might serve as a good test-selection strategy. In the following sections, the results for the two analyses will be discussed.

4.1 Rank-approximation

In the first method of evaluation, it is assessed how well a test-selection strategy can indicate which variable in the set of MAP-dependent variables is the best variable to observe. It can be investigated in how many simulation runs the best variable to observe is included in the top n variables according to the test-selection strategies. In each simulation run, the variable with the largest expected number of MAP-independent variables is identified as the best variable to observe in that simulation run. The n best variables to observe according to the test-selection strategies are identified based on maximizing or minimizing the test-selection strategy value. For the distance to the explanation variable, the n variables with the smallest distance are considered the best variables, because the distance to the explanation variable should be minimized. For all other test-selection strategies, the n variables with the highest value are considered the best variables, as those test-selection strategy values should be maximized. When variables are not included in the top n but do have the same value for the test-selection strategy as a variable that is included in the top n, we decided to add this variable to the top n variables too, even though that meant exceeding n.

We obtain the theoretical chance level for the number of simulation runs in which the best variable to observe is included in the top n variables in the following way: we calculate the probability that the best variable to observe is included in n randomly selected variables from a set of k MAP-dependent variables. Subsequently, we multiply this probability by the number of sim-

Table 4.2
Overview of the number and percentage of simulation runs in which the best variable to observe is included in the top one, top two and top three variables respectively for each test-selection strategy, based on minimizing (distance) or maximizing (other strategies) the test-selection strategy values. The best variable to observe is defined as the variable with the largest expected number of MAP-independent variables.

Test-selection strategy	**Nr of sims in top 1**	**Nr of sims in top 2**	**Nr of sims in top 3**	**% of sims in top 1**	**% of sims in top 2**	**% of sims in top 3**
In-degree	34	66	90	29.310	56.897	77.586
Out-degree	41	79	95	35.345	68.103	81.897
Total degree	50	77	96	43.103	66.379	82.759
Distance	103	105	110	88.793	90.517	94.828
Expected utility	70	105	108	60.345	90.517	93.103
Expected Gini index	75	103	110	64.655	88.793	94.828
Mutual information	69	96	104	59.483	82.759	89.655

ulation runs with k MAP-dependent variables. Calculating this for all k and adding the results together then gives the number of simulation runs in which the best variable to observe is included in a set of n randomly selected variables. As such, we can see that the theoretical chance levels are 24.94 (21.5%), 49.88 (43.0%) and 66.82 (57.6%) simulation runs for the top one, top two and top three variables respectively.

The results of the rank-approximation for the test-selection strategies are summarized in Table 4.2 and Figure 4.2. The bar plots in Figure 4.2 show the percentage of simulations in which the best variable to observe is included in the top n variables according to the different test-selection strategies, where n is varied between one (indicating that the best variable to observe is also the best variable to observe according to the test-selection strategy) and three (indicating that the best variable to observe is in the top three best variables to observe according to the test-selection strategy).

In general, it can be seen that all test-selection strategies perform better than chance for different values of n. The distance between an intermediate variable and the explanation variable seems to be the best indicator of which variable to observe, where the test-selection strategy could correctly identify the best variable to observe in 88.8% of the simulation runs. However, this difference decreases when looking at the top two and top three best variables according to the test-selection strategies, where the expected utility, expected Gini index and mutual information give similar results. Results suggest that the in-degree, out-degree and total degree are least good at indicating which variable to observe next, where the percentages are structurally lower than for the other test-selection strategies.

Table 4.3 shows the number of simulation runs in which there is more than one best variable to observe according to a test-selection strategy. Test-selection strategies should aim to keep this number as low as possible, because having multiple best variables means that a choice still has to be made about which variable to select out of all variables with the best value ac-

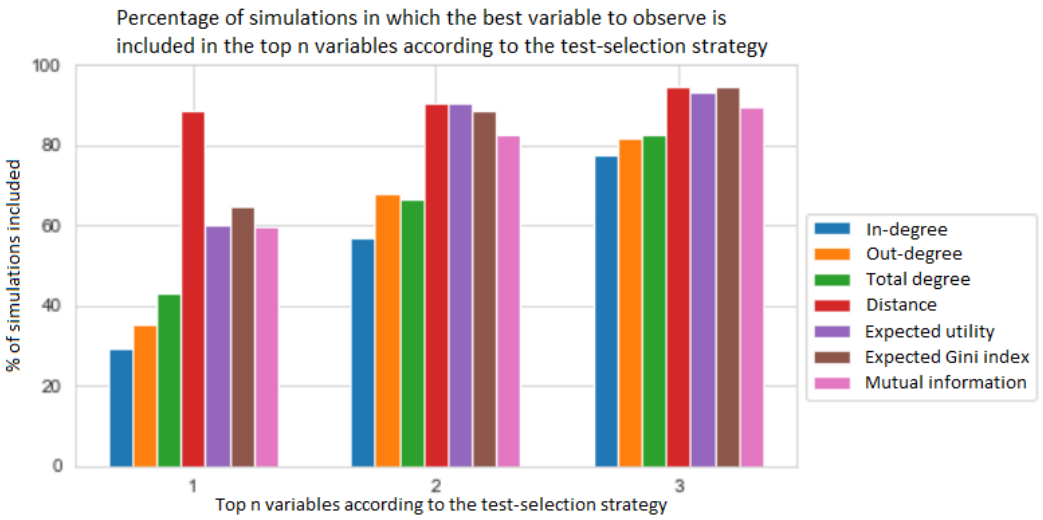

Figure 4.2
Bar plots indicating for each test-selection strategy the percentage of simulations in which the best variable to observe is included in the top one, top two and top three variables according to the test-selection strategy. The top one, top two and top three variables according to the test-selection strategies are selected based on minimizing (distance) or maximizing (other strategies) the test-selection strategy values. It can be seen that the in-degree, out-degree and total degree are performing structurally worse than the other test-selection strategies. The distance to the explanation variable seems to be the best indicator of the best variable to observe, although this is only the case when the best variable to observe is the top one variable according to the test-selection strategy.

Table 4.3
Overview of the number and percentage of simulations in which there are multiple best variables to observe according to the test-selection strategies, based on minimizing (distance) or maximizing (other strategies) the test-selection strategy values.

Test-selection strategy	**Nr of sims with multiple best vars**	**%sims with multiple best vars**
In-degree	34	0.293
Out-degree	11	0.095
Total degree	33	0.284
Distance	92	0.793
Expected utility	4	0.034
Expected Gini index	4	0.034
Mutual information	4	0.034

cording to the test-selection strategy. This is generally undesirable. As can be seen in the table, especially the qualitative test-selection strategies (in which the distance to the explanation variable stands out most) have many simulation runs in which there is no single best variable to observe. On the other hand, the expected utility, expected Gini index and mutual information seem to perform better in this regard, with only four simulation runs in which there are multiple variables with the same highest test-selection strategy value.

As the analysis above is carried out with simulation data for which there is more than one MAP-dependent variable in a simulation run, this data also contains simulation runs in which there are two or three MAP-dependent variables. This might influence the results summarized in Table 4.2 and Figure 4.2 since the best variable to observe is always included in the top three variables for those simulation runs. Hence, to see whether this has any effect on the results of the study, a second analysis is carried out in which all these simulation runs are removed from the data. Furthermore, simulation runs with less than seven MAP-dependent variables are removed from the data, such that the probability of the best variable being included in a randomly selected set of three variables is always smaller than 0.5. Using this additional analysis, something can be said about whether or not the usefulness of test-selection strategies depends on the number of MAP-dependent variables in a simulation run.

While the full analysis that is carried out on this data is added to Appendix A.1, we will here give a short summary of the findings. To start, all test-selection strategies are still performing better than chance, although the percentage of simulations in which the best variable to observe is included in the top n variables according to the test-selection strategies decreased compared to the analysis laid out above. For the in-degree, out-degree and total degree, this difference is quite substantial. For the other test-selection strategies, the difference is only minor. The general trend of the data is still the same, which means that the in-degree, out-degree and total degree still perform structurally worse than the other test-selection strategies and

distance is still the best indicator of the best variable to observe.

Regarding the number of simulations in which there are multiple best variables to observe according to the test-selection strategies, the relative number of simulation runs in which this occurs for the in-degree, total degree and distance has increased, which means that there are still many simulations in which there is no single best variable to observe according to these strategies. For the expected utility, expected Gini index and mutual information, there are no simulation runs anymore in which there are multiple best variables, which means that these selection strategies are able to select a single best variable to observe in every simulation run with more than six MAP-dependent variables.

4.2 Value-approximation

While the rank-approximation analysis as done in Section 4.1 focuses exclusively on whether or not a test-selection strategy is able to identify the best variable to observe in a simulation run, it can also be tested whether or not the values of the test-selection strategies correlate with the expected number of MAP-independent variables. After all, the values for the test-selection strategies need to be maximized or minimized to select the variable which is expected to have the largest number of MAP-independent variables. Therefore, there could be a correlation between the value of a test-selection strategy and the expected number of MAP-independent variables. Take for example the expected Gini index. Earlier, we argued that variables with a higher expected Gini index are better able to decrease the uncertainty in the explanation variable and because of this, more variables should become MAP-independent. Using value-approximation, we can see whether or not this is actually the case and whether indeed a variable with a higher expected Gini index is associated with a larger expected number of MAP-independent variables. For the other test-selection strategies, a similar argument can be made.

Testing whether or not the values of the test-selection strategies correlate with the expected number of MAP-independent variables can serve as an additional check to see whether the test-selection strategies are consistently evaluated across methods. Furthermore, when a test-selection strategy is able to value-approximate the expected number of weak MAP-independent variables well, this may also provide a useful indication about whether or not it is even preferable to observe another variable. Namely, when the value of a test-selection strategy for the best variable is associated with an expected number of MAP-independent variables that is lower than the current number of MAP-independent variables, it might not be beneficial to observe an additional variable, even though this is the best variable to observe according to the test-selection strategy.

In order to understand how well the different test-selection strategies can value-approximate the expected number of MAP-independent variables, we look for the Pearson correlation between the values of a test-selection strategy and the expected number of MAP-dependent variables. Note that MAP-dependence is the complement of MAP-independence and therefore, this will give us the same information. However, using MAP-dependence instead of MAP-independence does not give the problem that the maximum expected number of MAP-independent variables can differ between simulation runs. After all, when the evidence set consists of zero evidence variables, there can be at most 36 MAP-independent variables in the ALARM network. On the other hand, when the evidence set consists of four evidence variables, there can be at most 32 MAP-independent variables[1].

Figure 4.3 shows the scatter plots in which the values for the different test-selection strategies are plotted against the expected number of MAP-dependent variables. Whereas no immediate relationship can be seen between the in-degree, out-degree or total degree and the expected number of MAP-dependent variables, a positive relationship can be identified between the distance of an intermediate variable to the hypothesis variable and the expected number of MAP-dependent variables, where a higher distance corresponds to a higher expected number of MAP-dependent variables. For the expected utility, expected Gini index and mutual information, negative relationships can be identified, although these relationships do not seem to be completely linear.

[1]Recall that the ALARM network consists of 37 variables in total. One variable constitutes the explanation set and four variables constitute the evidence set. This leaves 32 intermediate variables that can be MAP-independent.

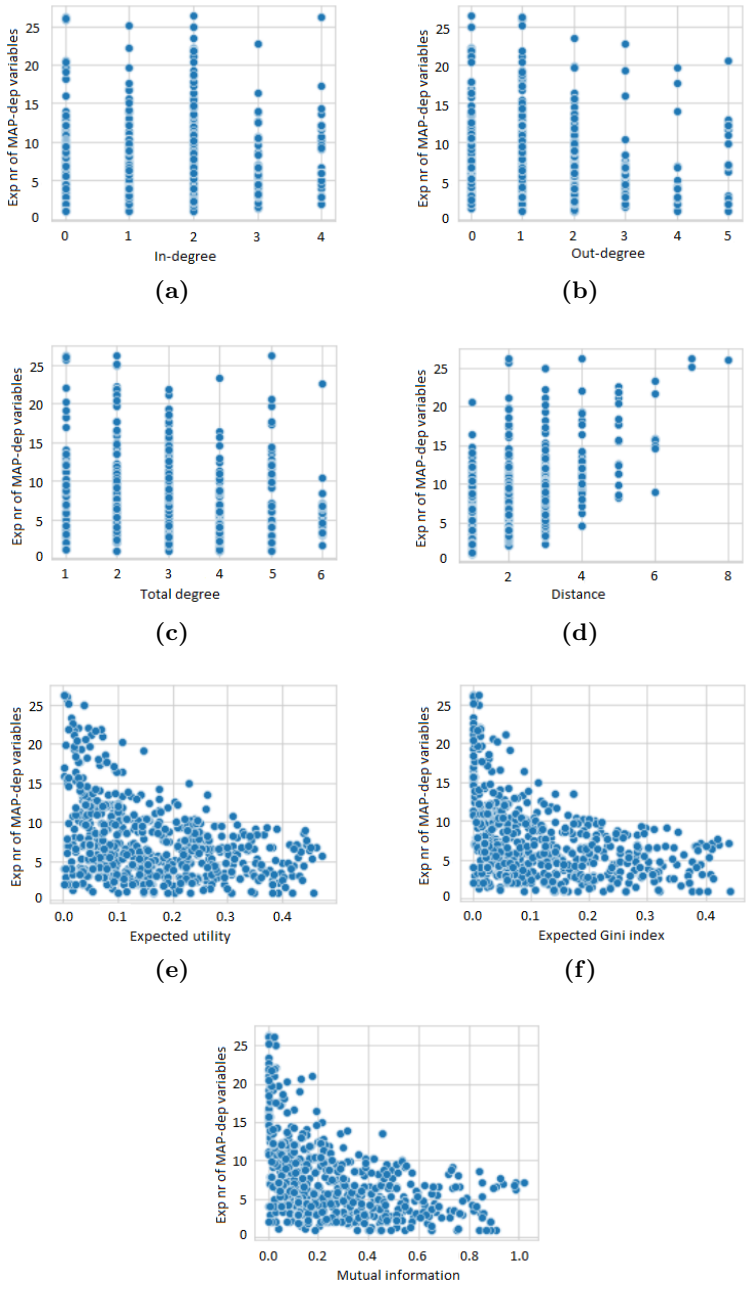

Figure 4.3
Scatter plots where the (a) in-degree, (b) out-degree, (c) total degree, (d) distance to the explanation variable, (e) expected utility, (f) expected Gini index and (g) mutual information are plotted on the x-axis against the expected number of MAP-dependent variables on the y-axis. In (a), (b) and (c), no clear relationship can be established. For (d), there seems to be a positive relationship between the variables. For (e), (f) and (g), the relationship seems to be negative.

Table 4.4
Overview of the correlations of the test-selection strategy values with the expected number of MAP-dependent variables and the fitted regression models.

Test-selection strategy	Transformation	Pearson's r	R-squared	Adj. R-squared	Intercept	Coef
In-degree	None	-0.038	0.001	0.000	7.492	-0.171
Out-degree**	None	-0.116	0.014	0.012	7.820	-0.448
Total degree**	None	-0.134	0.018	0.017	8.579	-0.463
Distance**	None	0.593	0.352	0.351	1.997	2.537
Expected utility**	Log	-0.435	0.189	0.188	2.580	-2.148
Expected Gini index**	Log	-0.566	0.321	0.320	1.962	-1.927
Mutual information**	Log	-0.592	0.353	0.352	3.471	-2.002

** Indicates significance at the $p < 0.01$ level.

To assess the strengths of the correlations between the values of the test-selection strategies (independent variable) and the expected number of MAP-dependent variables (dependent variable), linear regression can be used. We have to note that linear regression might be a sub-optimal model to use in the current context, as the assumption of normally distributed residuals and equal variance are violated by some test-selection strategies. However, using linear regression can still provide us with useful information, although results need to be interpreted with care. For the analysis, the expected utility, expected Gini index and mutual information have been log-transformed to improve linearity with the expected number of MAP-dependent variables.

Results of the regression models are shown in Table 4.4. It can be seen that the in-degree of a variable has no statistically significant correlation with the expected number of MAP-dependent variables at the $p < 0.01$ level (also not at the $p < 0.05$ level). For all other test-selection strategies, a statistically significant correlation is found at the $p < 0.01$ level. When looking at Pearson's r, it can be seen that both out-degree and total degree have a very weak negative correlation with the expected number of MAP-dependent variables. On the other hand, there seems to be a moderately strong positive correlation between the distance and the expected number of MAP-dependent variables, where a larger distance between the intermediate and explanation variable is associated with a larger expected number of MAP-dependent variables (and therefore a smaller expected number of MAP-independent variables). For the expected utility, expected Gini index and mutual information, there seems to be a moderately strong negative correlation with the expected number of MAP-dependent variables.

Similar to the rank-approximation, it would be interesting to see whether the usefulness of the test-selection strategies depends on the number of MAP-dependent variables in a simulation run. Therefore, again a second analysis is carried out in which only simulation runs with more than six MAP-dependent variables are considered to test whether or not the correlation with the expected number of MAP-dependent variables changes.

The full analysis is added to the Appendix A.2. However, we will summarize the results here. It can be seen that all test-selection strategies now show a statistically significant correlation with the expected number of MAP-dependent variables at the $p < 0.05$ level, where all strategies apart from in-degree also show significance at the $p < 0.01$ level. Furthermore, the correlations seem to have increased slightly compared to the analysis laid out above, except for the distance, which has a correlation that has slightly decreased. This suggests that in general, the number of MAP-dependent variables in a simulation run has almost no negative influence on the correlations of the test-selection strategy values with the expected number of MAP-(in)dependent variables. In fact, the correlations tend to increase rather than decrease when there are more MAP-dependent variables in a simulation run, although the strengths of the correlations have not significantly improved.

Chapter 5

Discussion

In this paper, the aim was to investigate which test-selection strategies could serve as a good heuristic to increase the number of MAP-independent variables. We specifically focused on weak MAP-independence, as strong MAP-independence is highly intractable and as such poses computational difficulties (Kwisthout, 2021). We defined the best variable to observe as the variable which has the largest expected number of MAP-independent variables when observed. Using this definition of the best variable to observe, we formally introduced MAP-SELECTION as the computational problem of interest. Using simulations on the ALARM network, it could be investigated which test-selection strategy could provide good solutions to the MAP-SELECTION problem.

In the current study, the test selection strategies that were investigated were the in-degree, out-degree and total degree of a variable, the distance to the explanation variable, the expected utility as defined by van der Gaag and Wessels (1993), the expected Gini index and the mutual information with the explanation variable. In order to test how well the test-selection strategies performed, it was evaluated whether or not the test-selection strategies were able to rank-approximate the best variable to observe and whether or not the test-selection strategies were able to value-approximate the expected number of MAP-(in)dependent variables. Furthermore, it was evaluated whether or not the performance of the test-selection strategies depended on the number of MAP-(in)dependent variables.

In general, we found that the performance of the test-selection strategies distance, expected utility, expected Gini index and mutual information is hardly affected by the number of MAP-(in)dependent variables. Results indicate that the percentage of simulation runs in which the best variable to observe is identified by these test-selection strategies slightly decreased and the correlation with the expected number of MAP-(in)dependent variables slightly increased when the number of MAP-dependent variables increased. This suggests that the usability of the test-selection strategies does not sig-

nificantly degrade when the number of MAP-dependent variables becomes larger. For the in-degree, out-degree and total degree of a variable, we found a more significant decrease in the percentage of simulations in which the best variable to observe is identified by the test-selection strategies. Therefore, we can conclude that these test-selection strategies are not able to handle scenarios with many MAP-dependent variables that well.

Based on the simulations, we can conclude that the in-degree, out-degree and total degree of a variable do not serve as a good heuristic to increase the number of MAP-independent variables. In real-world settings, observing additional variables might be costly, damaging or otherwise unwished for (Van der Gaag, 1996). Therefore, the general goal is to observe only the most useful variables to the problem under consideration. Results indicate that the in-degree, out-degree and total degree are able to identify the best variable in less than 45% of the simulations when considering all simulation runs where a choice has to be made about which variable to select. For the particularly interesting simulations in which the number of MAP-independent variables is initially low, this is even less than 30%. Although both are better than chance, these percentages might be too low to be effectively used in real-world settings. Also with respect to value-approximation, we see no apparent correlation between the in-degree, out-degree and total degree on the one hand and the expected number of MAP-(in)dependent variables on the other hand, which further supports the conclusion that the three selection strategies are not well-suited as a heuristic.

A possible explanation for the limited use of the in-degree, out-degree and total degree of a variable as a test-selection strategy might be that we did not consider the fact that many variables which are parents (part of the in-degree) or children (part of the out-degree) of the intermediate variable could potentially already be MAP-independent. This means that the in-degree, out-degree and total degree as investigated in this study might give a wrong indication about the actual importance of an intermediate variable, therefore ultimately resulting in a sub-optimal test-selection strategy. In this regard, adjusting the test-selection strategies to take into account that variables may already be MAP-independent might improve their usability.

Whereas the test-selection strategies for the in-degree, out-degree and total degree of a variable do not serve as good heuristics for increasing the number of MAP-independent variables, simulation results suggest that the distance to the explanation variable is already a better test-selection strategy. Apart from the fact that the best variable to observe can be indicated by distance to the explanation variable in more than 80% of the simulation runs, we also see that distance correlates moderately strong with the expected number of MAP-(in)dependent variables, where a smaller distance is associated with a larger expected number of MAP-independent variables. Hence, the evaluations suggest that distance could provide a good test-selection strategy to identify which variable is best to observe, as well as to

give an indication on the expected number of MAP-(in)dependent variables that can be used to signify whether observing an additional variable is useful at all. However, a disadvantage of using the distance as a test-selection strategy is that there is often no single variable with the smallest distance to the explanation variable (see Table 4.3 and Table A.2). This makes using the distance as a test-selection strategy more difficult, as the user is still left with multiple variables to decide between. Therefore, it might be concluded that the distance to the explanation variable could be a good first step in selecting the best variable to observe. However, additional steps have to be taken to arrive at a single best variable to observe.

The intuition behind the fact that the distance to the explanation variable might be a good test-selection strategy is that intermediate variables with a smaller distance to the explanation variable are more likely to contain much information about the explanation variable, thereby reducing or blocking the influence of other variables on the explanation variable. Simulation results support this intuition, where Figure 4.3d shows that typically MAP-dependent variables have a small distance to the explanation variable. It can be argued that MAP-dependent variables contain much information about the explanation variable since these variables are able to change the best explanation of the explanation variable. Therefore, it can also be argued that variables closer to the explanation variable contain more information about the explanation variable.

Regarding the quantitative test-selection strategies, it can be concluded that the expected utility, expected Gini index and mutual information serve as good heuristics to increase the number of MAP-independent variables. From the simulation results, we observe that the three test-selection strategies have approximately the same performance. This is probably a result of the fact that the strategies all have essentially the same goal of minimizing the uncertainty in the explanation variable (Sent & Van der Gaag, 2007; Van der Gaag & Wessels, 1993). That is, the only difference between the strategies is in the way information from the posterior distributions of the explanation variable is used. The three test-selection strategies are able to indicate the best variable to observe in approximately 60% of the simulations, which increases to approximately 90% when allowing the best variable to be part of the top two variables according to the test-selection strategies. In real-world settings, this may be very reasonable. Furthermore, all test-selection strategies show a moderately strong correlation with the expected number of MAP-(in)dependent variables, suggesting that the value of the test-selection strategies can provide a reasonably good indication of the expected number of MAP-(in)dependent variables. This also contributes to the conclusion that the test-selection strategies may serve as good heuristics.

In a previous study, it has already been shown that the Gini index is well-suited for decreasing the uncertainty in explanation variable (Sent & Van der Gaag, 2007), thereby making the best explanation more robust and stable.

Additionally, it has been argued that the expected utility may be a good starting point for research in selective evidence gathering to make the best explanation more robust and stable (Van der Gaag & Wessels, 1993). In the current study, it has been shown that these test-selection strategies serve as good heuristics to increase the expected number of MAP-independent variables with respect to the explanation variable. This suggests that making the best explanation more robust and stable may implicitly be the underlying objective of increasing the (expected) number of MAP-independent variables with respect to the explanation variable. As hypothesized earlier, making the best explanation more robust and stable also means that the best explanation is less likely to be influenced by other variables, hence the number of MAP-independent variables should increase. This intuition seems to be supported by the simulation results.

In the current study, we made the claim that the justification of a MAP explanation based on the concept of MAP-independence may be inadequate when the number of MAP-independent variables is small. That is, when there are many relevant variables according to MAP-independence and thus many variables can influence the decision, the user can only gain a limited understanding of the decision. In order to make this claim, we implicitly assumed that there were scenarios with many relevant MAP-dependent variables. Figure 4.1 illustrates that this assumption is valid in the current study, where we can see a long tail in the distribution of MAP-dependent variables. In approximately a quarter of the simulations, the number of MAP-dependent variables is larger than six, which should already make it more difficult for a user to understand the decision. Therefore, we can conclude that research is really needed into test-selection strategies that can increase the number of MAP-independent variables, thereby making the decision better to motivate and easier to understand for the user.

Although this study has shown that distance to the explanation variable, expected utility, expected Gini index and mutual information may be good test-selection strategies to use in order to increase the number of MAP-independent variables with respect to the explanation variable, results need to be interpreted with care. First of all, we acknowledge that the model we used to test whether or not the test-selection strategy values correlate with the expected number of MAP-(in)dependent variables is sub-optimal. Not all assumptions for using linear regression were satisfied by the test-selection strategies, which means that the regression models might be inaccurate and therefore the correlations might also be imprecise.

Furthermore, we tested the test-selection strategies specifically on the ALARM network in the current study. This means that the simulation results might be an artifact of the characteristics of the network and therefore, the results obtained in the current study may not generalise to other Bayesian networks. More research is needed to verify whether the test-selection strategies investigated in the current study are able to solve the

MAP-SELECTION problem for a larger population of Bayesian networks. With the results of these studies, findings about the usability of the test-selection strategies can subsequently be generalised.

Lastly, we note that the list of test-selection strategies that were investigated in the current study is by no means exhaustive. Therefore, it might be worthwhile to investigate also different test-selection strategies that could potentially be used as a heuristic to increase the number of MAP-dependent variables. It may be that one or multiple test-selection strategies are a significant improvement over the current test-selection strategies that were used, which means that our conclusions about the usability of these test-selection strategies might have to be adjusted in that case.

It can also be investigated which test-selection strategies can be used as a heuristic to increase the number of MAP-independent variables when some or all of the simplifying assumptions that we made in the current study are relaxed. Namely, in the current study we took a myopic approach toward evidence gathering, which means that only one intermediate variable is selected at a time. Secondly, the explanation set consisted of only one explanation variable in this study. While these simplifications assured that the study was manageable and simulations did not take a disproportional amount of time, we note that in real-world applications, these simplifying assumptions will most likely not hold. Research has shown that experts often select tests to observe variables in packages, instead of one after the other (Sent et al., 2005). This means that experts will be more inclined to use a non-myopic approach instead of a myopic approach. Furthermore, experts might be interested in multiple explanation variables at the same time (Van der Gaag, 1996). Therefore, real-world applications might benefit from a study that investigates which test-selection strategies are useful in these settings without using the assumptions that we made in the current study.

Finally, future work may focus on finding test-selection strategies that could be used as a heuristic to increase the number of strong MAP-independent variables. In the current study, we only looked at weak MAP-independence, where variables in the set could not interact with each other. It might also be interesting to see whether there are good test-selection strategies when variables do interact with each other. We note that this might require a more sophisticated test-selection strategy that can better capture the interactions between variables.

Chapter 6

Acknowledgement

The author would like to thank dr. Kwisthout for providing detailed and valuable feedback as well as ideas that led to this Bachelor thesis. The author would also like to thank the other members of the thesis group for their contributions to the discussions in meetings, with a special thanks to Merlijn van Elteren and Luuk Jacobs for contributing to the implementation that led to the program on which the simulations could be run.

References

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Beinlich, I. A., Suermondt, H. J., Chavez, R. M., & Cooper, G. F. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In J. Hunter, J. Cookson, & J. Wyatt (Eds.), *Aime 89* (pp. 247–256). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-93437-7_28

Ben-Bassat, M. (1978). Myopic policies in sequential classification. *IEEE Transactions on Computers*, *27*(2), 170–174. https://doi.org/10.1109/TC.1978.1675054

Cofiño, A. S., Cano, R., Sordo, C., & Gutiérrez, J. M. (2002). Bayesian networks for probabilistic weather prediction. *15th European Conference on Artificial Intelligence*, 695–700.

Gallego, M. J. F. (2005). *Bayesian networks inference: Advanced algorithms for triangulation and partial abduction.* [Doctoral dissertation, Ph. D. dissertation, Universidad de Castilla La Mancha].

Glasziou, P., & Hilden, R. (1989). Test selection measures. *Medical Decision Making*, *9*, 133–141.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI-explainable artificial intelligence. *Science Robotics*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

Kwisthout, J. (2021). Explainable AI using MAP-independence. In J. Vejnarová & N. Wilson (Eds.), *Symbolic and quantitative approaches to reasoning with uncertainty* (pp. 243–254). Springer International Publishing. https://doi.org/10.1007/978-3-030-86772-0_18

Kyrimi, E., & Marsh, W. (2016). A progressive explanation of inference in 'hybrid' Bayesian networks for supporting clinical decision making. In A. Antonucci, G. Corani, & C. P. Campos (Eds.), *Proceedings of the eighth international conference on probabilistic graphical models* (pp. 275–286). PMLR.

Kyrimi, E., McLachlan, S., Dube, K., Neves, M. R., Fahmi, A., & Fenton, N. (2021). A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future. *Artificial Intelligence in Medicine*, *117*, 102108. https://doi.org/10.1016/j.artmed.2021.102108

Lacave, C., & Díez, F. J. (2002). A review of explanation methods for Bayesian networks. *Knowledge Engineering Review*, *17*(2), 107–127. https://doi.org/10.1017/S026988890200019X

Madigan, D., & Almond, R. (1996). On test selection strategies for belief networks. In D. Fisher & H.-J. Lenz (Eds.), *Learning from data: Artificial intelligence and statistics v* (pp. 89–98). Springer New York. https://doi.org/10.1007/978-1-4612-2404-4_9

Meekes, M., Renooij, S., & Van der Gaag, L. C. (2015). Relevance of evidence in Bayesian networks. In S. Destercke & T. Denoeux (Eds.), *Symbolic and quantitative approaches to reasoning with uncertainty* (pp. 366–375). Springer International Publishing. https://doi.org/10.1007/978-3-319-20807-7_33

Pearl, J., & Paz, A. (1987). *Graphoids: A graph-based logic for reasoning about relevance relations*. Tech. Rep. R-53-L, UCLA Computer Science Department.

Sent, D., Van der Gaag, L. C., Witteman, C. L., Aleman, B. M., & Taal, B. G. (2005). Eliciting test-selection strategies for a decision-support system in oncology. *AISB Journal*, *1*(6), 543–561.

Sent, D., & Van der Gaag, L. C. (2006). Automated test selection in decision-support systems: A case study in oncology. *Studies in Health Technology and Informatics*, *124*, 491.

Sent, D., & Van der Gaag, L. C. (2007). On the behaviour of information measures for test selection. In R. Bellazzi, A. Abu-Hanna, & J. Hunter (Eds.), *Artificial intelligence in medicine* (pp. 316–325). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73599-1_42

Suermondt, H. J., & Cooper, G. F. (1993). An evaluation of explanations of probabilistic inference. *Computers and Biomedical Research*, *26*(3), 242–254. https://doi.org/10.1006/cbmr.1993.1017

Van der Gaag, L. C. (1996). Bayesian belief networks: Odds and ends. *The Computer Journal*, *39*(2), 97–113. https://doi.org/10.1093/comjnl/39.2.97

Van der Gaag, L. C., & Bodlaender, H. L. (2011). On stopping evidence gathering for diagnostic Bayesian networks. In W. Liu (Ed.), *Symbolic and quantitative approaches to reasoning with uncertainty* (pp. 170–181). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-22152-1_15

Van der Gaag, L. C., & Wessels, M. (1993). Selective evidence gathering for diagnostic belief networks. *AISB Quarterly*, *86*, 23–34.

Vlek, C. S., Prakken, H., Renooij, S., & Verheij, B. (2016). A method for explaining Bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law*, *24*, 285–324. https://doi.org/10.1007/s10506-016-9183-4

Yuan, C., Lim, H., & Lu, T. (2011). Most relevant explanation in Bayesian networks. *Journal of Artificial Intelligence Research*, *42*, 309–352. https://doi.org/10.1613/jair.3301

Appendix A

Appendix

In the Appendix, we show the additional analyses that were carried out in more detail, which have as goal to evaluate whether or not the usefulness of test-selection strategies depends on the number of MAP-dependent variables in a simulation run. That is, we will show the results for the (1) rank-approximation and (2) value-approximation to see how well test-selection strategies perform in simulation runs with more than six MAP-dependent variables. In the final section, we provide a link to a GitHub page containing all code and data that is used in order to run simulations and analyse the results.

A.1 Rank-approximation second analysis

To evaluate how well a test-selection strategy can indicate which variable in the set of MAP-dependent variables is the best variable to observe, the same procedure is used as described in Section 4.1. Namely, in each simulation run, the variable with the largest expected number of MAP-independent variables is identified as the best variable to observe and the top one, top two and top three variables according to the test-selection strategies are based on minimizing (distance) or maximizing (other strategies) the test-selection strategy value. If variables are not included in the top n but do have the same test-selection strategy value as another variable that is included in the top n, we added this variable to the top n variables too.

Using the data containing simulation runs with more than six MAP-dependent variables, the theoretical chance levels for the number of simulation runs in which the best variable to observe is included in the top n variables are 4.89 (4.2%), 9.78 (8.4%) and 14.67 (12.6%) for the top one, top two and top three respectively. Table A.1 and Figure A.1 show the results for the analysis. As can be seen, all test-selection strategies perform better than chance for different values of n. Furthermore, it can be observed that the distance to the explanation variable is the best indicator of the best vari-

Table A.1
Overview of the number and percentage of simulation runs is in which the best variable to observe is included in the top one, top two and top three variables respectively for each test-selection strategy, based on minimizing (distance) or maximizing (other strategies) the test-selection strategy values. The best variable to observe is defined as the variable with the largest expected number of MAP-independent variables. For this analysis, only simulation runs with more than six MAP-dependent variables are considered.

Test-selection strategy	**Nr of sims in top 1**	**Nr of sims in top 2**	**Nr of sims in top 3**	**% of sims in top 1**	**% of sims in top 2**	**% of sims in top 3**
In-degree	6	20	38	12.766	42.553	80.851[a]
Out-degree	6	26	33	12.766	55.319	70.213
Total degree	12	23	31	25.532	48.936	65.957
Distance	39	39	43	82.979	82.979	91.489
Expected utility	26	45	45	55.319	95.745[a]	95.745[a]
Expected Gini index	29	42	45	61.702	89.362[a]	95.745[a]
Mutual information	26	36	41	55.319	76.596	87.234

[a] Indicates improvement compared to the analysis containing all simulation runs with more than one MAP-dependent variable.

able to observe and that the in-degree, out-degree and total degree perform structurally worse than the other test-selection strategies.

In general, it can be seen that the percentage of simulation runs in which the best variable is included in the top n variables according to the test-selection strategies decreased compared to the previous analysis, except for a few situations indicated in Table A.1. This suggests that having more MAP-dependent variables tends to negatively impact the performance of the test-selection strategies. However, for the distance to the explanation variable, the expected utility, expected Gini index and mutual information, this impact is not very large as the percentages only slightly decreased compared to the previous analysis (approximately 5% for the top one results). For the test-selection strategies in-degree, out-degree and total degree, the impact seems to be more substantial (more than 10% decrease for most conditions), indicating that these test-selection strategies are affected more by the number of MAP-dependent variables.

When looking at the number of simulation runs in which there is more than one best variable to observe according to the test-selection strategy as summarized in Table A.2, it can be seen that especially distance to the explanation variable has a much higher percentage of simulation runs in which there is no single best variable to observe (93.6% compared to 79.3%). However, also the in-degree and total degree show an increase compared to the previous analysis. On the other hand, the expected utility, expected Gini index and mutual information have no simulation runs anymore in which multiple variables have the same test-selection strategy value.

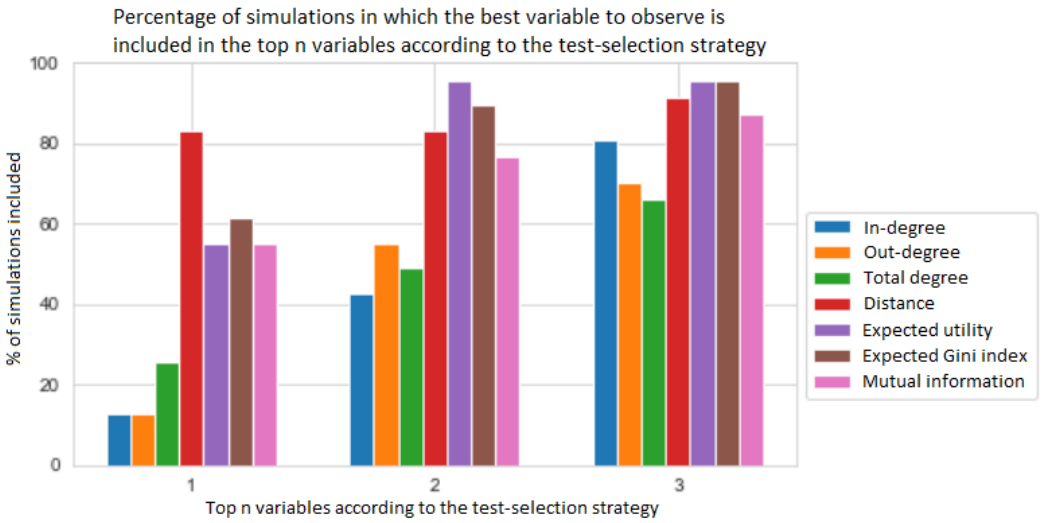

Figure A.1
Similar to Figure 4.2, where only simulation runs with more than six MAP-dependent variables are considered. The bar plots indicate for each test-selection strategy the percentage of simulation runs in which the best variable to observe is included in the top one, top two and top three results according to the test-selection strategy. Similar to Figure 4.2, the in-degree, out-degree and total degree are performing structurally worse than the other test-selection strategies. Furthermore, the distance to the explanation variable seems to be the best indicator of the best variable to observe again, although for the top two and top three variables, the expected utility and expected Gini index seem to be performing better.

Table A.2
Overview of the number and percentage of simulation runs in which there are multiple best variables to observe according to the test-selection strategies, based on minimizing (distance) or maximizing (other strategies) the test-selection strategy value. For this analysis, only the simulation runs with more than six MAP-dependent variables are considered.

Test-selection strategy	**Nr of sims with multiple best vars**	**%sims with multiple best vars**
In-degree	14	0.298[a]
Out-degree	1	0.021
Total degree	16	0.340[a]
Distance	44	0.936[a]
Expected utility	0	0.000
Expected Gini index	0	0.000
Mutual information	0	0.000

[a] Indicates a higher percentage compared to the analysis containing all simulation runs with more than one MAP-independent variable.

A.2 Value-approximation second analysis

To evaluate how well the test-selection strategy values (independent variable) correlate with the expected number of MAP-dependent variables (dependent variable), linear regression is used similar to the analysis carried out in Section 4.2. We note again that the assumptions of normally distributed residuals and equal variance are violated by some test-selection strategies, so the results have to be interpreted with care. For the analysis, the expected utility, expected Gini index and mutual information have been log-transformed again.

Table A.3 summarizes the results for the regression models where only simulation runs with more than six MAP-dependent variables are considered. As can be seen in the results, all test-selection strategies have a statistically significant correlation with the expected number of MAP-dependent variables, where the in-degree has a statistically significant correlation with the expected number of MAP-dependent variables at the $p < 0.05$ level and the other test-selection strategies have a statistically significant correlation at the $p < 0.01$ level.

When looking at Pearson's r, Table A.3 shows that the in-degree, out-degree and total degree have almost no correlation with the expected number of MAP-dependent variables. The other test-selection strategies show a moderately strong correlation with the expected number of MAP-dependent variables, where the correlation is positive for distance, indicating that a larger distance tends to result in a larger expected number of MAP-dependent variables and hence a smaller expected number of MAP-independent variables. On the other hand, the expected utility, expected Gini index and mutual information have a negative correlation with the expected number of MAP-dependent variables, indicating that the expected number of MAP-dependent variables tends to decrease and hence the expected number of MAP-independent variables tends to increase when these test-selection strategies show a higher value. Compared to the previous analysis, we see that the correlations are in general (except for the distance) higher than before, although the strengths of the correlations have not significantly improved.

A.3 Code

This section contains the link to the GitHub page which contains all code and data that is used to analyse the test-selection strategies. The page contains both the java program that is used for the simulation runs, the data that was obtained from running 164 simulation runs as well as the python file that is used to analyse the data.

Table A.3
Overview of the correlations of the test-selection strategy values with the expected number of MAP-dependent variables and the fitted regression models. For this analysis, only simulation runs with more than six MAP-dependent variables are considered.

Test-selection strategy	Transformation	Pearson's r	R-squared	Adj. R-squared	Intercept	Coef
In-degree*	None	-0.100	0.010	0.008	9.921	-0.464
Out-degree**	None	-0.135	0.018	0.016	9.885	-0.541
Total degree**	None	-0.194	0.038	0.036	11.184	-0.685
Distance**	None	0.549	0.301	0.300	4.156	2.166
Expected utility**	Log	-0.614	0.377	0.376	2.891	-2.914
Expected Gini index**	Log	-0.681	0.464	0.463	3.276	-2.087
Mutual information**	Log	-0.667	0.452	0.451	5.116	-2.020

* Indicates significance at the $p < 0.05$ level.
** Indicates significance at the $p < 0.01$ level.

https://github.com/SimonJanssen1/MAP-independence_Thesis_Simon_Janssen.git