Explainable Bayesian networks: taxonomy, properties and approximation methods

Iena Petronella Derks

Explainable Bayesian networks: taxonomy, properties and approximation methods

By

Iena Petronella Derks

13075782

Supervisor: Dr A. de Waal

Submitted in partial fulfilment of the requirements for the degree

PhD (Mathematical Statistics)

in the

Faculty of Economic and Management Sciences

University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

July 22, 2024

Declaration of Originality

I declare that the thesis, which I hereby submit for the degree PhD (Mathematical Statistics) at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: IP Derks

Date: 2024/07/22

Ethics statement

The author, whose name appears on the title page of this dissertation, has obtained the required research ethics approval/exemption for the research described in this work. The author declares that he/she has observed the ethical standards required in terms of the University of Pretoria's Code of ethics for scholarly activities.

This research is approved by the University of Pretoria Economic and Management Science ethics committee under the ethics number EMS114/23.

Abstract

Technological advances have integrated artificial intelligence (AI) into various scientific fields, necessitating understanding AI-derived decisions. The field of explainable artificial intelligence (XAI) has emerged to address transparency concerns, offering both transparent models and post-hoc explanation techniques. Recent research emphasises the importance of developing transparent models, with a focus on enhancing the interpretability of these models. An example of a transparent model that would benefit from enhanced post-hoc explainability is Bayesian networks. This research investigates the current state of explainability in Bayesian networks. Literature includes three categories of explanation: explanation of the model, reasoning, and evidence. Drawing upon these categories, we formulate a taxonomy of explainable Bayesian networks. Following this, we extend the taxonomy to include explanation of decisions, an area recognised as neglected within the broader XAI research field. This includes using the same-decision probability, a threshold-based confidence measure, as a stopping and selection criteria for decision-making. Additionally, acknowledging computational efficiency as a concern in XAI, we introduce an approximate forward-gLasso algorithm as a solution for efficiently solving the most relevant explanation. We compare the proposed algorithm with a local, exhaustive forward search. The forward-gLasso algorithm demonstrates accuracy comparable to the forward search while reducing the average neighbourhood size, leading to computationally efficient explanations. All coding was done in `R`, building on existing packages for Bayesian networks. As a result, we develop an open-source `R` package capable of generating explanations of evidence for Bayesian networks. Lastly, we demonstrate the practical insights gained from applying post-hoc explanations on real-world data, such as the South African Victims of Crime Survey 2016 - 2017.

Acknowledgements

> "*What is the bravest thing you've ever said?*" asked the boy.
> "*Help,*" said the horse. "*Asking for help isn't giving up,*" said the horse. "*It's refusing to give up.*"
>
> Charlie Mackesy,
> The Boy, the Mole, the Fox and the Horse.

First and foremost, I would like to thank God. He has given me strength and encouragement throughout all the challenging moments. I am truly grateful for His unconditional and endless love, mercy, and grace.

I express my deepest gratitude to my supervisor, Dr A. de Waal, for her constant guidance, mentorship, and valuable insights during my PhD journey. Her expertise, dedication, and unwavering support have played a crucial role in shaping this research and my academic development. I am also particularly grateful to Professor I. Fabris-Rotelli for her valuable comments and feedback. I am deeply grateful to the intellectual contributions of Jarod, JP, and Dr T. Loots in the conceptualisation of the Forward-gLasso algorithm.

I would be remiss in not mentioning my family. Their consistent encouragement and unwavering belief in me have kept my spirits and motivation high during this process. I want to express my sincere gratitude to Peet for his support and encouragement throughout this journey. He always believed in me, even when I doubted myself. Special thanks to Renate for her support throughout this journey, especially for not tripping us up during the three-legged race. We conquered it! I am also incredibly grateful for Gandhi's friendship. Her laughter during challenging moments brought a much-needed sense of relief and made the journey all the more enjoyable. I would also like to acknowledge my cats for the entertainment and emotional support they have provided.

Lastly, I acknowledge the CSIR-DSI Inter-bursary Support Programme (IBS) for their financial support.

Research Outputs

Publications:

- Iena Petronella Derks and Alta de Waal. A Taxonomy of Explainable Bayesian networks. In *Southern African Conference for Artificial Intelligence Research*, pages 220–235. Springer, 2020. Derks & de Waal (2020)

- Gandhi Jafta, Alta de Waal, Iena Derks, and Emma Ruttkamp-Bloem. Evaluation of XAI as an Enabler for Fairness, Accountability and Transparency. In *Proceedings of the Second Southern African Conference for Artificial Intelligence Research*, 2021. Jafta et al. (2021)

Conference and Symposium presentations:

- A Taxonomy of Explainable Bayesian networks, Southern African Conference for Artificial Intelligence Research 2020.

- Evaluation of XAI as an Enabler for Fairness, Accountability and Transparency, Southern African Conference for Artificial Intelligence Research 2021.

- A Statistical Approach to Explainability in Artificial Intelligence, International Symposium on Modern Biostatistics and Statistics 2022.

- A Glasso-Forward Search for Solving Most Relevant Explanation in Bayesian Networks, International Symposium on Modern Biostatistics and Statistics 2023.

Articles submitted:

- A Forward-gLasso Search for Solving Most Relevant Explanation in Bayesian Networks.

Contents

List of Figures iii

List of Tables v

1 Introduction 1

1.1 Motivation . . . 2

1.1.1 Transparency . . . 4

1.1.2 The performance-explainability trade-off . . . 5

1.1.3 Can we measure explanation quality? . . . 5

1.1.4 Efficient computation of explanations . . . 7

1.2 Bayesian networks as a proposed solution . . . 7

1.3 Research aims and objectives . . . 9

1.4 Contribution to scientific research . . . 10

1.5 Overview of thesis . . . 11

1.6 Data and resources . . . 12

2 BNs as inherently explainable models 13

2.1 Introduction . . . 13

2.2 Overview of Bayesian Networks . . . 14

2.2.1 Conditional independence in Bayesian networks . . . 14

2.2.2 Inference in Bayesian networks . . . 15

2.2.3 Decision problems in Bayesian networks . . . 17

2.2.4 Bayesian network software . . . 19

2.3 Explanation of the model . . . 19

2.4 Explanation of reasoning . . . 23

2.4.1 Chains of reasoning . . . 23
2.4.2 Variable importance . . . 24
2.4.3 Counterfactual and contrastive explanations . . . 25
2.4.4 Scenario-based explanations . . . 25
2.5 Conclusion . . . 26

3 Post-hoc explanation in Bayesian networks 28
3.1 Introduction . . . 28
3.2 Explanation of evidence . . . 29
3.2.1 Running example . . . 31
3.2.2 The most probable explanation . . . 32
3.2.3 The most relevant explanation . . . 34
3.3 Explanation of decisions . . . 39
3.3.1 Concepts in statistical decision theory . . . 40
3.3.2 Same-decision probability . . . 46
3.4 Conclusion . . . 50

4 Forward-gLasso search for solving the most relevant explanation 51
4.1 Introduction . . . 51
4.2 Graphical Lasso . . . 52
4.2.1 Search strategy . . . 53
4.3 Experimental design . . . 57
4.4 Experimental results . . . 58
4.4.1 Neighbourhood reduction . . . 58
4.4.2 Computational efficiency . . . 59
4.4.3 Most relevant explanation according to forward-gLasso . . . 60
4.5 Testing robustness of the MRE with the SDP . . . 62
4.6 Conclusion . . . 64

5 Taxonomy of explainable Bayesian networks 66
5.1 Introduction . . . 66
5.2 Taxonomy of explainable Bayesian networks . . . 67
5.2.1 Reasoning . . . 68

5.2.2 Evidence . . . 70
5.2.3 Decisions . . . 70
5.3 A package for solving the most relevant explanation . . . 73
5.3.1 Installation . . . 73
5.3.2 Specifying the parameters . . . 73
5.3.3 Practical demonstration . . . 74
5.4 Conclusion . . . 80

6 Explainable Bayesian networks in action: South African VCS 81
6.1 Introduction . . . 81
6.2 Data preparation . . . 82
6.3 Actionable insights: MRE . . . 83
6.3.1 Case study 1: rising crime in Mpumalanga . . . 84
6.3.2 Case study 2: victim perception . . . 86
6.4 Actionable insights: SDP . . . 90
6.4.1 Case study 3: public perception of the SAPS . . . 91
6.5 Conclusion . . . 93

7 Conclusion 94
7.1 Contributions to scientific research . . . 95
7.2 Future Work . . . 97
7.3 Limitations . . . 98

Bibliography 99

A List of abbreviations and symbols A1

B Description of variables used B1

List of Figures

1.1 Machine learning models and their respective performance vs explainability (Gunning & Aha 2019). 6

2.1 Types of reasoning in Bayesian networks (adapted from Korb & Nicholson (2010)). 18

2.2 Graphical display of the arc strengths as measured by the Bayesian information criterion in the Insurance network from Binder et al. (1997). 21

2.3 Graphical display of the marginal probabilities and the `fdp` layout from `Rgraphviz`. 22

3.1 Graphical illustration of the Insurance Bayesian network from Binder et al. (1997). 33

3.2 Solution space for the three target variables of interest, *Antilock* (A), *OtherCar* (B), and *Airbag* (C), in the Insurance (Binder et al. 1997) example. *Antilock*, *OtherCar*, and *Airbag* take states $\{True, False\}$, where state *False* is indicated as $\overline{a}, \overline{b}, \overline{c}$. 37

3.3 Illustration of the search path for *OtherCar* (B) through the forward search algorithm. 40

3.4 A flowchart for decision-readiness. 45

3.5 A naïve Bayesian network with a hypothesis variable D and four features $H_1, \cdots, H_4$. 47

3.6 The Asia Bayesian network from Lauritzen & Spiegelhalter (1988). 49

4.1 Path for *OtherCar* (**B**) through forward-gLasso search. 56

5.1 A schematic view of explainable Bayesian networks. 68

6.1 Graphical display of the learned structure for the South African VCS data set using a hill-climbing search algorithm. 84

List of Tables

3.1 MAP-generated variable instantiations for the Insurance running example. . 34
3.2 Brute-force MRE-generated variable instantiations for the Insurance running example scenario. 37
3.3 The conditional probability tables associated with the naïve Bayesian network in Figure 3.5. 47
3.4 Scenarios for the latent evidence variables for the naïve Bayesian network in Figure 3.5. 48

4.1 Summary of the benchmark networks used in the experiments. 58
4.2 Comparison of the average neighbourhood size of each algorithm. 59
4.3 Comparison of test cases solved exactly (CSE) and the average execution time (AET) of each algorithm in seconds. 60
4.4 Set of explanations for the Insurance network using the forward-gLasso algorithm. 60
4.5 Set of explanations for scenario 2 of the Insurance network using the forward-gLasso algorithm. 61
4.6 Description of variables of interest in the Win95pts Bayesian network. . . . 63
4.7 Most relevant explanation for initial evidence set. 63
4.8 Most relevant explanation for updated evidence set. 64

6.1 Variable encoding for target variables in case study 1. 85
6.2 Most relevant explanations for case 1. 86
6.3 Variable encoding for target variables in case study 2. 87
6.4 Most relevant explanations for the victim profiles in case 2. 88

6.5 Same-decision probabilities for case study 3. 92

B.1 Description of variables included in the Insurance Bayesian network from Binder et al. (1997). B2

B.2 Description of variables included in the South African VCS 2017 - 2018. Variables acc - pol. B3

B.3 Description of variables included in the South African VCS 2017 - 2018. Variables pr - why. B4

Chapter 1

Introduction

Technological advances have brought artificial intelligence (AI) closer to humans, transforming how we approach everyday tasks, from AI-driven virtual assistants to autonomous vehicles navigating our streets. However, despite these advancements, a growing concern remains regarding the lack of transparency and interpretability in AI models and algorithms, especially when applied to sensitive applications (Barredo Arrieta et al. 2020, Longo et al. 2020). This lack of transparency often manifests in what's known as the "black-box" nature of AI – where the model operates without explicitly showing *how* or *why* it arrives at a particular outcome.

Recognising the need for transparency and interpretability in AI models, the field of explainable artificial intelligence (XAI) has developed. XAI includes models explainable-by-design, featuring inherently transparent structures that facilitate intuitive understanding, as well as post-hoc explanation techniques aimed at explaining model outputs (Guidotti et al. 2018, Lipton 2018, Barredo Arrieta et al. 2020), such as SHapley Additive exPlanations (SHAP) (Lundberg & Lee 2017) and LIME (Ribeiro et al. 2016). These models go beyond just making accurate predictions or decisions. They can also provide clear and understandable explanations for their reasoning process (Escalante et al. 2018). Given the broad scope of XAI, reviewing all methods is beyond the scope of this work. Barredo Arrieta et al. (2020) provides a comprehensive review of XAI methods. Consequently, XAI acts as a tool that answers critical how and why questions, facilitating verification, improvement, and responsible management of AI models, fostering a fair, accountable, and transparent human-centred approach, and ultimately, enabling users to trust AI-derived

results and decisions (Cath 2018, Greene et al. 2019, Leslie 2019). Ribeiro et al. (2016) emphasises the role of *trust* in human interaction with AI models. The level of trust is closely tied to understanding the model's behaviour.

The literature on XAI identifies several motivations for building explainable models. Although these reasons do not occur in isolation and may overlap, they capture different motivations. One of the main reasons is to *justify* AI-derived predictions and decisions (Adadi & Berrada 2018). Rather than only explaining the inner workings or reasoning of the model, it is important to use XAI to justify an outcome or decision of the AI (Saeed & Omlin 2023). In other words, to show that it is reasonable. Another motivation for XAI is to *control* AI systems. Understanding a system's behaviour and reasoning allows users to intervene and adjust the model to align with requirements (Keane & Smyth 2020, Ghai et al. 2021). A third motivation for XAI is to *improve* AI systems. If one can explain the model, it can also be easier to improve (Adadi & Berrada 2018). Finally, XAI can be used to *discover* new knowledge (Saeed & Omlin 2023). Often, explanations include information which humans might find counterintuitive or wrong. These explanations can aid in knowledge discovery since they can point us to new areas for research.

Lacave & Díez (2002) highlights three aspects that must be explained within the framework of any expert system: the knowledge base, the reasoning process, and the evidence propagated.

1.1 Motivation

Despite significant research on explainability, the XAI community has yet to reach a consensus on the definition of "explanation". It lacks a standardised framework for assessing the quality of different explanation methods. While numerous efforts have been made to define explanation and explainability, (Doshi-Velez & Kim 2017, Lipton 2018, Gilpin et al. 2019), none of these definitions incorporate mathematical formalism (Barredo Arrieta et al. 2020). Rosenfeld & Richardson (2019) defines explainability as "*the ability for the human user to understand the agent's logic*", whereas Das & Rad (2020) define explanation as "*additional meta information, generated by an external algorithm or by the machine learning (ML) model itself, to describe the feature importance or relevance of an input instance towards a particular output classification*". Lacave & Díez (2002) offers

another perspective, stating that explanation entails "*exposing something in such a way that is understandable for the receiver of the explanation, which implies that he/she improves his/her knowledge about the object of the explanation; and is satisfactory as far as it covers the receiver's expectations*". Lastly, Nauta et al. (2023) defines an explanation as "*a presentation of (aspects of) the reasoning, functioning and/or behaviour of a machine learning model in human-understandable terms*".

The lack of consensus on the composition of an explanation has led to a gap between users' requirements and what AI researchers are producing. According to Ras et al. (2018), XAI explanations must cater to a diverse audience beyond technical experts, including consumers, regulators, and business executives. Differing levels of expertise and context-specific requirements across stakeholders can make it challenging to provide explanations that resonate with everyone. This suggests that explanations are versatile and can be used to explain the model as a whole, i.e., global explanations, or to explain single instances or decisions, i.e., local explanations (Das & Rad 2020). This is further reflected in the definition by Nauta et al. (2023). Moreover, Lacave & Díez (2002) identify two primary objectives of explanation: *description*, which provides insight into the underlying knowledge base, conclusions or intermediate results, and *comprehension*, which aims to cultivate user understanding of model implications, the system conclusions, as well as the relationship between them.

As such, explanations can take various formats; for example, explanations can take the form of visualisations or natural language (Goebel et al. 2018, Mittelstadt et al. 2019, Barredo Arrieta et al. 2020). Another explanation format is case-based reasoning in which the current prediction is compared to similar historical cases (Leake & Mcsherry 2005, Kolodner 2014). According to the definition of explanation by Das & Rad (2020), explanations in this context would highlight the individual features that contribute most to the model's decision. Whereas for Pearl (1988), an explanation consists of the most probable assignment of variables for some observed evidence.

Accordingly, the target audience is a pivotal aspect to consider when generating an explanation. Because AI systems are employed in various sectors with different goals, it is reasonable to expect several distinct user communities (Wick 1989). Although there may be some overlap, these audiences are not identical in their intent, requirements,

expectations, and demands from explainability (Preece et al. 2018, Langer et al. 2021, Barredo Arrieta et al. 2020). In recent work, Dwivedi et al. (2023) provides an overview of the various types of stakeholders involved in the XAI process and divides them into two phases: the *understanding* phase and the *explaining* phase. During the understanding phase, stakeholders are involved with improving the model before deploying it to the explaining phase. These stakeholders are categorised as *developers* in Barredo Arrieta et al. (2020), Langer et al. (2021), but also include *theorists* and *data scientists*. The explaining phase involves four main stakeholders: *users*, *consumers*, *businesses*, and *regulators*.

Having gained a foundational understanding of explainability, we now explore specific facets of explainability, such as transparency, the trade-off between performance and explainability, methods of measuring explanation quality, and strategies for efficiently computing explanations. We are particularly interested in transparency and the performance-explainability trade-off. A less transparent model, while potentially more accurate, might lead to opaque decisions that users struggle to understand and trust. Transparency fosters collaboration and allows humans to intervene when necessary. Furthermore, computational efficiency is a critical concern in XAI. Real-time applications and resource limitations necessitate efficient explanation generation. Beyond efficiency, understanding what constitutes a good explanation allows for clear communication between AI and humans, leading to better decision-making.

1.1.1 Transparency

Researchers often refer to inherently explainable models as transparent, i.e., we can easily trace how the model works and understand how each feature contributes to the prediction. Here, transparency is considered in three levels, based on the functional domain, namely *algorithmic transparency*, *decomposability*, and *simulatability* (Lipton 2018, Mittelstadt et al. 2019, Futia & Vetrò 2020). Algorithmic transparency means that the model's decisions should be visible and understandable to those affected (Diakopoulos & Koliska 2017). This includes understanding the steps used to make a decision and the rationale behind those steps. Decomposability refers to the property that each part or component of the model, i.e., model parameters and calculations, should have an intuitive interpretation (Lipton 2018, Lepri et al. 2018). This allows users to analyse the individual contributions

of features to overall predictions, which proves valuable for understanding the relative importance of factors in the decision-making process (Minh et al. 2022). Lastly, simulatability refers to the ease with which a human can mimic (or simulate) the decision process of a model (Barredo Arrieta et al. 2020). Simulatable models encompass both algorithmic transparency and decomposability. Accordingly, Lipton (2018) notions that a model is considered transparent if it can be contemplated in its entirety.

1.1.2 The performance-explainability trade-off

A majority of researchers acknowledge a supposed "trade-off" between explainability and performance in AI models, where a higher prediction accuracy is often obtained by a less explainable model (Xu et al. 2019). Figure 1.1 illustrates the performance-explainability trade-off of well-known statistical models. This has received criticism, stemming from the *Explainable Machine Learning Challenge* at the annual Neural Information Processing Systems (NeurIPS) conference 2018. The challenge required teams to create black-box models for the given data set and explain how the model works. Instead, one team developed a fully interpretable model. Rudin & Radin (2019) question the overuse of black-box models, arguing explanations for these models may be misleading and difficult to understand. They advocate for data scientists to consider a broader spectrum of models, including inherently interpretable models. While researchers have explored using a second model to explain a black-box model (post-hoc explanations), Rudin (2019) argue that explanations generated this way might be unreliable or difficult to understand. Therefore, focusing on developing inherently transparent models from the outset can be more beneficial. Minh et al. (2022) advocates for designing inherently explainable models rather than focusing on black-box models that can cause harm to society.

1.1.3 Can we measure explanation quality?

Despite ongoing discussion surrounding the definition of explanation in XAI, numerous researchers have proposed characteristics that they consider essential for a *"good"* explanation. One common evaluation strategy involves presenting individual examples that appear plausible (Murdoch et al. 2019). However, many researchers caution against solely relying on such anecdotal evidence (Adebayo et al. 2018). According to Miller (2019), the

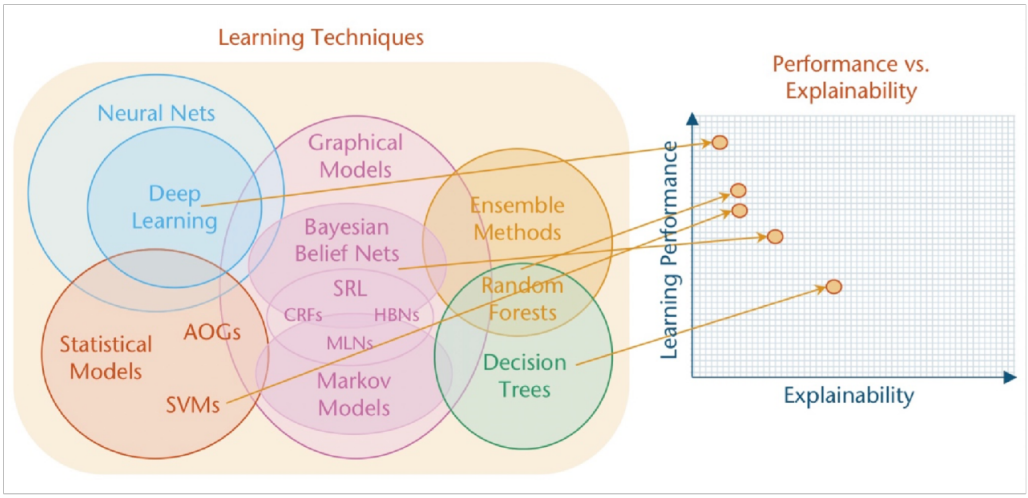

Figure 1.1: Machine learning models and their respective performance vs explainability (Gunning & Aha 2019).

majority of XAI research is guided by the researcher's intuition regarding what constitutes a good explanation. Whereas Longo et al. (2020) suggests that the quality of an explanation is influenced, to some extent, by the recipient thereof. Moreover, Linardatos et al. (2020) notes the absence of a universally accepted metric to measure explanation quality. Leake & Mcsherry (2005) underscores this problem, arguing that the lack of such a qualitative measure hinders XAI development.

Srinivasan & Chander (2021) define four primary features of a good explanation: *simplicity*, *robustness*, explanations should *bridge the knowledge-understanding gap*, and *self-evidencing*. Whereas Nauta et al. (2023) proposes measuring explainability by evaluating the degree to which 12 quality properties are met, these include *completeness*, *consistency*, and *confidence*. Arya et al. (2021) presents two qualitative metrics that serve as indicators of explanation quality: *faithfulness* and *monotonicity*. The former examines the accuracy of feature importance while the latter assesses whether adding more positive evidence increases the probability of classification for that class. Lastly, Yuan, Lim & Lu (2011) define two properties for a good explanation: *precise* and *concise*. Where *precise* entails minimising the surprise value and regards *high explanatory power* as a metric for preciseness. While a *concise* explanation contains only the most relevant variables.

1.1.4 Efficient computation of explanations

One of the goals of XAI is to develop explanation techniques that provide meaningful explanations in a computationally efficient manner (Barredo Arrieta et al. 2020). This was also highlighted in the "Great AI Debate" at the annual NeurIPS conference in 2017, where the importance of interpretability in machine learning was discussed. Several methods have been developed to provide efficient explanations, for example, Kernel SHAP offers computational efficiency and accurate approximation (Lundberg & Lee 2017, Dwivedi et al. 2023). Whereas Artelt & Hammer (2021) explore model-specific methods for generating computationally efficient contrastive explanations. Chuang et al. (2023) study efficient XAI and categorise existing techniques into two categories: *efficient non-amortised* and *efficient amortised* methods.

1.2 Bayesian networks as a proposed solution

Following the recommendations from Rudin (2019) and Minh et al. (2022), we will use a more transparent model and focus on enhancing the explainability thereof. Bayesian networks, a probabilistic graphical model, lie at the intersection of AI, ML and statistics. The graphical, qualitative structure of Bayesian networks allows the end-user to visually note the relationships among the variables, which promotes transparency (Chen & Pollino 2012, Moe et al. 2021). Furthermore, the Bayesian network framework allows us to perform probabilistic queries that can be framed as simple explanatory questions, such as, "*What is the probability of event X occurring, given the chain of events?*"

However, the inner workings – such as independence-dependence relationships and probabilistic belief updating – can present challenges for intuitive understanding (Korb & Nicholson 2010). Accordingly, these probabilistic graphical models fall somewhere in the middle of the performance-explainability trade-off (as illustrated in Figure 1.1). Literature categorises Bayesian networks as inherently transparent: they are simulatable, decomposable, and algorithmically transparent (Barredo Arrieta et al. 2020). Nonetheless, they lose their simulatability and decomposability properties when they become more complex. Therefore this suggests that Bayesian networks, although considered explainable-by-design (de Waal & Joubert 2022), can benefit from additional post-hoc explainability. As such,

according to the definition of explanation by Nauta et al. (2023), we can formulate these explanations to support the reasoning, function or behaviour of the Bayesian network.

Lacave & Díez (2002) categorise explanation tasks into three distinct categories: *explanation of the model*, *explanation of reasoning*, and *explanation of evidence*. These methods are categorised based on the focus of explanation and further subcategorised based on three properties: *content*, *communication*, and *adaption*. Not included in these categories are methods that describe whether the user is ready to make a decision, and if not, what additional information is required to better prepare for decision-making. A notable omission, since Bayesian networks are often utilised as decision-support tools (Druzdzel 1993). According to Främling (2020), decision theory has been neglected in the broader explainable artificial intelligence research field. Furthermore, explaining model decisions under uncertainty is difficult due to the lack of a formal methodology for the treatment of important variables (Guidotti et al. 2018). Lastly, the presence of uncertainty can greatly impact a decision-maker's ability to reach and appropriately *trust* the output obtained.

Within the Bayesian network domain, explanation of evidence involves finding the configuration of variables most likely to explain the observed evidence, a concept commonly referred to as abductive inference (Flores et al. 2005, Gámez 2004). Various methodologies have been developed to find the most likely configuration of variables that optimally explain the observed evidence. Kwisthout (2015) propose the concept of the most frugal explanation in Bayesian networks, a heuristic approach to the maximum-a-posteriori problem, and highlight its inherent computational intractability and the possibility of tractable approximation when subjected to specific situational constraints. Yuan, Lim & Littman (2011) propose several local algorithms to solve explanation of evidence. However, these algorithms involve an exhaustive search of all solutions, which can be computationally expensive. Akin to how the number of possible structures increases exponentially as more variables are added (Jensen & Nielsen 2007), the search space encompassing all possible configurations capable of explaining the evidence also increases exponentially. Notably, including at least one additional variable increases the complexity of the search space (Gelsema 1995). Therefore, an exhaustive enumeration of all possible configurations, as done in Yuan, Lim & Littman (2011), becomes impractical. Hence, the need arises for an efficient search algorithm to identify the optimal configuration of target nodes within a

reasonable timeframe.

Lastly, it is important to highlight that although there are numerous software applications, such as BayesiaLab[1] and HuginExpert[2], available for providing explanations within Bayesian networks, these software applications are not open-source and, as such, provide users with limited access to these important tools. Current open-source Bayesian network packages in `R` (R Core Team 2020) include `bnlearn` (Scutari 2010) and `gRain` (Højsgaard 2012). However, they do not offer explanation functionalities; instead, they offer functionalities such as structure learning, parameter estimation, and inferences.

1.3 Research aims and objectives

Three main objectives drive this research: 1) extend the current classification of explanation methods in Bayesian networks to include decision-theoretic methods that support decision-readiness and integrate them into a user-friendly and intuitive taxonomy; 2) develop an efficient search algorithm capable of providing explanations for some observed evidence; and 3) develop an open-source `R` package dedicated to explanation of evidence methods. To accomplish these goals, we aim to:

- Investigate and review the current state of explainability in Bayesian networks to organise the existing state-of-the-art methods according to our proposed taxonomy.
- Explore neighbourhood pruning algorithms, such as the graphical Lasso algorithm, to address computational challenges faced while searching for the most likely (relevant) variable configuration.
- Explore Decision Theory and decision-making in Bayesian networks to develop a new category in the proposed taxonomy: explanation of decisions.
- Investigate the current software tools to implement explainability methods in Bayesian networks.
- Illustrate the potential of the proposed taxonomy with experiments on well-known, established Bayesian networks.

[1] BayesiaLab: https://www.bayesia.com/bayesia
[2] HuginExpert: https://www.hugin.com/

- Implement the explainability methods captured in the open-source package to show it is computationally applicable and useful in real-world scenarios.
- Discuss the potential and shortcomings of this research.

1.4 Contribution to scientific research

Although Bayesian networks are powerful tools for decision-making (Jensen & Nielsen 2007), there is limited research on *explaining* decision-readiness in Bayesian networks. Consequently, while we endorse the explanation categories based on the focus of explanation as proposed by Lacave & Díez (2002), they do not account for these types of explanations. Hence, we propose a user-friendly and intuitive explanation taxonomy, encompassing existing explanation methods while introducing a fourth category: *explanation of decisions*. To our knowledge, we have yet to find any peer-reviewed work on an explanation facility for decision-readiness in Bayesian networks. We are particularly interested in the same-decision probability, which is a confidence measure that represents the probability that a specific threshold-based decision would be made if information about unobserved variables had been made available. While the same-decision probability allows us to explore two queries related to statistical decision theory, we utilise the same-decision probability to explore the potential impact of unobserved variables. For example, what if, upon observing such a variable, the explanation obtained from, say, the most relevant explanation changes? Our proposed taxonomy has already gained some attention from fellow researchers, where Valero Leal (2022) extended our taxonomy to include an additional category focused on providing explanation support.

Next, this research addresses one of the challenges of explainable artificial intelligence, i.e., providing meaningful explanations in a computationally efficient manner. Previous work has focused on developing mostly exhaustive search algorithms to solve the most relevant explanation in Bayesian networks. We propose incorporating the graphical Lasso, a statistical neighbourhood selection method, with a classic forward search algorithm to prune the search space. We evaluate the performance and computational efficiency of the proposed algorithm on a set of benchmark Bayesian networks. Furthermore, building on existing Bayesian network packages in `R`, such as `gRain` (Højsgaard 2012), we develop

an `R` package dedicated to providing explanations for observed evidence. The `R` package includes three search algorithms: an exhaustive brute-force search, a local (exhaustive) forward search algorithm, and the proposed approximate forward-gLasso search algorithm.

Lastly, this research extends the current literature and implementations in explainable Bayesian networks. Although various fields of science have highlighted the importance of explainability, limited real-world applications have been implemented. Explainability methods have often been the topic of theoretical discussions and toy implementations. As such, we apply the most relevant explanation and same-decision probability to a real-world publicly available data set to showcase these methods as well as the actionable insights obtained from these.

1.5 Overview of thesis

The document is structured as follows:

- Chapter 2 presents Bayesian networks as inherently transparent models. We provide a brief overview of Bayesian networks in Section 2.2. Section 2.3 presents methods of explainability focused on providing insights into the knowledge base of the model. Lastly, we explore explanation of reasoning techniques in Section 2.4.

- Chapter 3 is focused on post-hoc explanation in Bayesian networks. In particular, Section 3.2 presents local explanations for observed instances, such as the most probable explanation and the most relevant explanation. Section 3.3 explores principles from statistical decision theory and methods, such as the same-decision probability, that act as a confidence measure for decision-readiness.

- We present the proposed forward-gLasso search to efficiently prune the search space for the most relevant explanation as an explanation of evidence method in Chapter 4. This includes computational experiments on established benchmark Bayesian networks. After that, Section 4.5 explores the dynamic nature of the most relevant explanation by including additional evidence. The additional evidence is determined using the same-decision probability as a selection criterion.

- We organise the existing explanation methods and the newly proposed explanation

of decisions arch into a taxonomy in Chapter 5. Additionally, this chapter includes a demonstration of the `XBN R` package developed in this research.

- In Chapter 6, we implement explanation of evidence and explanation of decision methods on real-world data to demonstrate the actionable insights one can obtain from these explanations. We provide three case studies. Section 6.3 includes two case studies focused on explanation of evidence, while Section 6.4 features a case study on explanation of decisions.

- Finally, in Chapter 7, we draw concluding remarks, discuss the potential and shortcomings of this research, and present our future endeavours.

1.6 Data and resources

This research will draw upon the wealth of open-source data available through repositories such as the Bayesian Network Repository[3] available through `bnlearn` (Scutari 2010), which offers extensive collections of data relevant to our research domain. The `bnlearn` repository offers a wide array of frequently utilised reference Bayesian networks that serve as benchmarks in academic literature. These networks span various sizes, ranging from small networks with fewer than 20 nodes to massive networks exceeding 1000 nodes. These reference networks will allow us to evaluate and compare methodologies across various network sizes and complexities.

In addition, we will use the South African Victims of Crime Survey (VCS) 2017 - 2018, which is a comprehensive nationwide household-based survey that collects data on the prevalence of certain types of crime (Statistics South Africa 2018). The VCS utilises a Master Sample frame derived from the South African Census 2011. These data sets are publicly available through Statistics South Africa.

To supplement the available data, this research will incorporate data simulation. As a result, we can manipulate and control various parameters and explore specific scenarios, allowing for a more comprehensive analysis. We can study particular hypotheses and test the performance of the proposed methods under controlled conditions. We will store the code and simulated data on GitHub[4] to ensure reproducibility and ease of collaboration.

[3] The contents of this page are licensed under the Creative Commons Attribution-Share Alike License.
[4] https://github.com/iEna101/XBN_experiments to access the `R` scripts used in this thesis.

Chapter 2

Bayesian networks as inherently explainable models

2.1 Introduction

Bayesian networks (BNs) (Pearl 1988) are probabilistic graphical models that serve as tools to manage uncertainty. They leverage the combined strengths of graph theory and probability theory to represent the relationships between variables visually. This graphical representation allows for an intuitive understanding of the dependencies (and independencies) among variables and facilitates reasoning under uncertainty. The strength of Bayesian networks lies in their seamless integration of modelling and inference within a single framework. This allows Bayesian networks to predict outcomes and explain their reasoning, making them valuable in applications requiring transparency and understanding.

Their ability to explain their reasoning, coupled with their transparent graphical structure, sets it apart from many other models that lack such inherent explainability. Moreover, Bayesian networks can address diagnostic and counterfactual questions, such as: "*Was it X that caused Y, or rather something different?*" or "*If I have evidence that X did not happen, what then was the most likely cause for Y?*" or "*What if I had acted differently?*" As a result, BNs are often referred to as "explainable-by-design" (de Waal & Joubert 2022).

This chapter is divided into two main components to facilitate the understanding of

explainability methods in Bayesian networks. Firstly, we explore fundamental concepts related to Bayesian networks, such as independence and inference, and decision problems in Bayesian networks. Section 2.2.4 provides an overview of existing Bayesian network software. Thereafter, considering that Bayesian networks are often recognised as inherently transparent models (Barredo Arrieta et al. 2020), and given that inherently transparent models offer explanations directly derived from their structure and reasoning process, we investigate the explanation of the model in Section 2.3 and explanation of reasoning in Section 2.4.

2.2 Overview of Bayesian Networks

More formally, a Bayesian network is a pair $\mathcal{B} = (\mathcal{G}r, \Theta)$, where $\mathcal{G}r$ consists of a directed acyclic graph whose nodes represent the random variables in the relevant universe. The arcs in $\mathcal{G}$ indicate the direct dependencies among variables. Θ represents the network parameters expressed as conditional probability tables. A key feature of Bayesian networks is the *Markov property*[1] (Korb & Nicholson 2010), which allows us to express the joint probability distribution for $\mathbf{V}$ as the product of conditional probability distributions for each variable V_i given its parents $\text{Pa}(V_i)$:

$$Pr(\mathbf{V}) = \prod_{i=1}^{n} Pr(V_i|\text{Pa}(V_i)). \tag{2.1}$$

2.2.1 Conditional independence in Bayesian networks

Independence among random variables within a domain is a fundamental concept in probability theory and serves as the basis for various areas of study (Dawid 1979), including probabilistic graphical models. Let's consider two events, $\mathbf{A}$ and $\mathbf{B}$. Events $\mathbf{A}$ and $\mathbf{B}$ are considered independent if observing $\mathbf{A}$ provides no additional information about $\mathbf{B}$ (Barber 2012). Mathematically, events $\mathbf{A}$ and $\mathbf{B}$ are considered independent (denoted as $\mathbf{A} \perp\!\!\!\perp \mathbf{B}$) whenever conditioning on one, say $\mathbf{B}$, leaves the probability of the other, $\mathbf{A}$ unchanged (Korb & Nicholson 2010):

$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \equiv Pr(\mathbf{A}|\mathbf{B}) = Pr(\mathbf{A}). \tag{2.2}$$

[1] each node is conditionally independent of its non-descendants given its parents

Furthermore, conditional independence between two events, **A** and **B**, given an additional event **C** can be defined as,

$$\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C} \equiv Pr(\mathbf{A}|\mathbf{B},\mathbf{C}) = Pr(\mathbf{A}|\mathbf{C}). \tag{2.3}$$

2.2.2 Inference in Bayesian networks

Bayesian networks are frequently used in practice due to their ability to reason under uncertainty (Barber 2012), which is facilitated by probabilistic inference (also referred to as belief updating). This task entails estimating the posterior probability distribution for a set of variables conditioned on the observed evidence (D'Ambrosio 1999, Korb & Nicholson 2010). Accordingly, the Bayesian network framework permits the conditioning of any set of variables, fostering various directions of reasoning. This conditioning is performed according to the 'flow of information' and is not limited to the direction of the arcs.

Performing inference in Bayesian networks

The Bayesian network framework allows us to directly perform inference queries using the distribution of our model (Koller & Friedman 2009). Two commonly used queries are *conditional probability queries* and *maximum a posteriori* (MAP) queries (Nagarajan et al. 2013). MAP queries entail determining the most likely instantiation of multiple variables, which we explore in Section 3.2.2. To illustrate conditional probability queries, i.e., $Pr(\mathbf{Y}|E = e)$, consider a medical practitioner observing a patient with a fever (evidence). Using a Bayesian network that includes variables like *fever*, *flu*, *infection*, and *common cold*, we can perform inference to calculate the posterior probability of each disease given the fever: $Pr(flu|fever = true)$, $Pr(infection|fever = true)$, and $Pr(common\,cold|fever = true)$. As such, we can investigate the effect of new evidence on the distribution of the model, using the knowledge encoded in the network.

There are two common types of evidence, *hard evidence* and *soft evidence* (Jensen & Nielsen 2007). If one or more variables are instantiated, we call it hard evidence; otherwise, it is referred to as soft (or virtual) evidence, i.e., we specify a probability distribution for the variables of interest that reflect our level of belief about the different states. Inference

can be performed using either exact or approximate methods. Regardless of the approach, the computational complexity of inference in Bayesian networks remains $\mathcal{NP}$-hard (Koller & Friedman 2009, Korb & Nicholson 2010).

Exact inference in Bayesian networks is often implemented using *variable elimination* or via *junction trees* (Nagarajan et al. 2013). The variable elimination algorithm starts by initialising factors for each node in the network, representing conditional probability distributions based on the network's structure and parameters. Thereafter, the algorithm proceeds to select an elimination order for the variables. The main step of the algorithm iteratively eliminates variables according to the chosen order. This choice significantly impacts the algorithm's efficiency (Darwiche 2009). For each variable in the elimination order, it multiplies all factors containing that variable and then sums out the variable, resulting in a new factor. This process continues until all variables are eliminated except for the query variables. Finally, the algorithm normalises the resulting factor to obtain the desired probability distribution. Another approach to inference involves transforming the network into a junction tree. This is achieved by clustering the nodes into cliques to reduce the network structure to a tree (Nagarajan et al. 2013). Once the junction tree is established, a Message-Passing algorithm Kim & Pearl (1983) is applied for inference.

Approximate inference algorithms in Bayesian networks are often employed when exact inference becomes computationally infeasible due to the complexity of the network. Two commonly used approximate inference methods are *Markov Chain Monte Carlo* (MCMC) and *variational inference* (Salimans et al. 2015). The MCMC algorithm generates samples from the joint distribution of variables in the Bayesian network, allowing the approximation of relevant conditional probabilities (Nagarajan et al. 2013). While variational inference (Xing et al. 2002) approximates the posterior distribution with a simpler, parameterised family of distributions.

`R` provides several packages to perform inference in Bayesian networks. The `gRain` (Højsgaard 2012) package includes the `setEvidence` and `querygrain` functions, which can be used to perform exact inference. Whereas the dbnR (Quesada 2022) package offers inference functions such as `exact_inference` and `approximate_inference`.

Reasoning patterns

There are four primary categories of reasoning: *diagnostic*, *predictive*, *intercausal*, and combined reasoning. Figure 2.1 provides a visual representation of these reasoning patterns. Diagnostic reasoning refers to inference performed in the opposite direction of the arcs, i.e., reasoning from *effects* to *cause*. In Figure 2.1a E is evidence for B. With predictive reasoning, reasoning occurs in the direction of the arcs, i.e., from *cause* to *effects* as new information becomes available. Figure 2.1b illustrates predictive reasoning where node **C** is our observed evidence. Intercausal reasoning is concerned with mutual causes of a common effect. Suppose we have two causes, **A** and **C**, of the effect, **B**, as shown in Figure 2.1c. Note that **A** and **C** are independent of one another unless **B** is observed. Suppose we observe **B** (the common effect) and **C** (one of the mutual causes). This new information explains the observed effect, **B**, which lowers the probability of the alternative cause, **A**. This type of reasoning captured the *explaining away* phenomenon, in which the effect is sufficiently explained by the confirmed cause, making the alternative cause less likely.

Since the framework allows for conditioning upon any set of variables, we do not restrict nodes to either query (hypothesis) or evidence nodes. Consequently, the above-mentioned reasoning patterns might not apply to all scenarios. Different scenarios may require a combination of reasoning patterns. Figure 2.1d illustrates the combination of diagnostic and predictive reasoning.

2.2.3 Decision problems in Bayesian networks

The Bayesian network framework allows one to move beyond drawing statistical inferences and facilitates decision-making under uncertainty, i.e., with limited information. When feasible, one would likely decide to search for more information. However, a common challenge decision-makers face is whether exploring new information is 'worth' it. The value of information (VOI), a concept introduced in economics (Raiffa & Schlaifer 1961) and related to decision theory, is a quantitative measure used to estimate the expected benefit of acquiring information before making a decision.

When assessing the value of information, an initial step involves defining a *value function*, which may be defined in terms of entropy, variance (Jensen & Nielsen 2007) or reward

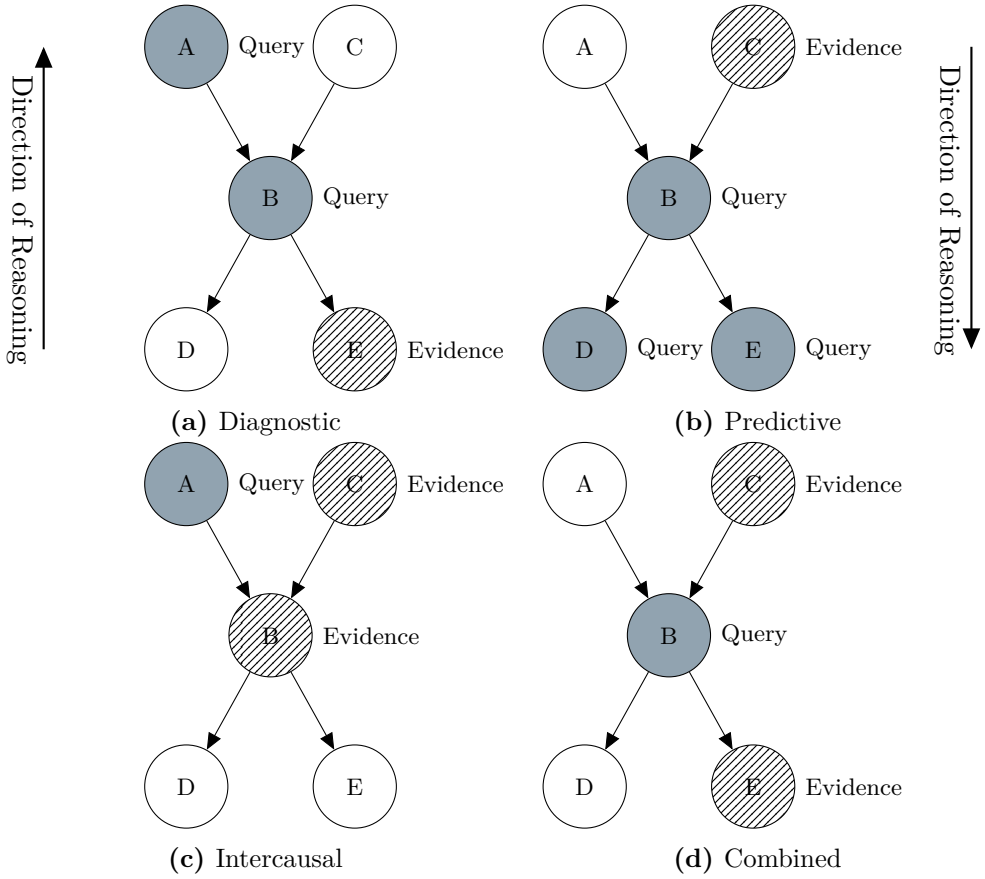

Figure 2.1: Types of reasoning in Bayesian networks (adapted from Korb & Nicholson (2010)).

(Krause & Guestrin 2009). Using a reward-based value function R, and given a hypothesis variable D, evidence e, and unobserved variables $\mathbf{H}$, we have an *expected reward*

$$\mathcal{ER}(R, D, \mathbf{H}, \mathbf{e}) = \sum_{h \in H} R(Pr(D|h, e))Pr(h|e), \tag{2.4}$$

since H is unobserved. The VOI (or *expected benefit*) of observing the variable $\mathbf{H}$ is then,

$$\mathcal{V}(R, D, \mathbf{H}, \mathbf{e}) = \mathcal{ER}(R, D, \mathbf{H}, \mathbf{e}) - R(Pr(D|\mathbf{e})), \tag{2.5}$$

where $R(Pr(D|\mathbf{e}))$ is the reward had we **not** observed the variables $\mathbf{H}$.

In situations where we need to choose which unobserved variable to observe next, limited to observing a single variable at a time, the *myopic approximation* is a popular choice (Jensen & Nielsen 2007).

2.2.4 Bayesian network software

Several software packages have been developed for building and evaluating probabilistic graphical models. These include BayesiaLab, BayesServer[2], HuginExpert, Netica[3], Elvira[4]. However, the majority of these tools are proprietary and, as such, provide users with limited open-source access.

In addition to these specialised tools, general-purpose languages, like `R` and `Python` provide dedicated packages for building and performing inference on probabilistic graphical models. For instance, `R` offers the `bnlearn` (Scutari 2010), `gRain` (Højsgaard 2012), and `bnstruct` Franzin et al. (2017), while `Python` boasts packages like `pgmpy` (Ankan & Panda 2015) and `pomegranate` (Schreiber 2018). However, these packages are not dedicated to providing explanations in Bayesian networks. For the remainder of this research, we will focus on implementations in `R` and build on top of existing packages.

Having gained an understanding of the foundational concepts underlying Bayesian networks, such as conditional independence and inference, we now focus on explainability methods that leverage the model's structure and reasoning process for direct explanation.

2.3 Explanation of the model

In essence, explanation of the model refers to the model's ability to provide an understandable summary of the relationships between variables in the domain. As such, it entails presenting the information in the knowledge base, which can be helpful in assisting application experts in the model construction phase and offering knowledge about the domain for instructional purposes. Recall that one of the goals of explanation is to provide insight into the workings of a model so that it can be better understood and improved if necessary. This aligns with *controlling* and *improving* AI systems. In the context of machine learning, explanation of the model may assist users in understanding how the model makes predictions and decisions, which can be useful for identifying biases or errors in the model. Within the broader landscape of XAI, the visual representation of decision trees directly aligns with the concept of explanation of the model as they depict the

[2]BayesServer:https://www.bayesserver.com/
[3]Netica: https://www.norsys.com/
[4]Elvira https://leo.ugr.es/elvira/

relationships between variables and their impact on the outcome.

Model explanations are considered static (Henrion & Druzdzel 1990) and, by definition, global since we are interested in explaining the model, i.e., the structure and the relation between variables, and not the reasoning process or some observed evidence. Generally, model explanations are visual representations and can be complemented with a verbal description of the nodes and arcs. For example, Henrion & Druzdzel (1990), Druzdzel (1993) proposed a method to translate the information contained in the network into natural language expressions. These explanations include phrases such as *impossible*, *very unlikely*, *unlikely*, *fairly likely*, *very likely*, and *certain* and are assigned based on probability ranges. For example, the expression *likely* is mapped to probabilities in the range $0.75 - 0.9$.

Given that Bayesian networks include a graphical representation (expressed as a directed acyclic graph $\mathcal{G}$), the most straightforward method of model explanation involves visualising this graphical model, allowing users to understand the nodes and the connections between them visually. In `R` , we can visualise the Bayesian network using the built-in `plot` function from `bnlearn` or through additional packages such as `Rgraphviz` (Hansen et al. 2023) and `visNetwork` (Almende B.V. and Contributors & Thieurmel 2022). Note that the `bnlearn` package incorporates some features of `Rgraphviz` for visualisation. This extends the visualisation capabilities by allowing users to visualise the Bayesian network in terms of the marginal probability distributions of each node. This type of plot provides a compact, visual summary that captures the structure and parameters of the network. This allows users to analyse how different inferences, i.e., evidence propagation scenarios impact the network's behaviour. Specialised Bayesian network software, like BayesiaLab, offers visual representations of network structures alongside conditional probability tables.

To gain a deeper understanding of the relationships between variables in a Bayesian network, we can visualise the arc strength. This metric measures the strength of the probabilistic relationship between nodes. Visualising these relationships helps us understand the variable influences at a glance, where the thickest arc represents the arc with the strongest strength of influence. Figure 2.2 depicts the Insurance Bayesian network (Binder et al. 1997) with associated arc strengths as measured by the Bayesian information criterion

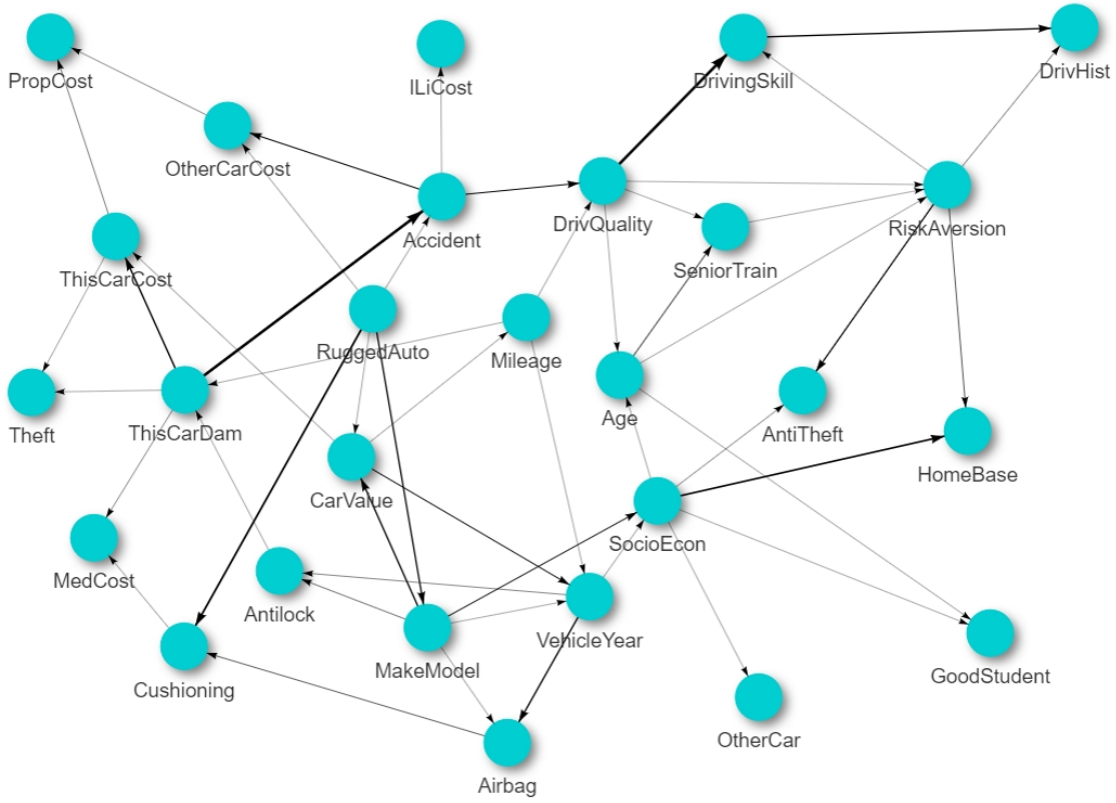

Figure 2.2: Graphical display of the arc strengths as measured by the Bayesian information criterion in the Insurance network from Binder et al. (1997).

(BIC). Please refer to Table B.1 in Appendix B for a description of the variables included in the Insurance network. During model construction, the arc strength can be used to build a network containing only significant arcs. This helps us create a more focused and interpretable model that captures the essential relationships between variables.

While Bayesian networks effectively represent the relationships between variables, their visual clarity may be compromised as the network grows in complexity. To improve readability, we can specify the layout using the `layout` argument in the `graphviz.chart` function from `bnlearn` – which incorporates features from `Rgraphviz`. The function offers basic layout options such as `dot`, `neato`, and `fdp`. `dot` positions nodes based on their topology (parents above, children below), while `neato` positions nodes based on an approximation of their path distance. `fdp` creates similar layouts as `neato`, but it prioritises keeping nodes further apart from one another. It is important to note that while the function offers some layout control, it does not currently allow for more intricate

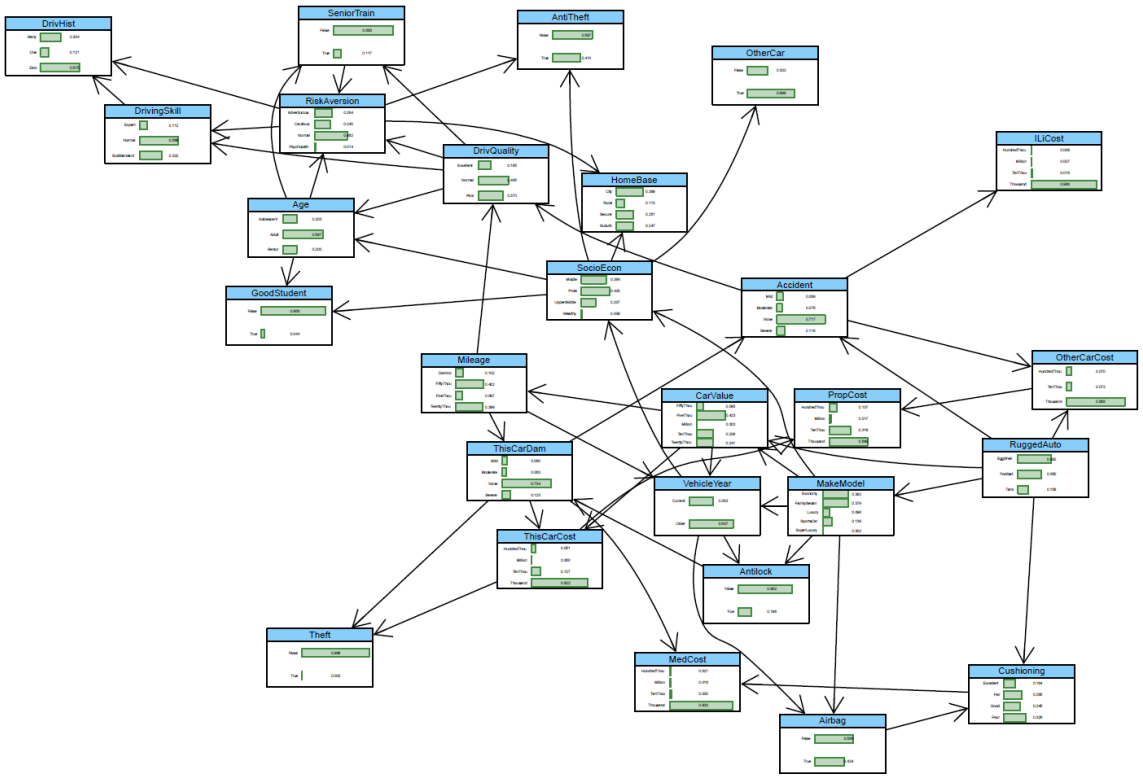

Figure 2.3: Graphical display of the marginal probabilities and the `fdp` layout from `Rgraphviz` .

customisation of fonts and font sizes. Figure 2.3 illustrates the `fdp` layout along with the marginal probabilities of the nodes in the Insurance Bayesian network. These basic layouts might not be sufficient for very complex networks where the graphical representation can become cluttered and difficult to interpret. To address this challenge, researchers have explored techniques for optimising the layout structure of these networks. For instance, Marriott et al. (2005) explores algorithms like horizontal layering and an extension thereof through an additional vertex coordinate assignment phase. These optimised layouts not only enhance visual clarity but also facilitate the process of understanding complex relationships within the network.

The transparency of Bayesian networks is further enhanced by their ability to show their probabilistic reasoning. The network structure encodes these relationships, allowing for explanations of how the model arrives at its conclusions based on probabilities. This leads us to the next explanation method: explanation of reasoning.

2.4 Explanation of reasoning

The overarching goal of explanation of reasoning aligns with one of the motivations for designing explainable models, which is to *justify* AI-derived conclusions. Here, explanation of reasoning aims to justify a particular instance in the network and how it was obtained, thereby allowing users to verify the results (Gallego 2005). Reasoning methods include extracting the chains of reasoning, measuring the impact of observed evidence (Kyrimi & Marsh 2016), counterfactual or contrastive explanations (Koopman & Renooij 2021), and scenario-based explanations (Druzdzel 1996). Since inherent transparency represents the model itself, explanation of reasoning can be considered in this category since the reasoning process is directly encoded within the network structure. In other words, the system enables users to directly explore "what-if" questions and perform contrastive explanations without the need for external assistance (Mittelstadt et al. 2019). Conversely, if we are interested in examining the reasoning process in more detail, explanation of reasoning would be considered post-hoc.

Within the broader XAI field, techniques like LIME (Ribeiro et al. 2016) and SHAP (Lundberg & Lee 2017) form part of explanation of reasoning, albeit post-hoc. For instance, SHAP employs an additive approach to explain individual predictions. Each feature receives a score reflecting its impact on the outcome, offering insights into the model's reasoning process.

2.4.1 Chains of reasoning

The chains of reasoning presented in Section 2.2.2 are encoded in the Bayesian network framework, allowing direct elementary explanation capabilities. Firstly, explanations through diagnostic reasoning may be "*E is evidence for B*". In other words, we are interested in finding an explanation to answer "*what went wrong*". Whereas explanations for predictive reasoning may be "*B may cause D*". Here, we are interested in questions concerning "*what will happen*" based on current conditions. Lastly, intercausal explanations may take the form "*A and C may each cause B; as C explains B, there is no evidence for A*". For example, a student received a low score on a test (observed outcome). Intercausal reasoning allows us to consider multiple causes, such as study time and test anxiety. Suppose the student mentions that they felt anxious before the test, then this observation

explains away the insufficient study time.

Henrion & Druzdzel (1990), Druzdzel (1993) introduced a qualitative analysis method implemented in the INSITE tool. The tool identifies reasoning chains linking evidence variables to target[5] variable. By examining how evidence influences variables in each chain, INSITE eliminates chains that impede evidence propagation. Whereas BANTER (Haddawy et al. 1997) selects those chains with the highest "strength" – the minimum impact of any variable within the chain. The selection is driven by analysing each variable's impact on the overall chain.

This type of explanation is a graphical display of the reasoning pattern and, therefore facilitates comprehension of the reasoning process. These explanations are inherently transparent and do not require expertise in probabilistic reasoning. That said, familiarity with evidence propagation methods can provide a deeper understanding.

2.4.2 Variable importance

Given some observed evidence, Bayesian networks use inference to make predictions on a variable of interest. A question that may arise here is "*which evidence variables supports or contradicts the prediction*"? In other words, we may be interested in finding explanations that indicate variable importance since not all evidence has an equal influence on the prediction. To facilitate this, we would need to determine the impact of an observed variable by analysing how the probability distribution of the variable of interest changes given the observed evidence. Suermondt's INSITE (Suermondt 1992) framework utilises the Kullback–Leibler divergence to quantify the difference between the posterior distribution of the variable of interest under the presence of all evidence and the distributions obtained by excluding specific evidence or subsets thereof. Whereas BANTER (Haddawy et al. 1997) quantifies the difference in the prior and posterior probability of the variable of interest based on each evidence variable. Madigan et al. (1997) evaluates the influence by continuously updating the *weight of evidence* (Good 1950) as each evidence variable is instantiated. Yap et al. (2008) propose a method, Explaining BN Inferences (EBI), that explains the prediction in terms of influential nodes and the variable of interests Markov blanket. These explanations promote transparency in the reasoning process. This allows

[5]Here, a target variable is the variable of interest.

users to understand how each piece of evidence affects the prediction.

2.4.3 Counterfactual and contrastive explanations

Counterfactual explanations have received significant attention from the XAI research community (Kenny & Keane 2021). Counterfactual explanations often take the form of "what-if" scenarios (Byrne 2016, Bica et al. 2021), where the input is changed. For example, "*A person is denied a loan because their credit score is too low. What if their credit score increased by 30 points? Would they then qualify for the loan?*" Accordingly, the decision is followed by a counterfactual statement (Wachter et al. 2017).

Pearl (2009) casts this in terms of probabilistic reasoning and Bayesian networks and defines a *counterfactual sentence* as "*Y would be y (in situation u), had X been x*". This can be evaluated using three steps: evidence propagation → action → prediction, where *evidence propagation* is based on the actual course of events, *action* refers to the formulation of the counterfactual and *prediction* to the computation of the counterfactual. Whereas Butz et al. (2024) investigate whether an actionable counterfactual explanation is perceived as a more useful explanation than a direct cause counterfactual explanation with a shorter chain.

According to Miller (2019), humans inherently prefer contrastive explanations. These explanations address why an alternative and preferred (or expected) prediction was not made instead and have been studied in the broader XAI research field (Lim & Dey 2009). As such, contrastive explanations clarify why the observed outcome r occurred instead of a different outcome r'. Koopman (2020) propose an algorithm that generates all explanations that are both contrastive and counterfactual to explain a particular prediction (target variable) using a Bayesian network. Focused only on contrastive explanations, Koopman & Renooij (2021) propose an algorithm for solving *pursuasive contrastive explanations* in Bayesian networks. Counterfactual and contrastive explanations provide a description of the reasoning process such that the user can understand the conclusion obtained.

2.4.4 Scenario-based explanations

Scenario-based explanations use a hypothetical situation, based on variables in the network, to illustrate how the network's reasoning process leads to conclusions based on evi-

dence. As such, scenarios can be envisaged as stories, each describing possible conditions (Parson 2008). Probabilistically, scenarios are represented by an assignment of values to relevant variables. These may be comprised of all variables or a subset thereof. In Druzdzel (1996), scenarios are extracted from the Bayesian network and are presented as a configuration of nodes in the network relevant to the prediction of the network. Vlek et al. (2015, 2016) proposed an approach to reasoning about legal evidence that merges Bayesian networks with scenario schemes, allowing for an integration of a narrative approach with a probabilistic approach. This framework allows for the construction of narrative explanations based on scenarios derived from the network. Furthermore, the authors introduce an approach for generating natural language scenario-based explanations as well as a format for alternative scenarios and their relation to the evidence.

Scenario-based explanations in Bayesian networks, while related, differ slightly from general case-based explanations in XAI. Where scenario-based explanations leverage the network structure and conditional probabilities to create scenarios, case-based explanations comprise various techniques, not necessarily specific to the model's internal structure, to create explanatory scenarios. Fundamentally, case-based explanations involve comparing a particular prediction to similar instances in the dataset and explaining the model's decision based on the outcomes of these similar cases. A notable research system is the CARES (Cancer Recurrence Support) System (Ong et al. 1997), which compares current and previous patient cases through case-based reasoning. Similar to counterfactual and contrastive explanations, scenario-based explanations describe the reasoning process. Here, explanations are provided in natural language with numerical probabilities.

2.5 Conclusion

As shown in this chapter, Bayesian networks offer a robust framework for modelling and reasoning under uncertainty. This chapter provided a brief overview of Bayesian networks, including concepts like conditional independence and inference, and how explanation methods (e.g., explanation of the model and reasoning) contribute to the network's inherent transparency. However, their inner workings can present challenges for intuitive understanding (Korb & Nicholson 2010). Acknowledging this limitation, researchers have developed a suite of post-hoc explainability techniques to make Bayesian networks more

interpretable.

While the methods discussed in this chapter concentrated on explaining the model or the reasoning process, these post-hoc techniques provide explanations about the domain using the Bayesian network. In the following chapter, we explore these post-hoc methods, detailing how these methods can be applied to explain observed evidence or decision-readiness in Bayesian networks.

Chapter 3

Post-hoc explanation in Bayesian networks

3.1 Introduction

Post-hoc explanations focus on *why* a model behaves in a certain way, rather than *how* it works (Mittelstadt et al. 2019). In this chapter, we explore post-hoc explanation in Bayesian networks and in particular, local explanation techniques which include explanation of evidence and explanation of decisions. The methods are considered local since we focus on explaining a specific instance of observed evidence or a particular decision.

The objective of explanation of evidence is to find the most likely configuration of variables that best explain the observed phenomena. Explanation of decisions calculates a confidence level in making a decision based on unobserved variables in the network. Essentially, it allows users to evaluate whether the decision would change if the true state of the variables were known. Should the decision confidence be low, it motivates a search for additional information. This involves selecting a variable from the set of unobserved variables for observation. While the literature refers to these variables as *hidden* variables, it is easy to confuse this terminology with the general statistical terminology for hidden variables, representing variables that are never observed (Elidan et al. 2000). Since these unobserved variables may be observed in a subsequent step, we can refer to these variables as *latent evidence variables*. For example, suppose a bank uses credit score and income for loan approvals, a high decision confidence indicates that knowing the applicants' debt-

to-income ratio would not have changed the outcome. Conversely, a low confidence level suggests the bank may benefit from considering the applicants' debt-to-income ratio. Here, the latent evidence variable is the applicants' debt-to-income ratio. Accordingly, we define the term *decision-readiness* to capture whether a user is ready to commit to a decision, given the information available.

The remainder of the chapter is structured as follows. Section 3.2 reviews explanation of evidence methods. In particular, we explore the most probable explanation (Section 3.2.2) and the most relevant explanation (Section 3.2.3). As our interest lies in finding computationally efficient explanations, we investigate existing algorithms, such as the forward search, to solve the most relevant explanation. Section 3.3 presents the proposed *explanation of decisions* category. Drawing on statistical decision theory, we explore concepts like stopping and selection criteria to support decision-readiness in the Bayesian network domain. This includes a brief illustration of the decision-readiness process. We then focus on the same-decision probability (Choi et al. 2012, Chen et al. 2012) in Section 3.3.2 and show how it can be used as both a stopping and selection criteria.

3.2 Explanation of evidence

Abductive inference, a concept derived from psychology and philosophy, involves reasoning and forming explanations based on uncertain information (Dew 2007, Peng & Reggia 2012). Within the Bayesian network domain, abductive inference refers to finding the configuration of variables that are most likely to explain the observed evidence (Gallego 2005, Gámez 2004). It is particularly useful in situations where more than one hypothesis can explain the observed phenomenon (Charniak & Shimony 1994) since abductive inference allows one to generate multiple hypotheses. Bayesian confirmation theory recognises the importance of considering alternative explanations and comparing these explanations with regard to the observed evidence. By providing multiple explanations, end-users are encouraged to explore different possible explanations that offer a broader perspective and allow for a comprehensive evaluation of evidence (Yuan, Lim & Lu 2011). By examining these alternatives, users gain deeper insight into how *good* the best hypothesis is or how sensitive the hypotheses are to parameter changes (Chan & Darwiche 2012). Anderson et al. (2020), Lim & Dey (2013) study the benefit of reasoning with multiple explanations.

Finding the most likely variable instantiation, given some observed evidence, has been a topic of interest for many years and is presented in various formats. For example, Pearl (1988) refers to finding the best variable instantiation given some observed variables as *belief revision*. While Shimony & Charniak (1990) refers to the problem as *Maximum A Posteriori*, Sy (1993) and Li & D'Ambrosio (1993) refer to it as *most probable explanations* – note that the majority of literature defines the most probable explanation (MPE) as a special case of maximum a posteriori (MAP). Other frequently used terms are *maximisation of a probabilistic expert system* (Dawid 1992). Yuan & Lu (2008) frames the problem as the most relevant explanation, which includes only the most relevant variables based on a relevance measure.

We define three variable types for explanation of evidence: *evidence*, *target*, and *intermediate* nodes. Evidence nodes represent observed evidence, this might be a test, a symptom, or even an error message displayed by a system. Target nodes are variables of interest, in other words, variables we would like to investigate that could provide a deeper understanding of the observed evidence, such as a patient's health states. The set of target variables form the hypothesis space. Intermediate, or auxiliary, nodes are those nodes in the network that are neither an observation nor a target variable. Depending on the set of target variables, abductive inference is presented in two variants: *total abduction* and *partial abduction* (Gámez 2004). If we are interested in finding an explanation based on a full set of target variables, i.e., an empty set of intermediate variables, we are interested in total abduction; otherwise, if we have a non-empty set of intermediate variables, we are interested in partial abduction.

Several approaches and algorithms (both local and approximations) have been proposed to generate a hypothesis capable of explaining the observed evidence (Santos Jr 1991, Seroussi & Golmard 1994, Park 2002). Explanation of evidence methods does not intend to predict future events. Instead, they reason backwards from observed evidence to identify the most likely circumstances that led to it. Though we explore both the most probable and most relevant explanations, our focus is on understanding the most relevant explanation. Mainly since MAP (and MPE) has been studied extensively in the literature (Koller & Friedman 2009, Korb & Nicholson 2010, Mengshoel et al. 2010, Castillo et al. 2012, Kwisthout 2013*b*). As such, our experiments will include a brute-force MAP

implementation along with the relevant MRE implementation, unless stated otherwise.

3.2.1 Running example

We illustrate the concepts discussed in this section with the *Insurance* Bayesian network developed by Binder et al. (1997). It is considered as a benchmark example to evaluate feature selection (Zeng et al. 2009, Broom et al. 2012) and structure learning algorithms (De Campos et al. 2003, Tsamardinos et al. 2006, Niculescu-Mizil & Caruana 2007). The network consists of 27 variables, with three designated output variables: *MedCost*, *ILiCost*, and *PropCost*, illustrated in Figure 3.1. Where *MedCost* refers to the cost of medical treatment, *ILiCost* the inspection cost, and *PropCost* the ratio of vehicle costs. The remaining variables are used to estimate the expected insurance claim cost for a policyholder. Please refer to Table B.1 in Appendix B for a description of the variables in the *Insurance* network.

By generating explanations, we can improve our understanding of attributes that contribute to observations such as high medical expenses, accidents or theft. Another example is the explanation of factors that influence premium calculations and claim approvals. Here, we illustrate how post-hoc explanations can be used to determine factors that led to the accident. For example, the network can be used to determine if the accident was due to attributes related to the driver, such as *SeniorTrain*, *DrivingSkill*, *RiskAversion*, etc. Another set of attributes worth investigating could be those related to the vehicle, such as *VehicleYear*, *RuggedAuto*, *Mileage*, etc. One could even consider a combination of these attributes to better understand the variables that contributed to the severity of the accident.

For this running example, we will use two observation variables, *Accident* and *RuggedAuto*. *Accident* represents the severity of the accident and *RuggedAuto* represents the ruggedness of the vehicle. Where the former takes states $\{None, Mild, Moderate, Severe\}$ and the latter takes states $\{Eggshell, Football, Tank\}$ – rugged vehicles, such as $Tank$, are often built with stronger materials and can withstand tougher driving conditions. Suppose we observe $\{RuggedAuto = Tank, Accident = Mild\}$. To assess the claim, the insurance company investigates three binary variables: *AntiLock*, *OtherCar*, and *Airbag*. *AntiLock* indicates if the vehicle has an anti-lock braking system installed, *Other-*

Car denotes whether a second vehicle was involved in the accident, and *Airbag* specifies if the vehicle is equipped with an airbag.

The presence of an anti-lock braking system in the vehicle may impact accident severity by preventing wheel lock-up during braking and maintaining vehicle control, thus potentially contributing to the mild nature of the accident. Understanding whether another vehicle was involved in the collision provide context about the accident dynamics and helps reconstruct the scenario, as multi-vehicle accidents can differ from single-vehicle incidents in terms of impact and severity. Additionally, the presence of airbags plays a role in mitigating injuries during an accident, offering insight into why the accident resulted in only mild consequences. By examining these variables, we gain an understanding of how safety features and accident dynamics interact to produce the observed outcome. Insights from this analysis can inform accident prevention strategies, promote the adoption of safety features like anti-lock braking systems and airbags, and guide policymakers and vehicle designers in enhancing vehicle safety standards and features.

3.2.2 The most probable explanation

Recall, from Equation 2.1, that a Bayesian network represents a distribution over the domain consisting of all possible variable instantiations. Essentially, MAP finds the configuration of the target set that maximises the posterior probability given the evidence. Suppose we have a set of n target variables $(X_1, X_2, \cdots, X_n)$, then MAP involves finding the variable instantiation such that $Pr(X_1 = x_1, X_2 = x_2, \cdots, X_n = X_n|E)$ is maximised, where E is the observed evidence (Korb & Nicholson 2010). The MPE, a special case of MAP, entails finding a full instantiation consisting of *all* target variables (Korb & Nicholson 2010, Helldin & Riveiro 2009).

Several search algorithms, such as best-first search (Marinescu & Dechter 2007), genetic algorithms (Mengshoel & Wilkins 1998), tabu search (Park & Darwiche 2004), ant colony optimisation algorithms (Guo et al. 2005), and modified max-product clique trees (Sun & Chang 2011), have been proposed to solve the MPE (and MAP) problem in Bayesian networks. Mengshoel et al. (2010) study various initialisation algorithms for generating initial explanations. Kwisthout (2013*a*) proposes an extension to MAP, named the *most inforbable explanation*, which integrates two fundamental properties of abduction: the

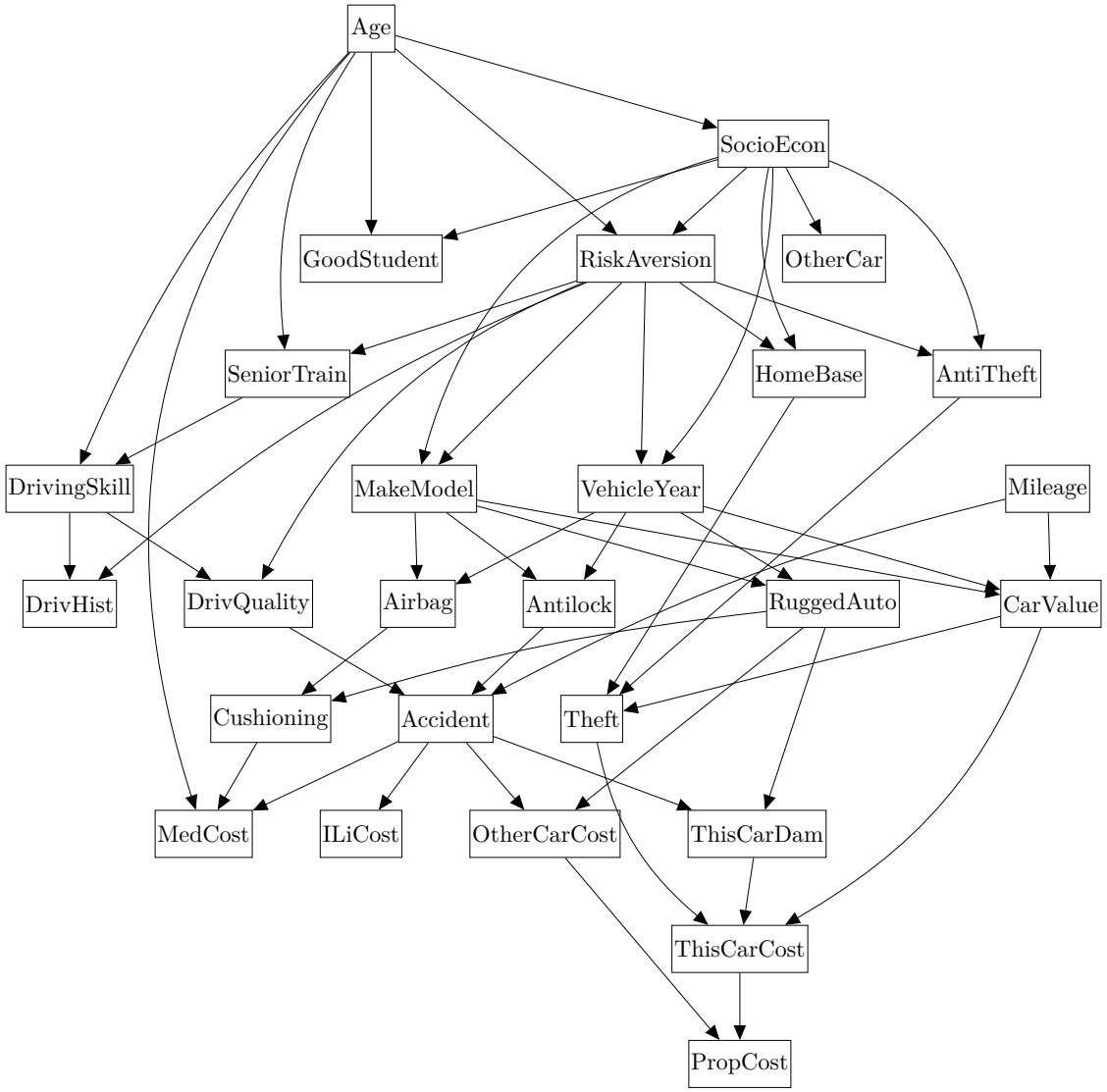

Figure 3.1: Graphical illustration of the Insurance Bayesian network from Binder et al. (1997).

selection of candidate hypotheses (and the determination of their granularity), along with the inference to the best explanation.

In principle, we can find the most probable configuration through a brute-force algorithm which generates the joint distribution and selects the configuration with maximum probability. However, this is intractable, as finding the most likely variable configuration, either through MAP or MPE, is shown to be $\mathcal{NP}$-hard (Shimony 1994, Abdelbar & Hedetniemi 1998), while Park (2002) extends this to show it $\mathcal{NP}$-complete. To illustrate the explanations obtained through the brute-force algorithm, consider the running example

with target nodes *AntiLock*, *OtherCar*, and *Airbag* and evidence nodes *RuggedAuto* and *Accident*. We can use the `setEvidence` and `querygrain` functions from the `gRain` package in `R` to compute the joint probabilities for all combinations. Table 3.1 presents the combinations of variable states and their associated joint probabilities, ranked from highest to lowest. Here, the *best* explanation for the observed *mild* accident of a *tank*-style vehicle is that the vehicle has no anti-lock braking system and no airbags and a second vehicle was involved in the accident. The second-best explanation is a vehicle with an anti-lock braking system and airbags were involved in an accident with a second vehicle. This allows users to consider various scenarios to explain the observed phenomena, effectively providing scenario-based explanations. Notice here that the most "unlikely" explanation, based on the three target variables, is a vehicle with an anti-lock braking system installed with no airbags and no other vehicle involved in the accident.

Table 3.1: MAP-generated variable instantiations for the Insurance running example.

AntiLock	OtherCar	Airbag	Joint Probabilities
False	True	False	0.335
True	True	True	0.243
False	True	True	0.198
False	False	False	0.127
False	False	True	0.056
True	False	True	0.036
True	True	False	0.004
True	False	False	0.001

3.2.3 The most relevant explanation

Since the explanations obtained from MPE consist of the full instantiation of target variables, the explanation may be *overspecified*. To address this, Yuan & Lu (2008) propose the most relevant explanation (MRE). In essence, MRE searches for and enumerates all possible partial instantiations of a set of target variables and finds the instantiations that maximise some relevance measure. Bayesian confirmation theory is a framework within Bayesian statistics and Philosophy of Science that addresses the problem of updating hypotheses when new evidence is presented. It offers a probabilistic manner to represent the degree of evidential support. One such probabilistic metric is the *Bayes factor* (Kass & Raftery 1995). Furthermore, according to Yuan, Lim & Lu (2011), the chosen relevance

measure should satisfy the properties of a good explanation, i.e., *conciseness* and *preciseness*. Proprietary software, such as BayesiaLab, offers functionality for most relevant explanation, but the `R` environment currently lacks this functionality.

Generalised Bayes factor

Given the assumption that not all target variables need to be included for a good explanation of observed evidence, the relevance measure should effectively prune irrelevant variables from the *best* hypothesis. *Bayes factor* (Kass & Raftery 1995) measures the strength of evidence among two competing hypotheses, i.e., two competing instantiations of target variables. Let's consider data D. We have multiple hypotheses, H_i, that could explain this data. Each hypothesis has a probability distribution, $Pr(D|H_i)$, representing the likelihood of the data occurring under that hypothesis. Suppose we are interested in comparing a specific hypothesis with one alternative hypothesis. Using Bayes's theorem, we obtain,

$$Pr(H_i|D) = \frac{Pr(D|H_i) \times Pr(H_i)}{Pr(D|H_1) \times Pr(H_1) + Pr(D|H_2) \times Pr(H_2)}, \tag{3.1}$$

so that,

$$\overbrace{\frac{Pr(H_1|D)}{Pr(H_2|D)}}^{\text{Posterior odds}} = \overbrace{\frac{Pr(D|H_1)}{Pr(D|H_2)}}^{\text{Bayes factor}} \times \overbrace{\frac{Pr(H_1)}{Pr(H_2)}}^{\text{Prior odds}}, \tag{3.2}$$

where the Bayes factor is given by,

$$B_{12} = \frac{Pr(D|H_1)}{Pr(D|H_2)}. \tag{3.3}$$

Therefore, the Bayes factor expresses the ratio between the posterior odds of H_1 to its prior odds, irrespective of the prior odds (Kass & Raftery 1995).

Similar to MPE and MAP, we are often presented with multiple possible hypotheses rather than just one hypothesis and its alternative. The generalisation of the Bayes factor allows for comparing multiple competing hypotheses (Fitelson 2001, Yuan, Lim & Lu 2011). The generalised Bayes factor (GBF), for observed evidence e and an explanation

x, is given by

$$GBF(x;e) = \frac{P(e|x)}{P(e|\bar{x})}\,, \tag{3.4}$$

where $\bar{x}$ denotes the set of all alternative hypotheses of x. The generalised Bayes factor penalises more complex explanations by considering the relative magnitude of variables, retaining only relevant variables in the explanation (Yuan & Lu 2008). Ranking the explanations by the generalised Bayes factor yields the most relevant explanation for the observed evidence. Therefore, the explanation x that maximises the generalised Bayes factor for the observed evidence e is considered the most relevant explanation. Mathematically, MRE is defined as

$$MRE(M;e) \equiv \underset{x,\emptyset\subset X\subseteq M}{\operatorname{argmax}}\, GBF(x;e)\,, \tag{3.5}$$

where M is the set of target nodes.

Multiple explanations: k-MRE

The solution space for the most relevant explanation contains all partial instantiations; two neighbouring explanations are connected if they have a local difference. In other words, both explanations have the same variable-state combinations except for one explanation having one less variable or the same variable with a single variable in a different state (Yuan et al. 2009). The solution space for three target variables of the Insurance network is illustrated in Figure 3.2. The lattice structure divides the nodes into layers. Nodes in layer 1 consist of singular instantiations, whereas nodes in the bottom layer consist of full instantiations.

To find all possible explanations, we can employ a brute-force search. Each explanation is ranked according to its generalised Bayes factor. Table 3.2 provides a summary of these explanations. Here, the MRE is a vehicle with an anti-lock braking system and an airbag with a GBF score of 1.657. Interestingly, the best explanation according to MAP $\{AntiLock = False,\ OtherCar = True\ Airbag = False,\}$ has a GBF score of 0.924 and is ranked 14^{th} according to MRE. Notice that the second-best explanation is a superset of the first explanation with a slightly lower GBF score. While the second explanation provides more details, it doesn't necessarily explain the observed phenomenon as effectively as the

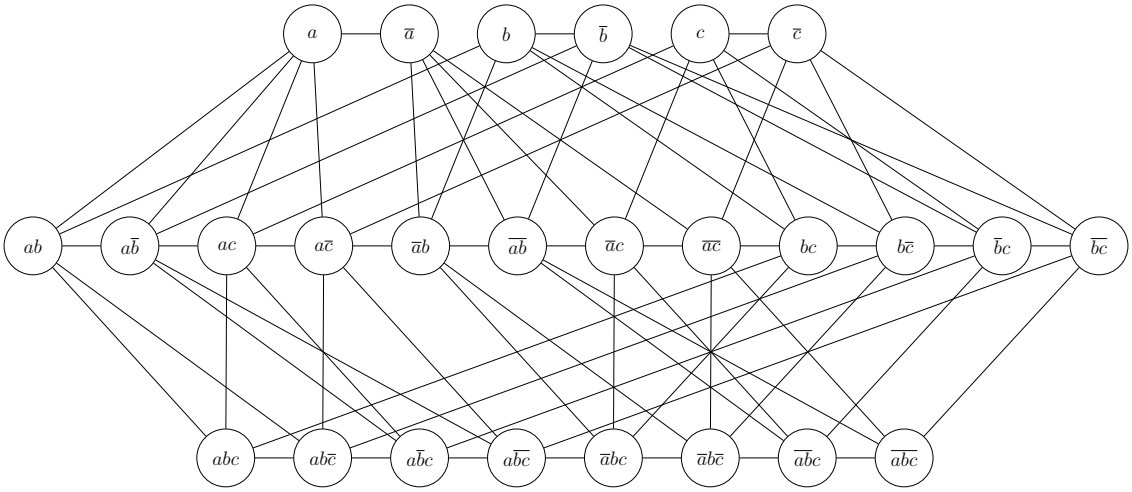

Figure 3.2: Solution space for the three target variables of interest, *Antilock* (A), *OtherCar* (B), and *Airbag* (C), in the Insurance (Binder et al. 1997) example. *Antilock*, *OtherCar*, and *Airbag* take states $\{True, False\}$, where state *False* is indicated as $\overline{a}, \overline{b}, \overline{c}$.

first. A concise yet diverse explanation set would be more insightful for understanding the observed evidence.

Table 3.2: Brute-force MRE-generated variable instantiations for the Insurance running example scenario.

	AntiLock	OtherCar	Airbag	GBF
1	True		True	1.657
2	True	True	True	1.655
3	True	True		1.613
⋮		⋮		⋮
14	False	True	False	0.924
⋮		⋮		⋮
26	True	False	False	0.237

Invoking Occam's Razor (Thorburn 1918), the set of explanations should be as simple as possible (Lötsch et al. 2022). Yuan, Lim & Lu (2011) propose filtering out explanations based on dominance relations such that the final explanation set is *minimal*. An explanation is considered minimal when there is no other explanation that either *strongly* or *weakly* dominates it. The remaining explanation set consists of k explanations that are diverse and representative. Computing these minimal explanations is computationally expensive since it involves an iterative comparison of each hypothesis, its neighbours, and all candidate hypotheses. As such, pruning the neighbourhood can improve the computational

efficiency.

Forward search to solve most relevant explanation

As illustrated previously, the solution space for MRE may be large since MRE is computed using all partial instantiations of a subset of unobserved variables. Several local and approximation methods have been developed for solving MRE (Yuan, Lim & Littman 2011). The inspiration behind these search methods stems from the similarity between MRE and feature selection, which aims to eliminate redundant and irrelevant characteristics from the set of features. The resulting subset comprises only relevant features (Chandrashekar & Sahin 2014). However, instead of only selecting features that are most relevant, MRE also entails the selection of the states of those features that will maximise the generalised Bayes factor. Therefore, the solution space for MRE will be more extensive than that of feature selection techniques. Yuan, Lim & Littman (2011) propose adapting existing feature selection techniques to solve MRE. Although various feature selection techniques exist, such as *forward* and *backward* search, we will focus on the forward search algorithm.

In essence, one or more starting solutions are invoked to initiate the forward search. For each initial solution, the solution is improved by either *adding* an additional feature or by *changing* the state of an existing feature in the solution (Yuan, Lim & Littman 2011). The former is defined as *add-one* neighbours and the latter as *change-one* neighbours. Although there are two ways to initialise a starting solution namely *empty initialisation* and *best pivot*, we focus on the best pivot starting solution. Here, we set the target features to their most likely state as a starting point. Figure 3.3 illustrates the forward search for three target variables in the Insurance Bayesian network (Binder et al. 1997). In particular, it shows the search path for *OtherCar*. The algorithm is given in Algorithm 1[1].

Exhaustive search algorithms are computationally feasible only for low-dimensional models. While the forward search algorithm is a helpful tool for exploring and identifying relevant features, it faces a notable challenge when applied to high-dimensional data (Meinshausen & Bühlmann 2006), particularly in the Bayesian network domain. As the complexity increases, the number of potential solutions that could explain the observed

[1] Algorithm adapted from Yuan, Lim & Littman (2011)

Algorithm 1 Forward-search algorithm

Input: Bayesian network $\mathcal{B}$, set of evidence variables E, and a set of target variables X.
Output: k-MRE solution.

Initialise the starting solution set I with the best pivot initialisation rule.
Initialise the current best solution, $y_{best} = \emptyset$ for each starting solution s
for each starting solution s in I **do**
    $y = s$
    repeat
        Find the neighbouring solution set N of y by either *changing* the state of a single variable or by *adding* an additional target variable with *any* state.
        Compute the GBF score for each solution in N.
        Filter the neighbouring solution set N based on dominance relations.
        Update y if the best solution in N yields a higher GBF score.
    until y stops updating
    if $GBF(y) > GBF(y_{best})$ **then**
        $y_{best} = y$
    end if
end for
return y_{best}

evidence grows rapidly. As a result, finding the optimal solution can be challenging, especially if the algorithm is not able to prune the solution space efficiently. We propose incorporating a statistical neighbourhood selection method, such as the Graphical Least Absolute Shrinkage and Selection Operator (graphical Lasso or gLasso) (Friedman et al. 2008), to prune the solution space. We present this in Chapter 4 and for the rest of this chapter turn our attention to methods that facilitate decision-readiness.

3.3 Explanation of decisions

Typically, Bayesian networks represent the relevant universe for a particular problem in which we have observed some evidence and want to draw inferences about the probability distribution of some other set of variables. Instead of focusing on statistical inferences, Wald (1949) proposed another framework, namely *statistical decision theory*, which is concerned with *statistical action*. Within the decision-making context, one would typically need to choose one action from a set of possible actions. Each potential action would lead to one of several outcomes, each of which is associated with a user preference. Decisions for a reasonable course of action are often made based on incomplete information since additional information may not be readily available, or be expensive to come by (Kochen-

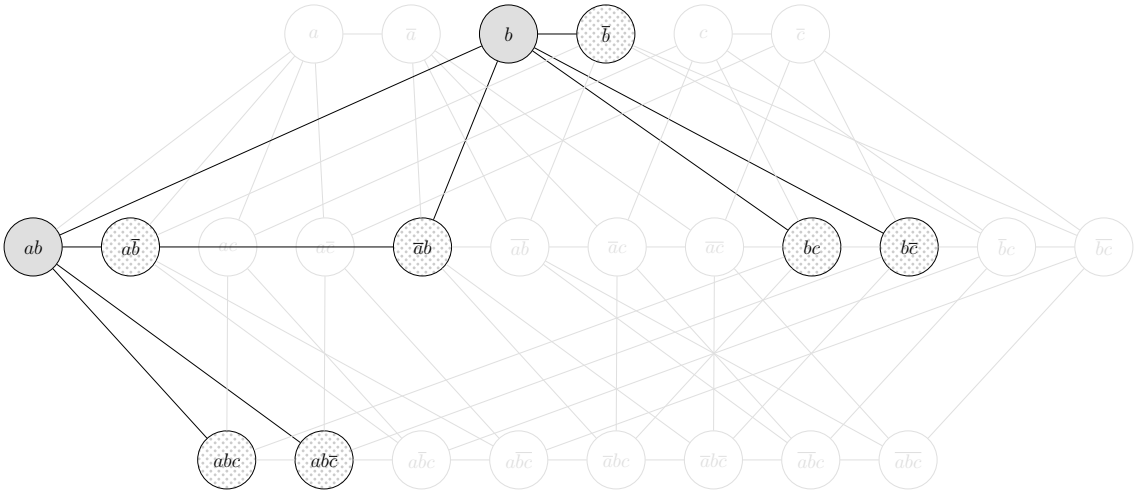

Figure 3.3: Illustration of the search path for *OtherCar* (B) through the forward search algorithm.

derfer 2015). This triggers an assessment of whether we have enough information to make an informed decision. If not, we need to identify what additional information is required to support an informed decision. In this context, we will *stop* information gathering if we can make an informed decision; otherwise, we need to identify and *select* the additional information required for informed decision-making. This aligns with a key motivation for explainable models: the ability to *discover* new information.

Given the discussion of Bayesian networks as inherently transparent models as well as existing post-hoc explanation techniques, statistical decision theory has been neglected. This is echoed by Främling (2020), who investigates Decision Theory notions in the broader XAI research field. Furthermore, generating explanations for decisions, especially under conditions of uncertainty, can be difficult due to the lack of a formal methodology for the treatment of important – potentially unobserved – variables (Guidotti et al. 2018). As such, we will explore the current state of decision theory, particularly focusing on *stopping* and *selection* criteria.

3.3.1 Concepts in statistical decision theory

As noted earlier, decisions are often made under conditions of uncertainty, and if feasible, one would typically seek out additional information. A common challenge for decision-makers is assessing whether the utility or benefit of additional information outweighs the cost, albeit monetary or otherwise, of the new information (Petitet et al. 2021). Conse-

quently, to make an informed decision, one would need to calculate the potential benefit of acquiring the new information. This requires quantifying the potential gains or losses, i.e., the utility of the decision outcome and comparing it with the cost of this new information. Bayesian decision theory (Savage 1972) provides a framework for calculating the expected value of information.

There are two primary types of decisions in the domain of statistical decision-making: *test decisions* and *action decisions* (Jensen & Nielsen 2007). The former refers to decisions to look for more evidence, while the latter refers to decisions that aim to change the state of the world. Two questions of interest typically stem from test decisions; "*given the available information, are we ready to make a decision?*" and "*if we are not yet ready to make a decision, what additional information do we require to make an informed decision?*" The first question relates to the *stopping criteria* of the decision-making process. Accordingly, the second question is associated with the *selection criteria*; when the stopping criteria are not met, one would need to acquire additional information to make a decision (Chen et al. 2012). The VOI (Raiffa & Schlaifer 1961), introduced in Section 2.2.3, can be used as a *selection criteria*. Stopping and selection criteria are not isolated techniques but rather integral parts of a broader decision-making framework. Various methods have been proposed in the literature to address either the stopping or selection criteria in decision-making.

Beyond these queries, exploring the potential impact of an unobserved variable presents a further opportunity in the realm of explanation of evidence. What if, upon observing a variable, the explanation obtained from MAP or MRE changes? In other words, how robust are the explanations obtained from MRE? Are these explanations sensitive to new explanation sets? Additionally, can we identify a subset of features that can *sufficiently* explain the decision while decreasing the impact of *irrelevant* features? That is, sufficient to provide strong probabilistic assurances that the model will exhibit similar behaviour even when all features are observed. The latter is out of scope for this research and is reserved for future research.

Stopping criteria in decision-making

Suppose a medical practitioner examines a patient with symptoms suggesting multiple possible diagnoses. Each diagnostic test offers valuable information but also costs time and resources. *Stopping criteria* acts as a guideline to help the practitioner determine when they have gathered sufficient information for making a confident decision (Saad & Russo 1996). Common stopping criteria in statistical decision theory include threshold-based criteria (Pauker & Kassirer 1980, Djulbegovic et al. 2015), resource-based criteria (Wang et al. 2015), and performance-based criteria (Zhu et al. 2010). In this research, we will focus on threshold-based criteria.

The concept of threshold-based notions is closely related to the idea of statistical action – when should the decision-maker act? This embodies decision-theoretic rationality, which suggests that the most rational course of action is to proceed when the expected benefits outweigh the expected harms. In other words, the decision-maker will commit to a decision once their belief about the event surpasses some predetermined threshold. These thresholds may be set based on, for example, user preference, expert domain knowledge, expected utility, information gain, computational analysis, or a combination of these.

For instance, in some clinical diagnosis models, thresholds depend on disease versus utility definitions and decision-maker preferences (Djulbegovic et al. 2019). Lu & Przytula (2006) defines probabilistic thresholds for multiple fault diagnosis. Decisions based on predetermined thresholds are sensitive to changes in the threshold. Renooij (2018) study the effect of changes in the threshold on decisions. Whereas Van Der Gaag & Bodlaender (2011) investigates, given the current observed evidence, the potential that future evidence may render another decision. Focusing on decision stability under uncertainty, Van Der Gaag & Coupé (1999) explored the robustness of Bayesian network outputs for threshold-based decision-making. The authors developed a sensitivity analysis method for computing bounds to which the network conditional probabilities can be changed while still resulting in the same decision.

Selection criteria in decision-making

Again, consider the medical practitioner examining a patient presenting symptoms indicative of various potential diagnoses, with each diagnostic test providing valuable informa-

tion. However, considering that these diagnostic tests usually involve some cost, whether monetary or otherwise, one would typically prioritise the tests that offer the greatest value. *Selection criteria* is then used to determine which variables should be selected for observation. The decision-theoretic framework provides a metric for measuring the value associated with making a particular observation (Koller & Friedman 2009). The VOI, a concept introduced in economics (Raiffa & Schlaifer 1961) and related to decision theory, is a quantitative measure used to estimate the expected benefit of acquiring information before making a decision. This can be done in two ways: making a single observation at a time or multiple observations at a time.

First, consider the case where we can select at most one observation at a time among a set of possible observations. In this case, we can compute the *myopic* value of information for each observation (Dittmer & Jensen 1997). Still, according to Koller & Friedman (2009), not all information is necessarily of value; information lacks value if it fails to change the optimal decision. This is related to the second question in test decisions, *if we are not yet ready to make a decision, what additional information do we require to make an informed decision.* For instance, whether we already have some observed evidence or not, an additional test should only be performed if it will change the diagnosis and improve our confidence in that decision.

Now, consider a more complex scenario in which multiple observations are made simultaneously. Here, we face the problem of which subset of variables to observe. If we have a set of m possible variables, the number of possible observation subsets is exponentially large, i.e., 2^m (Koller & Friedman 2009). Instead of observing all variables simultaneously, we can sequentially approach the problem, adding one observation at a time. Yet, the optimal choice of the next observation generally depends on the outcome of the previous selection. A common approximate approach is using the myopic value of information (Jensen & Nielsen 2007) discussed previously. Another solution is to extend the Bayesian network to an influence diagram, although this may increase the complexity of the model, which can influence the transparency thereof. Krause & Guestrin (2009) presents an algorithm for selecting observations in probabilistic graphical models.

Decision-readiness cycle

To illustrate the decision-readiness cycle, consider the flowchart depicted in Figure 3.4. We start the process by defining our decision, initial evidence, latent evidence variables, and decision threshold. Thereafter, we use a threshold-based stopping metric to determine whether the current evidence is enough to make a confident decision. One such threshold-based metric is the same-decision probability (Choi et al. 2012). This step will tell us whether the decision is likely to stay the same or change had we observe the latent evidence variables. If the stopping criteria are met, we can stop information gathering and commit to a decision since the decision is less likely to change even if we have observed the latent evidence variables. If not, we proceed with information gathering using the selection criteria. This involves computing the expected benefit of observing each of the latent evidence variables. We include the latent evidence variable that will, on average, lead to a more robust decision. Since discrete variables in Bayesian networks consist of at least two states, we include the state that maximises the threshold-based metric as evidence. This leads to updated evidence and latent evidence variable sets. The expected benefit (VOI) computation includes the threshold-based stopping criteria for the updated evidence – refer to Equation 2.4. As a result, we can use this to determine whether we now have sufficient evidence to make a decision. If not, we repeat the selection criteria process until we can commit to a more robust decision. Consequently, the decision-readiness cycle allows us to determine whether we have enough information to commit to a decision in light of incomplete information.

To put this in context, consider a baseline patient form. The form contains information on patient demographics, social circumstances, medical history, immunisations, current symptoms, exposure status, and clinical examination. Assume we have a Bayesian network for this along with nodes that represent the probability of a patient having a particular disease. Suppose we select the *tuberculosis* node as the decision node and enter current symptoms, exposure status, and clinical examination results as evidence. The patient withheld certain information on their social circumstances, medical history, and immunisations. These variables which the patient did not disclose are then seen as our latent evidence variables. For simplicity's sake, we will see these as three variables. However, in practice, there will be more variables to represent different social circumstances,

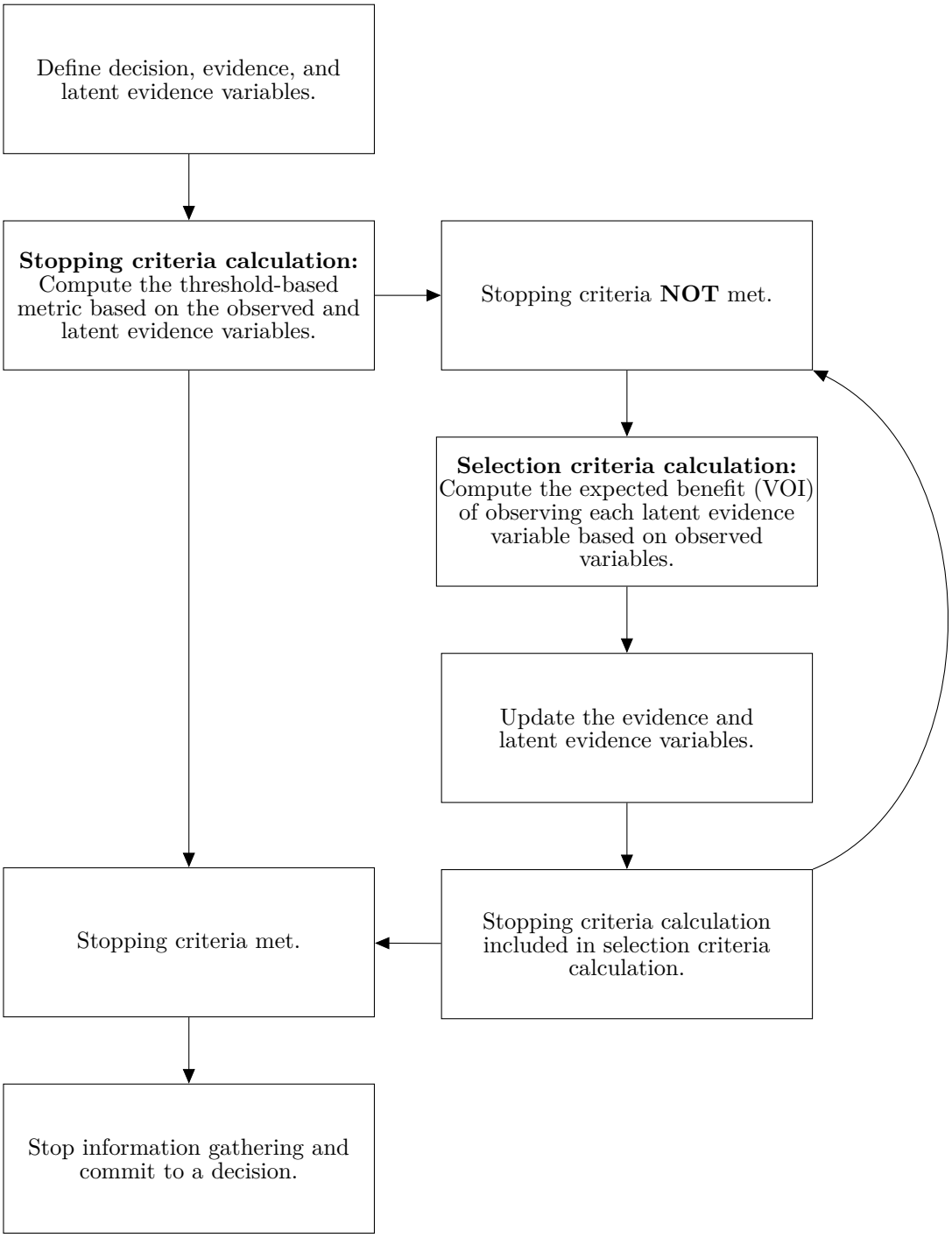

Figure 3.4: A flowchart for decision-readiness.

medical history, and immunisations. We proceed to compute the stopping criteria based on the current observed evidence and latent evidence variables. If this criteria is not met, we proceed with the selection criteria calculation. Suppose this highlights that, on av-

erage, observing the patient's medical history will lead to a more robust decision. Since the selection criteria use the expected benefit, it includes the stopping criteria calculation. This allows us to update the evidence to now include the patient's medical history. Of course, this means that the set of latent evidence variables is reduced to the patient's social circumstances and immunisations. Suppose that, even with the updated evidence the stopping criteria are not met. In this case, we will repeat the selection criteria process to determine which latent evidence variable to observe next since we use a myopic approach. This process is repeated until we are confident that the decision will not change had we observe the remaining latent evidence variables.

3.3.2 Same-decision probability

The same-decision probability (SDP), introduced by Choi et al. (2012), is a threshold-based confidence measure for decision-making with probabilistic graphical models. Suppose we want to make a decision d given some observed evidence e based on a threshold T. The decision is confirmed by $Pr(d|e) \geq T$. Then, the SDP can be defined as

$$SDP(d, e, H, T) = \sum_{h} [Pr(d|e, h) \geq T] Pr(h|e), \tag{3.6}$$

where H is a set of latent evidence variables and an indicator function $[Pr(d|e, h) \geq T]$ described by

$$[Pr(d|e, h) \geq T] = \begin{cases} 1 & \text{if } Pr(d|e, h) \geq T \\ 0 & \text{otherwise.} \end{cases}$$

Hence, SDP represents the expected probability that we would make the same decision even if we were to observe the latent evidence variables. Consequently, we treat SDP as a robustness measure for decision-making under uncertainty. Computing the same-decision probability is proven to be PP^{PP}-complete (Choi et al. 2012). Accordingly, (Chen et al. 2014) propose an approximate algorithm based on variable elimination. However, for this work, we use a brute-force algorithm that enumerates all possible instantiations. Next, we consider the same-decision probability as a stopping criterion for decision-making.

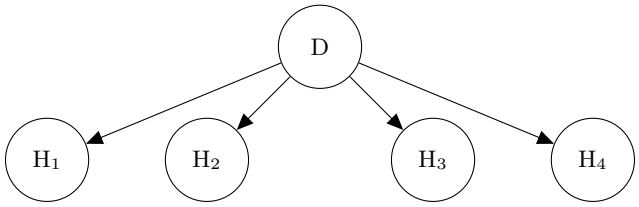

Figure 3.5: A naïve Bayesian network with a hypothesis variable D and four features $H_1, \cdots, H_4$.

Table 3.3: The conditional probability tables associated with the naïve Bayesian network in Figure 3.5.

D	H_1	$Pr(H_1\|D)$
$+$	$+$	0.55
$+$	$-$	0.45
$-$	$+$	0.45
$-$	$-$	0.55

D	H_2	$Pr(H_2\|D)$
$+$	$+$	0.55
$+$	$-$	0.45
$-$	$+$	0.45
$-$	$-$	0.55

D	H_3	$Pr(H_3\|D)$
$+$	$+$	0.60
$+$	$-$	0.40
$-$	$+$	0.40
$-$	$-$	0.60

D	H_4	$Pr(H_4\|D)$
$+$	$+$	0.65
$+$	$-$	0.35
$-$	$+$	0.35
$-$	$-$	0.65

Stopping criteria

By definition, SDP provides us with a confidence measure for decision-making. If we have a high SDP, we can confidently make a decision based on the available evidence since the likelihood of our decision changing based on additional information is low. Hence, SDP can be used as a stopping criterion to determine whether we have enough information. To illustrate this, we will consider two examples. The first reflects the work by Chen et al. (2012, 2014), while the second applies the same-decision probability to the Asia (Lauritzen & Spiegelhalter 1988) Bayesian network.

Consider the Bayesian network in Figure 3.5, with hypothesis variable D, where $Pr(D = +) = 0.5$, and four feature variables whose readings may affect our decision. Table 3.3 gives the conditional probability tables. Suppose we commit to a decision when $Pr(D = +|e) \geq 0.55$.

Suppose we observe $H_1 = +$ and $H_2 = +$, then the set of latent evidence variables consists of H_3 and H_4. Using `setEvidence` and `querygrain` in `R`, we can compute the hypothesis probability $Pr(d|e) = 0.599$. Since this is greater than the threshold, we can

Table 3.4: Scenarios for the latent evidence variables for the naïve Bayesian network in Figure 3.5.

H_3	H_4	$Pr(h\|e)$	$Pr(d\|h,e)$
+	+	0.290	**0.806**
−	+	0.240	**0.649**
+	−	0.230	0.547
−	−	0.240	0.349

conclude that our computed belief confirms the decision. However, this decision is made without the consideration of the latent evidence variables. Observing these variables may contradict the decision. Therefore, we would calculate the same-decision probability as a confidence measure to confirm our decision. Table 3.4 provides the probabilities, $Pr(h|e)$ and $Pr(d|e,h)$, for the various scenarios. Our SDP for this scenario is $0.290+0.240 = 0.53$, which indicates that there is a 47% chance that we would make a different decision had we observed H_3 and H_4. As a result, we should not yet commit to a decision but rather continue with information gathering.

Consider now the Asia Bayesian network in Figure 3.6 from Lauritzen & Spiegelhalter (1988). The network consists of eight variables, *visit to Asia* (A), *smoking* (S), *tuberculosis* T, *cancer* (C), *bronchitis* (B), *tuberculosis or cancer* (P), *abnormal x-ray* (X), and *dyspnoea* (D). Suppose we commit to a decision when $Pr(C = yes|e) \geq 0.6$, with evidence $S = yes$ and $X = yes$. While $Pr(C = yes|e) = 0.646 > 0.6$, the patient may withhold some information, such as a recent visit to Asia. The true state knowledge of A may confirm or contradict our decision. Using A as the latent evidence variable, the same-decision probability is 0.988, indicating that even if the patient had disclosed this information, there is still a 98.8% chance that we would make the same decision.

Selection criteria

Recall the same-decision probability for the example based on the naïve Bayesian network in Figure 3.5 indicated that there is a 47% chance that we would make a different decision if we had observed the two latent evidence variables. In this case, we need to decide which variable(s) to observe next such that we can make a more informed decision. As pointed out in Section 3.3.1, the selection criteria allows for a choice between observing one variable at a time or observing multiple variables simultaneously. We will proceed

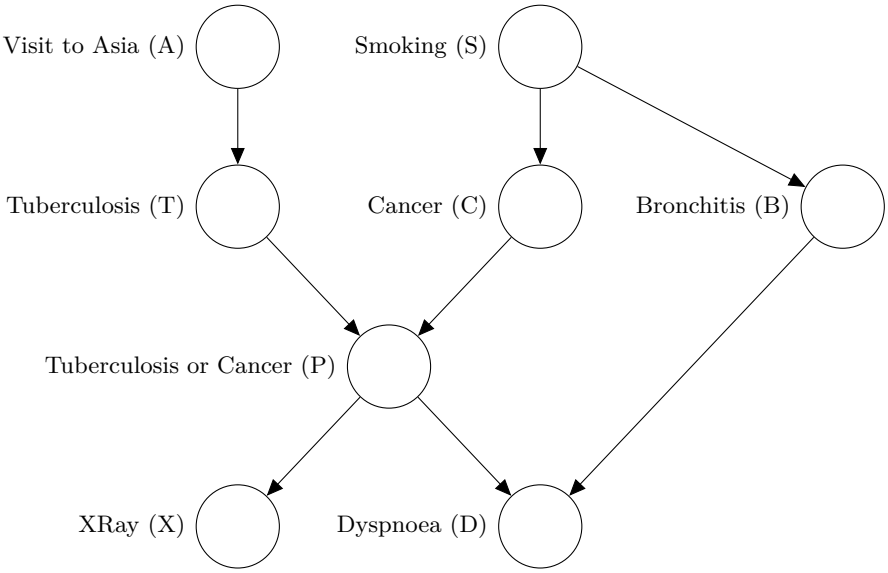

Figure 3.6: The Asia Bayesian network from Lauritzen & Spiegelhalter (1988).

with a myopic approach, concentrating on observing one variable at a time. Since we have two latent evidence variables, H_3 and H_4, we need to determine the expected benefit of observing each of these variables.

From Section 2.2.3 and 3.3.1, it follows that the VOI can be used as a selection criteria. Chen et al. (2012) propose defining the SDP as the reward function in VOI. We can then rewrite Equation 2.5 to obtain the *SDP gain* of observing variables G out of latent evidence variables H,

$$\mathcal{G}(G) = \mathcal{E}(G, H, e, T) - SDP(d, H, e, T), \tag{3.7}$$

where $SDP(d, H, e, T)$ is the SDP over latent evidence variables H and, from Equation 2.4, the expected SDP (also referred to as decision robustness) is given by

$$\mathcal{E}(G, H, e, T) = \sum_{g} SDP(d, H\ G, ge, T) \cdot Pr(g|e). \tag{3.8}$$

SDP gain serves as a selection criteria to prioritise variables that, on average, lead to the most robust decision given the observed phenomena.

Let us now compute the SDP gains for $\mathcal{G}(H_3)$ and $\mathcal{G}(H_4)$. Observing H_3 will give us an SDP of either 0.557 (if we observe $H_3 = positive$) or 0.5 for an expected SDP of 0.53, whereas observing H_4 will give us an expected SDP of 1. Therefore, $\mathcal{G}(H_3) = 0$ and $\mathcal{G}(H_4) = 0.47$. Hence, observing H_4 will, on average, allow us to make a more robust

decision that is less likely to change due to additional information.

3.4 Conclusion

This chapter highlights the theoretical components of post-hoc explanations in Bayesian networks that facilitate abductive inference and decision-readiness. Explanation of evidence methods is particularly useful in situations where we want to explain the observed evidence. This allows us to understand the factors that influence it. By leveraging this knowledge, we can improve existing measures, develop preventative measures, inform policymaking, and ultimately, drive positive user-centric outcomes. Explanation of decision methods facilitates decision-readiness. This is useful in situations where we do not have immediate access to all the necessary information to make a decision. The decision-readiness process informs the user whether they can make a decision based on the current available evidence. If not, it assists the user in identifying the information required to make a robust decision.

While this chapter has explored post-hoc explanations in Bayesian networks, a key challenge remains: efficiency. Computing explanations can be computationally expensive. The next chapter introduces an approximate approach to solving the most relevant explanation in Bayesian networks to address this challenge. The next chapter highlights the statistical contributions of this research: a computationally efficient algorithm for generating explanations that retain the characteristics of a good explanation. Additionally, we investigate the dynamic nature of explanations obtained through the MRE.

Chapter 4

Forward-gLasso search for solving the most relevant explanation

4.1 Introduction

Given that the solution space for the most relevant explanation can consist of full and partial instantiations, it becomes computationally infeasible for moderate to larger Bayesian networks. As such, it makes sense to prune the solution space to a subset of instantiations that could explain the observed evidence. By doing so, we can ensure that the current best solution is compared to a smaller, more concise, set of neighbours. To achieve this goal, we propose leveraging the gLasso algorithm. Integrating gLasso into the forward search algorithm is a symbiotic, approximate approach that combines the strengths of both methods. The forward search algorithm systematically explores the solution space, while gLasso helps identify the most relevant dependencies. Moreover, the sparsity induced by gLasso allows us to identify and select a reduced set of approximated relevant instantiations that exhibit dependencies, thus reducing the computational burden and enhancing the explainability of the results.

This chapter proposes a novel approximation algorithm, forward-gLasso, developed to generate computationally efficient explanations for observed evidence in Bayesian networks. The remainder of this chapter is structured as follows. Section 4.2 first explores the theoretical properties of the gLasso. After that, we incorporate the gLasso algorithm into the search strategy of the forward search algorithm. Here, we discuss the neighbourhood

selection process and the precision matrix estimation. Section 4.3 describes the experimental design while Section 4.4 provides the experimental results of the forward-gLasso algorithm compared to the forward search algorithm. Lastly, in Section 4.5, we conduct an experiment to test the most relevant explanation's sensitivity to additional evidence. This experiment demonstrates the dynamic nature of the most relevant explanation.

4.2 Graphical Lasso

Several researchers have explored using L_1 regularisation in estimating sparse undirected graphical models (Meinshausen & Bühlmann 2006, Banerjee et al. 2008, Dahl et al. 2008, Friedman et al. 2008). These models employ undirected graphs to define the conditional independence relationships between the variables. The default model assumes a multivariate Gaussian distribution. Consider the matrix $X_{n\times p}$, with n observations from p features with mean μ and covariance matrix Σ. Note that a zero element in Σ^{-1} indicates conditional independence between the two variables. Accordingly, sparser graphs are yielded when zero off-diagonal elements in Σ^{-1} increases. Define the precision matrix as $\Theta = \Sigma^{-1}$ and let S denote the empirical covariance matrix, with $S = \frac{1}{n}X^TX$. The gLasso problem is then defined as the maximisation of the penalised log-likelihood over non-negative definite matrix Θ,

$$\log \det\Theta - \text{tr}(S\Theta) - \lambda||\Theta||_1, \tag{4.1}$$

where $||\Theta||_1$ is the L_1 norm, λ is a tuning parameter controlling matrix sparsity, and tr represents the trace of a matrix.

The gradient for Eq. 4.1 is given by,

$$\Theta^{-1} - S - \lambda \cdot \text{sign}(\Theta) = 0. \tag{4.2}$$

Since $W = \Theta^{-1}$, we have,

$$W - S - \lambda \cdot (\Theta) = 0. \tag{4.3}$$

Graphical lasso solves the optimisation problem given in Equation 4.3 using a block-

coordinate descent. Consequently, the upper-right block of Equation 4.3 is,

$$w_{12} - s_{12} - \lambda \cdot \text{sign}(\theta_{12}) = 0. \tag{4.4}$$

Using the relationship $W\Theta = I$, we have

$$W \times \Theta = \begin{pmatrix} I & 0 \\ 0^T & 1 \end{pmatrix}. \tag{4.5}$$

From this, we can derive,

$$w_{12} = -W_{11}\frac{\theta_{12}}{\theta_{22}}. \tag{4.6}$$

Substituting Equation 4.6 into Equation 4.4 gives

$$-W_{11}\frac{\theta_{12}}{\theta_{22}} - s_{12} - \lambda \cdot \text{sign}(\theta_{12}) = 0. \tag{4.7}$$

Using $\beta = -\frac{\theta_{12}}{\theta_{22}}$ and $\text{sign}(\theta_{jk}) = \text{sign}(\theta_{jk})$ if $\theta_{jk} \neq 0$, else $\text{sign}(\theta_{jk}) \in [-1, 1]$ if $\theta_{jk} = 0$, we get

$$W_{11}\beta - s_{12} + \lambda \cdot sign(\beta) = 0. \tag{4.8}$$

While the forward search algorithm has been shown to be effective in generating the most relevant explanation, its exhaustive search nature can be computationally expensive. The gLasso offers a compelling solution to this limitation. By incorporating the gLasso into the search strategy of the forward search, we can leverage its ability to promote sparsity in the precision matrix. In essence, the gLasso acts as a pruning tool, guiding the search and reducing the need to explore all possible neighbours. This combined approach improves computational efficiency while maintaining the characteristics of a good explanation captured in MRE.

4.2.1 Search strategy

The forward search algorithm is the foundational framework for the proposed forward-gLasso algorithm. The forward search, introduced in Section 3.2.3, provides the essential architecture that guides the proposed algorithm through the solution space. We can think

of the forward search as a puzzle where we need to assemble many pieces to obtain the full picture. We start with one piece and gradually add one at a time to find the best-fitting combination. Referring to Figure 3.3, we initiate the search process by assigning the most likely state to each target variable. In this case, *OtherCar* is set to *True*, i.e., b – node filled with a solid grey in layer 1. We attempt to gradually improve the solution through the addition of one variable to the solution or by altering the state of one variable. These paths are indicated by solid lines. The addition of one variable is shown in layer 2 of the lattice structure. We evaluate the solution against the neighbouring solutions at each iteration according to the generalised Bayes factor. We stop the process if the current solution has a higher generalised Bayes factor than the best neighbour. In contrast, if the best neighbour has a higher generalised Bayes factor than the current best solution, we update the current best solution and repeat the process. Eventually, we reach a point where adding more variables (or changing the state of a single variable) does not improve the explanatory power of the explanation. Here, the algorithm visits all possible instantiations of variables, which can be computationally inefficient.

Neighbourhood selection

To avoid visiting all possible instantiations of variables and the possibility of overspecified explanations, we apply the gLasso at each iteration of the forward search to prune the neighbourhood, N. In general, gLasso adds variables to the solution with the strongest connections to the already included variables. In other words, gLasso helps us identify which variables (and their states) are closely related. The gLasso can contribute to variable selection through two approaches. Firstly, it facilitates the precision matrix estimation, which contains information about the dependencies among the variables. Variables that display strong connections in the precision matrix will likely hold significance in the model. Secondly, variables can be selected with gLasso by using the penalised log-likelihood score. The penalised log-likelihood measures how well the model fits the data. The variables with the highest penalised log-likelihood score are the most likely to be important for the explanation. We will focus on estimating the precision matrix to select the neighbours. Using gLasso to select the neighbourhood ensures that the variables we include are not only individually relevant but also contribute meaningfully to the explanatory power of

Algorithm 2 Forward-gLasso search algorithm

Input: Bayesian network $\mathcal{B}$, set of observation variables E, a set of target variables X, and sparsity parameter λ.
Output: k-MRE solution.

Initialise the starting solution set I with the best pivot initialisation rule.
Initialise the current best solution, $y_{best} = \emptyset$ for each starting solution s
for each starting solution s in I do **do**
    $y = s$
    repeat
        Find the neighbouring solution set N of y by either *changing* the state of a single variable or by *adding* an additional target variable with *any* state.
        Compute the GBF score for each solution in N.
        Construct the precision matrix based on the GBF scores.
        Apply gLasso to prune the neighbouring solutions set N.
        Filter the neighbouring solution set N based on dominance relations.
        Update y if the best solution in N has a *higher* GBF score.
    until y stops updating
    if $GBF(y) > GBF(y_{best})$ **then**
        $y_{best} = y$
    end if
end for
return y_{best}

the explanation.

Once the neighbourhood, N, has been pruned, we can continue the search by comparing the current best solution with the updated neighbouring set. The forward-gLasso algorithm is presented in Algorithm 2. Figure 4.1 illustrates the path for *OtherCar* using the forward-gLasso algorithm. At initialisation, *OtherCar* is set to *True*, i.e., b. Notice here that the path to $b\bar{c}$ and $\bar{a}b$ are dashed. During the forward search, these paths were solid. However, applying gLasso highlighted that these variables are conditionally independent of the current best solution and can be eliminated from the neighbouring set. We see here that the neighbouring set for the first iteration was reduced from the initial five to three.

Estimating the precision matrix

We will use a generalised Bayes factor score matrix to approximate the precision matrix used as input for the gLasso algorithm. Each cell in the score matrix corresponds to a neighbour's generalised Bayes factor score in the neighbouring solution set N. The

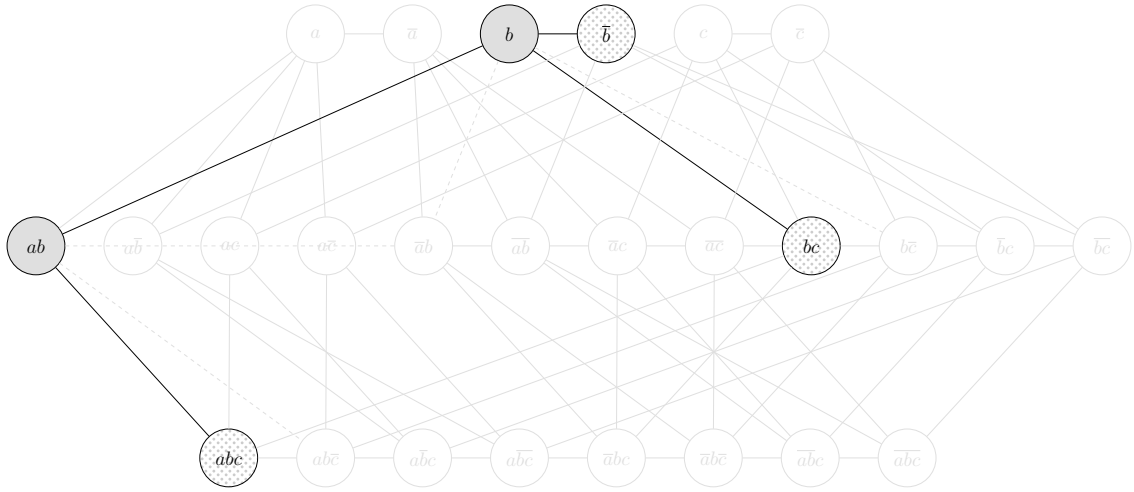

Figure 4.1: Path for *OtherCar* (**B**) through forward-gLasso search.

generalised Bayes factors are calculated using Equation 3.4. We split each neighbour into two components based on neighbour type to get off-diagonal scores. For *add-one* neighbours, the first component consists of the current best solution, and the second component consists of the additional variable. Then, for *change-one* neighbours, the first component consists of all variables in that neighbour for which the state did not change, whereas the second component consists of the variable for which the state changed.

To illustrate this, let's consider the current best solution, ab, in Figure 4.1. Our *add-one* neighbours are $\{abc, ab\bar{c}\}$ and the *change-one* neighbours are $\{a\bar{b}, \bar{a}b\}$. We can split the neighbours as follows: $ab\bar{c} = \{ab, \bar{c}\}$, $abc = \{ab, c\}$, $\bar{a}b = \{\bar{a}, b\}$, and $a\bar{b} = \{\bar{b}, a\}$. The resulting score matrix is symmetric with the following row and column names: ab, $\bar{c}$, c, $\bar{a}$, b, $\bar{b}$, a:

$$\Theta = \begin{array}{c} \\ ab \\ \bar{c} \\ c \\ \bar{a} \\ b \\ \bar{b} \\ a \end{array} \begin{array}{c} \begin{array}{ccccccc} ab & \bar{c} & c & \bar{a} & b & \bar{b} & a \end{array} \\ \begin{pmatrix} 1.61 & \mathbf{0.65} & \mathbf{1.66} & 0 & 1.61 & 0 & 1.61 \\ \mathbf{0.65} & 0.67 & 0 & 0.68 & 0.91 & 0.56 & 0.54 \\ \mathbf{1.66} & 0 & 1.49 & 1.05 & 1.54 & 0.97 & 1.66 \\ 0 & 0.68 & 1.05 & 0.63 & \mathbf{1.01} & 0.60 & 0 \\ 1.61 & 0.91 & 1.54 & \mathbf{1.01} & 1.53 & 0 & 1.61 \\ 0 & 0.56 & 0.97 & 0.60 & 0 & 0.65 & \mathbf{1.25} \\ 1.61 & 0.54 & 1.66 & 0 & 1.61 & \mathbf{1.25} & 1.60 \end{pmatrix} \end{array},$$

where boldface values represent the generalised Bayes factor scores for the neighbours of

ab.

Decoupling from directed structure

Although the graphical representation of Bayesian networks involves directed arcs that indicate conditional dependencies between variables, when solving the most relevant explanation in a Bayesian network, it should be noted that the resulting set of possible configurations is not directed. Instead, it represents the universe of undirected variable assignments that can potentially explain the observed evidence. This allows us to use the gLasso to prune the universe such that only the most relevant variable instantiations remain.

4.3 Experimental design

We evaluate the performance of the forward-gLasso algorithm on a set of benchmark Bayesian networks: Asia (Lauritzen & Spiegelhalter 1988), Alarm (Beinlich et al. 1989), Circuit (Poole & Provan 1990), Hepar2 (Kłopotek et al. 2000), and Insurance (Binder et al. 1997). Asia, Alarm, Hepar2, and Insurance are included in the Bayesian network data repository (Scutari 2010). We used the target and observation nodes described in the relevant literature for each network. If the diagnostic node groupings were not available for a particular network, we utilised alternative groupings. Table 4.1 provides an overview of the benchmark Bayesian networks. Following a similar approach as Yuan, Lim & Littman (2011), we will use each benchmark network as a generative model to generate test cases for evaluation. We limit the generated test cases to only include unique observations, as these observation nodes will serve as evidence. For smaller networks, such as Asia and Circuit, where we have a single binary observation, we will have two unique test cases. Since the Asia network allows for two possible diagnostics, we will run the experiments on both, therefore we will have four unique test cases.

We first report on the forward-gLasso search algorithm's ability to prune the neighbourhood size. Thereafter, we evaluate the forward-gLasso according to two performance indicators, namely *cases solved exactly* and *average execution time*. Since the graphical Lasso includes a non-negative tuning parameter to encourage sparsity, we implemented the forward-gLasso algorithm with three different tuning values: $\lambda = 0.01$, $\lambda = 0.001$,

Table 4.1: Summary of the benchmark networks used in the experiments.

Network	Type	Nodes	Arcs	Targets	Observations
Asia	Small	8	8	3	1
Circuit	Small	10	11	4	1
Insurance	Medium	27	52	6	5
Alarm	Medium	37	46	8	16
Hepar2	Large	70	123	7	63

$\lambda = 0.0001$. The results obtained from a brute-force approach implemented in `R` were used as ground truth. Lastly, we provide the explanations obtained through the forward-gLasso for the running example introduced in Section 3.2.1. The experiments were performed in a `R` environment with `R` version 4.3.1 for programming. The Bayesian networks were loaded through the `bnlearn` (Scutari 2010) and `gRain` (Højsgaard 2012) packages. Graphical Lasso was performed through the `glassoFast` (Sustik et al. 2023) package. The test cases were generated using the `rbn` function from `bnlearn`, with the loaded network structure and the number of samples to generate as input. We restricted the generated cases to only include unique observations, as previously described. Where applicable, we sampled 200 of these test cases without replacement, with a seed of 13.

4.4 Experimental results

4.4.1 Neighbourhood reduction

Table 4.2 displays the average number of neighbours visited for each algorithm. Boldface entries indicate the smallest average neighbourhood size among the algorithms. Overall, the results demonstrate the efficiency of the proposed forward-gLasso algorithm in reducing neighbourhood sizes compared to the forward search algorithm. This is particularly significant in scenarios where computational efficiency and resource utilisation are important. By reducing the number of neighbours to explore, the forward-gLasso algorithms simplify the process of finding the most relevant explanation within Bayesian networks. It is important to acknowledge the exception found in the Asia network, which is considered a small network. In this specific instance, the forward-gLasso$_{\lambda_{0.001}}$ variant displayed a larger average neighbourhood size in contrast to other versions of the algorithm as well as the forward search algorithm. This highlights the sensitivity of the choice of the regulari-

sation parameter, λ, as it determines the level of sparsity in the estimated matrix, thereby influencing the inclusion or exclusion of neighbours.

Table 4.2: Comparison of the average neighbourhood size of each algorithm.

	Forward	F-gLasso$_{\lambda_{0.01}}$	F-gLasso$_{\lambda_{0.001}}$	F-gLasso$_{\lambda_{0.0001}}$
Asia	13.75	**13.25**	14.25	13.75
Circuit	25.50	**24**	25.50	25.50
Insurance	120.76	**92.27**	112.05	118.83
Alarm	123.14	**88.02**	106.23	118.28
Hepar2	105.48	**74.79**	92.94	102.79

4.4.2 Computational efficiency

Table 4.3 provides a detailed breakdown of the computational results. The table includes the Bayesian network, the total test cases solved exactly (CSE), and the average execution time in seconds (AET). Overall, the forward-gLasso$_{\lambda=0.0001}$ stands out as it successfully solves the same number of test cases as the forward search, but in less time. This finding underscores the computational advantages of including gLasso in the forward search. However, it is worthwhile to note that while the average execution time for the forward-gLasso implementations is generally quicker than the forward search, the forward-gLasso$_{\lambda=0.01}$ and forward-gLasso$_{\lambda=0.001}$ variants tend to solve fewer cases exactly, particularly in larger networks. This observation sheds light on the trade-off between computational efficiency and accuracy. When using a smaller value of the regularisation parameter λ, the algorithms achieve higher computational efficiency but sacrifice some degree of accuracy in the results.

Let's look closer at the Alarm network, with 16 observation nodes and 8 target nodes. Here, the forward-gLasso$_{\lambda=0.0001}$ solves cases efficiently while maintaining high accuracy. In contrast, the forward-gLasso$_{\lambda=0.01}$ and forward-gLasso$_{\lambda=0.001}$ variants, while faster, do not attain the same level of accuracy. This observation echoes our earlier observation while comparing the neighbourhood pruning capabilities of each variant, as presented in Table 4.2. Despite a reduction in average neighbourhood size from 118.28 to 106.225, the forward-gLasso$_{\lambda=0.001}$ solves one less case than the forward-gLasso$_{\lambda=0.0001}$.

Table 4.3: Comparison of test cases solved exactly (CSE) and the average execution time (AET) of each algorithm in seconds.

	Forward		F-gLasso$_{\lambda_{0.01}}$		F-gLasso$_{\lambda_{0.001}}$		F-gLasso$_{\lambda_{0.0001}}$	
	CSE	AET	CSE	AET	CSE	AET	CSE	AET
Asia	4	0.236	4	0.159	4	0.155	4	0.166
Circuit	2	0.678	2	0.469	2	0.494	2	0.511
Insurance	199	14.887	195	9.663	192	10.375	199	11.873
Alarm	200	16.943	193	7.307	199	9.952	200	12.051
Hepar2	199	10.719	195	4.882	195	6.779	199	7.985

4.4.3 Most relevant explanation according to forward-gLasso

We also report the explanations obtained through the forward-gLasso search algorithm on our running example and a second scenario based on the Insurance network with 6 targets and 5 observation variables.

Scenario 1: running example

In this example, a policyholder submitted a claim after their vehicle, with tank-level ruggedness, was involved in a mild accident. The three target variables are *Antilock*, *OtherCar*, and *Airbag*. The minimal explanations, using $\lambda = 0.2$, are given in Table 4.4. In this scenario, the most relevant explanation for the observed evidence is the presence of both an anti-lock braking system and an airbag with a GBF score of 1.657, followed by the presence of an anti-lock braking system and the involvement of a second vehicle, and lastly the presence of an airbag and the involvement of a second vehicle.

Table 4.4: Set of explanations for the Insurance network using the forward-gLasso algorithm.

AntiLock	OtherCar	Airbag	GBF
True		True	1.657
True	True		1.613
	True	True	1.539

Note that in the running example, we only include binary target variables. We experiment on a larger set of observations and target nodes from the Insurance network to further investigate the most relevant explanations obtained with the forward-gLasso algorithm.

Scenario 2: including socio-economic and demographic factors

Suppose we observe a *moderate* accident and include additional socio-economic and demographic factors related to the driver as evidence, such as *Age*, *SocioEcon*, *HomeBase*, and *DrivHist*. In this scenario, the policyholder submitting a claim is a middle-class adult based in a rural neighbourhood with zero prior accidents. We are interested in attributes of the vehicle, such as *Antilock*, *Airbag*, *VehicleYear*, *MakeModel*, and *RuggedAuto* to explain the observed evidence.

The minimal explanations, using $\lambda = 0.001$, obtained through the proposed forward-gLasso are given in Table 4.5. The first explanation consists of a singular instantiation, *MakeModel* = *FamilySedan* with a generalised Bayes factor of 3.074. The remaining explanations are partial instantiations of the target variables and no explanation consists of a full instantiation, i.e., all five target variables. In this scenario, the simplest explanation carries the highest explanatory power and more complex explanations have lower explanatory power. This observation agrees with Lötsch et al. (2022) who argue that explanations should be simple since this is a requirement of comprehensibility. Accordingly, the set of explanations provided are considered *minimal* explanations, since neither of the explanations are either strongly or weakly dominated by another explanation. As such, the explanations obtained through the forward-gLasso algorithm retain the properties of a good explanation, namely *preciseness* and *conciseness*.

Table 4.5: Set of explanations for scenario 2 of the Insurance network using the forward-gLasso algorithm.

AntiLock	Airbag	VehicleYear	MakeModel	RuggedAuto	GBF
			FamilySedan		3.074
False		Older		Tank	2.173
False	False			Tank	2.167
		Older		Football	1.480
	False			Football	1.449

Using the developed forward-gLasso algorithm, we can now explore additional points of interest. For example, is the explanation set obtained from the most relevant explanation static or dynamic? Does the best explanation stay consistent or change in light of new information? We expect the explanation set to be updated with new evidence to reflect the dynamic nature of human reasoning.

4.5 Testing robustness of the most relevant explanation with the same-decision probability

In this section, we explore the impact of new evidence on the most relevant explanation. Instead of haphazardly selecting new evidence to observe, we make use of the same-decision probability as a selection criterion. As such, we first compute the most relevant explanation for a set of observed evidence. Thereafter, we employ the same-decision probability to determine, from a set of latent evidence variables, which variable we should observe next using a myopic approach. Having determined the next variable to observe, we update the explanation set and recompute the most relevant explanation. To illustrate this, we use the Win95pts Bayesian network provided by `bnlearn` (Scutari 2010). The Win95pts network consists of 76 binary nodes with 112 arcs and is used for troubleshooting print-related problems. For example, it can be used to troubleshoot "no output", "garbled output" or slow printing.

Suppose we want to understand why we observed no output when we know the printer is switched on with the correct application data. To troubleshoot, we investigate the printer paper supply, whether the correct printer was selected, whether the printer timed out, and the toner supply. The initial evidence and target variables are described in Table 4.6, where the variable type, variable name, a short description, and the observed state (where applicable) are displayed. Using this, the most relevant explanation for the observed evidence is presented in Table 4.7. Notice that these explanations consist of full instantiations of the target set. Out of interest sake, we also compute the most probable explanation as $\{PrtPaper = Has\_Paper, PrtTimeOut = Long\_Enough, TnsSpply = Adequate, PrtSel = Yes\}$.

Since the explanations for the initial evidence provide similar generalised Bayes factor scores, we implement the same-decision probability to determine which variable we should observe next, based on a decision variable *PC2PRT*, a threshold of 0.5, and latent evidence variables *PrtDataOut* and *PrtCbl*. The variable descriptions are provided in Table 4.6. For this experiment, we separate the set of target variables from the latent evidence variables such that the target set remains constant. The same-decision probability for this scenario is 0.655. Although this is greater than the threshold, there is still a 34.5% chance that

Table 4.6: Description of variables of interest in the Win95pts Bayesian network.

Variable type	Variable name	Short description	State
evidence	`Problem1`	No output	No_Output
	`PrtOn`	Printer on and online	Yes
	`AppData`	Application data	Correct
target	`PrtPaper`	Printer paper supply	
	`PrtSel`	Correct printer selected	
	`PrtTimeOut`	Printer timeouts	
	`TnrSpply`	Toner supply	
decision	`PC2PRT`	PC to PRT Transport	Yes
hidden unobserved	`PrtDataOut`	Print data out	
	`PrtCbl`	Local printer cable	

Table 4.7: Most relevant explanation for initial evidence set.

TnrSupply	PrtPaper	PrtTimeOut	PrtSel	GBF
Adequate	Has_Paper	Too_Short	Yes	2.295
Adequate	No_Paper	Long_Enough	Yes	2.182
Low	Has_Paper	Long_Enough	Yes	2.129
Adequate	Has_Paper	Long_Enough	No	2.034

we would make a different decision had we observed the latent evidence variables. At this point, we can either decide to commit to the decision or continue information gathering.

Suppose we continue information gathering, we now need to determine which variable, *PrtDataOut* and *PrtCbl*, to observe next such that we can make a more informed decision. Therefore, the next step is to determine the expected benefit of observing each of these variables. Observing *PrtDataOut* will result in an SDP of either 0.95 (if we observe $PrtDataOut = Yes$ or 1 with an expected SDP of 0.965, whereas observing *PrtCbl* will give us an SDP of either 0.684 (if we observe $PrtCbl = Connected$) or 1 for an expected SDP of 0.697. Therefore, the corresponding SDP gains are: $\mathcal{G}(PrtDataOut) = 0.31$ and $\mathcal{G}(PrtCbl) = 0.041$. Hence, observing *PrtDataOut* will on average allow us to make a more robust decision that is less likely to change due to additional information.

Consider the scenario where we observe $PrtDataOut = Yes$. The most relevant explanation for the updated evidence is given in Table 4.8. Notice here the most relevant explanation changes in light of this new evidence. Previously, each explanation consisted of a full instantiation of the target variables. Now, however, the "best" explanation is a singleton explanation consisting of only *PrtSel*, indicating that the most relevant cause

Table 4.8: Most relevant explanation for updated evidence set.

TnrSupply	PrtPaper	PrtTimeOut	PrtSel	GBF
			Yes	12.761
Adequate	Has_Paper	Too_Short		3.111
Adequate	No_Paper	Long_Enough		2.866
Low	Has_Paper	Long_Enough		2.768

for the observed evidence is whether the correct printer is selected and not a combination of factors. Furthermore, this variable is excluded from the remaining explanations. If we focus on target variables *TnrSupply*, *PrtPaper*, and *PrtTimeOut*, for explanations two, three, and four, we notice the states are the same as before (Table 4.6), except for the exclusion of *PrtSel*. This may be attributed to the fact that, in light of the new evidence, selecting the correct printer holds more explanatory power than the remaining target variables.

This experiment demonstrates that the most relevant explanation is dynamic and can adapt to new evidence, reflecting the dynamic nature of real-world decision-making. Furthermore, the experiment suggests that including more evidence can elevate the explanatory power of certain variables, potentially leading to a more comprehensive understanding.

4.6 Conclusion

A key challenge in XAI is developing methods capable of producing computationally efficient explanations. Local search algorithms, such as the forward search algorithm, are exhaustive and can be computationally inefficient, especially for larger networks. These algorithms visit all variable instantiations to obtain the set of most relevant explanations for the observed evidence. This motivated the development of a novel algorithm capable of efficiently pruning the search space.

This chapter introduced a forward-gLasso search algorithm as an approximate search algorithm to solve the most relevant explanation. The forward-gLasso search algorithm builds upon the forward search by incorporating the neighbourhood selection capabilities of gLasso. This combined approach results in a more computationally efficient algorithm. We compared the proposed algorithm with the forward search algorithm in terms of com-

putational efficiency and the number of cases solved exactly, with the results of the brute-force algorithm as our ground truth. After that, we illustrated the minimal explanations obtained from forward-gLasso for the running example from Section 3.2.1 as well as the explanation set from another scenario in the Insurance Bayesian network. Thereafter, we showed the dynamic nature of the most relevant explanation in light of new evidence obtained through the same-decision probability.

Having established the theoretical foundation and methodology, we can now turn our attention to the practicality of these methods. The following chapter presents a taxonomy of explainable Bayesian networks and an `R` package to support the work illustrated in this research.

Chapter 5

Taxonomy of explainable Bayesian networks

5.1 Introduction

After exploring existing explanation methods in Bayesian networks and identifying a critical research gap of limited focus on decision-readiness, it becomes clear that a standardised taxonomy for explanations in Bayesian networks is also missing. This absence hinders researchers' ability to assess how well explanations translate into actionable insights. To address this challenge, we propose a taxonomy of explainable Bayesian networks. This framework will serve as a foundation for improved communication among users, facilitating a more nuanced understanding of explanation types and their potential to support informed analysis and decision-making. Furthermore, recognising the lack of open-source software to generate explanations for Bayesian networks, we develop an `R` package. This package facilitates the use of the most relevant explanations and includes three algorithms: a brute-force search, a classic forward search, and the proposed forward-gLasso search.

This chapter is structured as follows. Section 5.2 presents the proposed taxonomy of explainable Bayesian networks. This includes a discussion on how we can utilise the taxonomy to address questions a decision-maker seeks to answer. Lastly, we demonstrate the developed `R` package in Section 5.3 using two benchmark networks included in the `bnlearn` package. It is important to note that the package is still in its early stages. Our future endeavours include expanding the package to support additional explanation

methods.

5.2 Taxonomy of explainable Bayesian networks

Building upon the explanation categories proposed by Lacave & Díez (2002) and incorporating insights from discussions in Sections 2.3, 2.4, and 3.2, we present a taxonomy for explainable Bayesian networks (XBN), which also incorporates the newly proposed category on decision-readiness discussed in Section 3.3. Figure 5.1 illustrates the proposed XBN taxonomy.

Recognising the importance of the target audience for explanations, XBNs prioritise a user-centred approach. Due to the diverse intent, requirements, and expectations of XAI communities (Preece et al. 2018, Langer et al. 2021, Barredo Arrieta et al. 2020), XBNs shift focus from technical details to explaining the specific task at hand. This approach enables XBNs to address user questions like "*why*", "*what*" or "*how*" by selecting the most suitable method based on the user's specific needs.

While we define the taxonomy along the four categories, we acknowledge that not all categories are necessarily of interest to the user. In the context of participatory modelling (Düspohl et al. 2012), for example, the emphasis will be on explaining the model. However, if the Bayesian network is used as a classifier, the focus shifts towards explanation of reasoning and explanation of decisions. For example, suppose the model prediction is unexpected or counterintuitive. Users might then seek to understand the reasoning behind the output. For example, encountering a "loan denied" prediction instead of "loan approved" might trigger questions like "*why was the loan denied instead of approved?*" or "*was it the client's low credit score or their unstable employment history that lead to the denied loan?*". However, a user might leverage the model to evaluate the confidence level associated with the current evidence before committing to a final decision. For example, "*beyond the symptoms presented, are there any environmental factors, such as recent travel history, exposure to allergens, etc, that will change the current diagnosis?*" The remainder of this section illustrates several questions or scenarios users may ask or investigate and assigns them to the relevant category in the XBN taxonomy. Given that explanation of the model is deemed static and primarily involves the display of the knowledge base, we opt to exclude it from this discussion.

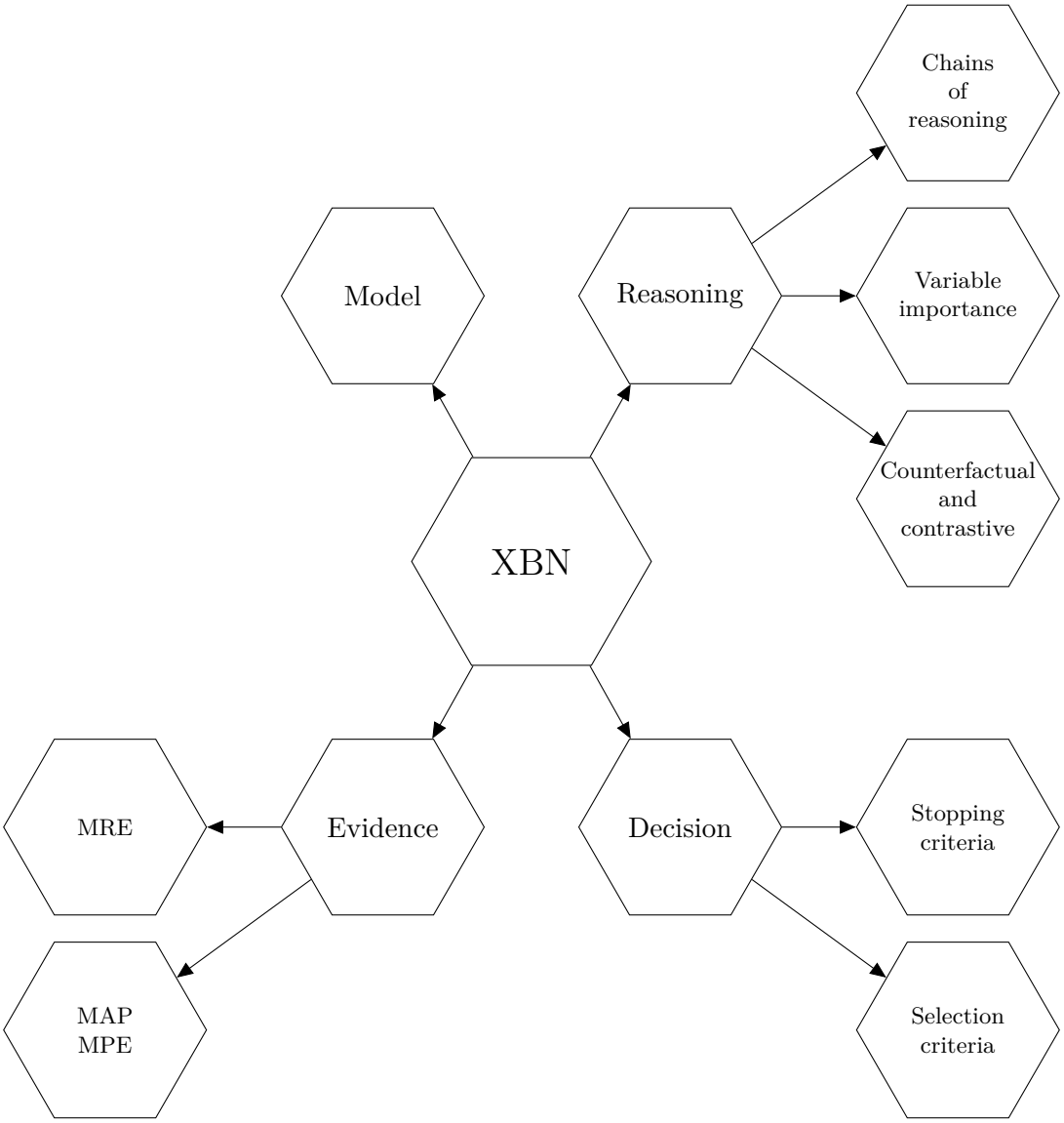

Figure 5.1: A schematic view of explainable Bayesian networks.

5.2.1 Reasoning

In the context of explanation of reasoning, different stakeholders have different expectations of justifications provided for predictions. For example, a data scientist may be interested in understanding how evidence propagates through the network to arrive at the prediction, or they may be interested in identifying which variables have the most significant influence on the prediction. At the same time, regulators would be interested in examining variable importance to ensure that sensitive features do not disproportionately influence decisions. On the other hand, consumers might be interested in counterfactual

explanations to understand why a specific prediction was made and how it can be changed.

Consider the Asia Bayesian network from (Lauritzen & Spiegelhalter 1988) (illustrated in Figure 3.6). A patient may ask "*given their history of smoking, how likely is an abnormal X-ray?*. In this case, the user is concerned with a single outcome, i.e., the X-ray result. Similarly, a healthcare provider may ask "*what is the probability of a patient being a smoker, given that they presented shortness of breath?*" or "*what if the patient had not visited Asia, how would the probability of tuberculosis change?*"

Next, consider a network that predicts a higher than usual customer churn rate for a subscription service. We might be interested in examining data on customer demographics, usage patterns, and reasons for cancellations to better understand what contributes most to churn. Identifying the variable that contributes the most can lead to actionable insight. Suppose we discover a significant number of cancellations are from customers who have not used the service features extensively. We can use this to create clear and engaging on-boarding tutorials to familiarise new customers with the features and benefits, implement a free trial period to allow users to experience the value proposition before committing to a subscription, or provide contextual in-app messaging highlighting features relevant to user behaviour, encouraging them to explore the service's full potential. In addition to this, a marketing manager might ask *what if they offer a discount on the subscription, would this influence the churn rate?*

Consider a network that predicts high-risk areas. Suppose the model identifies a specific neighbourhood with a high risk of property crime over the next week. Using variable importance in explanation of reasoning, the system may provide justifications for the prediction. For example, it may highlight recent crime statistics, upcoming events, or weather patterns. This may then motivate law enforcement to focus their resources on high-risk areas and allocate more manpower strategically based on the predicted activity. Moreover, this could highlight potential risk factors and work with community leaders to implement preventative measures, fostering cooperation and proactive approaches to crime reduction.

5.2.2 Evidence

When users are interested in finding a hypothesis that best describes specific observed phenomena, we use explanation of evidence methods. For example, a medical professional might ask "*which diseases are the most likely cause of the symptoms presented?*" or "*which diseases are most relevant in explaining the symptoms presented by the patient?*". Similarly, a patient may ask "*given their shortness of breath and* " Additionally, hospitals can analyse patient data to identify factors contributing to patient readmission within a specific timeframe. This may include diagnosis, severity of illness, patient demographics, social support network, medication adherence and adverse events. By understanding the factors most likely associated with readmission, hospitals can identify high-risk patients who might benefit from additional support after discharge.

In the context of an insurance claim, the service provider might ask "*what subset of variables would be most relevant in explaining a high inspection cost?*" or "*given a newer vehicle was involved in a severe accident, what aspects related to driver behaviour are most likely to explain the accident?*" Alternatively, insurance companies can analyse historical claims data to identify patterns that suggest potential fraud. This might involve looking at inconsistencies between a claim and a policyholder's past information, unusual claim filing times, or suspicious third-party details. By understanding the factors most likely associated with fraudulent claims, insurers can flag high-risk cases for further investigation. This helps them allocate resources efficiently and potentially save in fraudulent payouts.

Similarly, a bank might analyse loan applications that were rejected. By using explanation of evidence, they can understand which factors in a borrower's risk profile (e.g., low credit score, high debt-to-income ratio, other economic indicators) have the greatest impact on loan rejection. This helps the bank refine its loan approval process and potentially offer alternative solutions to those with less severe risk factors.

5.2.3 Decisions

In the context of decision-readiness, users would typically ask "*do we have enough evidence to make a decision?*" and if not, "*what additional evidence do we need in order to make an informed decision?*". A medical professional might ask "*do we have enough evidence on the symptoms presented to make a decision on the disease?*" or "*since we cannot*

yet make a decision, what additional information – tests, comorbidities, other patient histories – is required to make a decision?." For example, consider a 35-year-old patient arriving at the emergency room with a fever, cough, and fatigue. These symptoms are common across various respiratory illnesses, making a conclusive diagnosis challenging based solely on this information. The medical professional might explain that while the symptoms are concerning, they do not have enough evidence to pinpoint a specific disease. However, the decision-readiness process could highlight information that will lead to a more robust decision. For example, a chest X-ray can help differentiate between pneumonia and other lung conditions. Blood tests can identify potential causes like influenza or bacterial infections. Additionally, understanding if a patient has been exposed to anyone with similar symptoms might provide insights about potential contagions. By identifying the specific tests and information needed, resources are used more efficiently which can lead to a more timely diagnosis and treatment plan.

Applied to forensic investigations, this can be used to answer questions relating to crime scene investigations. The analyst may ask questions regarding the actual evidence collected from the crime scene, i.e., whether enough evidence is collected to rule a crime a homicide or what additional evidence is required to rule the crime a homicide. Should they investigate further, or is the evidence already collected enough to make an informed decision? For example, suppose a body is found at a crime scene. The cause of death appears to be a gunshot wound. Is there enough evidence at the scene to determine if this is a homicide or not? The analyst examines the collected evidence, which might include witness statements mentioning arguments or suspicious activity, location and trajectory of the gunshot wound, gunshot residue patterns on the victim and surrounding area, and fingerprints collected from the scene. Based on the initial evidence, decision-readiness might determine there is not enough evidence to definitively rule this as a homicide. While the gunshot wound suggests foul play, they need more information to build a stronger case. Further investigation may include analysing the bullet recovered from the scene to potentially link to a specific firearm, conducting a thorough DNA analysis of the crime scene to identify any suspects or trace evidence and re-interviewing witnesses to gather more details about the events leading up to the victim's death. By highlighting what evidence is missing, it guides investigators to focus their efforts on collecting the most

crucial pieces. Furthermore, explanation of decisions provides a clear rationale behind the need for further investigation, fostering better communication between the investigative team and legal authorities. Lastly, a targeted approach to evidence collection strengthens the case and may increase the likelihood of a successful prosecution. By following these, the forensic team can gather the necessary evidence to reach a more definitive conclusion. This could be a homicide, an accident, or even suicide, depending on the additional information collected.

Suppose an e-commerce company is experiencing a decline in online sales. The company would like to decide whether to implement a new marketing campaign to boost sales or is there another underlying issue causing the decline? The marketing team has gathered data on website traffic, conversion rates, and customer demographics. The marketing manager might explain that while a new marketing campaign could potentially increase sales, we do not have enough evidence to pinpoint the exact cause of the decline. Based on the decision-readiness analysis, they decided to investigate further before allocating the budget to a new campaign. They may analyse customer reviews and social media sentiment, which can reveal if there are product quality issues, website usability problems, or competitor offerings impacting customer satisfaction. Digging deeper into website traffic data might identify a drop in organic traffic due to search engine algorithm changes or a decrease in paid advertising effectiveness. Furthermore, analysing sales data by product category can highlight if specific products are underperforming due to pricing issues, lack of proper marketing, or changing customer preferences. Decision-readiness encourages a data-centric approach, prioritising investigation to understand the root cause before allocating resources to potential solutions. Furthermore, by identifying the core problem, the company can implement targeted solutions rather than launching a potentially ineffective marketing campaign. Lastly, understanding customer behaviour and market trends can help the company make adjustments to its products, marketing strategies, and overall business model for sustainable growth. Based on the insights gained through the analysis, the company can decide on the most effective course of action. This might be revamping the website for better user experience, improving product quality based on customer feedback, or adjusting marketing strategies to target the right audience.

While the taxonomy is focused on all four arches in explainable Bayesian networks,

there is a lack of open-source software supporting these methods. Hence, one of the objectives of this research is to develop an open-source package to address this. Though it is still in its early stages, we now explore the package's functionalities.

5.3 A package for solving the most relevant explanation

Although numerous specialised Bayesian network software are available, these applications are not open-source. Furthermore, open-source software such as `R` have dedicated Bayesian network packages but do not support generating explanations. Instead, these packages are focused on structure learning and inferences. As such, one of the aims of this research is to develop an open-source `R` package that offers explanation facilities in Bayesian networks. The main function of this package is to solve the most relevant explanation in Bayesian networks and supports three algorithms: a brute-force search, a forward search, and the forward-gLasso proposed in this research. The `XBN` package builds upon existing `R` packages such as `gRain`. The `XBN` package is available on GitHub.

5.3.1 Installation

The package can be downloaded and installed from the GitHub repository using,

```
devtools::install_github('iEna101/XBN')
```

The `XBN` package depends on the `gRain` package for evidence propagation and the `glassoFast` package for implementation of the graphical Lasso. Other packages called in the `XBN` package include: `dplyr`, `gtools`, `magrittr`, `plyr`, `stringi`, `stringr`, and `tidyr`.

5.3.2 Specifying the parameters

Since the focus of this research is on post-hoc explanations, we assume the Bayesian network has already been specified in `R`. To solve the most relevant explanation, we need to specify four primary input parameters:

- `target_set` : a character vector that specifies the set of hypothesis variables. In other words, the variables you want to investigate to explain the observed evidence.
- `evidence_set` : a character vector defining the node names of the observed evidence.

- `evidence_states` : a character vector that contains the observed states for each variable in the `evidence_set` .
- `bn_grain` : a Bayesian network object of class `grain` .

Since `XBN` expects the Bayesian network to be of class `grain` , we can convert an object `bn` of class `bn.fit` can be converted to a `grain` object as follows,

```
bn_grain <- compile(as.grain(bn))
```

The forward-gLasso search introduced in this research, requires two additional parameters, `bn_rho` and `score_scale` , where the former refers to the tuning parameter controlling matrix sparsity and the latter to indicate whether random noise should be added to the score matrix for gLasso.

5.3.3 Practical demonstration

We illustrate the main functionality of the `XBN` package by analysing two benchmark Bayesian networks, namely the Asia network from Lauritzen & Spiegelhalter (1988) and the Insurance network from Binder et al. (1997). The two networks are included in the Bayesian network repository available through `bnlearn` (Scutari 2010). For each implementation, we provide the `R` syntax along with the output.

For the Asia network, we consider the scenario where a patient presents a shortness of breath, i.e., dyspnoea. The Bayesian network `bn_asia` is specified as a `grain` object. Suppose the medical practitioner is interested in understanding whether the cause of the dyspnoea is cancer, tuberculosis, or bronchitis. For this demonstration, we will specify the character vectors for `target_set` , `evidence_set` , and `evidence_states` within the function.

For the Insurance network, we consider the scenario where an adult with no advanced training was involved in a moderate accident. To explain this observation, we consider a combination of driver and vehicle attributes, such as *AntiLock*, *DrivHist*, *DrivQuality*, *RiskAversion*, and *RuggedAuto*. Previously, our running example (Section 3.2.1) based on the Insurance network considered only binary target variables for simplicity. However, in this scenario, we consider target variables with at least two states. The Bayesian network `bn_insurance` is specified as a `grain` object. We can specify `target_set` ,

`evidence_set`, and `evidence_states` as follows,

```
target_set <- c("Antilock", "DrivHist", "DrivQuality",
                "RiskAversion", "RuggedAuto")
evidence_set <- c("Accident", "Age", "SeniorTrain")
evidence_states <- c("Moderate", "Adult", "False")
```

Initialisation

The function `init_gbf` implements the best pivot as an initialisation rule for the most relevant explanation. In essence, the function computes the most likely starting solution based on the highest generalised Bayes factor score for each target.

Using the Asia network,

```
init_gbf(target_set = c("tub", "lung", "bronc"),
         evidence_set = c("dysp"),
         evidence_states = c("yes"),
         bn_grain = bn_asia)

tub     lung    bronc   GBF
<NA>    <NA>    yes     6.139114
<NA>    yes     <NA>    1.967800
yes     <NA>    <NA>    1.827646
```

The column names reflect the variable names as given in `target_set` along with the generalised Bayes factor score `GBF` for each initialisation. From this, we see that *bronchitis = yes*, *lung cancer = yes*, and *tuberculosis = yes* are the best starting solutions for each of the three target variables. Similarly, we can implement this for the Insurance network with parameters specified previously,

```
init_gbf(target_set = target_set,
         evidence_set = evidence_set,
         evidence_states = evidence_states,
         bn_grain = bn_insurance)

Antilock DrivHist DrivQuality RiskAversion RuggedAuto GBF
<NA>     <NA>     Poor        <NA>          <NA>       18.067803
<NA>     Many     <NA>        <NA>          <NA>       4.575542
```

```
<NA>     <NA>     <NA>        Psychopath   <NA>       1.547200
False    <NA>     <NA>        <NA>         <NA>       1.189034
<NA>     <NA>     <NA>        <NA>         Tank       1.168337
```

Notice the difference in generalised Bayes factor scores for *DrivQuality = Poor* and the remaining starting solutions. This highlights the explanatory power of *DrivQuality*.

Brute-force search

The function `mre_brute` implements the brute-force search and returns the full set of explanations, ordered from the highest generalised Bayes factor score to the lowest. Similar to the output from `init_gbf` , the column names reflect the variable names as given in `target_set` along with the generalised Bayes factor score `GBF` . It also includes the hypothesis size `mre_size` , which indicates the number of target variables included in the hypothesis. Using the Asia network, we implement `mre_brute` as follows,

```
mre_brute(target_set = c("tub", "lung", "bronc"),
          evidence_set = c("dysp"),
          evidence_states = c("yes"),
          bn_grain = bn_asia)

bronc   lung    tub     GBF         mre_size
yes     <NA>    <NA>    6.1391138   1
yes     <NA>    no      5.8438928   2
yes     no      <NA>    4.6240442   2
yes     no      no      4.4784580   3
...     ...     ...     ...         ...
...     ...     ...     ...         ...
no      <NA>    <NA>    0.1628900   1
no      <NA>    no      0.1557581   2
no      no      <NA>    0.1323684   2
no      no      no      0.1247761   3
```

Kindly note that the output presented here displays a summary of the hypotheses with the highest and lowest generalised Bayes factor scores. This summary is provided to avoid overwhelming the page with extensive data. Similarly, using the parameters previously specified, the brute force algorithm for the Insurance network can be implemented as

follows,

```
mre_brute(target_set = target_set,
          evidence_set = evidence_set,
          evidence_states = evidence_states,
          bn_grain = bn_insurance)

Antilock DrivHist DrivQuality RiskAversion RuggedAuto GBF
<NA>     <NA>     Poor         <NA>          <NA>       18.067803
False    <NA>     Poor         <NA>          <NA>       7.084173
<NA>     Many     Poor         <NA>          <NA>       5.321693
...      ...      ...          ...           ...        ...
...      ...      ...          ...           ...        ...
True     One      Excellent    Cautious      EggShell   0.02253219
True     Zero     Excellent    Cautious      EggShell   0.02240894
True     <NA>     Excellent    Cautious      EggShell   0.02240628
```

Note the `mre_size` column is excluded from this display for illustration purposes.

Forward-search

The function `mre_fwd` implements the forward search algorithm and returns a set of minimal explanations, ordered from the highest generalised Bayes factor score to the lowest. As with `mre_brute`, the output columns reflect the variables specified in `target_set` along with the generalised Bayes factor score and the hypothesis size. `mre_fwd` can be implemented as follows for the Asia network,

```
mre_fwd(target_set = c("tub", "lung", "bronc"),
        evidence_set = c("dysp"),
        evidence_states = c("yes"),
        bn_grain = bn_asia)

tub      lung     bronc    GBF          mre_size
<NA>     <NA>     yes      6.139114     1
<NA>     yes      <NA>     1.967800     1
yes      <NA>     <NA>     1.827646     1
```

Similarly, we can implement this for the Insurance network using the previously specified parameters,

```
mre_fwd(target_set = target_set,
        evidence_set = evidence_set,
        evidence_states = evidence_states,
        bn_grain = bn_insurance)

Antilock DrivHist DrivQuality RiskAversion RuggedAuto GBF
<NA>     <NA>     Poor         <NA>          <NA>       18.067803
<NA>     Many     <NA>         <NA>          <NA>       4.575542
False    One      <NA>         Cautious      Tank       2.254417
True     <NA>     <NA>         Psychopath    Tank       1.723377
False    <NA>     <NA>         Psychopath    EggShell   1.599162
```

Again, the `mre_size` column is excluded from the display.

Forward-gLasso search

`mre_fwd_glasso` implements the forward-gLasso search proposed in this research. As mentioned, this function takes two additional parameters `bn_rho` and `score_scale`. The output obtained from `mre_fwd_glasso` produces the set of minimal explanations that best explain the observed evidence. For the Asia network with a tuning parameter of $\lambda = 0.001$, we have

```
mre_fwd_glasso(target_set = c("tub", "lung", "bronc"),
               evidence_set = c("dysp"),
               evidence_states = c("yes"),
               bn_grain = bn_asia,
               bn_rho = 0.001,
               score_scale = TRUE)

tub      lung     bronc    GBF          mre_size
<NA>     <NA>     yes      6.139114     1
<NA>     yes      <NA>     1.967800     1
yes      <NA>     <NA>     1.827646     1
```

The output of `mre_fwd_glasso` provides the set of minimal explanations that best explain the observed evidence. This output matches the output from `mre_fwd`. Similarly, using the Insurance network with a tuning parameter of $\lambda = 0.001$,

```
mre_fwd_glasso(target_set = target_set,
               evidence_set = evidence_set,
               evidence_states = evidence_states,
               bn_grain = bn_insurance,
               bn_rho = 0.001,
               score_scale = TRUE)

Antilock DrivHist DrivQuality RiskAversion RuggedAuto GBF
<NA>     <NA>     Poor        <NA>         <NA>       18.067803
<NA>     Many     <NA>        <NA>         <NA>       4.575542
False    One      <NA>        Cautious     Tank       2.254417
True     <NA>     <NA>        Psychopath   Tank       1.723377
False    <NA>     <NA>        Psychopath   EggShell   1.599162
```

Although not displayed here, the output in `R` includes the `mre_size`. Note that, according to both `mre_fwd` and `mre_fwd_glasso`, the best explanation according to the generalised Bayes factor is also the simplest since it consists of only one variable, i.e, *DrivQuality = Poor*. This is in line with Occam's razor and the definition of a *good* explanation, i.e., *preciseness* and *conciseness*. It is also excluded from the remaining instantiations.

The package makes use of several internal functions not illustrated here. For example, `minimal_exp` is a function used to return a minimal explanation based on dominance relations. This function filters both strongly and weakly dominated explanations and is used in both `mre_fwd` and `mre_fwd_glasso`. As such, the explanations obtained from these functions are diverse and representative. At this time, there is no provision for a "switch" that would allow users to filter both strongly and weakly dominated explanations, applying a single dominance relation, or opting for no filtering. Nevertheless, we anticipate integrating such a switch in a forthcoming update.

5.4 Conclusion

This chapter presented two contributions of this research. The first, less technical contribution is the taxonomy of explainable Bayesian networks. Instead of focusing on the technical details, the taxonomy is focused on explaining a specific task or question. As such, we illustrated the practicality of the taxonomy through a series of scenarios a specific end-user might be interested in. Thereafter, we demonstrated the `XBN R` package that stems from this research. The `XBN` package is available on GitHub. We are now in a position to apply the concepts discussed and developed in this research to real-world data sets.

Chapter 6

Explainable Bayesian networks in action: South African VCS

6.1 Introduction

Existing research on explanation methods is mostly limited to benchmark models, such as the Asia network from Lauritzen & Spiegelhalter (1988). One of the objectives of this research is to bridge the gap between theory and practice by demonstrating real-world applications of these methods. By doing so, we will showcase the power of these explanation methods to reveal actionable insights in real-world scenarios. This chapter presents applications of explanation of evidence and explanation of decisions on the South African Victims of Crime Survey (VCS) 2017 - 2018 (Statistics South Africa 2018).

The South African VCS 2017 - 2018 is a nationwide household-based survey capturing data on the prevalence of specific crimes in South Africa. Its primary aim is to establish the prevalence of crime within certain groups in the population. The objectives include providing insights about crime dynamics from a household and victim perspective and exploring public perceptions of law enforcement's role in preventing crime and victimisation. The survey focuses on various aspects, including people's perceptions and experiences of crime, their views on the police service and the criminal justice system, and community responses to crime. The data, which is publicly available[1], profiles different characteristics of crime, such as the location and timing of the crimes and the nature and extent of vio-

[1] Available on Statistics South Africa and various other online platforms.

lence involved. The survey's geographic coverage spans all nine provinces in South Africa, with data aggregated at the provincial level. However, it does not cover institutionalised or military persons or households. The South African VCS utilises a Master sample frame derived from the South African Census 2011.

The remainder of the chapter is structured as follows. Section 6.2 provides a brief description of the data preparation process. In Section 6.3, we apply the most relevant explanation to two scenarios. The first scenario is focused on a specific crime committed in a particular province in South Africa. This allows us to identify areas for intervention or resource allocation. The second scenario is concerned with victim vulnerability. The results obtained from this analysis can steer targeted crime prevention strategies and the development of support services for victims. Lastly, we apply the same-decision probability to assess respondent confidence in the South African Police Service (SAPS) based on observed evidence, such as police presence and specialised operations and unobserved variables such as respondents' satisfaction with police services and the way courts deal with perpetrators, whether they have been asked to pay a bribe, and their perception on violent crime sentencing. While the same-decision probability is typically applied as a threshold-based confidence measure, it can be leveraged to identify unobserved factors that will have the greatest impact on improving the public's confidence in the SAPS.

6.2 Data preparation

The South African VCS included two data sets of interest, the first is focused on person-level data and the second on household-level data. From the person-level data, we extract data such as *age*, *gender*, *education*, and *economic_activity*. We exclude persons younger than 18 and create four categories for *age*: *young*, *early_career*, *late_career*, and *retired*. The *education* variable consists of 7 categories: *primary_school*, *high_school*, *matric*, *vocational_technical_training*, *higher_education*, *unspecified*, and *other*. While *economic_activity* reflects the nature of an individual's work, for example, *permanent* employee.

We aggregate the person-level data with the household-level data, which consists of 21190 cases and 779 variables. We select a subset of variables from the household-level data and, where applicable, create new variables based on these. This includes the type of crime that occurs mostly *type_crime_occur*, the type of crime feared mostly *type_crime_afraid*,

the daily activities a household avoids due to crime *fear_crime_prevent_actions*, why the household believes people commit crimes *why_crime_commit*, measures taken to protect themselves *protection_measures_indiv*, groups that provide protection *protection_groups*, access to institutions *access_instituitions*, the type of crime the household has experienced *crime_experienced*.

The aggregated data set contains several missing values. Instead of imputing missing values, we removed incomplete cases as the number of cases allow for this. Furthermore, several variables contained minimal unspecified values. We removed these cases as well. The final data set consists of 39562 cases, with 21265 females and 18761 early career individuals, 11504 late career individuals, and 2988 retired individuals. We use a score-based hill-climbing search algorithm from `bnlearn` (Scutari 2010) to learn the structure of the network and use `bn.fit` to fit the parameters of the network based on the aggregated data set. The network consists of 49 nodes and is illustrated in Figure 6.1. Please refer to Tables B.2 and B.3 in Appendix B for a description of the variables included in the network.

6.3 Actionable insights: most relevant explanation

Explanation of evidence methods, such as the most probable explanation and the most relevant explanation, can be implemented to understand the causes of crime and victim perceptions. This goes beyond just knowing that a crime happened or that a victim feels unsafe. By identifying factors like police response times, lack of trust in police, or ineffective neighbourhood watch programs, we can develop targeted interventions to address those specific issues. In addition, knowing the most relevant explanations allows for a more focused and efficient allocation of resources. Efforts can be directed towards addressing the factors that have the most impact on victim perceptions and overall crime rates. Furthermore, when analysing explanations from a large dataset like the South African VCS 2017 - 2018 data, common patterns might emerge. These patterns can highlight potential systemic issues with law enforcement, social services, or environmental factors that contribute to crime and distrust.

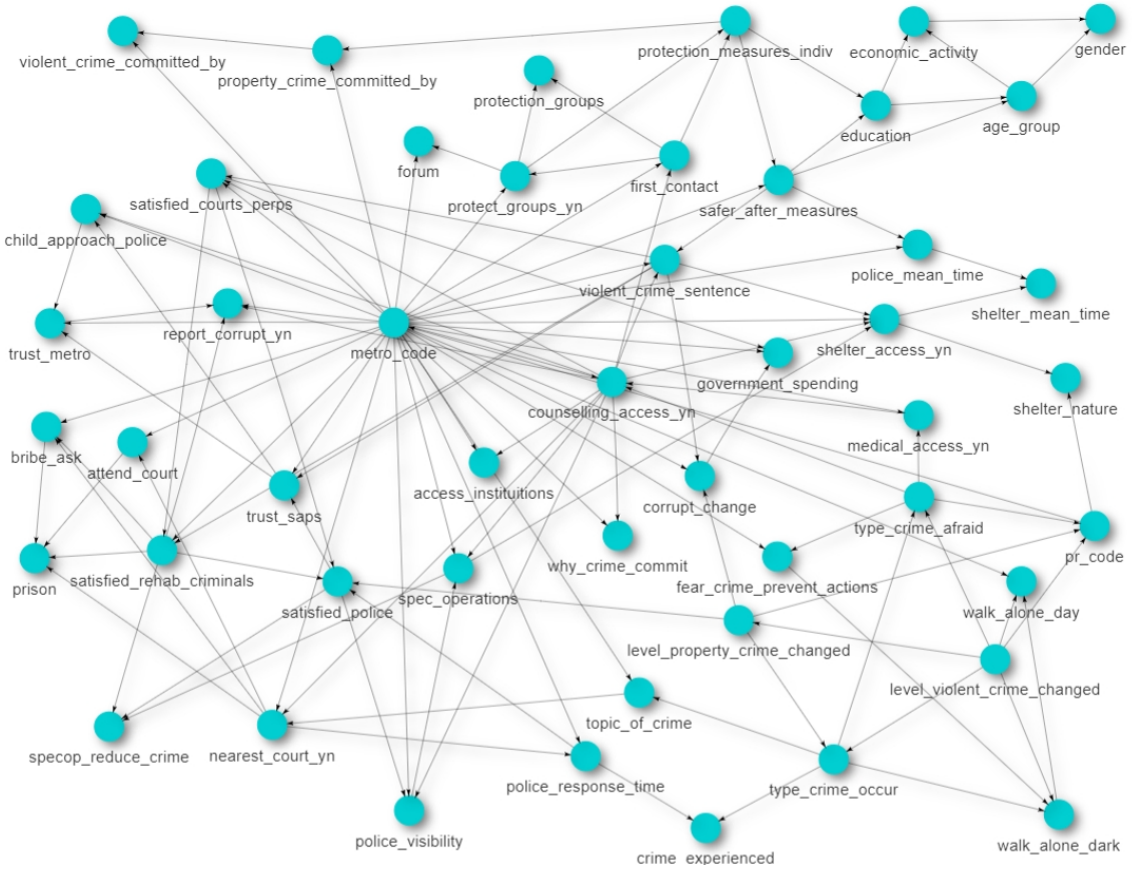

Figure 6.1: Graphical display of the learned structure for the South African VCS data set using a hill-climbing search algorithm.

6.3.1 Case study 1: rising crime in Mpumalanga

The first case study is concerned with a specific crime committed in a particular province with no specialised police operations. Suppose we want to investigate "*given property crime is on the rise in Mpumalanga and there are no specialised police operations, what factors, such as the nearest police station, police response time, and police presence are most relevant in explaining this observation?*" The most relevant explanation can help identify areas for intervention or resource allocation, i.e., increased police presence or response times. The variable encoding for the target set is given in Table 6.1.

Before computing the most relevant explanation, let us first obtain the most probable explanation. For this scenario, the most probable explanation is the nearest police station is less than 30 minutes away, with an unspecified police response time and police presence of less than once a month. This suggests a potential link between infrequent police presence

Table 6.1: Variable encoding for target variables in case study 1.

Variable	Encoding
`police_mean_time`	1: less than 30 minutes
	2: 31 - 60 minutes
	3: 61 - 120 minutes
	4: more than 2 hours
`police_response_time`	1: less than 30 minutes
	2: 31 - 60 minutes
	3: 61 - 120 minutes
	4: more than 2 hours
	5: never arrived
	9: unspecified
`police_visibility`	1: at least once a day
	2: at least once a week
	3: at least once a month
	4: less than once a month
	5: never

and the rise in property crimes in Mpumalanga. While a nearby police station might provide a sense of security, slow or unpredictable response times could create a perception of low risk among criminals. If they believe police won't arrive promptly, they might feel emboldened to commit crimes. Let us now explore the most relevant explanations.

The explanations obtained through MRE are given in Table 6.2. The most relevant explanation for the recent rise in property crimes in Mpumalanga, with no specialised police operations, is despite the nearest police station being less than 30 minutes away, the police response time is more than two hours and there is low police presence (less than once a month). All of the explanations highlight low police presence, with either a presence of less than once a month or at least once a month (but not as frequent as once a week). Furthermore, two of the explanations include a police response time of more than two hours while the last explanation reports an unspecified response time. As mentioned earlier, unpredictable police response times and infrequent police patrols might create windows of opportunity for criminals to commit property crimes with a lower perceived chance of being apprehended.

Actionable insights include evaluating resource allocation, such as police officers and available police vehicles and considering reallocating them to areas with higher crime rates and slow response times. This could involve strategically deploying officers or opti-

Table 6.2: Most relevant explanations for case 1.

police_mean_time	police_response_time	police_visibility	GBF
1	4	4	2.336
1	4	3	1.937
4	9	3	1.174

mising patrol routes. Furthermore, one could explore cost-effective technology solutions like automated dispatch systems to streamline call routing and location tracking for police vehicles. Lastly, if resources allow, authorities can increase the frequency of patrols, particularly in high-risk areas and incorporate foot patrols in neighbourhoods to build trust and encourage resident interaction with police officers.

6.3.2 Case study 2: victim perception

Let us now focus on victim vulnerability, i.e., given a specific victim profile and perception of crime, what factors have led to this perception? Suppose we investigate "*consider a fixed-contract late-career female with a higher education degree, who harbours dissatisfaction and distrust toward the police service. She believes the level of violent crime in the area has increased, impeding her from participating in public activities. What factors are most relevant in explaining this observation?*" We will focus on factors such as the type of crime experienced (if any), who she will contact in a time of need, the average time to the nearest police station, police response time, police presence, the type of measures taken to protect herself against crime and violence, and her perception on the perpetrators of violent crimes. Table 6.3 gives these target variables' variable encoding and codes.

By understanding the factors influencing a victim's perception of crime, interventions can be tailored to address those vulnerabilities and deter similar crimes against young women. Analysing factors like police response times and visibility in the area can highlight areas of improvement. This can lead to increased police patrols, particularly in areas frequented by young women, and faster response times, fostering a greater sense of security. Additionally, exploring why the victim wouldn't necessarily contact the police in a time of need can reveal a gap in trust. This could be addressed through community policing initiatives that build trust and encourage residents to report crimes.

As before, we start with the most probable explanation for the specific victim profile

Table 6.3: Variable encoding for target variables in case study 2.

Variable	Code	Encoding
`crime_experienced`	A	1: property crime
		2: violent crime
		3: other
		9: no crime experienced previously
`first_contact`	B	1: nobody
		2: relative/friend
		3: private security companies
		4: community group/organisation
		5: religious/traditional
		6: South African Police Service
		7: metro police
		8: community policing forum
`police_mean_time`	C	1: less than 30 minutes
		2: 31 - 60 minutes
		3: 61 - 120 minutes
		4: more than 2 hours
`police_response_time`	D	1: less than 30 minutes
		2: 31 - 60 minutes
		3: 61 - 120 minutes
		4: more than 2 hours
		5: never arrived
		9: unspecified
`police_visibility`	E	1: at least once a day
		2: at least once a week
		3: at least once a month
		4: less than once a month
		5: never
`protection_measures_indiv`	F	1: private security
		2: selfhelp group
		3: weapon
		3: other
		9: no protection measures taken
`violent_crime_committed_by`	G	1: people from this area
		2: people from other areas in SA
		3: people from outside SA
		9: unspecified

Table 6.4: Most relevant explanations for the victim profiles in case 2.

(a) Explanations for a late-career female victim.

A	B	C	D	E	F	G	GBF
3	1	1	5	1	1	2	15.779
3	3	1	5	1	1	2	14.926
2	1	1	5	1	1	2	13.755
2	3	1	5	1	1	2	13.035
2	3	1	4	4	1	2	10.689
3	3	1	5	3	3	2	9.659
3	3	1	5	1	3	2	9.387

(b) Explanations for an early-career male victim.

A	B	C	D	E	F	G	GBF
3	3	1	5	3	3	2	5.162
3	3	1	5	1	3	2	5.028
3	4	1	5	1	3	2	4.760
3	3	1	5	2	3	2	4.740
2	3	1	5	3	3	2	4.501
2	3	1	4	3	3	2	4.120
2	3	1	4	4	3	2	3.783

and perception of crime. Here, the MPE indicates that the late-career female perceives violent crime to be committed by locals, yet there is a strong police presence and a nearby station (within 30 minutes). Furthermore, she will seek help from the SAPS first. Also included in the explanation is no crime experienced previously, which leads to an unspecified response time and no additional protection measures taken. This presents a seemingly contradictory situation. Despite positive factors like police presence and accessibility, there's underlying dissatisfaction and distrust toward the SAPS. This leads us to explore the most relevant explanation.

The set of most relevant explanations is presented in Table 6.4a. The column names correspond to the variable codes provided in Table 6.3. The best explanation for the victim's profile, with a GBF score of 15.779, is that the victim was previously affected by crime (other types of crime not captured in the survey). Despite a strong police presence and the nearest police station being less than 30 minutes away, a previous call for help went unanswered. Although the victim utilises private security measures, she does not rely on them in times of need. Instead, the explanation indicates that she would not contact anyone – this may be because she does not have anyone to contact. The victim perceives perpetrators of violent crimes as coming from other areas in South Africa. This explanation gives us a better understanding of the victim's dissatisfaction and distrust in the SAPS as opposed to the MPE since it is not counterintuitive. The explanation highlights the contrast between the strong perceived police presence and the negative past experience, explaining how this could lead to dissatisfaction.

Let us inspect the states of the variables in the explanation set. Notice how all of the

explanations include a perception that violent crimes in the victim's area are committed by people from other areas in South Africa. Using this information as a starting point, authorities can develop targeted community outreach programs that promote social cohesion and understanding between residents from different areas. This can help dispel stereotypes and foster a sense of collective safety. It is important to note that this variable is based on the victim's perception and one should not make generalisations. As such, this perception can be used by authorities in combination with other statistics. Consequently, if it is found that the victim's perception corresponds to actual crime patterns, then this can help inform resource allocation and police collaboration strategies.

Consequently, the widespread use of private security (as highlighted in five of the seven explanations) suggests a lack of confidence in the police's ability to deter crime or respond effectively to incidents. Residents are likely taking matters into their own hands because they do not feel adequately protected by official authorities. To address this, authorities can implement community engagement initiatives that foster positive interactions between police officers and residents. This can involve community policing forums, neighbourhood watch groups, or open forums for residents to voice their concerns and suggestions. Furthermore, they can explore potential partnerships with private security companies to supplement police presence in areas where residents rely on private security measures. This would require clear communication and collaboration between public and private security forces.

Next, we focus on *police_mean_time* (C), *police_response_time* (D), and *police_visibility* (E). In six of the seven explanations, police never respond to a call for help despite there being a nearby police station. Five of the explanations include a police presence at least once a day, one explanation highlights a police presence at least once a month, and the remaining explanation shows a police presence of less than once a month. This lack of response, despite being near a police station, underscores the importance of addressing the ineffectiveness of the police. This could involve investigating the reasons behind the lack of response and implementing measures to ensure calls are responded to promptly and effectively. For example, the lack of response could be due to a lack of resources, such as insufficient manpower or vehicles, inefficient dispatch systems may contribute to delays in assigning officers to calls, and high call volume might result in hindered response times

for lower-priority incidents.

Lastly, two of the seven explanations include "nobody" as the first point of contact in times of need. This paints a concerning picture of victim isolation and a breakdown in trust mechanisms. None of the explanations include the SAPS or metro police as a first point of contact. To address this, authorities can develop targeted outreach programs that connect with vulnerable populations and educate them about available resources for help. This could involve collaborating with community centres, religious or traditional leaders, or social service agencies. Additionally, authorities can invest in robust victim support services that provide immediate assistance, emotional support, and information on legal rights. This can help victims feel less isolated and empower them to navigate the aftermath of crime.

Suppose we adjust the victim profile to a permanently employed, early-career male with matric. The most relevant explanations for the new victim profile are shown in Table 6.4b. The best explanation here differs in terms of *first_contact*, *police_visibility*, and *protection_measures_indiv*. The remaining variables take the same values as the best explanation for our late-career female victim as presented in Table 6.4a. Where the late-career female victim would not contact anyone, the male victim would first contact a private security company. Furthermore, the male victim reported a less frequent police presence of at least once a month as opposed to the female victim who reported daily police presence. Lastly, it should be noted that the male victim possesses a weapon for personal security purposes. The consistent variable instantiation of *police_response_time* is worrying. For both victims, the police never showed up. As such, the explanations capture systemic issues with public safety in the area. Factors like nearby police stations without effective response and the perception of outside perpetrators are likely not specific to one group but likely reflect broader problems within the community.

6.4 Actionable insights: same-decision probability

The same-decision probability is typically used to understand confidence in decisions. Recall that the SDP is a threshold-based confidence measure that indicates the probability that we would make the same decision even if we had more information. A high SDP leads to a more robust decision that is less likely to change with new evidence. A low SDP might

require further information gathering before committing to a decision. Beyond this, we can leverage the same-decision probability to understand how latent evidence variables, might influence a decision, i.e., increase confidence in a decision. By identifying the variables that, on average, lead to a more robust decision, we can prioritise addressing those first.

6.4.1 Case study 3: public perception of the SAPS

Suppose we use the public's confidence in the SAPS as a decision variable, where trust in the SAPS is the positive decision. While this is not a conventional decision variable, we can use it as a *proxy* decision variable. This will help us understand the public's confidence in the SAPS and identify hidden factors that can improve trust in the SAPS. To achieve this, we provide the following steps to identify the latent evidence variable that, on average, will improve the public's confidence in the SAPS. These steps follow the decision-readiness cycle described in Section 3.3.1.

- Define the decision variable, evidence variables, latent evidence variables, and decision threshold.
- Calculate the SDP for the decision based on the observed evidence and latent evidence variables.
- Calculate the expected SDP and SDP gains for each latent evidence variable.
- The latent evidence variable with the highest SDP gain will, on average, lead to a more robust decision.

Suppose we investigate the City of Cape Town and we know there is a police presence at least once a week and no specialised police operations in the area. We want to investigate the public's confidence based on hidden binary variables such as respondents' satisfaction with the police (*satisfied_police*) and the way courts deal with perpetrators (*satisfied_courts_perps*), their experience with bribery (*bribe_ask*), and their belief in the effectiveness of court sentencing (*violent_crime_sentence*). Based on a decision threshold of 0.55, the same-decision probability is 0.474. This suggests there is a 52.6% chance that observing respondents' satisfaction with the police and the way courts deal with perpetrators, their experience with bribery, and their belief in the effectiveness of court sentencing could sway the public's confidence in the SAPS.

Table 6.5: Same-decision probabilities for case study 3.

Variable	SDP	Expected SDP
`bribe_ask`	0.195, 0.689	0.638
`satisfied_courts_perps`	0.797, 0.433	0.474
`satisfied_police`	1, 0.885	0.932
`violent_crime_sentence`	1, 0.604	0.666

Interestingly, if we change the scenario to areas in Limpopo or even Nelson Mandela Bay in the Eastern Cape, we get a same-decision probability of 1. This indicates a very high level of confidence. In these areas, police presence at least once a week and no specialised police operations in the area seem to be a strong indicator of the public's confidence in the SAPS, and the respondents' satisfaction with the police and the way courts deal with perpetrators, their experience with bribery, and their perception of the effectiveness of court sentencing might not have such a noteworthy influence in improving the public's confidence in the SAPS. However, this does not necessarily mean these resources are irrelevant in Limpopo or Nelson Mandela Bay in the Eastern Cape. They might still play a role, but their impact might be smaller.

Given the SDP of 0.474, we can leverage the SDP gain as a selection criteria to identify which latent evidence variable has, on average, the largest impact on the public's confidence in the SAPS: improving the public's satisfaction with the police or with the way courts deal with perpetrators, tackling bribery, or strengthening the public's belief in court sentencing. To do this, we need to determine the expected benefit of observing each of these variables. The same-decision probabilities and the expected same-decision probabilities for each latent evidence variable are given in Table 6.5. The corresponding SDP gains are $\mathcal{G}(bribe\_ask) = 0.192$, $\mathcal{G}(satisfied\_courts\_perps) = 0$, $\mathcal{G}(satisfied\_police) = 0.458$, and $\mathcal{G}(violent\_crime\_sentence) = 0.181$. Observing the public's satisfaction with the police will allow us to make a more robust decision that is less likely to change due to additional information. Since the same-decision probability for both states in *satisfied_police* are high, i.e., 1 and 0.885, we can motivate that interventions focused on improving the public's satisfaction with the police will increase the public's confidence in the SAPS. Consequently, the high same-decision probabilities allow us to stop information gathering since there is a low chance that observing *bribe_ask*, *satisfied_courts_perps*, and *violent_crime_sentence* will change the decision, i.e., sway the public's confidence in the SAPS.

6.5 Conclusion

This chapter demonstrated the power of explanation methods like explanation of evidence and explanation of decisions. By uncovering factors that influence victim perceptions, rising crime rates, and the public's confidence in the SAPS, these methods provide actionable insights. Explanation of evidence methods helps us trace and understand the causes of these issues, enabling the development of targeted interventions. Whereas explanation of decision methods helps identify latent evidence variables that can change a decision, i.e., which latent evidence variable will lead to a positive decision when the confidence in a decision is low. Beyond the specific examples provided, explanation methods in Bayesian networks offer a powerful tool for policymakers and law enforcement agencies across the board. Combined with other data sources like geographical information, these insights can further guide policymakers and law enforcement agencies in their efforts to develop targeted interventions. As these methods continue to evolve, they have the potential to improve crime prevention strategies, leading to safer communities.

Chapter 7

Conclusion

Explainable AI includes inherently transparent models and post-hoc explanation techniques aimed at explaining model reasoning and outputs. While a less transparent model may be potentially more accurate according to the performance-explainability trade-off (Gunning & Aha 2019), the output generated might lead to opaque decisions. This often requires a second, more transparent model to explain the black-box model.

Following the recommendations from Rudin (2019), Minh et al. (2022) this research leveraged a more transparent model and focused on enhancing the explainability thereof. Bayesian networks are considered transparent and explainable-by-design, offering insights into the model and reasoning process. However, their inner workings can present challenges for intuitive understanding, especially in more complex networks. Existing explanation methods in Bayesian networks include explanation of the model, reasoning, and evidence. To this extent, we investigate the current state of techniques associated with these three explanation categories. In contrast to previous work on explanations in Bayesian networks that utilised specialised software with built-in explanation functionalities, this thesis explores the potential of `R` packages. While these `R` packages are not explicitly designed to generate explanations, some functions can be used as a foundation for generating explanations.

7.1 Contributions to scientific research

One of the contributions of this research is to develop a taxonomy of explainable Bayesian networks. We use the existing categories defined in the literature and expand on these. For example, for the explanation of evidence category, we include methods such as the most relevant explanation as a method of abduction instead of the usual most probable explanation and maximum-a-posteriori. We then extended the categories to include explanation of decisions, which was not previously included in the literature and is considered a neglected research area in the broader XAI field. Explanation of decisions is concerned with decision-readiness; in other words, given the current evidence, are we ready to make a decision, and if not, what additional observations do we need to make an informed decision? The same-decision probability is a threshold-based measure that quantifies decision confidence in light of unobserved variables. It can also be used as a reward function in the value of information to select the following observation for decision-making. We present the XBN taxonomy along with typical questions a user may ask to emphasise the benefits of each category given a specific usage of the Bayesian network. Our prospects lead to a future where the XBN taxonomy empowers end-users. This framework will serve as a guideline, enabling end-users to understand the "how" and "why" behind predictions or observations. Notably, Valero Leal (2022) have built upon our proposed taxonomy by incorporating an explanation support category.

Secondly, this research introduces an approximate forward-gLasso search algorithm to solve the most relevant explanation in Bayesian networks. Using the gLasso, we can identify and select a reduced set of neighbours based on the most relevant dependencies, thus pruning the neighbourhood and decreasing the complexity of the most relevant explanation. Previous work shows the most relevant explanation exhibits the conditions of a *good* explanation, i.e., preciseness and conciseness. Integrating the gLasso into the forward search algorithm retains these attributes and enhances the explanations by focusing on key relationships. Bayesian networks inherently handle uncertainty, and the gLasso can capture uncertainty associated with conditional dependencies. The zero entries in the sparse inverse covariance matrix obtained from the gLasso represent conditional independence between variables, indicating uncertainty in their relationships given the observed evidence. We implemented the forward-gLasso using three different regularisation param-

eters: $\lambda = 0.01$, $\lambda = 0.001$, and $\lambda = 0.0001$. The computational efficiency of the three implementations is compared to the benchmark forward search algorithm while using the explanations obtained through a brute-force search as ground truth. The experimental results show an improvement in execution time compared to the benchmark algorithm. At the same time, the accuracy is preserved, particularly in the case of the forward-gLasso$_{\lambda_{0.0001}}$. Furthermore, our results emphasise the potential benefits of harnessing the neighbourhood pruning capabilities of the gLasso.

Thirdly, this research presents an open-source `R` package for solving explanations in Bayesian networks. While software exists for generating explanations in Bayesian networks, these are primarily proprietary and limit access. In contrast, the `XBN` `R` package not only enhances accessibility but also fosters collaboration and innovation within the research community. The `XBN` package is available for download on GitHub. We demonstrate the package use using two benchmark Bayesian networks, namely the Asia network from Lauritzen & Spiegelhalter (1988) and the insurance network from Binder et al. (1997).

Finally, this research demonstrated the usefulness of post-hoc explanation techniques in Bayesian networks on real-world data sets. The literature on post-hoc explanation techniques includes mostly implementations on benchmark Bayesian networks and synthetic data sets. This research extends the current literature by applying explanation of evidence and explanation of decision methods to the South African Victims of Crime Survey 2017 - 2018. In particular, we investigate two case studies to demonstrate the practical applicability of the most relevant explanation and one case study to demonstrate the same-decision probability. The first case study is focused on property crime in Mpumalanga. The most relevant explanation demonstrated that although there is a nearby police station, police response time is more than two hours and there is a police presence of less than once a month. The second case study is focused on a specific victim profile and perception of crime. Although this case study included more target variables, the following is most concerning. Although there is a frequent police presence in the area and a nearby police station, police never respond to a call for help. Actionable insights from these two case studies indicate that authorities should focus on improving police response time in times of need. For case study 3, we used a proxy decision variable based on the public's confidence in the SAPS to understand which latent evidence variable would, on average, improve the

public's trust in the SAPS. This analysis revealed that targeted interventions focused on the public's satisfaction with the police will have the greatest impact on increasing the public's confidence.

Other contributions of this work include exploring the impact of new evidence on the most relevant explanation. It was shown that, by adding new evidence based on the same-decision probability, the most relevant explanation is dynamic and can adapt to new evidence.

7.2 Future Work

Given the broad scope of explainability, there are several areas for future work.

- Our research highlights the neighbourhood pruning capabilities of the gLasso, thereby paving the way for further exploration into alternative neighbourhood pruning techniques that may enhance computational efficiency.

- Given the nature of the same-decision probability, can we identify a subset of features that can *sufficiently* explain the decision while decreasing the impact of *irrelevant* features? In other words, do we need to observe the complete set of evidence variables to make an informed decision, or is there a minimal set of variables that will give us probabilistic assurances that the model will behave similarly even when all variables are observed?

- Currently, the package is focused on providing post-hoc explanations, particularly the most relevant explanation. Not included in the current version is a feature to limit the search space to include only the Markov Blanket of a target node. We envisage a future update to include this feature.

- Our prospects lead to the inclusion of other explanation methods in the `R` package, such as the same-decision probability as both a stopping and selection criteria and providing natural language-based explanations of probabilities.

7.3 Limitations

While XAI literature advocates for a human-centred approach, there's a disconnect – current explanations often fail to consider user type and competency. Ideally, explanations should be adaptable, catering to diverse user needs with varying levels of complexity. To bridge this gap and achieve user-centric explanations, we believe interdisciplinary collaboration is key. By incorporating insights from AI, computer science, statistics, cognitive science, and social science, we can move towards explanations tailored to user expertise and background.

Bibliography

Abdelbar, A. M. & Hedetniemi, S. M. (1998), 'Approximating MAPs for belief networks is NP-hard and other theorems', *Artificial Intelligence* **102**(1), 21–38.

Adadi, A. & Berrada, M. (2018), 'Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)', *IEEE Access* **6**, 52138–52160.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. & Kim, B. (2018), 'Sanity checks for saliency maps', *Advances in Neural Information Processing Systems* **31**.

Almende B.V. and Contributors & Thieurmel, B. (2022), *visNetwork: network visualization using 'vis.js' library*. R package version 2.1.2.
URL: *https://CRAN.R-project.org/package=visNetwork*

Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., Chattopadhyay, S., Olson, M., Fern, A. & Burnett, M. (2020), 'Mental models of mere mortals with explanations of reinforcement learning', *ACM Transactions on Interactive Intelligent Systems (TiiS)* **10**(2), 1–37.

Ankan, A. & Panda, A. (2015), pgmpy: Probabilistic graphical models using Python, *in* 'Proceedings of the 14th Python in Science Conference (SCIPY 2015)', Citeseer.

Artelt, A. & Hammer, B. (2021), Efficient computation of contrastive explanations, *in* '2021 International Joint Conference on Neural Networks (IJCNN)', IEEE, pp. 1–9.

Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R. & Mojsilovic, A. (2021), One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques, *in* 'INFORMS 2021'.

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008), 'Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data', *The Journal of Machine Learning Research* **9**, 485–516.

Barber, D. (2012), *Bayesian Reasoning and Machine Learning*, Cambridge University Press.

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.

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* 'AIME 89: Second European Conference on Artificial Intelligence in Medicine, London, August 29th–31st 1989. Proceedings', Springer, pp. 247–256.

Bica, I., Jarrett, D., Hüyük, A. & van der Schaar, M. (2021), Learning "what-if" explanations for sequential decision-making, *in* 'International Conference on Learning Representations'.

Binder, J., Koller, D., Russell, S. & Kanazawa, K. (1997), 'Adaptive probabilistic networks with hidden variables', *Machine Learning* **29**, 213–244.

Broom, B. M., Do, K.-A. & Subramanian, D. (2012), 'Model averaging strategies for structure learning in Bayesian networks with limited data', *BMC Bioinformatics* **13**, 1–18.

Butz, R., Hommersom, A., Schulz, R. & van Ditmarsch, H. (2024), 'Evaluating the usefulness of counterfactual explanations from Bayesian networks', *Human-Centric Intelligent Systems* pp. 1–13.

Byrne, R. M. (2016), 'Counterfactual thought', *Annual Review of Psychology* **67**, 135–157.

Castillo, E., Gutierrez, J. M. & Hadi, A. S. (2012), *Expert Systems and Probabilistic Network Models*, Springer Science & Business Media.

Cath, C. (2018), 'Governing artificial intelligence: Ethical, legal and technical opportunities and challenges', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376**(2133).

Chan, H. & Darwiche, A. (2012), 'On the robustness of most probable explanations', *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2006* .

Chandrashekar, G. & Sahin, F. (2014), 'A survey on feature selection methods', *Computers & Electrical Engineering* **40**(1), 16–28.

Charniak, E. & Shimony, S. E. (1994), 'Cost-based abduction and MAP explanation', *Artificial Intelligence* **66**(2), 345–374.

Chen, S., Choi, A. & Darwiche, A. (2012), The same-decision probability: A new tool for decision making, *in* 'Proceedings of the Sixth European Workshop on Probabilistic Graphical Models', Citeseer, pp. 51–58.

Chen, S. H. & Pollino, C. A. (2012), 'Good practice in Bayesian network modelling', *Environmental Modelling & Software* **37**, 134–145.

Chen, S. J., Choi, A. & Darwiche, A. (2014), 'Algorithms and applications for the same-decision probability', *Journal of Artificial Intelligence Research* **49**, 601–633.

Choi, A., Xue, Y. & Darwiche, A. (2012), 'Same-decision probability: A confidence measure for threshold-based decisions', *International Journal of Approximate Reasoning* **53**(9), 1415–1428.

Chuang, Y.-N., Wang, G., Yang, F., Liu, Z., Cai, X., Du, M. & Hu, X. (2023), 'Efficient XAI techniques: A taxonomic survey', *arXiv preprint arXiv:2302.03225* .

Dahl, J., Vandenberghe, L. & Roychowdhury, V. (2008), 'Covariance selection for non-chordal graphs via chordal embedding', *Optimization Methods & Software* **23**(4), 501–520.

D'Ambrosio, B. (1999), 'Inference in Bayesian networks', *AI Magazine* **20**(2), 21–21.

Darwiche, A. (2009), *Inference by Variable Elimination*, Cambridge University Press, p. 126–151.

Das, A. & Rad, P. (2020), 'Opportunities and challenges in explainable artificial intelligence (XAI): A survey', *arXiv preprint arXiv:2006.11371* .

Dawid, A. P. (1979), 'Conditional independence in statistical theory', *Journal of the Royal Statistical Society: Series B (Methodological)* **41**(1), 1–15.

Dawid, A. P. (1992), 'Applications of a general propagation algorithm for probabilistic expert systems', *Statistics and Computing* **2**(1), 25–36.

De Campos, L. M., Fernández-Luna, J. M. & Puerta, J. M. (2003), 'An iterated local search algorithm for learning Bayesian networks with restarts based on conditional independence tests', *International Journal of Intelligent Systems* **18**(2), 221–235.

de Waal, A. & Joubert, J. W. (2022), 'Explainable Bayesian networks applied to transport vulnerability', *Expert Systems with Applications* **209**, 118348.

Derks, I. P. & de Waal, A. (2020), A taxonomy of explainable Bayesian networks, *in* A. Gerber, ed., 'Artificial Intelligence Research', Springer International Publishing, Cham, pp. 220–235.

Dew, N. (2007), 'Abduction: A pre-condition for the intelligent design of strategy', *Journal of Business Strategy* **28**(4), 38–45.

Diakopoulos, N. & Koliska, M. (2017), 'Algorithmic transparency in the news media', *Digital Journalism* **5**(7), 809–828.

Dittmer, S. L. & Jensen, F. V. (1997), Myopic value of information for influence diagrams, *in* 'Myopic Value of Information for Influence Diagrams', Morgan Kaufmann, pp. 142–149.

Djulbegovic, B., Hozo, I., Mayrhofer, T., van den Ende, J. & Guyatt, G. (2019), 'The threshold model revisited', *Journal of Evaluation in Clinical Practice* **25**(2), 186–195.

Djulbegovic, B., van den Ende, J., Hamm, R. M., Mayrhofer, T., Hozo, I., Pauker, S. G. & (ITWG), I. T. W. G. (2015), 'When is rational to order a diagnostic test, or prescribe

treatment: The threshold model as an explanation of practice variation', *European Journal of Clinical Investigation* **45**(5), 485–493.

Doshi-Velez, F. & Kim, B. (2017), 'Towards a rigorous science of interpretable machine learning', *arXiv preprint arXiv:1702.08608* .

Druzdzel, M. J. (1993), *Probabilistic Reasoning in Decision Support Systems: From Computation to Common Sense*, Carnegie Mellon University.

Druzdzel, M. J. (1996), 'Qualitative verbal explanations in Bayesian belief networks', *Artificial Intelligence and Simulation of Behaviour Quarterly* **94**(July), 43–54.

Düspohl, M., Frank, S. & Döll, P. (2012), 'A review of Bayesian networks as a participatory modeling approach in support of sustainable environmental management', *Journal of Sustainable Development* **5**(12).

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G. et al. (2023), 'Explainable AI (XAI): Core ideas, techniques, and solutions', *ACM Computing Surveys* **55**(9), 1–33.

Elidan, G., Lotner, N., Friedman, N. & Koller, D. (2000), 'Discovering hidden variables: a structure-based approach', *Advances in Neural Information Processing Systems* **13**.

Escalante, H. J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., van Gerven, M. & van Lier, R. (2018), *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer.

Fitelson, B. (2001), Studies in Bayesian Confirmation Theory, PhD thesis, University of Wisconsin, Madison.

Flores, M. J., Gámez, J. A. & Moral, S. (2005), Abductive inference in Bayesian networks: finding a partition of the explanation space, *in* 'Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 8th European Conference, ECSQARU 2005, Barcelona, Spain, July 6-8, 2005. Proceedings 8', Springer, pp. 63–75.

Främling, K. (2020), Decision theory meets explainable AI, *in* 'International Workshop on Explainable, Transparent Autonomous Agents and Multi-agent Systems', Springer, pp. 57–74.

Franzin, Alberto, Sambo, Francesco, Camillo, D. & Barbara (2017), 'bnstruct: an R package for Bayesian network structure learning in the presence of missing data', *Bioinformatics* **38**(8), 1250–1252.

Friedman, J., Hastie, T. & Tibshirani, R. (2008), 'Sparse inverse covariance estimation with the graphical Lasso', *Biostatistics* **9**(3), 432–441.

Futia, G. & Vetrò, A. (2020), 'On the integration of knowledge graphs into deep learning models for a more comprehensible AI — Three challenges for future research', *Information* **11**(2), 122.

Gallego, M. J. F. (2005), Bayesian networks inference: Advanced algorithms for triangulation and partial abduction, PhD thesis, Universidad de Castilla La Mancha.

Gámez, J. A. (2004), 'Abductive inference in Bayesian networks: A review', *Advances in Bayesian Networks* pp. 101–120.

Gelsema, E. S. (1995), Abductive reasoning in Bayesian belief networks using a genetic algorithm, *in* 'Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics', PMLR, pp. 245–251.

Ghai, B., Liao, Q. V., Zhang, Y., Bellamy, R. & Mueller, K. (2021), 'Explainable active learning (XAL) toward AI explanations as interfaces for machine teachers', *Proceedings of the ACM on Human-Computer Interaction* **4**(CSCW3), 1–28.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. & Kagal, L. (2019), 'Explaining explanations: An overview of interpretability of machine learning', *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018* pp. 80–89.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P. & Holzinger, A. (2018), Explainable AI: the new 42?, *in* 'International Cross-domain Conference for Machine Learning and Knowledge Extraction', Springer, pp. 295–303.

Good, I. J. (1950), Probability and the weighing of evidence, Technical report, C. Griffin London.

Greene, D., Hoffmann, A. L. & Stark, L. (2019), 'Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning', *Proceedings of the 52nd Hawaii International Conference on System Sciences* pp. 2122–2131.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. & Pedreschi, D. (2018), 'A survey of methods for explaining black box models', *ACM Computing Surveys (CSUR)* **51**(5), 1–42.

Gunning, D. & Aha, D. W. (2019), 'DARPA's explainable artificial intelligence program', *AI Magazine* **40**(2), 44–58.

Guo, H., Boddhireddy, P. R. & Hsu, W. H. (2005), An ACO algorithm for the most probable explanation problem, *in* 'AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17', Springer, pp. 778–790.

Haddawy, P., Jacobson, J. & Kahn Jr, C. E. (1997), 'BANTER: A Bayesian network tutoring shell', *Artificial Intelligence in Medicine* **10**(2), 177–200.

Hansen, K. D., Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F. & Sarkar, D. (2023), *Rgraphviz: Provides plotting capabilities for R graph objects.* R package version 2.44.0.
URL: *https://bioconductor.org/packages/Rgraphviz*

Helldin, T. & Riveiro, M. (2009), Explanation methods for Bayesian networks: Review and application to a maritime scenario, *in* 'Proceedings of The 3rd Annual Skövde Workshop on Information Fusion Topics (SWIFT 2009)', pp. 11–16.

Henrion, M. & Druzdzel, M. J. (1990), Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning, *in* 'Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence, UAI 1990', Elsevier, pp. 10–20.

Højsgaard, S. (2012), 'Graphical independence networks with the gRain package for R', *Journal of Statistical Software* **46**(10), 1–26.

Jafta, G., de Waal, A., Derks, I. & Ruttkamp-Bloem, E. (2021), Evaluation of XAI as an enabler for fairness, accountability and transparency, *in* 'Proceedings of the Second Southern African Conference for Artificial Intelligence Research: SACAIR 2021'.

Jensen, F. V. & Nielsen, T. D. (2007), *Bayesian Networks and Decision Graphs*, Vol. 2, Springer.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Keane, M. T. & Smyth, B. (2020), Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI), *in* 'Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings 28', Springer, pp. 163–178.

Kenny, E. M. & Keane, M. T. (2021), On generating plausible counterfactual and semi-factual explanations for deep learning, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 35, pp. 11575–11585.

Kim, J. H. & Pearl, J. (1983), A computational model for causal and diagnostic reasoning in inference systems, *in* 'International Joint Conference on Artificial Intelligence'.

Kłopotek, M., Michalewicz, M., Wierzchoń, S. T., Oniśko, A., Druzdzel, M. J. & Wasyluk, H. (2000), Extension of the Hepar II model to multiple-disorder diagnosis, *in* 'Intelligent Information Systems: Proceedings of the IIS'2000 Symposium, Bystra, Poland, June 12–16, 2000', Springer, pp. 303–313.

Kochenderfer, M. J. (2015), *Decision Making Under Uncertainty: Theory and Application*, MIT press.

Koller, D. & Friedman, N. (2009), *Probabilistic Graphical Models: Principles and Techniques*, MIT press.

Kolodner, J. (2014), *Case-based Reasoning*, Morgan Kaufmann.

Koopman, T. (2020), Computing contrastive, counterfactual explanations for Bayesian networks, Master's thesis, Utrecht University.

Koopman, T. & Renooij, S. (2021), Persuasive contrastive explanations for Bayesian networks, *in* 'Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21–24, 2021, Proceedings 16', Springer, pp. 229–242.

Korb, K. B. & Nicholson, A. E. (2010), *Bayesian Artificial Intelligence*, CRC press.

Krause, A. & Guestrin, C. (2009), 'Optimal value of information in graphical models', *Journal of Artificial Intelligence Research* **35**, 557–591.

Kwisthout, J. (2013*a*), Most inforbable explanations: finding explanations in Bayesian networks that are both probable and informative, *in* 'European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty', Springer, pp. 328–339.

Kwisthout, J. (2013*b*), Structure approximation of most probable explanations in Bayesian networks, *in* 'European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty', Springer, pp. 340–351.

Kwisthout, J. (2015), 'Most frugal explanations in Bayesian networks', *Artificial Intelligence* **218**, 56–73.

Kyrimi, E. & Marsh, W. (2016), A progressive explanation of inference in 'hybrid' Bayesian networks for supporting clinical decision making, *in* 'Conference on Probabilistic Graphical Models', PMLR, pp. 275–286.

Lacave, C. & Díez, F. J. (2002), 'A review of explanation methods for Bayesian networks', *Knowledge Engineering Review* **17**(2), 107–127.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A. & Baum, K. (2021), 'What do we want from explainable artificial intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research', *Artificial Intelligence* **296**, 103473.

Lauritzen, S. L. & Spiegelhalter, D. J. (1988), 'Local computations with probabilities on graphical structures and their application to expert systems', *Journal of the Royal Statistical Society: Series B (Methodological)* **50**(2), 157–194.

Leake, D. & Mcsherry, D. (2005), 'Introduction to the special issue on explanation in case-based reasoning', *The Artificial Intelligence Review* **24**(2), 103.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A. & Vinck, P. (2018), 'Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges', *Philosophy & Technology* **31**, 611–627.

Leslie, D. (2019), 'Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector', *SSRN* .

Li, Z. & D'Ambrosio, B. (1993), An efficient approach for finding the MPE in belief networks, *in* 'Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence, UAI 1993', Morgan Kaufmann, pp. 342–349.

Lim, B. Y. & Dey, A. K. (2009), Assessing demand for intelligibility in context-aware applications, *in* 'Proceedings of the 11th International Conference on Ubiquitous Computing', pp. 195–204.

Lim, B. Y. & Dey, A. K. (2013), Evaluating intelligibility usage and usefulness in a context-aware application, *in* 'Human-Computer Interaction. Towards Intelligent and Implicit Interaction: 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part V 15', Springer, pp. 92–101.

Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. (2020), 'Explainable AI: A review of machine learning interpretability methods', *Entropy* **23**(1), 18.

Lipton, Z. C. (2018), 'The mythos of model interpretability', *Queue* **16**(3), 31–57.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P. & Holzinger, A. (2020), Explainable artificial intelligence: Concepts, applications, research challenges and visions, *in* 'International Cross-domain Conference for Machine Learning and Knowledge Extraction', Springer, pp. 1–16.

Lötsch, J., Kringel, D. & Ultsch, A. (2022), 'Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients', *BioMedInformatics* **2**(1), 1–17.

Lu, T.-C. & Przytula, K. W. (2006), Focusing strategies for multiple fault diagnosis, *in* 'FLAIRS Conference', Citeseer, pp. 842–847.

Lundberg, S. M. & Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Proceedings of the 31st International Conference on Neural Information Processing Systems', pp. 4768–4777.

Madigan, D., Mosurski, K. & Almond, R. G. (1997), 'Graphical explanation in belief networks', *Journal of Computational and Graphical Statistics* **6**(2), 160–181.

Marinescu, R. & Dechter, R. (2007), Best-first AND/OR search for most probable explanations, *in* 'Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2007', AUAI Press, pp. 259–266.

Marriott, K., Moulder, P., Hope, L. & Twardy, C. (2005), Layout of Bayesian networks, *in* 'ACM International Conference Proceeding Series', Vol. 102, Citeseer, pp. 97–106.

Meinshausen, N. & Bühlmann, P. (2006), 'High-dimensional graphs and variable selection with the Lasso', *The Annals of Statistics* **34**(3), 1436 – 1462.

Mengshoel, O. J. & Wilkins, D. C. (1998), Abstraction for belief revision: Using a genetic algorithm to compute the most probable explanation, *in* 'Proceedings AAAI Spring Symposium Series on Satisficing Models'.

Mengshoel, O. J., Wilkins, D. C. & Roth, D. (2010), 'Initialization and restart in stochastic local search: Computing a most probable explanation in Bayesian networks', *IEEE Transactions on Knowledge and Data Engineering* **23**(2), 235–247.

Miller, T. (2019), 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence* **267**, 1–38.

Minh, D., Wang, H. X., Li, Y. F. & Nguyen, T. N. (2022), 'Explainable artificial intelligence: A comprehensive review', *Artificial Intelligence Review* pp. 1–66.

Mittelstadt, B., Russell, C. & Wachter, S. (2019), Explaining explanations in AI, *in* 'Proceedings of the Conference on Fairness, Accountability, and Transparency', pp. 279–288.

Moe, S. J., Carriger, J. F. & Glendell, M. (2021), 'Increased use of Bayesian network models has improved environmental risk assessments', *Integrated Environmental Assessment and Management* **17**(1), 53–61.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. (2019), 'Definitions, methods, and applications in interpretable machine learning', *Proceedings of the National Academy of Sciences* **116**(44), 22071–22080.

Nagarajan, R., Scutari, M. & Lèbre, S. (2013), 'Bayesian networks in R', *Springer* **122**, 125–127.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M. & Seifert, C. (2023), 'From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI', *ACM Computing Surveys* **55**(13s), 1–42.

Niculescu-Mizil, A. & Caruana, R. (2007), Inductive transfer for Bayesian network structure learning, *in* 'Artificial intelligence and statistics', PMLR, pp. 339–346.

Ong, L. S., Shepherd, B., Tong, L. C., Seow-Choen, F., Ho, Y. H., Tang, C. L., Ho, Y. S. & Tan, K. (1997), 'The colorectal cancer recurrence support (CARES) system', *Artificial Intelligence in Medicine* **11**(3), 175–188.

Park, J. D. (2002), MAP complexity results and approximation methods, *in* 'Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, UAI 2002', Morgan Kaufmann, pp. 388–396.

Park, J. D. & Darwiche, A. (2004), 'Complexity results and approximation strategies for MAP explanations', *Journal of Artificial Intelligence Research* **21**, 101–133.

Parson, E. A. (2008), 'Useful global-change scenarios: Current issues and challenges', *Environmental Research Letters* **3**(4), 045016.

Pauker, S. G. & Kassirer, J. P. (1980), 'The threshold approach to clinical decision making', *New England Journal of Medicine* **302**(20), 1109–1117.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann.

Pearl, J. (2009), *Causality: Models, Reasoning and Inference*, 2nd edn, Cambridge University Press, USA.

Peng, Y. & Reggia, J. A. (2012), *Abductive Inference Models for Diagnostic Problem-solving*, Springer Science & Business Media.

Petitet, P., Attaallah, B., Manohar, S. G. & Husain, M. (2021), 'The computational cost of active information sampling before decision-making under uncertainty', *Nature Human Behaviour* **5**(7), 935–946.

Poole, D. L. & Provan, G. M. (1990), What is an optimal diagnosis?, *in* 'Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI 1990', Elsevier, pp. 46–53.

Preece, A., Harborne, D., Braines, D., Tomsett, R. & Chakraborty, S. (2018), Stakeholders in explainable AI, *in* 'Proceedings of the AAAI FSS-18: Artificial Intelligence in Government and Public Sector'.

Quesada, D. (2022), *dbnR: dynamic Bayesian network learning and inference*. R package version 0.7.8.
URL: *https://CRAN.R-project.org/package=dbnR*

R Core Team (2020), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Raiffa, H. & Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Harvard University Graduate School of Business Administration.

Ras, G., van Gerven, M. & Haselager, P. (2018), 'Explanation methods in deep learning: users, values, concerns and challenges', *Explainable and Interpretable Models in Computer Vision and Machine Learning* pp. 19–36.

Renooij, S. (2018), Same-decision probability: Threshold robustness and application to explanation, *in* 'International Conference on Probabilistic Graphical Models', PMLR, pp. 368–379.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016), "Why should I trust you?" Explaining the predictions of any classifier, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 1135–1144.

Rosenfeld, A. & Richardson, A. (2019), 'Explainability in human-agent systems', *Autonomous Agents and Multi-Agent Systems* **33**, 673–705.

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* **1**(5), 206–215.

Rudin, C. & Radin, J. (2019), 'Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition', *Harvard Data Science Review* **1**(2), 10–1162.

Saad, G. & Russo, J. E. (1996), 'Stopping criteria in sequential choice', *Organizational Behavior and Human Decision Processes* **67**(3), 258–270.

Saeed, W. & Omlin, C. (2023), 'Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities', *Knowledge-Based Systems* **263**, 110273.

Salimans, T., Kingma, D. & Welling, M. (2015), Markov chain Monte Carlo and variational inference: Bridging the gap, *in* 'International Conference on Machine Learning', PMLR, pp. 1218–1226.

Santos Jr, E. (1991), On the generation of alternative explanations with implications for belief revision, *in* 'Uncertainty Proceedings 1991', Elsevier, pp. 339–347.

Savage, L. J. (1972), *The Foundations of Statistics*, Courier Corporation.

Schreiber, J. (2018), 'Pomegranate: Fast and flexible probabilistic modeling in Python', *Journal of Machine Learning Research* **18**(164), 1–6.

Scutari, M. (2010), 'Learning Bayesian networks with the bnlearn R package', *Journal of Statistical Software* **35**(3), 1–22.

Seroussi, B. & Golmard, J.-L. (1994), 'An algorithm directly finding the k most probable configurations in Bayesian networks', *International Journal of Approximate Reasoning* **11**(3), 205–233.

Shimony, S. E. (1994), 'Finding MAPs for belief networks is NP-hard', *Artificial Intelligence* **68**(2), 399–410.

Shimony, S. E. & Charniak, E. (1990), A new algorithm for finding MAP assignments to belief networks, *in* 'Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI 1990', Elsevier, pp. 185–196.

Srinivasan, R. & Chander, A. (2021), Explanation perspectives from the cognitive sciences — A survey, *in* 'Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence', pp. 4812–4818.

Statistics South Africa (2018), 'Victims of Crime Survey 2017-2018 [dataset]'.

Suermondt, H. J. (1992), *Explanation in Bayesian Belief Networks*, Stanford University.

Sun, W. & Chang, K. (2011), Study of the most probable explanation in hybrid Bayesian networks, *in* 'Signal Processing, Sensor Fusion, and Target Recognition XX', Vol. 8050, SPIE, pp. 338–345.

Sustik, M. A., Calderhead, B. & Clavel, J. (2023), *glassoFast: fast graphical Lasso*. R package version 1.0.1.
URL: *https://CRAN.R-project.org/package=glassoFast*

Sy, B. K. (1993), 'A recurrence local computation approach towards ordering composite beliefs in Bayesian belief networks', *International Journal of Approximate Reasoning* **8**(1), 17–50.

Thorburn, W. M. (1918), 'The myth of Occam's razor', *Mind* **27**(107), 345–353.

Tsamardinos, I., Brown, L. E. & Aliferis, C. F. (2006), 'The max-min hill-climbing Bayesian network structure learning algorithm', *Machine Learning* **65**, 31–78.

Valero Leal, E. (2022), Explanations for dynamic Bayesian networks: A case study in climate science, PhD thesis, ETSI_Informatica.

Van Der Gaag, L. C. & Bodlaender, H. L. (2011), On stopping evidence gathering for diagnostic Bayesian networks, *in* 'Symbolic and Quantitative Approaches to Reasoning

with Uncertainty: 11th European Conference, ECSQARU 2011, Belfast, UK, June 29–July 1, 2011. Proceedings 11', Springer, pp. 170–181.

Van Der Gaag, L. C. & Coupé, V. M. (1999), Sensitivity analysis for threshold decision making with Bayesian belief networks, *in* 'Congress of the Italian Association for Artificial Intelligence', Springer, pp. 37–48.

Vlek, C., Prakken, H., Renooij, S. & Verheij, B. (2015), Constructing and understanding Bayesian networks for legal evidence with scenario schemes, *in* 'Proceedings of the 15th International Conference on Artificial Intelligence and Law', pp. 128–137.

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**(3), 285–324.

Wachter, S., Mittelstadt, B. & Russell, C. (2017), 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR', *Harvard Journal of Law & Technology* **31**, 841.

Wald, A. (1949), 'Statistical decision functions', *The Annals of Mathematical Statistics* **20**(2), 165–205.

Wang, S.-d., Wang, X.-c. & Zhang, H.-b. (2015), 'Simulation on optimized allocation of land resource based on DE-CA model', *Ecological Modelling* **314**, 135–144.

Wick, M. R. (1989), 'The 1988 AAAI workshop on explanation', *AI Magazine* **10**(3), 22–22.

Xing, E. P., Jordan, M. I. & Russell, S. (2002), A generalized mean field algorithm for variational inference in exponential families, *in* 'Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence, UAI 2003', Morgan Kaufmann, pp. 583–591.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D. & Zhu, J. (2019), 'Explainable AI: A brief survey on history, research areas, approaches and challenges', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11839 LNAI**, 563–574.

Yap, G.-E., Tan, A.-H. & Pang, H.-H. (2008), 'Explaining inferences in Bayesian networks', *Applied Intelligence* **29**, 263–278.

Yuan, C., Lim, H. & Littman, M. L. (2011), 'Most relevant explanation: Computational complexity and approximation methods', *Annals of Mathematics and Artificial Intelligence* **61**(3), 159–183.

Yuan, C., Lim, H. & Lu, T. C. (2011), 'Most relevant explanation in Bayesian networks', *Journal of Artificial Intelligence Research* **42**, 309–352.

Yuan, C., Liu, X., Lu, T. C. & Lim, H. (2009), 'Most relevant explanation: Properties, algorithms, and evaluations', *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2009* pp. 631–638.

Yuan, C. & Lu, T.-C. (2008), A general framework for generating multivariate explanations in Bayesian networks, *in* 'AAAI', pp. 1119–1124.

Zeng, Y., Luo, J. & Lin, S. (2009), Classification using Markov blanket for feature selection, *in* '2009 IEEE International Conference on Granular Computing', IEEE, pp. 743–747.

Zhu, J., Wang, H., Hovy, E. & Ma, M. (2010), 'Confidence-based stopping criteria for active learning for data annotation', *ACM Transactions on Speech and Language Processing (TSLP)* **6**(3), 1–24.

Appendix A

List of abbreviations and symbols

This section offers a summary of the key abbreviations and notations frequently used in this thesis.

AET	Average execution time in seconds.
AI	Artificial Intelligence.
BNs	Bayesian networks.
CSE	Total test cases solved exactly.
GBF	Generalised Bayes factor.
gLasso	Graphical Least Absolute Shrinkage and Selection Operator.
MAP	Maximum a posteriori.
ML	Machine learning.
MPE	Most probable explanation.
MRE	Most relevant explanation.
SAPS	South African Police Service.
SDP	Same-decision probability.
VCS	Victims of Crime Survey.
VOI	Value of information, also referred to as expected benefit.
XAI	Explainable Artificial Intelligence.
XBN	Explainable Bayesian networks.

$\mathcal{E}(G, H, e, T)$	Expected SDP of observing variables G out of latent evidence H.
$\mathcal{ER}(R, D, \mathbf{H}, \mathbf{e})$	Expected reward using a reward-based value function R, a hypothesis variable D, evidence e, and unobserved variables $\mathbf{H}$.
$\mathcal{G}(G)$	SDP gain of observing variables G out of latent evidence variables H.
$GBF(x, e)$	Generalised Bayes factor for observed evidence e and explanation x.
$MRE(M, e)$	Most relevant explanation for observed evidence e and set of target variables M.
$\mathcal{V}(R, D, \mathbf{H}, \mathbf{e})$	Value of information or expected benefit observing the variable $\mathbf{H}$.
$SDP(d, e, H, T)$	Same-decision probability for decision d, evidence e, threshold T, and latent evidence variables H.

Appendix B

Description of variables used

Table B.1: Description of variables included in the Insurance Bayesian network from Binder et al. (1997).

Variable name	Description
`Accident`	Severity of the accident.
`Age`	Age group.
`Airbag`	Vehicle equipped with an airbag.
`Antilock`	Vehicle equipped with an anti-lock braking system.
`AntiTheft`	Vehicle equipped with an anti-theft system.
`CarValue`	Vehicle value.
`Cushioning`	Impact absorption.
`DrivHist`	Driver accident history.
`DrivingSkill`	Driver driving skill.
`DrivQuality`	Driver driving quality.
`GoodStudent`	Is driver a good student.
`HomeBase`	Neighbourhood type.
`ILiCost`	Inspection cost
`MakeModel`	Vehicle model.
`MedCost`	Cost of medical treatment.
`Mileage`	Vehicle mileage.
`OtherCar`	Was another vehicle involved in the accident.
`OtherCarCost`	Cost of other vehicle.
`PropCost`	Cost ratio of the vehicles involved.
`RiskAversion`	Drivers' risk aversion.
`RuggedAuto`	Ruggedness of vehicle.
`SeniorTrain`	Advances driving course taken.
`SocioEcon`	Socio-economic status.
`Theft`	Theft of vehicle.
`ThisCarCost`	Cost of the insured vehicle.
`ThisCarDam`	Damage cost of this vehicle.
`VehicleYear`	Vehicle age.

Table B.2: Description of variables included in the South African VCS 2017 - 2018. Variables acc - pol.

Variable name	Description
access_instituitions	Respondent access to institutions in time of need.
age_group	Respondents' age group.
attend_court	Did the respondent visit the court in the past 12 months?
bribe_ask	Was the respondent personally asked to pay a bribe?
child_approach_police	Would the respondent teach their child to approach the police in time of need?
corrupt_change	Respondent perception of change in corruption levels.
counselling_access_yn	Respondents' access to counselling services.
crime_experienced	Respondents' experience of crime in the last five years.
economic_activity	Respondents' nature of work.
education	Respondents' education.
fear_crime_prevent_actions	Fear of crime prevents respondent from participating in activities.
first_contact	First person to call for help in time of need.
forum	Access to a forum that discusses crime in the neighbourhood.
gender	Respondents' gender
government_spending	Respondents' perception of government spending to reduce crime.
level_property_crime_changed	Changes in the level of property crime in the area.
level_violent_crime_changed	Changes in the level of violent crime in the area.
medical_access_yn	Respondents' access to medical services.
metro_code	Metro code
nearest_court_yn	Does the respondent know where the nearest magistrate court is?
police_mean_time	Average time to get to the nearest police station.
police_response_time	Average response time to an emergency call.
police_visibility	Extent of visible policing in the area.

Table B.3: Description of variables included in the South African VCS 2017 - 2018. Variables pr - why.

Variable name	Description
pr_code	South African provinces.
prison	Whether the respondent has been to prison for any reason in the past 12 months.
property_crime_committed_by	Respondents' perception about the origin of property crime perpetrators.
protect_groups_yn	Access to organisations or groups other than the SAPS for protection.
protection_groups	Type of organisation or group that provides protection.
protection_measures_indiv	Respondents' measures to protect themselves against crime.
report_corrupt_yn	Does the respondent know where to report corruption?
safer_after_measures	Does the respondent feel safer after taking measures to protect themselves after crime?
satisfied_courts_perps	Is the respondent satisfied with the way courts deal with perpetrators of crime?
satisfied_police	Is the respondent satisfied with the SAPS?
satisfied_rehab_criminals	Is the respondent satisfied with the way correctional services rehabilitates criminals?
shelter_access_yn	Access to a place of safety.
shelter_mean_time	Average time to place of safety.
shelter_nature	Nature of the place of safety.
spec_operations	Specialised police operations in the last 12 months.
specop_reduce_crime	Respondents' perception of specialised police operations to reduce crime.
topic_of_crime	Has the topic of crime come up in conversation in the last 2 weeks?
trust_metro	Public confidence in the metro/traffic police.
trust_saps	Public confidence in the SAPS.
type_crime_afraid	Crimes that are mostly feared.
type_crime_occur	Crimes that occur most in the area.
violent_crime_committed_by	Respondents' perception about the origin of violent crime perpetrators.
violent_crime_sentence	Respondents' perception on sentences served for violent crimes.
walk_alone_dark	Respondents' attitude toward walking alone when it is dark.
walk_alone_day	Respondents' attitude toward walking alone during the day.
why_crime_commit	Respondents' perception about the cause of people committing a crime.