June 2022

Extending MAP-independence for Bayesian network explainability

Enrique VALERO-LEAL [a], Pedro LARRAÑAGA [a] and Concha BIELZA [a]
[a] *Computational Intelligence Group, Universidad Politécnica de Madrid*

Abstract. One of the open challenges of explainable AI is to ensure quality in the explanations. In this work, we expand the idea of MAP-independence in Bayesian networks and explore its properties. We believe that this concept is related to explanation stability and that MAP-independence can potentially be used to measure it and improve explanations of an evidence in those probabilistic graphical models. Stability is a measure of quality in explanations that refers to maintain a similar explanation for similar cases.

Keywords. Bayesian networks, explainable AI, robustness, stability

1. Introduction

In the past years there has been an increasing interest in explainable AI (XAI), since it can be a potential solution to the performance, ethical and legal concerns of the new obscure complex models such as neural networks. Selecting transparent models over top performing ones can be a better option in terms of both performance and explainability [1]. As such, in this work we use Bayesian networks.

Stability is a desirable property of explanations [2] which consists of avoiding that a minimal change in the input data leads to a significant modification in the explanation. We review MAP-independence [3] as a measure related to stability for Bayesian network explanations and formulate properties that improve computational efficiency and expand the concept to continuous domains.

In Section 2, we review the basic concepts concerning our proposal, which is presented in Section 3. In Section 4, conclusions about our work are drawn.

2. Literature review

2.1. Bayesian networks

A Bayesian network [4,5] $\mathscr{B} = (\mathscr{G}, \theta)$ is a probabilistic graphical model that encodes a joint probability distribution (JPD) $P(X_1, ..., X_n)$ over a set of variables $X = \{X_1, ..., X_n\}$. Qualitatively, a Bayesian network is a directed acyclic graph $\mathscr{G}$ that represents the conditional (in)dependences between the variables X and, quantitatively, Bayesian networks factorise the JPD in the vector of parameters $\theta = (\theta_1, ..., \theta_n)$, storing only lo-

cal conditional probability distributions (CPDs) of each node X_i given its parents, $\theta_i = P(X_i|Pa_{X_i}), \forall X_i \in X$.

An interesting type of query is the partial abduction: given an evidence e that is desired to be explained, the maximum a posteriori (MAP) h^* of a subset H of the set unobserved nodes is computed, $h^* = arg\,max_H P(H|e)$ with $H \subset X \setminus E$.

2.2. Bayesian network explanations

Bayesian network explanations can be classified according to what they are explaining. Lacave and Díez [6] distinguish three types of explanations (model, reasoning and evidence), and Derkas and De Waal [7] add a fourth one (decision):

1. Model explanations aim to explain the graph topology and JPD.
2. Reasoning explanations try to measure how likely is a conclusion given a certain evidence by means of explaining the reasoning followed.
3. Evidence explanations answer why some variables are in a particular state (evidence) using other variables. A MPE query falls under this category.
4. Decision explanation refers to answering whether we have enough information to make a decision or not.

2.3. MAP-independence

In MAP queries, intuitively, we consider an unobserved node to be relevant if observing it may alter the value assignment of the set H. Relevance in Bayesian networks is traditionally modeled as (in)conditional independence [8], where a node is said to be irrelevant if it is conditionally independent of H given e.

According to Kwisthout [3], this definition of relevance is too strict because there can be context-specific irrelevances that are not found using conditional independence and, as such, introduces the concept of MAP-independence, $MInd(h^*, e, R)$ to verify if a subset of unobserved R are relevant or irrelevant. Given a MAP query, an explanation $h^* = arg\,max_H P(H|e)$ is MAP-independent of the subset R if, for every value assignment $r \in \Omega(R)$, the explanation h^* remains unchanged, i.e. $\forall r \in \Omega(R),\ h^* = arg\,max_H P(H, r|e)$.

3. Proposal

3.1. Limitation of the original proposal

We claim that the notion of MAP-independence can potentially be very useful to verify whether the explanation is more or less stable. However, we still find important limiting factors that we aim to solve in the next Subsections.

First, while the intention of MAP-independence is to relax the strict criteria to render a node as (ir)relevant, the original proposal is in itself still too strict. It is necessary that all possible value assignments r of R hold the MAP-independence condition: $arg\,max_H P(H|e) = arg\,max_H P(H, r|e)$. However, there might be value assignments r of R that do not hold the condition and they have a very low probability given the evidence (i.e. very unlikely).

Second, some properties of MAP-independence have not been explored yet. They could potentially prune the search space if we are looking for MAP-independences of a given explanation and many different sets of variables R.

Finally, the original proposal is unsuitable for networks that contain continuous variables due to two reasons. First, comparing the equality of the value assignment with maximum density for H for the two densities $P(H|e)$ and $P(H,r|e)$ is difficult, since it will be a real number (specifically, the mode of the densities). Second, we need to verify such equality for every value assignment r of R, which is impossible when the domain of R is infinite, as in the continuous case.

3.2. MAP-independence strength

We introduce the concept of MAP-independence strength to address the problem of the original proposal being too strict. The MAP-independence strength can be defined as the sum of the probability of all value assignments r of R in a MAP query such that hold the MAP-independence condition holds, $\Omega^{MInd}_{H,e}(R)$. MAP-independence strength measures how (ir)relevant a set of nodes R is, rather than simply indicating whether it is completely irrelevant or not. The concept provides the probability of an observation over R (to not) to alter the explanation h^*.

This concept reminds to same-decision probability (SDP) [9], but there are some key differences: Here we are interested in h^*, whereas in SDP the value of interest h is user-defined. In addition, we check the MAP-independence of h^* and r and SDP checks if $P(h|r,e)$ is a higher than a threshold.

Let $\Omega^{MInd}_{H,e}(R)$ be the set of all value assignments r of R that hold MAP-independence. Then, we formally define the MAP-independence strength:

$$MIndStrength(h^*,R,e) = \sum_{r\in\Omega^{MInd}(R,H,e)} P(r|e), \text{ where}$$

$$\Omega^{MInd}_{H,e}(R) = \{r \in \Omega(R) \mid h^* = arg\,max_H P(H|e) = arg\,max_H P(H,r|e)\}.$$

3.3. Properties

Theorem 3.1. *Given a Bayesian network, a MAP query and two subsets of intermediate nodes R_i and R, such that $R_i \subseteq R$; if h^* is MAP-independent from R, then h^* is also MAP-independent from R_i.*

Proof. We denote $r_j \in \Omega(R_j)$, with $R_j = R \setminus R_i$. As $h^* = arg\,max_H P(H,r_i,r_j|e)$, then $P(h^*,r_i,r_j|e) > P(h,r_i,r_j|e)$, for all $h \neq h^*, h \in \Omega(H)$. Then, by marginalising over R_j, we also have $P(h^*,r_i|e) > P(h,r_i|e)$, and thus, $arg\,max_H P(H,r_i|e) = h^*$, which proves the result. □

Theorem 3.2. *Given a Bayesian network, a MAP query and two subsets of nodes, R_i and R_j; if h^* is MAP-independent from R_i and R_j is conditionally independent of H given R_i and e then h^* is MAP-independent from R_j.*

Proof. We have $h^* = arg\,max_H P(H,r_i|e)$ and also $h^* = arg\,max_H P(H|r_i,e)$, since $P(H,r_i|e) \propto P(H|r_i,e)$. As in Theorem 3.1, this gives that $P(h^*|r_i,e) > P(h|r_i,e)$, for all $h \neq h^*, h \in \Omega(H)$.

As R_j is conditionally independent of H given R_i and e, we have that $P(H|r_i,e) = P(H|r_i,r_j,e)$, $\forall r_j \in \Omega(R_j)$, and hence $P(h^*|r_i,r_j,e) > P(h|r_i,r_j,e)$, for all $h \neq h^*, h \in \Omega(H)$.
Furthermore, it also holds that $P(h^*,r_i,r_j|e) > P(h,r_i,r_j|e)$, for all $h \neq h^*$, since $P(H,r_i,r_j|e) \propto P(H|r_i,r_j,e)$. Now by marginalizing over R_i, we have $P(h^*,r_j|e) > P(h,r_j|e)$, for all $h \neq h^*$. Finally, $h^* = arg\,max_H P(H,r_j|e)$, which proves the result. □

3.4. MAP-independence with continuous variables

First, we are going to use the conditional density $f(H|r,e)$ instead of the joint $f(H,r|e)$ for checking MAP-independence, in order to work with a distribution in the same feature space as $f(H|e)$. Then we will compare both density functions $f(H|e)$ and $f(H|r,e)$ using a distance measure $dist()$. We approximate MAP-independence checking if such distance is lower than an user-specified threshold ε. Some examples of distance measures are the Manhattan or Euclidean distance of the modes or the Jensen-Shannon divergence.

Furthermore, to get rid of the infinite value assignments of R, we can take a sample $\mathscr{S}$ from $f(R|e)$ and consider it an approximation of $\Omega(R)$. Since it is unrealistic to assume that for every single sample the MAP-independence condition will hold, we will compute the average distance of all pairs of densities when each s of the sample $\mathscr{S}$ is varied and verify if it is below ε:

$$\varepsilon MInd(h^*,R,e,\varepsilon) = \frac{\sum_{s\in\mathscr{S}} dist(f(H|e), f(H|R=s,e))}{|\mathscr{S}|} \leq \varepsilon, \text{ with } \mathscr{S} \sim f(R|e).$$

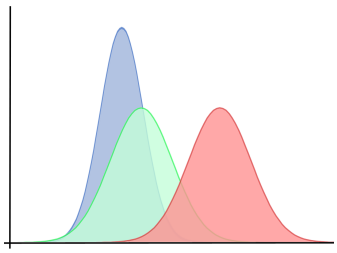
Analogously, we can define ε-MAP-independence strength as the ratio of samples that follow the condition of distance proximity given a threshold ε:

$$\varepsilon MIndStrength(h^*,R,e,\varepsilon) = \frac{|\{s \in \mathscr{S} | dist(f(H|e), f(H|R=s,e)) \leq \varepsilon\}|}{|\mathscr{S}|}$$

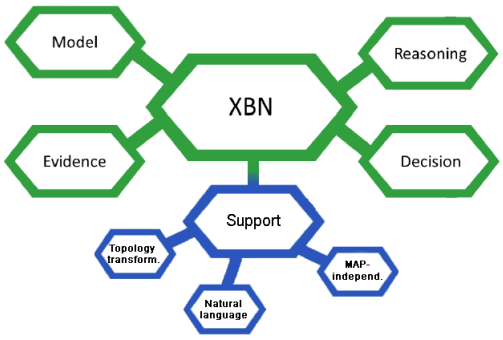
3.5. MAP-independence in Gaussian Bayesian networks

In the case of Gaussian Bayesian networks (GBNs), sampling can be avoided by exploiting the linearity in the relations. Consider two nodes of a GBN, X and Y, whose marginal densities are $f(X) \sim \mathscr{N}(\mu_X, \sigma_X)$ and $f(Y) \sim \mathscr{N}(\mu_Y, \sigma_Y)$.

If we make a random observation x over X, the variance of the posterior marginal density of Y, $f(Y|x)$ would be reduced, regardless of the actual value of x (we will refer to the a posteriori variance as $\sigma_{Y|X}$ instead of $\sigma_{Y|x}$, as it is not affected by the concrete value x). Formally, $\sigma_{Y|X} \leq \sigma_Y$. The mode ($\mu_{Y|x}$) of the posterior density $f(Y|x)$ will be conditioned by the value of the observation x. If $x = arg\,max_X f(X)$ (i.e., the mode of $f(X)$), then $\mu_{Y|x} = \mu_Y$. With a different arbitrary observation x', the mode will be different, although the variance will be the same. As such, we can consider $f(Y|x')$ a "displaced" version of $f(Y|x)$ such that $dist(f(Y), f(Y|x)) < dist(f(Y), f(Y|x'))$. This holds in the aforementioned distance examples. As the relations are linear, if we have a third observation x'' that is even further from the mean than x', then we know that $dist(f(Y), f(Y|x'')) > dist(f(Y), f(Y|x'))$ (illustrated in Figure 1a).



(a) In blue, $f(Y)$; in green, $f(Y|x)$ and in red, $f(Y|x')$, where x is closer to the mode of $f(X)$ than x'



(b) Taxonomy of explanations in Bayesian networks [7]. In blue, our contributions

Figure 1. Miscellaneous illustrations: To the left, posterior distributions in GBNs and, to the right, the expanded taxonomy of Bayesian network explanations

Theorem 3.3. *Given a MAP query and a Gaussian node R, if for an observation r at a normalised distance d from the mean of $f(R|e)$, $d = \frac{|\mu_{R|e} - r|}{\sigma_{R|E}}$, the ε-MAP-independence condition is met, then for any observation r' that is in range $[\mu_{R|e} - d\sigma_{R|E}, \mu_{R|e} + d\sigma_{R|E}]$ the condition will be met as well. If r does not meet the condition, some values in $[\mu_{R|e} - d\sigma_{R|E}, \mu_{R|e} + d\sigma_{R|E}]$ will not follow it either.*

Proof. Let $r' \in [\mu_{R|e} - d\sigma_{R|E}, \mu_{R|e} + d\sigma_{R|E}]$.
We know then that $dist(f(H|e), f(H|r',e)) \leq dist(f(H|e), f(H|r,e))$.
If with r the ε-MAP-independence condition holds, then $dist(f(H|e), f(H|r,e)) \leq \varepsilon$.
Therefore, by transitivity $dist(f(H|e), f(H|r',e)) \leq \varepsilon$ (i.e., the ε-MAP-independence holds also with r'). □

Therefore, if we select an appropriate d, we can verify if the range of values with probability $P(\mu_{R|e} - d\sigma_{R|E} < R < \mu_{R|e} + d\sigma_{R|E}|e)$ satisfy the ε-MAP-independence condition, or not. The higher the distance d, the wider the range (and probability) of R, but the stricter the condition. In case of having a set R, we check the point b of Mahalanobis distance d from the mean $\mu_{R|e}$ that maximises $dist(f(H|e), f(H, R = b|e))$. If said point follows the condition, a certain probability range of R will follow it as well:

$$\varepsilon MInd(h^*, R, e, \varepsilon) = dist(P(H|e), P(H|R = b, e)) \leq \varepsilon$$

3.6. MAP-independence in the Bayesian network explanations taxonomy

The main potential of MAP-independence is to study the quality of an already existing explanation h^* with node relevance, checking if additional observations may alter it. We consider that this action can be related to verifying stability, as we study if a reduction in uncertainty of the evidence changes the explanation. Stability is a desirable property that states that the explanation should not vary when two similar instances are explained [2].

Indeed, MAP-independence does not generate an explanation, but rather complements an existing one. Therefore, we claim that MAP-independence and existing techniques in the literature with similar purposes (like natural language generation or simpli-

fying the topology) belong to a newly introduced category that we named *support methods* (see Figure 1b). This is more of an umbrella term that refers to methodologies aimed to improve and measure the quality of explanations in Bayesian networks. In the specific case of MAP-independence, the goal is to check if a change in the uncertainty of the model will affect the explanation.

4. Conclusions

In this work we have theoretically presented some advances in MAP-independence that will be helpful to improve Bayesian network explainability. In addition, we relate the concepts of node relevance in MAP queries and stability of explanations and classify MAP-independence into a new category that compiles methods that aim towards improving the quality of explanations in Bayesian networks.

In the future, we will study our approximations to MAP-independence in practice, what values for threshold ε and distance d yield good results and how the distance function $dist()$ impact them. In addition, we see potential for checking stability and implement it into the existing explainability methods for Bayesian networks. One of the challenges will be the time complexity of this new refined concept of MAP-independence and heuristics may be used to solve the problem.

Acknowledgments

This research has been partially funded by the Spanish Ministry of Science and Innovation through the PID 2019-109247GB-I00 project and to the BBVA Foundation (2019 Call) through the "Score-based nonstationary temporal Bayesian networks. Applications in climate and neuroscience" project.

References

[1] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15.
[2] Molnar C. Interpretable Machine Learning. Lulu.com; 2020.
[3] Kwisthout J. Explainable AI using MAP-independence. In: European Conference on Symbolic and Quantitative Approaches with Uncertainty. Springer; 2021. p. 243-54.
[4] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann; 1988.
[5] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. The MIT press; 2009.
[6] Lacave C, Díez FJ. A review of explanation methods for Bayesian networks. The Knowledge Engineering Review. 2002;17(2):107-27.
[7] Derks IP, De Waal A. A taxonomy of explainable Bayesian networks. In: Southern African Conference for Artificial Intelligence Research. Springer; 2021. p. 220-35.
[8] Pearl J, Paz A. GRAPHOIDS: A graph-based logic for reasoning about relevance relations. UCLA Computer Science Department; 1985. R-53-L.
[9] Choi A, Xue Y, Darwiche A. Same-decision probability: A confidence measure for threshold-based decisions. International Journal of Approximate Reasoning. 2012;53(9):1415-28.