Constraint-Based Bayesian Network Structure Learning using Uncertain Experts' Knowledge

Christophe Gonzales[1] and **Axel Journe**[2] and **Ahmed Mabrouk**[2]
[1] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[2] CSAI Engie Lab, France
christophe.gonzales@lis-lab.fr, axel.journe@engie.com, ahmed.mabrouk@engie.com

Abstract

Exploiting experts' knowledge can significantly increase the quality of the Bayesian network (BN) structures produced by learning algorithms. However, in practice, experts may not be 100% confident about the opinions they provide. Worst, the latter can also be conflicting. Including such specific knowledge in learning algorithms is therefore complex. In the literature, there exist a few score-based algorithms that can exploit both data and the knowledge about the existence/absence of arcs in the BN. But, as far as we know, no constraint-based learning algorithm is capable of exploiting such knowledge. In this paper, we fill this gap by introducing the mathematical foundations for new independence tests including this kind of information. We provide a new constraint-based algorithm relying on these tests as well as experiments that highlight the robustness of our method and its benefits compared to other constraint-based learning algorithms.

1 Introduction

Bayesian networks (BN) are popular graphical models designed for compactly encoding joint probabilities. Their graphical structures represent sets of conditional independences. They can be either constructed by expert knowledge when the number of variables is limited or, more often than not, they are learnt from data. Although structure learning is known to be NP-hard (Chickering 1996), there exist numerous exact and approximate learning algorithms in the literature. Basically, they can be divided into four types of approaches: i) score-based (Heckerman, Geiger, and Chickering 1995); ii) constraint-based (Spirtes and Glymour 1991); iii) variable ordering-based (Teyssier and Koller 2005); iv) hybrid algorithms that combine the above methods to take advantage of their best (Tsamardinos, Brown, and Aliferis 2006). Each approach has its own pros and cons. In this paper, we focus on constraint-based learning algorithms. Actually, this type of learning exhibits appealing properties in terms of the accuracy of the structures learnt (Scutari, Graafland, and Gutiérrez 2019), and also in terms of causal discovery and overfitting avoidance.

Although BNs are seldom constructed entirely from expert knowledge, exploiting the latter in learning algorithms can still prove useful in order to increase the accuracy of the output structures. In a sense, score-based approaches natively include expert knowledge through their priors. But their expressive power is not very high in practice. One step toward more sophisticated knowledge, expressed in the form of ancestral sets constraints, was introduced in (Borboudakis and Tsamardinos 2012; Chen et al. 2016). Sets of possibly conflicting uncertain knowledge provided by several experts about the existence or absence of arcs in the structure were addressed in (Richardson and Domingos 2003; Amirkhani et al. 2017). This consisted essentially in modifying the scores classically used for learning.

In this paper, we are interested in exploiting possibly conflicting uncertain knowledge within a constraint-based approach. As far as we know, only knowledge about the existence or absence of arcs expressed as hard structural constraints have been introduced in the constraint-based framework (de Campos and Castellano 2007). This rules out both conflicting and uncertain experts' knowledge, which are often available in practical situations. We therefore propose a novel algorithm filling this gap. More precisely, constraint-based approaches consist of exploiting statistical independence tests in order to determine the set of conditional independences underlying the joint distribution of the BN's random variables. The main issue is therefore to introduce such experts' knowledge into the statistical tests. In Section 2, we introduce the mathematical foundation for such tests. Relying on tailored BNs, it results in new G-tests that precisely take into account sets of possibly conflicting assertions about the conditional dependence or independence between pairs of random variables. The confidence the experts have in these assertions as well as the trust we have in the experts are also taken into account. In Section 3, a variant of PC-stable (Colombo and Maathuis 2014) is proposed that includes these new tests. Finally, in Section 4, some experiments are provided to highlight the behavior of this new algorithm, and Section 5 concludes the paper.

2 Novel Independence Tests

The goal of this section is to introduce Rule 1, a new independence test taking into account multiple experts' uncertain knowledge. It relies on the exploitation of BNs:

Definition 1 (Bayesian network). *A BN is a pair* $(G, \mathbf{\Theta})$ *where* $G = (\mathbf{V}, \mathbf{E})$ *is a directed acyclic graph (DAG),* $\mathbf{V}$ *represents a set of random variables*[1]*,* $\mathbf{E}$ *is a set of arcs, and* $\mathbf{\Theta} = \{P(X|\mathbf{Pa}(X))\}_{X\in\mathbf{V}}$ *is the set of the conditional probability distributions (CPD) of the nodes / random variables* X *in* G *given their parents* $\mathbf{Pa}(X)$ *in* G*. The BN encodes the joint probability over* $\mathbf{V}$ *as* $P(\mathbf{V}) = \prod_{X\in\mathbf{V}} P(X|\mathbf{Pa}(X))$.

For our new independence test, consider two random variables $\mathbb{X}$ and $\mathbb{Y}$ whose domains are $\Omega_{\mathbb{X}} = \{x_1, \ldots, x_r\}$ and $\Omega_{\mathbb{Y}} = \{y_1, \ldots, y_s\}$ respectively and whose joint distribution is $P(\mathbb{X}, \mathbb{Y}) = \{\theta_{x_i y_j} : x_i \in \Omega_{\mathbb{X}}, y_j \in \Omega_{\mathbb{Y}}\}$. For all $x_i \in \Omega_{\mathbb{X}}$, let $\theta_{x_i \cdot} = \sum_{y_j\in\Omega_{\mathbb{Y}}} \theta_{x_i y_j}$ and, for all $y_j \in \Omega_{\mathbb{Y}}$, let $\theta_{\cdot y_j} = \sum_{x_i\in\Omega_{\mathbb{X}}} \theta_{x_i y_j}$. Let $\mathbf{S} = \{(XY)^{(n)}\}_{n=1}^N$ be a set of mutually independent variables distributed w.r.t. $P(\mathbb{X}, \mathbb{Y})$. Hence, $P((XY)^{(n)} = x_i y_j) = \theta_{x_i y_j}$ for all $x_i y_j \in \Omega_{\mathbb{X}\mathbb{Y}} = \Omega_{\mathbb{X}} \times \Omega_{\mathbb{Y}}$. A set $\mathbf{s} = \{(xy)^{(n)}\}_{n=1}^N$ of observations of the $(XY)^{(n)}$ is therefore an i.i.d. sample of observations of $(\mathbb{X}, \mathbb{Y})$. Finally, let $\mathbb{I}$ be the unobserved random variable indicating whether $\mathbb{X}$ and $\mathbb{Y}$ are independent ($\mathbb{I} = 1$) or not ($\mathbb{I} = 0$). We model the relationships between the $(XY)^{(n)}$ and $\mathbb{I}$ using the BN of Fig. 1, that is, the $(XY)^{(n)}$ are mutually independent but become dependent whenever $\mathbb{I}$ is observed. Indeed, if $\mathbb{I} = 1$ and we observe a subset $\mathbf{S}' \subset \mathbf{S}$ incompatible with $\mathbb{I} = 1$, $\mathbf{S}\backslash\mathbf{S}'$ will most likely account for the independence between $\mathbb{X}$ and $\mathbb{Y}$.

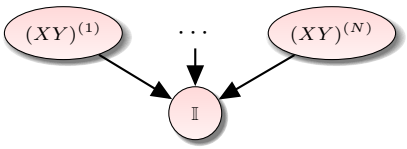

Figure 1: Relationship between the $(XY)^{(n)}$ and Variable $\mathbb{I}$.

Proposition 1. *Let* $\mathbf{s} = \{(xy)^{(n)}\}_{n=1}^N$ *be a sample of observations of* $\{(XY)^{(n)}\}_{n=1}^N$*. Let* $p_{\mathbb{I}} = P(\mathbb{I} = 1)$*. Assume that the conditional joint distribution of* $\{(XY)^{(n)}\}_{n=1}^N$ *given* $\mathbb{I} = 1$ *is approximately equal to* $\prod_{n=1}^N P((XY)^{(n)}|\mathbb{I} = 1)$*. Then, we have that:*

$$P(\mathbb{I} = i|\mathbf{s}) \approx \begin{cases} 1 - P(\mathbb{I} = 1|\mathbf{s}) & \text{if } i = 0, \\ p_{\mathbb{I}} \times \prod\limits_{xy\in\Omega_{\mathbb{X}\mathbb{Y}}} \left[\dfrac{\theta_{x\cdot} \times \theta_{\cdot y}}{\theta_{xy}}\right]^{N_{xy}} & \text{if } i = 1, \end{cases} \quad (1)$$

where N_{xy} *refers to the number of elements in* $\mathbf{s}$ *equal to* xy*.*

Note that the above assumption precisely corresponds to that which is used in the classical χ^2 and G^2 independence tests. Fig. 1, the values of $p_{\mathbb{I}}$ and of $\{\theta_{xy}\}$ and the above proposition fully characterize the BN of the joint of $\mathbb{I}$ and $\{(XY)^{(n)}\}_{n=1}^N$. Parameters $\{\theta_{xy}\}$ can simply be estimated by Maximum Likelihood. It is easy to see that Equation (1) does not constrain the value of $p_{\mathbb{I}}$. This one should therefore reflect our background knowledge about the independence of $\mathbb{X}$ and $\mathbb{Y}$. We will provide some reasonable possible values in the experimental section.

[1]By abuse of notation, we use interchangeably $X \in \mathbf{V}$ to denote a node in the BN and its corresponding random variable.

Now, assume that K experts provide their knowledge about the state of $\mathbb{I}$, i.e., about whether $\mathbb{X}$ and $\mathbb{Y}$ are independent, through assertions e_k similar to "I think that $\mathbb{X}$ and $\mathbb{Y}$ are independent" or "I believe that $\mathbb{X}$ directly influences $\mathbb{Y}$". Let O_k be a random variable representing the opinion of the kth expert ($O_k = 0$ and $O_k = 1$ mean dependence and independence respectively). Hence statement e_k is an observation of Variable O_k and $P(e_k|O_k)$ is a vector of size $|O_k|$ with one value equal to 1 and the other equal to 0. In addition, assume that the experts are able to estimate their confidence, say γ_k, in their correct identification of the true state of $\mathbb{I}$. For instance, the kth expert may estimate that she is 70% confident that her judgment is right. In this case, $\gamma_k = 0.7$. Then $P(O_k = i|\mathbb{I} = i) = \gamma_k$ for $i \in \{0, 1\}$. The experts' opinions are based on previous experiences with $\mathbb{X}$ and $\mathbb{Y}$. Hence they are not independent. However, whenever the value of $\mathbb{I}$ is known, the knowledge of some experts should not provide new information to the others. So the relations between $\mathbb{I}$ and the O_k's can be encoded by a BN containing only arcs $\mathbb{I} \to O_k$. Finally, it is well-known that people have trouble to accurately estimate probabilities (Tversky and Kahneman 1992). So, we should not directly use γ_k in $P(O_k = i|\mathbb{I} = i)$ but rather a transform of γ_k. To avoid ambiguities, we introduce a new variable E_k using this transform, i.e., $P(E_k = i|\mathbb{I} = i) = \varphi_k(\gamma_k)$. Note that $\varphi_k(\cdot)$ can also include our own perception about the accuracy of the kth expert. In this paper, we suggest using the following parameterized logistic function mapping $[0, 1]$ to $[0, 1]$:

$$\varphi_k(\gamma_k) = \frac{1}{\beta_k - 1}\left[\frac{1 + \beta_k}{1 + \beta_k^{1-2\gamma_k}} - 1\right]. \quad (2)$$

This transform is quite general, as shown in Fig. 2 in which, for simplicity, $\rho_k = 2\log\beta_k$.

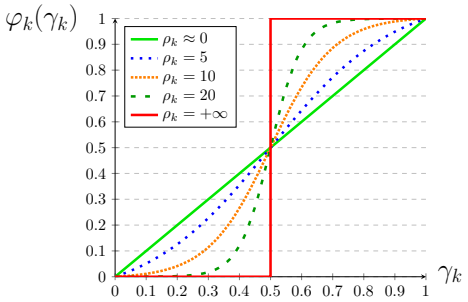

Figure 2: Probability transform $\varphi_k(\cdot)$.

Overall, to take into account the uncertain experts knowledge, the BN of Fig. 1 can be extended as that of Fig. 3. Evidence e_k can then be entered as $P(e_k|E_k) = P(e_k|O_k)$.

Using this Bayesian network, it is possible to compute the likelihood ratios used in G^2-type independence tests:

Proposition 2. *Let* $\mathbf{s} = \{(xy)^{(n)}\}_{n=1}^N$ *be a sample of observations of* $\{(XY)^{(n)}\}_{n=1}^N$ *and* $\mathbf{e} = \{e_k\}_{k=1}^K$ *be some expert knowledge. Then, the likelihoods* $\mathcal{L}(\mathbf{s}|\mathbb{I} = 1, \mathbf{e}) = P(\mathbf{s}|\mathbb{I} = 1, \mathbf{e}) = \mathcal{L}_{P_1}(\mathbf{s}|\mathbf{e})$ *and* $\mathcal{L}(\mathbf{s}|\mathbf{e}) = P(\mathbf{s}|\mathbf{e})$ *are such that:*

$$\mathcal{L}(\mathbf{s}|\mathbf{e}) = \frac{\epsilon_0}{P(\mathbf{e})} \times \mathcal{L}_{P_0}(\mathbf{s}|\mathbf{e}) + \left[1 - \frac{\epsilon_0}{P(\mathbf{e})}\right] \times \mathcal{L}_{P_1}(\mathbf{s}|\mathbf{e}), \quad (3)$$

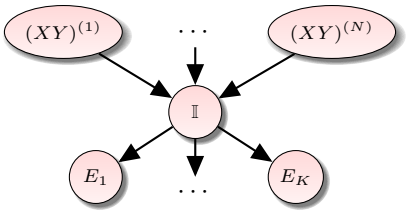

Figure 3: Relationships including experts' knowledge.

$$\textit{with } \mathcal{L}_{P_0}(\mathbf{s}|\mathbf{e}) = \prod_{xy\in\Omega_{\mathbb{XY}}} \theta_{xy}^{N_{xy}},\ \mathcal{L}_{P_1}(\mathbf{s}|\mathbf{e}) = \prod_{xy\in\Omega_{\mathbb{XY}}} (\theta_{x.}\theta_{.y})^{N_{xy}},$$

$$P(\mathbf{e}) = \epsilon_0 + P(\mathbb{I}=1)\times[\epsilon_1-\epsilon_0],$$

$$\epsilon_0 = P(\mathbf{e}|\mathbb{I}=0) = \sum_{E_1,\ldots,E_K}\prod_{k=1}^{K} P(E_k|\mathbb{I}=0)P(e_k|E_k),$$

$$\epsilon_1 = P(\mathbf{e}|\mathbb{I}=1) = \sum_{E_1,\ldots,E_K}\prod_{k=1}^{K} P(E_k|\mathbb{I}=1)P(e_k|E_k),$$

N_{xy} *= the number of elements of* $\mathbf{s}$ *equal to* xy*.*

Equation (3) can be interpreted as follows: there is an uncertainty concerning Probability distribution P of the population, which can be equal either to $P_0(\mathbb{X},\mathbb{Y}) = \{\theta_{xy} : x \in \Omega_{\mathbb{X}}, y \in \Omega_{\mathbb{Y}}\}$ or to $P_1(\mathbb{X},\mathbb{Y}) = \{\theta_{x.}\theta_{.y} : x \in \Omega_{\mathbb{X}}, y \in \Omega_{\mathbb{Y}}\}$. $\mathcal{L}$ can therefore be modeled as a mixture of the likelihoods of two populations, weighted w.r.t. the expert knowledge confidence and our perception of its accuracy.

When no expert knowledge e_k is available, we trivially have $\epsilon_0 = \epsilon_1 = P(\mathbf{e}) = 1$. Similarly, when we trust equally the expertise of the experts w.r.t. dependence and independence of $\mathbb{X}$ and $\mathbb{Y}$ and when the expert has no knowledge and just says that there is a 50-50% chance that there is an independence between $\mathbb{X}$ and $\mathbb{Y}$, then $\epsilon_0 = \epsilon_1$. Note that, whenever $\epsilon_0 = \epsilon_1$, $P(\mathbf{s},\mathbf{e}) = \mathcal{L}_{P_0}(\mathbf{s}|\mathbf{e})$, which is the likelihood used at the numerator of the classical G-test.

Let $\mathcal{LR}(s|e)$ denote Likelihood ratio $\mathcal{L}(s|e)/\mathcal{L}(s|\mathbb{I}=1,e)$. Then, by the above proposition, it holds that:

$$\mathcal{LR}(\mathbf{s}|\mathbf{e}) = \frac{\epsilon_0}{P(\mathbf{e})}\times\mathcal{LR}_{P_0}(\mathbf{s}|\mathbf{e}) + \left[1-\frac{\epsilon_0}{P(\mathbf{e})}\right]\times\mathcal{LR}_{P_1}(\mathbf{s}|\mathbf{e}),$$

with $\mathcal{LR}_{P_0}(\mathbf{s}|\mathbf{e}) = \mathcal{L}_{P_0}(\mathbf{s}|\mathbf{e})/\mathcal{L}_{P_1}(\mathbf{s}|\mathbf{e})$ and $\mathcal{LR}_{P_1}(\mathbf{s}|\mathbf{e}) = 1$. This likelihood ratio is therefore the mixture of two likelihoods, one over the population related to P_0 and the other one over the population related to P_1. Now, in the classical G-test, under the assumption that the likelihood under the alternative hypothesis is close to that under the null hypothesis, their ratio should be close to 1 and, therefore, at significance level α, the G-test consists of computing a threshold above which only a percentage α of the set of all possible samples would produce a ratio greater than this threshold (given that the aforementioned assumption holds). Here, the α-percentage over the whole population is logically split into an $\alpha\times\eta$ percentage over the population related to P_0 plus an $\alpha\times(1-\eta)$ percentage over the population related to P_1, with $\eta = \epsilon_0/P(\mathbf{e})$. As likelihood ratio $\mathcal{LR}_{P_1}(\mathbf{s}|\mathbf{e})$ is always equal to 1, the $\alpha\times(1-\eta)$ percentage is uninformative for the G-test. So, the test relies only on $\mathcal{LR}_{P_0}(\mathbf{s}|\mathbf{e})$.

Proposition 3. *Assuming that the likelihood under the alternative hypothesis is close to that under the null hypothesis,* $2\ln(\mathcal{LR}_{P_0}(\mathbf{S}|\mathbf{e})) \sim \chi^2_k$*, where* $k = (|\Omega_{\mathbb{X}}|-1)\times(|\Omega_{\mathbb{Y}}|-1)$*.*

Rule 1. *Our independence test at significance level* α *given sample* $\mathbf{s}$ *and expert knowledge* $\mathbf{e}$ *consists of accepting independence between* $\mathbb{X}$ *and* $\mathbb{Y}$ *if and only if* $2\ln(\mathcal{LR}_{P_0}(\mathbf{s}|\mathbf{e})) < c_{k,\alpha\eta}$ *or, equivalently, if the p-value of* $2\ln(\mathcal{LR}_{P_0}(\mathbf{s}|\mathbf{e}))$ *is greater than* $\alpha\times\eta$*, where* $\eta = \epsilon_0/P(\mathbf{e})$ *and, for any* δ*,* $c_{k,\delta}$ *denotes the* $(1-\delta)$ *quantile of the* χ^2 *distribution with* k *degrees of freedom.*

In other words, our test consists of applying the classical G-test with significance level $\alpha\times\eta$ instead of α. It can be easily extended to cope with conditional independence testing given a set $\mathbb{Z}$ of random variables. For this purpose, it is sufficient to extend the network of Fig. 3 to that of Fig. 4. In this BN, $Z^{(1)},\ldots,Z^{(N)}$ are random variables distributed w.r.t. $P(\mathbb{Z}) = \{\theta_z : z \in \Omega_{\mathbb{Z}}\}$ and $\mathbb{I}$ is a random variable indicating whether $\mathbb{X}$ and $\mathbb{Y}$ are conditionally independent ($\mathbb{I}=1$) or not ($\mathbb{I}=0$) given $\mathbb{Z}$.

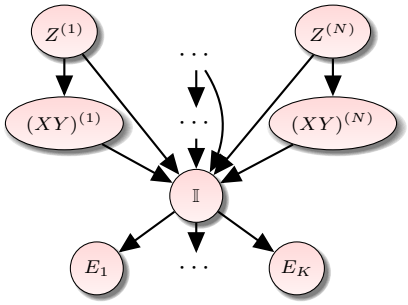

Figure 4: The BN for conditional independence testing.

Proposition 4. *Let* $\mathbf{s} = \{((xy)^{(n)}, z^{(n)})\}_{n=1}^N$ *be a sample of observations of* $\mathbf{S} = \{((XY)^{(n)}, Z^{(n)})\}_{n=1}^N$*. For any* $z \in \Omega_{\mathbb{Z}}$*, assume that* $P(Z^{(n)}|\mathbb{I}) \approx P(Z^{(n)})$ *and that:*

$$P(\mathbb{I}=1|\mathbf{s}) \approx p(\mathbb{I}=1)\times\prod_{xyz\in\Omega_{\mathbb{XYZ}}}\left[\frac{\theta_{x.|z}\theta_{.y|z}}{\theta_{xy|z}}\right]^{N_{xyz}}, \quad (4)$$

where $\theta_{x.|z} = \sum_{y\in\Omega_{\mathbb{Y}}}\theta_{xy|z}$*,* $\theta_{.y|z} = \sum_{x\in\Omega_{\mathbb{X}}}\theta_{xy|z}$ *and* $\theta_{xy|z} = P(\mathbb{X}=x,\mathbb{Y}=y|\mathbb{Z}=z)$*. Then, denoting by* $\mathcal{L}_{|\mathbf{z}}(\cdot)$ *the conditional likelihoods given* $\{z^{(n)}\}_{n=1}^N$*, i.e.,* $\mathcal{L}_{|\mathbf{z}}(\mathbf{s}|\mathbf{e}) = P(\{(xy)^{(n)}\}_{n=1}^N|\mathbf{e},\{z^{(n)}\}_{n=1}^N)$*, it holds that:*

$$\begin{aligned}\mathcal{LR}_{|\mathbf{z}}(\mathbf{s}|\mathbf{e}) &= \frac{\mathcal{L}_{|\mathbf{z}}(\mathbf{s}|\mathbf{e})}{\mathcal{L}_{|\mathbf{z}}(\mathbf{s}|\mathbb{I}=1,\mathbf{e})}\\ &= \frac{\epsilon_0}{P(\mathbf{e})}\times\mathcal{LR}_{P_0|\mathbf{z}}(\mathbf{s}|\mathbf{e}) + \left[1-\frac{\epsilon_0}{P(\mathbf{e})}\right],\end{aligned}$$

where $P(\mathbf{e}) = \epsilon_0 + P(\mathbb{I}=1)\times[\epsilon_1-\epsilon_0]$,

$$\begin{aligned}\mathcal{LR}_{P_0|\mathbf{z}}(\mathbf{s}|\mathbf{e}) &= \mathcal{L}_{P_0|\mathbf{z}}(\mathbf{s}|\mathbf{e})/\mathcal{LR}_{P_1|\mathbf{z}}(\mathbf{s}|\mathbf{e}), \quad (5)\\ \mathcal{L}_{P_0|\mathbf{z}}(\mathbf{s}|\mathbf{e}) &= \prod_{xyz\in\Omega_{\mathbb{XYZ}}}\theta_{xy|z}^{N_{xyz}},\\ \mathcal{L}_{P_1|\mathbf{z}}(\mathbf{s}|\mathbf{e}) &= \prod_{xyz\in\Omega_{\mathbb{XYZ}}}\left(\theta_{x.|z}\theta_{.y|z}\right)^{N_{xyz}},\end{aligned}$$

N_{xyz} *is the number of records in* $\mathbf{s}$ *equal to* xyz*.*

Assuming that the likelihoods under the alternative hypothesis and the null hypothesis are close, $2\ln(\mathcal{LR}_{P_0|\mathbf{z}}(\mathbf{S}|\mathbf{e})) \sim \chi^2_k$*, where* $k = (|\Omega_{\mathbb{X}}| - 1) \times (|\Omega_{\mathbb{Y}}| - 1) \times \Omega_{\mathbb{Z}}$.

Our conditional independence test is thus the same as that of Rule 1, except that conditional probabilities are used instead of unconditional ones. Note that all the θ's can be estimated by Maximum Likelihood, as in classical G-tests.

3 Constraint-based Learning with Multiple Uncertain Experts' Knowledge

In theory, we can exploit directly our new independence tests within the well-known PC, IC or PC-stable algorithms. However, using uncertain experts' knowledge raises two issues: i) what if the dataset is too small to perform a given G-test? and ii) what if the G-tests with and without experts' knowledge result in different decisions?

As for the first issue, usually, the learning algorithm simply does not take into account the G-test. But in our situation, expert's knowledge might still be exploitable. For this purpose, let d denote the decision to consider that $\mathbb{X}$ and $\mathbb{Y}$ are conditionally independent given $\mathbb{Z}$ and let u_1 and u_0 denote the utilities of making decision d while being right and wrong respectively, i.e., they represent our perception about the gain and loss of making decision d (we therefore assume that $u_1 > 0$ and $u_0 < 0$). Utilities u_1 and u_0 may not be easy to elicit but their ratio $T = -u_0/u_1$ is (Keeney and Raiffa 1976). Finally, note that ϵ_1 and ϵ_0 represent the probabilities, according to the experts, of being right and wrong. Therefore, from an expected utility point of view, we should consider $\mathbb{X}$ and $\mathbb{Y}$ conditionally independent given $\mathbb{Z}$ if and only if $\epsilon_1 u_1 + \epsilon_0 u_0 > 0$ or, equivalently, if $\epsilon_1/\epsilon_0 > T$. We use precisely this rule in Algorithm 1.

For the second issue, note that removing or keeping an edge at one step of the constraint-based learning can have a significant impact on the subsequent steps. So such decisions must be taken with care. When both G-tests with and without experts' knowledge agree, we can be confident in the resulting decision. However, when they disagree, we should follow the test with expert knowledge (since it takes into account more information) but we should also not definitely rule out the information provided by the G-test without experts' knowledge. To do so, in Algorithm 1, we construct a graph G which will be the learnt skeleton, but we also maintain another graphical structure named G_{adj}, which is used to determine the candidate Sepsets (separating sets). Graph G_{adj} takes into account the disagreements between the G-tests with and without experts' knowledge.

These variants combined with PC-stable (Colombo and Maathuis 2014) result in Algorithm 1, called PCSe.

Algorithm 1 Algorithm PCSe

Input: dataset $\mathbf{D}$, experts' knowledge $\mathbf{K}$, significance level α, *a priori* knowledge $P(\mathbb{I})$, threshold T
Output: an undirected graph G and a set of $Sepsets$

$\mathbf{V} \leftarrow$ all the variables/columns of $\mathbf{D}$
$G = (\mathbf{V}, \mathbf{E}) \leftarrow$ complete undirected graph
$G_{adj} = (\mathbf{V}, \mathbf{E}_{adj}) \leftarrow$ complete undirected graph
$m \leftarrow -1$; $Sepset \leftarrow \emptyset$
repeat
  $m \leftarrow m + 1$
  for all vertices X_i in $\mathbf{V}$ **do**
    $\mathbf{adj}_i \leftarrow \{X_j \in \mathbf{V} : (X_i, X_j) \in \mathbf{E}_{adj}\}$
  end for
  for all $(X_i, X_j) \in \mathbf{E}_{adj}$ s.t. $|\mathbf{adj}_i| \geq m + 1$ **do**
    repeat
      Choose a new set $\mathbf{Z} \subseteq \mathbf{adj}_i \backslash \{X_j\}$ s.t. $|\mathbf{Z}| = m$
      Extract from $\mathbf{K}$ the knowledge on $(X_i, X_j|\mathbf{Z})$
      Create the BN of Fig. 4 and compute its θ-parameters, ϵ_0, ϵ_1, $P(\mathbf{e})$ and $\eta = \epsilon_0/P(\mathbf{e})$
      if $\mathbf{D}$ is too small for performing G-test **then**
        if $\epsilon_1/\epsilon_0 > T$ **then**
          Remove (X_i, X_j) from both $\mathbf{E}$ and $\mathbf{E}_{adj}$
          $Sepset(\{X_i, X_j\}) \leftarrow \mathbf{Z}$
        end if
      else
        $p \leftarrow$ p-value of $2\ln(\mathcal{LR}_{P_0|\mathbf{z}}(\mathbf{s}|\mathbf{e}))$ of Eq. (5)
        if $p \geq \alpha$ **then**
          Remove (X_i, X_j) from $\mathbf{E}_{adj}$
        end if
        if $p \geq \alpha \times \eta$ **then**
          Remove (X_i, X_j) from $\mathbf{E}$
          $Sepset(\{X_i, X_j\}) \leftarrow \mathbf{Z}$
        end if
      end if
    until either $p \geq \alpha$ and $p \geq \alpha \times \eta$ or all $\mathbf{Z}$ s.t. $|\mathbf{Z}| = m$ have been considered
  end for
until all vertices $X_i \in \mathbf{V}$ are such that $|\mathbf{adj}_i| \leq m$
return G, $Sepset$

4 Experimentations

We now highlight the effectiveness of PCSe by comparing it with PC stable and the algorithm proposed in (de Campos and Castellano 2007), hereafter called PCS and PCSDC respectively. For this purpose, we randomly generate datasets from benchmark networks taken from a BN repository[2]. Samples of sizes ranging from 500 to 10000 records are thus generated. To provide experts' knowledge, we first select randomly 50%[3] of the structural information encoded in the original network. Then, to highlight the robustness of PCSe w.r.t. errors, part of this information is included as is into the experts' knowledge whereas the other part is transformed into erroneous information, i.e., the presence (resp. absence) of an edge is included as an absence (resp. presence) of this edge. In order to enable comparisons with PCSDC, only one expert is taken into account. Finally, for all the experiments, Threshold T is fixed at 0.5, and we set $P(\mathbb{I} = 1) = 0.95$ as this translates well the fact that, in practice, BNs are "sparse" and, as such, represent a majority of independences.

PCSe, PCS and PCSDC are compared on the basis of the

[2]https://www.bnlearn.com/bnrepository/

[3]We assume that experts' knowledge cannot exceed 50% of the structural information encoded in the BN.

Sample size	1000			5000			10000		
Correctness (%)	PCS	PCSDC	PCSe	PCS	PCSDC	PCSe	PCS	PCSDC	PCSe
0	**0.703**	0.000	0.695	**0.836**	0.000	0.822	**0.889**	0.000	0.873
25	0.703	0.174	**0.719**	0.836	0.174	**0.837**	**0.889**	0.174	0.881
50	0.703	0.347	**0.742**	0.836	0.347	**0.848**	0.889	0.347	**0.891**
75	0.703	0.521	**0.762**	0.836	0.521	**0.862**	0.889	0.521	**0.897**
100	0.703	**0.863**	0.779	0.836	**0.913**	0.859	0.889	**0.939**	0.909

Table 1: F-scores of PCS, PCSDC and PCSe for the Alarm BN in function of the correctness percentage.

	PCS	PCSDC	PCSe	PCS	PCSDC	PCSe	PCS	PCSDC	PCSe
Correctness (%)	Asia			Andes			Alarm		
0	**0.662**	0.000	0.591	**0.639**	0.000	0.599	**0.623**	0.000	0.615
25	**0.662**	0.166	0.624	0.639	0.164	**0.635**	0.623	0.166	**0.634**
50	0.662	0.399	**0.671**	0.639	0.326	**0.667**	0.623	0.333	**0.663**
75	0.662	0.674	**0.729**	0.639	0.487	**0.706**	0.623	0.503	**0.694**
100	0.662	**0.844**	0.778	0.639	**0.838**	0.752	0.623	**0.831**	0.736
Correctness (%)	Child			Hailfinder			Insurance		
0	**0.833**	0.000	0.768	**0.522**	0.000	0.495	**0.608**	0.000	0.602
25	**0.833**	0.199	0.785	**0.522**	0.117	0.499	**0.608**	0.177	0.634
50	**0.833**	0.350	0.802	0.522	0.233	**0.523**	0.608	0.329	**0.640**
75	0.833	0.503	**0.854**	0.522	0.358	**0.534**	0.608	0.506	**0.646**
100	0.833	**0.914**	0.877	0.522	**0.725**	0.540	0.608	**0.815**	0.664

Table 2: F-scores of PCS, PCSDC and PCSe for benchmark datasets with 500 records in function of the correctness percentage.

skeletons they produce, i.e., the BNs' graphs in which arcs are substituted by (undirected) edges. To do so, these skeletons are compared against those of the "true" BNs used to generate the datasets on the basis of the F-score criterion. The latter is defined as $(2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$, where the Recall and Precision metrics are defined as ratios $\text{TP}/(\text{TP}+\text{FN})$ and $\text{TP}/(\text{TP}+\text{FP})$ respectively, and TP, FP and FN refer to true positives, false positives and false negatives respectively. As such, the F-score metric provides a general overview on the reliability of each studied algorithm. To make the comparison study more thorough, in the experiments, we vary the following parameters:

- the size of the dataset,
- the percentage of correct expert's assertions,
- the expert's confidence γ_k of giving a correct assertion,
- the parameter ρ_k of the expert's confidence transform (see Fig. 2 and Eq. (2)).

Finally, for each set of the above parameters, 100 tests are performed and Tables 1 to 4 display the average F-scores over these 100 tests.

All experiments are performed on a 1.9GHz Intel Core i7 computer with 16GB of memory running Windows 10. The studied learning algorithms have been implemented using the pyAgrum library (Gonzales, Torti, and Wuillemin 2017).

Table 1 displays the average F-scores for the Alarm BN, with $\gamma_k = 0.8$ and $\rho_k = 14$, in function of the dataset size and the percentage of correct expert's assertions. In the case of PCSDC, although $\gamma_k = 0.8$, all the expert's opinions are taken into account since PCSDC cannot take into account the confidence of the expert. As can be observed, except for the case where the correctness is equal to 100%, our algorithm always outperforms the PCSDC method, whatever the size of the dataset: with a dataset size equal to 1000, the F-score of our algorithm is actually greater than that of PCSDC by 24% to 69%. And the larger the dataset size, the larger the difference between PCSe and PCSDC. This results from the fact that our new independence tests are able to aggregate the valuable information provided by the experts with those contained in the training dataset. On the contrary, by constraining the learning algorithm to comply with the expert's opinions without taking into account the dataset, PCSDC is doomed to be ineffective when the expert makes too many mistakes. The differences w.r.t. PCS are slightly lower compared with those w.r.t. PCSDC. Still, we always outperform PCS when the correctness percentage lies between 50% and 100%. Between 25% and 50%, even if the expert makes many mistakes, PCSe often outperforms PCS, which makes PCSe quite robust to experts' errors. But when this percentage is below 25%, the expert introduces too many mistakes, which has too negative an impact on the F-score of PCSe. This makes the latter decrease slightly below that of PCS. The same results can be observed with other benchmark BNs, see Table 2, in which γ_k and ρ_k are still equal to 0.8 and 14 respectively.

The accuracy of our algorithm can also be demonstrated by varying Parameters γ_k and ρ_k related to the confidence of the expert. In Table 3, the correctness percentage is fixed

at 75% and ρ_k is fixed at 14. Parameter γ_k varies and the F-scores are computed on datasets of size 1000 generated from the Insurance BN. As can be seen, PCSe significantly outperforms all the other algorithms. Note that, for $\gamma_k = 0.5$, i.e., the expert has no clue about whether her assertion is right or not, we enforced PCSDC not to take into account the expert knowledge, hence the 0.581 F-score. For PCSe, $\gamma_k = 0.5$ implies that $\epsilon_0 = \epsilon_1$, hence that the expert knowledge must not be taken into account. Note also that, for $\gamma_k = 0.8$, the F-scores of both PCS and PCSe are lower than those for Insurance, with a correctness equal to 75%, given in Table 2, although the sizes of the datasets used in Table 3 are twice higher those used in Table 2. This is due to the G-tests being performed only when the datasets are large enough (see Lines 15–19 of Algorithm 1).

γ_k	PCS	PCSDC	PCSe
0.5	0.581	**0.581**	**0.581**
0.6	0.581	0.513	**0.592**
0.7	0.581	0.513	**0.606**
0.8	0.581	0.513	**0.620**
0.9	0.581	0.513	**0.639**

Table 3: F-scores of PCS, PCSDC and PCSe for the Insurance BN in function of γ_k.

Finally, in Table 4, the percentage of correct assertions is fixed at 100%, and the datasets are of size 1000 and are generated from the Alarm BN. The higher the value of ρ_k, the more the expert's opinion is taken into account. Hence if her assertions are correct, the accuracy of the produced skeleton tends also to increase. For $\gamma_k < 1$, when the confidence transform parameter ρ_k increases from 5 to 20, we observe an increase from about 2% to 7% of the F-score. All these results clearly highlight the reliability of our approach, which is suited to cope in an efficient way with many important parameters associated with the expert's knowledge.

γ_k	$\rho_k = 5$	$\rho_k = 10$	$\rho_k = 15$	$\rho_k = 20$
0.6	0.711	0.718	0.726	0.738
0.7	0.720	0.739	0.757	0.774
0.8	0.736	0.760	0.783	0.802
0.9	0.755	0.782	0.805	0.822
1.0	0.873	0.873	0.873	0.873

Table 4: F-scores of PCSe for the Alarm BN in function to γ_k and ρ_k.

5 Conclusion

In this paper, we proposed a new constraint-based algorithm for learning the structure of BNs in the presence of conflicting and uncertain experts' knowledge. This algorithm relies on a very effective novel independence test whose mathematical correctness has been proven. As shown in the experimentations, our method significantly outperforms other constraint-based learning approaches. For future works, we plan to develop a new hybrid BN learning approach capable of integrating in a coherent way precisely the same experts' knowledge in its score-based refinement phase.

References

Amirkhani, H.; Rahmati, M.; Lucas, P. J. F.; and Hommersom, A. 2017. Exploiting experts' knowledge for structure learning of Bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(11):2154–2170.

Borboudakis, G., and Tsamardinos, I. 2012. Incorporating causal prior knowledge as path-constraints in Bayesian networks and maximal ancestral graphs. In *Proc. of the International Conference on Machine Learning*, 427–434.

Chen, E. Y.-J.; Shen, Y.; Choi, A.; and Darwiche, A. 2016. Learning Bayesian networks with ancestral constraints. In *Proc. of the International Conference on Neural Information Processing Systems*, 2333–2341.

Chickering, D. 1996. Learning Bayesian networks is NP-complete. In Fisher D., L. H., ed., *Learning from Data*, volume 112 of *Lecture Notes in Statistics*. Springer. 121–130.

Colombo, D., and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research* 15:3921–3962.

de Campos, L. M., and Castellano, J. G. 2007. Bayesian network learning algorithm using structural restrictions. *International Journal of Approximate Reasoning* 45:233–254.

Gonzales, C.; Torti, L.; and Wuillemin, P.-H. 2017. aGrUM: A graphical universal model framework. In *Proc. of the International Conference on Industrial Engineering, Other Applications of Applied Intelligent Systems*, 171–177.

Heckerman, D.; Geiger, D.; and Chickering, D. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20:197–243.

Keeney, R. L., and Raiffa, H. 1976. *Decisions with Multiple Objectives - Preferences and Value Tradeoffs*. Cambridge university Press.

Richardson, M., and Domingos, P. 2003. Learning with knowledge from multiple experts. *Proc. of the International Conference on Machine Learning* 2:624–631.

Scutari, M.; Graafland, C. E.; and Gutiérrez, J. 2019. Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning* 115:235–253.

Spirtes, P., and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9(1):62–72.

Teyssier, M., and Koller, D. 2005. Ordering-based search: A simple and effective algorithm for learning Bayesian networks. In *Proc. of the Conference on Uncertainty in Artificial Intelligence*, 584–590.

Tsamardinos, I.; Brown, L.; and Aliferis, C. 2006. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning* 65.

Tversky, A., and Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5(4):297–323.