Applied Intelligence (2024) 54:4668–4688
https://doi.org/10.1007/s10489-023-05153-8

Conditional probability table limit-based quantization for Bayesian networks: model quality, data fidelity and structure score

Rafael Rodrigues Mendes Ribeiro[1] · Jordão Natal[1] · Cassio Polpo de Campos[2] · Carlos Dias Maciel[1]

Accepted: 1 November 2023 / Published online: 3 April 2024

Abstract
Bayesian Networks (BN) are robust probabilistic graphical models mainly used with discrete random variables requiring discretization and quantization of continuous data. Quantization is known to affect model accuracy, speed and interpretability, and there are various quantization methods and performance comparisons proposed in literature. Therefore, this paper introduces a novel approach called CPT limit-based quantization (CLBQ) aimed to address the trade-off among model quality, data fidelity and structure score. CLBQ sets CPT size limitation based on how large the dataset is so as to optimize the balance between the structure score of BNs and mean squared error. For such a purpose, a range of quantization values for each variable was evaluated and a Pareto set was designed considering structure score and mean squared error (MSE). A quantization value was selected from the Pareto set in order to balance MSE and structure score, and the method's effectiveness was tested using different datasets, such as discrete variables with added noise, continuous variables and real continuous data. In all tests, CLBQ was compared to another quantization method known as Dynamic Discretization. Moreover, this study assesses the suitability of CLBQ for the search and score of BN structure learning, in addition to examining the landscape of BN structures while varying dataset sizes and confirming its consistency. It was sought to find the expected structure location through a landscape analysis and optimal BNs on it so as to confirm whether the expected results were actually achieved in the search and score of BN structure learning. Results demonstrate that CLBQ is quite capable of striking a balance between model quality, data fidelity and structure score, in addition to evidencing its potential application in the search and score of BN structure learning, thus further research should explore different structure scores and quantization methods through CLBQ. Furthermore, its code and used datasets have all been made available.

Keywords Quantization · Bayesian network · BIC · CPT

1 Introduction

Bayesian Networks (BN) stand out as robust probabilistic graphical models for modelling and reasoning [1, 7, 30], in addition to being extensively employed in diverse domains such as medical problems [30], water quality [1], risk assessment [33], and network traffic prediction [26]. A BN comprises a directed acyclic graph (DAG) and associated parameters able to capture probabilistic dependencies among random variables represented as nodes [9, 12, 15]. BNs exhibit versatility in handling both continuous and discrete random variables [10]. However, an application of continuous random variables demands assumptions regarding parametric statistical distributions, leading to its infrequent use [10]. Conversely, discrete random variables necessitate data discretization by employing Conditional Probability Tables (CPTs) to quantify relationships between variables [7, 12], enabling BNs to effectively model intricate and nonlinear relationships [10]. Such a discrete approach has been widely embraced by authors across numerous applications [10]. Additionally, prominent algorithms for BN model learning predominantly rely on discrete and quantized time series data [2], given that discretization transforms continuous signals into sampled representations while quantization diminishes precision in specific data points [21]. It is known that a quantization of variables substantially affects the accuracy, computational efficiency and interpretability of resulting BN models [4].

The significance of quantization has led to the development of numerous quantization techniques for Bayesian

✉ Rafael Rodrigues Mendes Ribeiro
rafael.mendes.ribeiro@usp.br

Extended author information available on the last page of the article

Networks (BNs). Friedman and Goldszmidt [8] introduced a quantization approach based on Minimal Description Length capable of balancing quantization levels using the learned (DAGs), thence accentuating its fidelity in modelling training data. Notably, its assessment is primarily centred on model prediction accuracy. Monti and Cooper [18] put forth a technique for multivariate discretization and quantization, and has thoughtfully considered inter-variable interactions. This approach introduces a Bayesian scoring metric aimed to evaluate variable quantization within the context of a BN structure. However, it is worth observing that an evaluation of its efficacy was notably absent in the original paper. Mabrouk et al. [16] proposed a multivariate quantization algorithm able to systematically assess learned dependencies while employing clustering techniques based on expectation maximization through the Gaussian mixture model. Furthermore, this algorithm outperformed the method devised by Friedman and Goldszmidt [8] in terms of prediction accuracy, quality of learned structures and computational efficiency. Moreover, Chen et al. [4] devised a quantization technique fashioned in such a way as to be used with the K2 algorithm in BN structure learning. Such an innovative quantization method optimizes quantization levels based on network structure by considering the relationships with quantified variable's parents, offspring and wives. The approach underwent a rigorous evaluation consisting in comparisons with a Bayesian quantization method utilizing datasets found in literature and mean cross-validated log-likelihood assessments. Additionally, Fang et al. [6] presented a unique quantization method grounded in matrix decomposition, specifically designed to enhance the accuracy of BN inference. Notably, this technique was tested using a two-node BN in order to exhibit its functional capabilities.

In recent years, Talvitie et al. [29] introduced an algorithm for Bayesian Network (BN) structure learning by incorporating an adaptive quantization approach for continuous variables. This method utilizes quantile quantization exhibiting considerable flexibility in adapting between 2 and 7 divisions based on the inherent structure parameters of the model. Comparative evaluations pitted this proposed algorithm against other structure learning methods in order to assess both the accuracy of discovered structures and predictive performance. The proposed approach outperformed its counterparts in these evaluations. Furthermore, Ciunkiewicz et al. [5] developed an open-source implementation of a dynamic quantization technique, leveraging relative entropy to intelligently select intervals that faithfully represent the underlying data distribution. Testing this methodology on datasets asserted its superiority in enhancing variable quantization when compared to static methods. Notably, while achieving improved quantization, the predictive performance exhibited no significant deviation from the performance of tested methods.

Numerous research publications have undertaken comprehensive comparisons of various discretization and quantization techniques for BNs. Notably, Nojavan et al. [20] conducted a meticulous comparative analysis of quantization methods within the context of BNs by focusing on a predetermined structure. Critical aspects were considered in their evaluation, such as CPTs, predictive modelling and practical management recommendations. Their findings revealed a nuanced landscape, where no single method emerged as definitively superior. In parallel, Beuzen et al. [2] undertook an assessment comparing manual, supervised and unsupervised quantization techniques employing a 4-node BN having a predetermined structure as testing ground. Outcomes were unsatisfactory, as each method possesses unique strengths and limitations. Moreover, Toropova and Tulupyeva [31] delved into the impact of diverse quantization approaches on BN performance, particularly while estimating behavioural rates. Results prominently highlighted that equal width quantization reached the highest and average levels of precision among methods under consideration.

It must also be noted that these studies should have taken into account the fundamental requirements for ample datasets while learning BNs so as to effectively model variable distributions [10, 32]. Insufficient data may lead to ill-informed or missing probabilities within CPTs, ultimately undermining the quality of model performance [17, 24]. Furthermore, the size of CPTs is intricately tied to the number of variable states, as well as its number of parents [22]. Thus, data adequacy depends on the BN structure.

To address the aforementioned data quantity limitation, Mayfield et al. [17] introduced the concept of Structure-Aware Discretization (SAD), a structure-aware quantization algorithm. SAD dynamically adjusts bin ranges and the number of bins to strike a balance between ensuring robust CPT coverage and providing sufficient bins for a reasonable resolution. If compared to Equal Cases Discretization (ECD) and the method proposed with no structure-unaware stage discretization (SUD), both SAD and SUD demonstrated comparable performance, notably surpassing ECD. A strategic reduction of bins in SAD ensures that each bin contains a predefined minimum number of instances while simultaneously diminishing the occurrence of CPT combinations with insufficient data. Depending on data distributions and sample sizes, rigidly adhering to a minimum limit for each category can potentially reveal modelling inaccuracies.

A notable omission in the majority of these studies pertains to the application of these quantization methods within the context of search and score of BNs structure learning, i.e. a pivotal technique for deriving BN structures from data

[19]. Interestingly, only Friedman and Goldszmidt [8], Chen et al. [4] and Talvitie et al. [29] extended their methods to the domain of BN structure learning. However, it is of paramount importance to observe that their evaluations primarily focused on the final structure attained at the end of the structure learning process. This approach may potentially constrain the scope of the analysis, given that these methods do not inherently ensure an identification of the global optimum [19]. A more comprehensive exploration of the entire landscape within the search space is warranted to adequately address such issue.

Our paper introduces a novel approach differing from Mayfield et al. [17], as the critical aspects of CPT completion are addressed along with BN structure-aware discretization and quantization. In our CPT limit-based quantization (CLBQ) method, a CPT size constraint derived from the dataset size was established so as to ensure a minimum level of CPT filling. Such limitation guarantees that CPT size remains within acceptable bounds for all CPTs in a given BN structure, meticulously examining the largest number of categories that system variables can accommodate. Subsequently, our algorithm systematically evaluates the number of bins within the range of 2 to this defined limit, while concurrently assessing BN structure scores and mean squared errors (MSE) between original and quantized data. The search process is concluded upon identifying a quantization value within the Pareto set [11], with a 2-degree divergence from the minimum score point. Such a meticulous approach ensures that variable quantization is conducted in a structure-aware manner, thus expertly balancing the trade-off between MSE and structure score while maintaining comprehensive CPT coverage.

Our method has been tested for the quantization of three different kinds of dataset: simulated discrete data with added noise, simulated continuous data and real data. CLBQ was compared to the Dynamic Discretization (DD) algorithm proposed in Ciunkiewicz et al. [5] considering the selected quantization and its mean squared error (MSE) in all datasets. Also, its suitability for employment in the search and score of BN structure learning was evaluated through landscape analysis. The key contributions of our proposed algorithm are:

- balance between data fidelity and structure score while at the same time maintaining CPT coverage during variable quantization;
- examining its potential influence on BN structure learning considering the reshaping the entire DAG search space.

This comprehensive assessment extends beyond the scope of prior studies, as unexplored facets of BN structure learning are addressed.

2 Related works

A detailed literature review conducted on Scopus in early 2023 resulted in the identification of a total of sixteen scholarly works. The investigation was centered on the domains of "Bayesian Network" and either "Discretization" or "Quantization". Information theory is employed to ascertain the appropriate quantization of variables for BNs. The determination of the quantization threshold in BN structural learning can be influenced by a metric derived from the Minimal Description Length principle (MDL) [8]. In the context of a specified structure, non-uniform partitions can be implemented as a strategy to mitigate the loss of information resulting from quantization, as suggested by Kozlov and Koller [13]. Furthermore, relative entropy error can be used to detect intervals that do not sufficiently capture the characteristics of the underlying distribution. This methodology makes it feasible to ascertain the intervals that ought to be merged or divided to get intervals that effectively represent the fundamental distribution [5].

Other studies utilize the expectation-maximization algorithm and likelihood functions to ascertain the appropriate quantization of variables for Bayesian networks. A scoring methodology can be implemented to assess the structure of BNs and the process of data quantization. This approach is a component of a multi-variable quantization technique, wherein each continuous variable is discretized based on its interactions (dependencies) with other variables [18]. An alternative methodology for BN structural learning involves the application of expectation maximization on the Gaussian mixture model, which is utilized to represent bins and determine quantizations by considering the clustering patterns within the distribution of variables [16]. In addition, an alternative approach in BN structural learning involves choosing the quantization with the highest probability based on the available data, considering the dependencies imposed by the BN structure. This method can be combined with the K2 algorithm for BN structural learning [4].

Furthermore, there are numerous varied approaches to tackle the matter of quantization. The approaches under consideration have distinct characteristics that set them apart from earlier methodologies. One possible approach for quantization is exploiting the significant correlation in high-dimensional data in conjunction with a non-parametric dimensionality reduction technique and a Gaussian mixture model, as Song et al. [27] suggested. Alternatively, a Genetic Algorithm can be employed to minimize the Normalized Root Mean Square Error (NRMSE) of a mainly selected output variable, considering the values of peaks and valleys [14].

In the study conducted by Fang et al. [6], an alternative approach to quantization is proposed, which involves employing matrix decomposition techniques to quantize variables that exist in different states with differing probabilities.

Alternatively, an adaptive quantization technique can be employed to optimize the selection of quantile bins for each variable within the range of 2 to 7 [29]. An alternative approach involves employing a structure-aware quantization technique that dynamically modifies the bin ranges and the number of bins to strike a suitable equilibrium between effectively completing the CPT and ensuring an adequate level of resolution. One approach to achieve this objective is to decrease the number of bins to ensure that each bin contains a pre-established minimum number of instances, hence diminishing the number of combinations in the CPT that have inadequate cases [17].

Various methods of quantization can be compared. A potential avenue of research is conducting a comparative study on quantization methods in BNs. This analysis would aim to evaluate the effects of different quantization methods and the number of intervals on the resulting BNs, employing a specified structure as Nojavan et al. [20] outlined. In their study, Nojavan et al. [20] found that no single technique simultaneously demonstrated superior performance in CPTs, prediction, and recommendation. Instead, each method shows excellence in one of these domains.

An alternative methodology for evaluating quantization approaches involves comparing manual, supervised, and unsupervised techniques within a pre-established framework while considering the model's accuracy and F-score accuracy [2]. The analysis conducted by Beuzen et al. [2] showed that manual quantization methods yielded BNs with greater semantic significance. On the other hand, supervised methods resulted in BNs with enhanced predictive capabilities. Additionally, unsupervised methods were deemed desirable due to their computational simplicity and versatility.

The quantization techniques for BN in a classification task can be examined to assess and contrast various quantization approaches, considering the classifier's performance [23]. In their study, Ropero et al. [23] compared the performance of four different methods: Equal Frequency, Equal Width, Chi-Merge, and Minimum Description Length principle. The findings of their study revealed that the Chi-Merge approach exhibited exceptional average performance across the conducted tests. In a study by Sari et al. [25], a comparison was made between equal-width, equal-frequency, and K-means methods for analyzing earthquake damage data. The results of the tests indicated that the K-means approach yielded the highest level of accuracy. Toropova and Tulupyeva [31] conducted a comparable study to estimate the behavioural rate and compare different quantization methods, including equal width, frequency, EF_Unique, and expert quantization. The findings revealed that equal-width quantization yielded the highest level of precision on average.

In contrast to the quantization methods discussed in this part, CLBQ can perform data quantization on a fixed BN structure or during BN structural learning. The utilization of structural learning, as observed in several literary works, is often limited to examining the final outcome, neglecting the comprehensive exploration of the entire search space, a distinctive feature of the present study. Moreover, the CLBQ technique also addresses the impact of quantization on the size of the CPT and the structural score. This aspect of CLBQ contributes to its originality and sets it apart from alternative approaches.

3 Material and methods

This section describes the proposed quantization method and its evaluations. The CLBQ algorithm was introduced and tested based on simulated and real datasets. A summary of test cases and datasets used on each of them can be seen in Table 1.

3.1 CPT limit-based quantization (CLBQ)

Figure 1 shows a flowchart of the proposed method and its key steps and components. Numerous samples are required to model the conditional distributions of variables. A lack

Table 1 Summary of test cases and datasets used in each of them by detailing their type, number of samples, number of variables and equations and figures used to find its generation equation, expected structure, and signal and distributions of its variables

Test case	Dataset	Type	Number of samples	Number of variables	Equation	Expected structure	Signal and distribution of variables
Functionality test	D4	Discrete with added noise	10^5	4	Equation (8)	Figure 3	Figure 6
Simulated continuous data test	XYZ	Continuous	10^6	3	Equation (9)	Figure 4	Figure 7
	XYZ3	Continuous	10^6	3	Equation (10)	Figure 5	Figure 8
Real-data landscape analysis	Weather	Real	107,802	4	–	–	Figure 9

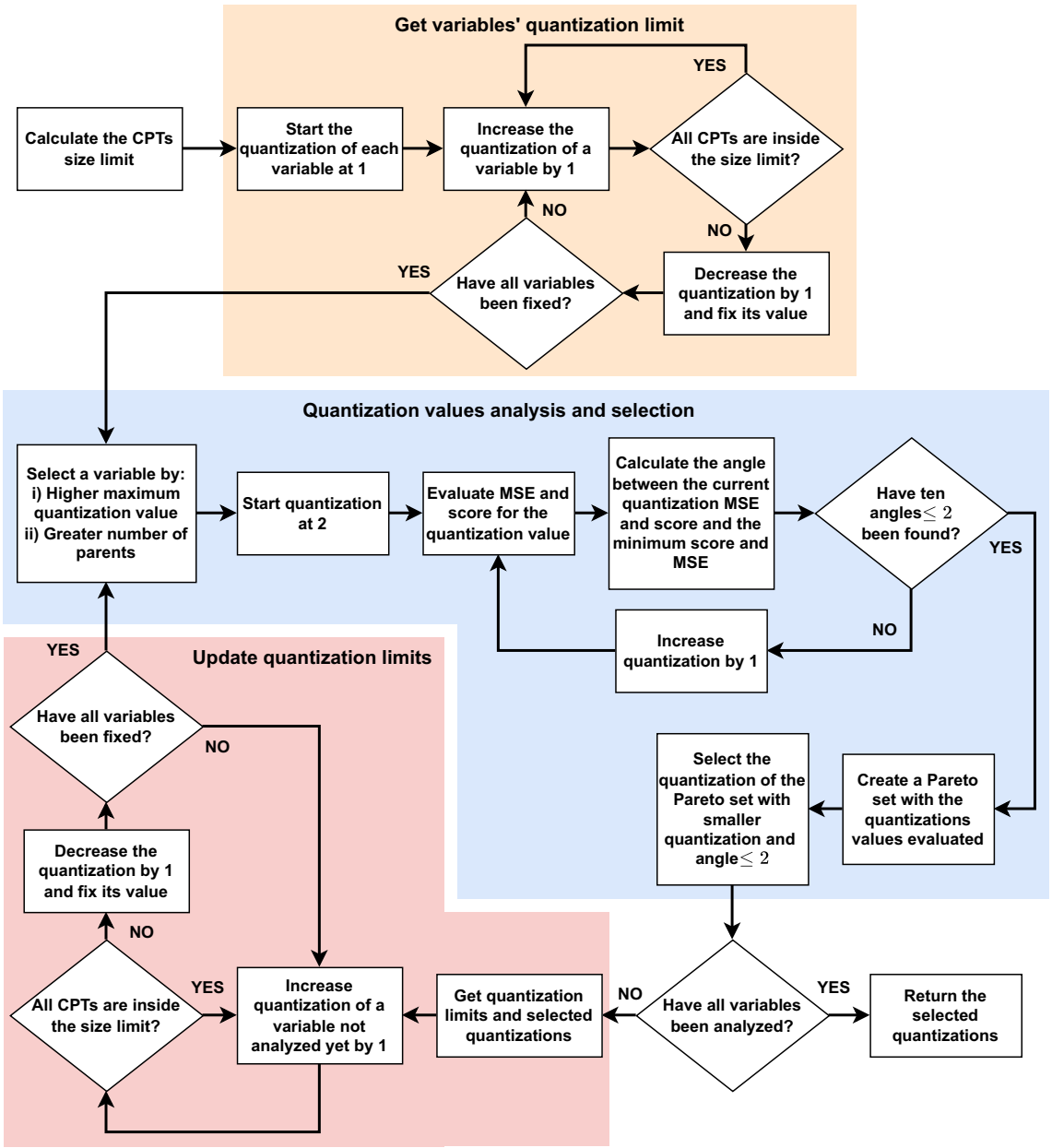


Fig. 1 Flowchart of CLBQ illustrating the steps taken to quantize data given a BN structure. It covers the discovery of the quantization limits, the order of selection of variables for analysis and the selection of quantization for each variable

of samples can result in missing or uninformed probabilities in the CPT, thus leading to poor model performance [10, 17, 32]. Given the above, the initial step of CPT limit-based quantization (CLBQ) is to identify a limit size for the CPT considering the number of samples the dataset has to enhance distribution modeling and achieve good model quality. This is performed according to the following equation.

$$\mathcal{M} = \frac{\text{Number of samples}}{\eta} \tag{1}$$

where $\mathcal{M}$ is the CPT size limit, and η is the number of desirable samples for each element of the CPT so as to achieve good model quality. This only defines the limit, thus CPTs can achieve better modelling than that at the end of the process. $\eta = 3$ was used in our test. This value only controls the upper quantization limit. Since BIC score increases as there is smaller quantization, this upper limit and value of η exert no influence on CLBQ quantization, once the dataset has enough data. When data is lacking, η ensures a certain level of model quality for all quantization values considered

by the CLBQ. The value of $\eta = 3$ was selected on account of the fact that the worst-case scenario, i.e. a uniform distribution, has three points for each element of the CPT, thus offering modelling of acceptable quality.

Once CPT size limit and BN structure are defined, it is possible to calculate the dimension of CPTs of each variable considering the number of states (quantization) each variable has. The CPT dimension of a variable is given by

$$\mathcal{L}_v = q_v \prod_{i=1}^{p} q_i \tag{2}$$

where $\mathcal{L}_v$ is the CPT dimension for a variable, q_v is the variable quantization, p is the number of parents of the variable and $[q_1, \ldots, q_p]$ are the quantization of parent variables.

Then, the operation of checking whether $\mathcal{L}_v \leq \mathcal{M}$ can be computed given each variable quantization. Thus, a heuristic method was used to compute the quantization limit of each variable considering their restrictions. The method consists in initializing the quantization of each variable in 1, increasing them one at a time by 1 and checking whether all $\mathcal{L}_v$ still comply with $\mathcal{L}_v \leq \mathcal{M}$. This is performed using the matrix product by creating a quantization matrix and a dependency matrix able to store which node is the parent of which node and the node itself, thus storing which quantization values are needed to compute each CPT dimension. This dependency matrix does not store correlation measurements, but only reveals if that node must be considered on the CPT dimension computation with a 1 or with a 0 if it is not needed.

$$\begin{aligned}\begin{bmatrix} q_1 & q_2 & \cdots & q_n \end{bmatrix} \times \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{nn} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \\ = \begin{bmatrix} q_1 a_{11} & q_1 a_{12} & \cdots & q_1 a_{1n} \\ q_2 a_{21} & q_2 a_{22} & \cdots & q_2 a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ q_n a_{n1} & q_n a_{n2} & \cdots & q_n a_{nn} \end{bmatrix}\end{aligned} \tag{3}$$

where $a_{ij} \in \{0, 1\}$ represents relationships between variables, thus selecting which q_i values must be included in the calculation of CPT size. Then, the values of $q_i a_{ij} = 0$ are replaced by 1, and a column-wise product is carried out so as to calculate all CPTs sizes.

$$\mathcal{L}_v = \prod_{i=1}^{n} q_i a_{iv} \tag{4}$$

When a variable reaches its limit (if increased, $\mathcal{L}_v \leq \mathcal{M}$ for $v \in [1, 2, \ldots, n]$ is broken) its quantization is fixed on the limit, and the process continues to increase the other variables and test the condition. The process is over when all variables have their quantization fixed on their limit value.

Afterwards, an analysis of quantization values between 2 and the previously found quantization limit is carried out so as to select a quantization value having good trade-off between the mean squared error (MSE) of original values and quantized values as well as the BN structure score. For such a purpose, variables are selected one at a time according to the priority list below: (i) higher maximum quantization value (ii) greater number of parents. Once a variable has been selected, quantization values, starting from $q_v = 2$, are used to evaluate the BN score considering quantized data and the MSE between the quantized data and original data. For each evaluation, an angle is calculated as follows.

$$\theta_q = \arctan\left(\frac{\Delta \text{MSE}_q}{\Delta \text{Score}_q}\right) \tag{5}$$

$$\Delta \text{MSE}_q = \frac{|\text{MSE}_{MIN} - \text{MSE}_q|}{|\text{MSE}_{MIN} - \text{MSE}_{MAX}|} \tag{6}$$

$$\Delta \text{Score}_q = \frac{|\text{Score}_{MIN} - \text{Score}_q|}{|\text{Score}_{MIN} - \text{Score}_{MAX}|} \tag{7}$$

MSE_{MIN} is found by considering the quantization limit of the variable. MSE_{MAX} and Score_{MAX} considers $q_v = 2$. Score_{MIN} is calculated considering all variables in their quantization limit. An example of what angle θ represents can be seen in Fig. 2.

The analysis of increasing values of q_v is stopped after ten increasing q_v values where $\theta \leq 2$. With these values, a Pareto set is created, and the smallest quantization value belonging to the Pareto set where $\theta \leq 2$ is selected as the variable quantization value.

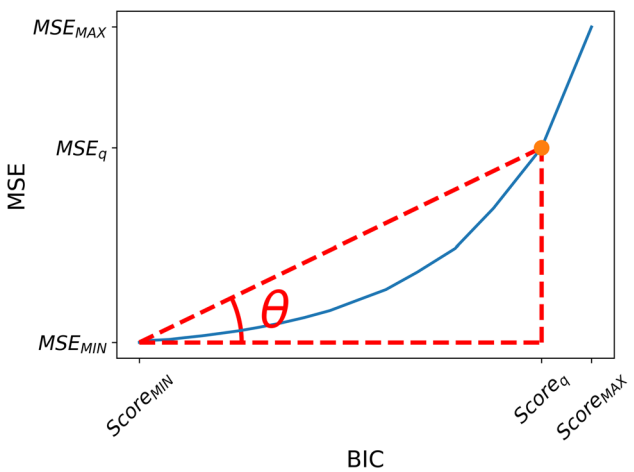


Fig. 2 Example of what angle θ represents. BIC is a structure score used in it. The quantization value under analysis is depicted in orange and the blue line represents the curve of all quantization values for the variable under analysis

After a quantization value is selected, it is fixed and the maximum quantization value of the other variables is updated accordingly. To evaluate the score and MSE, quantizations of variables that are not under analysis are set to the highest quantization value found or their selected quantization value. This process is repeated until all system variables have been analyzed. The algorithm of this process can be seen in Algorithms 1 and 2. The code of CLBQ is available at https://www.doi.org/10.5281/zenodo.8368057.

Algorithm 1 CLBQ - Part 1.

```
function CLBQ(dag, data, score, η)
    M = int(len(data)/η)
    Q_MAX = get_quantization_limit(dag, len(data), η)
    fixed_values = [False, False, ..., False]
    Score_MIN = score(dag, equal_quantization(data, Q_MAX))
    while not(all(fixed_values==True)) do
        v = select_node(fixed_values, Q_MAX, dag)        ▷ Selects the variable to be analysed
        MSE_MIN = mse(data, equal_quantization(data, Q_MAX))
        Q_MAX[v] = 2
        MSE_MAX = mse(data, equal_quantization(data, Q_MAX))
        MSE_MAX = score(dag, equal_quantization(data, Q_MAX))
        score_list = []
        mse_list = []
        angle_list = []
        q_list = []
        i = 0
        for q_v in range(2, q_vMAX) do
            q_list ← q_v
            Q_MAX[v] = q_v
            Score_v = score(dag, equal_quantization(data, Q_MAX))
            score_list ← Score_v
            MSE_v = mse(data, equal_quantization(data, Q_MAX))
            mse_list ← MSE_v
            if MSE_MAX==MSE_MIN then
                ΔMSE = 0
            else
                ΔMSE = |MSE_MIN − MSE_v|/|MSE_MIN − MSE_MAX|
            end if
            ΔScore = |Score_MIN − Score_v|/|Score_MIN − Score_MAX|
            if ΔScore == 0 then
                θ = 0
            else
                θ = arctan(ΔMSE/ΔScore)
            end if
            angle_list ← θ
            if θ ≤ 2 then
                i = i + 1
                if i ≥ 10 then
                    break
                end if
            end if
        end for
```

Algorithm 2 CLBQ - Part 2.

```
        pareto = get_pareto(score_list, mse_list)
        q_c = ∞
        if len(pareto)==1 then
            q_c = q_list[pareto[0]]
        else
            for j in pareto do
                if angle_list[j]≤ 2 then
                    if q_list[j]< q_c then
                        q_c =q_list[j]
                    end if
                end if
            end for
        end if
        if q_c == ∞ then
            q_c = q_list[pareto[-1]]
        end if
        Q_MAX[v] = q_c
        fixed_values[v] = True
        Q_MAX = adjust_quantizations(Q_MAX, dag, len(data), η, fixed_values)
    end while
end function
```

3.2 Datasets

This section describes all datasets used for testing. The first one was a discrete simulated dataset with added noise so as to identify the ideal quantization and enhance and ease the analysis of its quantization results. Afterwards, two continuous simulated datasets were used to analyze how CLBQ would perform with no ideal quantization. Finally, a continuous real data dataset using variables with different characteristics, ranges and distributions was used to confirm whether results would remain true to real data.

3.2.1 D4

D4 is a simulated dataset comprising four nodes (A, B, C, and D) generated by

$$\begin{cases} A' = \text{rand\_int}(0, 9) \\ B' = \text{rand\_int}(0, 9) \\ C' = A' + B' \\ D' = A' + \text{rand\_int}(0, 9) \\ A = A' + \mathcal{N}(\mu = 0, \sigma = 0.4) \\ B = B' + \mathcal{N}(\mu = 0, \sigma = 0.4) \\ C = C' + \mathcal{N}(\mu = 0, \sigma = 0.4) \\ D = D' + \mathcal{N}(\mu = 0, \sigma = 0.4) \end{cases} \tag{8}$$

where rand_int(0, 9) is a uniform sample of integers in [0, 9] and $\mathcal{N}(\mu = 0, \sigma = 0.4)$ is a sample of a normal distribution with $\mu = 0$ and $\sigma = 0.4$. Thus, it is observed that the ideal quantization of variables would be $q_A = 10, q_B = 10, q_C = 19$, and $q_D = 19$. This dataset has 10^5 samples. A figure showing the dataset variables and their distribution can be seen in Fig. 6. The expected structure for this dataset can be seen in Fig. 3.

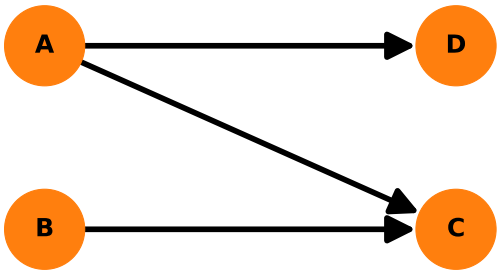


Fig. 3 Expected BN structure for D4 dataset obtained from its generation equation found in (8)

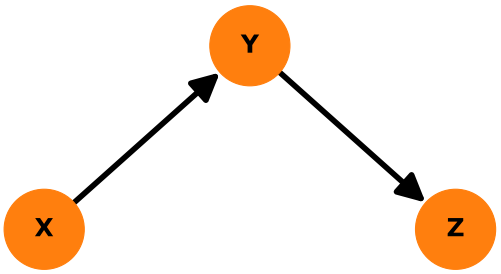


Fig. 5 Expected BN structure for the XYZ3 dataset obtained from its generation equation found in (10)

3.2.2 XYZ

The XYZ dataset is a three-variable dataset (X, Y, and Z) generated by

$$\begin{cases} X = \mathcal{N}(\mu = 0, \sigma = 1) \\ Y = 3 \cdot X + 1 + \mathcal{N}(\mu = 0, \sigma = 1) \\ Z = 2 \cdot X + 2 + \mathcal{N}(\mu = 0, \sigma = 1) \end{cases} \tag{9}$$

where $\mathcal{N}(\mu, \sigma)$ is a sample of normal distribution. This dataset has 10^6 samples. A section of dataset variables and their distribution can be seen in Fig. 7. The expected structure for this dataset can be seen in Fig. 4.

3.2.3 XYZ3

The XYZ3 dataset is composed of three variables (X, Y, and Z) generated by

$$\begin{cases} X = \mathcal{N}(\mu = -2, \sigma = 0.1) \cup \mathcal{N}(\mu = 0, \sigma = 0.1) \\ \quad \cup \mathcal{N}(\mu = 2, \sigma = 0.1) \\ Y = 3 \cdot X + 1 + 0.7 * \mathcal{N}(\mu = 0, \sigma = 1) \\ Z = 2 \cdot X + 2 + 0.7 * \mathcal{N}(\mu = 0, \sigma = 1) \end{cases} \tag{10}$$

where $\mathcal{N}(\mu, \sigma)$ is a sample of normal distribution. This dataset has 10^6 samples. A section of dataset variables and their distribution can be seen in Fig. 8. The expected structure for this dataset can be seen in Fig. 5.

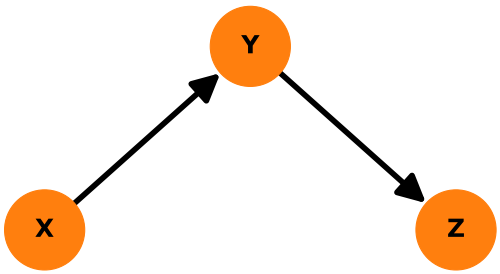


Fig. 4 Expected BN structure for the XYZ dataset obtained from its generation equation found in (9)

3.2.4 Weather

Weather dataset was acquired in Imperatriz (5° 31′ 33″ S, 47° 28′ 33″ W), a Brazilian city in the state of Maranhão, containing hourly measurements of temperature, humidity, radiation and wind speed. Measurements started at 23:00 on 02/03/2008 and ended at 14:00 on 21/01/2022. Such data was obtained from the National Institute of Meteorology (INMET), a public and open dataset, able to be accessed at https://tempo.inmet.gov.br/TabelaEstacoes/A225. A section of dataset variables and their distribution can be seen in Fig. 9.

3.3 Functionality test

Dataset D4 was used to test the functionality of CLBQ. It was performed by following a step-by-step quantization process considering the expected structure for the dataset. In this process, tested quantization values, their score, MSE and angle θ were saved and plotted so as to better understand the algorithm dynamics. An example of how data results are presented can be seen in Fig. 10. In addition, the result was compared to the ideal quantization from D4, as it is a discrete system with added noise. Ideal and CLBQ quantizations were compared considering the MSE of each variable between original values and quantized values, as well as the total MSE, i.e. a sum of the MSE of all variables. CLBQ was also compared to the Dynamic Discretization (DD) algorithm proposed in Ciunkiewicz et al. [5] considering its quantization and MSE for each variable. DD was selected, as its code was provided. The quantization selected by DD was also plotted in a histogram with the distributions of variables so as to observe how well the selected quantization modelled the distributions.

D4 was also used to test the fitness of CLBQ to be used on the search and score BN structure learning. For such a

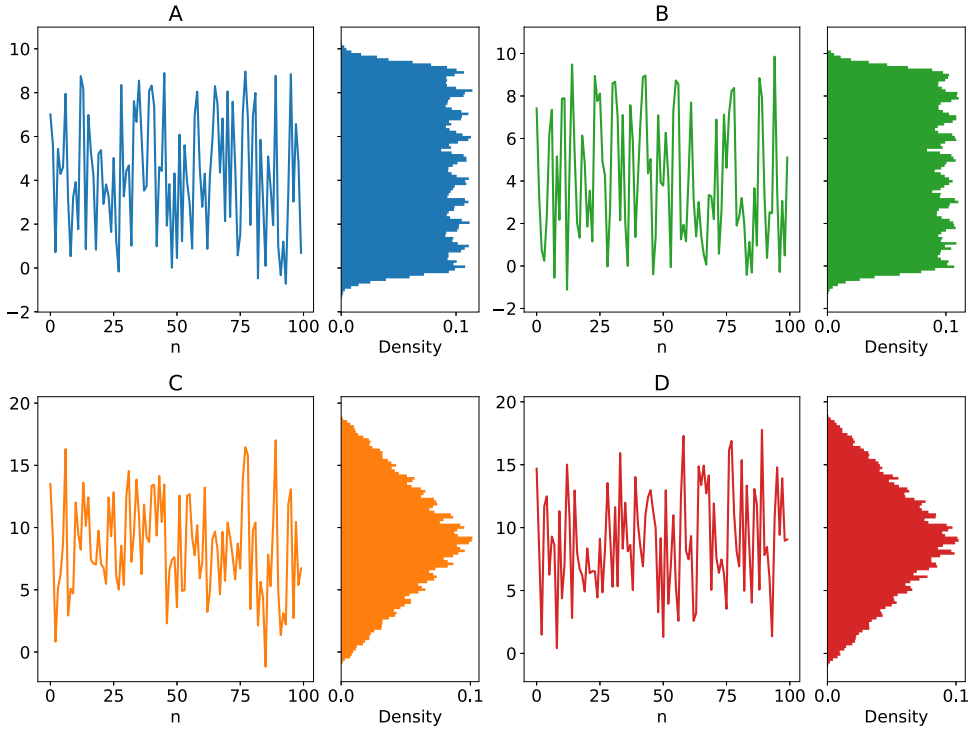


Fig. 6 Plot of a section of the signal of each variable and the distribution of each variable for dataset D4. Each variable is represented using different colors and their name can be seen at the top of each subplot. The y-axis shows the value of this variable for all plots. The x-axis shows the sample index for the signal plot and density for the distribution plot

purpose, a landscape analysis was performed to investigate how the search space of BN structures would change using CLBQ, and all 543 possible DAGs were scored considering the data quantized by CLBQ. It was carried out for 10^3, 10^4, and 10^5 samples. The time taken by CLBQ to select a quantization for each structure was saved and also analysed. Numeric values were assigned to each DAG to plot the results in a figure. Thus, DAGs were sorted by the number of edges, and they were sorted by score inside each group using the same number of edges. The score used for sorting was the one using 10^5 samples. Each group was then equally spaced in the interval [number of edges, number of edges+1). By

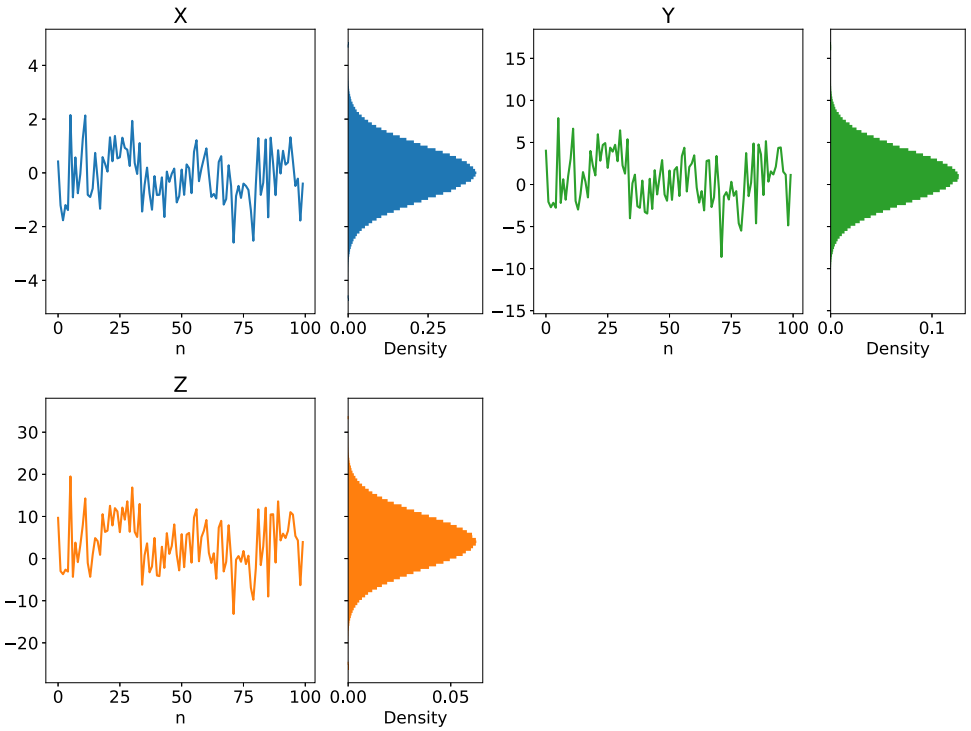


Fig. 7 Plot of a section of the signal of each variable and the distribution of each variable for dataset XYZ. Each variable is represented using different colors and their name can be seen at the top of each subplot. The y-axis shows the value of this variable for all plots. The x-axis is the index of the sample for the signal plot and density for the distribution plot

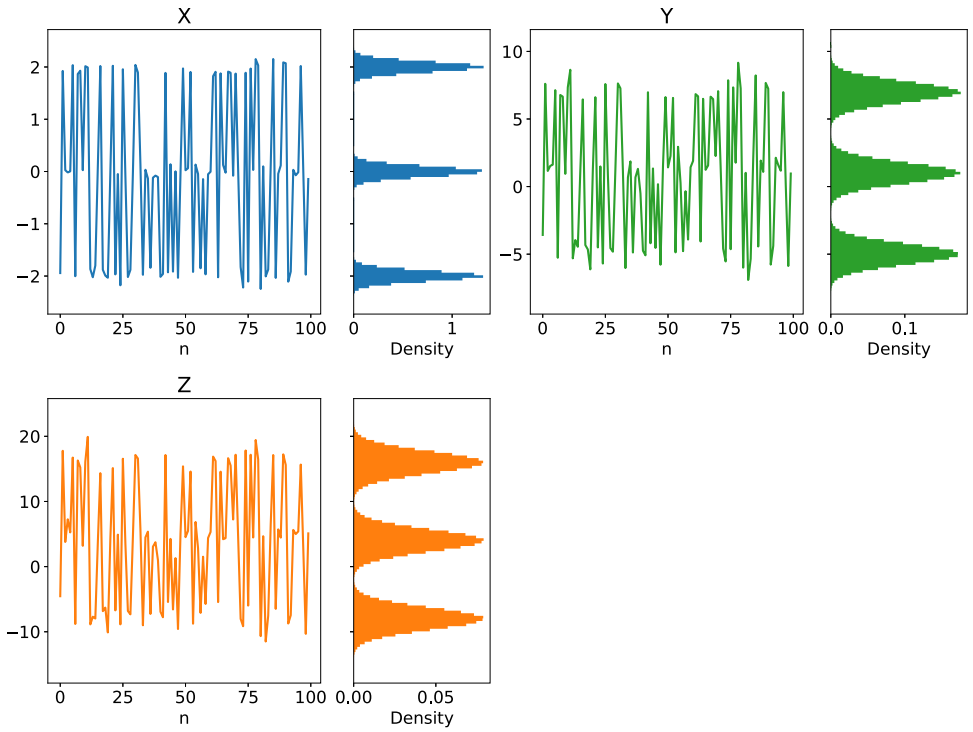


Fig. 8 Plot of a section of the signal of each variable and the distribution of each variable for the XYZ3 dataset. Each variable is represented by a different color and their names can be seen at the top of each subplot. The y-axis shows the value of this variable for all plots. The x-axis is the sample index for the signal plot and density for the distribution plot

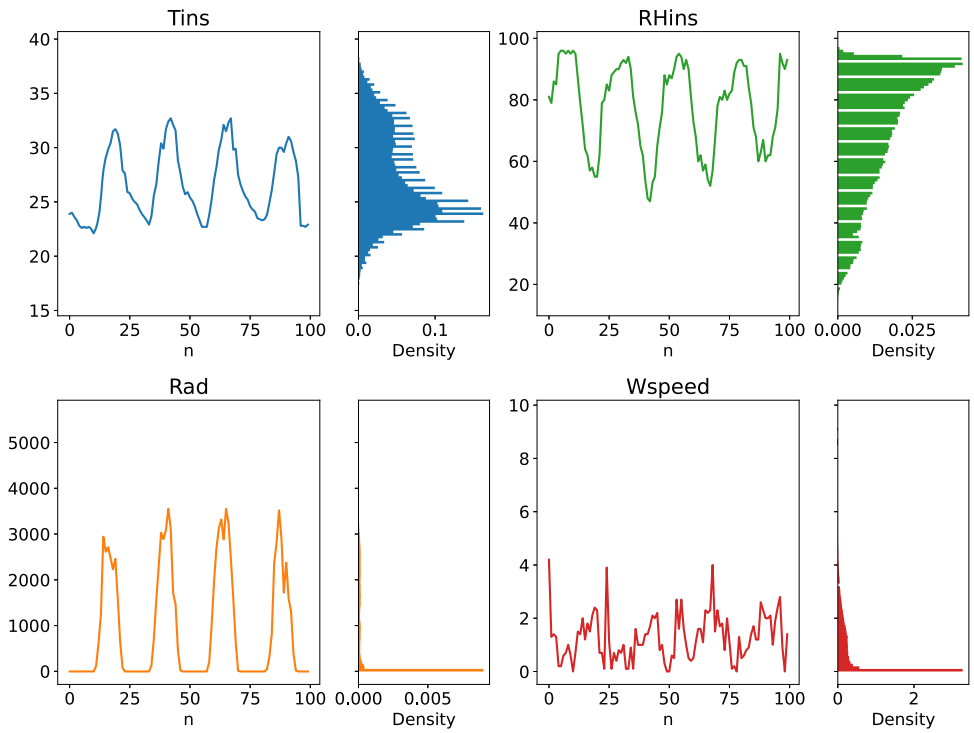


Fig. 9 Plot of a section of the signal of each variable and the distribution of each variable for the Weather dataset. Each variable is represented using different colors and their names can be seen at the top of each subplot. The y-axis is the value of this variable for all plots. The x-axis is the sample index for the signal plot and density for the distribution plot

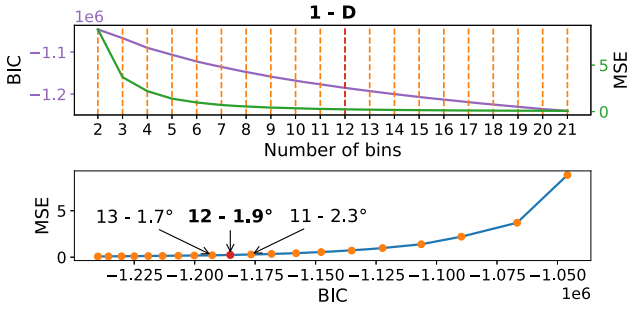


Fig. 10 Example of how data on BIC, MSE and θ are going to be presented in the functionality test. The number at the top in bold letters indicates which step it is and the name of the variable under analysis on it is on its side. BIC is plotted in purple and MSE in green. Number of bins is the variable quantization. The Pareto front is plotted in blue. The Pareto set quantizations are depicted in orange on both plots. The quantization selected by the CLBQ is in red. Angle θ is indicated by annotation at the bottom plot for the selected quantization and its neighbors. This indication contains the quantization value - θ

doing so, the differences and similarities between the search spaces considering different numbers of samples could be observed. One example of how the numbering of DAGs was performed can be seen in Fig. 11.

3.4 Simulated continuous data test

As means to further test the CLBQ, two simulated continuous datasets were used to observe its application using continuous data. The initial test was the quantization found by CLBQ for datasets XYZ and XYZ3, which was plotted against the actual variable distribution in order to analyze if the quantization was able to model well the distribution of the variables. CLBQ was again compared to DD by considering the chosen quantization and its MSE. DD quantization was also plotted against the variables' distribution to find if quantization has well modelled the distribution of variables.

The second test was a landscape analysis to pinpoint where the expected structure is in order to reveal if the search algorithm could actually find the expected structure. Lastly, if DAGs were better than those expected for the same number of edges, the quantization found using the CLBQ, the score found by using it and that found using the same quantization as that in the expected structure were analyzed.

3.5 Real-data landscape analysis

The same tests were performed once more by using the Weather dataset to confirm whether the results obtained from the functionality and simulated continuous data tests still held in real data. To discover an expected structure, the PC algorithm was used [28], given that it is a constraint-based BN structure learning algorithm [19]. In addition, the quantization found for the best structures in the landscape analysis were also compared considering MSE. DD was once more compared to CLBQ considering its quantization and MSE, and its histogram was plotted against the variables' distribution for analysis.

4 Results

The results of tests and analyses follow the same structure as that presented in the Material and Methods section.

4.1 Functionality test

A step-by-step analysis of CLBQ for dataset D4 can be seen in Fig. 12. It shows the evaluations made by plotting the score and MSE for each evaluation, the Pareto set chosen during each step, and the Pareto front found. In the Pareto front, angle θ for the chosen quantization and the previous and next values on the Pareto set can be seen. This analysis allowed clarifying the functioning of CLBQ.

Table 2 shows the ideal quantization, the one found by CLBQ and that found by DD. It also shows the MSE of each quantization for each variable. Figure 13a shows the actual distributions of variables in a dashed line and a histogram of the quantization found by CLBQ. Figure 14a shows the actual distributions of variables in a dashed line and a histogram of the quantization found by DD.

Results of the landscape analysis considering different quantities of samples for dataset D4 can be seen in Fig. 15a.

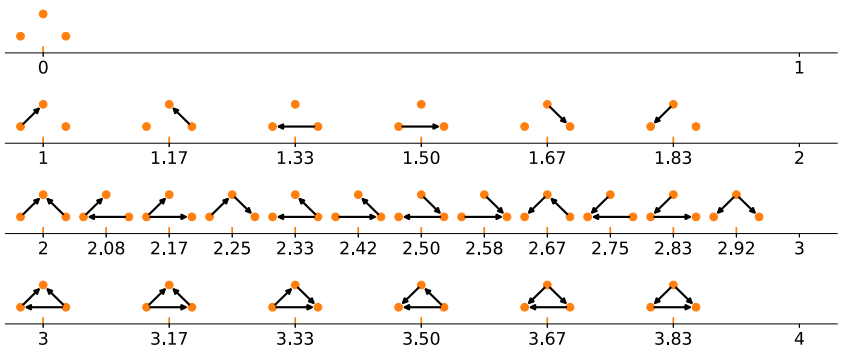


Fig. 11 Example of how the numbering of DAGs was performed for landscape plots, i.e. by converting BN structures into numbers to be placed on the x-axis. A system with 3 variables was considered for such example, since it only has 25 possible DAGs, to ease visualization

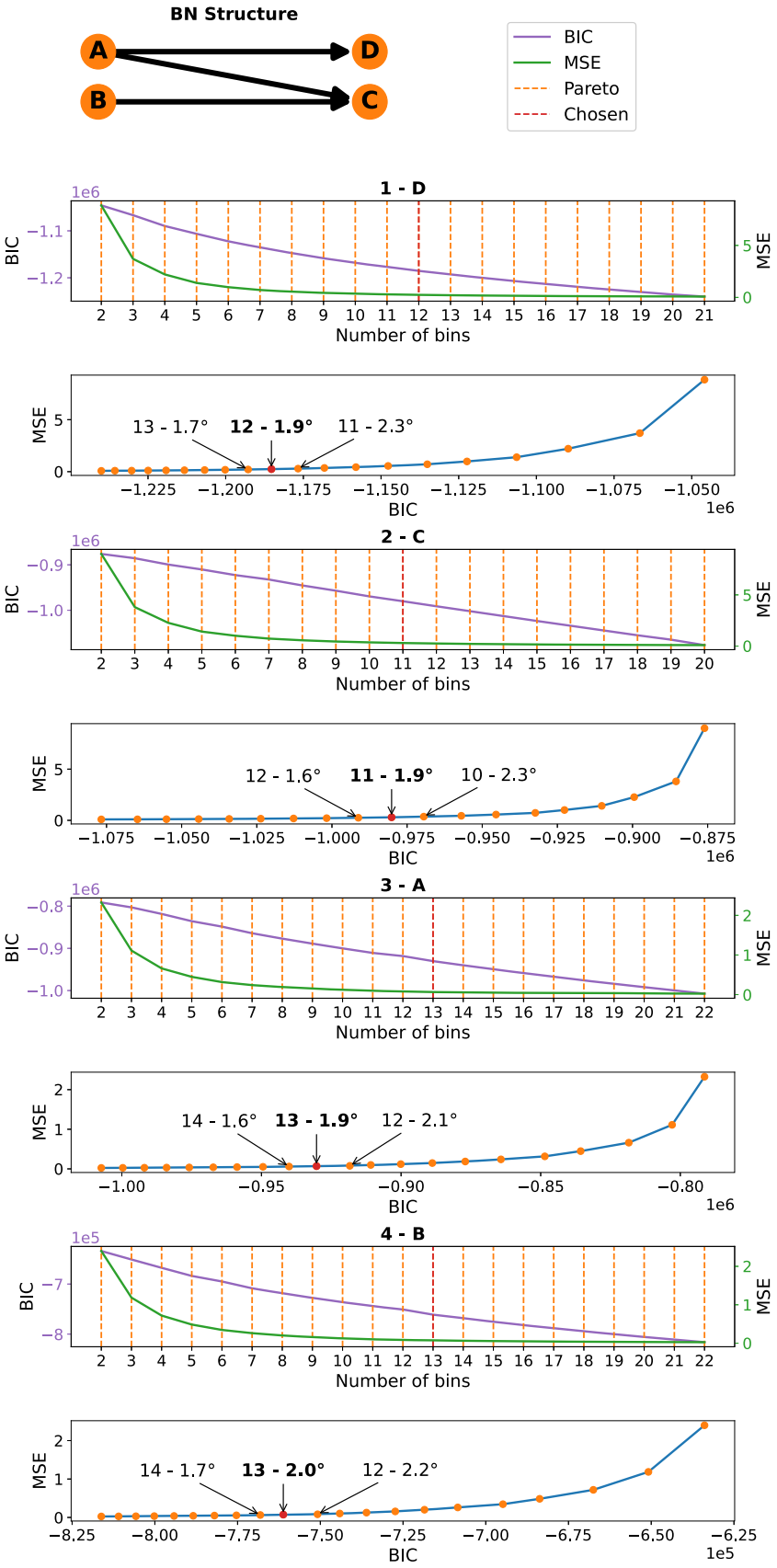


Fig. 12 Step-by-step analysis of CLBQ for the D4 dataset and its used structure. On the BICxMSE plot, the Pareto front found by CLBQ is plotted, and the number of bins and angle θ are presented for the selected quantization (highlighted in bold letters) and its first neighbours. At the top of each chart, the number of each step is shown in bold letters and which variable is being analyzed on it. BIC is the structure score, MSE is the mean squared error between quantized data for a given number of bins and the original data. Number of bins refers to the amount used to quantize the variable. Pareto refers to the members of the pareto set found. Chosen refers to the selected quantization value from the pareto set

Table 2 Quantization and their respective MSE for the functionality test using dataset D4

	q_A	q_B	q_C	q_D	MSE_A	MSE_B	MSE_C	MSE_D
Ideal	10	10	19	19	0.120	0.126	**0.100**	**0.097**
CLBQ	13	13	11	12	**0.069**	**0.074**	0.298	0.245
DD	9	13	11	4	5.763	0.988	1.334	2.208

The ideal quantization values, the quantization values found by CLBQ and quantization values obtained from DD are shown. Values with q are quantization values (the number of values able to be assumed by the variable) and MSE values are the mean squared error of original and quantized data. The bold values are the best values for each of those measures

The mean and STD time of execution of CLBQ when making the landscape analysis can be seen in Table 3.

Finally, in Fig. 16a, the expected structure location found on the landscape analysis is shown. Thenceforth, it can be observed that the expected structure achieved the best score for BNs with 3 edges. In Fig. 17a, the BNs whose score equals or is higher than that found for the expected structure with the same number of edges can be observed.

4.2 Simulated continuous data test

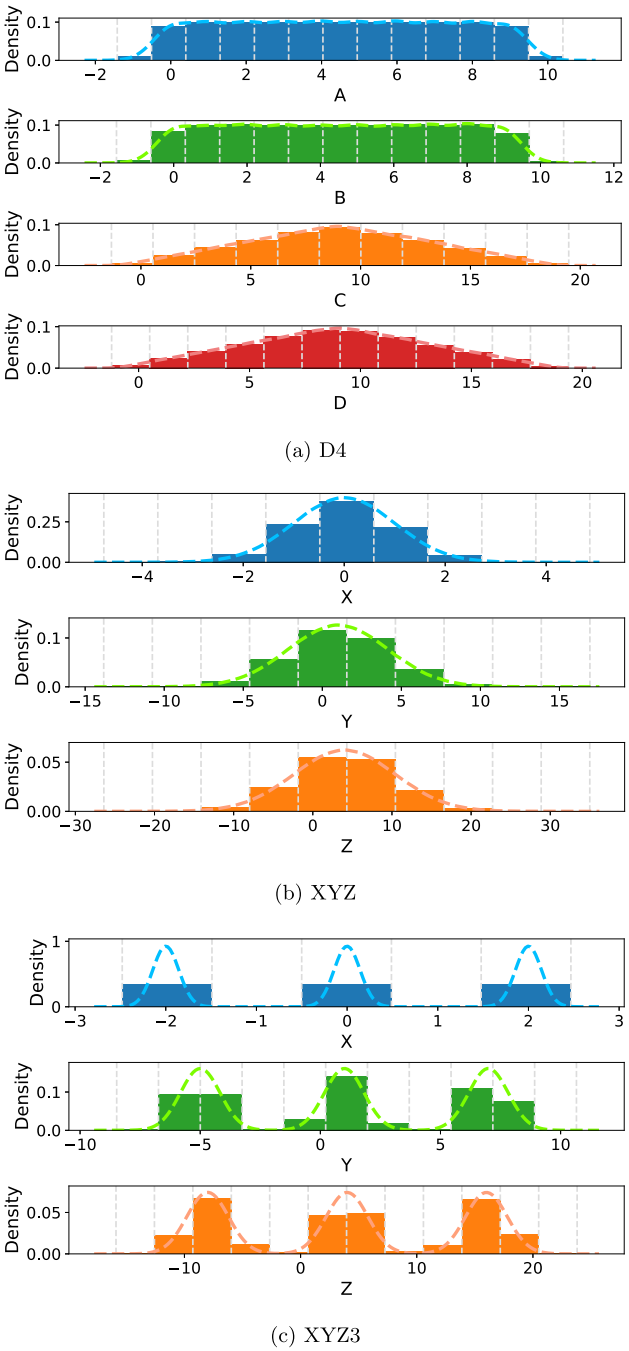
The result of CLBQ for dataset XYZ was $q_X = 9$, $q_Y = 10$ and $q_Z = 10$. The result of CLBQ for dataset XYZ3 was $q_X = 5$, $q_Y = 11$, and $q_Z = 12$ (Table 4). A comparison between its quantization histogram and the actual distribution found for both cases can be seen in Fig. 13. Figure 15 shows the landscapes for different sample sizes of XYZ and XYZ3 datasets. The mean and STD time of execution of CLBQ when making the landscape analysis can be seen in Table 3.

The landscape analysis for datasets XYZ and XYZ3 can be seen in Fig. 16. Both show that, although the expected structure is among the best scores, it is not intrinsically the best. To understand the reason for such, an analysis of structures whose scores were higher than or equal to the expected structure was performed. Structures achieving a better score, their quantization, score using its quantization and the score using the expected structure quantization can be seen in Fig. 18.

4.3 Real-data landscape analysis

Results of the landscape analysis can be seen in Fig. 15b. The mean and STD time of execution of CLBQ when making the landscape analysis can be seen in Table 3. Figure 19 shows the structure obtained using the PC algorithm. The landscape analysis and the structure location on it can be seen in Fig. 16b. The best BNs with 3 and 4 edges can be seen in Fig. 17b.

As it can be seen, CLBQ selected the same quantizations for all the best structures that have the same number of edges.



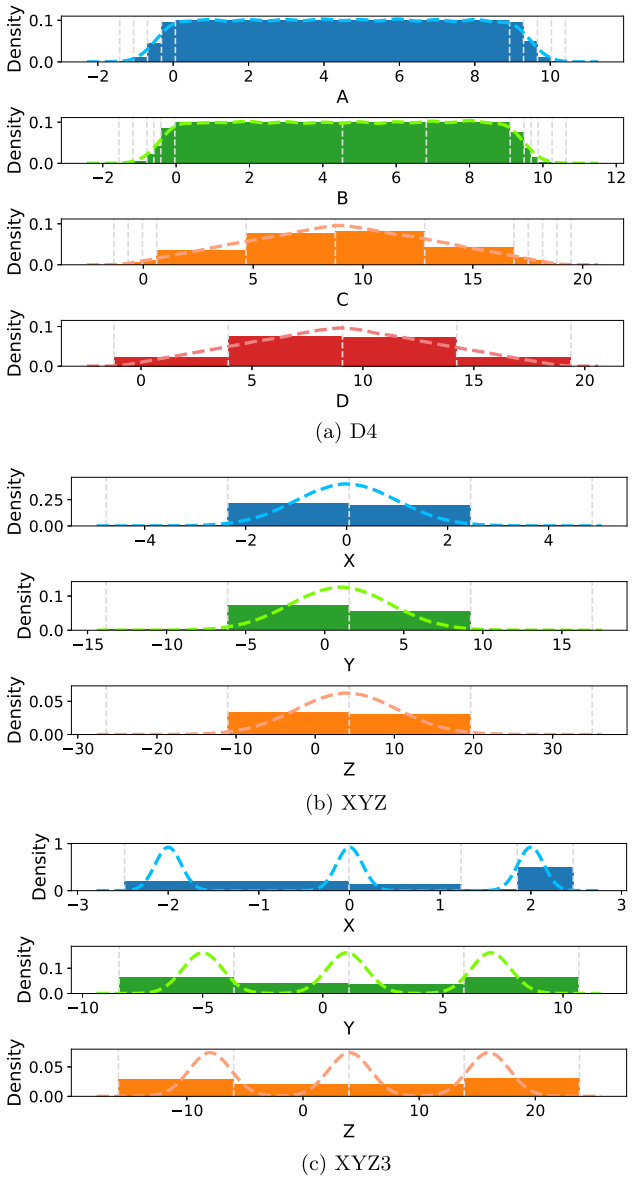
(a) D4

(b) XYZ

(c) XYZ3

Fig. 13 Comparison between the distribution of variables and a histogram of the variables of the datasets with the quantization found by CLBQ considering the expected structure for the datasets D4, XYZ and XYZ3. The distributions of the variables are shown in lighter-coloured dashed lines

Table 5 shows quantizations and their MSE for the best BNs found and the best BN found by PC for comparison purposes. It also shows the quantization selected by DD and its MSE. These quantizations can also be seen in Figs. 20 and 21, where their histograms are compared to the distributions of variables.



(a) D4

(b) XYZ

(c) XYZ3

Fig. 14 Comparison between the distribution of variables and a histogram of variables showing the quantization found by DD for datasets D4, XYZ and XYZ3. Distributions of variables are depicted in lighter-coloured dashed lines

5 Discussion

Regarding quantization and histogram results, Table 2 reveals that although CLBQ has not selected an ideal quantization, it achieved a quantization with a good MSE. Moreover, Fig. 13 shows that despite not being an ideal quantization, the quantization found by CLBQ for D4 also has good representation of the distributions of the variables. As for the histograms for datasets XYZ and XYZ3, it can be seen that the selected values result in a good modelling of distributions for both cases. The result achieved for variable X on dataset XYZ3 is rather interesting, as it presents three separated peaks CLBQ chose to model using one bin for each, and having one bin between each peak. In Table 5 and Fig. 20, it is evidenced that all three different quantizations selected by CLBQ for the two best structures and the structure obtained for the PC presented good modelling of the distributions of the variables. Finally, given a comparison with DD, CLBQ performed better if considering MSE for all test cases. Moreover, histograms reveal that the modelling made by DD was insufficient if compared to CLBQ.

Considering the landscapes for different volumes of samples shown in Fig. 15, its results indicate that CLBQ can be applied in the search and score of BN structure learning, as the best structures are the same for different volumes of samples. Thus, despite a change in the score range and slight changes in the search space, the best BNs are still the same and the search space presents the same behaviour. Such consistency of best structures and behaviour is observed for all datasets and test cases.

Considering the execution time of CLBQ shown in Table 3, it can be seen that the execution time varies a lot between dataset sizes and between datasets. This is probably caused by the evaluation of the score and the variables' distributions. The BIC score used counts the number of points that each element of each CPT has, thus its evaluation time is dependent on the dataset size. Moreover, the variables' distributions could generate the difference between datasets, as the number of quantization values needed to be evaluated until a trade-off is found varies according to the variables' distribution that affects the structural score and the MSE values used on CLBQ. When looking at the time CLBQ needs to select quantizations, although it may not be as small as you may want it, for a one-structure evaluation it is very quick and for use in search and score BN structural learning, it may add a considerable running time, however, since the structural learning process is already very time intensive, adding this time in exchange for having the quantization done for each structure is something that could be of value.

Considering the expected BN location on the landscape analysis shown in Fig. 16, it is found that the expected BN for dataset D4 was the best structure with three edges. Meanwhile, the expected BN was not the most desirable for datasets XYZ and XYZ3, but had a good score value. As for the Weather dataset, the expected BN found using PC is not the best structure found in the landscape analysis. This can be explained by the fact that the BIC score has a function that penalizes the addition of edges on the structure if it does not lead to a good increase in the description of data [3]. This means that, for the dataset being used, just 3 or 4 edges were enough to balance model complexity and data description. Thus, BNs with 5 and 6 edges had higher penalization as a result of model complexity and achieved a worse score than that found for the best BNs with 3 and 4 edges. An analysis

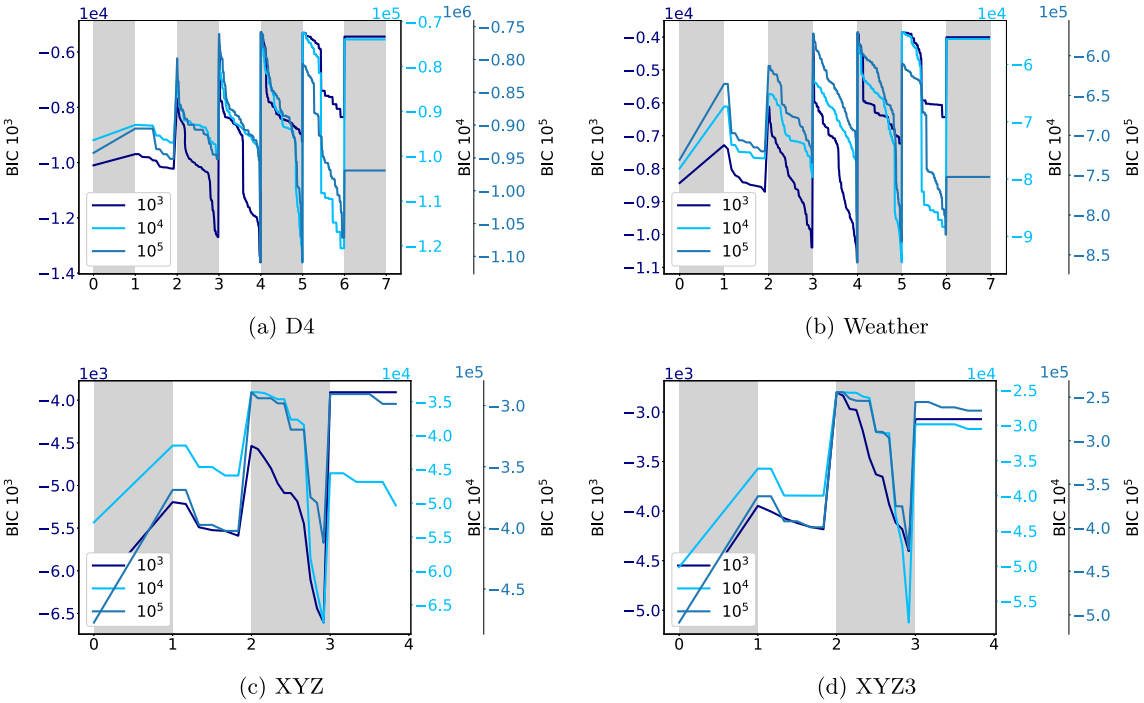
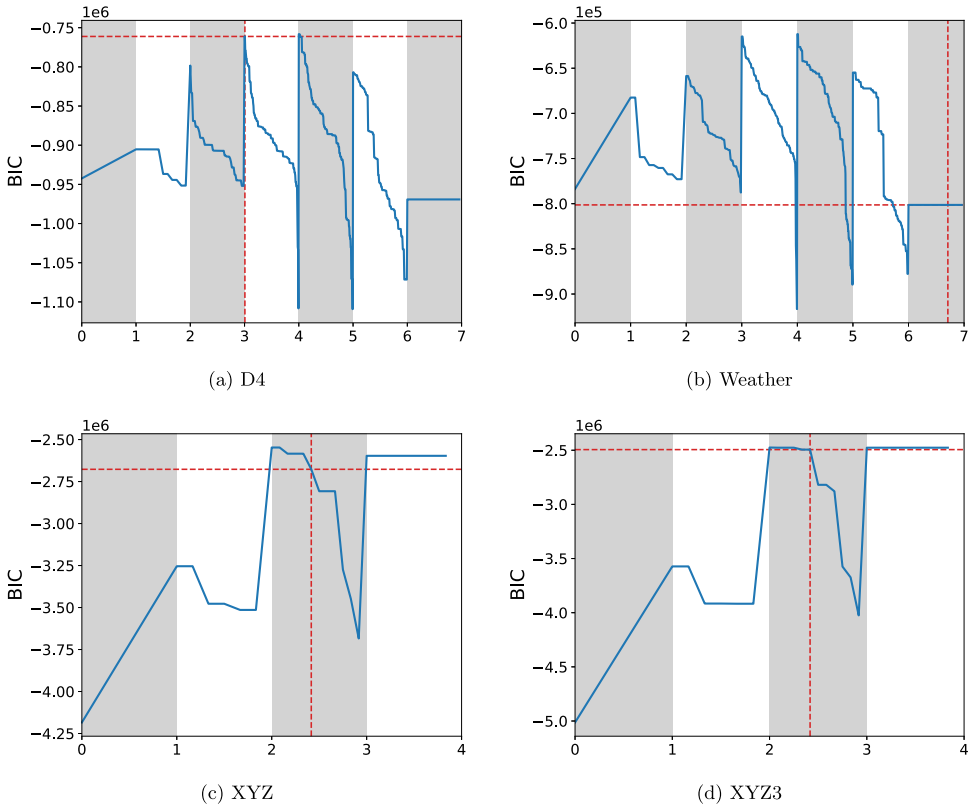
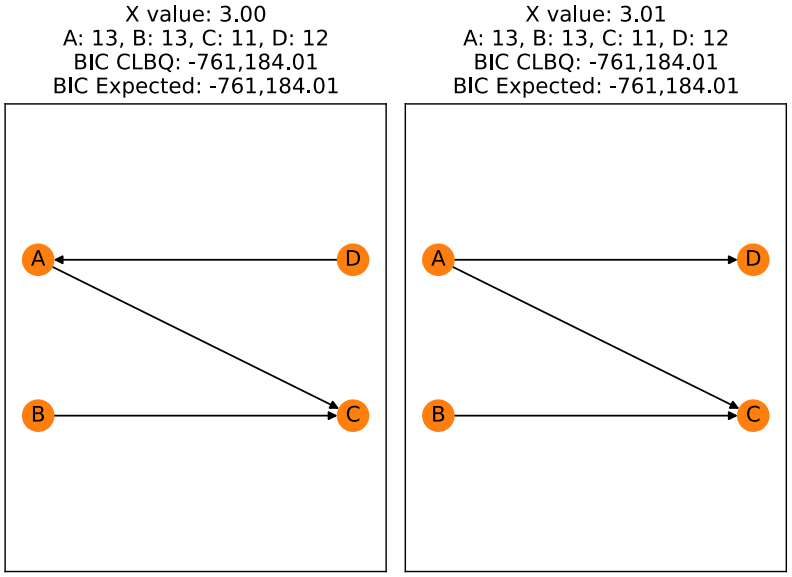


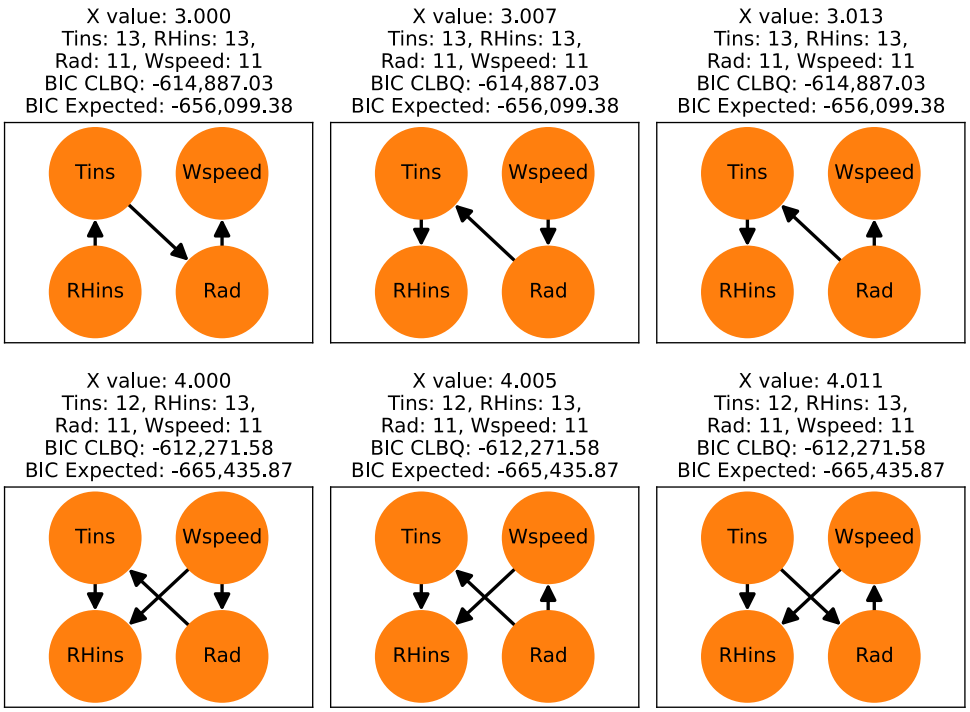
Fig. 15 Landscape plots for different dataset sizes of the BIC score of BNs using CLBQ for all datasets. They were devised considering 10^3, 10^4, and 10^5 samples. Each dataset size was plotted with different colours explained on the legend of each subfigure

Fig. 16 Landscape analysis for all datasets with the expected BN value and score depicted in dashed red lines. For the Weather dataset, the expected BN is the one found using the PC algorithm. The landscape analysis was performed using the whole dataset and the CLBQ was used to quantize the data





(a) D4

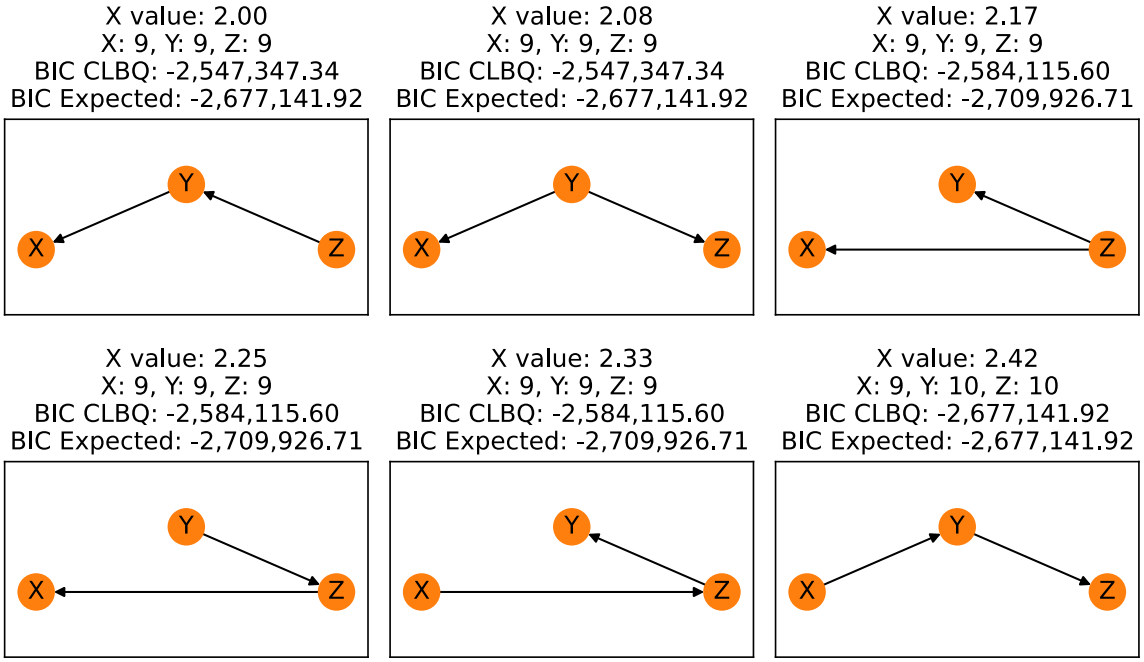


(b) Weather

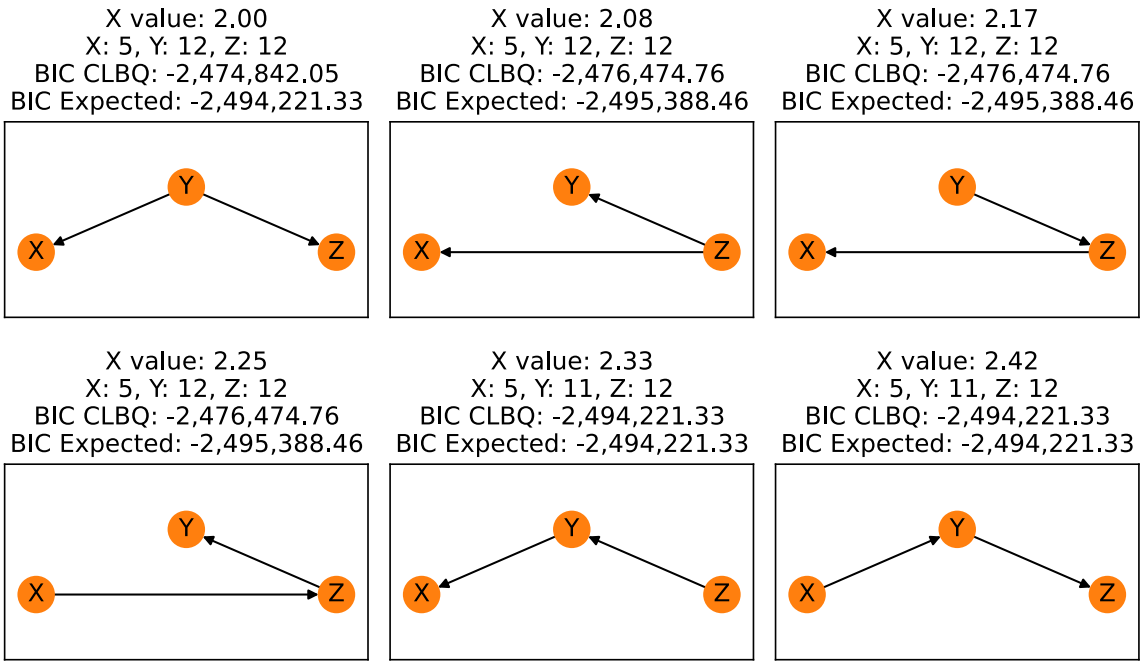
Fig. 17 Analysis of BNs with higher or equal scores than the expected BN. X is the numeric value on the x-axis assigned to the structure for landscape plotting. The quantization selected is shown beside the names of variables. BIC CLBQ is the score found by using CLBQ quantization. BIC Expected is the score found by using the quantization of expected structure

of the best structures was performed in order to understand which structures were better than the expected ones.

Considering the data shown in Fig. 18, it can be observed that, for dataset D4, the only BN with the same score as that of the expected structure is the one with a permutation of one of the edges of the expected structure. This evidences that using CLBQ in the search and score of BN structure learning is appropriate, as the expected structure could be selected if a good search algorithm was used. As for XYZ and XYZ3, it it found that, considering the same quantization, these structures have either the same or very close scores than those for the expected structure. Also, it is worth mentioning that quantization values are very similar, which was expected considering that structure limitations are minor and dataset size is large. This indicates that these slight variations in the quantization of variables affect the BIC score and mix the best structures. In addition to that, many of the structures having an equal to or higher score are permutations of edges of the



(a) XYZ



(b) XYZ3

Fig. 18 Analysis of BNs with higher or equal scores than the expected BN. X is the numeric value on the x-axis assigned to the structure for landscape plotting. The quantization selected is shown beside the names of variables. BIC CLBQ is the score found by using the CLBQ quantization. BIC Expected is the score using the quantization of the expected structure

Table 3 Mean and STD execution time of CLBQ for each structure on the landscape analysis

Dataset	Size = 10^3	Size = 10^4	Size = 10^5
D4	1.55 (0.66)	6.69 (1.87)	55.88 (9.98)
Weather	1.32 (0.66)	2.93 (0.86)	10.56 (1.92)
XYZ	1.29 (0.53)	4.02 (0.45)	29.73 (2.38)
XYZ3	1.17 (0.36)	3.70 (0.35)	28.48 (2.30)

The times are separated by dataset and dataset size. All times are shown in seconds. The results are shown as mean (std)

Table 4 Quantization and their respective MSE for the simulated continuous data test using datasets XYZ and XYZ3 showing the values for CLBQ and DD

XYZ	q_X	q_Y	q_Z	MSE_X	MSE_Y	MSE_Z
CLBQ	9	10	10	**0.095**	**0.788**	**3.140**
DD	4	4	4	0.502	5.118	20.403
XYZ3	q_X	q_Y	q_Z	MSE_X	MSE_Y	MSE_Z
CLBQ	5	11	12	**0.010**	**0.254**	**0.912**
DD	4	4	4	0.470	2.323	10.997

q are quantization values (the number of values the variable can assume) and MSE values are the mean squared error of the original and quantized data. The bold values are the best values for each of those measures

expected structure. This bodes well, as the direction of the edge on BN does not necessarily follow the causal relation [15]. Thus, although it may not be the best result, as that found for dataset D4, it still reveals that it is appropriate to use CLBQ for the search and score of BN structure learning. Finally, for the Weather dataset, all the best structures have edges marking dependencies included in the PC structure, but with varying orientations of edges. This indicates that these connections are the most important dependencies for describing data according to BIC scores. Considering the characteristic of BIC score and the best BNs found, these results are promising and indicate a promising possibility of using CLBQ for BN structure learning.

6 Conclusion

There are four most relevant results achieved herein: quantization and histogram, landscapes with different sample sizes, location of the expected structure, and the best structure analysis. As for quantization and histogram, it is found that CLBQ has not exactly selected the ideal quantization, however, it chose values that had good MSE and balanced MSE and BIC score. These results reveal the capacity of the CLBQ method to balance model quality, data fidelity and structure score. Results also indicate that, considering the tests performed and metrics used, CLBQ performed better than DD in all tests.

Regarding the landscapes with different sample sizes, it is observed that although there were small differences in behaviour, landscapes maintained the same general behaviour considering different dataset sizes, and the best structures followed the same trend. This evidences that a variation in dataset size, which affects CLBQ, has not significantly altered the search space of BN structures. Thus, indicating that CLBQ could be used in the search and score of BN of structure learning.

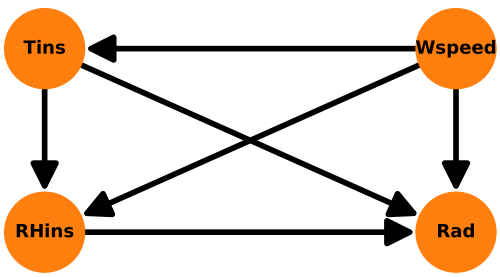


Fig. 19 Expected BN structure found for the Weather dataset using the PC algorithm, i.e. a constraint-based algorithm for BN structure learning

Regarding the CLBQ execution time, it was observed that it varies with the dataset size and between datasets. Its execution time is generally considerably small and can easily be applied to one structure. When considering the application of structural learning, the trade-off between the added execution time and the benefit of having the quantization done has to be considered to decide if using it would be beneficial.

Concerning the expected structure location, it is observed that the expected structure was not always the best, although the best structure analysis revealed that, when the exact expected structure was not the best, the best structure had dependencies in different directions, which is not a issue since BN does not ensure a proper direction of dependencies. Moreover, as for the real case data in which the expected structure was obtained using the PC algorithm, the best structures found when using CLBQ were subsets of dependencies of the expected structure that the BIC score considered as sufficient to represent data, which is beneficial as a smaller structure is generally easier to understand.

Considering all tests and analyses, CLBQ is an excellent method to quantize variables while using BN. In addition, it was observed that it can be used on the search and score of BN structure learning. The CLBQ limitations detected in the tests by analysing the CLBQ method are that it does not ensure an ideal quantization in addition to the fact that it is a method dependent on data volume for achieving good performance. When there is little data, CLBQ limits quantization to guarantee good model quality, although it can result in poor data fidelity.

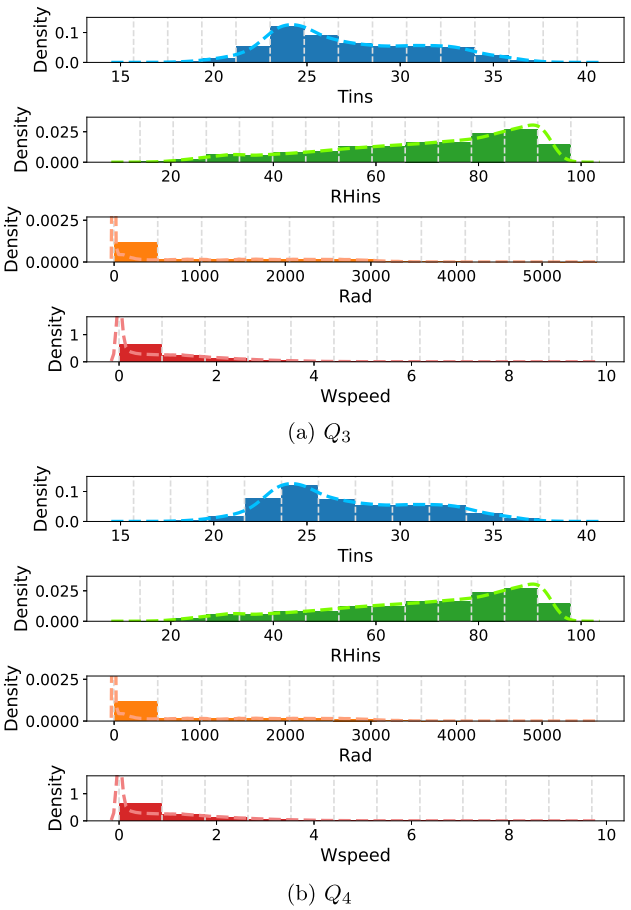
Table 5 Quantization and their respective MSE for the real-data test using the Weather dataset

	q_{Tins}	q_{RHins}	q_{Rad}	q_{Wspeed}	MSE_{Tins}	$\text{MSE}_{\text{RHins}}$	MSE_{Rad}	$\text{MSE}_{\text{Wspeed}}$
Q_{PC}	14	14	14	13	**0.241**	**3.261**	**26,498.219**	**0.075**
Q_3	13	13	11	11	0.281	3.557	43,433.104	0.105
Q_4	12	13	11	11	0.327	3.557	43,433.104	0.105
DD	4	6	11	8	4.759	32.588	73,400.493	0.118

Values obtained from the CLBQ for the best BNs, the BN found by PC and for DD are shown. q are quantization values (the number of values the variable can assume) and MSE values are the mean squared error of the original data and the quantized data. The bold values are the best values for each of those measures

7 Future Works

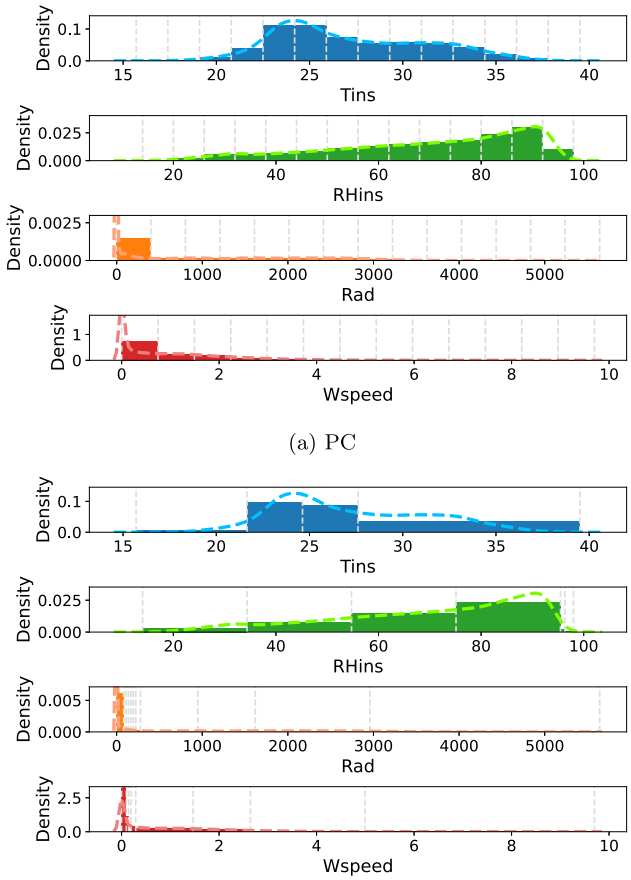
Additional studies should focus on using different types of quantization than the equal width one on the CLBQ. Also, using CLBQ with different scores ought to be analyzed, especially considering an BN structure score unaffected (or at least less affected) by the quantization of variables, as it could be a solution to guarantee that the expected structure is the best in its group. Another possibility would be to explore mechanisms to make CLBQ return the same quantization for the same dependencies, despite the variation in score. Furthermore, the burden of execution time of CLBQ when used for the search and score of BN structure learning should be futher analysed, since structure learning is already a time-demanding process and adding CLBQ to the process would increase its execution time. There are mechanisms commonly used to reduce computational time of algorithms



(a) Q_3

(b) Q_4

Fig. 20 Comparison between the distribution of variables and a histogram of variables of datasets with the quantization found by CLBQ considering the best structures found with 3 and 4 edges for the Weather dataset. The distributions of variables are shown in lighter-coloured dashed lines



(a) PC

(b) DD

Fig. 21 Comparison between the distribution of variables and a histogram of variables of datasets with the quantization found by CLBQ and DD considering the structure found by the PC algorithm for the Weather dataset. The distributions of variables are shown in lighter-coloured dashed lines

that also demand analysis, as their use would be beneficial for reducing CLBQ running time, such as parallelization and code optimization. Other options such as using CLBQ to quantize the data based on one structure and then using that quantization for the structural learning process and redoing the CLBQ quantization for the final structure should also be considered.

Acknowledgements This work was supported by Sao Paulo Research Foundation (FAPESP) [grant number #2021/09396-0, #2018/23139-8].

Author Contributions Conceptualization: Rafael Rodrigues Mendes Ribeiro, Cassio Polpo de Campos and Carlos Dias Maciel; Methodology: Rafael Rodrigues Mendes Ribeiro and Carlos Dias Maciel; Writing - original draft preparation: Rafael Rodrigues Mendes Ribeiro and Jordão Natal de Oliveira Junior; Writing - review and editing: Rafael Rodrigues Mendes Ribeiro, Jordão Natal de Oliveira Junior, Cassio Polpo de Campos and Carlos Dias Maciel; Supervision: Cassio Polpo de Campos and Carlos Dias Maciel.

Data Availability All datasets used and the code of CLBQ are available at https://www.doi.org/10.5281/zenodo.8368057.

Declarations

Ethical and informed consent for data used The data used did not involve human or animal participants. Thus, no ethical or informed consent was required.

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bertone E, Rousso BZ, Kufeji D (2023) A probabilistic decision support tool for prediction and management of rainfall-related poor water quality events for a drinking water treatment plant. J Environ Manag 332:117209. https://doi.org/10.1016/j.jenvman.2022.117209
2. Beuzen T, Marshall L, Splinter KD (2018) A comparison of methods for discretizing continuous variables in bayesian networks. Environ Model Softw 108:61–66. https://doi.org/10.1016/j.envsoft.2018.07.007
3. de Campos LM (2006) A scoring function for learning bayesian networks based on mutual information and conditional independence tests. J Mach Learn Res 2149–2187
4. Chen YC, Wheeler T, Kochenderfer M (2015) Learning discrete bayesian networks from continuous data. J Artif Intell Res 59. https://doi.org/10.1613/jair.5371
5. Ciunkiewicz P, Yanushkevich S, Roumeliotis M et al (2022) Improved design of bayesian networks for modelling toxicity risk in breast radiotherapy using dynamic discretization. In: 2022 International joint conference on neural networks (IJCNN), pp 01–08. https://doi.org/10.1109/IJCNN55064.2022.9892531
6. Fang H, Xu H, Yuan H et al (2017) Discretization of continuous variables in bayesian networks based on matrix decomposition. In: 2017 International conference on computing intelligence and information system (CIIS), pp 184–187. https://doi.org/10.1109/CIIS.2017.36
7. Fang W, Zhang W, Ma L et al (2023) An efficient bayesian network structure learning algorithm based on structural information. Swarm Evol Comput 76:101224. https://doi.org/10.1016/j.swevo.2022.101224
8. Friedman N, Goldszmidt M (1996) Discretizing continuous attributes while learning bayesian networks. In: Proceedings of the thirteenth international conference on international conference on machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML'96, p 157–165
9. Hao J, Yue K, Zhang B et al (2021) Transfer learning of bayesian network for measuring qos of virtual machines. Appl Intell 51:8641–8660. https://doi.org/10.1007/s10489-021-02362-x
10. Jackson-Blake LA, Clayer F, Haande S et al (2022) Seasonal forecasting of lake water quality and algal bloom risk using a continuous gaussian bayesian network. Hydrol Earth Syst Sci 26(12):3103–3124. https://doi.org/10.5194/hess-26-3103-2022
11. Jahan A, Edwards KL, Bahraminasab M (2016) 4 - multi-criteria decision-making for materials selection. In: Jahan A, Edwards KL, Bahraminasab M (eds) Multi-criteria decision analysis for supporting the selection of engineering materials in product design (Second Edition), second edition edn. Butterworth-Heinemann, p 63–80. https://doi.org/10.1016/B978-0-08-100536-1.00004-7
12. Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT press
13. Kozlov AV, Koller D (1997) Nonuniform dynamic discretization in hybrid networks. In: Proceedings of the thirteenth conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, UAI'97, pp 314–325
14. Lima MD, Nassar SM, Rodrigues PIR et al (2014) Heuristic discretization method for bayesian networks. J Comput Sci 5:869–878. https://doi.org/10.3844/jcssp.2014.869.878
15. Luo G, Zhao B, Du S (2019) Causal inference and bayesian network structure learning from nominal data. Appl Intell 49:253–264. https://doi.org/10.1007/s10489-018-1274-3
16. Mabrouk A, Gonzales C, Jabet-Chevalier K et al (2015) Multivariate cluster-based discretization for bayesian network structure learning. In: Beierle C, Dekhtyar A (eds) Scalable Uncertainty Management. Springer International Publishing, Cham, pp 155–169
17. Mayfield H, Bertone E, Sahin O et al (2017) Structurally aware discretisation for bayesian networks
18. Monti S, Cooper GF (1998) A multivariate discretization method for learning bayesian networks from mixed data. In: Proceedings of the fourteenth conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, UAI'98, pp 404–413
19. Neapolitan R (2003) Learning Bayesian. Networks. https://doi.org/10.1145/1327942.1327961

20. Nojavan AF, Qian SS, Stow CA (2017) Comparative analysis of discretization methods in bayesian networks. Environ Model Softw 87:64–71. https://doi.org/10.1016/j.envsoft.2016.10.007
21. Oppenheim AV, Schafer RW, Buck JR (1999) Discrete-Time Signal Processing, 2nd edn. Prentice-Hall Inc, USA
22. Rohmer J (2020) Uncertainties in conditional probability tables of discrete bayesian belief networks: A comprehensive review. Eng Appl Artif Intell 88:103384. https://doi.org/10.1016/j.engappai.2019.103384
23. Ropero R, Renooij S, van der Gaag L (2018) Discretizing environmental data for learning bayesian-network classifiers. Ecol Model 368:391–403. https://doi.org/10.1016/j.ecolmodel.2017.12.015. https://www.sciencedirect.com/science/article/pii/S0304380016308377
24. Ru X, Gao X, Wang Y et al (2023) Bayesian network parameter learning using constraint-based data extension method. Appl Intell 53:9958–9977. https://doi.org/10.1007/s10489-022-03941-2
25. Sari D, Rosadi D, Effendie A et al (2021) Discretization methods for bayesian networks in the case of the earthquake. Bull Electric Eng Inform 10(1):299–307. https://doi.org/10.11591/eei.v10i1.2007. https://beei.org/index.php/EEI/article/view/2007
26. Shiomoto K, Otoshi T, Murata M (2023) A novel network traffic prediction method based on a bayesian network model for establishing the relationship between traffic and population. Ann Telecommun 78:53–70. https://doi.org/10.1007/s12243-022-00940-9
27. Song D, Ek CH, Huebner K et al (2011) Multivariate discretization for bayesian network structure learning in robot grasping. In: 2011 IEEE International conference on robotics and automation, pp 1944–1950. https://doi.org/10.1109/ICRA.2011.5979666
28. Spirtes P, Glymour C, Scheines R (1993) Causation, Prediction, and Search 81. https://doi.org/10.1007/978-1-4612-2748-9
29. Talvitie T, Eggeling R, Koivisto M (2019) Learning bayesian networks with local structure, mixed variables, and exact algorithms. Int J Approx Reason 115:69–95. https://doi.org/10.1016/j.ijar.2019.09.002
30. Tian T, Kong F, Yang R et al (2023) A bayesian network model for prediction of low or failed fertilization in assisted reproductive technology based on a large clinical real-world data. Reprod Biol Endocrinol 21:8. https://doi.org/10.1186/s12958-023-01065-x
31. Toropova AV, Tulupyeva TV (2022) Discretization of a continuous frequency value in a model of socially significant behavior. In: 2022 XXV International conference on soft computing and measurements (SCM), pp 28–30. https://doi.org/10.1109/SCM55405.2022.9794892
32. Wilson SF, Nietvelt C, Taylor S et al (2022) Using bayesian networks to map winter habitat for mountain goats in coastal british columbia, canada. Frontiers Environ Sci 10. https://doi.org/10.3389/fenvs.2022.958596
33. Xu Q, Liu H, Song Z et al (2023) Dynamic risk assessment for underground gas storage facilities based on bayesian network. J Loss Prev Process Ind 82. https://doi.org/10.1016/j.jlp.2022.104961

Authors and Affiliations

Rafael Rodrigues Mendes Ribeiro[1] · Jordão Natal[1] · Cassio Polpo de Campos[2] · Carlos Dias Maciel[1]

Jordão Natal
jordao.oliveira@usp.br

Cassio Polpo de Campos
c.decampos@tue.nl

Carlos Dias Maciel
carlos.maciel@usp.br

[1] Department of Electrical and Computing Engineering, University of São Paulo, São Carlos, São Paulo, Brazil

[2] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, North Brabant, The Netherlands