2015 IEEE 27th International Conference on Tools with Artificial Intelligence

A new equilibrium criterion for learning the cardinality of latent variables

Hasna Njah
MIRACL laboratory
FSEGS, University of Sfax
Sfax, Tunisia
njah.hasna@gmail.com

Salma Jamoussi
MIRACL laboratory
ISIMS, University of Sfax
Sfax, Tunisia
salma.jamoussi@isimsf.rnu.tn

Walid Mahdi
College of Computers
and Information Technology
Taif University
Saudi Arabia
wmahdifr@gmail.com

Afif Masmoudi
Laboratory of Probability
and Statistics
FSS, University of Sfax
Sfax, Tunisia
Afif.Masmoudi@fss.rnu.tn

***Abstract*—Mining high dimensional data-sets extracted from real world problems is a challenging task due to the large features' space. The latent variables are used to reduce the dimensions of this space by representing highly dependent features. They simplify the creation of probabilistic models and they clarify the semantic of the inferred knowledge. Learning these variables for Bayesian network, as the most generic probabilistic model, is problematic. Actually, there is not a direct way that leads to finding their cardinalities. The precision of the inferred model is highly dependent on the accuracy of the latent variable's cardinality. Therefore, choosing a small value leads to a generalized model having a high rate of information loss. Moreover, a high cardinality tend to over-fit the data, to generate complex latent variables and to burden the parameter learning of the probabilistic model. In this paper, we propose a new criterion based on the mutual information and the log likelihood, called the equilibrium criterion. We mathematically and experimentally validate its efficiency for estimating the cardinality of the latent variable. We also demonstrate its performance in finding the hidden cause of a set of observed variables. The experimental analysis shows that our method succeeded in restoring the original cardinality of intentionally deleted variables in known networks.**

***Index Terms*—Latent variable, Bayesian network, mutual information, log likelihood**

I. Introduction

With the development of data acquisition systems, several real-world problems are represented by high amounts of features (attributes, descriptors or variables). Using the whole set of attributes for modeling, predicting and analyzing these high dimensional problems is challenging in terms of complexity and intelligibility [1]. Selecting the most representative features and/or reducing their dimensionality is not always feasible, semantically correct and axiomatic. Adding to that, some groups of attributes are correlated. The dependency between them is usually controlled by some unobserved set of features. This problem is concretized in the biological applications where the behavior of tens of thousands of genes is controlled by hidden factors (e.g. gene regulations and protein interactions). Hence, it is important to mine these factors in the raw data sets and correctly define them. From a technical viewpoint, it is primordial to identify the latent variables of the data-set used for modeling a given real world problem.

The latent variable (LV), also known as the hidden variable, is generally integrated in probabilistic graphical models such as the Bayesian Network (BN) [2] and the Hidden Markov Model (HMM) [3]. The determination of the LV when dealing with uncertainty, generally, and with BN, particularly, is beneficial from different perspectives [4]. Indeed, the basic advantage of integrating LVs is to highight the latent cause of a set of highly dependent observed variables. In this paper, we will be concentrating on exploring the hidden variables in BN as the most basic probabilistic model. We recall that the BN is a Directed Acyclic Graph (DAG) where each node is represented by a random variable. The conditional probability distribution of each variable depends on the probability distribution of its parents; it is indicated in the Conditional Probability Table (CPT). We choose the BN, in our analysis, because it is the most basic and generic probabilistic model. It is the simplest way to spot out the major challenges of learning the LV.

We present, through Fig.1 a classic example of the use of the LV. It consists in observing the symptoms fever (F), nausea (N), diarrhea (D) and headache (H) while the real cause, which is the virus (V) is unknown. The states of each symptom are conditioned by the states of the other symptoms. Therefore, a straightforward way to model these symptoms and the relations between them is to elaborate a fully (or partially full) connected BN of their corresponding variables (see Fig.1 (A)). However, the estimation of each CPT and the semantic analysis of the obtained BN are complicated. This problem is more and more accentuated when the set of the observed symptoms is voluminous (i.e. dozens or hundreds of variables). Therefore, the integration of their real cause, which is the virus, in the network solves this issue. Hence, the obtained network is no longer a complete graph but a hierarchical structure (see Fig.1(B)). The root node is the hidden cause, the virus, and the leaf nodes are the observed symptoms.

The above explained abstraction simplifies the complexity of the CPT learning step. Suppose that all the variables of BN are bi-valuated and the initial complete DAG contains n nodes. The specification of each CPT takes the complexity's order of 2^{n-1} considering that each node has at maximum $n-1$ parents. This complexity decreases significantly by adding the LV. In this case, each variable has only one parent which is the LV

1082-3409/15 $31.00 © 2015 IEEE
DOI 10.1109/ICTAI.2015.138

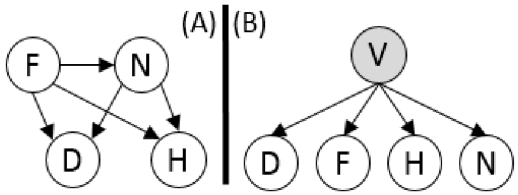

Fig. 1. The difference between a BN without the LV and a BN with the LV
The symptoms are represented by the set of nodes {D, F, H, N} and the hidden cause, which is the virus, is represented by the shaded variable V.

and the CPT learning takes the order of 2. Furthermore, the semantic interpretation of the dependencies between all the observed variables (OVs) is feasible and understandable even by non-computer scientists. Consequently, the use of the LV is demanded in wide applications of knowledge discovery such as the analysis of social networks and the study of genetic diseases.

We explained, so far, the importance of learning the LV. Yet, the process of finding the value of this particular variable is a challenging task and has not aroused a significant interest by researchers. An arbitrary choice of the cardinality of the LV does not lead to consistent model. In fact, a small value of cardinality leads to generalized models where the rate of information loss is high. Contrariwise, a LV with high cardinality leads to complex parameter learning, to data over-fitting and to biased inference. As it is stated by [5] in their discussion about the tradeoff between the complexity of the LVs and the latent structure, it is implacable that the LV reaches a considerable complexity when its cardinality is high.

In this paper, we address the problem of choosing the optimal cardinality of the LV while ensuring the low complexity of the variable and its suitability with respect to the model. In order to produce a fair study, we were based on several assumptions. First, we suppose that each set of data can be represented by only one LV. In that manner, we will be able to focus our study on the cardinality of the LV and not on the optimal number of LVs that represent the dataset. Second, we adopt a bottom-up strategy for creating the LV. Given that the prior knowledge about the LV is not provided, we established our method on extracting the cardinality of the LV from the examined set of the OVs. Besides these assumptions, our method is founded on the scoring functions that are usually used to quantify the quality of the BN regarding the training dataset.

This paper exhibits the related work to the problematic of finding the optimal cardinality of the LV in BN, in the second section. It gives, in the third section, the fundamental background of the BN and the BN's scoring functions as the starting point of our analysis. In the fourth section, a detailed explanation of the proposed method is provided. It is based on the semantic and the statistic analysis of the LV's integration in the BN. Subsequently, in the fifth section, an extensive experimental analysis is exhibited. Finally, the last section concludes the present work and gives its perspectives.

II. Related work

Integrating the LV in a BN raises three major challenges: finding the optimal number of LV, learning the structure of the BN with the presence of LV and finding the cardinality of the LVs. The first two problems were surmounted through wide variety of philosophies naming the cluster-based models such as the latent tree models [5] and the clique-based models such as the work presented in [6]. On the other hand, the problem of determining the cardinality of the LV is still under discussion. Obviously, the simplest solution consists on using binary LV such as in [7] and [6]. This choice is justified by the fact that the parameter learning algorithms are time consuming when the number of the LV is large and the respective cardinalities of the LVs are high. Hence, researchers opt for setting a minimized and constant cardinality of the LV. However, this assumption leads to overly generalized models and to information loss.

The most straightforward approach to overcome this problem is to explore the networks with all the possible cardinalities of the added LV. At each iteration, it is possible to apply any parameter estimation algorithm. This approach can be agglomerative [8] or divisive. The former starts by the maximal number of possible states and merges them in a greedy way. The latter starts by the minimal cardinality (i.e. the LV has two states) and increments the cardinality. In both methods, a BN scoring function is computed for each obtained network. The optimal values of this score corresponds to the model with the best cardinality. The used scoring functions are generally decomposable since their values are modified when the cardinality of is changed. Hence, they are suitable for reducing the execution time of evaluating the BN with LV.

Selecting the cardinality that optimizes the chosen score is evident, when applying these methods. However, the main problem of this approach is its exhaustiveness. For instance, if all the n variables of a given data-set are bi-valuated, then there are 2^{n-1} possible models. Therefore, finding the optimal model reaches a computational burden. This is explained by the fact that the dimension of the search space augments exponentially with the number of the OVs. To tackle this problem, [9] proposed a mathematical proof for a stopping criterion. The author applies a greedy search approach, starting by the minimal value of cardinality. At each iteration, this value is incremented so as to meet the optimal *regularity criterion* of a hierarchical latent class. Although this solution reduces the search space, it remains computationally demanding.

Advanced searching algorithms can be applied for decreasing the complexity of the search procedure. Zhang *et al.* used a hill climbing routine [10] and applied advanced greedy search algorithm [11]. Chen *et al.* experimentally proved that the grow restructure thin strategy reduces the complexity of the search space [12][13]. Optimization methods, such as the simulated annealing, can enhance their complexity while ensuring to escape the local optimum. Moreover, it is possible to adopt the variable clustering for determining the cardinality of the LV. For instance, [14] used the conceptual clustering in order to find the best cardinality.

It is stated in most of the studies of the state-of-the-art that in many cases the scoring functions of the BN are not suitable for identifying the best cardinality of the LV. Adding to that, these scores are not applicable when the OVs are not completely dependent. They might lead to puzzling values. Basically, the variability of these scores is not monotone all through the cardinalities, in one hand, and the optimized values of one score may correspond to mediocre values of another score, on the other hand. [9] improves the Bayesian Information Criterion (BIC) score so as it takes into account the complexity of the model when adding the LV. However, the author stated that the perfect value of a given score can correspond to the model with a LV having a high cardinality. Yet, adding the LV may lead to over-fitted model and to high complexity parameter learning procedure.

III. Background

Bayesian Network (BN) is a graphical probabilistic model. The graphical perspective of the BN is represented by a DAG where the nodes stand for the features. The probabilistic aspect of the BN is represented by the CPT of the random variable that corresponds to each node. The probability of each node's is computed using the probability of its parents' in the DAG. This probability is computed by applying the Bayes theorem (1) on the corresponding variables. Conventionally, the DAG is called the structure of the BN and the CPTs are called the parameters of the BN.

A. Bayesian Network's related concepts

The particularity of BN is that directed edges are supposed to represent causal relation between nodes. Two major concepts participate in the organization of the directionality affectation task for each node: *D-separation* and *Causal Markov Condition*.

Directional separation [15] called d-separation defines whereas two variables are independent in a BN or not. A formal definition of d-separation is presented as follows: Let T be an undirected path (i.e. trail) from X to Y. T is d-separated by a set of nodes S if and only if at least one of these conditions holds:

- T contains a chain ($X \rightarrow Z \rightarrow Y$ or $X \leftarrow Z \leftarrow Y$) such that the middle node Z is in S.
- T contains a fork ($X \leftarrow Z \rightarrow Y$) such that the middle node Z is in S.
- T contains a collider (a.k.a V-structure) ($X \rightarrow Z \leftarrow Y$) such that the middle node Z and its descendants are not in S.

It is substantial to affirm that all edges are placed by respecting *Causal Markov Condition* [16]. It is defined as the conditional independence of a node from its non-descendant given its parents. It is formally represented as: $X \perp\!\!\!\perp Y | Parents(X)$ where $X \in V$ (the nodes' set) and $Y = V - (Descendant(X) \cup Parents(X))$.

It happens that many DAGs have the same conditional dependence [17]. This assertion leads to define *Markov's Class of Equivalence* which is based on d-separation and causal Markov condition. For instance, the Markov's class of equivalence of $X \perp\!\!\!\perp Y | Z$ could be graphically represented by three possible graphs which are $X \rightarrow Z \rightarrow Y$, $X \leftarrow Z \leftarrow Y$ and $X \leftarrow Z \rightarrow Y$. Hence, these three equivalent structures represent the same probabilistic distributions:

$$\begin{aligned} P(X,Y,Z) &= P(X)P(Z|X)P(Y|Z) \\ &= P(X|Z)P(Z)P(Y|Z) \\ &= P(X|Z)P(Z|Y)P(Y) \end{aligned}$$

The proof of this assertion is ensured by the application of *Bayes Theorem* where A and B are mutually exclusive and exhaustive events such as P(B) $\neq$ 0.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

The main advantage of Markov's Class of Equivalence is that same BN skeletons could present different BNs which allow comparison between graphs and detection of similarity between them.

B. Bayesian Network's scoring

The BN is established through a structure learning algorithm, which finds the DAG of the BN, and a parameter learning algorithm, which estimates the CPT of each node. There is a panoply of structure and parameter learning algorithms. Hence, it is important to quantify the quality of the obtained BN given a training data-set. The most intuitive way is to use the BN's scoring functions. Two major BN's score families exist: the *information theory based scores* and the *Bayesian scores*.

The first family of scores is established through the adequacy between the data describing the BN and the BN model itself. It penalizes complex models and promotes simple ones [4]. These scores are based on the Logarithm of Likelihood (LL). The LL score tends to accentuate complete network. Other LL-based scores can be used such as Akaike Information Criterion (AIC) [18], Bayesian Information Criterion (BIC) [19] and the Mutual Information Tests (MIT) [20]. These scores penalize the LL by the complexity of the BN through the application of *Occam's razor* principle: *"Given two equally predictive theories, choose the simpler"*.

Bayesian scoring functions compute the posterior probability distribution of the parameters in a BN based on their prior probability distribution with regards to the data. The better is the BN, the higher is the posterior probability. Generally, Bayesian scoring functions are based on the Bayesian Dirichlet (BD) score [21]. The most used ones are the K2 [22], BDe [21] and BDeu [23] scores.

An important property of the BN's scoring functions is considering the graph equivalence [24] which was proven by the theorem of the *Markov's class of equivalence* [17]. This theorem indicates that two DAGs are equivalent if they have the same skeletons and the same v-structures. It was proven that two equivalent BN's structures have approximately the

same score [25]. Hence, the information theory based scores and the Bayesian scores are equivalent, except for the BD score.

IV. Estimating the cardinality of the LV

Our analysis is inspired by the families of BN's structure learning algorithms which are the score-based methods and dependency-based methods. Both of them tend to find the optimal BN's structure that represents a given data-set. The former assigns a score to the candidate structures and chooses the structure that optimizes this score. The latter looks for the conditional dependencies between the features and return the most relevant structure.

A. Score-based evaluation

Despite the blurriness and the ineffectiveness of using the BN's scores for finding the optimal cardinality of the LV, these metrics are maintained and utilized in most cases. The intuitive use of these scores is to study their evolution among the possible cardinalities. Since the structures of the obtained networks are the same, these score are measuring in reality the adequacy between the training data and the obtained structure. Generally, the information theory based scores (i.e. the BIC and the AIC scores) are computed based on the Log of Likelihood (LL) of the given BN. For a given BN where X is the set of its nodes, the expression of the LL is explained in the equation (2) where N is the number of instances in the dataset, r_i is the number of the possible values of a node X_i, q_i is the number of the possible combinations of the parents of the node X_i and N_{ijk} (respectively N_{ij}) is the number of instances of the ijk (respectively ij) configuration.

$$LL(BN) = \sum_{i=1}^{N} \sum_{k=1}^{r_i} \sum_{j=1}^{q_i} N_{ijk} \log(\frac{N_{ijk}}{N_{ij}}) \tag{2}$$

The calculation of the LL-based scores follows the equation (3) where $LL(BN)$ is the LL of the network and $Dim(BN)$ is the BN's complexity measure and is called the dimension of the BN. In fact, as we mentioned in the previous sections, the LL tends to promote the complete networks. Hence, the LL-based scores tend to quantify the optimal complexity-fitness trade-off. This is ensured by the term $f(N)Dim(BN)$ which penalizes the score $S(BN)$ so as to take into account the complexity of the BN's structure. For instance, for the BIC score, $f(N) = 0.5\log(N)$ and for the AIC score, $f(N) = 1$.

$$S(BN) = LL(BN) - f(N)Dim(BN) \tag{3}$$

It is noteworthy that all the obtained structures, in our case, are similar because all of them follow the abstraction of Fig.2. Hence, the term $f(N)Dim(BN)$ has not a significance in our case. Therefore, we focus our study on the LL measure.

Regarding the optimal cardinality estimation problem, one possible solution is to measure the LL of the obtained network without introducing the LV, BN_0, and to measure the LL of the candidate networks by integrating the LV, BN_k, with different cardinalities, k. The best cardinality, $k*$, corresponds to the network whose LL is approximately equal to the LL of BN_0 which means $LL(BN_{k*}) \simeq LL(BN_0)$. Although this idea seems promising and exploits the equivalence property of the BN, the obtained N_{k*} is not really the optimal network. In fact, this demarche is absurd since the equivalence of scores is not useful unless BN_0 and BN_k have the same skeleton, consequently, the same set of variables. This is not the case since (i) the LV is not included in BN_0 and (ii) the LVs of each candidate BN_k are not the same variables (the cardinalities are different). This particular ascertainment explains why the evolution of the LL, particularly, and the BN's scoring functions, generally, all over the possible cardinalities has a random aspect.

B. Dependence-based evaluation

The conditional dependence relations between variables reflect a consistent semantic measure for quantifying the quality of a given BN. These relations can be measured by the conditional Mutual Information (MI) [26] computed between each pair of nodes (variables). The overall MI between all the variables in the BN is computed as in equation (4).

$$MI(BN) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{r_i} \sum_{j=1}^{q_i} N_{ijk} \log(\frac{NN_{ijk}}{N_{ik}N_{ij}}) \tag{4}$$

Applied to the optimal cardinality estimation problem, the study of the MI's evolution over the possible cardinalities, k, reveals that this measure increases significantly for small values of k and tends to stabilize for high values of k. We consider this stabilization as the beginning of data over-fitting. The obtained network, in this case, tends to learn by heart all the possible combinations of the training data-set D. Hence, the question is for which cardinality, $k*$, the LV represents the most the data without over-fitting all the instances? The simplest answer is to keep the cardinality k where the difference between obtained MI_k and MI_{k+1} is smaller than a fixed ε. This is explained by the fact that the rarely observed instances will be represented by the LV which means that the LV will tend to cover all the possible combinations of the OV's values. Therefore, the conditional probability of observing these instances given the value of the LV will tend to 0. Unfortunately, this cardinality is usually high and perplexes the subsequent use of the LV.

Similarly to our discussion in the previous subsection, a possible way to find the optimal cardinality $k*$ is to approximate the MI_k of the network having a LV with cardinality k and the MI_0 of the network that does not include the LV. However, this reasoning is not coherent enough to find the best cardinality $k*$ since the measures of the MI tend to have high values when the network has more nodes. As it was explained in [20], the more is the number of parents of a given node, the higher is the MI.

V. Proposed method

Since the use of traditional BN scoring functions lead to puzzling results, we propose a new consistent metric called

the *Equilibrium Criterion* (EC). It finds the best compromise that guides to the optimal cardinality. Therefore, our main goal is to approach, as much as possible, to the optimal value of the LV's cardinality.

A. Framework

We use the bottom-up strategy to infer the LV. So, the structure of the obtained BN, noted G, is a two-levelled tree-like structure where the root is the new LV and the leaves are the set of the OVs. The directionality of edges is oriented from the LV to the OV (see Fig.2).

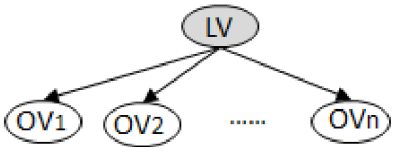

Fig. 2. The main graphical abstraction

The root of the BN (shaded in the gray color) is the latent variable (LV) and the leaves represent the set of the n observed variables (OVs)

We adopt the generic algorithm for learning the LV that represents a given set of variables. It consists on enumerating all the possible cardinalities, k, of the LV and applying the following three steps for each k:

1) learn the LV that has k discrete values,
2) create the corresponding BN structure, noted S_k, by following the abstraction of Fig.2 and
3) learn the parameters of the obtained S_k.

This process starts by the minimal cardinality ($k = 2$). It is iterative until reaching the optimal value of the EC. The best cardinality, $k*$, corresponds to the structure, S_{k*}, that satisfies the EC. Note that learning the LV (step 1) is typically ensured by the Expectation Maximization (EM) algorithm [27]. Other algorithms can be applied such as the LCM-based EM[28] and the spectral methods [29].

B. The equilibrium criterion

From a semantic point of view, highly dependent OVs tend to have a common cause. They present a significant conditional dependency between them. This dependency can be quantified by high values of MI between the set of the OV. Likewise, the conditional dependency between these variables is visualized through a highly connected BN's structure. Therefore, the determination and the introduction of the LV that causes this phenomenon is important. Mathematically speaking, the higher is the overall conditional MI between the nodes in the obtained structure, the more informative is the LV (5).

$$k* = \arg\max_{k\geq 2} MI(BN_k) \tag{5}$$

From a statistic point of view, the cardinality of the latent variable has a direct impact on the obtained structure with respect to the data. The idea here is to find the cardinality where the probability distribution of the OVs is quite similar to their probability distribution when conditioned by the LV. In other words, we look for the cardinality where the insertion of the LV will not affect the prior distribution of the OV. Mathematically speaking, the optimal cardinality $k*$ corresponds to the network where the LL is optimized.

$$k* = \arg\max_{k\geq 2} LL(BN_k) \tag{6}$$

To summarize, the "best" cardinality can be determined through semantic and statistic viewpoints. The MI by itself does not produce an exact value of the $k*$ since the choice of the ε is not supervised. Adding to that, the LL do not converge to an optimal value since the evolution of this score as a function of the cardinality is not monotone. Therefore, searching the best LL score can easily fall in local optimal.

Our idea is inspired by the work of [30] where a new objective function was derived from the conditional likelihood in order to propose a new feature selection method. Our contribution consists on finding a compromise between the MI and the LL score of the variables in the BN. In other words, the optimal cardinality is reached when the dependency between variables is ensured without over-fitting the data, from one hand, and the score of the obtained network with respect of data is optimized, on the other hand. This compromise leads us to establish the Equilibrium Criterion. It corresponds to the cardinality, $k*$, where LL_{k*} is approximated to MI_{k*}. Hence, putting (5) and (6) together, the problem of finding the optimal cardinality of the LV, for a structure S_k, is formally explained in equation (7).

$$k* = \arg\min_{k\geq 2} |MI(S_k) - LL(S_k)| \tag{7}$$

VI. Results and Discussion

We propose a two-stepped framework for validating the performance of using the EC for estimating the optimal cardinality of the LV. We start by verifying the correctness of our reasoning. We describe our verification process and we apply it on a benchmark data-set in order to visualize the evolution of the used functions. We wrap up this section by applying our method on different known BNs and by comparing our output with the output of some competitive algorithms.

A. Verification

1) Process description: The development of our experimental analysis proceeds as follows.

We select the most connected node in a given BN. We consider that this node is the LV and that the descendants of this node are the set of the OVs.

We run the proposed algorithm on the set of the OVs and we compute the EC at each iteration. The algorithm terminates at the cardinality $k*$ that corresponds to the minimized value of $|LL(BN_{k*}) - MI(BN_{k*})|$.

Our method succeeds when the obtained $k*$ coincides with the real cardinality of the chosen node.

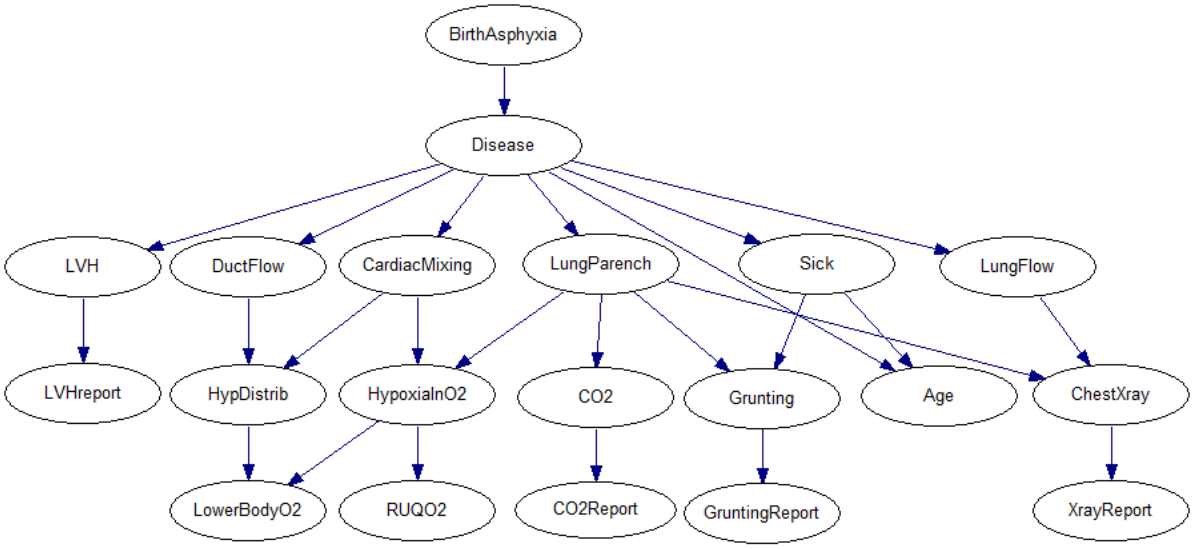

Fig. 3. The Child network

2) Empirical validation: In order to validate our choice of the EC, we propose to examine the "Child" network [31] as a known BN (see Fig.3). This network models the effect of the disease "Blue-baby syndrome" on the values of routine biological tests.

As it is remarkable in Fig.3, the respective values of the nodes LVH, DuctFlow, CardiacMixing, LungParench, LungFlow, Sick and Age are conditioned by the value of the node Disease. Hence, this latter is considered as the direct cause of this set of nodes.

The objective of this experiment is to test the reliability of our proposed EC. We remove the cardinality of the node Disease, which is equals to six, and we try to recover it by using our method.

In analogy to the abstraction of Fig.2, our LV is the node Disease and the OVs are the nodes LVH, DuctFlow, CardiacMixing, LungParench, LungFlow, Sick and Age.

We tested three assumptions:

- The BN's scoring assessment. In this case, $k*$ corresponds to the network where $LL(BN_{k*})$ is maximized. We visualized the $LL(BN_k)$ in Fig.4. The LL function evolves randomly over the cardinalities. The obtained curve presents local maxima at $k = 3$ and $k = 9$ and could present higher values beyond $k = 14$. Hence, the LL score itself cannot reveal $k*$.

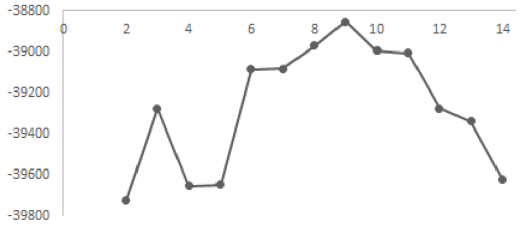

Fig. 4. The evolution of $-LL(BN_k)$ over the possible cardinalities k

- The MI conservation between variables. In this case, $k*$ corresponds to the network where $MI(BN_{k*})$ is maximized. We visualize the evolution of the $MI(BN_k)$ over the possible cardinalities k in Fig.5. We remark that this function is increasing, in a monotonous way, when incrementing the cardinality of the LV. Theoretically, this function keeps increasing until reaching a certain stabilization. Whether this is achieved or not, the maintained cardinality will be high and the obtained LV will not be useful for further manipulations.

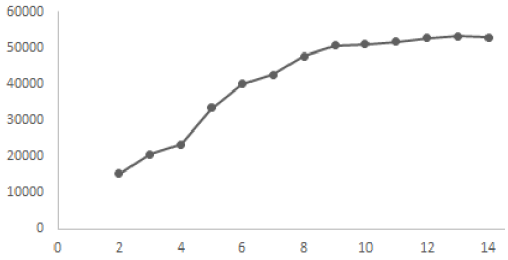

Fig. 5. The evolution of the $MI(BN_k)$ over the possible cardinalities k

- The equilibrium criterion. In this case, $k*$ corresponds to the network where $MI(BN_{k*}) - LL(BN_{k*})$ is minimized. We visualize in Fig.6 the evolution of the -LL and the MI as functions of the LV's cardinality. The obtained curves intersect within the value of 6 which is effectively the real cardinality of the node Disease. Consequently, the compromise between the semantic aspect of the LV (expressed through the MI) and the statistical aspect of the LV (expressed through the LL score) leads to the appropriate estimation of the LV's cardinality. Hence, the chosen EC is adequate for determining the latent cause of highly correlated set of observed variables.

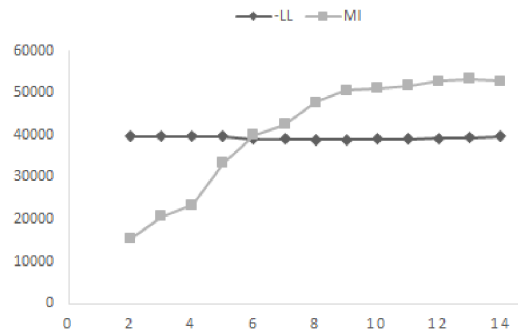

Fig. 6. The evolution of $-LL(BN_k)$ and $MI(BN_k)$ over the possible cardinalities k

B. Comparison

We perform the same steps of the previous experimental analysis (see section 5.1.1) on the whole set of significant nodes in four state of the art BNs[1] (Hepar2, Insurance, Child and Hailfinder). We hide, for each network, the most connected nodes and we try to restore their original cardinality. We compare our method with three of the state-of-the-art algorithms which are mainly based on the abstraction that we used for creating the LV in this paper. These algorithms[2] are:

- The Bridged Island (BI) algorithm proposed by [32]. It uses the uni-dimensionality test in order to quantify the dependencies among the set of variables that will be represented by a LV.
- LTM-EAST proposed by [13] and based on the EAST algorithm proposed by [10]. It uses the Hill-climbing routine in order to find the optimal cardinality of the LV. The algorithm adds or removes the states of the LV so as to get a simple model (i.e. structure and parameter). The optimal model is determined by a modified version of the BIC score.
- LCM (Latent Class Model) which is an adapted version of the LCMB-LTM algorithm proposed by [28]. It uses the MI in order to group the most dependent couple of variables so as to create the optimal structure.

In order to conduct a fair analysis, we compute the accuracy of each method for the four chosen data-sets. The accuracy is the number of times an algorithm correctly predicted the cardinalities divided by the number of the chosen set of nodes. We remark through Table 1 that our method (EC) succeeded in finding the real cardinalities by an overall accuracy on the four chosen databases of 0.49. This is caused by degree of dependence between the nodes. The more they are conditionally dependent, the more probable that our method finds the real cardinality. Furthermore, this result is promising when compared with the three other algorithms (BI: 0.19, LTM-EAST: 0.15 and LCM: 0.4). In fact, it was remarkable that when the real cardinality of the variable is high, these algorithms tend to use more than one LV for representing the set of the OVs. Adding to that, we remark that our algorithm (besides the LCM algorithm) are adequate for finding the cardinalities of large sets of data. Contrariwise, the BI and the LTM-EAST failed in treating moderately large datasets. They also do not tend to converge to a final cardinality when the LV supports high number of states.

[1]These networks are found in the BN repository of the bnlearn R package: http://www.bnlearn.com/bnrepository/

[2]An implementation of these algorithms is found in the Lantern software: http://www.cse.ust.hk/faculty/lzhang/ltm/index.htm

TABLE I
THE ACCURACIES OF BI, LTM-EAST, LCM AND OUR EC-OC ALGORITHM

Bayesian network	BI	LTM-EAST	LCM	EC
Hepar2	0.22	0.11	0.37	0.46
Insurance	0.19	0.19	0.49	0.54
Child	0.25	0.20	0.36	0.42
Hailfinder	0.11	0.12	0.39	0.55
Average accuracy	0.19	0.15	0.40	0.49

VII. CONCLUSION AND PERSPECTIVES

In this paper, we treated the problem of finding the cardinality of the LV. In fact, exhaustively testing all the possible cardinalities of a LV is a laborious task especially when the number of the OV is high and/or their cardinalities are multiple. Moreover, high LV's cardinality leads to over-fitted variable which learns by heart all the possible combinations of the set of the OVs. Such situation corresponds to a model where the prediction ratio is high for the training dataset and very likely low for test dataset. Furthermore, the high LV's cardinality mitigate the parameter learning (i.e. CPT computation) step when using the LV in other contexts. To tackle these problems, we proposed a study on the most basic BN's evaluation score which is the Log Likelihood. We also use the Mutual Information for quantifying the degree of dependency between the LV and the set of OVs. We used these scores in order to provide a new equilibrium criterion. We also provided an exhaustive experimental analysis in order to assess the quality of the proposed criterion. Our analysis is oriented to the main objective of the introduction of the LV which is the determination of the latent cause of a given phenomenon. As future work, we aim at applying our method on real world problems. We opt to use our algorithm in data clustering by automatically determining of the number of clusters in a given dataset.

REFERENCES

[1] H. Njah and S. Jamoussi, "Weighted ensemble learning of bayesian network for gene regulatory networks," *Neurocomputing*, vol. 150, pp. 404–416, 2015.

[2] N. Friedman, I. Nachman, and D. Peér, "Learning bayesian network structure from massive datasets: the sparse candidate algorithm," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 206–215.

[3] S. R. Eddy, "Hidden markov models," *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996.

[4] B. Grossmann-Hutter, A. Jameson, and F. Wittig, "Learning bayesian networks with hidden variables for user modeling," in *IJCAI-99 Workshop" Learning About Users*, 1999, pp. 29–34.

[5] R. Mourad, C. Sinoquet, N. L. Zhang, T. Liu, P. Leray *et al.*, "A survey on latent tree models and applications." *J. Artif. Intell. Res.(JAIR)*, vol. 47, pp. 157–203, 2013.

[6] C. He, K. Yue, H. Wu, and W. Liu, "Structure learning of bayesian network with latent variables by weight-induced refinement," in *Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning*. ACM, 2014, pp. 37–44.

[7] X. Huang, L. A. Stefanski, and M. Davidian, "Latent-model robustness in structural measurement error models," *Biometrika*, vol. 93, no. 1, pp. 53–64, 2006.

[8] G. Elidan and N. Friedman, "Learning the dimensionality of hidden variables," in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 144–151.

[9] N. L. Zhang, "Hierarchical latent class models for cluster analysis," *The Journal of Machine Learning Research*, vol. 5, pp. 697–723, 2004.

[10] N. L. Zhang and T. Kocka, "Effective dimensions of hierarchical latent class models," *J. Artif. Intell. Res. (JAIR)*, vol. 21, pp. 1–17, 2004.

[11] ——, "Efficient learning of hierarchical latent class models," in *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*. IEEE, 2004, pp. 585–593.

[12] T. Chen, N. L. Zhang, and Y. Wang, "Efficient model evaluation in the search-based approach to latent structure discovery," in *4th European Workshop on Probabilistic Graphical Models*. Citeseer, 2008, pp. 57–64.

[13] T. Chen, N. L. Zhang, T. Liu, K. M. Poon, and Y. Wang, "Model-based multidimensional clustering of categorical data," *Artificial Intelligence*, vol. 176, no. 1, pp. 2246–2269, 2012.

[14] D. Connolly, "Constructing hidden variables in baysesian networks via conceptual clustering," in *Proceedings of the Tenth International Conference on Machine Learning*, 2014, pp. 65–72.

[15] W. C. Gogel and R. D. Sturm, "Directional separation and the size cue to distance," *Psychologische Forschung*, vol. 35, no. 1, pp. 57–80, 1971.

[16] D. M. Hausman and J. Woodward, "Independence, invariance and the causal markov condition," *The British journal for the philosophy of science*, vol. 50, no. 4, pp. 521–583, 1999.

[17] J. Pearl, T. Verma *et al.*, *A theory of inferred causation*. Morgan Kaufmann San Mateo, CA, 1991.

[18] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[19] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[20] L. M. De Campos, "A scoring function for learning bayesian networks based on mutual information and conditional independence tests," *The Journal of Machine Learning Research*, vol. 7, pp. 2149–2187, 2006.

[21] D. Heckerman, "Bayesian networks for knowledge discovery." *Advances in knowledge discovery and data mining*, vol. 11, pp. 273–305, 1996.

[22] G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine learning*, vol. 9, no. 4, pp. 309–347, 1992.

[23] W. Buntine, "Theory refinement on bayesian networks," in *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1991, pp. 52–60.

[24] D. M. Chickering, "A transformational characterization of equivalent bayesian network structures," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 87–98.

[25] S. Acid and L. M. de Campos, "Searching for bayesian network structures in the space of restricted acyclic partially directed graphs," *Journal of Artificial Intelligence Research*, pp. 445–490, 2003.

[26] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[27] T. K. Moon, "The expectation-maximization algorithm," *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[28] S. Harmeling and C. K. Williams, "Greedy learning of binary latent trees," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 6, pp. 1087–1097, 2011.

[29] L. Song, E. P. Xing, and A. P. Parikh, "A spectral algorithm for latent tree graphical models," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1065–1072.

[30] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 27–66, 2012.

[31] D. J. Spiegelhalter and R. G. Cowell, "Learning in probabilistic expert systems," *Bayesian statistics*, vol. 4, pp. 447–465, 1992.

[32] T. Liu, N. Zhang, A. Liu, and L. Poon, "A novel ltm-based method for multidimensional clustering," in *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models (PGM-12)*, 2012.