ROMA TRE
UNIVERSITÀ DEGLI STUDI

COLLANA DEL
DIPARTIMENTO DI ECONOMIA

IMPROVING DISCRETIZATION EXPLOITING DEPENDENCE STRUCTURE

Daniela Marella - Mauro Mezzini - Paola Vicard

ISSN 2279-6916 Working papers
(Dipartimento di Economia Università degli studi Roma Tre) (online)

Working Paper n° 199, 2015

I Working Papers del Dipartimento di Economia svolgono la funzione di divulgare tempestivamente, in forma definitiva o provvisoria, i risultati di ricerche scientifiche originali. La loro pubblicazione è soggetta all'approvazione del Comitato Scientifico.

Per ciascuna pubblicazione vengono soddisfatti gli obblighi previsti dall'art. l del D.L.L. 31.8.1945, n. 660 e successive modifiche.

Copie della presente pubblicazione possono essere richieste alla Redazione.

REDAZIONE:
Dipartimento di Economia
Università degli Studi Roma Tre
Via Silvio D'Amico, 77 - 00145 Roma
Tel. 0039-06-57335655 fax 0039-06-57335771
E-mail: dip_eco@uniroma3.it
http://dipeco.uniroma3.it

ROMA TRE
UNIVERSITÀ DEGLI STUDI

DIPARTIMENTO DI ECONOMIA

IMPROVING DISCRETIZATION EXPLOITING DEPENDENCE STRUCTURE

Daniela Marella - Mauro Mezzini - Paola Vicard

Improving discretization exploiting dependence structure

Daniela Marella and Mauro Mezzini
Dipartimento di Scienze dell'Educazione
Università Roma Tre, Italy

Paola Vicard
Dipartimento di Economia
Università Roma Tre, Italy

Abstract

Bayesian networks are multivariate statistical models using a directed acyclic graph to represent statistical dependencies among variables. When dealing with Bayesian Networks it is common to assume that all the variables are discrete. This is not often the case in many real contexts where also continuous variables are observed. A common solution consists in discretizing the continuous variables. In this paper we propose a discretization algorithm based on the Kullback-Leibler divergence measure. Formally, we deal with the problem of discretizing a continuous variable Y conditionally on its parents. We show that such a problem is polynomially solvable. A simulation study is finally performed.

JEL Classification: C100, C180.

Keywords: Discretization, Kullback-Leibler divergence measure, Bayesian Networks.

1 Introduction

Bayesian networks (BN) are multivariate statistical models satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG) (Pearl (1988); Cowell et al. (2007); Neapolitan (2003)). The advantage of using BNs to represent joint probability distributions is that they graphically show causal and conditional independence relations between random variables making easier for decision makers to evaluate and use the model. Augmented with a table of parameters representing marginal and conditional probabilities, BNs are capable of representing the probabilities over any discrete sample space: the probability of any sample point in that space can be computed from the probabilities in the BN.

Building BNs by hand can be a difficult and time-consuming task. The association structure (i.e. presence/absence of direct edges in the DAG and conditional probability distributions) can be known in advance by subject matter knowledge or have to be learned from a database of statistical units. In the second case, it is crucial to appropriately estimate the dependence model. To this aim, there are two main approaches (Cooper and Herskovits (1992); Lam and Bacchus (1994); Bouckaert (1994); Neapolitan (2003); Kjræulff and Madsen (2012)): constraint based learning, where the structure is inferred by a sequence of independence and conditional independence tests (for instance the PC algorithm, Spirtes et al. (2000)); score-plus-search algorithms (also known as Bayesian approach) that is an optimization based search requiring a scoring function or metric and a search strategy (Cooper and Herskovits (1992)).

As far as score-plus-search algorithms are concerned, there are a variety of possible score functions. Among them, the likelihood function which is then penalized to avoid overfitting giving rise to AIC or BIC depending on the penalization used.

Other score functions can be derived using the *minimum description length* (MDL) approach that is based on the MDL principle. In this approach a BN is selected if the sum of the description length of the BN and the encoding length of the database, given the BN, is minimized (Bouckaert (1993)). Clearly, in both constraint-based and score-plus-search approaches,

the primary and unique goal is to learn a BN such that the probability distribution modeled by it is an accurate approximation of the true probability distribution. On the other hand, in order to guarantee an adequate efficiency of the probability table estimators, a desirable property of the learned BN is that the network structure is as simple as possible, i.e. the tables are as small as possible.

When dealing with the problem of learning a BN from a database it is common to assume that all the variables involved are discrete, in particular that each variable has a finite domain. This is not often the case in many real contexts (such as economics, finance, medicine, engineering, etc.) where also continuous variables are observed.

When the analyzed variables are both discrete and continuous, a common solution consists in discretizing the continuous variables. In this way on one hand there is a loss of information, but on the other hand there is no need to resort to strong distributional assumptions that may be completely unrealistic. Suppose, for example, that the variable of interest is a financial asset; in this case assuming the normality of the distribution would be wrong and would consequently generate huge biases in the statistical data analysis.

It is also important to notice that even when the variables are continuous, national statistical institutes deliver information on the corresponding discretized variable where the classes are either estimated by researches or *a priori* fixed by the final user, i.e. the decision makers.

For the reasons above, it is often convenient (if not necessary) to discretize. This operation can be done in advance, i.e. before learning the dependence model. However in this case the true dependence structure could be dramatically altered. That is, the approximation of a continuous variable with a discrete one can affect the dependence or independence relationships between variables modifying, in turn, the BN structure.

Alternatively, continuous variables can be discretized while learning the BN model (Friedman and Goldszmidt (1996); Monti and Cooper (1998)).

Friedman and Goldszmidt (1996) use the MDL approach in discretizing continuous variables while learning the BN. More specifically, they propose to include in their metric the length of the encoding of the original database. This choice, they argue, is naturally based on the MDL principle.

In this paper we propose a discretization algorithm based on the Kullback-Leibler (KL) divergence measure (Cover and Thomas (2006)). The basic assumption is that the variables directly influencing Y (named parents of Y) are known.

The paper is organized as follows. In Section 2 the discretization algorithm is described and an heuristic is proposed. In Section 3 its performance is evaluated through a simulation study. Finally, potentialities and possible extensions of the algorithm both to the multivariate case and to the BN learning process are discussed in Section 4.

2 Discretization algorithm and an heuristic search

A DAG is a pair $G = (V, E)$ where V is the set of nodes and E is the set of directed edges (arrows) between pairs of nodes. Each node represents a random variable, while missing arrows between nodes imply independence or conditional independence between the corresponding variables. Given a variable Y, its parents are all the variables that directly influence the state of Y, *i.e* linked to Y by arrows pointing to it.

In a BN, each node in a DAG is associated with the distribution of the corresponding variable given its parents (if a node has no parents, it is associated with its marginal distribution). Formally speaking a BN is a pair DAG/joint probability distribution satisfying the Markov properties (see Cowell et al. (2007)). On the basis of these probabilistic conditional independence properties, *i.e.* according to the DAG, the multivariate distribution of all the variables can be decomposed in the product of the conditional probability distribution of each variable given its parents (*chain rule*). Suppose that $(X_1, \ldots, X_p)$ are the variables of interest, then by the chain rule we can write

$$f(x_1, \ldots, x_q) = \prod_{i=1}^{q} f(x_i | pa(x_i)) \tag{1}$$

where $pa(X_i)$ denotes the parent set of the variable X_i, $i = 1, \ldots, q$.

From expression (1), it derives that the *nucleus* of a discretization procedure when a BN is considered, consists in discretizing a continuous variable given a discrete one. Note that the results can be easily extended to the case of multiple discrete parents.

Let X be a discrete variable with states t, $t = 1, 2, \ldots, T$ and let Y be a continuous variable; denote by $f(x, y)$ the joint mixed probability density function for (X, Y) given by

$$f(x, y) = p(x)f(y|x) \tag{2}$$

where $p(x)$ is the marginal distribution function of X and $f(y|x)$ is the conditional distribution of Y given X. Without loss of generality, we suppose that X is a parent of Y.

Here we discuss the problem of discretizing the continuous variable Y. A *discretization sequence* $\beta = (d_0, d_1, \ldots, d_h)$ with $d_0 < d_1 < \cdots < d_h$ is an ordered set of real values partitioning the real line into a finite number of intervals. More specifically, each element d_i, $i = 0, 1, \ldots, h$ is called *midpoint*. The *length* of β is denoted as $|\beta| = h + 1$. Roughly speaking, to discretize a continuous variable means to partition its support into a set of intervals identified by β followed by the definition of a random variable having constant probability function $f_\beta(x, y)$ (*e.g.* the discretized version of $f(x, y)$ based on the sequence β) on the partitioning intervals.

The aim of the paper is to identify the discretization sequence β according to the Kullback-Leibler divergence measure between $f(x, y)$ and $f_\beta(x, y)$. The KL divergence is utilized to measure the information loss in approximating the joint density $f(x, y)$ with piecewise-constant functions based on the sequence β.

Alternative methods can be used to determine the sequence β. A first way to proceed is to identify β by minimizing the Kullback-Leibler measure between the marginal density function $f(y)$ and the locally constant function $f_\beta(y)$, *i.e.* ignoring the information provided by X. Let us denote by β_m such a sequence and by $KL(\beta_m)$ the corresponding Kullback-Leibler divergence $KL(f(x, y)||f_{\beta_m}(x, y))$. Clearly, such a procedure is justified only under the assumption that X and Y are independent. In fact, if X affects Y the discretization procedure should take into account the dependence structure

between the two variables.

Let β^t be the sequence minimizing $KL(f(y|t)||f^t_\beta(y|t))$, for each state t, $t = 1, \ldots, T$, of the discrete variable X, then

$$\begin{aligned} KL(\beta^1, \ldots, \beta^T) =& KL(f(x,y)||f_{\beta^1,\ldots,\beta^T}(x,y)) \\ =& \sum_t \int_y f(x,y) \log \frac{f(x,y)}{f_{\beta^1,\ldots,\beta^T}(x,y)} \\ =& \sum_t \int_y f(y|t)f(t) \log \frac{f(y|t)f(t)}{f_{\beta^1,\ldots,\beta^T}(y|t)f(t)} \\ =& \sum_t f(t) \int_y f(y|t) \log \frac{f(y|t)}{f_{\beta^1,\ldots,\beta^T}(y|t)} \\ =& E_X \left[KL(f(y|t)||f_{\beta^t}(y|t)) \right]. \end{aligned} \tag{3}$$

Clearly, $KL(\beta^1, \cdots, \beta^T) \leq KL(\beta_m)$, the equality holding if and only if X and Y are independent.

We next propose a discretization algorithm able to identify a sequence β from the T sequences β^t such that

$$KL(\beta^1, \cdots, \beta^T) \leq KL(\beta) \leq KL(\beta_m). \tag{4}$$

The algorithm is organized in the following three steps:

Step 1 Choose, for each conditional distribution function $f(y|t)$, $t = 1, 2, \ldots T$, an initial discretization sequence $\delta^t = (d^t_0, d^t_1, \ldots, d^t_{k_t})$. Compute the discretized version $f_{\delta^t}(y|t)$ of $f(y|t)$ based on sequence δ^t.

Step 2 Find a new discretization sequence $\beta^t = (d^t_0 = b^t_0 < b^t_1 \cdots < b^t_{h-1} < b^t_{h_t} = d^t_{k_t})$ shorter than δ^t (*i.e* with $h_t < k_t$) minimizing the Kullback-Leibler (KL) divergence measure between $f_{\delta^t}(y|t)$ and $f_{\beta^t}(y|t)$.

Step 3 Define a unique discretization sequence from the T discretization sequences β^t.

2.1 Step 1

For each conditional distribution $f(y|t)$ an initial discretization sequence $\delta^t = (d_0^t, d_1^t, \ldots, d_{k_t}^t)$ can be defined such that

$$\int_{-\infty}^{d_0^t} f(y|t)\, dy = \int_{d_{k_t}^t}^{\infty} f(y|t)\, dy = 0$$

then d_0^t and $d_{k_t}^t$ can be interpreted as domain specific bounds of $f(y|t)$. Let $I_{(\delta^t,i)} = \{y \in \mathbb{R} | d_{i-1}^t \leq y < d_i^t\}$ and $|I_{(\delta^t,i)}| = d_i^t - d_{i-1}^t$ for $i = 1, \ldots, k$ be the ith interval and its length, respectively. The function $f(y|t)$ can be estimated by

$$f_{\delta^t}(y|t) = \begin{cases} 0 & y < d_0^t \text{ or } y \geq d_{k_t}^t \\ \dfrac{N(t, I_{(\delta^t,i)})}{|I_{(\delta^t,i)}| N(t)} & d_{i-1}^t \leq y < d_i^t, \\ & i = 1, \ldots, k \end{cases} \tag{5}$$

where $N(t)$ and $N(t, y \in I_{(\delta^t,i)})$ represent the number of units of y such that $X = t$ and $(X = t, d_{i-1}^t \leq y < d_i^t)$, respectively.

2.2 Step 2

Given the initial sequence δ^t derived in *Step 1* conditionally on t, *Step 2* finds a new sequence $\beta^t = (d_0^t = b_0^t < b_1^t \cdots < b_{h_t-1}^t < b_{h_t}^t = d_{k_t}^t)$ shorter than δ^t minimizing the Kullback-Leibler divergence between $f_{\delta^t}(y|t)$ and its approximation $f_{\beta^t}(y|t)$ induced by the discretization sequence β^t, given by

$$f_{\beta^t}(y|t) = \begin{cases} 0 & y < b_0^t \text{ or } y \geq b_{h_t}^t \\ \dfrac{1}{|I_{(\beta^t,i)}|} \displaystyle\int_{b_{i-1}^t}^{b_i^t} f_{\delta^t}(y|t)\, dy & b_{i-1}^t \leq y < b_i^t, \\ & i = 1, \ldots, h_t \end{cases} \tag{6}$$

where $I_{(\beta^t,i)} = b_i^t - b_{i-1}^t$. Therefore if we measure the difference between f_{δ^t} and f_{β^t} using the Kullback-Leibler distance we have

$$KL(f_{\delta^t}(y|t) || f_{\beta^t}(y|t)) = \int_{b_0^t}^{b_h^t} f_{\delta^t}(y|t) \log \frac{f_{\delta^t}(y|t)}{f_{\beta^t}(y|t)}\, dy \tag{7}$$

and by (7) and by the definition of f_{β^t} we have

$$
\begin{aligned}
KL(f_{\delta^t}(y|t)||f_{\beta^t}(y|t)) &= \int_{b_0^t}^{b_h^t} f_{\delta^t}(y|t) \log \frac{f_{\delta^t}(y|t)}{f_{\beta^t}(y|t)}\, dy \\
&= \int_{b_0^t}^{b_h^t} f_{\delta^t}(y|t) \log f_{\delta^t}(y|t)\, dy - \int_{b_0^t}^{b_h^t} f_{\delta^t}(y|t) \log f_{\beta^t}(y|t)\, dy \\
&= -\mathbb{H}(f_{\delta^t}(y|t)) - \sum_{i=1}^{h} \int_{b_{i-1}}^{b_i} f_{\delta^t}(y|t) \log f_{\beta^t}(y|t)\, dy \\
&= -\mathbb{H}(f_{\delta^t}(y|t)) - \sum_{i=1}^{h} \log \left(\frac{\int_{b_{i-1}}^{b_i} f_{\delta^t}(y|t) dy}{|I_{(\beta^t,i)}|} \right) \int_{b_{i-1}}^{b_i} f_{\delta^t}(y|t) dy
\end{aligned}
\tag{8}
$$

where $\mathbb{H}(f_{\delta^t}(y|t))$ is the entropy of $f_{\delta^t}(y|t)$. Given the initial discretization sequence δ^t, the term $\mathbb{H}(f_{\delta^t}(y|t))$ in equation (8) is constant. Hence the distance between f_{δ^t} and f_{β^t} only depends on the term

$$
\mathbb{H}(f_{\beta^t}(y|t)) = \sum_{i=1}^{h} \log \left(\frac{\int_{b_{i-1}}^{b_i} f_{\delta^t}(y|t) dy}{|I_{(\beta^t,i)}|} \right) \int_{b_{i-1}}^{b_i} f_{\delta^t}(y|t) dy \tag{9}
$$

which is the entropy of the variable Y conditioned on t, induced by the discretization sequence β^t.

If for all discretization sequences β_2^t with $\beta_2^t \subset \beta_1^t$ we have that $\mathbb{H}(f_{\beta_2^t}(y|t)) > \mathbb{H}(f_{\beta_1^t}(y|t))$, we say that the discretization sequence β_1^t is *minimal*. The main results are in Theorem 1 and Corollary 2.

Theorem 1. *Let β^t be a minimal discretization sequence of length h^t minimizing the KL divergence (8) then β^t is a proper subset of δ^t, $\beta^t \subset \delta^t$.*

Proof. In order to simplify the notation in the proof we drop everywhere the reference to the state t of the discrete variable, thus considering all the density functions as univariate density of the continuous variable Y. Let $\delta = (d_0, d_1, \ldots, d_k)$ and $\beta = (b_0, b_1, \ldots, b_h)$. Suppose that b_j falls somewhere between d_r and d_{r+1}. We distinguish three cases: (i) $b_{j-1} < d_r$ and $d_{r+1} < b_{j+1}$, (ii) $d_r < b_{j-1}$ and $b_{j+1} < d_{r+1}$ and (iii) $b_{j-1} < d_r$ and $b_{j+1} < d_{r+1}$ (the case $b_{j-1} > d_r$ and $b_{j+1} > d_{r+1}$ is symmetric and is omitted). We begin with case (i). Let

$$\begin{aligned} v =& v_{r+1} = f_\delta(d_r) \\ z =& b_j - d_r \\ I =& d_{r+1} - d_r \\ P_L =& \int_{b_{j-1}}^{d_r} f_\delta(z)dz \\ P_R =& \int_{d_r}^{b_{j+1}} f_\delta(z)dz \\ I_L =& d_r - b_{j-1} \\ I_R =& b_{j+1} - d_r \end{aligned}$$

We can rewrite (9) as a function on z as follows

$$\begin{aligned} \mathbb{H}(f_{\beta^t}(y)) = h(z) = \sum_{\substack{i \neq j \\ i \neq j+1}} \log \left(\frac{|I_{(\beta,i)}|}{\int_{b_{i-1}}^{b_i} f_\delta(y)dy} \right) \int_{b_{i-1}}^{b_i} f_\delta(y)dy \\ + (P_L + vz) \log \frac{I_L + z}{P_L + vz} + (P_R - vz) \log \frac{I_R - z}{P_R - vz} \end{aligned} \tag{10}$$

If we take the derivative of (10) with respect to z we have

$$\begin{aligned} h'(z) = \frac{d}{dz} \left[\mathbb{H}(f_{\beta^t}(y)) \right] =& v \log \frac{I_L + z}{P_L + vz} + \frac{P_L - vI_L}{I_L + z} - v \log \frac{I_R - z}{P_R - vz} \\ & - \frac{P_R - vI_R}{I_R - z} \end{aligned} \tag{11}$$

Now if we take the second derivative of (10) we have

$$\begin{aligned}\frac{d}{dz}\left[h'(z)\right] = & v\frac{P_L - vI_L}{(I_L+z)(P_L+vz)} - \frac{P_L - vI_L}{(I_L+z)^2} \\ & + v\frac{P_R - vI_R}{(I_R-z)(P_R-vz)} - \frac{P_R - vI_R}{(I_R-z)^2} \\ = & \frac{P_L - vI_L}{I_L+z}\left(\frac{v}{P_L+vz} - \frac{1}{I_L+z}\right) \\ & + \frac{P_R - vI_R}{I_R-z}\left(\frac{v}{P_R-vz} - \frac{1}{I_R-z}\right) \\ = & -\frac{(P_L - vI_L)^2}{(P_L+vz)(I_L+z)^2} - \frac{(P_R - vI_R)^2}{(P_R-vz)(I_R-z)^2} \end{aligned} \tag{12}$$

By (12) the second derivative of (10) is negative. This means that (10) is a concave function of z in the interval $[0, (d_{r+1} - d_r)]$ and the minimum can be only attained in the extremes of the interval, which implies that b_j must be equal either to d_r or to d_{r+1}.

As for case (ii) suppose that $b_{j-1} > d_r$ and $b_{j+1} < d_{r+1}$ then we can rewrite (9) as

$$\begin{aligned}\mathbb{H}(f_{\beta^t}(y)) = & \sum_{\substack{i\neq j \\ i\neq j+1}} \log\left(\frac{|I_{(\beta,i)}|}{\int_{b_{i-1}}^{b_i} f_\delta(y)dy}\right)\int_{b_{i-1}}^{b_i} f_\delta(y)dy \\ & - v(b_{j+1} - b_{j-1})\log v \end{aligned} \tag{13}$$

which is the entropy of the density $g_{\beta'}$ induced by the discretization sequence β' obtained from β by deleting b_j. Therefore β was not minimal as supposed (contradiction).

The proof of (iii) is similar to the proof of (i) and is omitted. □

To solve the discretization problem, we have to search over all the possible discretization sequences of length h in order to find the one that minimizes the KL distance. Thanks to Theorem 1, the number of the discretization sequences to check is finite. Furthermore, the set of all possible discretization sequences of length h is obtained by generating all possible subsets of length

$h-1$ chosen in a set of $k-1$, where $k+1$ is the length of the initial discretization sequence. Since h is fixed in advance (and in practical application it is generally very small), we have the following

Corollary 2. *The optimal discretization sequance problem is polynomially solvable.*

Proof. By Theorem 1 and by the fact that $O(\binom{k-1}{h-1}) = O((k+1)^h)$. □

In case the length $k+1$ of the initial discretization sequence δ is extremely large, the computational burden can become unbearable although the search of β is a polynomial time solvable problem. Then the following heuristic can be applied. We start with the initial discretization sequence and we remove one midpoint at a time until the number of midpoints is equal to h. The midpoint to be removed is chosen so that the entropy increase due to the new (and one midpoint shorter) discretization sequence is minimized.

In this way we can effectively reduce the length of the discretization sequence up to one hundred or less without loosing much information. At this point the computational cost of the algorithm becomes feasible.

2.3 Step 3

From the union of the T discretization sequences β^t obtained in *Step 2* a unique sequence β is identified. Formally, from (4) we have

$$\begin{aligned} KL(\beta) &= KL(f_{\delta^t}(y|t)||f_\beta(y|t)) = \mathbb{H}(f_{\delta^t}(y|t)) + \int f_{\delta^t}(y|t) ln f_\beta(y|t)\, dy \\ &\leq H(f_{\delta^t}(y|t)) + \int f_{\bigcup \beta^t}(y|t) ln f_\beta(y|t)\, dy \end{aligned} \tag{14}$$

Note that, if the length of $\bigcup \beta^t$ is much larger then length of β the heuristic introduced in *Step 2* can be applied, otherwise an exhaustive search over the space generated by the vectors β of length h from $\bigcup \beta^t$ can be performed.

3 A Simulation Study

The performance of the algorithm has been tested through a simulation study. Five populations of size $N = 90000$ have been generated, where X is a

categorical variable assuming the states $1, 2, 3$ with probability $1/3$ while Y is a mixture of Gaussian distributions given by

$$f(y) = \frac{1}{3}N(10, 1) + \frac{1}{3}N(20, 2) + \frac{1}{3}N(15, .8). \tag{15}$$

The five populations differ in terms of association strength between X and Y. The association is measured by the global correlation coefficient λ which takes values between 0 and 1 and captures the overall dependence, both linear and nonlinear. $\lambda = 0$ if and only if X contains no information on Y and $\lambda = 1$ if there exists a perfect relationship between X and Y. The number of initial bins is equal to the square root of the number of observations. The length of sequences β^t is equal to 10, for each $t = 1, 2, 3$. Then, the union of β^t provides a sequence of length 30 and the final sequence β has been obtained through an exhaustive search over the space generated by the vectors of length 10.

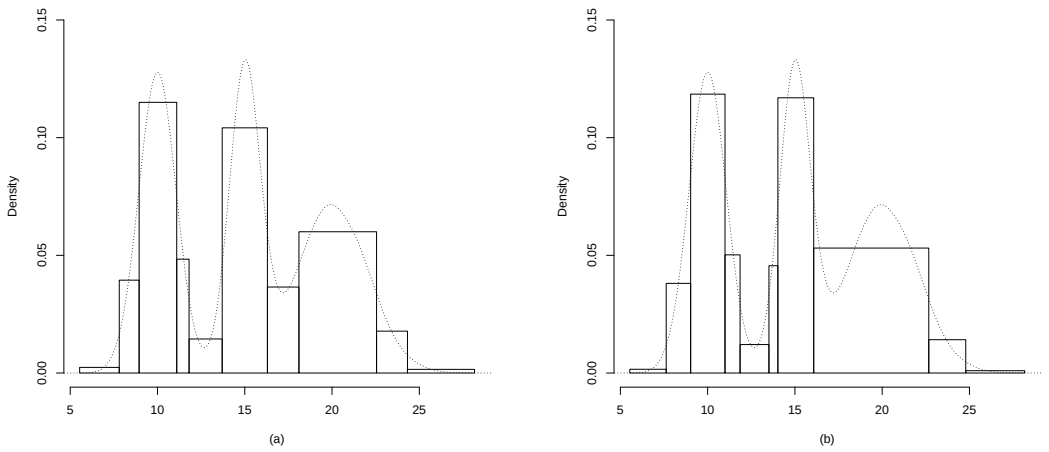
The results are reported in Table 1, where PV represents the percentage variation between the upper bound $KL(\beta_m)$ and the (14).

Population	λ	$KL(\beta^1, \cdots, \beta^T)$	$KL(\beta)$	$KL(\beta_m)$	PV
1	0.01	0.033	0.033	0.033	0
2	0.29	0.034	0.039	0.051	23
3	0.48	0.031	0.047	0.063	26
4	0.70	0.029	0.055	0.103	46
5	0.92	0.012	0.063	0.177	65

Table 1: Performance of the discretization algorithm

If X and Y are independent then $\beta^t \cong \beta$, for each t, the $KL(\beta^1, \cdots, \beta^T) \cong KL(\beta_m)$ and the $PV \cong 0$. The stronger is the dependence between X and Y as measured by λ, the better is the performance of the discretization algorithm. In Figure 1 the histograms corresponding to the sequence β and β_m and the true density are reported.

Figure 1: *Histogram and True Density for* $\lambda = 0.48$ *using* β *(a) and* β_m *(b)*



4 Conclusions and future work

In this paper we have proposed an algorithm that can be generally applied to discretize a continuous variable Y in case an auxiliary discrete variable affecting Y is known. It is also straightforward to extend this algorithm to the case where more than one auxiliary variable is available.

We have shown that when X affects Y the discretization takes advantage of the dependence structure. Since BNs are a natural tool to represent the dependence structure among variables, this algorithm can be usefully applied to discretize continuous variable(s) while keeping a known association structure. In fact, the auxiliary variables for Y constitute the parent set of Y in the DAG; moreover the joint probability distribution of all the variable in the DAG can be factorized according to the chain rule (1). Therefore an algorithm to discretize a continuous variable conditionally on its parents represents the basic and fundamental building-block to deal with discretization in BNs.

Hence, the natural directions for future research include (i) the extension to the multivariate case (more than one continuous variable) assuming that the BN association structure is known; (ii) the discretization process assuming that the BN association structure is unknown.

As far as the first point is concerned, the problem with discretization in BNs is that two continuous variables Y_1 and Y_2 which are conditionally independent given X may have discretized counterparts Y_1^d and Y_2^d that are not independent conditionally on X, and viceversa. That is the relationship between Y_1 and Y_2 can not be properly represented by the Bayesian network learned from Y_1^d and Y_2^d.

Then, preserving the BN association structure at least approximately is important if the probability distribution modeled by BN must be an accurate approximation of the true probability distribution. Let β^1, β^2 be the optimal discretization sequences corresponding to the variables Y_1 and Y_2, respectively. Then, β^1, β^2 must be chosen in such a way that the association structure remains unchanged. This can be checked incorporating an independence test between Y_1^d and Y_2^d in the discretization algorithm.

As far as the second point is concerned, when the BN structure is unknown the discretization can be done in advance, *i.e.* before learning the dependence model. However in this case the true dependence structure could be dramatically altered. An alternative way to proceed consists in discretizing while learning BN as in Friedman and Goldszmidt (1996) and Monti and Cooper (1998), using a score function based on the Kullback-Leibler divergence measure. In particular Friedman and Goldszmidt (1996) proposed to alternate the BN learning step and the discretization step. A different way to work could consist in the following two steps:

Step 1 the network is learned on the data base where the continuous variable is replaced by its discrete approximation in (5);

Step 2 the minimal discretization sequence is searched compatibly with the structure resulting from step 1 or with structures independence equivalent to it.

Acknowledgements

We would like to thank Fabio Corradi for useful comments and remarks, which improved the presentation of the paper.

References

Bouckaert, R. R. (1993). Probalistic network construction using the minimum description length principle. In Clarke, M., Kruse, R., and Moral, S., editors, *ECSQARU*, volume 747 of *Lecture Notes in Computer Science*, pages 41–48. Springer.

Bouckaert, R. R. (1994). Properties of bayesian belief network learning algorithms. In de Mántaras, R. L. and Poole, D., editors, *UAI*, pages 102–109. Morgan Kaufmann.

Cooper, G. F. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, 9(4):309–347.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory (2. ed.)*. Wiley.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (2007). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer Publishing Company, Incorporated, 1st edition.

Friedman, N. and Goldszmidt, M. (1996). Discretizing continuous attributes while learning bayesian networks. In Saitta, L., editor, *ICML*, pages 157–165. Morgan Kaufmann.

Kjræulff, U. and Madsen, A. (2012). *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis: A Guide to Construction and Analysis*. Information science and statistics. Springer.

Lam, W. and Bacchus, F. (1994). Learning bayesian belief networks: An approach based on the mdl principle. *Computational Intelligence*, 10:269–294.

Monti, S. and Cooper, G. F. (1998). A multivariate discretization method for learning bayesian networks from mixed data. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, UAI'98, pages 404–413, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Neapolitan, R. E. (2003). *Learning Bayesian Networks*. Prentice Hall.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, 2 edition.