Unified Conditional Probability Density Functions for Hybrid Bayesian Networks

Mohadeseh Delavarian[1], Mahmoud Naghibzadeh[2], and Mahdi Emadi[3]
[1,2]Computer Engineering Department, [3]Department of Statistics
Ferdowsi University of Mashhad
Mashhad, Iran
[1]delavarian.mohadeseh@stu-mail.um.ac.ir, [2]naghibzadeh@um.ac.ir, [3]emadi@um.ac.ir

Abstract— **Bayesian Network is a significant graphical model that is used to do probabilistic inference and reasoning under uncertainty circumstances. In many applications, existence of discrete and continuous variables in the model are inevitable which has lead to high amount of researches on hybrid Bayesian networks in the recent years. Nevertheless, one of the challenges in inference in hybrid BNs is the difference between conditional probability density functions of different types of variables. In this paper, we propose an approach to construct a Unified Conditional Probability Density function (UCPD) that can represent probability distribution for both types of variables. No limitation is considered in the topology of the network. Hence, the construction of the unified CPD is developed for all pairs of nodes. We take use from mixture of Gaussians in the UCPD construct. Additionally, we utilize Kullback–Liebler divergence to measure the accuracy of our estimations.**

Keywords- hybrid bayesian network; mixture of Gaussians; unified conditional probability density function

I. INTRODUCTION

Bayesian Network (BN) is one of the most prominent graphical models that can handle reasoning under uncertainty. Besides, it is a powerful model to represent the joint probability distribution of the random variables of the model. This graphical model is presented by a Directed Acyclic Graph (DAG) where nodes are random variables of the model and edges represent conditional dependencies. Associated with each node are parameters that represent the conditional probability densities. Each conditional probability density is a function of a random variable, with respect to its parents in the topology of the graph. Random variables in the graph can be either discrete or continuous.

Two main aspects exist for BNs: construction methods, such as [1,2], and inference in it. The goal of constructing BNs is to use this model to do probabilistic inference. Inference in BNs is NP-Hard [3]. Therefore, exact and approximate inference algorithms have been developed for different classes of BNs. Here, our attention is on hybrid BNs and more details will be discussed in the proceeding sections.

We propose a unified conditional probability density function for all types of nodes in the hybrid BNs with the aid of mixture of Gaussians.

The rest of the paper is organized as following. Chapter 2 is focused on hybrid BNs; limitations that are stipulated in this class of BNs and inference in it. Chapter 3 introduces the proposed unified probability distribution function and the details of the specific definition for each pair of node types.

II. HYBRID BAYESIAN NETWORK: DEFINITION AND INFERENCE

Hybrid Bayesian networks contain discrete and continuous nodes simultaneously. These networks are essential to model many applications such as target tracking, speech recognition and fault diagnosis where existence of both types of nodes is inevitable.

During the definition of hybrid BNs, some limitations are also defined. These limitations are mostly categorized into three classes. First limitation is considered in the topology of the graph and its random variables, that no continuous parent can have any discrete child. This limitation implies that none of the descendents of any continuous node can be a discrete node. The second limitation is about continuous nodes and their probability distributions that can be Gaussian or non-Gaussian (any arbitrary function). The third limitation is about the relationship between continuous nodes that can be linear or nonlinear. If we allow each one of these limitations, we have a restricted class of hybrid BNs. In this paper, we have the only limitation that the probability distributions for continuous nodes are Gaussian. But this does not affect the generality of our discussion because it can be extended for non-Gaussian distribution functions.

Inference in hybrid BNs is more complicated than in other classes of BNs. This is because of different kinds of nodes and their probability distributions. The computations become more complicated such that calculations of the posterior probability on hybrid BNs mostly result in approximate answers.

Different researches have been done for presenting the probability distribution and inference in hybrid BNs. Some of the inference algorithms work only for a special class. One class is CLG (Conditional Linear Gaussian). In CLGs a continuous node cannot have any discrete child and linear relationships exist between continuous nodes. Lauritzen algorithm is an exact method for inference in CLGs [4]. In [5] an architecture that is an extension of lazy propagation is extended for CLGs. For general hybrid BNs, different approaches have been investigated. In inference algorithms, one approach is to approximate the probability distribution function. In [6] Gaussian and Dirac functions are utilized to do inference and compute the required messages. In [7] the

978-1-4673-1460-2/12/$31.00 ©2012 IEEE

author approximates hybrid BN by MoG (mixture of Gaussians) BNs. In [8] inference has been done by the aid of approximating probability density function by mixture of truncated exponentials. In [9] by approximating the probability density function with mixture of polynomials, inference has been done in hybrid BNs. In [10], authors use arc reversal method and describe it between all possible node types.

III. Unified Conditional Probability Density Function

As mentioned earlier, one of the most important characteristic of hybrid BNs is inclusion of discrete and continuous random variables. Here, we do not consider any limitation in the topology of the network graph; i.e. child and parent nodes can be either discrete or continuous. Accordingly, eight different cases of child and parent node types occur. Whether the child node is discrete or continuous, a parent has four possible cases. These cases are as follows: 1) discrete node without any parent, 2) discrete node with discrete parents, 3) discrete node with continuous parents, 4) discrete node with discrete and continuous parents, 5) continuous node without any parent, 6) continuous node with continuous parents, 7) continuous node with discrete parents, and 8) continuous node with discrete and continuous parents.

In each case according to the node and parent type conditional probability density is presented by a probability table called conditional probability table (CPT) or by a probability function that is called conditional probability density function (CPD). In these cases, a discrete node with continuous parents, a discrete node with discrete and continuous parents, a continuous node with discrete parents, and a continuous node with discrete and continuous parents, the number of CPD functions is more than one. More explanation can be found in the rest of the paper.

In this paper, we propose a unified conditional probability density function. Our goal is to define a function that represents the whole CPT and CPD functions for all node types. We call this function Unified Probability Density function (UCPD). As a result, a UCPD can handle differences between CPD types in inference algorithms.

We further illustrate each case in details where for each case UCPD is defined separately. This UCPD is approximately equal to the actual CPDs or CPT. For the purpose of defining this UCPD; we take use from mixture of Gaussian functions. In addition, Kullback–Liebler (KL) divergence is used as a measure of goodness of estimation in each case [11].

For the rest of the paper, we use σ to denote standard deviation, Σ to denote covariance matrix, μ to denote mean, p to denote the probability of states of a discrete random variable, c to denote coefficient, and finally s_X to denote the number of states of a discrete random variable X. We use uppercase letters to represent the random variables, lowercase letters to represent the states of those variables, and finally Pa(X) denotes parents of variable X. The function $N(x,\mu,\sigma)$ is normal or Gaussian function with mean μ and standard deviation σ.

A. Discrete node without any parent

For this type of node, its probability density function is presented by a CPT. For each state of the random variable, one probability value exists. We define the UCPD as follow, for each state of the random variable one Gaussian function is defined. Mean μ of each state is equal to the value of that state. As very small standard deviation is required for accurate estimation, we define σ equal to 0.001. The coefficient c of each Gaussian is the product of the probability of that state and 2.5σ. The UCPD is as below:

$$UCPD(X)=\sum_{i=1}^{s_X} c_i N(x,\mu_i,\sigma), \qquad c_i=2.5\times\sigma\times p_i \tag{1}$$

where, X is a discrete random variable without any parent.

Example 1: Consider node X to be a discrete variable with states 1 and 3. A CPT that defines probabilities of this variable is shown in Table I; also the value of UCPD is shown, where the UCPD is as in (2). The KL divergence between two densities is shown in Table III.

TABLE I. CPT and UCPD for discrete variable X

States of X	CPT	UCDP(rounded by 2 d.p.)
1	0.3	0.3
3	0.7	0.7

$$\begin{aligned}&\sigma=0.001,\ \mu_1=1,\ \mu_2=3,\ p_1=0.3,\ p_2=0.7\\ &UCPD(X)=\sum_{i=1}^{2} c_i N(x,\mu_i,\sigma), \qquad c_i=2.5\times\sigma\times p_i\end{aligned} \tag{2}$$

B. Discrete node with discrete parents

For this case, the states are combinations of states of child and parent nodes. This case is also presented by a CPT. UCPD is defined as follow: first a Gaussian function is defined for each state, then for each parent we add one dimension to the density function. Mean μ for each Gaussian is equal to the value of that state. Σ is a diagonal square matrix with the size of the state dimension and the diagonal elements are equal to 0.01. The coefficient c is the probability of that state multiply by square of covariance determinant and a constant α. The UCPD for this case is as:

$$UCPD(X\mid Pa(X))=\sum_{i=1}^{s_X} c_i N([x\ Pa(x)],\mu_i,\Sigma),\ \ c_i=\alpha\times\sqrt{|\Sigma|}\times p_i \tag{3}$$

Where, X is a discrete random variable that has at least one parent. α is very small and should be determined for each function separately. For these two recent cases, if one considers states as data and estimates density function by non parametric methods, this method can resemble kernel density estimation (KDE) method with Gaussian kernel [12]. However, the bandwidth for the kernel is predetermined and no extra computation is essential.

Example 2: Consider two discrete nodes U and V where V is the parent of U and P(U|V) is defined by a CPT as is shown in Table II. The values of the UCPD are also shown in that Table. The UCPD is as in (4). The KL divergence between two densities is shown in Table III. Constant α for this function is 0.063.

TABLE II. CPT AND UCPD FOR DISCRETE VARIABLES (U|V)

States of U|V		CPT	UCDP(rounded by 2 d.p.)
U=1	*V=1*	0.5	0.5
	V=3	0.5	0.5
U=2	*V=1*	0.3	0.3
	V=3	0.7	0.7

$$\mu_1 = [1 \quad 1]', \mu_2 = [1 \quad 3]', \mu_3 = [2 \quad 1]', \mu_4 = [2 \quad 3]'$$
$$c_i = 0.063 \times \sqrt{|\Sigma|} \times p_i \qquad (4)$$
$$UCPD(U \mid Pa(U)) = \sum_{i=1}^{4} c_i N([u \;\; Pa(u)], \mu_i, \Sigma)$$

C. Continuous node without/with continuous parents

It is assumed that continuous nodes have Gaussian distribution. So the density distribution for these two cases is Gaussian, i.e. no changes for the actual functions are needed and hence computation of KL divergence is meaningless.

D. Discrete node with continuous parents

This case is one of the limitations that is considered for hybrid BNs. If this limitation holds, probability distribution for every discrete node is represented by a CPT. Hence, inference algorithms could deal with all discrete nodes the same. However, if this type of node exists, the probability distribution is presented by a function for each state of a discrete node with respect to its continuous parent, instead of a table. As the number of states for discrete variable increase, the numbers of functions to represent the distribution also increase. We use *f* to denote this function. We need to sum up these functions so that it becomes one probability function for all states. The UCPD is defined as follows, for each state a Gaussian function is defined. Mean μ of each Gaussians is equal to the value of that state; σ should be very small so it is predetermined and is set to 0.001. The coefficient *c* of each Gaussian is the product of 2.5σ and function *f*. The resulted UCPD is shown in (5).

$$UCPD(X \mid Pa(X)) = \sum_{i=1}^{s_X} c_i N([x \;\; Pa(x)], \mu_i, \sigma), \; c_i = 2.5 \times \sigma \times f_i \qquad (5)$$

Where, discrete node X can have one or more continuous parent; this relation is embedded in function *f*.

Example 3: Consider discrete node X and continuous node Y; where Y is the parent of X. X has two states 1 and 9. For each one of these states, as mentioned earlier, one function is defined as the probability distribution with respect to its continuous parent Y. These probability functions and UCPD are defined in (6). The KL divergence between the actual CPDs and UCPD is shown in Table III.

$$f_1 = P(X = 1 \mid Y) = \frac{1}{1 + e^{-2y}}, \; f_2 = P(X = 9 \mid Y) = \frac{e^{-2y}}{1 + e^{-2y}}$$
$$\mu_1 = 1, \mu_2 = 9, \qquad c_i = 2.5 \times \sigma \times f_i \qquad (6)$$
$$UCPD(X \mid Pa(x)) = \sum_{i=1}^{2} c_i N([x \;\; Pa(x)], \mu_i, \sigma)$$

E. Continuous node with discrete parents

In this case, conditional probability for every combination of states of discrete parents is represented by a function, here with a Gaussian function. It means that if discrete parents have more than one state, more than one CPD function represent the actual probability distribution. Now we want to sum up these functions so that there is only one probability function. For the purpose of defining the UCPD, we add dimensions to the actual function with respect to the number of discrete parents. We want to sum up all of them, so we should define the new mean and covariance, instead of standard deviation, for the UCPD. As mentioned before, we add dimensions for the new function so the new mean μ is the mean of the actual probability function and every state of the parents. Covariance matrix Σ is a diagonal one with the diagonal elements equal to 0.001 except for the first element that is the standard deviation of the actual probability function of that state. The coefficient *c* of each Gaussian is the product of square of covariance determinant and a constant α. The UCPD is shown as in (7):

$$UCPD(X \mid Pa(X)) = \sum_{i=1}^{s_{Pa(x)}} c_i N([x \; Pa(x)], \mu_i, \Sigma_i), \;\; c_i = \alpha \times \sqrt{|\Sigma_i|} \qquad (7)$$

where, X is a continuous node and Pa(X) represents discrete parent nodes.

Example 4: Consider a continuous node Y with its discrete parent X. X has two states 1 and 2. As shown in (8) P(Y|X) has two Gaussian functions since parent of Y has two states. The node Y has one discrete parent therefore the UCPD has mean and covariance of dimension two. The UCPD is also shown in (8). The constant α for this case is equal to 0.079. The KL divergence between the actual CPDs and UCPD is shown in Table III.

$$f_1 = P(Y \mid x = 1) = N(y, -2, 1)$$
$$f_2 = P(Y \mid x = 2) = N(y, 2, 1)$$
$$\mu_1 = [-2 \quad 1]', \mu_2 = [2 \quad 2]' \qquad (8)$$
$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0.001 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0.001 \end{bmatrix}, c_i = 0.079 \times \sqrt{|\Sigma_i|}$$
$$UCPD(X \mid Pa(X)) = \sum_{i=1}^{2} c_i N([x \; Pa(x)], \mu_i, \Sigma_i)$$

F. Continuous node with discrete and continuous parents

For this node type, the assumptions are the same as the previous case. We have one function as probability distribution for every state of discrete parents. The procedure for defining the UCPD is the same as the previous one. However, we should pay attention to the fact that in this case

there is a relationship between every continuous node and its parent. This relationship can be linear or nonlinear and is embedded in the probability function, which is maintained in defining UCPD. So the UCPD is as follows:

$$c_i = \alpha \times \sqrt{|\Sigma_i|}$$
$$UCPD(X \mid Pa(X)) = \sum_{i=1}^{s_{discrete-Pa(x)}} c_i N([x\ Pa(x)], \mu_i, \Sigma_i) \quad (9)$$

where, X is a continuous node and Pa(X) consist of discrete and continuous parents.

Example 5: Consider two continuous nodes Z and Y and discrete node U, where Y and U are the parents of Z. The node U has two states, 1 and 2. As shown below, the relationship between Z and U is embedded in the corresponding functions: Linear (LR) or Nonlinear (NLR) relationships. When the UCPD is defined these relationships are also considered. The UCPD for linear and nonlinear cases is shown below and the KL divergence values are shown in Table III. Here, the constant α is equal to 0.079.

$$LR: \begin{cases} f_1 = P(Z \mid Y, u=1) = N(z, 2+y, 1) \\ f_2 = P(Z \mid Y, u=2) = N(z, -2+y, 1) \end{cases}$$
$$NLR: \begin{cases} f_1 = P(Z \mid Y, u=1) = N(z, 0.01y^3, 1) \\ f_2 = P(Z \mid Y, u=2) = N(z, 0.01(-y)^3, 1) \end{cases}$$
$$Linear: \ \mu_1 = [2+y \ \ 1]', \mu_2 = [-2+y \ \ 2]' \quad (10)$$
$$Nonlinear: \ \mu_1 = [0.01y^3 \ \ 1]', \mu_2 = [0.01(-y)^3 \ \ 1]$$
$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0.001 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0.001 \end{bmatrix}, c_i = 0.079 \times \sqrt{|\Sigma_i|}$$
$$UCPD(Z \mid Pa(Z)) = \sum_{i=1}^{2} c_i N([z\ Pa(z)], \mu_i, \Sigma_i)$$

G. Discrete node with discrete and continuous parents

This case is similar to the case *D* except that here we have discrete parents, too. Hence, the number of states is more than the case with no discrete parent. The relationship between the child node and continuous parent is embedded in function *f*. If we have more than one parent the covariance matrix is a diagonal one with all elements equal to 0.001. So the UCPD is shown as:

$$UCPD(X \mid Pa(X)) = \sum_{i=1}^{s_x} c_i N([x\ Pa(x)], \mu_i, \Sigma), \quad c_i = \alpha \times \sqrt{|\Sigma|} \times f_i \quad (11)$$

where, X is a discrete node. If we have one discrete parent we should replace constant α with 2.5 and the covariance matrix with standard deviation σ. Based on the experiments, constant α is always less than 0.1.

IV. Conclusion and Future Work

In this paper, we focus on hybrid BNs and the complex task of inference. One reason is the difference between node types and the representation of their probability distribution. Consequently we propose a unified probability distribution to represent the CPT and CPD functions of random variables. This UCPD can be extended to consider non-Gaussian functions for continuous nodes. Also it can be used in inference algorithm as a preprocessing phase to change the CPTs and CPDs to one unified form. Currently, we are working to introduce a new algorithm for inference in hybrid BNs with the aid of the proposed UCPD and Loopy Belief Propagation algorithm.

TABLE III. KL Divergence comparison

Case	KL Divergence			
III.A	2.3938e-10			
III.B	1.36181e-09			
III.D	***f1:*** 2.3951e-17		***f2:*** 3.3742e-16	
III.E	***f1:*** 7.6719e-16		***f2:*** 9.6338e-16	
III.F	***Linear Relationship***		***Non- Linear Relationship***	
	f1:5.8312e-16	***f2***:6.9374e-16	***f1***:6.2440e-17	***f2***:2.9825e-17

References

[1] D.M. Chickering, "Optimal Structure Identification with Greedy Search," Journal of Machine Learning Research, vol. 3, pp. 507-554, 2002.

[2] K. Etminani, M. Naghibzadeh,and A.R. Razavi, "Globally Optimal Structure Learning of Bayesian Networks from Data," ICANN, pp.101-106, 2010.

[3] P. Dagum and M. Luby, "Approximating probabilistic inference in Bayesian belief networks is NP-hard," Artifcial Intelligence, vol. 60, pp. 141-153, 1993.

[4] S. L. Lauritzen and F. Jensen, "Stable local computations with conditional Gaussian distributions," Statistics and Computing, vol. 11, pp. 191-203, 2001.

[5] A. L. Madsen, "Belief update in CLG Bayesian networks with lazy propagation," International Journal of Approximate Reasoning, vol. 49, pp. 503–521, 2008.

[6] O. C. Schrempf and U. D. Hanebeck, "A New Approach for Hybrid Bayesian Networks Using Full Densities," In Proceedings of the 6th International Workshop on Computer Science and Information Technologies CSIT, Budapest, Hungary, 2004.

[7] P. P. Shenoy, "Inference in Hybrid Bayesian Networks Using Mixtures of Gaussians," In UAI, 2006.

[8] R. Rumí and A. Salmerón, "Approximate probability propagation with mixtures of truncated exponentials," International Journal of Approximate Reasoning, vol. 45, pp. 191-210, 2007.

[9] P. P. Shenoy and J. C. West, "Inference in hybrid Bayesian networks using mixtures of polynomials," International Journal of Approximate Reasoning, vol. 52, pp. 641-657, July 2011.

[10] E. N. Cinicioglu and P. P. Shenoy, "Arc reversals in hybrid Bayesian networks with deterministic variables," International Journal of Approximate Reasoning, vol. 50, pp. 763-777, 2009.

[11] S. Kullback and R.A. Leibler, "On information and sufficiency," Annals of Mathematical Statistics,vol. 22, pp. 76–86, 1951.

[12] E. Parzen, "On estimation of a probability density function and mode," Annals of Mathematical Stattistics vol. 33, pp. 1065–1076, 1962.