JMLR: Workshop and Conference Proceedings vol 52, 240-251, 2016 PGM 2016

Hybrid Copula Bayesian Networks

Kiran Karra KIRAN.KARRA@VT.EDU
Lamine Mili LMILI@VT.EDU
Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University
Arlington, VA (USA)

Abstract

This paper introduces the hybrid copula Bayesian network (HCBN) model, a generalization of the copula Bayesian network (CBN) model developed by Elidan (2010) for continuous random variables to multivariate mixed probability distributions of discrete and continuous random variables. To this end, we extend the theorems proved by Nešlehová (2007) from bivariate to multivariate copulas with discrete and continuous marginal distributions. Using the multivariate copula with discrete and continuous marginal distributions as a theoretical basis, we construct an HCBN that can model all possible permutations of discrete and continuous random variables for parent and child nodes, unlike the popular conditional linear Gaussian network model. Finally, we demonstrate on a numerous synthetic datasets and a real life dataset that our HCBN compares favorably, from a modeling and flexibility viewpoint, to other hybrid models including the conditional linear Gaussian and the mixture of truncated exponentials models.

Keywords: Hybrid networks; Bayesian; copula.

1. Introduction and Motivation

Modeling large multivariate probability distributions is an essential research activity in many fields, ranging from engineering to computational biology to economics. Their advantage is that they allow us to understand the dependencies and interactions between random variables, which is arguably instrumental in making intelligent decisions from data. Although modeling large multivariate distributions has many advantages, it is a difficult problem from an analytical and computational perspective. Closed form expressions for multivariate continuous random variables are limited in number; common ones include the multivariate Gaussian, t, and chi-squared distributions. Although analytically tractable, these standard models do not fully capture neither the true marginal distribution nor the statistical dependency for datasets which do not follow these parametric models. Computationally, one runs into the curse of dimensionality with many real-life datasets.

For discrete random variables, the situation is better because multinomial random variables can represent arbitrary discrete distributions. However, a difficulty arises when the data is modeled using multivariate probability distributions of both discrete and continuous random variables, referred to henceforth as mixed random variables. One approach to modeling them is to factorize the joint distribution into a conditional distribution of one set of outcomes and a marginal distribution of the other set (Koller and Friedman, 2009; de Leon and Wu, 2011). While this is a valid approach, it suffers from the problem of identifying probability distributions for each discrete outcome. When multiple discrete random variables are combined, the number of conditional distributions to identify explodes combinatorially.

1.1 Prior Work

Bayesian networks are graphical models used to estimate large dimensional joint probability distribution functions by estimating products of smaller joint probability distributions under conditional independence assumptions (Koller and Friedman, 2009). In an all continuous network, these joint probability distribution functions are typically chosen from parametric models such as the multivariate normal probability distribution function. For hybrid networks, where nodes can be modeled as both discrete and continuous random variables, popular models such as the conditional linear Gaussian (CLG) network have additional limitations; in particular, they do not allow a continuous parent to have a discrete child (Koller and Friedman, 2009). By contrast, the mixture of truncated exponentials (MTE) approach removes the parent/child random variable type restrictions by effectively piecewise surface-fitting the underlying distribution (Moral et al., 2001); although effective from a modeling perspective, it still retains the combinatorial problem of identifying conditional distributions, and could introduce inaccuracies at the inflection points of the multivariate distribution due to the nature of the exponential distribution. This motivates the need for new hybrid models.

Elidan (2010) overcomes the limitations of analytically tractable models by using copula theory to model Bayesian networks. Copulas are multivariate joint probability distribution functions for which the marginal distributions are uniform (Nelsen, 2006). Formally, we have

$$H(x_1, \dots, x_n) = C(F_{X_1}(x_1), \dots, F_{X_n}(x_n)). \tag{1}$$

where $H(\cdot)$ is the joint distribution function of $X_1, \dots, X_n$, $C(\cdot)$ is the copula that captures the dependency structure between random variables $X_1, \dots, X_n$, and $F_{X_i}(x_i)$ is the univariate cumulative distribution function of the random variables X_i. Elidan (2010) uses this relationship to create CBNs, and shows that they can effectively model the underlying data while reducing the number of variables to be estimated jointly.

1.2 Limitations of the CBN Approach

The main limitation of the CBN framework in its current form is that it can only model nodes in the Bayesian network as continuous random variables. This is due to the fact that Sklar's theorem, given by (1), guarantees the existence of a unique copula only when the marginal distribution functions represent continuous random variables. For a given joint distribution function with marginal distribution functions that represent discrete random variables, there exist many copulas that satisfy (1). Copula network classifiers (CNCs) (Elidan, 2012) build upon the CBN model and define conditional copulas, allowing for copula networks of mixed random variables. However, CNCs have the same structure limitations as the CLG model.

To overcome this difficulty, several methods have been proposed in literature to construct copulas for discrete probability distributions. For instance, following Schweizer and Sklar (1974), Denuit and Lambert (2005) and Nešlehová (2007) develop continuous extensions and transformations, respectively, of discrete random variables to define unique copulas with discrete marginal distributions that retain many of the familiar properties of copulas with continuous marginal distributions. By contrast, de Leon and Wu (2011) develop a conditional probability distribution-based methodology, while Smith and Khaled (2012) propose a latent variable approach based on MCMC. As for Kolesarova et al. (2006), they introduce a method based on sub-copulas over domains defined by uniform discrete distributions. For reasons to be explained later, we concentrate on the approach investigated by Nešlehová (2007).

1.3 Goals

In this paper, we develop the theoretical framework for the HCBN model. We begin by extending the bivariate discrete random variable copula construction and proofs provided by Nešlehová (2007) to the multivariate ($n \geq 2$) hybrid copula in Section 2 and discuss the applicability of these constructions. Section 3 then applies these copulas to CBNs, originally developed by Elidan (2010), to create the HCBN framework. We then show the expressive modeling power of the HCBN in Section 4 with a synthetic and real-life dataset, and compare its performance to the CLG, MTE, and multinomial BN models for hybrid data. Concluding remarks are provided in Section 5.

2. Hybrid Copulas

The building block of our HCBN construction will be the hybrid copula, which is a copula function that joins both continuous and discrete random variables. The discrete random variables under consideration are count or ordinal discrete random variables, with some ordering between values in the domain of the discrete random variable (this point is discussed in further detail in Section 2.2). The difficulty here stems from the lack of uniqueness of the copula function, since the associated cumulative probability distribution function will exhibit discontinuities at discrete values where it has non-zero probability. Consequently, the associated copula expressed in terms of the joint cumulative distribution function $H(\cdot)$ is given by

$$C(u_1, \ldots, u_n) = H(F_{X_1}^{(-1)}(u_1), \ldots, F_{X_n}^{(-1)}(u_n)) \tag{2}$$

where u_i for all $i = 1, \ldots n$ is a uniform random variable taking on values in the unit interval $[0, 1]$. In this representation of Sklar's theorem, $H(\cdot)$ is a function of u_i for $i = 1 \ldots n$ that takes discrete values over the unit interval $[0, 1]$. Therefore, there exist many copulas that satisfy the relations given by (1) and (2), i.e. the copula is not unique. In order to address this issue, Denuit and Lambert (2005) propose continuing the discrete random variable X with a random variable U valued on $(0, 1)$ independent of X with a strictly increasing cumulative distribution function $L_U(u)$, on $(0, 1)$ and sharing no parameters with X. Nešlehová (2007) generalizes the foregoing construction by defining the transformation $\psi : [-\infty, \infty] \times [0, 1] \rightarrow [0, 1]$

$$\psi(x, u) = P[X < x] + uP[X = x], \tag{3}$$

where X denotes a random variable that can be continuous or discrete and U denotes a uniform continuous random variable that is independent of X, and u is a realization of U. It is then shown in the bivariate case that transforming the discrete random vector (X, Y) to the vector $\mathbf{\Psi}(\mathbf{X}, \mathbf{U}) = (\psi(X, U), \psi(Y, V))$ yields a possible copula describing the original joint discrete distribution function of (X, Y). Additionally, if U and V are independent, the copula so defined is again the standard extension copula of Schweizer and Sklar (1974). Consequently, it preserves the concordance properties of the original discrete distribution function of (X, Y) (Denuit and Lambert, 2005; Nešlehová, 2007). Concordance is a form of dependence, which measures the degree to which two random variables are associated with each other; a pair of random variables are said to be concordant if large values of one random variable tend to be associated with large values of the other random variable, and small values of one random variable are associated with small values of the other variable (Nelsen, 2006). Due to this and other properties and generality, we use the ψ transformation defined by Nešlehová (2007) and extend the proofs to the multivariate mixed copula.

In this paper, we consider the ψ function to take any admissible form; this will allow us to encompass not only the continuous extension defined by Denuit and Lambert (2005), but also the case of continuous random variables following any probability distribution. For instance, if X is a continuous random variable, then we have

$$\psi(X, U) = X, \tag{4}$$

because $P[X \leq x] = P[X < x] + uP[X = x] = P[X < x]$ since $P[X = x] = 0$. In summary, the transformation $\psi(X, U)$ preserves continuous random variables and transforms discrete random variables into continuous uniform random variables over $[0, 1]$. In the next section, we apply (4) and the results in Nešlehová (2007) to define the multivariate mixed copula.

2.1 Multivariate Mixed Copulas

Define the random vector $\mathbf{X} = (X_1, \ldots, X_n)$, and without loss of generality let $X_1, \ldots, X_k$, $k \leq n$ be continuous random variables, and $X_{k+1}, \ldots, X_n$ be ordinal or count discrete random variables. Following the construction of Nešlehová (2007), let us define the transformed random vector as

$$\begin{aligned} \Psi(\mathbf{X}, \mathbf{U}) &= (\psi(X_1, U_1), \ldots, \psi(X_k, U_k), \psi(X_{k+1}, U_{k+1}), \ldots, \psi(X_n, U_n)) \\ &= (X_1, \ldots, X_k, \psi(X_{k+1}, U_{k+1}), \ldots, \psi(X_n, U_n)). \end{aligned} \tag{5}$$

Proposition 1 *The unique copula* $C_{\mathbf{\Psi}(\mathbf{X},\mathbf{U})}$ *of* $\mathbf{\Psi}(\mathbf{X}, \mathbf{U})$ *is a possible copula of* $\mathbf{X}$.

Proof: See Appendix A.

Proposition 2 *If* $U_1 \ldots U_n$ *are independent, then* $C_{\mathbf{\Psi}(\mathbf{X},\mathbf{U})}$ *is the standard extension copula of the form specified by Schweizer and Sklar, extended to a k-linear interpolation.*

Proof: Follows directly from Proposition 1 and Nešlehová (2007).
Additionally, due to (4), we can directly apply the results of Mesfioui and Quessy (2010) to obtain the concordance properties of the multivariate mixed copula. More formally, we have

$$\tilde{Q}(H_1, H_2) = \mathbf{T}Q(H_1, H_2), \tag{6}$$

where $\mathbf{T} = \prod_{l=1}^{d} \left(\frac{T_l+1}{2}\right)$, $T_l P[\mathbf{X}_1 < \mathbf{X}_2] = P(X_{11} < X_{21}, \ldots, X_{1l} < X_{2l}, \ldots, X_{1n} < X_{2n}]$, $\mathbf{X}_1 = (X_{11}, \ldots, X_{1d})$, $\mathbf{X}_2 = (X_{21}, \ldots, X_{2d})$, $\tilde{Q}(H_1, H_2) = Q(C_1^*, C_2^*)$, and C_1^* and C_2^* are the unique copulas associated with $\Psi(\mathbf{X_1}, \mathbf{U_1})$ and $\Psi(\mathbf{X_2}, \mathbf{U_2})$ respectively. Equation 6 shows us that the multivariate mixed copula has the desirable consequence of retaining the concordance properties of the original joint distribution $\mathbf{X} = (X_1, \ldots, X_k, X_{k+1}, \ldots, X_n)$.

2.2 Applicability of the Multivariate Extension Framework

The mixed copulas and their properties derived above in Section 2.1 apply to count or ordinal discrete data. This is because copulas contain information about the dependency structure between two (or more) random variables. If there is no ordering of the discrete data, then the concept of two random variables behaving together in either a concordant or discordant way (in order words, having some dependency structure) does not exist. This however does not prevent one from constructing copulas (details of copula construction are provided in Section 3.2) with mixed random variables where the discrete random variables are unordered, such as categorical data. To construct a copula with categorical data, one assigns arbitrary ordinal values to the categories. This works because the continued random variable CDFs given by (5) agree with the original discrete CDF in the discrete domain. However, interpretation of any dependence structure from the computed copula only makes sense and is valid if the discrete random variables are count data or ordinal data. This is not a limitation of the HCBN framework but rather an inherent property of the discrete data being modeled.

3. Hybrid Copula Bayesian Networks

Having established the theoretical properties and framework for the multivariate hybrid copula, we are now ready to formally define the HCBN. An HCBN is a tuple $\mathcal{C}_{HCBN} = (\mathcal{G}, \Theta_C, \Theta_f)$ that encodes the joint density $f_{\mathcal{X}}(x)$. $\mathcal{G}$ is the graph structure, Θ_C is the set of mixed copula functions for each copula family (i.e. a child and its parents), and Θ_f is the set of parameters representing the marginal densities $f_{X_i}(x_i)$. This is similar to the CBN definition in Elidan (2010), with the difference that Θ_C is a matrix in which the i^{th} column represents the parameters of the i^{th} copula, rather than a 1-D vector with the i^{th} element describing the dependence parameter of the i^{th} copula. This is because the copulas joining all continuous variables only require one parameter (assuming one parameter families are used) to describe the copula, but copulas joining continuous and discrete variables will need more.

To express the overall joint density $\mathcal{X}$ represented by the HCBN framework, let us first define the local density of a family via the HCBN framework. In a local density containing n nodes, without loss of generality, define k to be the continuous marginals, and the remaining $n-k$ to be discrete marginals. The local joint density of the i^{th} family can then be written as

$$f_i(\mathbf{x_i}) = \prod_{l=1}^{k} f_{X_l}(x_l) \times \sum_{j_{k+1}=1}^{2} \cdots \sum_{j_n=1}^{2} (-1)^{j_{k+1}+\cdots+j_n} \times \\ C_i^k(F_{X_1}(x_1), \ldots, F_{X_k}(x_k), u_{k+1,j_n+1}, \ldots, u_{n,j_n}) \tag{7}$$

where

$$C_i^k = \frac{\partial^k}{\partial u_1 \partial u_2 \ldots \partial u_k} C_i(u_1, \ldots, u_n) \tag{8}$$

, $u_{j,1} = F_{X_j}(x_j^-)$, $u_{j,2} = F_{X_j}(x_j)$. The local density in (7) is a product of the continuous marginal distributions, the partial derivative of the encompassing copula function with respect to the continuous variables, and the C-volume of the discrete marginal distributions; the C-volume is computed via the summation of the encompassing copula function in the discrete dimensions over {1,2}, following (2.2.3) in Nelsen (2006). By defining the local density, and defining that $C_i^k = 1$ in the case of root nodes, we derive the full joint density, $\mathcal{X}$, using the conditional independence assumptions of Bayesian networks as follows

$$f_{\mathcal{X}}(\mathbf{x}) = \prod_{i=1}^{D} f_i(\mathbf{x_i}) \tag{9}$$

This is different than the CBN parametrization of $\mathcal{X}$ as it does not rely on the copula densities alone (Elidan, 2010). This is because the CBN model depends upon the fact that for continuous distributions, the relation given by

$$f(x_1, \ldots, x_N) = c(F_{X_1}(x_1), \ldots, F_{X_N}(x_N)) \prod_{i=1}^{N} f_{X_i}(x_i) \tag{10}$$

is true. However, (10) does **not** hold true for discrete or mixed distributions (Panagiotelis et al., 2012). This is the main reason why the CBN framework cannot be directly used for hybrid data. The local density described by (7) is a generalization of (10), and thus applicable to the mixed random variable scenario.

3.1 HCBN Construction

The steps common to any existing structure learning algorithm required to construct the HCBN are:

1. Preprocess each discrete random variable X_i with the transformation $\psi(X_i, U_i)$.
2. Compute empirical marginal distributions for each node in the Bayesian network.

If score-based approaches such as greedy hill climbing are used for learning, candidate structures scores would be computed using the scoring function given by given by

$$\mathcal{L}(\mathcal{D} : \mathcal{C}_{HCBN}) = \sum_{m=1}^{M} \sum_{i} \log f_i(\mathbf{x_i}[\mathbf{m}]),$$

where f_i is the local density defined above in (7) and the best scoring structure taking into account graph complexity would be chosen. The HCBN framework is also compatible with constraint based approaches, although statistical tests of independence and conditional independence for mixed random variables would need to be explored.

It is worth noting the flexibility of the ψ transform in the context of HCBN construction; the transformation ψ is applied without any a-priori knowledge of the Bayesian network structure. One consequence of this is that for each node to which the transformation $\psi(X_i, U_i)$ is applied, an independent U_i is used. The advantage of using an independent U_i for each discrete node is that the constructed copula inherits the dependence properties of the data it is trying to model, as shown by (6). The other advantage is that continuing the discrete random variables and the structure learning are independent, allowing for computational efficiency.

3.2 Copula Density Estimation

In the CBN construction described by Elidan (2010), standard copula models including the Frank and Gaussian copulas are used to join the continuous marginal distributions. The HCBN construction can also use standard copula models when all the nodes of a copula family are modeled as continuous random variables. In the continuous marginals case, this approach works well because the chosen copula's dependency parameter θ can be set such that the empirical Kendall's $\hat{\tau}$ of the dataset matches the Kendall's τ of the copula, using known relationships between τ and θ and thereby capturing the underlying concordance properties of the data (although copula model selection itself is another problem). However, using the relationship between a copula's dependency parameter θ and empirical Kendall's $\hat{\tau}$ with discrete marginal distributions leads to inconsistent and biased estimates of the dependence parameter (Genest and Nešlehová, 2007).

Although algorithms to estimate parameters for standard copulas with discrete marginals without bias have been explored (Smith and Khaled, 2012), to the best of the authors' knowledge, model selection for copulas in the hybrid scenario has not yet been explored. For this reason, we do not use standard copulas for modeling mixed random variables. Propositions 1 and 2 show that copulas constructed by transforming the discrete random variables with ψ and independent U_i's yield unique copulas that follow the concordance properties of the underlying data. Although this amounts to creating a unique copula to model each local mixed distribution, we take this approach for it's theoretical soundness. Equivalently, this unique copula model can be viewed as a nonparametric approach to estimating the dependency structure of the corresponding copula family. This nonparametric approach also avoids the copula compatibility problem, because overlapping copula marginals remain the same due to them being estimated directly from the data, avoiding the marginal skewness problem outlined in the CBN construction (Elidan, 2010). However, this advantage comes at the computational cost of estimating the copula density nonparametrically.

To estimate the multivariate empirical copula density, the beta-kernel approach is taken (Charpentier et al., 2007). Beta-kernels are an ideal way to estimate copula density functions due to their bounded support. For the purposes of HCBN, we estimate the copula density from the data $\mathbf{x}$ directly with

$$\hat{c}_h(\mathbf{u}) = \frac{1}{M} \sum_{m=1}^{M} \prod_{d=1}^{D} \beta(F_{X_d}(x_d(m)), \frac{u}{h}+1, \frac{1-u}{h}+1), \tag{11}$$

where h is a bandwidth parameter, β is the probability distribution function of the Beta distribution, and $F_{X_d}(x_d(m))$ is the m^{th} pseudo-observation for the d^{th} dimensional data point. Equation (11) is a multivariate extension of the bivariate copula density estimated by beta-kernals (Charpentier et al., 2007).

In the context of HCBN, estimating the copula density directly is preferred to finding the copula function and then differentiating. This is due to two primary reasons: 1.) although the empirical copula function estimate introduced by Deheuvels (1979) provides a robust and universal way to compute an empirical copula estimate, the discontinuous features of the estimator introduce difficulties when differentiating the empirical copula, even with respect to the continuous random variables, and 2.) the inverse function of H in (2) at some point $\frac{i}{T}$ may be chosen arbitrarily between the defined points of the dataset from which the empirical copula is constructed (Charpentier et al., 2007). Once the copula density is estimated, the discrete dimensions are numerically integrated to find C_i^k for the i^{th} family.

3.3 Accuracy of Density Estimation in the Hybrid Context

In the hybrid random variable context, (11) becomes a better estimator of the true copula density as the number of discrete outcomes increases. This is because in this scenario, less volume in the unit hypercube is filled in with the uniform random variable in $\psi(X, U)$. Figure 1 shows pseudo-observations of two different bivariate mixed random variables, with the discrete random variable continued by (5). Both have a dependency structure described by the same copula whose pseudo-observations are shown in the middle plot; the mixed random variable with two discrete outcomes is shown in the left plot and the mixed random variable with ten discrete outcomes is shown in the right plot. It is seen that as as the number of outcomes for the discrete random variable increases, the empirical pseudo-observations are closer to the true pseudo-observations, and thus the copula and dependency structure is better estimated.

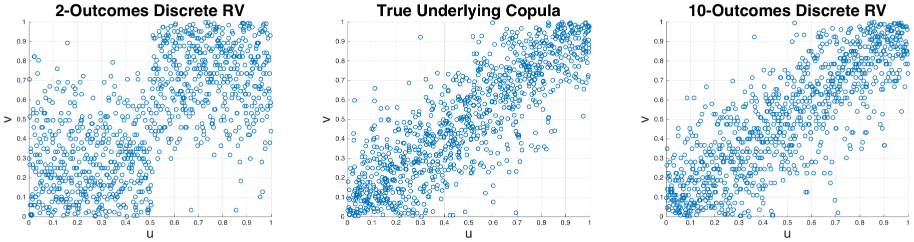
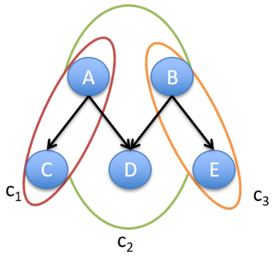


Figure 1: Pseudo-Observations plot for different discrete random variables with the same copula compared to true pseudo-observations of copula.

This intuition is also corroborated by theory in Genest and Nešlehová (2007); Nešlehová (2007); Denuit and Lambert (2005). Conversely, as the number of discrete outcomes increases, the MTE and CLG models need to define an exponentially growing number of conditional distributions with a decreasing number of samples per conditional distribution. Therefore, from a computational and modeling perspective, the HCBN can be recommended over the CLG or MTE models as the number of discrete outcomes per local density (copula family in the HCBN context) increases.

4. Experimental Evaluation

We test the HCBN construction with various experiments on various synthetic datasets and a real-life dataset. In order to compare the HCBN's performance against the CLG, MTE, and multinomial BN models, we first create a synthetic data-set generated from a structure with the same I-map as the Bayesian Network structure showed in Figure 2, where nodes A and B are multinomial discrete random variables, and C, D, and E are continuous random variables. This structure restriction allows us to compare the CLG model fairly to the other models.



Nodes A/B Multinomial Probabilities	Nodes C/D/E PDF Types	Dependency Structure Models (C1/C2/C3)	Dependency Strengths for C1/C2/C3
[0.5 0.5]/[0.5 0.5]	N(2,0.5)/N(2,0.5)/N(2,0.5)	**Linear Dependency** Gaussian/ Gaussian/ Gaussian	**Strong**
	U(-2,2)/U(-2,2)/U(-2,2)		
[0.25 0.25 0.25 0.25]/ [0.25 0.25 0.25 0.25]	N(-2,0.3)+N(2,0.8)/ N(-2,0.3)+N(2,0.8)/ N(-2,0.3)+N(2,0.8)		
	T(3)/T(3)/T(3)		
	N(-2,0.3)+N(2,0.8)/ U(-2,2)/ N(-2,0.3)+N(2,0.8)	**Non-Linear Dependency** Frank/ Gaussian/ Frank	**Weak**
	N(2,0.5)/ N(-2,0.3)+N(2,0.8)/ U(-2,2)		
	U(-2,2)/U(-2,2)/T(3)		
	N(-2,0.3)+N(2,0.8)/ N(2,0.5)/U(-2,2)		

Figure 2: Bayesian Network used for synthetic data testing.

To test each model, every permutation of the different configurations for the discrete nodes, continuous nodes, and the dependencies between them outlined in Figure 2 were tested for the synthetic data evaluation for the CLG, MTE, multinomial BN (data was discretized using a histogram approach), and HCBN models. All models were given the true structure in Figure 2, in order to isolate and quantify the expressiveness of each model. Figure 3 shows the experimental bias and variance of each model for strong and weak, linear and nonlinear dependency structures for the graphical model with copulas C_1, C_2, C_3 in Figure 2.

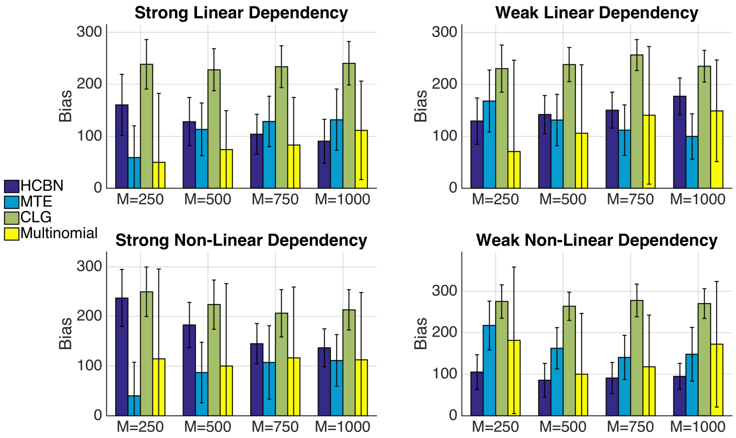


Figure 3: Monte Carlo simulations of Hybrid Models against synthetic data.

The results show that in general, for weak dependencies, the copula approach is on par with or exceeds the performance of the MTE models. For strong dependencies, it is seen that the MTE model has better performance characteristics than the HCBN model. The strong dependency results are explained by the reasoning provided in Section 3.3. The weak dependency results can be explained in a similar light; in a weak dependency scenario, the pseudo-observations are more spread out and resultantly, the domain of the discrete random variable has less impact in the copula estimation of the true dependency structure. As expected, the CLG model displays high bias due to cases in the experiments where the data does not follow the Gaussian distribution. It is seen that the multinomial approach exhibits relatively low bias but high variance for each category tested, which can be explained by the discretization process. It remains to be explored how sensitive the multinomial approach is to other discretization approaches.

In addition to the synthetic data experiments described above, we model the 1994 Census dataset (Lichman, 2013) using the HCBN. Figure 4 shows samples generated from the HCBN Bayesian network, versus the collected census data. The left handed columns in Figure 4 are scatter plots of two random variables from the actual data, and the right handed columns are samples from the HCBN generative model of that data. It is seen from Figure 4 heuristically, that both continuous and discrete random variables are modeled expressively and accurately with the HCBN model.

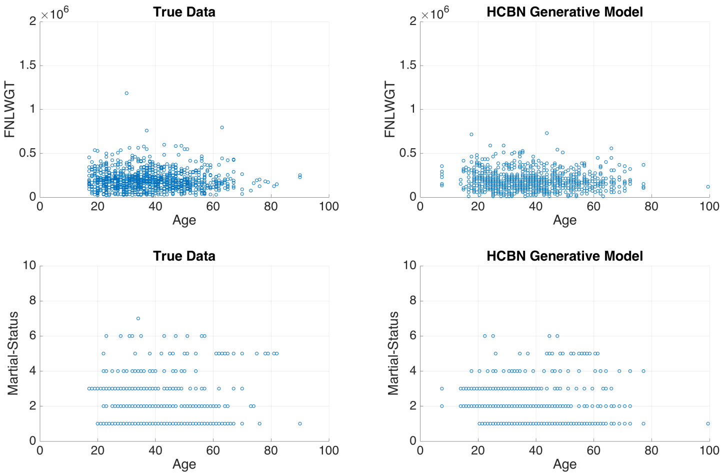


Figure 4: The top left figure shows a scatter plot of Age (Continuous) vs. FNLWGT (Continuous) in the 1994 Census data. The top right figure shows samples produced from the HCBN generative model of Age vs. FNLWGT. The bottom left figure shows a scatter plot of Age (Continuous) vs Martial-Status (Discrete-Categorical) in the 1994 Census data. The bottom right figure shows samples produced from the HCBN generative model of Age vs. Martial-Status.

5. Conclusion

In this paper, we have shown a new method for constructing multivariate copulas with discrete and continuous random variables, and applied this construction to Bayesian networks to create the HCBN. The main contribution is the extension of proofs by Nešlehová (2007) to the multivariate case to ensure the validity of the constructed hybrid copulas and incorporating them into the CBN framework proposed by Elidan (2010). The defined framework has the ability to:

1. Effectively model multivariate hybrid distributions, while removing the restrictions of graph structure imposed by the CLG model (Koller and Friedman, 2009).
2. Avoid defining a combinatorial number of conditional distributions.
3. Efficiently sample the Bayesian network through copula sampling algorithms.

The empirical evaluation shows that the HCBN construction compares favorably to the CLG model and performs similarly to the MTE model. The estimated copula contains information about the underlying dependency structure, which may eventually be useful for causality related studies. Although both of these topics are out of the scope of this paper, they motivate research activities in estimating the copula density of the family via the beta-kernel based technique.

Additional research that will be conducted is to investigate the h parameter in the beta-kernel methods given by 11. Furthermore, more research into copula constructions for hybrid random variables will be conducted to address the issue of underestimating strong dependencies when the domain of the discrete random variables is small, as was seen in Figure 1 and the experimental results shown in Figure 3. Finally, we aim to address the approximate inference and structure learning problem that remains to be solved in the HCBN model.

Acknowledgments

We would like to acknowledge the Hume Center at Virginia Tech for its support. We would also like to acknowledge the anonymous reviewers and editors who provided many helpful comments to improve the quality of this paper.

Appendix A.

Proof As in Nešlehová (2007), let $w \in [0,1]$ and set

$$x(w) := F_X^{(-1)}(w+) \qquad\qquad u(w) = \begin{cases} 1 & \text{if } P[X = x(w)] = 0 \\ \frac{w - P[X < x(w)]}{P[X = x(w)]} & \text{else ,} \end{cases} ,$$

where $F_X^{(-1)}(x+)$ denotes the right hand side limit of the generalized inverse of F_X. For $(w_1, \dots, w_n) \in [0,1]^n$ one can construct points $(x_1(w_1), \dots, x_n(w_n))$ such that

$$\begin{aligned} P[\Psi(X_1{,}U_1) \leq w_1, \dots, \Psi(X_n, U_n) \leq w_n] \\ = \sum_{b \in S} C_{\mathbf{U}}(u_1(w_1), \dots, u_n(w_n)) \times \\ P[X_1 \Box x_1(w_1), \dots, X_n \Box x_n(w_n)], \end{aligned} \tag{12}$$

where S is the set of all possible combinations of binary vectors of length n,

$$u_i(w_i) = \begin{cases} 1 & \text{if } S_b(i) = 1 \\ u_i(w_i) & \text{else} \end{cases} , \qquad\qquad \Box = \begin{cases} = & \text{if } S_b(i) = 1 \\ < & \text{else .} \end{cases} ,$$

$\Box$ stands for the mathematical symbol defined above, and $S_b(i)$ is the i^{th} element of the b^{th} binary vector in the set S. Now suppose that $(w_1, \dots, w_n) \in RANGE\{F_1\} \times \dots \times RANGE\{F_n\}$. As in Nešlehová (2007, Proof of Proposition 4), we have that $x_i(w_i) := F_{X_i}^{(-1)}(w_i)$ or $x_i > F_{X_i}^{(-1)}(w_i)$, which implies that

$$u_i(w_i) = \begin{cases} 0 & \text{if } P[X_i = x_i(w_i)] > 0 \\ 1 & \text{if } P[X_i = x_i(w_i)] = 0. \end{cases}$$

Substituting the appropriate values of $u_i(w_i)$ depending on $P[X_i = x_i(w_i)$, we arrive at

$$\begin{aligned} C_{\Psi(\mathbf{X},\mathbf{U})}(w_1, \dots, w_n) &= P[\psi(X_1, U_1), \dots, \psi(X_n, U_n)] \\ &= P[X_1 \leq F_{X_1}^{(-1)}(w_1), \dots, X_1 \leq F_{X_n}^{(-1)}(w_n)] \\ &= C_{\mathbf{X}}(w_1, \dots, w_n). \end{aligned}$$

■

References

A. Charpentier, J.-D. Fermanian, and O. Scaillet. The Estimation of Copulas: Theory and Practice. *Copulas: from theory to applications in finance*, pages 35–62, 2007.

A. de Leon and B. Wu. Copula-based Regression Models for a Bivariate Mixed Discrete and Continuous Outcome. *Statistics in Medicine*, 2011.

P. Deheuvels. La fonction de dépendance empirique et ses propriétés. Académie Royale de Belgique. *Bulletin de la Classe des Sciences*, 65(5):274–292, 1979.

M. Denuit and P. Lambert. Constraints on Concordance Measures in Bivariate Discrete Data. *Journal of Multivariate Analysis*, 2005.

G. Elidan. Copula Bayesian Networks. In *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010.

G. Elidan. Copula Network Classifiers. In *15th International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2012.

C. Genest and J. Nešlehová. A Primer on Copulas for Count Data. *ASTIN Bulletin*, 2007.

A. Kolesarova, R. Mesiar, J. Mordelova, and C. Sempi. Discrete Copulas. *Fuzzy Systems, IEEE Transactions on*, 2006.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

M. Lichman. UCI machine learning repository - census+income dataset, 2013. URL `http://archive.ics.uci.edu/ml/datasets/Census+Income`.

M. Mesfioui and J.-F. Quessy. Concordance Measures for Multivariate Non-Continuous Random Vectors. *Journal of Multivariate Analysis*, 2010.

S. Moral, R. Rumi, and A. Salmern. Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Springer Berlin Heidelberg, 2001.

R. B. Nelsen. *An Introduction to Copulas*. Springer-Verlag New York, 2006.

J. Nešlehová. On Rank Correlation Measures for Non-Continuous Random Variables. *Journal of Multivariate Analysis*, 2007.

A. Panagiotelis, C. Czado, and H. Joe. Pair Copula Constructions for Multivariate Discrete Data. *Journal of the American Statistical Association*, 107(499):1063–1072, 2012.

B. Schweizer and A. Sklar. On Nonparametric Measures of Dependence for Random Variables. *Studia Mathematica*, 1974.

M. S. Smith and M. A. Khaled. Estimation of Copula Models with Discrete Margins via Bayesian Data Augmentation. *Journal of the American Statistical Association*, 2012.