Hybrid Copula Bayesian Networks

Kiran Karra
kiran.karra@vt.edu

Hume Center
Electrical and Computer Engineering
Virginia Polytechnic Institute and State University

September 7, 2016

Outline

Introduction

- Graphical models can model datasets as large dimensional probability distributions.
- Real world data typically consist of both discrete and continuous random variables.
- Often, simplifying assumptions are made either in modeling the individual marginal distributions, or the dependency structure.

Introduction

- Graphical models can model datasets as large dimensional probability distributions.
- Real world data typically consist of both discrete and continuous random variables.
- Often, simplifying assumptions are made either in modeling the individual marginal distributions, or the dependency structure.

- We present a new model for representing mixed random variables in graphical models using hybrid copulas, based on Copula Bayesian Networks [Eli10].

Introduction to Copulas

Sklar's Theorem[Nel06]

$$C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)) = F(x_1, \ldots, x_n)$$

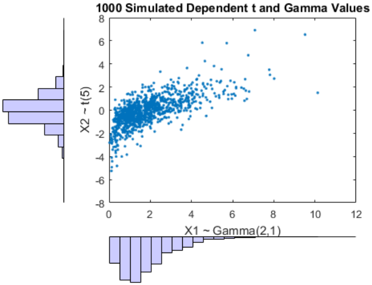
- Any joint distribution can be generated from its marginal distributions and copula.
  - Allows for heterogeneous marginals in joint distribution.
  - Dependency structure is independent of marginal distributions.

Introduction to Copulas

Sklar's Theorem[Nel06]

$$C(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)) = F(x_1, \ldots, x_n)$$

- Any joint distribution can be generated from its marginal distributions and copula.
  - Allows for heterogeneous marginals in joint distribution.
  - Dependency structure is independent of marginal distributions.



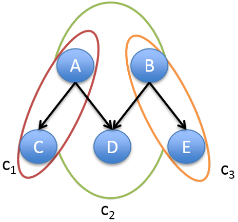
Copula Bayesian Networks (CBN) [Eli10]

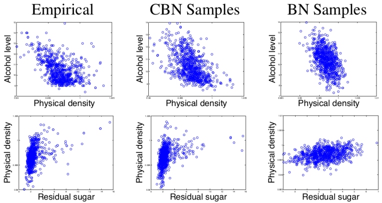
- Copula ratio defines relationship between a child node and it's parents.

$$R_c(F_X(x), \mathbf{F}_{\mathbf{pa}}(y_{\mathbf{pa}})) = \frac{c(F_X(x), \mathbf{F}_{\mathbf{pa}}(y_{\mathbf{pa}}))}{\frac{\partial^K C(1, \mathbf{F}_{\mathbf{pa}}(y_{\mathbf{pa}}))}{\partial \mathbf{F}_{\mathbf{pa}}(y_{\mathbf{pa}})}}$$

- The density χ factorizes over the graph as

$$f_\chi(\mathbf{x}) = \prod_i R_{c_i}(F_{X_i}(x_i), \mathbf{F}_{\mathbf{pa_i}}(y_{\mathbf{pa_i}})) f_{X_i}(x_i)$$





Limitations of CBN Approach

- Uniqueness of Sklar's theorem is only guaranteed for continuous marginal distributions.

Limitations of CBN Approach

- Uniqueness of Sklar's theorem is only guaranteed for continuous marginal distributions.
- Computationally, Gaussian and Archimedean copula densities (and many other families) follow the grounded property. (i.e. $c(\mathbf{0}, v) = c(u, \mathbf{0}) = c(\mathbf{1}, v) = c(v, \mathbf{1}) = 0$).

Limitations of CBN Approach

- Uniqueness of Sklar's theorem is only guaranteed for continuous marginal distributions.
- Computationally, Gaussian and Archimedean copula densities (and many other families) follow the grounded property. (i.e. $c(\mathbf{0}, v) = c(u, \mathbf{0}) = c(\mathbf{1}, v) = c(v, \mathbf{1}) = 0$).
  - For discrete RV's, $F_{X_i}(x_{i_{max}}) = 1 \implies R_c = 0$.

Limitations of CBN Approach

- Uniqueness of Sklar's theorem is only guaranteed for continuous marginal distributions.
- Computationally, Gaussian and Archimedean copula densities (and many other families) follow the grounded property. (i.e. $c(\mathbf{0}, v) = c(u, \mathbf{0}) = c(\mathbf{1}, v) = c(v, \mathbf{1}) = 0$).
  - For discrete RV's, $F_{X_i}(x_{i_{max}}) = 1 \implies R_c = 0$.
  - $\therefore$ We can't directly apply CBNs to mixed data.

Limitations of CBN Approach

- Uniqueness of Sklar's theorem is only guaranteed for continuous marginal distributions.
- Computationally, Gaussian and Archimedean copula densities (and many other families) follow the grounded property. (i.e. $c(\mathbf{0}, v) = c(u, \mathbf{0}) = c(\mathbf{1}, v) = c(v, \mathbf{1}) = 0$).
  - For discrete RV's, $F_{X_i}(x_{i_{max}}) = 1 \implies R_c = 0$.
  - $\therefore$ We can't directly apply CBNs to mixed data.
- Current approaches for mixed networks
  - Conditional Linear Gaussian (CLG) Model
    - Reverts to assumptions of Gaussianity
    - Imposes restrictions on parent/child node variable types
  - Mixture of Truncated Exponentials (MTE) Model [MRS01]
    - Piecewise fitting of marginal distributions.

Hybrid Copulas

Can we use copulas to capture the dependency between arbitrary continuous and discrete random variables?

- Schweizer and Sklar's extension copula [SS74]
  - Denuit and Lambert's continuing transform, $X^* = X + (U - 1)$ [DL05]
  - Nešlehová's transform $X^* = \psi(X, U)$ [Ne7]
- de Leon and Wu's conditional distribution approach [dLW11]
- Smith and Khaled's MCMC latent variable approach [SK12]

Hybrid Copulas

Can we use copulas to capture the dependency between arbitrary continuous and discrete random variables?

- Schweizer and Sklar's extension copula [SS74]
  - Denuit and Lambert's continuing transform, $X^* = X + (U - 1)$ [DL05]
  - Nešlehová's transform $X^* = \psi(X, U)$ [Ne7]
- de Leon and Wu's conditional distribution approach [dLW11]
- Smith and Khaled's MCMC latent variable approach [SK12]

- Apply Nešlehová's transform $X^* = \psi(X, U)$ to discrete RVs.
- $\mathbf{X}^*$ corresponds to a unique copula, C^*.
- C^* captures and preserves the dependence and concordance properties of the underlying mixed vector $\mathbf{X} = (X_1, \ldots, X_n)$ [MQ10, Ne7].

Applicability of Hybrid Copulas

- Types of Discrete Random Variables

Applicability of Hybrid Copulas

- Types of Discrete Random Variables
  - Ordinal Random Variables
  - Count Random Variables
    - Have concept of dependency between events in sample space.

Applicability of Hybrid Copulas

- Types of Discrete Random Variables
  - Ordinal Random Variables
  - Count Random Variables
    - Have concept of dependency between events in sample space.
  - Categorical Random Variables
    - Do not have concept of dependency between events in sample space.

Applicability of Hybrid Copulas

- Types of Discrete Random Variables
  - Ordinal Random Variables
  - Count Random Variables
    - Have concept of dependency between events in sample space.
  - Categorical Random Variables
    - Do not have concept of dependency between events in sample space.

- Are categorical random variables allowed in this model?

Applicability of Hybrid Copulas

- Types of Discrete Random Variables
  - Ordinal Random Variables
  - Count Random Variables
    - Have concept of dependency between events in sample space.
  - Categorical Random Variables
    - Do not have concept of dependency between events in sample space.

- Are categorical random variables allowed in this model?
- Yes!
  - Convert categorical to ordinal arbitrarily to avoid defining conditional distributions.
  - Copulas can no longer be interpreted as dependence structures.

Hybrid Copula Bayesian Networks

Extend framework of CBN to incorporate discrete and continuous random variables.

- Construction
  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.
  3. Estimate structure of Bayesian network. [1]
  4. Estimate the copula density capturing the dependency between each node and its parents.

[1] Iterate between steps 3 and 4 in score based approach.

Copula Density Estimation

Copula family with all continuous nodes

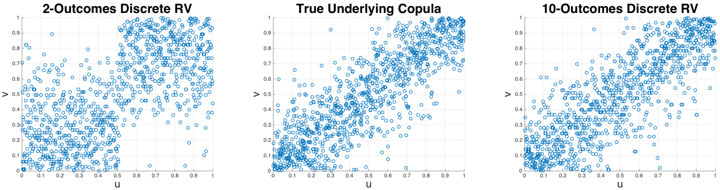
1. Use copula model selection algorithms to select copula family (Gaussian, Archimedean, etc...).
2. Estimate copula dependency θ parameter(s) by inverting $\tau = f(\theta)$.

Copula family with continuous and discrete nodes

1. Compute pseudosamples $\mathbf{U}^*$ from modified dataset $\mathbf{X}^*$.
2. Estimate copula density that captures underlying dependency properties of $\mathbf{X}$ using beta-kernels [CFS07].

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

Accuracy of Hybrid Copula Density Estimation



- As discrete outcomes increase, pseudo observations of transformed discrete random variables are closer to underlying copula's pseudo observations.
- Conversely, as discrete outcomes increase, CLG and MTE have to define an exponentially growing number of conditional distributions.
- Hybrid copulas recommended for large numbers of discrete outcomes. MTE recommended for smaller number.

HCBN Factorization

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

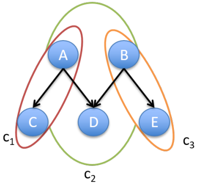
$$f_{\mathcal{X}}(\mathbf{x}) = \prod_{i=1}^{D} f_i(\mathbf{x_i})$$

$$u_{j,1} = F_{X_j}(x_j^-), u_{j,2} = F_{X_j}(x_j)$$

$$C_i^k = \frac{\partial^k}{\partial u_1 \partial u_2 \ldots \partial u_k} C_i(u_1, \ldots, u_n)$$
$$= \int_{k+1} \ldots \int_n c_i(\mathbf{u})$$

- $X_1, \ldots, X_k$ are continuous random variables.
- $X_{k+1}, \ldots, X_n$ are ordinal or count discrete random variables.
- i represents i^{th} family.

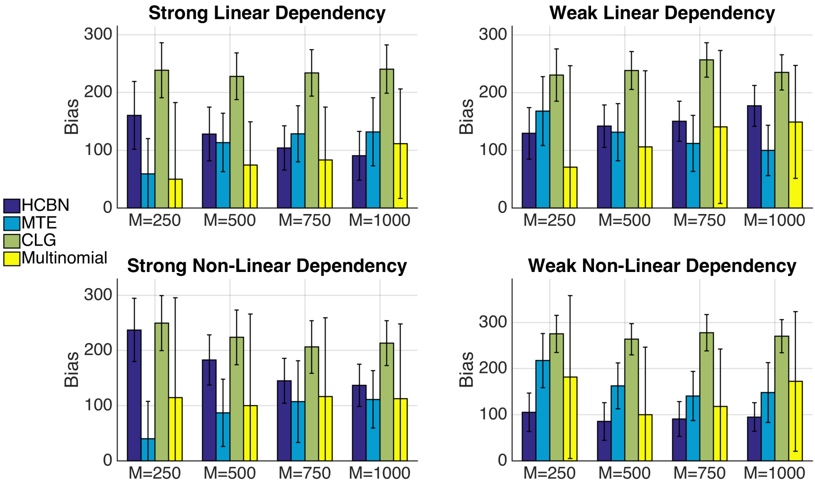
Experimental Evaluation - Synthetic Data Set Generation



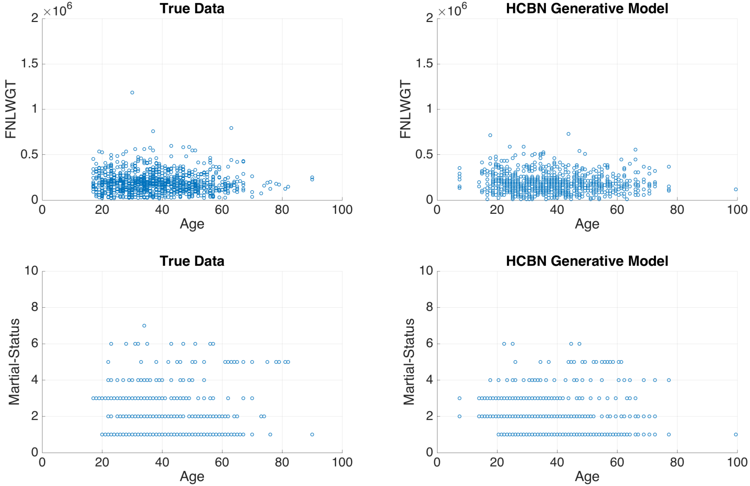
Nodes A/B Multinomial Probabilities	Nodes C/D/E PDF Types	Dependency Structure Models (C1/C2/C3)	Dependency Strengths for C1/C2/C3
[0.5 0.5]/[0.5 0.5]	N(2,0.5)/N(2,0.5)/N(2,0.5)	**Linear Dependency** Gaussian/ Gaussian/ Gaussian	**Strong**
	U(-2,2)/U(-2,2)/U(-2,2)		
[0.25 0.25 0.25 0.25]/ [0.25 0.25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

Synthetic BN MC Data Generation

Experimental Evaluation - Synthetic Data Set Results



Experimental Evaluation - Synthetic Data Set



Conclusion

- HCBN framework allows for expressive modeling of large discrete and continuous RV's.
- Performance compares favorably to both MTE and CLG models in synthetic data experiments.
- Good approach when there are high numbers of discrete outcomes.
- Future Work
  - Approximate Inference
  - Large scale structure learning taking advantage of copula theory.
  - Further experimentation with real-life datasets.
- Code available at `https://github.com/stochasticresearch/copula`
- Questions?

References I

- Arthur Charpentier, Jean-David Fermanian, and Olivier Scaillet, *The Estimation of Copulas: Theory and Practice*, Copulas: from theory to applications in finance (2007), 35–62.
- Michel Denuit and Philippe Lambert, *Constraints on Concordance Measures in Bivariate Discrete Data*, Journal of Multivariate Analysis (2005).
- A.R. de Leon and B. Wu, *Copula-based Regression Models for a Bivariate Mixed Discrete and Continuous Outcome*, Statistics in Medicine (2011).
- Gal Elidan, *Copula Bayesian Networks*, Advances in Neural Information Processing Systems 23, Curran Associates, Inc., 2010.

References II

- Mhamed Mesfioui and Jean-Franois Quessy, *Concordance Measures for Multivariate Non-Continuous Random Vectors*, Journal of Multivariate Analysis (2010).
- Serafn Moral, Rafael Rumi, and Antonio Salmern, *Mixtures of Truncated Exponentials in Hybrid Bayesian Networks*, Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer Berlin Heidelberg, 2001.
- Johanna Nešlehová, *On Rank Correlation Measures for Non-Continuous Random Variables*, Journal of Multivariate Analysis (2007).
- Roger B. Nelsen, *An introduction to copulas*, Springer-Verlag New York, 2006.

References III

- Michael S. Smith and Mohamad A. Khaled, *Estimation of Copula Models with Discrete Margins via Bayesian Data Augmentation*, Journal of the American Statistical Association (2012).
- B Schweizer and Abe Sklar, *On Nonparametric Measures of Dependence for Random Variables*, Studia Mathematica (1974).