Latent Tree Copulas

Sergey Kirshner
Department of Statistics, Purdue University, West Lafayette, IN, USA
skirshne@purdue.edu

Learning Multivariate Density from Data

\[ D \] – a complete real-valued i.i.d. data sets

\[ N \] – number of examples (rows)

\[ d \] – number of dimensions (columns)

Want to fit a pdf \( p(x) \) to \( D \). Difficult because

- Non-parametric methods (e.g., kernel density estimators, or KDE) need the number of samples exponential in \( d \) (curse of dimensionality).

- Few families have a canonical multivariate form (e.g., Gaussian, t-distribution), and these families may not be supported by the data.

Univariate marginals – easy to optimize, not subject to curse of dimensionality

\[
\ln p(D) = \sum_{n=1}^{N} \sum_{i=1}^{d} \ln p(x_{n}^{i}) + \sum_{n=1}^{N} \sum_{i=1}^{d} \frac{p(a^{n})}{\prod_{j=1}^{d} p(x_{n}^{j})}
\]

Tree-Structured Copula

\[ \pi^{(s)}(\varepsilon) = \prod_{(u,v) \in \varepsilon} \pi(u,v) \]

Needs only bivariate copulas to be specified!

Latent Tree Copula

Latent tree copulas are an extension of latent tree models to continuous variables.

Bottom-up Structure Learning

Similar to Bin-G (Harremoëls and Williams 11)

- Start with trees consisting of individual variables
- Repeat until only one subtree is left
- Estimate posterior bivariate mutual informations for all pairs of root nodes
- Join two subtrees with the highest mutual information
- Re-estimate the parameters for the new trees

Inference requires integrating out latent variables: easy for Gaussian copulas; hard for all others.

Can use variational approach

- approximation the posterior distribution using a tree-structured distribution over piece-wise uniform variables
- Essentially, approximate using the tree over categorical variables

- Requires numerical integration of double-integrals

Parameter Estimation

Gaussian case: parameter estimation using EM:

- E-step: closed form inference, \( O(N) \) per iteration
- M-step: closed form maximization, \( O(NE) \) per iteration (need to estimate parameters for edges
- Log-likelihood increases at each iteration

Non-Gaussian case: parameter estimation using variational EM:

- E-step: approximate inference, \( O(NE|\lambda|) + |\lambda| \) bivariate integrals per iteration
- M-step: approximate maximization, need to update \( |\lambda| \) bivariate copula parameters
- Lower-bound on log-likelihood increases at each iteration

Unknown Structure

- Gaussian LTCs: same as for tree-structured Gaussians (size of possible trees can be limited)
- Non-Gaussian LTCs: need to restrict the space of possible models (very large space of structures/copula families):
  - Fix the bivariate copula family
  - Consider only binary latent tree copulas

Copula

What is a copula? Copula is a multivariate distribution (cdf) defined on \([0,1]^d\) where univariate marginal distributions for all variables are uniform.

\[
F(x) = C(F_1(x_1), \ldots, F_d(x_d)), \quad x \in \mathbb{R}^d
\]

\[
C(a) = \Phi^d(\Phi^{-1}(a_1), \ldots, \Phi^{-1}(a_d))
\]

\[
\pi(a) = \prod_{i=1}^{d} \pi_i(a_i)
\]

Distribution = Marginals + Copula

Application: Regionalization According to Rainfall

State of Indiana (USA)

15 stations, average rainfall, 12 months X 47 years (1951-1996)

Used Gaussian KDEs to fit the marginals. Recovered LTC structure appears consistent with the geography.

Gaussian Copula Example

\[
F(x, y) = \Phi^2\left(\frac{x - \mu_x}{\sigma_x}, \frac{y - \mu_y}{\sigma_y}\right),
\]

\[
C(a, b) = \Phi(\Phi^{-1}(a_1) - \Phi^{-1}(a_2), \Phi^{-1}(b_1) - \Phi^{-1}(b_2))
\]

\[
\pi(a, b) = \pi_1(a_1, b_1) \pi_2(a_2, b_2)
\]

Illustration: MAGiC Gamma Telescope Data Set from UCI ML Repository

12000 10-dimensional examples (signal), 10 random partitions into training set (10000) and test set (2000).

Training set sizes (from 10000): 50, 100, 200, 500, 1000, 2000, 5000, 10000. For copula models, marginals are estimated using Gaussian KDEs.