Latent Tree Copulas

Sergey Kirshner

Department of Statistics, Purdue University, West Lafayette, IN, USA

skirshne@purdue.edu

Learning Multivariate Density from Data

$\mathcal{D}$ – a complete real-valued i.i.d. data sets

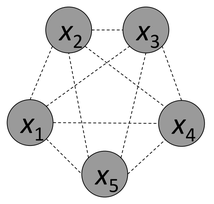
N – number of examples (rows)
d – number of dimensions (columns)

Want to fit a pdf ***p(x)*** to $\mathcal{D}$. 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.

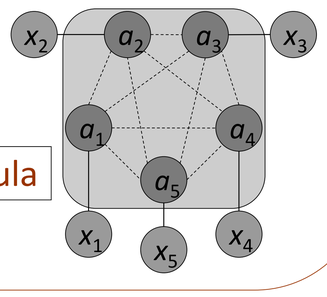
Copula

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



$$F(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d)), \ \boldsymbol{x} \in \mathbb{R}^d$$
$$C(\boldsymbol{a}) = F\left(F_1^{-1}(a_1), \ldots, F_d^{-1}(a_d)\right)$$
$$a_i = F_i(x_i), \quad x_i = F_i^{-1}(a_i), \quad i = 1, \ldots, d$$

$$c(\boldsymbol{a}) = \frac{\partial^d C(\boldsymbol{a})}{\partial a_1 \ldots \partial a_d} = \frac{p(\boldsymbol{x})}{\prod_{i=1}^d p_i(x_i)}$$



Distribution = Marginals + Copula

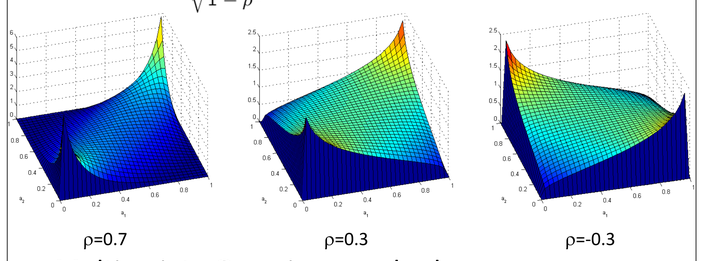
Gaussian Copula Example

$$F(x_1, x_2) = \mathcal{N}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \rho\sigma_{11}\sigma_{22} \\ \rho\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{pmatrix}\right) = \Phi_\rho\left(\frac{x_1 - \mu_1}{\sigma_{11}}, \frac{x_2 - \mu_2}{\sigma_{22}}\right)$$

$$a_1 = F_1(x_1) = \Phi\left(\frac{x_1 - \mu_1}{\sigma_{11}}\right), \quad a_2 = F_2(x_2) = \Phi\left(\frac{x_2 - \mu_2}{\sigma_{22}}\right)$$

$$C_{\mathcal{N}}(a_1, a_2) = \Phi_\rho\left(\Phi^{-1}(a_1), \Phi^{-1}(a_2)\right)$$

$$c_{\mathcal{N}}(a_1, a_2) = \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{\rho^2\Phi^{-1}(a_1)^2 + \rho^2\Phi^{-1}(a_2)^2 - 2\rho\Phi^{-1}(a_1)\Phi^{-1}(a+2)}{2(1-\rho^2)}}$$



ρ=0.7 ρ=0.3 ρ=-0.3

Multivariate Gaussian copula shares many properties with multivariate normal distribution:

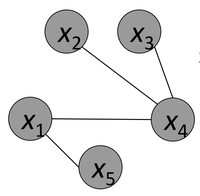
- generalizes to d dimensions,
- marginal and conditional copulas are also normal.

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

$$\ln p(\mathcal{D}) = \sum_{n=1}^N \sum_{i=1}^d \ln p(x_i^n) + \sum_{n=1}^N \ln \frac{p(\boldsymbol{x}^n)}{\prod_{i=1}^d p(x_i^n)}$$

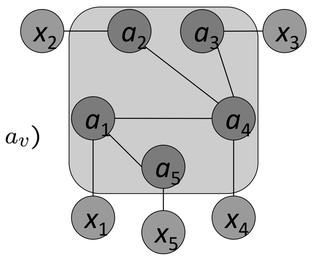
1. Estimate the marginals $p_i(x_i)$.
2. Replace each feature x_i^n with corresponding probability value $a_i^n = F_i(x_i^n)$.
3. Estimate the parameters of the copula density *c* by maximizing the log-likelihood of the transformed data $A = \{a^1, \ldots, a^N\}$.

Tree-Structured Copula



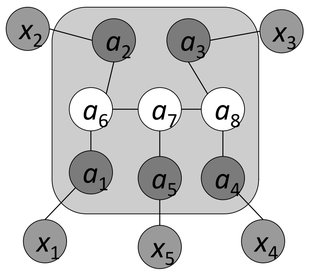
$$p(\boldsymbol{x}) = \left[\prod_{v \in \mathcal{V}} p_v(x_v)\right] \left[\prod_{\{u,v\} \in \mathcal{E}} \frac{p_{uv}(x_u, x_v)}{p_u(x_u)\, p_v(x_v)}\right]$$

$$c_{\mathcal{E}}(\boldsymbol{a}) = \prod_{\{u,v\} \in \mathcal{E}} c_{uv}(a_u, a_v)$$



Needs only bivariate copulas to be specified!

Latent Tree Copula



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

$$c_{LT}(\boldsymbol{a}) = \iint_{\mathbb{R}^{t-d}} \prod_{\{u,v\} \in \mathcal{E}} c_{uv}(a_u, a_v)\, \mathrm{d}a_{d+1} \ldots \mathrm{d}a_t$$

Inference

Inference requires integrating out latent variables:

- easy for Gaussian copulas;
- hard for all other copulas.

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

$$\ln c_{LT}\left(\boldsymbol{a}_O^n | \boldsymbol{\theta}'\right) = \int_{\mathbb{I}^{t-d}} q^n(\boldsymbol{a}_H^n) \ln \frac{c_{LT}\left(\boldsymbol{a}_O^n, \boldsymbol{a}_H^n | \boldsymbol{\theta}'\right)}{q^n\left(\boldsymbol{a}_H^n\right)} \mathrm{d}\boldsymbol{a}_H^n + D\left(q^n(\boldsymbol{a}_H^n) \,\|\, c_{LT}\left(\boldsymbol{a}_H^n | \boldsymbol{a}_O^n, \boldsymbol{\theta}'\right)\right);$$

$$q(\boldsymbol{a}_H) = \prod_{u \in H} q_u(a_u) \left[\prod_{\{u,v\} \in \mathcal{E}_H} \frac{q_{uv}(a_u, a_v)}{q_u(a_u)\, q_v(a_v)}\right]$$

$$q_{uv}(a_u, a_v) = p_{uv}(i, j) \geq 0 \text{ for } a_u \in \mathbb{I}_i, \ a_v \in \mathbb{I}_j,$$

$$q_u(a_u) = p_u(i) \text{ for } a_u \in \mathbb{I}_i,$$

where $\mathbb{I}_i = \left(\frac{i-1}{K}, \frac{i}{K}\right]$.

Parameter Estimation

Gaussian case: parameter estimation using EM:

- E-step: closed form inference, $O(Nt)$ per iteration
- M-step: closed form maximization, $O(N|E|)$ 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, $O(sN|E|k^2) + |E|$ bivariate integrals per iteration
- M-step: approximate maximization, need to update $|E|$ 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

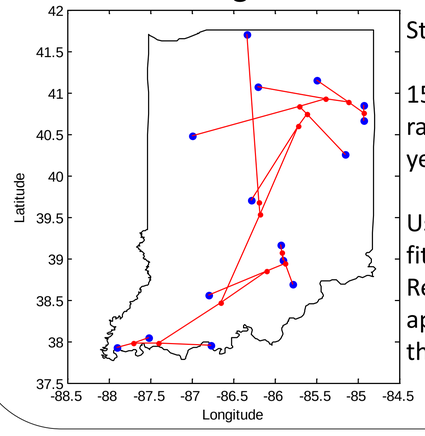
Bottom-up Structure Learning

Similar to Bin-G [Harmeling 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
- Create a new latent node and join it to the roots of the two subtrees
- Re-estimate the parameters for the new trees

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

Application: Regionalization According to Rainfall



State of Indiana (USA)

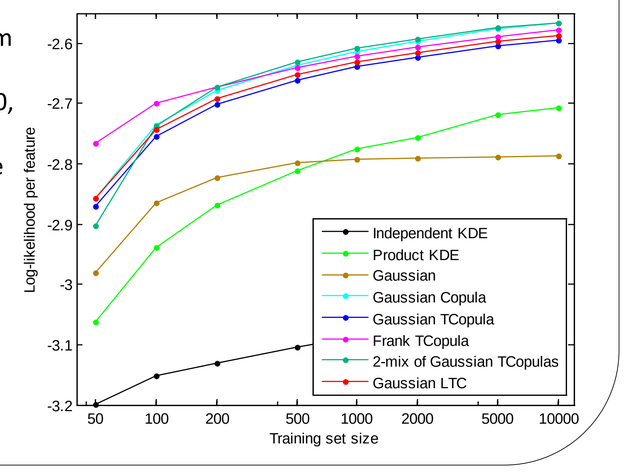
15 stations, average rainfall, 12 months X 47 years (1951-1996)

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

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.



Research Supported by US NSF Award AGS-1025430