Latent Tree Copulas

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

Granada, Spain,
September 19, 2012
Coming Attractions

• Want to fit density to model multivariate data?
  - and organize real-valued data into a hierarchy of features?

• New density estimation model based on tree-structured dependence with latent variables
  - Distribution = Univariate Marginals + Copula
  - Hierarchy of variables as a latent tree-copula
  - Parameter estimation and structure learning
    - Efficient inference for Gaussian copulas (100s of variables), several structure learning approaches
    - Variational inference for other copulas (10-30 variables)
Building a Hierarchy of Rainfall Stations

State of Indiana (USA)

Average monthly observations for 15 rainfall stations 1951-1996 (47 years)
Most Popular Distribution...

- Interpretable
- Closed under taking marginals
- Generalizes to multiple dimensions
- Models pairwise dependence
- Tractable

- 245 pages out of 691 from *Continuous Multivariate Distributions* by Kotz, Balakrishnan, and Johnson
What If the Data Is NOT Gaussian?
Separating Univariate Marginals

\[ \ln p(x) = \sum_{i=1}^{d} \ln p_i(x_i) + \ln \frac{p(x)}{p_1(x_1) \cdots p_d(x_d)} \]

- Univariate marginals, independent variables,
- Multivariate dependence term, copula
Monotonic Transformation of the Variables
Copula

**Copula** $C$ is a multivariate distribution (cdf) defined on a unit hypercube with uniform univariate marginals:

$$C : [0, 1]^d \rightarrow [0, 1]$$

$$C_i (a_i) = a_i, \; i = 1, \ldots, d$$

$$F (\mathbf{x}) = C \left( F_1 (x_1), \ldots, F_d (x_d) \right), \; \mathbf{x} \in \mathbb{R}^d$$

$$C (\mathbf{a}) = F \left( F_1^{-1} (a_1), \ldots, F_d^{-1} (a_d) \right)$$

$$a_i = F_i (x_i), \; i = 1, \ldots, d$$

$$x_i = F_i^{-1} (a_i), \; i = 1, \ldots, d$$

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

\[ \text{[Sklar 59]} \]
Example: Multivariate Gaussian Copula

\[ F(x) = \mathcal{N}(x; \mu, \Sigma) \]

\[ \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11}^2 & \cdots & \sigma_{11}\sigma_{dd}\rho_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{11}\sigma_{dd}\rho_{d1} & \cdots & \sigma_{dd}^2 \end{pmatrix} \]

\[ F(x_1, \ldots, x_d) = \Phi_R \left( \frac{x_1 - \mu_1}{\sigma_{11}}, \ldots, \frac{x_d - \mu_d}{\sigma_{dd}} \right) \]

\[ a_1 = F_1(x_1) = \Phi \left( \frac{x_1 - \mu_1}{\sigma_{11}} \right), \ldots, a_d = F_d(x_d) = \Phi \left( \frac{x_d - \mu_d}{\sigma_{dd}} \right) \]

\[ C(a) = F\left( F_1^{-1}(a_1), \ldots, F_d^{-1}(a_d) \right) = \Phi_R \left( \Phi^{-1}(a_1), \ldots, \Phi^{-1}(a_d) \right) \]
Separating Univariate Marginals

\[
\ln p(x) = \sum_{i=1}^{d} \ln p_i(x_i) + \ln \frac{p(x)}{p_1(x_1) \cdots p_d(x_d)}
\]

\[
\ln p(D) = \sum_{n=1}^{N} \sum_{i=1}^{d} \ln p_i(x_i^n) + \sum_{n=1}^{N} \ln c(F_1(x_1^n), \ldots, F_d(x_d^n))
\]

1. Fit univariate marginals (parametric or non-parametric)

2. Replace data points with cdf’s of the marginals

3. Estimate copula density

Inference for the margins [Joe and Xu 96]; canonical maximum likelihood [Genest et al 95]
Graphical Model Using a Copula

\[ f(x_1, \ldots, x_d) = \left[ \prod_{i=1}^{d} f_i(x_i) \right] c\left(F_1(x_1), \ldots, F_d(x_d)\right) \]

\[ = \left[ \prod_{i=1}^{d} \phi_i(x_i, a_i) \right] c(a_1, \ldots, a_d) \]
Graphical Model Approaches to Estimating Copulas

- **Vines** [Bedford and Cooke 02]
- **Trees** [Kirshner 08]
- **Nonparanormal model** [Liu et al. 09]
- **Copula Bayesian networks** [Elidan 10]
Tree-Structured Densities

\[ p_\mathcal{E}(x) = \left[ \prod_{v \in \mathcal{V}} p_v(x_v) \right] \left[ \prod_{\{u,v\} \in \mathcal{E}} \frac{p_{uv}(x_u, x_v)}{p_u(x_u) p_v(x_v)} \right] \]
Tree-Structured Copulas

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

[Kirshner 08]
Using Tree-Structured Copulas

• Tree-structured copulas are convenient, but are restrictive
  – True distribution may require much larger cliques to decompose

• Can approximate other dependencies using latent variables
  – Mixtures [Kirchner 08]: discrete latent variables
  – Latent tree copulas: continuous random variables embedded in copula trees
Latent Tree Copulas

- Defined as a tuple of variables, tree structure, and bivariate copulas

\[
c_{LT}(a) = \int \int \prod_{t-d \in E} c_{uv}(a_u, a_v) \, da_{d+1} \ldots da_t
\]
Latent Tree Copulas

- Defined as a tuple of variables, tree structure, and bivariate copulas

- “Siblings” of latent tree models (LTMs) for categorical variables [e.g., Zhang 02, 04]
Inference

• **Good news:** posterior distribution is also tree-structured
  - Fairly easy to carry out inference for LTMs

• **Bad news:** Latent variables are continuous: infinite number of possible values
  - Need to estimate the joint posterior densities
Inference

- Easy for Gaussian copulas
  - Apply inverse standard normal CDF; use belief propagation on jointly Gaussian distribution

- Difficult for non-Gaussian copulas
  - May have no exact form for the posterior!
Inference for non-Gaussian Case

• Variational approach:
  - Approximate the posterior distribution using a tree-structured distribution over piece-wise uniform variables
  - Essentially, approximate using the tree over categorical variables

\[
q(a_H) = \prod_{u \in H} q_u(a_u) \left[ \prod_{\{u,v\} \in E_H} \frac{q_{uv}(a_u, a_v)}{q_u(a_u) q_v(a_v)} \right]
\]

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

\[
q_u(a_u) = p_u(i) \text{ for } a_u \in \mathbb{I}_i, \quad \text{where } \mathbb{I}_i = \left[ \frac{i-1}{K}, \frac{i}{K} \right]
\]

\[
\arg\min_{q \in Q} D \left( q^n(a_H^n) \mid\mid c_{LT} \left( a_H^n | a_O^n, \theta' \right) \right)
\]
Parameter Estimation with Known Structure

- (Variational) EM
  - E-step: minimize KL divergence
  - M-step: maximize the expected compete-data log-likelihood

\[
l (\theta) = \sum_{i=1}^{n} \int_{I_{t-d}} q^n (a^n_H) \ln \frac{c_{LT} (a^n_O, a^n_H | \theta')}{{q^n} (a^n_H)} \, da^n_H \\
+ \sum_{i=1}^{n} D \left( q^n (a^n_H) \parallel c_{LT} \left( a^n_H | a^n_O, \theta' \right) \right)
\]
Parameter Estimation with Known Structure

- Gaussian copula case: EM  
  \[ q^n(a^n_H) = c(a^n_H | a^n_O | \theta) \]
  - E-step: closed form inference, \( O(NT) \) per iteration
  - M-step: maximize the expected complete-data log-likelihood

\[
l(\theta) = \sum_{i=1}^{n} \int_{I_{t-d}} q^n(a^n_H) \ln \frac{c_{LT}(a^n_O, a^n_H | \theta')}{q^n(a^n_H)} \, da^n_H \]
\[
+ \sum_{i=1}^{n} D \left( q^n(a^n_H) \parallel c_{LT} \left( a^n_H | a^n_O, \theta' \right) \right)\]
Parameter Estimation with Known Structure

- Non-gaussian copula case: 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

$$ l(\theta) = \sum_{i=1}^{n} \int_{t-d} q^n (a^n_H) \ln \frac{c_{LT} (a^n_O, a^n_H | \theta')} {q^n (a^n_H)} da^n_H $$

$$ + \sum_{i=1}^{n} D (q^n (a^n_H) \| c_{LT} (a^n_H | a^n_O, \theta')) $$
Unknown Structure

• Gaussian LTCs: same as for tree-structured Gaussians
  - Size of possible trees can be limited
  - e.g., can use information distances [Choi et al 2011]

• 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
    - Observed nodes = leaves
    - Motivation: Any Gaussian LTC is equivalent to some binary LTC
Bottom-up Binary LTC Learning

[Similar to Harmeling and Williams 11]

- Initialize the subtrees to consist of individual variables (variable = root of a subtree)
- Iterate until all variables are in one tree
  - Estimate mutual informations (MIs) between the root nodes
  - Pick the pair of roots with the largest MI
  - Merge the subtrees by creating a new latent root node
  - Re-estimate parameters (EM)
Bottom-up Binary LTC Learning

[Similar to Harmeling and Williams 11]
Illustration for Building of Hierarchy of Rainfall Stations

State of Indiana (USA)

Average monthly observations for 15 rainfall stations 1951-1996 (47 years)
Experiments: Log-Likelihood on Test Data

UCI ML Repository
MAGIC data set
12000 10-dimensional vectors
2000 examples in test sets
Average over 10 partitions
Summary

• Multivariate distribution = univariate marginals + copula

• New model: tree-structured multivariate distribution with marginally uniform latent variables (latent tree copula, LTC)
  – Sufficient to employ only bivariate copula families!

• Closed form inference for Gaussian copulas (efficient)

• Variational inference for non-Gaussian copulas (slow)

• Parameter estimation using the EM algorithm

• Bottom-up structure learning for bivariate LTCs

• Can be used for parsimonious multivariate density estimation or to structure variables into hierarchies
Thank you!

Software:

http://www.stat.purdue.edu/~skirshne/LTC/index.html

Support:

US National Science Foundation Award AGS-1025430

Questions?

See me at the poster tonight for more details