GAUSSIAN COPULA MODELS
UBC Machine Learning Group

July 20th, 2016

Steven Spielberg Pon Kumar, Tingke (Kevin) Shen
University of British Columbia
Overview

1. Motivating example
2. UGM and Gaussian graphical
3. Copula model
4. Copula inference
5. Case Study
6. Closing remarks
The Copula model is a joint probability distribution...
Motivating example
A Motivating Example
A Motivating Example

SELL!

SELL!
A Motivating Example
UGM and Gaussian graphical
Graph with nodes $V$ and edges $E$. 

$$G = (V, E)$$

$$p(x) \sim \prod_{j=1}^{d} \phi_j(x_j) \prod_{(i,j) \in E} \phi_{ij}(x_i, x_j)$$
UGM and Multivariate Gaussian

\[ p(x) \sim \exp\left(-\frac{1}{2}(x - \mu)^T \sum^{-1}(x - \mu)\right) \]

\[ p(x) \sim \left( \prod_{i=1}^{d} \prod_{j=1}^{d} \frac{\exp\left(-\frac{1}{2}x_i x_j \Sigma^{-1}_{ij}\right)}{\phi_{ij}(x_i, x_j)} \right) \left( \prod_{i=1}^{d} \frac{\exp(x_i v_i)}{\phi_i(x_i)} \right) \]

- Pair-wise Markov property holds iff \( \Sigma^{-1}_{v1, v2} = 0 \)
- Edges of G correspond with off-diagonol non-zero elements
Advantages of Gaussian Graphical Model:
- Covariance matrix conjugate with G-Wishart prior.
- Relatively easy to sample.
- Overall cheap and simple.

Disadvantages of Gaussian Graphical Model:
- Unimodal joint distribution
- Marginals are Gaussian
- Random variables must be continuous
Limitations of M. Gaussian and Motivation for Copula
Limitations of M. Gaussian and Motivation for Copula
Limitations of M. Gaussian and Motivation for Copula

Disadvantages of Gaussian Graphical Model:
- Uni-modal joint distribution
- Marginals are Gaussian
- Random variables must be continuous

Solved by Copula Model:
- Multi-modal joint distribution
- Marginals can be arbitrary functions
- Both discrete and continuous variables
Copula model
The Copula Model

If we have $d$ random variables and we want to satisfy the following conditions:

- Marginals can be arbitrary functions
- Both discrete and continuous variables

Then what is the natural way to combine the random variables into a joint distribution?

Answer: use their CDF’s
Mapping the CDF

In order to allow continuous and discrete variables to "communicate," we consider a joint distribution as a function of marginal CDF’s.

\[ F(F_1(x_1), F_2(x_2), ..., F_d(x_d)) \]

But working in CDF space is not nice.

Idea: we map the marginals CDF’s back into a latent variable.

\[ F(\phi^{-1}[F_1(x_1)], \phi^{-1}[F_2(x_2)], ..., \phi^{-1}[F_d(x_d)]) \]
Figure: Mapping from observed to latent variable via CDF. Multimodal to unimodal.
Figure: Mapping from observed to latent variable via CDF. Multimodal to unimodal.
Figure: Mapping from observed to latent variable via CDF. Multimodal to unimodal.
We’ve been talking about mapping the marginal of $x$ to a latent variable but do we know the marginals?

Yes! Given a set of data, we can approximate marginals.
Gaussian Copula

Notation:

- \( \phi(x) \) - standard normal density (PDF)
- \( \Phi(x) \) - standard normal Cumulative Distribution Function (CDF)
- \( \Phi^{-1}(x) \) - Inverse CDF
- Latent random variable \( Z \)
- CDF \( F_1(x_1) = \Phi(z_1) \)
- PDF \( f_1(x_1) = \frac{1}{\sigma_1} \phi(z_1) \)
Gaussian Copula

For any multivariate distribution, with CDF $F$ and marginal CDF’s $F_i$, copula $C$ is such distribution on $[0, 1]^d$ s.t.

$$F(x_1, x_2 \ldots, x_d) = C(F_1(x_1), \ldots, F_1(x_d))$$

$$= C(\phi^{-1}[F_1(x_1)], \phi^{-1}[F_2(x_2)], \ldots, \phi^{-1}[F_d(x_d)])$$

$$= C(z_1, z_2, \ldots, z_d)$$

$$= \Phi_d(z_1, z_2, \ldots, z_d)$$  (1)

We picked $\phi$ and $\Phi_d$ to be Gaussian but they could be Student-t, Laplace, etc.
Gaussian Copula

**CDF**

\[ F(x) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \]

**PDF**

\[ f(x) = c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \prod_{i=1}^{d} f_i(x_i) \]

where \( f_i(x_i) \) is the marginal PDF.

**Copula density** \( c \) is defined by:

\[ c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) = \frac{\partial^d C}{\partial F_1 \ldots \partial F_d} \]
2-D case

\[ f(x, y) = \frac{\partial^2 C(F_x(x), F_y(y))}{\partial X \partial Y} \]

\[ = \frac{\partial}{\partial X} \left( \frac{\partial}{\partial Y} \left( C(F_x(x), F_x(y)) \right) \right) \]

\[ = \frac{\partial}{\partial X} \left( \frac{\partial C}{\partial F_y} \frac{dF_y}{dy} \right) \]

\[ = \frac{\partial^2 C}{\partial F_x \partial F_y} \cdot \frac{dF_x}{dX} \frac{dF_y}{dY} \]

\[ = \text{copula density } \times \text{ product of marginal pdf} \]
Gaussian Copula

PDF can be written with a correlation matrix $K$:

$$f(x) = \frac{1}{|K|^\frac{1}{2}} \exp\left\{-\frac{1}{2} z(K^{-1} - I)z^T\right\} \prod_{i=1}^{d} \frac{1}{\sigma_i} \phi(z_i)$$

where

$$z_i = \Phi^{-1} [F_i(x_i)]$$

Density of copula:

$$c(x) = \frac{1}{|K|^\frac{1}{2}} \exp\left(-\frac{1}{2} z(K^{-1} - I)z^T\right)$$
Special Case: Uniform Correlation Structure

\[ K = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}, \quad \rho \in \left(\frac{-1}{d-1}, 1\right] \]

Solving for \( K^{-1} \) and \(|K|\),

\[ c(x) = k_1(\rho, d) \times \exp\left\{ k_2(\rho, d) \left( (d - 1) \rho \sum_{i=1}^{d} z_i^2 - 2 \sum_{j=1}^{d} \sum_{i<j} z_i z_j \right) \right\} \]
Special Case: Serial Correlation Structure

\[
K = \begin{pmatrix}
1 & \rho & \ldots & \rho^{d-1} \\
\rho & 1 & \ldots & \rho^{d-2} \\
\vdots & \vdots & \ddots & \vdots \\
\rho^{d-1} & \rho^{d-2} & \ldots & 1
\end{pmatrix}, \quad \rho \in \left(\frac{-1}{d-1}, 1\right)
\]

Solving for \(K^{-1}\) and \(|K|\),

\[
c(x) = k_3(\rho, d) \ast \exp \left\{ k_4(\rho, d) \left( 2\rho \sum_{i=1}^{d} z_i^2 - \rho (z_1^2 + z_d^2) - 2 \sum_{i=1}^{d-1} z_i z_{i+1} \right) \right\}
\]
Copula inference
Given $n$ points of $d$ dimensional data $x^{1:n}$, we would like to find the relationship between pairs of random variables.

$$G = (V, E) \Rightarrow \{ K | K_{ij} = 0 \text{ if } (i, j) \notin E \}$$

$$P(Z_i \perp Z_j | Z_{-i-j}) = 1 - \frac{1}{T} \sum_{t=1}^{T} I_{ij}(G^t)$$
Markov properties associated with UGM for Z translate into Markov properties for X [proof omitted]:

\[
P(X_i \perp X_j | X_{-i-j}) = P(Z_i \perp Z_j | Z_{-i-j}) = 1 - \frac{1}{T} \sum_{t=1}^{T} I_{ij}(G^t)
\]

\[
I_{ij}(G) = \begin{cases} 
1, & \text{if } (i, j) \in E \\
0, & \text{otherwise}
\end{cases}
\]
Inference can be done independent of marginals

Given \( x^{(1:n)} \), any set of marginal CDF’s will obey the following constraint A on \( z^{(1:n)} \):

\[
A(x^{(1:n)}) = [l_v^i < z_v^i < u_v^i : 1 \leq i \leq n, 1 \leq v \leq d]
\]

\[
l_v^i = \max \{z_v^k : x_v^k < x_v^i\}, \quad u_v^i = \min \{z_v^k : x_v^i < x_v^k\}
\]

If \( z^{(1:n)} \) obey constraint A, no need for marginals.
Inference can be done independent of marginals

Idea: Only order of $z_i$ matter because choosing $F_i$ is simply choosing a way to "connect-the-dots" in marginal CDF’s of $x$. 
G be a graph defining a gaussian graphical model for the latent variables $Z_v$

Joint posterior distribution of $K$, the latent data $z^{(1:n)}$ and the Graph is,

$$p(K, z^{1:n}, G|C) \propto p(z^{1:n}|K, C) \times p(K|G) \times p(G)$$

$C$ is the event that $z^{(1:n)}$ obeys constraint $A(x^{(1:n)})$

Joint distribution is not defined if $K \notin P_G$. $P_G$ is the set of symmetric, positive, definite matrices "obeying" graph $G$
Since joint distribution is not defined for $K \not\in P_G$, construct Gibbs sampling algorithm for the marginal:

$$p(z^{(1:n)}, G|C) = \int_{K \in P_G} p(K, z^{(1:n)}, G|C) dK$$
Gibbs sampling Algorithm

We have a joint density,

\[ f(x, y_1, \ldots, y_k) \]

and we are interested in the marginal density,

\[ f(x) = \int \int \ldots \int f(x, y_1, \ldots, y_k) dy_1, dy_2, \ldots dy_k \]

Assume we can sample the \( k + 1 \)-many univariate conditional densities:

\[
\begin{align*}
    f(X|y_1\ldots, y_k) \\
    f(Y_1|x, y_2\ldots, y_k) \\
    f(Y_2|x, y_1, y_3\ldots, y_k) \\
    \vdots \\
    f(Y_k|x, y_1, y_3\ldots, y_{k-1})
\end{align*}
\]
Choose arbitrarily, \( k \) initial values: \( Y_1 = y_1^0, Y_2 = y_2^0, Y_3 = y_3^0 \ldots Y_k = y_k^0 \)

\[ x^1 \text{ by a draw from } f(X|y_1^0, \ldots, y_k^0) \]
\[ y_1^1 \text{ by a draw from } f(Y_1|x^1, y_2^0, \ldots, y_k^0) \]
\[ y_2^1 \text{ by a draw from } f(Y_2|x^1, y_1^0, y_3^0, \ldots, y_k^0) \]
\[ \ldots \]
\[ y_k^1 \text{ by a draw from } f(Y_k|x^1, y_1^1, \ldots, y_{k-1}^1) \]

This constitutes one Gibbs "pass" through \( k+1 \) conditional distributions, yielding samples: \((x^1, y_1^1, y_2^1, \ldots y_k^1), (x^2, y_1^2, y_2^2, \ldots y_k^2)\)...
The average of the conditional densities \( f(X|y_1, \ldots y_k) \) will be a close approximation to \( f(X) \)
Inference Algorithm

We would like to solve for the marginal:

\[ p(z^{(1:n)}, G|C) = \int_{K \in P_G} p(K, z^{(1:n)}, G|C) dK \]

- Initialize variables to \( G^0, K^0, z^{(1:n)} \) where \( z^{(1:n)} \) obeys \( C \)
- Sample \( G \) using Metropolis-Hastings
- Sample \( K \) using block Gibbs-sampling
- Sample \( z^{(1:n)} \) using Gibbs-sampling
- Repeat for \( T \) iterations
Inference Algorithm: Metropolis-Hasting

1. Sample $G$ from the conditional
   \[ p(G|z^{(1:n)}, C) = p(G|z^{(1:n)}) \propto p(z^{(1:n)}|G)p(G) \]
2. Propose $G^{\text{new}}$ where $G^{\text{new}} \in \text{nbd}(G)^*$
3. Generate $u$ from $U(0,1)$
4. Move to $G^{\text{new}}$ if
   \[ u < \frac{p(G^{\text{new}}|z^{(1:n)})p(G|G^{\text{new}})}{p(G|z^{(1:n)})p(G^{\text{new}}|G)} \]

* $\text{nbd}(G)$ are all the graphs $G^*$ s.t. $G^*$ differs from $G$ by the addition or subtraction of one edge.

- $p(G^{\text{new}}|G)$ is the proposal function and is chosen.
Inference Algorithm:

\[ p(G|z^{(1:n)}) = \frac{p(z^{(1:n)}|G)p(G)}{p(z^{(1:n)})} \]

- We need to solve for \( p(z^{(1:n)}|G) \), the marginal likelihood
- Solving the numerator gives estimate for \( p(z^{(1:n)}, G|C) \)
We need to solve for $p(z^{(1:n)}|G)$, the marginal likelihood

- $Z$ is Gaussian with G-Wishart prior

$$p(K) = \frac{1}{I_G(\delta, D)} |K|^{\delta-2/2} \exp\left(\frac{-1}{2} \langle K, D \rangle \right)$$

where $\langle K, D \rangle = tr(K^T D)$ is the trace inner product.
The marginal is the ratio of normalizing constants

\[ p(z^{(1:n)}|G) = \frac{I_G(\delta + n, D + U)}{I_G(\delta, D)} \]

where \( U = \sum_{i=1}^{n} (z^i)^T(z^i) \)

- If \( G \) is decomposable, then this can be solved explicitly, else use numerical integration
- Laplace approximation, other methods
Inference Algorithm: Block-Gibbs

We sample $K$ from the posterior:

$$p(K|G, z^{(1:n)}, C) = p(K|G, z^{(1:n)})$$

Again exploit the conjugacy of Gaussians:

$$p(K|G, z^{(1:n)}) = W_G(\delta + n, D + \sum_{i=1}^{n} (z^i)^T z^i)$$
1. Choose some block $b$ from $K$

2. Set $K_{-b}^{t+1} = K_{-b}^{t}$ and sample $K_b$ from conditional

$$K_{b}^{t+1} \sim p(K_b|K_{-b}^{t}, G, z^{(1:n)})$$

3. Repeat for $S$ iterations

Block Gibbs sampling for G-Wishart: projecteuclid.org/euclid.ejs/1328280902
Inference Algorithm: Gibbs

Having $K$, we sample $z^{(1:n)}$, noting independence between samples:

$$p(z^{(1:n)}|K, G, C) = \prod_{i=1}^{n} p(z^{(i)}|K, C)$$

We sample $z^{(i)}$ independently and employ a Gibbs sampler with conditional:

$$p(z_{v}^{(i)}|z_{-v}^{(i)}, K, C)$$

The conditional is truncated Gaussian. To impose $C$, we require

$$z_{v}^{(i)} \in [l_{v}^{i}, u_{v}^{i}]$$
Estimate of conditional independence:

\[
P(X_i \perp X_j | X_{-i-j}) = 1 - \frac{1}{T} \sum_{t=1}^{T} I_{ij}(G^t)
\]

\[
I_{ij}(G) = \begin{cases} 
1, & \text{if } (i,j) \in E \\
0, & \text{otherwise}
\end{cases}
\]

Estimate of correlation matrix:

\[
\tilde{K} = \frac{1}{T} \sum_{t=1}^{T} K^t
\]
Estimate of correlation matrix:

\[
\tilde{K} = \frac{1}{T} \sum_{t=1}^{T} K^t
\]

Estimate of Gaussian Copula CDF:

\[
F(X) = C(\hat{F}_1(x_1) \ldots \hat{F}_p(x_p)|\tilde{K})
\]

where \( \hat{F}_1(x_1) \) is the empirical marginal distribution.
Sampling x is easy once the correlation matrix is known.

- Sample Gaussian latent variable $z$
- Map $z$ back to $x$ using empirical marginal distributions

$$u_v = \phi(\tilde{z}_v)$$
$$\tilde{x}_v = \tilde{F}_v^{-1}(u_v)$$
Case Study
Case Study - Labor Force Survey Data

- Considers dependencies among income levels, educational attainment, fertility and family background
- Link: http://webapp.icpsr.umich.edu/GSSS/
<table>
<thead>
<tr>
<th>Variables</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEC - Income of the Respondent</td>
<td>ordinal variable (21C)</td>
</tr>
<tr>
<td>(INC)</td>
<td></td>
</tr>
<tr>
<td>DEG - Highest degree obtained</td>
<td>ordinal variable (5C)</td>
</tr>
<tr>
<td>by the respondent</td>
<td></td>
</tr>
<tr>
<td>CHILD - number of children of</td>
<td>count variable</td>
</tr>
<tr>
<td>the respondent</td>
<td></td>
</tr>
<tr>
<td>PINC - financial status of the</td>
<td>ordinal variable (5c)</td>
</tr>
<tr>
<td>parents of the respondent</td>
<td></td>
</tr>
<tr>
<td>PDEG - highest degree obtained</td>
<td>an ordinal variable (5c)</td>
</tr>
<tr>
<td>by the respondent’s parents</td>
<td></td>
</tr>
<tr>
<td>PCHILD - number of children of</td>
<td>count variable</td>
</tr>
<tr>
<td>the respondent’s parents</td>
<td></td>
</tr>
<tr>
<td>AGE - respondent’s age in years</td>
<td>count variable</td>
</tr>
</tbody>
</table>
Results

Fig: Estimates of the posterior inclusion probability of edges (CHILD, PINC) and (DEG, PCHILD) across iterations.

Table 4.1: Posterior estimates of the off-diagonal elements of $\Upsilon$ and posterior inclusion probability of edges for the labor force data.

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Entry in $\Upsilon$</th>
<th>Edge Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHILD</td>
<td>INC</td>
<td>0.292</td>
<td>0.997</td>
</tr>
<tr>
<td>CHILD</td>
<td>PCHILD</td>
<td>0.22</td>
<td>0.999</td>
</tr>
<tr>
<td>CHILD</td>
<td>PDEG</td>
<td>-0.262</td>
<td>0.953</td>
</tr>
<tr>
<td>CHILD</td>
<td>AGE</td>
<td>0.599</td>
<td>1</td>
</tr>
<tr>
<td>INC</td>
<td>DEG</td>
<td>0.489</td>
<td>1</td>
</tr>
<tr>
<td>INC</td>
<td>AGE</td>
<td>0.34</td>
<td>1</td>
</tr>
<tr>
<td>DEG</td>
<td>PCHILD</td>
<td>-0.187</td>
<td>0.668</td>
</tr>
<tr>
<td>DEG</td>
<td>PDEG</td>
<td>0.473</td>
<td>1</td>
</tr>
<tr>
<td>PCHILD</td>
<td>PDEG</td>
<td>-0.303</td>
<td>0.991</td>
</tr>
<tr>
<td>PINC</td>
<td>PDEG</td>
<td>0.453</td>
<td>1</td>
</tr>
<tr>
<td>PDEG</td>
<td>AGE</td>
<td>-0.232</td>
<td>0.988</td>
</tr>
</tbody>
</table>
Results

Table: Posterior estimates of the off-diagonal elements of $\Upsilon$ and posterior inclusion probability of edges for the labor force data

<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Entry in $\Upsilon$</th>
<th>Edge Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHILD</td>
<td>INC</td>
<td>0.292</td>
<td>0.997</td>
</tr>
<tr>
<td>CHILD</td>
<td>PCHILD</td>
<td>0.22</td>
<td>0.999</td>
</tr>
<tr>
<td>CHILD</td>
<td>PDEG</td>
<td>-0.262</td>
<td>0.953</td>
</tr>
<tr>
<td>CHILD</td>
<td>AGE</td>
<td>0.599</td>
<td>1</td>
</tr>
<tr>
<td>INC</td>
<td>DEG</td>
<td>0.489</td>
<td>1</td>
</tr>
<tr>
<td>INC</td>
<td>AGE</td>
<td>0.34</td>
<td>1</td>
</tr>
<tr>
<td>DEG</td>
<td>PCHILD</td>
<td>-0.187</td>
<td>0.668</td>
</tr>
<tr>
<td>DEG</td>
<td>PDEG</td>
<td>0.473</td>
<td>1</td>
</tr>
<tr>
<td>PCHILD</td>
<td>PDEG</td>
<td>-0.303</td>
<td>0.991</td>
</tr>
<tr>
<td>PINC</td>
<td>PDEG</td>
<td>0.453</td>
<td>1</td>
</tr>
<tr>
<td>PDEG</td>
<td>AGE</td>
<td>-0.232</td>
<td>0.988</td>
</tr>
</tbody>
</table>

Figure 4.1: Estimates of the posterior inclusion probability of edges (CHILD, PINC) and (DEG, PCHILD) across iterations.

Hoﬀ (9) assessed links between variables in this dataset by inspecting the 95% credible intervals for the regression coefficients to see if they have spanned zero. The main conclusions resulting from our copula Gaussian graphical models approach are shared by Hoﬀ (9) though we differ in two instances. First, our method shows a high probability (essentially one) of an edge between CHILD and PCHILD, while this link was absent in Hoﬀ (9). Such an inclusion seems sensible, as individual fertility levels are likely to be related to historical fertility in a given family. Furthermore, we place only a 20% inclusion probability on an edge between PINC and PCHILD, though this connection was displayed in Hoﬀ (9).
Closing remarks
We’ve covered an introduction to Copula Models:

- Advantages of Copula models over traditional Gaussian
- The Gaussian Copula model
- Inference using Copula model
Further Readings


Recent paper and accompanying code:

- Variational Gaussian Copula Inference
  people.ee.duke.edu/~lcarin/VGC_AISTATS2016.pdf
- github.com/shaobohanan/VariationalGaussianCopula

Copula models in a financial setting: