Multivariate density estimation via copulas

Peter Hoff

Statistics, Biostatistics and the CSSS

19th May 2006
Introduction to Copulas

Parameterization of Copulas

Parameter estimation

Example: Imputation of Pima diabetes data

Discussion
What is a copula?

A *copula density* is a multivariate probability density on \([0, 1]^2\) having uniform marginals:

\[
p_1(u) = \int_0^1 p(u_1, u_2) \, du_2 = 1 \quad p_2(u) = \int_0^1 p(u_1, u_2) \, du_1 = 1
\]

More generally, a copula refers to the CDF of such a density: \(C : [0, 1]^p \rightarrow [0, 1]\) is a copula if

- \(C\) is increasing.
- \(C(1, \ldots, 1, u_k, 1, \ldots, 1) = u_k\); 
- \(C(u_1, \ldots, u_p) = 0\) if \(\min\{u_1, \ldots, u_p\} = 0\);
What do they look like?
Why use copulas?

Any multivariate distribution can be completely described by its copula and its univariate distributions:

**Sklar’s Theorem:** Let $F$ be a $p$-dimensional CDF and $F_1, \ldots, F_p$ the univariate margins. Then there exists a copula $C$ such that

$$F(y_1, \ldots, y_p) = C(F_1(y_1), \ldots, F_p(y_p))$$

Think in terms of changes of variables. If $F$ is continuous,

$$(y_1, \ldots, y_p) \sim F \leftrightarrow \left\{ \begin{array}{l} u_k = F_k(y_k) \\ y_k = F_k^{-1}(u_k) \end{array} \right\} \leftrightarrow (u_1, \ldots, u_p) \sim C$$

“Copulas are of interest for two main reasons (N. Fisher)”:

1. a way of studying scale-free measures of dependence;
2. a starting point for constructing families of multivariate distributions.

they also allow us to divide multivariate density estimation into two parts:

univariate density estimation and copula estimation
"Kendall’s τ": \((y_{i,1}, y_{i,2})\) and \((y_{j,1}, y_{j,2})\) are a concordant pair if 
\((y_{i,1} - y_{j,1}) \times (y_{i,2} - y_{j,2}) > 0\), otherwise they are discordant.

\[ \hat{\tau} = \frac{1}{\binom{n}{2}} (c - d) \]

"Spearman’s ρ": Let \(r_{i,j}\) be the rank of \(y_{i,j}\) among variable \(\{y_{1,j}, \ldots, y_{n,j}\}\), \(i = \{1, \ldots, n\}\), \(j \in \{1, 2\}\).

\[ \hat{\rho} = \text{Cor}[(r_{1,1}, \ldots, r_{n,1}), (r_{1,2}, \ldots, r_{n,2})] \]

Both of these are invariant to monotone transformations of the variables, and thus depend on the copula and not the marginals. A variety of other dependence measures are derivable from the copula.
Consider the following model: Let $c$ be a copula density, and $G_1, \ldots, G_p$ be increasing functions with domain $[0, 1]$.

1. $u = \{u_1, \ldots, u_p\} \sim c$ be latent variables.
2. $y = \{y_1, \ldots, y_p\} = \{G_1(u_1), \ldots, G_p(u_p)\}$ be the observed data.

Then

- $c$ models the multivariate dependence,
- $G_1 = F_1^{-1}, \ldots, G_p = F_p^{-1}$ model the univariate distributions.
**Discrete copulas**

**Doubly stochastic:** A $K \times K$ matrix $\mathbf{M}$ is called **doubly stochastic** if it is positive and $\mathbf{M} \mathbf{1} = \mathbf{M}^T \mathbf{1} = \mathbf{1}$.

**Discrete copula:** If $\mathbf{M}$ is doubly stochastic then $\mathbf{M}/K$ is a **discrete copula**, a distribution on $\left\{ \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K}{K} \right\}^2$ with uniform marginals.
Smoothed copulas

A discrete copula can be smoothed out: \( f = (f_1, \ldots, f_K)^T : [0, 1] \rightarrow \mathbb{R}^K \) such that

1. each \( f_k \) is a probability density on \([0, 1]\), and
2. \( \sum_{k=1}^{K} f_k(u) = 1 \) for all \( u \in [0, 1] \).

By straightforward integration it can be shown that the function

\[
p(u_1, u_2|K, M) = \frac{1}{K} f(u_1)^T M f(u_2)
\]

is a copula density on \([0, 1]^2\) for any doubly stochastic matrix \( M \).

One such \( f \) is the set of beta densities with integer \((a, b)\), \( a + b = K + 1 \):

\[
f(u) = \{ \text{dbeta}(u, 1, K), \text{dbeta}(u, 2, K - 1), \ldots, \text{dbeta}(u, K, 1) \}
\]

Such an \( f \) is essentially a Bernstein polynomial, and the resulting copula is called a Bernstein copula.
How things get smoothed

- Introduction to Copulas
- Parameterization of Copulas
- Parameter estimation
- Example: Imputation of Pima diabetes data
- Discussion
Another way to write out the model is

\[ p(u_1, u_2 | M) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} M_{k_1,k_2} f_{k_1}(u_1) f_{k_2}(u_2) \]

This extends to higher dimensional densities as

\[ p(u | M) = \sum_{k_1=1}^{K} \cdots \sum_{k_p=1}^{K} M_{k_1,\ldots,k_p} \prod_{j=1}^{p} f_{k_j}(u_j) \]

This can be seen as a latent class model:

1. Sample a latent class vector \( k \in \{1, \ldots, K\}^p \) according to \( M \);
2. Sample \( u | k \sim \prod_{j=1}^{p} f_{k_j}(u_j) \).

Then \( u \) is a sample from \( p(u | M) \).
Estimation I

- Sancetta and Satchell (2004):
  1. Pick $K$ as a function of $n$, based on an asymptotic result;
  2. Let $\hat{M}$ be the empirical proportions in the $K \times K$ bins;
  3. Let $\hat{p}(u_1, u_2) = \frac{1}{K} f(u_1)^T \hat{M} f(u_2)$.
     Warning: not actually a copula density!

- Maximum likelihood:
  1. The parameter space for $M$ is a compact convex set.
  2. Use Newton’s method with a logarithmic barrier to minimize
     $-\sum_{i=1}^{n} \log p(u_{i,1}, u_{i,2}|M)$.
  3. Compare values of $K$ using AIC, BIC or something similar.

Question: Wait a minute, are $(u_{1,1}, u_{1,2}), \ldots, (u_{n,1}, u_{n,2})$ actually observed?

Answer: No. People generally plug-in $\hat{u}_{i,j} = \hat{F}(y_{i,j})$. 
Problems with the aforementioned approaches:

- The $u_{i,j}$'s not actually observed - uncertainty in their value is not accounted for (this is primarily a concern if the $y_{i,j}$'s are discrete).
- In S&S’s approach the estimate isn't actually a copula.
- In the MLE approach things get pretty messy in higher dimensions.
- In some cases we may want the coarseness of $\mathbf{M}$ to be different across the $p$ variables.

Maybe we can solve these problems and/or make everything more complicated. I propose to do this by constructing a mixture model for copula densities that mix over simple Bernstein copulas of varying coarseness.
Bivariate Bernstein copula: \( \mathbf{M} \in \mathcal{M}_K = \{\text{stochastic } K \times K \text{ matrices}\} \).

Choquet’s theorem: \( \mathcal{M}_K \) is a compact, convex set. Every \( \mathbf{M} \in \mathcal{M}_K \) can therefore be expressed as a mixture over the extreme points (vertices) of \( \mathcal{M}_K \).

\[
\mathbf{M} \in \mathcal{M}_K \iff \mathbf{M} = \sum w(S)S
\]

In other words, the

constrained estimation problem (estimation of \( \mathbf{M} \in \mathcal{M}_K \))

can be re-expressed as an

unconstrained mixture estimation problem (estimation of \( w(S) \)).
Permutation matrices

The extreme points of \( \mathcal{M}_K \) are the permutation matrices: stochastic matrices consisting of zeros and ones.

\[
S_1 = \begin{pmatrix}
0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0
\end{pmatrix} \quad S_2 = \begin{pmatrix}
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix} \quad S_3 = \begin{pmatrix}
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0
\end{pmatrix}
\]

Note that these can be expressed as \( K \times 2 \) matrices

\[
A_1 = \begin{pmatrix}
1 & 1 \\
2 & 2 \\
3 & 3 \\
4 & 4
\end{pmatrix} \quad A_2 = \begin{pmatrix}
1 & 4 \\
2 & 3 \\
3 & 1 \\
4 & 2
\end{pmatrix} \quad A_3 = \begin{pmatrix}
1 & 1 \\
2 & 3 \\
3 & 4 \\
4 & 2
\end{pmatrix}
\]

Extreme Bernstein density:

\[
p(u_1, u_2 | S) = \frac{1}{K} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} S_{k_1, k_2} f_{k_1}(u_1)f_{k_2}(u_2) = \frac{1}{K} \sum_{k=1}^{K} f_{A_{k,1}}(u_1)f_{A_{k,2}}(u_2)
\]
Permutation arrays

This idea can be extended beyond two dimensions:

**Stochastic arrays:** Let $\mathcal{M}_K$ be the set of $K^p$-dimensional stochastic arrays

**Extreme points:** Stochastic arrays consisting of ones and zeros $\Rightarrow$ multivariate permutation arrays.

\[
\begin{pmatrix}
1 & 1 & 4 \\
2 & 2 & 3 \\
3 & 3 & 2 \\
4 & 4 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 4 & 1 \\
2 & 3 & 2 \\
3 & 1 & 4 \\
4 & 2 & 3
\end{pmatrix}
\begin{pmatrix}
1 & 1 & 2 \\
2 & 3 & 1 \\
3 & 4 & 3 \\
4 & 2 & 4
\end{pmatrix}
\]

**Extreme Bernstein density:**
\[
p(u|A) = \frac{1}{K} \sum_{k=1}^{K} \left\{ \prod_{j=1}^{p} f_{A_{k,j}}(u_j) \right\}
\]
Suppose $u_1$ and $u_2$ are highly dependent, but independent of $u_3$. It will take a mixture of many extreme Bernstein copulas to represent this.

To obtain a more efficient representation, consider permutation arrays of the form

$$
\begin{pmatrix}
1 & 1 & 1 \\
2 & 2 & 1 \\
3 & 3 & 1 \\
4 & 4 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 1 & 1 \\
2 & 1 & 1 \\
3 & 2 & 1 \\
4 & 2 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 4 & 1 \\
2 & 2 & 2 \\
3 & 3 & 2 \\
4 & 1 & 1
\end{pmatrix}
$$

Each column of $A$ is a permutation of $K/K_j$ copies of $\{1, \ldots, K_j\}$. We have restricted $K/K_j$ to be a power of 2. The density

$$
p(u|A) = \frac{1}{K} \sum_{k=1}^{K} \left\{ \prod_{j=1}^{p} f_{A_{k,j}}(u_j) \right\}
$$

is still a copula density.
Mixtures of Bernstein copulas

**Mixture model:** Any copula density can be approximated by a mixture of the form

\[ p(u|q) = \int_A p(u|A)q(dA). \]

The mixing measure \( q \) is unknown and to be estimated. It is a measure over rectangular permutation arrays of various shapes and sizes.
We generally don’t observe data $u_1, \ldots, u_n$ with uniform marginals. What does the observed data tell us about $u_1 \ldots, u_n$?

$$D = \{u_1, \ldots, u_n: u_{i_1,j} < u_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j}\}$$

The “partial likelihood” is

$$p(D|q) = \int_D \left\{ \prod_{i=1}^{n} p(u_i|q) \, du_i \right\}$$

We will consider penalbayesized estimates which achieve high values of the following objective function

$$f(q|u_1, \ldots, u_n) = \log p(D|q) + \log \pi(q)$$

As you might guess, there are MCMC schemes to obtain such estimates.
**Imputation experiment**

**Experiment:** Given data on $p = 8$ variables for $n = 532$ women,

1. replace 10% of data with missing values;
2. obtain posterior mean $\hat{y}_{i,j}^b$ of $y_{i,j}$ for each missing value;
3. obtain 5-nearest-neighbor estimate $\hat{y}_{i,j}^n$ of $y_{i,j}$ for each missing value;
4. compare $\hat{y}^b$ and $\hat{y}^n$ to actual values.

<table>
<thead>
<tr>
<th>variable</th>
<th>Bayes error$^{1/2}_B$</th>
<th>Bayes error$^{1/2}_G$</th>
<th>KNN error$^{1/2}$</th>
<th>MSE$^{1/2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>npreg</td>
<td>0.92</td>
<td>0.95</td>
<td>0.90</td>
<td>1.25</td>
</tr>
<tr>
<td>glu</td>
<td>1.01</td>
<td>0.99</td>
<td>1.07</td>
<td>1.08</td>
</tr>
<tr>
<td>bp</td>
<td>0.97</td>
<td>0.97</td>
<td>0.90</td>
<td>1.07</td>
</tr>
<tr>
<td>skin</td>
<td>0.66</td>
<td>0.67</td>
<td>0.82</td>
<td>0.88</td>
</tr>
<tr>
<td>bmi</td>
<td>0.73</td>
<td>0.74</td>
<td>0.74</td>
<td>0.88</td>
</tr>
<tr>
<td>ped</td>
<td>0.82</td>
<td>0.82</td>
<td>0.96</td>
<td>0.86</td>
</tr>
<tr>
<td>age</td>
<td>0.96</td>
<td>0.98</td>
<td>1.15</td>
<td>1.16</td>
</tr>
<tr>
<td>type</td>
<td>0.97</td>
<td>0.93</td>
<td>0.99</td>
<td>1.04</td>
</tr>
</tbody>
</table>
Example: Imputation of Pima diabetes data
Estimates of bivariate marginals: Raw data

- bp vs npreg
- age vs npreg
- type vs glu
- age vs ped
- bp vs bmi
- bp vs skin
Estimates of bivariate marginals: Bernstein copula
Estimates of bivariate marginals: Gaussian copula
Summary
Future work