Multivariate density estimation via copulas

Peter Hoff

Statistics, Biostatistics and the CSSS

May 31st 2006
Outline

Multivariate Data

The Gaussian copula

Diabetes example

Nonparametric copulas
Social science studies often yield multivariate data of varied types.

**Hypothetical survey data:** A vector of responses $y_i = (y_{i,1}, \ldots, y_{i,p})$ for each person $i$ in a sample of survey respondents, $i \in \{1, \ldots, n\}$.

- $y_{i,1} =$ sex
- $y_{i,3} =$ age
- $y_{i,4} =$ income
- $y_{i,2} =$ education level
- $y_{i,5} =$ number of children
- $y_{i,6} =$ attitude (likert scale)

Often of interest are the potential associations among these variables.
Measures of association

“Pearson’s $\rho$”: Measures the linear association between two data vectors, or more precisely, the angle between the data vectors:

$$\hat{\rho} = \frac{\sum (y_{i,1} - \bar{y}_{.,1})(y_{i,2} - \bar{y}_{.,2})}{\sqrt{\sum (y_{i,1} - \bar{y}_{.,1})^2 \sum (y_{i,2} - \bar{y}_{.,2})^2}}$$

“Kendall’s $\tau$”: $(y_{i,1}, y_{i,2})$ 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 $\rho$”: Let $r_{i,j}$ be the rank of $y_{i,j}$ among variable $\{y_{1,j}, \ldots, y_{n,j}\}$, $i = \{1, \ldots, n\}$, $j \in \{1, 2\}$.

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

All are between -1 and +1. The latter two are invariant to monotone transformations, and so are “scale free”. The moment correlation is not.
Monotone transformations

<table>
<thead>
<tr>
<th>variables</th>
<th>moment</th>
<th>rank</th>
<th>concordance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1, y_2$</td>
<td>.28</td>
<td>.39</td>
<td>.27</td>
</tr>
<tr>
<td>$\log y_1, y_2$</td>
<td>.26</td>
<td>.39</td>
<td>.27</td>
</tr>
<tr>
<td>$y_1, \log y_2$</td>
<td>.42</td>
<td>.39</td>
<td>.27</td>
</tr>
<tr>
<td>$\log y_1, \log y_2$</td>
<td>.44</td>
<td>.39</td>
<td>.27</td>
</tr>
</tbody>
</table>
Pros and cons

Scale free measures:

• **Pros**: They are scale free.

• **Cons**: Bivariate, limited in term of inference, not sure what to do with ties, not related to a statistical model, . . .

Product correlation:

• **Pros**: Measures the linear association, which may be of interest. Also, if the population has a multivariate normal distribution, the moment correlations make up the *sufficient statistics*: If you’ve reported the sample correlations (and the means and variances), you’ve reported everything there is to report.

• **Cons**: Data are often not normally distributed.
Univariate models

One approach is to fit several univariate models, tinkering around with transformations or generalized linear models:

\[
\frac{y_{npreg}^{1/2}}{2} = \beta_{npreg,1} + \beta_{npreg,2}y_{glu} + \beta_{npreg,3}y_{bp} + \cdots \\
\log \text{odds}(y_{type} = 1) = \beta_{type,1} + \beta_{type,2}y_{glu} + \beta_{type,3}y_{bp} + \cdots
\]

Problems:

- Typical univariate models are parametric, too restrictive.
- Univariate model selection is subject to sampling variability, and typically ad-hoc.
- Univariate models may be inconsistent with each other, in that they do not correspond to any multivariate model (for example, the different sets of \(\beta\)'s could represent inconsistent relationships).
Inverse normal model

Suppose we could transform each variable so that it had a standard normal distribution. This is possible in the following sense:

If $F$ is a distribution there exists a nondecreasing function $G$ such that

1. if $z \sim \text{normal}(0,1)$,
2. and $y = G(z)$,

then $y \sim F$.

If $F$ is continuous then $G(z) = F^{-1}(\Phi(z))$, $G^{-1}$ is a function and $G^{-1}(y)$ is standard normal. If $F$ is not continuous then $G^{-1}$ maps to a set.
Inverse normal model

Probit models are included in this formulation:

\[
\begin{cases}
  z \sim \Phi \\
y = 1_{(c,\infty)}(z) = G(z)
\end{cases}
\Leftrightarrow p(y) = (1 - \Phi(c))^y \Phi(c)^{1-y}
\]

In general,

- The marginal CDF of \( y_j \) is \( F_j(y) = \Phi[\max G^{-1}(y)] \)
- If \( F_j, G_j \) are invertible then
  - \( F_j(y) = \Phi[G_j^{-1}(y)] \) mapping \( Y \)-space to \([0,1]\)
  - \( G_j(z) = F_j^{-1}[\Phi(z)] \) mapping \( Z \)-space to \( Y \)-space
Multivariate normal copula model

This idea motivates the following “latent variable” model:

\[
(z_1, \ldots, z_p) \sim \text{mvn}(\mathbf{0}, \Sigma) \\
(y_1, \ldots, y_p) = (G_1(z_1), \ldots, G_p(z_p))
\]

\(\Sigma\) parameterizes the dependence, \(G_1, \ldots, G_p\) parameterize the marginal distributions.

**Estimation strategies:**

- estimation of \(\Sigma\) conditional on plug-in estimates of \(G_1, \ldots, G_p\);
- joint estimation of \(\Sigma\) and \(G_1, \ldots, G_p\);
- partial information estimation.
Given observed data $y_1, \ldots, y_n$, the “full likelihood” is

$$p(y_1, \ldots, y_n | \Sigma, G_1 \ldots, G_p) = \prod_{i=1}^{n} p(y_i | \Sigma, G_1, \ldots, G_p)$$

If the $G$’s are invertible then

$$p(y_i | \Sigma, G_1, \ldots, G_p) = \frac{\text{mvnorm}[z(y) | \Sigma]}{\prod_{j=1}^{p} \text{norm}[z_j(y_j)]},$$

otherwise, the likelihood is more complicated. In any case:

- Estimation of $G_j$’s with parametric families may be too restrictive.
- Nonparametric estimation of $G_j$’s jointly with $\Sigma$, is very hard.
Partial information estimation

What information do the $y$’s give us about the $z$’s?

$$y_{i_1,j} < y_{i_2,j} \Rightarrow z_{i_1,j} < z_{i_2,j}$$

A component of the observed data is that the $z$’s lie in the following set:

$$D(y_1, \ldots, y_n) = \{z_1, \ldots, z_n : y_{i_1,j} < y_{i_2,j} \Rightarrow z_{i_1,j} < z_{i_2,j} \}$$

We can base inference about $\Sigma$ on this information, without making any parametric assumptions about the marginal distributions of the $y$’s.

$$p(D|\Sigma, G_1, \ldots, G_p) = p(D|\Sigma) = \int_D \prod_{i=1}^n \{\text{mvnorm}(z_i|\Sigma) \} \, dz_i$$
Diabetes example

Survey of 532 women of Pima Indian heritage living near Phoenix.

\texttt{npreg} : number of pregnancies
\texttt{glu} : plasma glucose concentration
\texttt{bp} : diastolic blood pressure
\texttt{skin} : triceps skin fold thickness
\texttt{bmi} : body mass index
\texttt{ped} : diabetes pedigree function
\texttt{age} : age in years
\texttt{type} : diabetic, yes or no

Using MCMC integration, we estimate
\[ \Sigma, \text{ the correlation matrix, and} \]
\[ \Sigma_{[j,-j]} \Sigma_{[-j,-j]}^{-1}, \text{ the regression coefficients.} \]
Diabetes example: Correlations and regressions
Joint distributions
## Imputation

<table>
<thead>
<tr>
<th>Method</th>
<th>npreg</th>
<th>glu</th>
<th>bp</th>
<th>skin</th>
<th>bmi</th>
<th>ped</th>
<th>age</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian copula</td>
<td>7.67</td>
<td>763.47</td>
<td>132.92</td>
<td>72.36</td>
<td>29.06</td>
<td>0.11</td>
<td>75.30</td>
<td>0.16</td>
</tr>
<tr>
<td>9-nearest neighbors</td>
<td>9.46</td>
<td>950.60</td>
<td>149.42</td>
<td>97.68</td>
<td>34.70</td>
<td>0.13</td>
<td>105.68</td>
<td>0.19</td>
</tr>
</tbody>
</table>
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: $C : [0, 1]^p \to [0, 1]$ 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
Multivariate normal copula

Let $H$ be the CDF of a $p$-variate multivariate normal population. Applying Sklar’s theorem to $H$ says there is a copula $C_H$ such that

$$H(z_1, \ldots, z_p) = C_H(\Phi(z_1), \ldots, \Phi(z_p))$$

1. $C_H$ is a density with uniform marginals.
   - Let $u_1 = \Phi(z_1), \ldots, u_p = \Phi(z_p)$.
   - Note that $u_j$ is marginally uniformly distributed.

2. Suppose the marginals CDF’s of our observed data are $F_1, \ldots, F_p$.
   - Let $y_1 = F_1^{-1}(u_1), \ldots, y_p = F_p^{-1}(u_p)$.
   - Note that $y_j$ is marginally distributed as $F_j$.

The model

$$(u_1, \ldots, u_p) \sim C_H, \quad (y_1, \ldots, y_p) = (F_1^{-1}(u_1), \ldots, F_p^{-1}(u_p))$$

is said to be a multivariate normal copula model. It is what we were doing in slides 9-17 (recall, $y = G(z) = F^{-1}(\Phi(z))$). But what if there is more to the dependence than just correlation (on some scale)?
Discrete copulas

**Doubly stochastic:** A $K \times K$ matrix $M$ is called **doubly stochastic** if it is positive and $M1 = M^T 1 = 1$.

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

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

(f1) each \( f_k \) is a probability density on \([0, 1]\), and

(f2) \( \sum_{k=1}^{K} f_k(u) = 1 \) 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
Another way to write out the model is

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

This extends to higher dimensional densities as

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

This can be seen as a latent class model:

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

Then \( u \) is a sample from \( p(u|\mathbf{M}) \).
Estimation

- 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 $\hat{u}_{i,j} = \hat{F}(y_{i,j})$. 
Research goals

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 am working on a method of doing this by constructing a mixture model for copula densities that mix over simple Bernstein copulas of varying coarseness.
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 \( \hat{y}_{i,j} \) of \( y_{i,j} \) for each missing value;
3. compare \( \hat{y} \) to actual values.

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

Multivariate Data  The Gaussian copula  Diabetes example  Nonparametric copulas
Estimates of bivariate marginals: Gaussian copula
Another example: Boston data

Data on 506 townships near Boston

**crim** : crime

**zn** : proportion of land zoned for large lots

**nox** : nitrogen oxides concentration (parts per 10 million)

**rm** : average rooms per dwelling

**lstat** : percent of pop in “lower status”

**medv** : median value of owner-occupied homes in $1000
Boston data
Nonparametric copula estimate
Summary

- Many multivariate normal datasets are not multivariate normal.
- A more flexible model class would be nonparametric marginals with a Gaussian copula.
- An even more flexible class would be nonparametric marginals and a nonparametric copula...but
- Gaussian stuff works pretty well considering how simple it is.