Marginal likelihood for Bayesian copula estimation

Peter Hoff

Statistics, Biostatistics and the CSSS
University of Washington

August 10, 2006
Outline

Multivariate Data

- Measures of association
- GSS data

The Gaussian copula

- Inverse normal model
- Marginal likelihood estimation
- GSS example

Nonparametric copulas

- Discrete copulas
- Bernstein copulas
- GSS example
Social science studies often yield multivariate data of varied types.

**Hypothetical survey data:** A vector of responses $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,p})$ 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 ρ”: 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 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 \(\{y_{1,j}, \ldots, y_{n,j}\}\), \(i = \{1, \ldots, n\}, j \in \{1, 2\}\).

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

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>
GSS data
Inverse normal model

Rank-based methods are problematic for discrete data because of ties, and such data cannot be transformed to have a normal distribution. However, such data can be viewed as a function of normal data.

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 (this includes probit models, for example).
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 \(G_1, \ldots, G_p\); 
  (gives inconsistent results for discrete data)
- joint estimation of \(\Sigma\) and \(G_1, \ldots, G_p\); 
  (parametric models too simple, nonparametric to complex)
- marginal likelihood estimation.
  (??)
Marginal likelihood estimation

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

\[ y_{i_1,j} < y_{i_2,j} \Rightarrow z_{i_1,j} < z_{i_2,j} \]

Part of the information in the data is that the $z$'s lie in the following set:

\[ D(y_1, \ldots, y_n) = \{ z_1, \ldots, z_n : z_{i_1,j} < z_{i_2,j} \text{ if } y_{i_1,j} < y_{i_2,j} \} \]

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

\[
p(Z \in D|\Sigma, G_1, \ldots, G_p) = p(Z \in D|\Sigma) = \int_D \prod_{i=1}^n \{ \text{mvnorm}(z_i|\Sigma) \} \, dz_i
\]
Marginal likelihood estimation

\[
p(y_1, \ldots, y_n|\Sigma, G) = p(Z \in D|\Sigma, G) \times p(y_1, \ldots, y_n|Z \in D, \Sigma, G) \\
= p(Z \in D|\Sigma) \times p(y_1, \ldots, y_n|Z \in D, \Sigma, G)
\]

So the marginal likelihood \( p(Z \in D|\Sigma) \) doesn’t depend on the nuisance parameters \( G_1, \ldots, G_p \). Using this likelihood for copula estimation is “optimal” if the \( G_j \)’s are assumed to be continuous:

- The model is a transformation model and \( \Sigma \) is a maximal invariant;
- \( D \) gives the same information as the ranks;
- \( D \) is \( G \)-sufficient with respect to \( \Sigma \) in the sense of Barnard (1963);
- \( D \) is \( L \)-sufficient with respect to \( \Sigma \) in the sense of Rémon (1984).

However, if some of the \( G_j \)’s are discrete then

- \( D \) does not contain the same information as the ranks;
- The ranks give information about \( G_1, \ldots, G_p \).
Marginal likelihood estimation

Inference about $\Sigma$ can be obtained via Gibbs sampling, an approximate EM-algorithm, or Monte-Carlo maximum likelihood:

1. for ( j in 1:p ) {
   for ( y in unique(Y[,j]) ) {
      for i's such that Y[i,j]=y, find the constraints on Z[i,j] imposed by D;
      sample each Z[i,j] from a constrained univariate normal distribution.
   }
}
2. Sample $\Sigma$ from its full conditional distribution.

- For large datasets with “continuous” data, these steps can be avoided by matching quantiles, i.e. setting $z_{i,j} = \Phi^{-1}[\hat{G}_j(y_{i,j})]$.
- For discrete data on the other hand, setting $G_j$ to $\hat{G}_j$ will give inconsistent estimators.
GSS Example

Data on 1819 respondents to the 1987 GSS.

SEX : sex of respondent
AGE : age of respondent
CHILDS : number of children ever had
DEGREE : highest degree
PADEG : father’s highest degree
MADEG : mother’s highest degree
WORDSUM : score of vocabulary test
FINCOME : family income
ATTEND : church attendance
PALEFULL : is the afterlife a pale, or full existence
NEARGOD : closeness to god
BIBLE : bible is word of god, inspired book, or book of fables

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

Predictive MSE comparisons to the multivariate normal copula model:

<table>
<thead>
<tr>
<th>Variable</th>
<th>ordinary least-squares imputation</th>
<th>9-nearest neighbor imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEX</td>
<td>1.09</td>
<td>1.09</td>
</tr>
<tr>
<td>AGE</td>
<td>1.25</td>
<td>1.43</td>
</tr>
<tr>
<td>CHILDS</td>
<td>1.03</td>
<td>1.49</td>
</tr>
<tr>
<td>DEGREE</td>
<td>1.02</td>
<td>1.53</td>
</tr>
<tr>
<td>PADEG</td>
<td>1.07</td>
<td>1.69</td>
</tr>
<tr>
<td>MADEG</td>
<td>1.03</td>
<td>1.50</td>
</tr>
<tr>
<td>WORDSUM</td>
<td>1.22</td>
<td>1.45</td>
</tr>
<tr>
<td>FINCOME</td>
<td>1.05</td>
<td>1.30</td>
</tr>
<tr>
<td>ATTEND</td>
<td>1.05</td>
<td>1.25</td>
</tr>
<tr>
<td>PALEFULL</td>
<td>1.28</td>
<td>1.29</td>
</tr>
<tr>
<td>NEARGOD</td>
<td>2.18</td>
<td>1.25</td>
</tr>
<tr>
<td>BIBLE</td>
<td>1.31</td>
<td>1.48</td>
</tr>
</tbody>
</table>
Copulas

The “model” we have been using could be called a semiparametric Gaussian copula model. Generally speaking, a copula model involves

- a vector of latent variables \((u_1, \ldots, u_p) = \mathbf{u} \sim p\)
- a vector of observed variables \((y_1, \ldots, y_p) = (G_1(u_1), \ldots, G_p(u_p))\).

The density \(p\) has fixed marginals, typically taken to be uniform (but were standard normal in the preceding slides)

Traditionally, the word “copula” refers to the CDF of such a density:

\[ C : [0, 1]^p \rightarrow [0, 1] \text{ is a copula if} \]

- \(C\) is increasing.
- \(C(1, \ldots, 1, u_k, 1, \ldots, 1) = u_k\)
- \(C(u_1, \ldots, u_p) = 0\) if \(\min\{u_1, \ldots, u_p\} = 0\);

If \(z \sim \text{mvn}(\mathbf{0}, \Sigma)\), the mvn copula is the joint CDF of \((\Phi(z_1), \ldots, \Phi(z_p))\).
What do they look like?
Discrete copulas

Idea: build a nonparametric class of copula densities out of smoothed versions of simple, discrete copulas.

Doubly stochastic: A $K \times K$ matrix $M$ is called doubly stochastic if it is positive and $M1 = M^T1 = 1$.

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: \( f = (f_1, \ldots, f_K)^T : [0, 1] \rightarrow \mathbb{R}^K \) 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

The diagrams illustrate the process of smoothing in a grid. The values in the grid represent the intensity or weight of the effect. As the smoothing progresses, the values in the grid become more uniform, indicating a smoother image. The transitions between grid states show how the smoothing algorithm reduces high-intensity areas, creating a more even distribution of values across the grid.
Multivariate extension

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

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

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 $p(u|M)$.

Parameters to estimate include $M$ and $K$
<table>
<thead>
<tr>
<th>AGE</th>
<th>FINCOME</th>
<th>AGE</th>
<th>FINCOME</th>
<th>DEGREE</th>
<th>WORDSUM</th>
<th>DEGREE</th>
<th>WORDSUM</th>
<th>DEGREE</th>
<th>FINCOME</th>
<th>NEARGOD</th>
<th>BIBLE</th>
<th>NEARGOD</th>
</tr>
</thead>
</table>

Bernstein - Gaussian comparison
Summary and future work

Summary

▶ Marginal likelihood provides inference for dependence parameters.
▶ Mixtures of Bernstein copulas provide a flexible class of densities.
▶ Mixtures of Bernstein copulas may be too flexible.

Future work

▶ Study the information properties/asymptotics of marginal likelihood.
▶ Identify principled ways of smoothing the Bernstein mixture model.