Extending the rank likelihood for semiparametric copula estimation

Peter Hoff

Statistics, Biostatistics and the CSSS
University of Washington
Outline

Multivariate Association
  Scale-free association measures
  Conditional modeling

The Gaussian copula
  Inverse normal model
  Rank likelihood estimation
  GSS example

Sampling properties
  Small sample behavior
  Asymptotic efficiency

Nonparametric copulas
  Discrete copulas
  Bernstein copulas
  GSS example
Survey data 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, $i \in \{1, \ldots, n\}$.

- $y_{i,1} =$ income
- $y_{i,2} =$ education level
- $y_{i,3} =$ number of children
- $y_{i,4} =$ age
- $y_{i,5} =$ 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:

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

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

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

“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.

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

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>
Conditional models

Interest is typically in the conditional relationship between pairs of variables, accounting for heterogeneity in other variables of less interest. Standard bivariate rank-based methods are inappropriate.

Model 1

\[
\text{INC}_i = \beta_0 + \beta_1 \text{CHILD}_i + \beta_2 \text{DEG}_i + \beta_3 \text{AGE}_i + \beta_4 \text{PCHILD}_i + \beta_5 \text{PINC}_i + \beta_6 \text{PDEG}_i + \epsilon_i
\]

p-value for \( \beta_1 \) is 0.11: “little evidence” that \( \beta_1 \neq 0 \)

Model 2

\[
\text{CHILD}_i \sim \text{Pois}(\exp\{\beta_0 + \beta_1 \text{INC}_i + \beta_2 \text{DEG}_i + \beta_3 \text{AGE}_i + \beta_4 \text{PCHILD}_i + \beta_5 \text{PINC}_i + \beta_6 \text{PDEG}_i\})
\]

p-value for \( \beta_1 \) is 0.01: “strong evidence” that \( \beta_1 \neq 0 \).

<table>
<thead>
<tr>
<th>Predictor</th>
<th>INC</th>
<th>CHILD</th>
<th>DEG</th>
<th>AGE</th>
<th>PCHILD</th>
<th>PINC</th>
<th>PDEG</th>
</tr>
</thead>
<tbody>
<tr>
<td>INC</td>
<td>NA</td>
<td>1.10 (.11)</td>
<td>7.03 (&lt;.01)</td>
<td>.34 (&lt;.01)</td>
<td>4.07 (&lt;.01)</td>
<td>.28 (.41)</td>
<td>1.40 (.12)</td>
</tr>
<tr>
<td>CHILD</td>
<td>.01 (.01)</td>
<td>NA</td>
<td>-.07 (.06)</td>
<td>.04 (&lt;.01)</td>
<td>-.06 (.20)</td>
<td>.02 (.08)</td>
<td>-.05 (.20)</td>
</tr>
</tbody>
</table>
Inverse normal model

One possibility would be to transform the data to have normal marginals, then fit a multivariate normal model. This cannot be done for discrete data, but 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).
Multivariate normal copula model

This idea motivates the following “latent variable” model:

\[(z_1, \ldots, z_p) \sim \text{mvn}(0, \Sigma)\]

\[(y_1, \ldots, y_p) = (G_1(z_1), \ldots, G_p(z_p))\]

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

- scale free
- appropriate for discrete and continuous data
- compatible full conditional distributions

**Estimation strategies:**

- estimation of \(\Sigma\) conditional on plug-in estimates of \(G_1, \ldots, G_p\);
  (procedures for continuous data gives inconsistent results for discrete data)

- joint estimation of \(\Sigma\) and \(G_1, \ldots, G_p\);
  (parametric models of \(G\) too simple, nonparametric too complex)

- marginal likelihood estimation.
  (how would that work?)
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

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

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;
- \( Z \in D \) gives the same information as the ranks;
- \( Z \in D \) is \( G \)- and \( L \)-sufficient (Barnard[1963], Rémon[1984]) for \( \Sigma \).

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

- \( Z \in D \) contains less information than the ranks;
- The distribution of the ranks depends on \( G_1, \ldots, G_p \);
- Perhaps \( \{Z \in D\} \) is a optimal insufficient statistic?
Marginal likelihood estimation

\[
\Pr(\Sigma \in A | \mathbf{Z} \in D) = \int_A \int p(\Sigma, \mathbf{Z} | \mathbf{Z} \in D) \, d\mathbf{Z} d\Sigma
\]

Inference about \( \Sigma \) can be obtained via iterative Gibbs sampling of

\[
\mathbf{Z} \sim p(\mathbf{Z} | \Sigma, \mathbf{Z} \in D) \text{ and } \Sigma \sim p(\Sigma | \mathbf{Z})
\]

More precisely,

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.
The actual R-code

Given \{Z,S\} and \{Ranks,n,p,S0,n0\}:

### update Z

for (j in 1:p) {

    Sjc <- S[j,-j] %*% solve(S[-j,-j])
    sdj <- sqrt( S[j,j] - S[j,-j] %*% solve(S[-j,-j]) %*% S[-j,j] )
    muj <- Z[,-j] %*% t(Sjc)

    for (r in unique(Ranks[,j])){

        ir <- (1:n)[Ranks[,j] == r & !is.na(Ranks[,j])]
        lb <- suppressWarnings(max( Z[ Ranks[,j] == r-1,j ], na.rm=TRUE ))
        ub <- suppressWarnings(min( Z[ Ranks[,j] == r+1,j ], na.rm=TRUE ))
        Z[ir,j] <- qnorm(runif(length(ir),
            pnorm(lb,muj[ir],sdj),pnorm(ub,muj[ir],sdj)),muj[ir],sdj)
    }

    ir <- (1:n)[is.na(Ranks[,j])]
    Z[ir,j] <- rnorm(length(ir),muj[ir],sdj)

}  

### update S

S <- solve(rwish(solve(S0*n0+t(Z) %*% Z),n0+n))
Prior distribution on $C$

The model and prior:

$$
\mathbf{V} \sim \text{inverse-Wishart}(\nu_0, \nu_0 \mathbf{V}_0)
$$

$$
\{C_{[j_1,j_2]}\} = \left\{\mathbf{V}_{[j_1,j_2]}/\sqrt{\mathbf{V}_{[j_1,j_1]} \mathbf{V}_{[j_2,j_2]}}\right\}
$$

$$
\mathbf{z}_1, \ldots, \mathbf{z}_n \sim \text{i.i.d. multivariate normal}(\mathbf{0}, \mathbf{C})
$$

$$
y_{i,j} = G_j(z_{i,j}),
$$

The model and prior for computation:

$$
\mathbf{V} \sim \text{inverse-Wishart}(\nu_0, \nu_0 \mathbf{V}_0)
$$

$$
\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n \sim \text{i.i.d. multivariate normal}(\mathbf{0}, \mathbf{V})
$$

$$
\mathbf{z}_{i,j} = \tilde{\mathbf{z}}_{i,j}/\sqrt{\mathbf{V}_{[j,j]}}, \quad \text{and let } \mathbf{C} = \text{Cov}(\mathbf{z})
$$

$$
y_{i,j} = G_j(\mathbf{z}_{i,j}) = \tilde{G}_j(\tilde{\mathbf{z}}_{i,j}).
$$

$p(\mathbf{Y}, \mathbf{C})$ is common across specifications.
GSS Example

Data on 1002 male respondents to the 1994 GSS.

\[ \text{INC} : \text{income of respondent} \]
\[ \text{DEG} : \text{highest degree obtained} \]
\[ \text{CHILD} : \text{number of children} \]
\[ \text{PINC} : \text{income category of parents} \]
\[ \text{PDEG} : \text{maximum of mother's and father's highest degree} \]
\[ \text{PCHILD} : \text{number of siblings plus one} \]
\[ \text{AGE} : \text{age in years} \]

Using MCMC integration, we estimate

\[ \Sigma, \text{the correlation matrix, and} \]
\[ \Sigma_{[j,-j]} \Sigma_{[-j,-j]}^{-1}, \text{the regression coefficients.} \]
MCMC diagnostics
Correlations and regressions
Correlations and regressions
Predictive conditional distributions
How can you make predictions without the margins? Consider sampling from the posterior predictive:

\[ \tilde{z} \sim \text{multivariate normal}(0, C) \]
\[ \tilde{y}_j = G_j(\tilde{z}_j) \]

We don't know \( G_j \), but monotonicity implies

\[
\max \{ y_{i,j} : z_{i,j} < \tilde{z}_j \} \leq \tilde{y}_j \leq \min \{ y_{i,j} : \tilde{z}_j < z_{i,j} \},
\]

which gives an interval (or point) prediction, shrinking with \( n \). This is essentially a multivariate version of Hill’s \( A_n \) prediction procedure.
### 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>
Small sample behavior

MSE from ordinary likelihood

0.03 0.05

MSE from s-likelihood

n=25

0.010 0.015 0.020

n=100

0.010 0.015 0.020

MSE from s-likelihood

n=50

0.015 0.025 0.035

MSE from s-likelihood

0.03 0.05

MSE from s-likelihood
Large sample behavior

For continuous data the rank likelihood gives AE MLE’s:

$$\log \frac{p(Z \in D|C + \frac{1}{\sqrt{n}}A)}{p(Z \in D|C)} \approx \log \frac{p(Y|G, C + \frac{1}{\sqrt{n}}A)}{p(Y|G, C)}$$

We are working on a proof of AE for the discrete case. If successful, then

- for small $n$, rank likelihood outperforms (nonparametric) likelihood;
- for large $n$, rank likelihood is efficient.

This suggests that rank likelihood is broadly applicable if interest is in $C$. 
Non-Gaussian copulas

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

- a vector of latent variables \((u_1, \ldots, u_p) = u \sim p\)
- 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.

Traditionally, the word “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 \(\min\{u_1, \ldots, u_p\} = 0\);

If \(z \sim \text{mvn}(0, \Sigma)\), 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 $\left\{ \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K}{K} \right\}^2$ 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

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

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

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
Multivariate extension

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

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

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 \( p(u|M) \).

Parameters to estimate include \( \mathbf{M} \) and \( K \).
Estimation

1. Obtain “data”: \( u_{i,j} = \hat{F}(y_{i,j}), j \in \{1, 2\}; \)
2. Sancetta and Satchell (2004):
   2.1 Pick \( K \) as a function of \( n \), based on an asymptotic result;
   2.2 Let \( \hat{M} \) be the empirical proportions in the \( K \times K \) bins;
   2.3 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!
3. Maximum likelihood:
   3.1 The parameter space for \( \mathbf{M} \) is a compact convex set.
   3.2 Use Newton’s method with a logarithmic barrier to minimize
   \( -\sum_{i=1}^{n} \log p(u_{i,1}, u_{i,2}|\mathbf{M}). \)
   3.3 Compare values of \( K \) using AIC, BIC or something similar.

Step 1 is problematic for discrete data.
The extended rank likelihood provides an alternative.
Bernstein - Gaussian comparison

![Heatmaps showing the comparison between different variables (AGE, FINCOME, DEGREE, WORDSUM, BIBLE, NEARGOD).](image-url)
Summary and future work

Summary

▶ Extended rank likelihood provides inference for dependence parameters, treating the marginals as nuisance parameters.
▶ Bayesian estimation for $C$ is easy to implement.
▶ Performance is good in small and large samples.
▶ The rank likelihood can be used for other copula models.

Future work

▶ Study the information properties/asymptotics of the likelihood.
▶ Combine this approach with one for (non-ordinal) categorical data.
▶ Identify principled ways of smoothing the Bernstein copula model.