Summary of *Extending the Rank Likelihood for Semiparametric Copula Estimation*, by Peter Hoff

David Gerard

Department of Statistics
University of Washington
gerard2@uw.edu

April 16, 2013

Setup

- Modeling conditional associations are very important.
- General Social Survey
- Standard practice is to use regression models.
- If the regression coefficient is not significantly non-zero, standard practice is to conclude the two variables are conditionally independent, given all the other variables.

INC	Respondent's Income
DEG	Highest degree obtained
CHILD	Number of children
PINC	Parent's income when respondent was 16
PDEG	Max(mother's degree, father's degree, na.rm = T)
PCHILD	Number of siblings + 1
AGE	Age of respondent

Which do we choose?

$$
\begin{aligned}
INC_i = &\beta_0 + \beta_1 CHILD_i + \beta_2 DEG_i + \beta_3 AGE_i \\
&+ \beta_4 PCHILD_i + \beta_5 PINC_i + \beta_6 PDEG_i + \epsilon_i
\end{aligned}
$$

or

$$
\begin{aligned}
CHILD_i \sim &\text{Pois}(\exp\{\beta_0 + \beta_1 INC_i + \beta_2 DEG_i + \beta_3 AGE_i \\
&+ \beta_4 PCHILD_i + \beta_5 PINC_i + \beta_6 PDEG_i\})
\end{aligned}
$$

It matters

Response	INC	CHILD	DEG	AGE
INC	NA	**1.103(0.112)**	7.025(<0.001)	0.335(<0.001)
CHILD	**0.005(0.009)**	NA	-0.068(0.056)	0.037(<0.001)

Response	PCHILD	PINC	PDEG
INC	0.284(0.407)	4.070(0.001)	1.399(0.115)
CHILD	0.021(0.080)	-0.063(0.195)	-0.051(0.204)

Which variable you choose as the response can lead to different conclusions!
Jointly modeling the variables of interest helps.

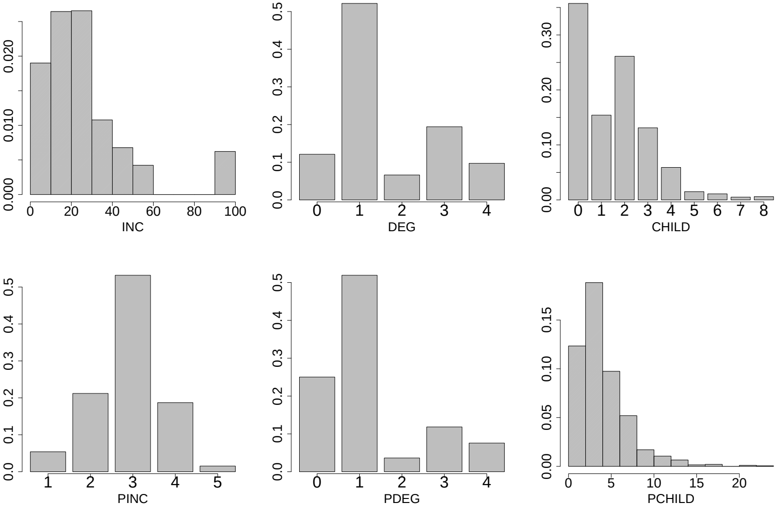
Copulas (Copulae)

- A copula is any multivariate distribution with uniform marginals.
- Sklar's Theorem: Any multivariate c.d.f. $H(x_1, \ldots, x_p) = Pr(X_1 \leq x_1, \ldots, X_p \leq x_p)$ of a random vector $(X_1, \ldots, X_p)$ with marginals $F_i(x_i) = Pr(X_i \leq x_i)$ can be written as $H(x_i, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p))$, where C is a copula. (C is unique if all marginals are continuous)

Problem

- Univariate marginals hard to estimate (i.e. don't belong to standard families).
- Still want to describe dependence structure.
- General Social Survey Example

Marginal Distributions of Variables in G.S.S.



Previous Work

- Genest, Ghoudi, and Rivest (1995) – semiparametric approach where they just plugged in empirical cdf's as the marginals
- Olsson (1979) – latent gaussian variables for ordinal data
- Both semi-parametric approaches (parametric in the copula, non-parametric in the marginals).

New Approach

Gaussian Copula Sampling Model

$$\mathbf{z_1}, \ldots, \mathbf{z_n} | \mathbf{C} \sim \textit{i.i.d. multivariate normal}(\mathbf{0}, \mathbf{C})$$

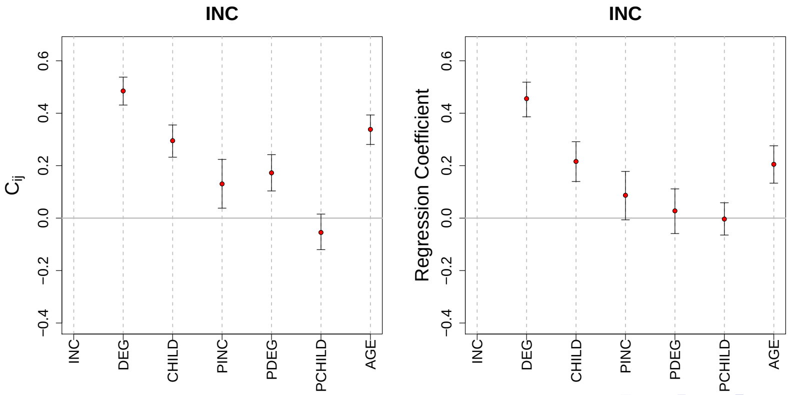
$$y_{i,j} = F_j^{-1}[\phi(z_{i,j})]$$

- Use only the partial ordering of the z's induced by the observed values of the y's. I.e., given $\mathbf{Y} = (\mathbf{y_i}, \ldots, \mathbf{y_n})^T$, $\mathbf{Z} = (\mathbf{z_1}, \ldots, \mathbf{z_n})^T$ is in the set

$$D := \left\{ \mathbf{Z} \in \mathbb{R}^{n \times p} : \max\{z_{k,j} : y_{k,j} < y_{i,j}\} < z_{i,j} < \min\{z_{k,j} : y_{i,j} < y_{k,j}\} \right\}$$

- And use the likelihood $P(\mathbf{Z} \in D | \mathbf{C})$, which depends only on the association parameters.
- Can use, e.g., maximum likelihood or Bayesian approaches using this likelihood.

- Using a Gibbs sampler, we can do inference on the z level about the correlation parameters.
- Or we can do inference on the "regression parameters", $\mathbf{C}_{[j,-j]}\mathbf{C}^{-1}_{[-j,-j]}$.

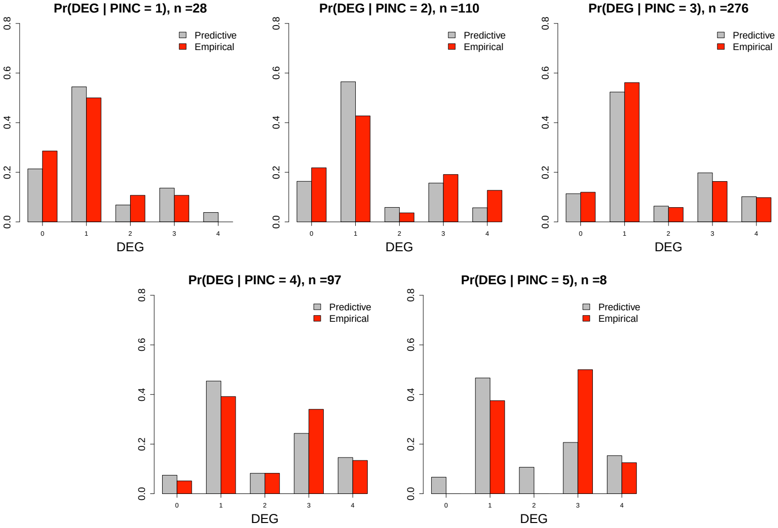


Posterior Predicitive

- We can also sample from a posterior predictive distribution to do inference on y's.
- Compare to empirical distributions.

1. sample $\mathbf{C} \sim p(\mathbf{C}|\mathbf{Z} \in D)$;
2. sample $\mathbf{z} \sim$ multivariate normal$(\mathbf{0}, \mathbf{C})$;
3. set $y_j = \hat{F}_j^{-1}(\Phi(z_j))$.

Posterior Predictive



Notes on the "Likelihood"

- Sort of like a marginal likelihood (c.f. Wakefield pp46-47).

$$
\begin{aligned}
P(\mathbf{Y}|\mathbf{C}, F_i, \ldots, F_p) &= P(\mathbf{Z} \in D, \mathbf{Y}|\mathbf{C}, F_i, \ldots, F_p) \\
&= P(\mathbf{Z} \in D|\mathbf{C}) \times P(\mathbf{Y}|\mathbf{Z} \in D, \mathbf{C}, F_i, \ldots, F_p)
\end{aligned}
$$

- Using this "marginal likelihood" means we don't have to estimate the nuisance parameters.
- Is there a cost? (Not using all of the data)
- The partial ordering is not sufficient, but perhaps "partially" sufficient.

References

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995).
A semiparametric estimation procedure of dependence parameters in multivariate families of distributions.
Biometrika, 82(3):543–552.

Hoff, P. D. (2007).
Extending the rank likelihood for semiparametric copula estimation.
The Annals of Applied Statistics, pages 265–283.

Olsson, U. (1979).
Maximum likelihood estimation of the polychoric correlation coefficient.
Psychometrika, 44(4):443–460.