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

David Gerard

Department of Statistics
University of Washington
gerard2@uw.edu

May 21, 2013

Setup

- Modeling conditional associations are very important.
- General Social Survey
- Standard practice is to use regression models.
- If p-value of regression coefficient is greater than 0.05, then we take that to mean there is not significant evidence against the regression coefficient being non-zero. Model implies two variables are conditionally independent.

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)

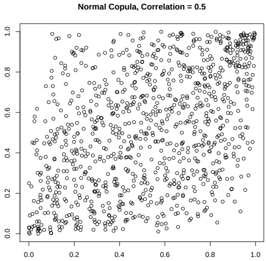
Response	PCHILD	PINC	PDEG
INC	0.284(0.407)	4.070(0.001)	1.399(0.115)
CHILD	0.021(0.080)	-0.063(0.195)	-0.051(0.204)

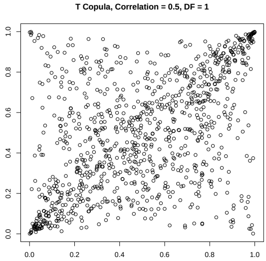
Which variable you choose as the response can lead to different conclusions!

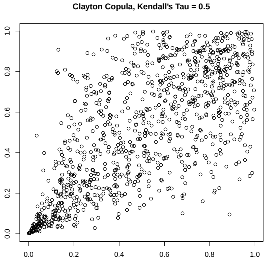
Jointly modeling the variables of interest helps.

Copulas (Copulae)

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



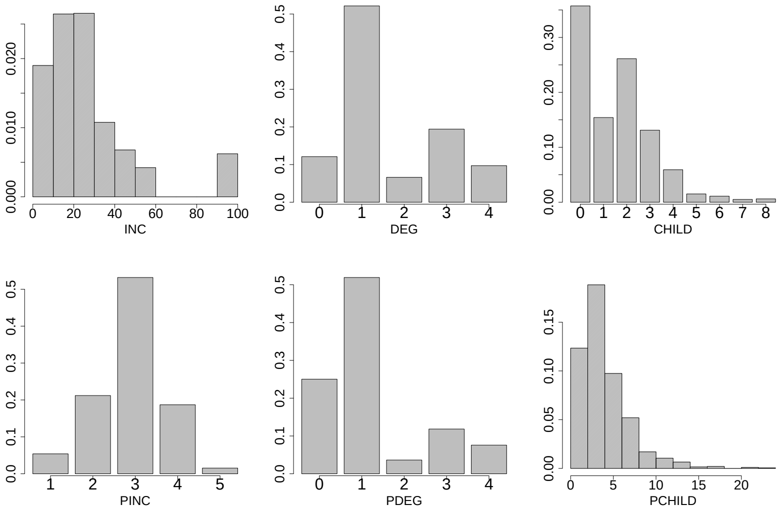




Problem

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

Marginal Distributions of Variables in G.S.S.



Previous Work

- [Genest et al., 1995] – semiparametric approach where they just plugged in empirical c.d.f.'s as the marginals
- [Olsson, 1979] – latent Gaussian variables for ordinal data
- [Poon and Lee, 1987] – extends [Olsson, 1979] to case when all marginals of continuous variables are normal.

Setup

Let $y_{i,j}$ be the value of the j^{th} variable taken on by the i^{th} observational unit.

Gaussian Copula Sampling Model

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

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

- Estimate association parameters without having to estimate marginals.
- Do this by only using the partial ordering induced by the data: $y_{k,j} < y_{i,j} \Rightarrow z_{k,j} < z_{i,j}$.
- A partial ordering is a total ordering without the *totality* condition (i.e., some elements may be incomparable).

The Extended Rank Likelihood

Let D be the set of $\mathbf{Z} := (z_{i,j})$'s that satisfy the partial ordering. Use the following "marginal likelihood" for inference:

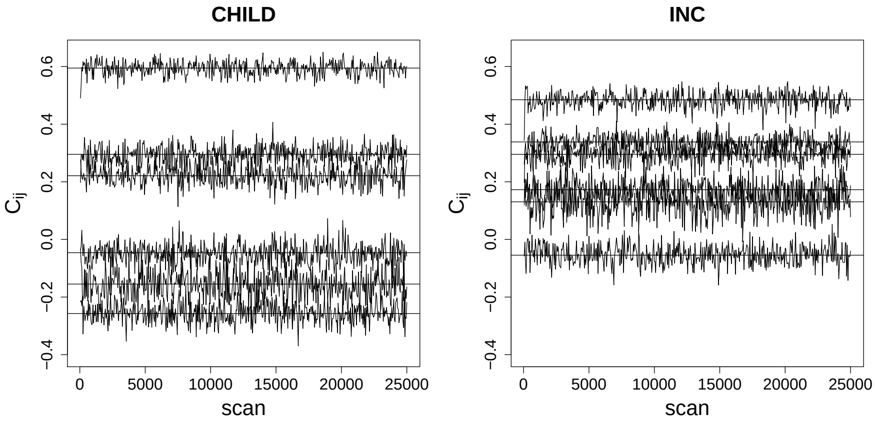
$$Pr(\mathbf{Z} \in D | \mathbf{C}, F_1, \ldots, F_p) = \int_D p(\mathbf{Z}|\mathbf{C}) d\mathbf{Z} = Pr(\mathbf{Z} \in D)$$

Gibbs Sampler

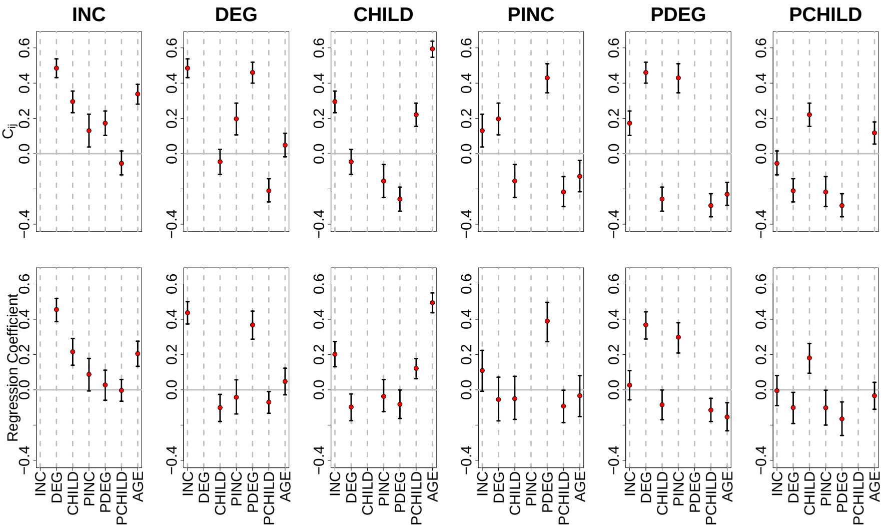
Full conditionals of the $z_{i,j}$'s are easy to derive, so we can implement a Gibb's sampler (using covariance matrix rather than correlation matrix, but doesn't matter for estimation).

1. Sample $z_{i,j}|\mathbf{Z}_{[-i,-j]}, \mathbf{V}$ from a truncated normal with the bounds set by the partial ordering and the conditional mean and variance found in the usual way: $\sigma_j^2 = \mathbf{V}_{[j,j]} - \mathbf{V}_{[j,-j]}\mathbf{V}_{[-j,-j]}^{-1}\mathbf{V}_{[-j,j]}$ and $\mu = \mathbf{V}_{[j,-j]}\mathbf{V}_{[-j,-j]}\mathbf{Z}_{[i,-j]}^{T}$.
2. Sample $\mathbf{V}$ from an inverse-Wishart distribution (if you use the conjugate prior).
3. Let $\mathbf{C}_{[i,j]} = \mathbf{V}_{[i,j]}/\sqrt{\mathbf{V}_{[i,i]}\mathbf{V}_{[j,j]}}$

Some Trace Plots



95% Posterior Credible Intervals



Visualizing Conditional Dependencies of GSS data.

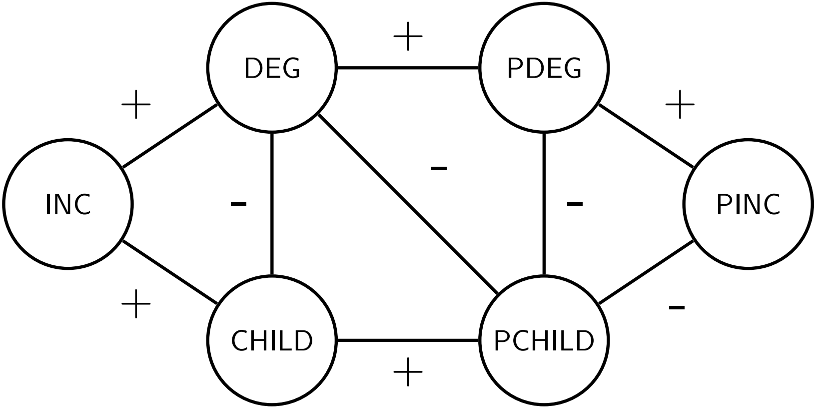


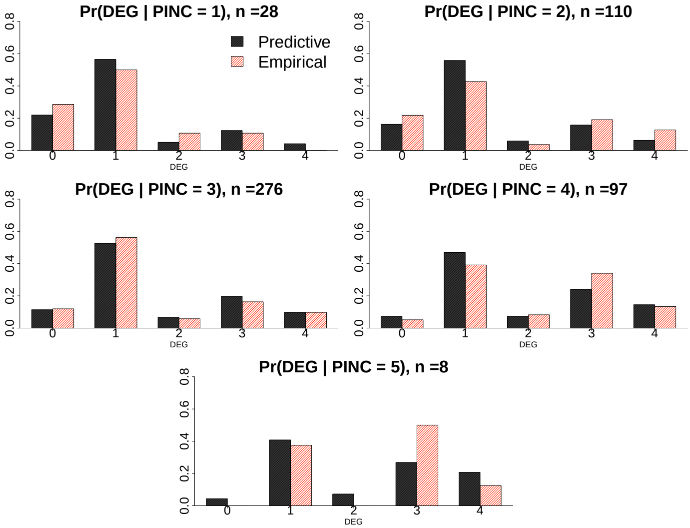
Figure: Reduced conditional dependence graph for the General Social Survey data.

Posterior Predictive

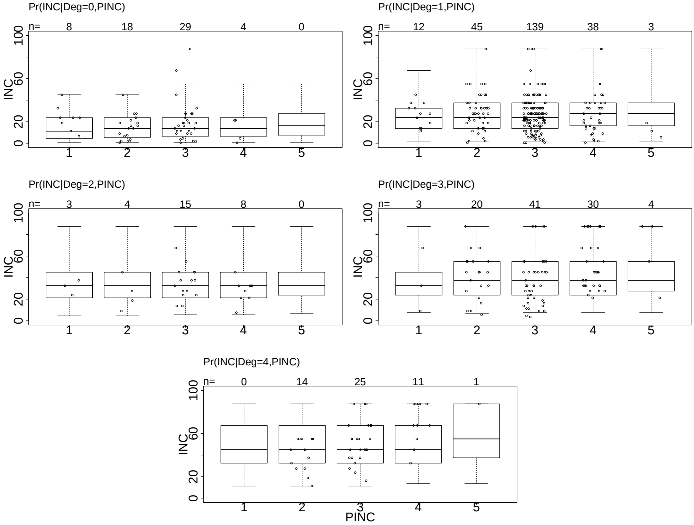
- We can also sample from a posterior predictive distribution to do inference on y's.
- Compare to empirical distributions for model checking.

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 and Empirical distributions of DEG given PINC



Posterior Predictive and Empirical distributions of INC given DEG and PINC



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?
- The partial ordering is not sufficient, but perhaps "partially" sufficient.

Partial Sufficiency

- The paper proves that ranks are "G-sufficient" and "L-Sufficient" when we have continuous marginals.
- However, when the data are discrete the partial ordering does not have either of these properties.

Groups

Definition

A collection $\mathcal{G}$ of 1-1 transformations of $\mathcal{X}$ (the sample space) is a group if

1. For all $g_1, g_2 \in \mathcal{G}$, $g_1 g_2 \in \mathcal{G}$
2. For all $g \in \mathcal{G}, g^{-1} \in \mathcal{G}$.

Definition

Two points $x_1, x_2 \in \mathcal{X}$ are equivalent if there exists a $g \in \mathcal{G}$ such that $x_1 = gx_2$. The sets of equivalent points are the orbits of $\mathcal{G}$.

Definition

A function M is said to be maximally invariant if it is in 1-1 correspondence with the orbits of $\mathcal{G}$. i.e. $M(gx) = M(x)$ for all $g \in \mathcal{G}$ and $M(x_1) = M(x_2) \Rightarrow x_2 = gx_1$ for some $g \in \mathcal{G}$.

Groups

- $\mathcal{G}$ induces a group $\bar{\mathcal{G}}$ over the parameter space.
- Let $\bar{\mathcal{G}} = \{\bar{g} : \Omega \to \Omega \text{ s.t. } \bar{g}\theta = \theta' \text{ if } X \sim P_\theta \text{ and } gX \sim P_{\theta'}\}$.
- There are also maximally invariant parameters for $\bar{\mathcal{G}}$.

G-Sufficiency

- Under a group of transformations, $\mathcal{G}$, the maximally invariant statistic is called "G-sufficient" for the maximally invariant parameter.
- The ranks are the maximally invariant statistics under the group of continuous strictly increasing functions [Lehmann and Romano, 2005, pp 215 - 216]. You can also put the correlation matrix in a 1-1 correspondence with the induced group $\bar{\mathcal{G}}$'s orbits (but only if the marginals are continuous).
- Hence, the ranks are "G-sufficient" for the correlation matrix.
- Intuition: if we assume the marginals are unknown, then applying strictly increasing continuous functions to the data should not change the estimation problem

L-Sufficiency

- For $\{F_1, \ldots, F_p\} \in \mathcal{F}$ the marginal distributions, a statistic $t(\mathbf{Y})$ is L-sufficient for $\mathbf{C}$ if
  1. $t(\mathbf{Y}_0) = t(\mathbf{Y}_1) \Rightarrow \sup_{\{F_1,\ldots,F_p\}\in\mathcal{F}} p(\mathbf{Y}_0|\mathbf{C}, F_1, \ldots, F_p) = \sup_{\{F_1,\ldots,F_p\}\in\mathcal{F}} p(\mathbf{Y}_1|\mathbf{C}, F_1, \ldots, F_p)$; and
  2. $p(t(\mathbf{Y})|\mathbf{C}, F_1, \ldots, F_p) = p(t(\mathbf{Y})|\mathbf{C})$
- If there are no nuisance parameters, L-sufficiency becomes "full sufficiency".
- Intuition: partition generated by L-sufficient statistic is "at least as fine" as the partition generated by the M.L.E. of $\mathbf{C}$ [Rémon, 1984] (i.e. MLE is function of ranks alone).

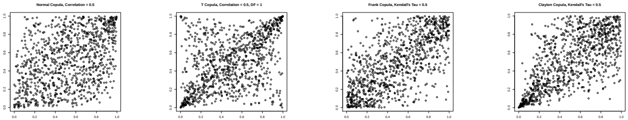
Posterior Consistency

- even though the partial ordering does not have any sufficiency results, it has the nice property of having its distribution be independent of nuisance parameters.
- A recent paper [Murray et al., 2013] also proved that using the extended rank likelihood will result in posterior consistency of C.

Simulations: Set Up

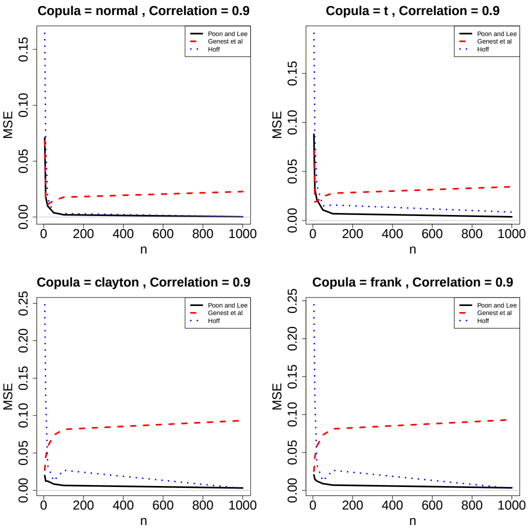
- Copula:

Normal T (df = 1) Frank and Clayton

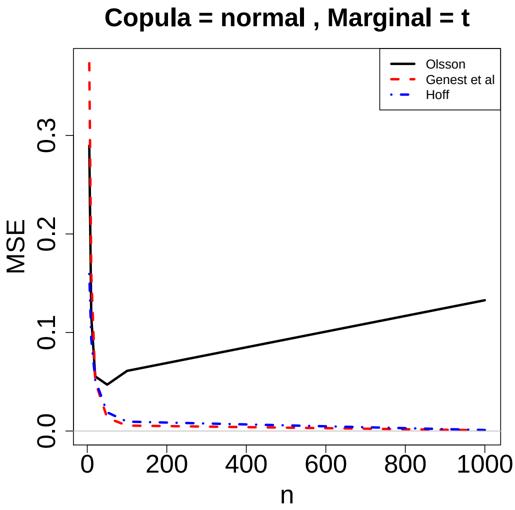


- Correlation: $\{0.0, 0.5, 0.9\}$
  Used Pearson correlation (ρ) for Normal and T copulas, and Kendall's Tau (τ) for Frank and Clayton Copulas (estimator then using all three methods becomes $\hat{\tau} = \frac{2}{\pi}\arcsin(\hat{\rho})$)
- Continuous marginals: Normal(0,1), T(df $=$ 1), Gamma(1,1)
- Discrete marginal set at Bern(1/2)
- n: $\{5, 10, 20, 50, 100, 1000\}$

Misspecified Copulas, ρ or $\tau = 0.9$



Misspecified Marginals, $\rho = 0.5$



Summary of Simulations

- Polyserial correlation worked extremely well when the copula was mispecified, but performed poorly when the continuous marginals were no longer normal.
- The method of [Genest et al., 1995] works poorly when the marginals are discrete.
- [Hoff, 2007]'s method under-performs in some scenarios, but does not have the horrible behavior that [Poon and Lee, 1987] and [Genest et al., 1995] showed (at least not under any of the scenarios I tried out).

References I

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.

Lehmann, E. E. L. and Romano, J. P. (2005).
Testing statistical hypotheses.
Springer Science+ Business Media.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013).
Bayesian gaussian copula factor models for mixed data.
Journal of the American Statistical Association, (just-accepted).

References II

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

Poon, W.-Y. and Lee, S.-Y. (1987).
Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients.
Psychometrika, 52(3):409–430.

Rémon, M. (1984).
On a concept of partial sufficiency: L-sufficiency.
International Statistical Review/Revue Internationale de Statistique, pages 127–135.