A goodness-of-fit test for Archimedean copula models in the presence of censoring

To appear in

Computational Statistics and Data Analysis

Takeshi Emura* (江村剛志)、Weijing Wang (王維菁)
Institute of Statistics, National Chiao Tung University

Chien-Wei Lin (林建威)
Institute of Statistical Science, Academia Sinica

Table of Contents

- Part I: Background (5 slides)
  - Bivariate lifetime data
  - Copula model
  - Goodness-of-fit for bivariate lifetimes
  - Shih (1998)'s goodness-of-fit test under Clayton model

- Part II: Proposed Method (10 slides)
  - Goodness-of-fit test under general AC model
  - Asymptotic analysis
  - Modification to censoring
  - Data analysis & Simulations
  - Concluding remarks

Background

- Australian Twin Study Data

(Duffy et al. 1990, downloadable from Web)

Bivariate lifetimes: $\{(X_j, Y_j)\,(j=1,\ldots,n)\}$

For Twin pair j:

$$\begin{cases} X_j : \text{Age at appendecectomy for one of twin pairs} \\ Y_j : \text{Age at appendecectomy for the other one of twin pairs} \end{cases}$$

*Appendecetomy: 盲腸手術

1. Correlation between X and Y may be of interest
2. Prentice & Hsu (1997) fitted Clayton model without model diagnostics
3. Some subject never experience appendecectomy (right-censoring)

Background - Copula -

- By Skalar (1959)'s theorem, any joint distribution of (X, Y) has a representation:

$$\Pr(X > x, Y > y) = C[S_X(x), S_Y(y)],$$

where $S_X(x) = \Pr(X > x)$, $S_Y(y) = \Pr(Y > y)$ and the function $C[u, v]$ is called "Copula".

- $C[u, v]$ characterize the association between X and Y

e.g., Clayton copula (Clayton, 1978):

$$C[u, v] = [u^{-(\alpha-1)} + v^{-(\alpha-1)} - 1]^{-\frac{1}{\alpha-1}} \quad \Rightarrow \text{Kendall's tau} = \frac{\alpha - 1}{\alpha + 1}$$

- Model selection: How to select $C[u, v]$ given data without specifying marginal functions

(Likelihood based approach, such as AIC do not apply).

Background -Archimedean copula -

- For some function $\phi_\alpha(\cdot)$, consider a subclass

$$C[u,v] = \phi_\alpha^{-1}[\phi_\alpha(u) + \phi_\alpha(v)]$$

, called Archimedean copula (AC) family.

Parameter α is called association parameter

$*$ Eample 1. Clayton copula :

$$\phi_\alpha(t) = (t^{-(\alpha-1)} - 1)/(\alpha - 1), \quad \text{Kendall's tau on (X, Y)} = \frac{\alpha - 1}{\alpha - 1}$$

$*$ Example 2 : Frank copula

$$\phi_\alpha(t) = \log\{(1-\alpha)/(1-\alpha^t)\}, \quad \text{Kendall's tau} = 1 - \frac{4\{D_1(-\log\alpha) - 1\}}{\log\alpha}$$

$*$ Example 3 : Gumbel copula

$$\phi_\alpha(t) = \{-\log(v)\}^\alpha, \quad \text{Kendall's tau} = \frac{\alpha\text{ -1}}{\alpha}$$

Background

- Our problem is to test wether a chosen ϕ_α fit data well under bivariate censored data.
- Many papers discuss this problem under complete data But, there is few paper on bivariate censored data

1.Model selection(Wang & Wells, 2000):

How to select the best $\phi_\alpha(\cdot)$ among several candidates.

2.Goodness-of-fit test (Andersen et al. 2005):

Statistical test for checking whether a selected $\phi_\alpha(\cdot)$ is correct or not.

* Ansersen et al.'s test is based on chi-square tests for comparing model based vs. model free estimates of the copula function.

Bootstrap is used for fining cutoff since the null distribution is unknown.

Background - Shih's idea -

- $\text{Early work starts from Clayton model:}$

$$\Pr(X > x, Y > y) = \{S_X(x)^{-(\alpha-1)} + S_X(x)^{-(\alpha-1)} - 1\}^{-1/(\alpha-1)}$$

- Clayton (1979) proposed a conditional likelihood estimator $\hat{\alpha}_1$ while S_X and S_Y to be completelyunspecified
- Oakes (1982) proposed a moment-type estimator $\hat{\alpha}_2$ while S_X and S_Y to be completelyunspecified
- Oakes (1986) showed $\hat{\alpha}_2$ and $\hat{\alpha}_1$ belong to the same estimating function, with the different weight
- Shih (1998)consider a distance: $|\hat{\alpha}_1 - \hat{\alpha}_2|$

Weighted　　　　Un-weighted

Then, reject the Clayton model if $|\hat{\alpha}_1 - \hat{\alpha}_2|$ is large

Proposed method: Setup

- Temporality ignore censoring so that we observe complete data : $\{(X_j, Y_j)\ (j = 1, \ldots, n)\}$
- (X_j, Y_j) are i.i.d. replica from

$$\Pr(X > x, Y > y) = C[S_X(x), S_Y(y)]$$

where parameter (S_X, S_Y) is unspecified

- We are interested in testing

$$\mathrm{H_0} : C[u, v] = \phi_\alpha^{-1}[\phi_\alpha(u) + \phi_\alpha(v)] \quad \text{for some } \alpha$$

versus

$$\mathrm{H_a} : C[u, v] = \text{any other copula.}$$

Here, α is unknown.

Proposed method: Basic Idea

α : Association parameter in $\phi_\alpha(\cdot)$

- Consider a distance : $|\hat{\alpha}_1 - \hat{\alpha}_2|$

Weighted Un-weighted

- $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are shown to belong to the same class, but differ only in weight (I explain later for details)

- Both $\hat{\alpha}_1$ and $\hat{\alpha}_2$ converges to the true α if $\phi_\alpha(\cdot)$ is correctly specified

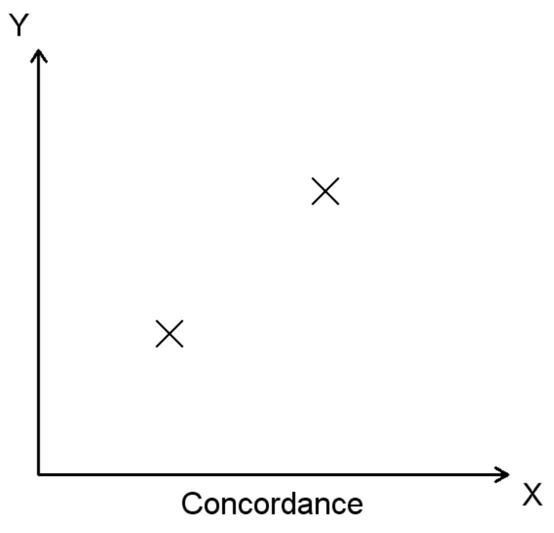
- Reject the model $\phi_\alpha(\cdot)$ if $|\hat{\alpha}_1 - \hat{\alpha}_2|$ is large

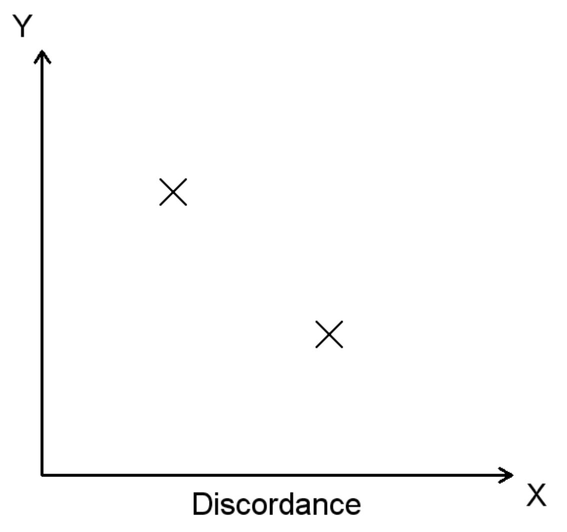
Proposed method

How to estimate α?

- Consider a concordance indicator

$$\Delta_{ij} = I\{(X_i - X_j)(Y_i - Y_j) > 0\}$$





- Information for α is contained in Δ_{ij}

$\Rightarrow$ Moment estimator based on Δ_{ij}

Proposed method

- Oakes (1989) show that, if $\phi_\alpha(\cdot)$ is correctly specified

$$E(\Delta_{ij} \mid \breve{X}_{ij} = x, \tilde{Y}_{ij} = y) = \frac{\theta_\alpha\{S(x,y)\}}{1+\theta_\alpha\{S(x,y)\}}$$

$$\text{where } \breve{X}_{ij} = \min(X_i, X_j),\ \tilde{Y}_{ij} = \min(Y_i, Y_j)$$

$$S(x,y) = \Pr(X > x, Y > y) \text{ and } \theta_\alpha(\eta) = -\eta \frac{\phi''_\alpha(\eta)}{\phi'_\alpha(\eta)}$$

- Estimating equation for α

$$U_2(\alpha) = \sum_{i<j} \left[\Delta_{ij} - \frac{\theta_\alpha\{\hat{S}(\breve{X}_{ij}, \tilde{Y}_{ij})\}}{1+\theta_\alpha\{\hat{S}(\breve{X}_{ij}, \tilde{Y}_{ij})\}} \right]$$

$$\text{where } \hat{S}(x,y) = n^{-1} \sum_i I(X_i \geq x, Y_i \geq y)$$

- Unweighted estimator $\hat{\alpha}_2 : U_2(\alpha) = 0$

Proposed method

- To derive weighted estimator, we extends Clayton (1979)'s likelihood principle (details, omitted)

- Estimating equation based on generalized Clayton's likelihood is

$$U_1(\alpha)=$$

$$\sum_{i<j}\frac{\dot{\theta}_\alpha\{\hat{S}(\tilde{X}_{ij},\tilde{Y}_{ij})\}[\theta_\alpha\{\hat{S}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1]}{\theta_\alpha\{\hat{S}(\tilde{X}_{ij},\tilde{Y}_{ij})\}[R_{ij}-1+\theta_\alpha\{\hat{S}(\tilde{X}_{ij},\tilde{Y}_{ij})\}]}\left[\Delta_{ij}-\frac{\theta_\alpha\{\hat{S}(\tilde{X}_{ij},\tilde{Y}_{ij})\}}{\theta_\alpha\{\hat{S}(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}\right]$$

where $\tilde{R}_{ij}=nS(\tilde{X}_{ij},\tilde{Y}_{ij})$

- Weighted estimator $\hat{\alpha}_1 : U_1(\alpha)=0$
- The above weight derivation is new result in this work. It gives smaller SD than unweighted estimator

Proposed method: Asymptotic Analysis

- Theorem 1: Under correct model and suitable conditions,

$$n^{1/2}(\log\hat{\alpha}_1 - \log\hat{\alpha}_2) \to N(0,\sigma^2)$$

where $\sigma^2 = 4E[h\{(X_1,Y_1),(X_2,Y_2)\}h\{(X_1,Y_1),(X_3,Y_3)\}],$

$$h\{(X_i,Y_i),(X_j,Y_j)\} \equiv \frac{1}{\alpha}\left(\frac{\dot{\theta}_\alpha\{S(\tilde{X}_{ij},\tilde{Y}_{ij})\}[\theta_\alpha\{S(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1]}{A_L\theta_\alpha\{S(\tilde{X}_{ij},\tilde{Y}_{ij})\}S(\tilde{X}_{ij},\tilde{Y}_{ij})} - \frac{1}{A}\right)\left[\Delta_{ij} - \frac{\theta_\alpha\{S(\tilde{X}_{ij},\tilde{Y}_{ij})\}}{\theta_\alpha\{S(\tilde{X}_{ij},\tilde{Y}_{ij})\}+1}\right],$$

$$A \equiv E\left(\frac{\dot{\theta}_\alpha\{S(\tilde{X}_{12},\tilde{Y}_{12})\}}{[\theta_\alpha\{S(\tilde{X}_{12},\tilde{Y}_{12})\}+1]^2}\right) \text{ and } A_L \equiv E\left(\frac{[\dot{\theta}_\alpha\{S(\tilde{X}_{12},\tilde{Y}_{12})\}]^2}{\theta_\alpha\{S(\tilde{X}_{12},\tilde{Y}_{12})\}[\theta_\alpha\{S(\tilde{X}_{12},\tilde{Y}_{12})\}+1]}\right)$$

- Reject $H_0 : C(u,v) = \phi_\alpha^{-1}[\phi_\alpha(u) + \phi_\alpha(v)]$

  if $|(\log\hat{\alpha}_1 - \log\hat{\alpha}_2)/\hat{\sigma}| > 1.96$

Proposed method: Adjustment for Censoring

- If lifetimes(X_j, Y_j) is censored by (A_j, B_j), we observe $(\breve{X}_j, \breve{Y}_j, \delta_j^X, \delta_j^Y)$

where $\breve{X}_j = \min(X_j, A_j), \breve{Y}_j = \min(Y_j, B_j), \delta_j^X = I(X_j \le A_j), \delta_j^Y = I(Y_j \le B_j)$

- Following Oakes (1986), the estimating functions can be modified as

$$U_1(\alpha) =$$

$$\sum_{i<j} Z_{ij} \frac{\dot{\theta}_\alpha\{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}[\theta_\alpha\{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}+1]}{\theta_\alpha\{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}[R_{ij} - 1 + \theta_\alpha\{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}]} \left[\Delta_{ij} - \frac{\theta_\alpha\{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}}{\theta_\alpha\{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}+1} \right],$$

$$U_2(\alpha) = \sum_{i<j} Z_{ij} \left[\Delta_{ij} - \frac{\theta_\alpha\{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}}{\theta_\alpha\{\hat{S}(\tilde{X}_{ij}, \tilde{Y}_{ij})\}+1} \right],$$

Estimating equations are unbiased
Under independent censoring assumption

where

$$\Delta_{ij} = I\{(\breve{X}_i - \breve{X}_j)(\breve{Y}_i - \breve{Y}_j) > 0\}, \tilde{X}_{ij} = \min(\breve{X}_i, \breve{X}_j), \tilde{Y}_{ij} = \min(\breve{Y}_i, \breve{Y}_j),$$

$$Z_{ij} = I(\tilde{X}_{ij} \le \tilde{A}_{ij}, \tilde{Y}_{ij} \le \tilde{B}_{ij})$$

- Reject $H_0 : C(u,v) = \phi_\alpha^{-1}[\phi_\alpha(u) + \phi_\alpha(v)]$

  if $|(\log \hat{\alpha}_1 - \log \hat{\alpha}_2)/\hat{\sigma}| > 1.96$

Proposed method: Data analysis

Table 3A: The Goodness-of-fit test results for four AC models based on Australian Twin Study (Duffy et al. 1990)

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$(\log\hat{\alpha}_1 - \log\hat{\alpha}_2)/\hat{\sigma}_{Jack}$	p-value
Clayton	1.446	1.717	-1.867	0.000
Frank	1.308	1.496	-1.090	0.117
Gumbel	0.115	0.114	0.084	0.497
Log-copula	1.447	1.147	1.351	0.034

*Gumbel copula is the best fitted model.

*Analysis of Prentice & Hsu (1998) under Clayton copula model is questionable. Re-anaysis under Gumbel model is suggested.

* P-values are not adjusted for multiple testing

Proposed method: Simulations

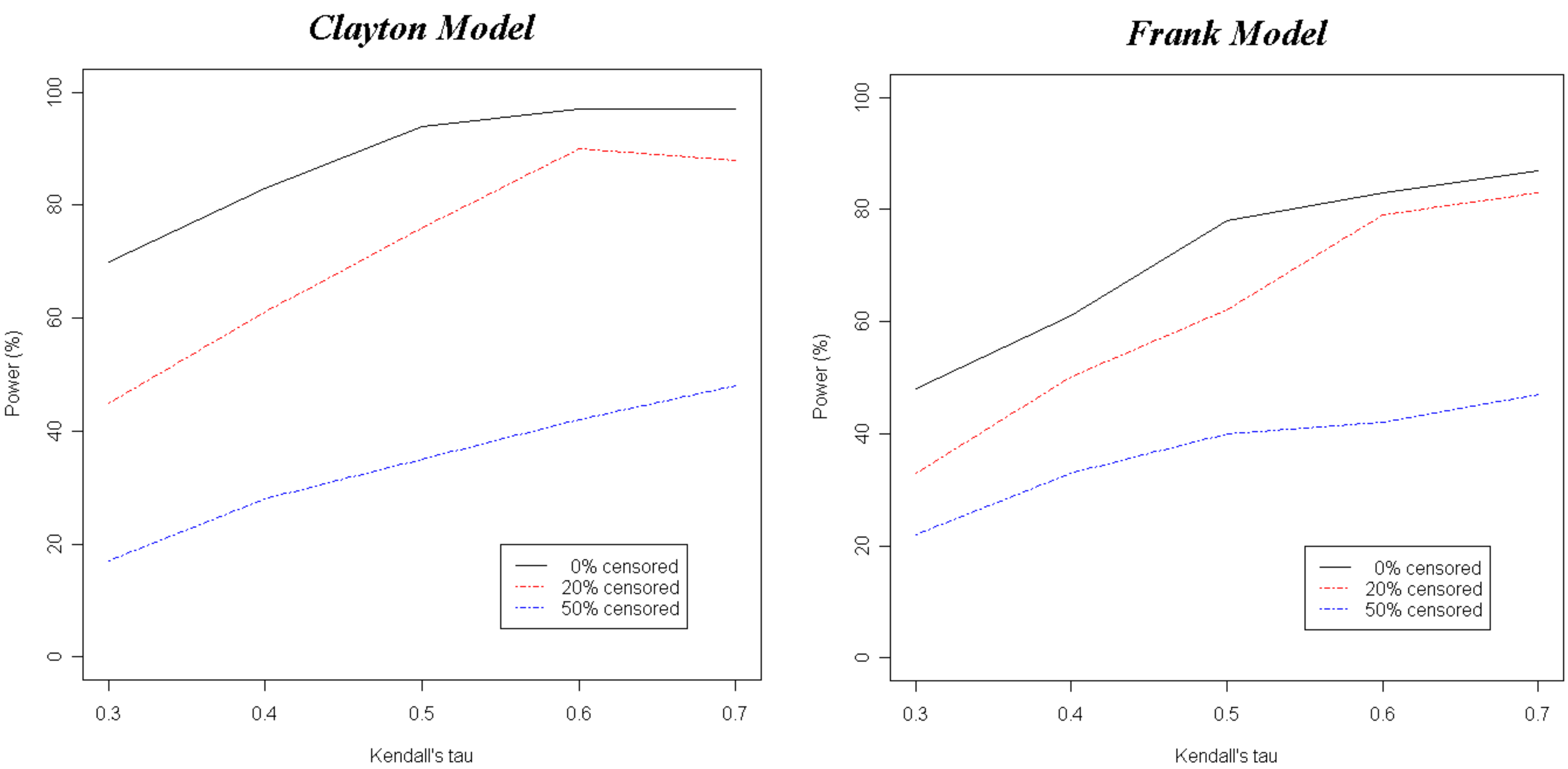


Fig. 1A: Empirical powers with n=100 under H_0 : Gumbel vs. H_a : Not Gumbel. Powers are the rates of rejecting H_0 with 5% significance during 100 replications.

Concluding remarks

- We proposed a goodness-of-fit test based on the distance between two points estimator
- Mean-zero property of the asymptotic null distribution lead to a simple test statistics
- The method can handle independent right-censoring, by applying Oakes (1986)'s idea
- The methods is empirically valid even under dependent right-censoring (robustness)

  - Under dependent censoring, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are *biased* estimator. Nevertheless,

    $$\log \hat{\alpha}_1 - \log \hat{\alpha}_2$$

    still follows zero-mean distribution since the bias cancel out (we prove this by simulations in the paper).

Thank you for your attention