第十六屆南區統計研討會議

A class of Log-rank test for quasi-independence of truncation variable

Takeshi Emura（江村剛志）
National Chiao Tung University

Table of Contents

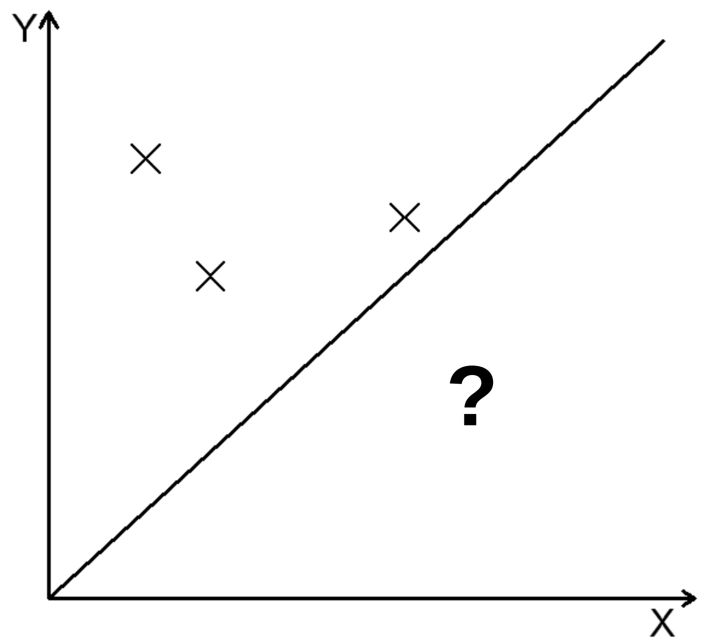
1. Truncation data: overview
2. Quasi-independence
3. Tsai's test & U-statistics test
4. Proposed log-rank test
5. Proposed conditional score test
6. Asymptotic analysis & simulation

Background

- **Truncation Data**:

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

subject to $X_j \leq Y_j$

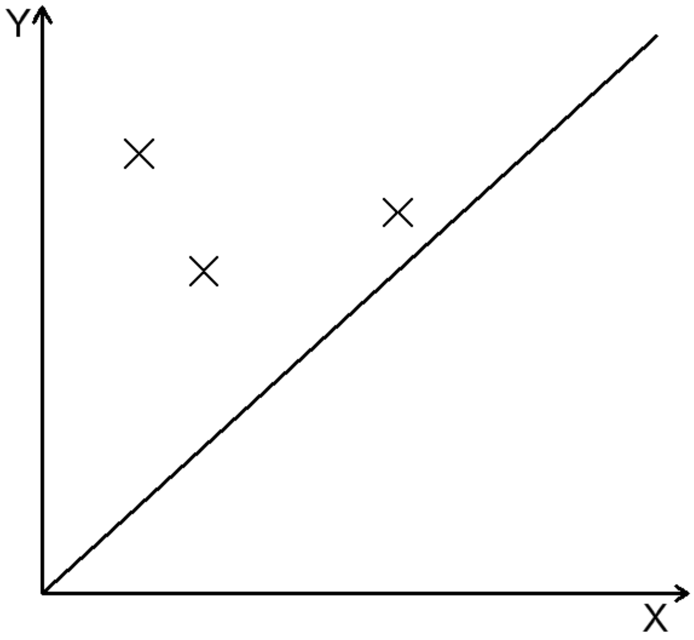


Independent truncation

- Traditionally, we assume

$$\Pr(X \le x, Y \le y) = \Pr(X \le x)\Pr(Y \le y)$$

(not testable by data)



Identifiability

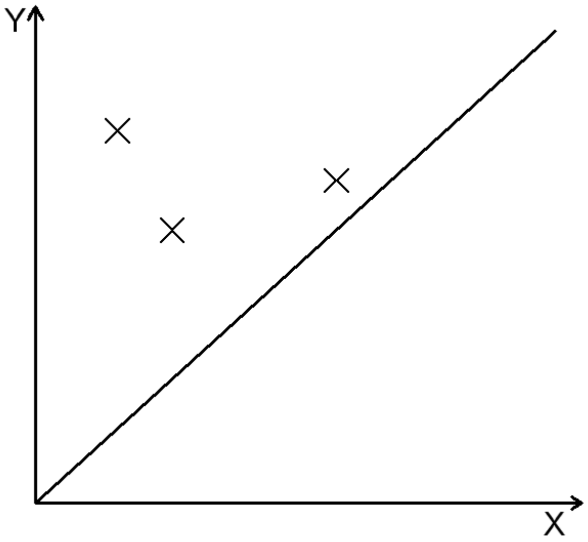
- Any assumption on

$$\Pr(X \le x, Y \le y)$$

$$\Pr(X \le x)$$

$$\Pr(Y \le y)$$

is not identifiable by data

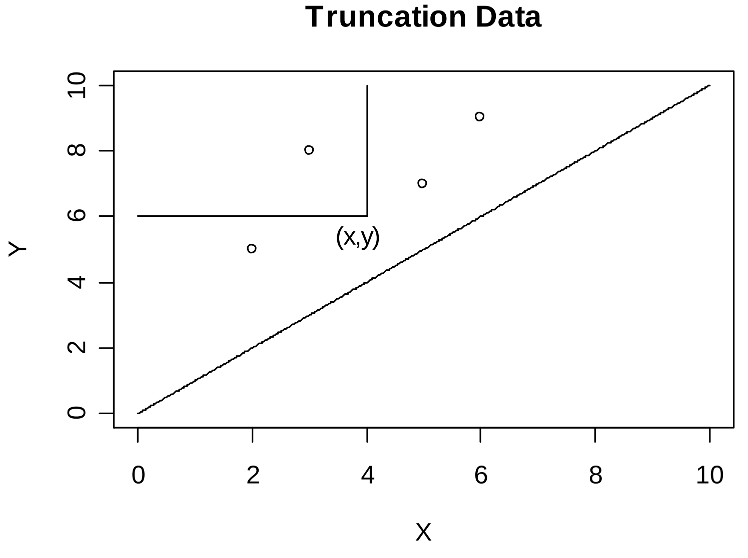


Truncated distribution

The identifiable function

$$\pi(x,y) = \Pr(X \le x, Y > y \mid X \le Y) \qquad (x \le y)$$

determines the distribution for (X, Y) on the upper wedge



Quasi-independence

- Definition: **quasi-independence**

$$H_0 : \pi(x, y) = F_X(x) S_Y(y) / c \qquad (x \le y)$$

where

$$\pi(x, y) = \Pr(X \le x, Y > y \mid X \le Y)$$

$$c = -\iint_{x \le y} dF_X(x) dS_Y(y)$$

- A testable condition
  (identifiable by data)

Previous Results

- Under assumption of quasi-independence between X and Y

  * Estimate $S_Y(t) = \Pr(Y > t)$ (Lynden-Bell's,1971)

  * Estimate $c = \Pr(X < Y)$ (He and Yang, 1998)

- Test of quasi-independence

  * Conditional Kendall's tau (Tsai, 1990)

  * U-statistics test (Martin & Betensky, 2005)

  * Product-moment correlation (Chen et al., 1996)

Testing Independence

- **Data**: $\{(X_j, Y_j)\,(j=1,\ldots,n)\}$

  subject to $X_j \le Y_j$

- **Interest:** Test independence of X and Y

  Existing Results:

  * Tsai test (1990),

  * U-statistics test (Martin & Betensky, 2005)

  * Product moment (Chen et al. 1996)

Construction of a Test Statistic

- Moment-based → nonparametric test
  - Tsai's test (Tsai, 1990)
  - U-statistics test (Martin & Betenski, 2005)
  - Two-by-two Table → log-rank test (Ours)
- Likelihood-based → score test (Ours)

* our proposal for power improvement

Conditional Kendall's tau (measure of dependence)

- Ref: Tsai, 1990

$$\tau_a = E\{\operatorname{sgn}(X_i - X_j)(Y_i - Y_j) \mid A_{ij}\}$$

where $A_{ij} = \{X_i \wedge X_j \le Y_i \wedge Y_j\}$

$$\operatorname{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

- under quasi-independence:

$$\tau_a = 0$$

Tsai's test (1990)

- Test statistics

$$K = \sum_{i<j} I\{A_{ij}\}\operatorname{sgn}\{(X_i - X_j)(Y_i - Y_j)\}$$

- Under quasi-independence $E(K) = 0$
- Write in the form of rank sums

$$K \sim N(0, AVar(K))$$

where, variance can be estimated by

$$AVar(K) \leftarrow \frac{1}{3}\sum_{j}\{R(X_j, X_j)^2 - 1\}$$

U-statistics test (Martin & Betensky, 2005)

- Define U-statistics

$$U_c = \binom{n}{2}^{-1} K = \binom{n}{2}^{-1} \sum_{i<j} I\{A_{ij}\} \mathrm{sgn}\{(X_i - X_j)(Y_i - Y_j)\}$$

- Under quasi-independence $E(U_c) = 0$
- U-statistics CLT

$$U_c \xrightarrow{d} N(0, AVar(U_c))$$

where, empirical variance estimator is available from the U-statistics CLT

Tsai’s test vs. U-statistics test

- Tsai’s test and U-statistics test only differ in the way we calculate the variance estimator
- Tsai’s test does not allow ties
- Tsai’s variance estimator based on exact expression is easily calculated
- U-statistics variance is computationally complicated

(*comments refer from Martin & Betensky, 2005)

Tsai's test & U-statistics test

- Advantages
  - easy to compute statistics
  - a non-parametric test
  - empirical variance estimator
  - asymptotic normality
- Drawbacks
  - power properties are completely unknown
  - not always powerful

Test based on two by two tables

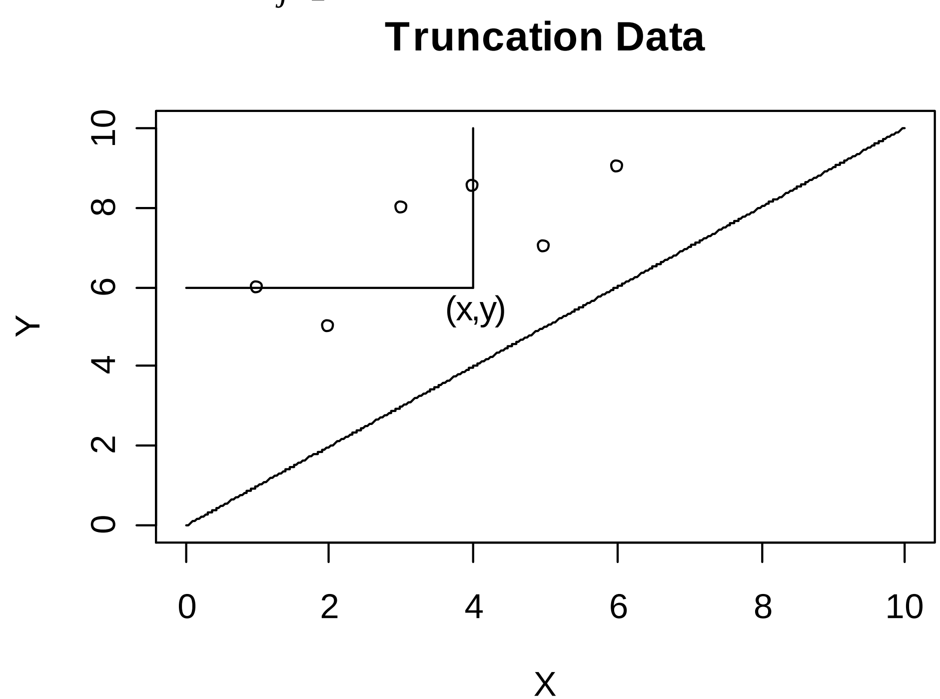
- Counting processes

$$\Delta(x,y)=\sum_{j=1}^{n} I(X_j = x, Y_j = y),\qquad N_{1\bullet}(x,dy)=\sum_{i=1}^{n} I(X_i = x, Y_i \ge y)$$

$$N_{\bullet 1}(x,dy)=\sum_{i=1}^{n} I(X_i \le x, Y_i = y)\qquad R(x,y)=\sum_{j=1}^{n} I(X_j \le x, Y_j \ge y)$$

- Two-by-two tables

	$Y = y$	$Y > y$	
$X = x$	$\Delta(x,y)$		$N_{1\bullet}(dx,y)$
$X < x$			
	$N_{\bullet 1}(x,dy)$		$R(x,y)$



Proposal: Log-rank type test

- Conditional expectation

$$E\{\Delta(x,y)\,|\,\text{margins}\} = \frac{N_{1\bullet}(dx,y)N_{\bullet 1}(x,dy)}{R(x,y)}$$

- Log-rank statistics

$$L_w = \iint_{x \le y} W(x,y)\left[\Delta(x,y) - \frac{N_{1\bullet}(dx,y)N_{\bullet 1}(x,dy)}{R(x,y)}\right]$$

Relationship with discordance test

- Algebraic relation

$$\iint_{x\le y} W(x,y)\left\{\Delta(x,y) - \frac{N_{1\bullet}(dx,y)N_{\bullet 1}(x,dy)}{R(x,y)}\right\}$$

$$= -\sum_{i<j} I\{A_{ij}\}\frac{W(\tilde{X}_{ij},\tilde{Y}_{ij})}{R(\tilde{X}_{ij},\tilde{Y}_{ij})}\mathrm{sgn}\{(X_i - X_j)(X_i - X_j)\}$$

(weighed sign-test)

Relationship with Tsai's test

- If choose "size of the risk set" as the weight

$$\iint_{x \le y} R(x,y)\left\{\Delta(x,y) - \frac{N_{1\bullet}(dx,y)N_{\bullet 1}(x,dy)}{R(x,y)}\right\}$$

$$= -\sum_{i<j} I\{A_{ij}\}\operatorname{sgn}\{(X_i - X_j)(Y_i - Y_j)\}$$

$$= -K$$

Theoretical properties

- Asymptotic Normality: under quasi-independence

$$L_W \xrightarrow{d} N(0, AVar(L_W))$$

- How to choose a good weight function $W(x, y)$
  - Purpose: improve the power
  - Requirement: information about the alternative hypothesis
  - Proposal: apply the **conditional likelihood approach**

Alternative structure

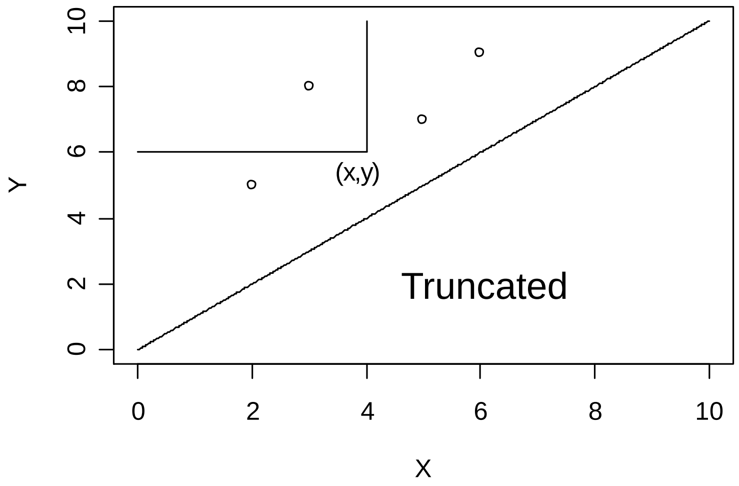
- Assumption: Semi-survival AC model (Chaieb et al. 2006)

$$\pi(x, y) = \Pr(X \le x, Y > y \mid X \le Y)$$
$$= \phi_\alpha^{-1}[\phi_\alpha\{F_X(x)\} + \phi_\alpha\{S_Y(y)\}] / c$$

$F_X(x)$:CDF

$S_Y(y)$:Survival function

Truncation Data



Example of models

- Clayton model: $\phi_\alpha(t) = t^{-(\alpha-1)} - 1$

$$\Pr(X \le x, Y > y \mid X \le Y)$$

$$= (1/c)[F_X(x)^{-(\alpha-1)} + S_Y(y)^{-(\alpha-1)} - 1]^{-\frac{1}{\alpha-1}}$$

- Large alpha ➔Large association
- (α, c, F_X, S_Y) are unknown
- Marginal distributions unspecified

Odds ratio

- All point (x,y) so that two by two tables are computed: $\varphi = \left\{ (x,y) \mid x \le y, N_{1\bullet}(dx,y) \ge 1, N_{\bullet 1}(x,dy) \ge 1 \right\}$

	$Y = y$	$Y > y$	
$X = x$	$\Delta(x,y)$		$N_{1\bullet}(dx,y)$
$X < x$			
	$N_{\bullet 1}(x,dy)$		$R(x,y)$

- Odds ratio= $\theta_\alpha \{c\pi(x,y)\}$

Here, $\theta_\alpha(v) = -v \cdot \phi''_\alpha(v) / \phi'_\alpha(v)$ (Chaieb, 2006)

Proposal: Conditional likelihood

- Bernoulli variable at a grid point: (x,y)

$$\Pr\{\Delta(x,y)=1 \mid R(x,y)=r,(x,y)\in\varphi\} = \frac{\theta_\alpha\{c\pi(x,y)\}}{r-1+\theta_\alpha\{c\pi(x,y)\}}$$

- Likelihood (under independence working assumption)

$$L(\alpha,\pi(x,y),c)$$
$$= \prod_{(x,y)\in\varphi}\left[\frac{\theta_\alpha\{c\pi(x,y)\}}{R(x,y)-1+\theta_\alpha\{c\pi(x,y)\}}\right]^{\Delta(x,y)}\left[\frac{r-1}{R(x,y)-1+\theta_\alpha\{c\pi(x,y)\}}\right]^{1-\Delta(x,y)}$$

- Score equation: $0=\partial \log L(\alpha,\hat{\pi}(x,y),c)/\partial\alpha$

Proposal: score test

- Under AC model, score equation from conditional likelihood is

$$U_L(\alpha,c) = \iint_{(x,y)\in\varphi} \frac{\dot{\theta}_\alpha\{c\hat{\pi}(x,y)\}}{\theta_\alpha\{c\hat{\pi}(x,y)\}}\left[\Delta(x,y) - \frac{N_{1\bullet}(dx,y)N_{\bullet 1}(x,dy)\theta_\alpha\{c\hat{\pi}(x,y)\}}{R(x,y)-1+\theta_\alpha\{c\hat{\pi}(x,y)\}}\right]$$

- Quasi-independence: $\alpha = 1$
- Take $\alpha \to 1$:score test

$$\lim_{\alpha\to 1} U_L(\alpha,c)$$
$$= \iint_{(x,y)\in\varphi} \lim_{\alpha\to 1} \dot{\theta}_\alpha\{c\hat{\pi}(x,y)\}\left[\Delta(x,y) - \frac{N_{1\bullet}(dx,y)N_{\bullet 1}(x,dy)}{R(x,y)}\right]$$

Proposed weight function

- Under semi-survival AC model

$$W(x,y) = \lim_{\alpha \to 1} \dot{\theta}_{\alpha}\{\hat{c}\hat{\pi}(x, y-)\}$$

- Example: Clayton alternative $w(x,y) = 1$
- Example: Frank alternative $w(x,y) = \hat{\pi}(x, y-)$

A general class of test statistics

- G^{ρ} class statistics

$$L_{\rho} = \iint_{x \le y} \hat{\pi}(x, y-)^{\rho} \left\{ N_{11}(dx, dy) - \frac{N_{1\bullet}(dx, y) N_{\bullet 1}(x, dy)}{R(x, y)} \right\}$$

- Efficiency

  - Clayton alternative $\rho = 0$

  - Frank alternative $\rho = 1$

Variance estimation

- Class G^{ρ}

$$L_{\rho} = \iint_{x \le y} \hat{\pi}(x, y-)^{\rho} \left\{ N_{11}(dx, dy) - \frac{N_{1\bullet}(dx, y) N_{\bullet 1}(x, dy)}{R(x, y)} \right\}$$

- Empirical variance estimator

  - apply the functional delta method

$$AVar(L_{\rho})$$

$$\approx \sum_{j} \left[\frac{1}{n} \sum_{k} I\{A_{jk}\} \hat{\pi}(\tilde{X}_{jk}, \tilde{Y}_{jk}-)^{\rho-1} \operatorname{sgn}\{(X_j - X_k)(Y_j - Y_k)\} + \frac{(\rho+1)L_{\rho}}{n} \right.$$

$$\left. + \frac{\rho-1}{n^2} \sum_{k<l} I\{A_{kl}\} \hat{\pi}(\tilde{X}_{kl}, \tilde{Y}_{kl}-)^{\rho-2} \operatorname{sgn}\{(X_k - X_l)(Y_k - Y_l)\} I(X_j \le \tilde{X}_{kl}, Y_j \ge \tilde{Y}_{kl}) \right]^2.$$

Sketch of asymptotic analysis

- Statistical functionals

$$L_w = \iint_{x \le y} W(x,y)\left[\Delta(x,y) - \frac{N_{1\bullet}(dx,y)N_{\bullet 1}(x,dy)}{R(x,y)}\right]$$

$$= \Phi(\hat{\pi})$$

$$\hat{\pi}(x,y) \equiv \frac{1}{n}\sum_j I(X_j \le x, Y_j > y)$$

$\Phi(\cdot)$: Hadamard differentiable

➔Asymptotic normal

$\Phi(\cdot)$: Continuously Gateaux differentiable

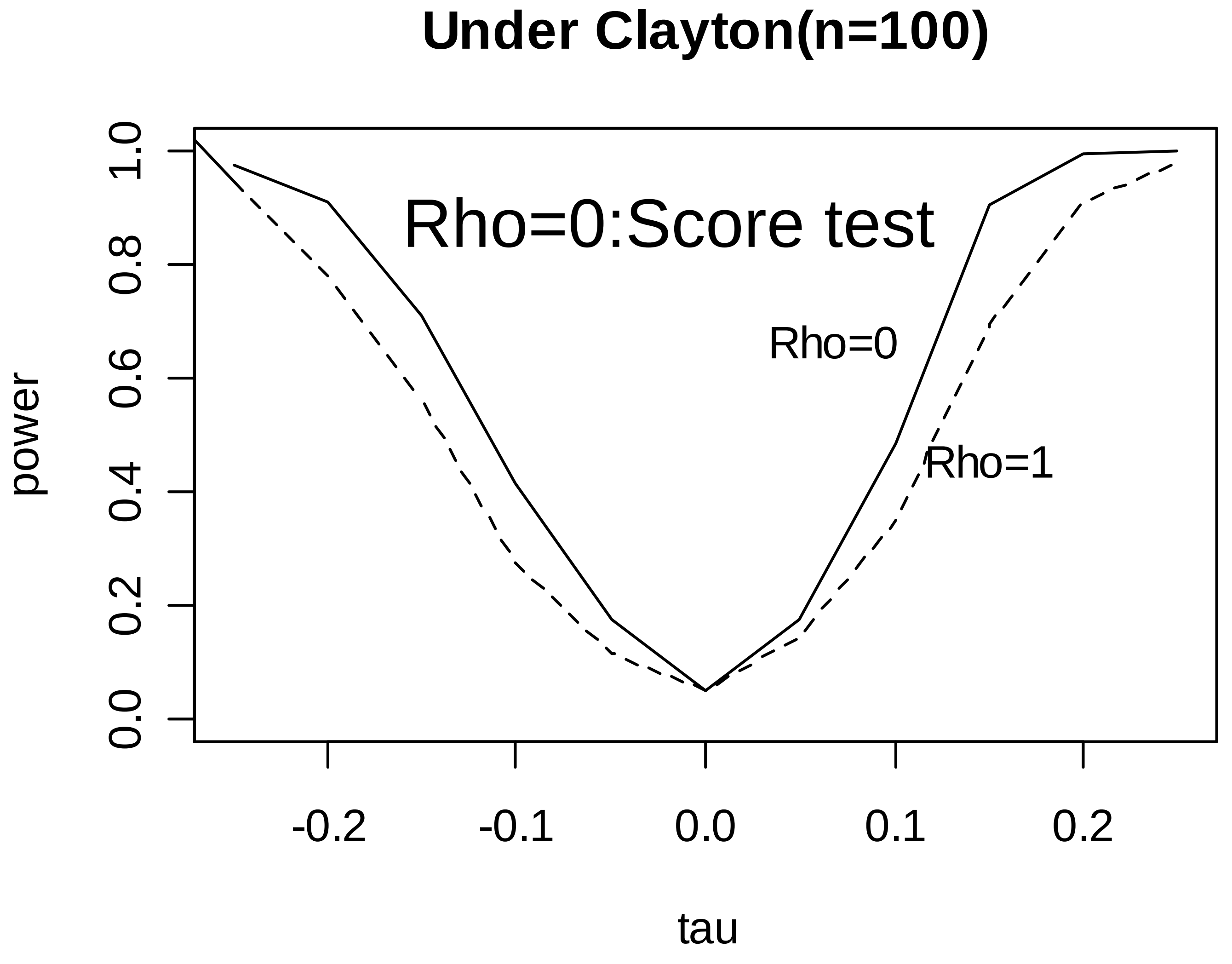
➔Consistency of Jackknife

Extension to right-censored data

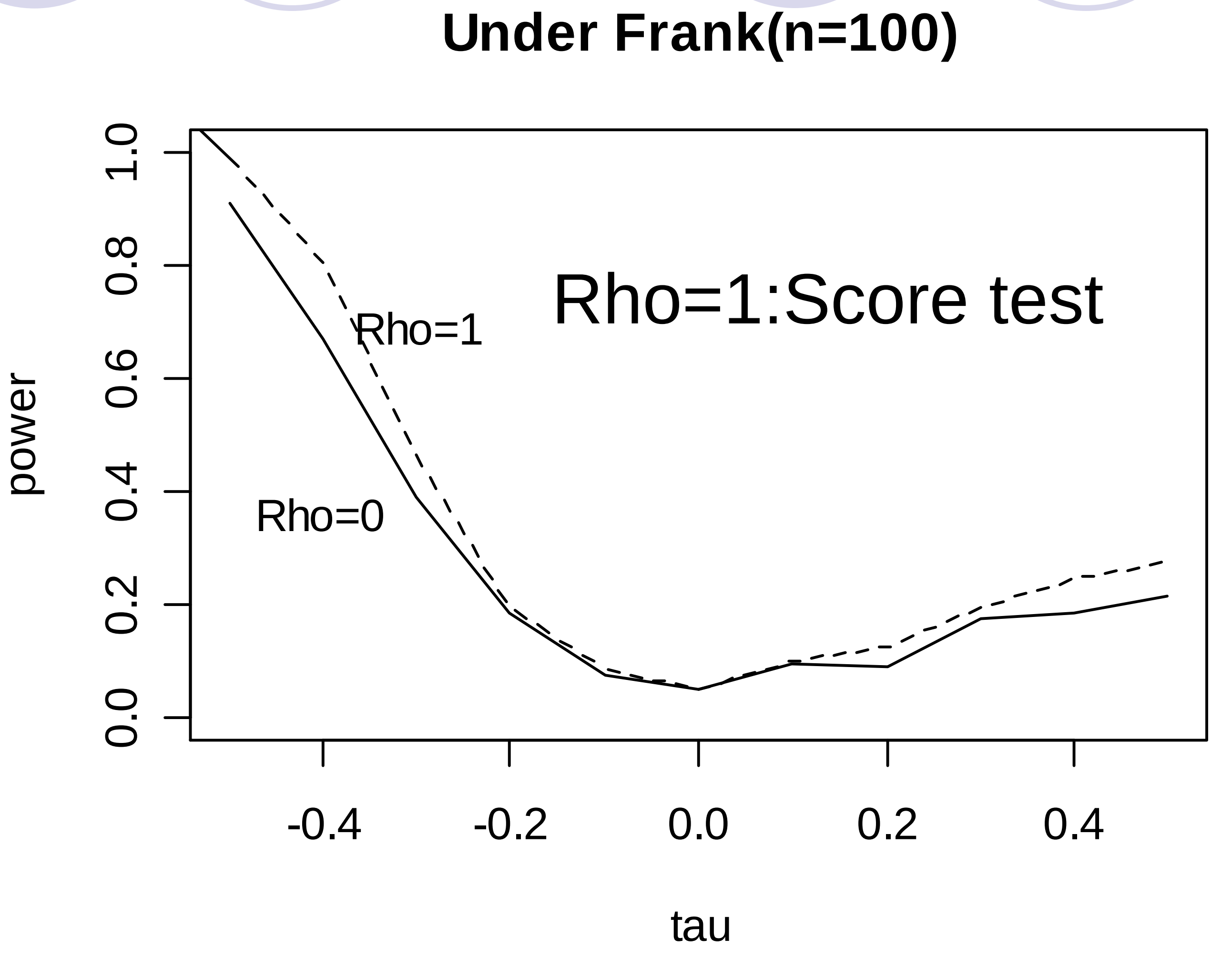
- Extension to left-truncation with right-censored data is easily obtained
- Details are available in pre-print

  by W. Wang & T. Emura

Power comparison



Power comparison



Future Works
– Testing independence

- Justification of Jackknife method under right-censoring
- Theoretical account for efficiency gain
- Supreme log-rank test, combination of several weighted log-rank test (versatile test)