Tung Hai University,
Department of Statistics
4/24, 2012

Copula-based inference for truncation models based on copulas

Takeshi Emura
Graduate Institute of Statistics, National Central University

Joint work with Weijing Wang
Institute of Statistics, National Chiao Tung Uiversity

Outlines

Review & Motivation

- Truncation data:
- Traditional approaches
- Copula-based approaches

Proposed method

- Proposed method
- Simulation, Data analysis
- Conclusion

Part I: Review & Motivation

Truncation:

- A pair of (X, Y) is observed only when X ≤ Y holds.
- If X > Y, nothing is observed ! (truncated)

Today's focus

A special case of doubly truncated data

- X is observed only when Z ≤ X ≤ Y, where (Z, Y) is random
- If Z=0, this reduces to truncation

See

Efron & Petrosian (1992 *JASA*), Wang & Stovring (2007, *BMC Medical Res.)*
Shen (2010 *AISM*), Moreira & Una-Alvares (2010 *J. of Nonpar*.)
Shen (2012 *J. of App. Stat*.), but name a few.

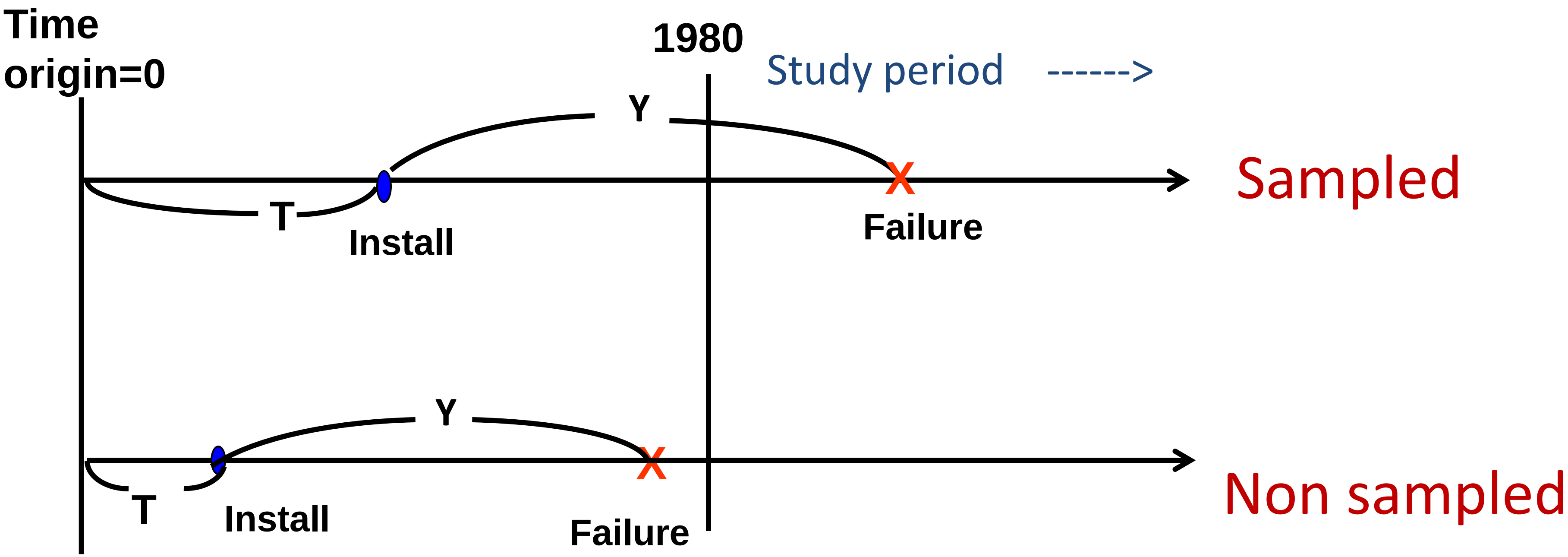
Truncation occurs in Survival analysis of Power Transformer (電力變壓器)

Fig. from Toshiba Corporation

What is Truncation?

- Power Transformer (電力變壓器) data

(Hong et al., 2009 *Ann. Applied Statistics*)

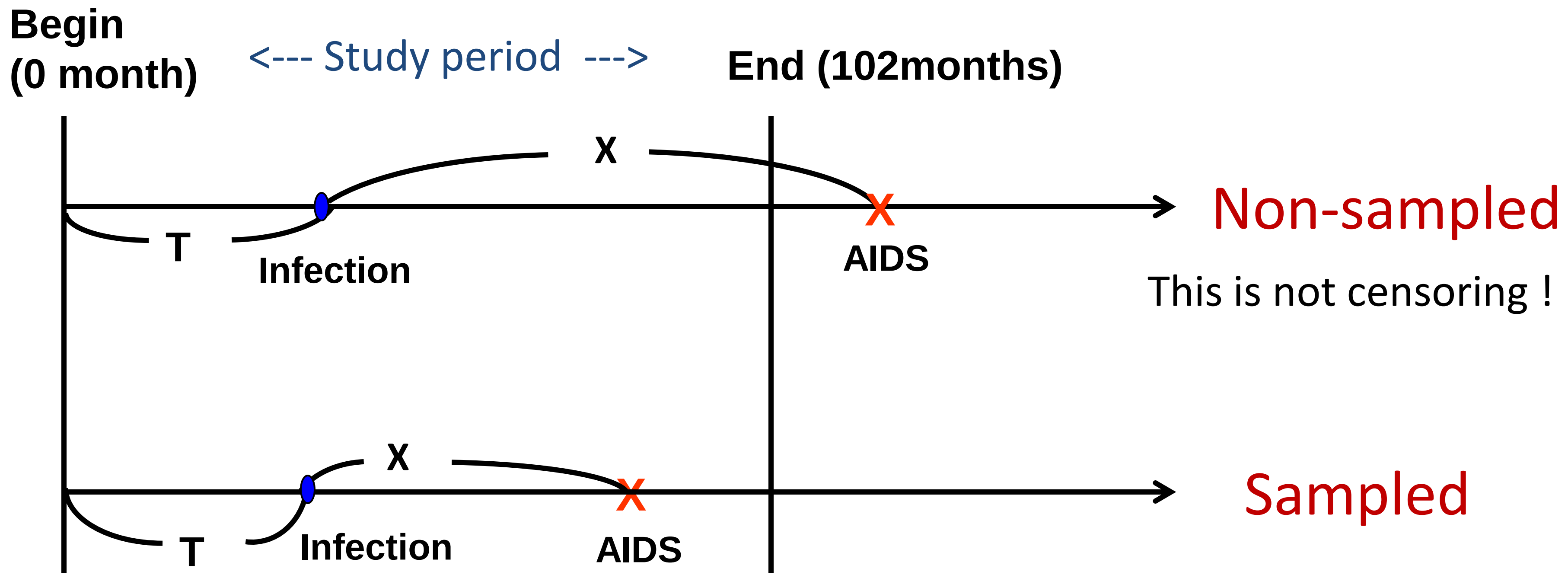


Inclusion criterion: $1980 \le T+Y$, i.e., $1980-T = X \le Y$

What is Truncation?

- Transfusion-related AIDS

(Lagakos et al., 1988 *Biometrika*)



Inclusion criterion: $T+X \leq 102$ i.e., $X \leq Y \equiv 102-T$

Truncation data

- Truncation data :

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

subject to $X_j \leq Y_j$

$$\Downarrow$$

i.i.d. from

$\Pr(X \leq x, Y \leq y \mid X \leq Y),$

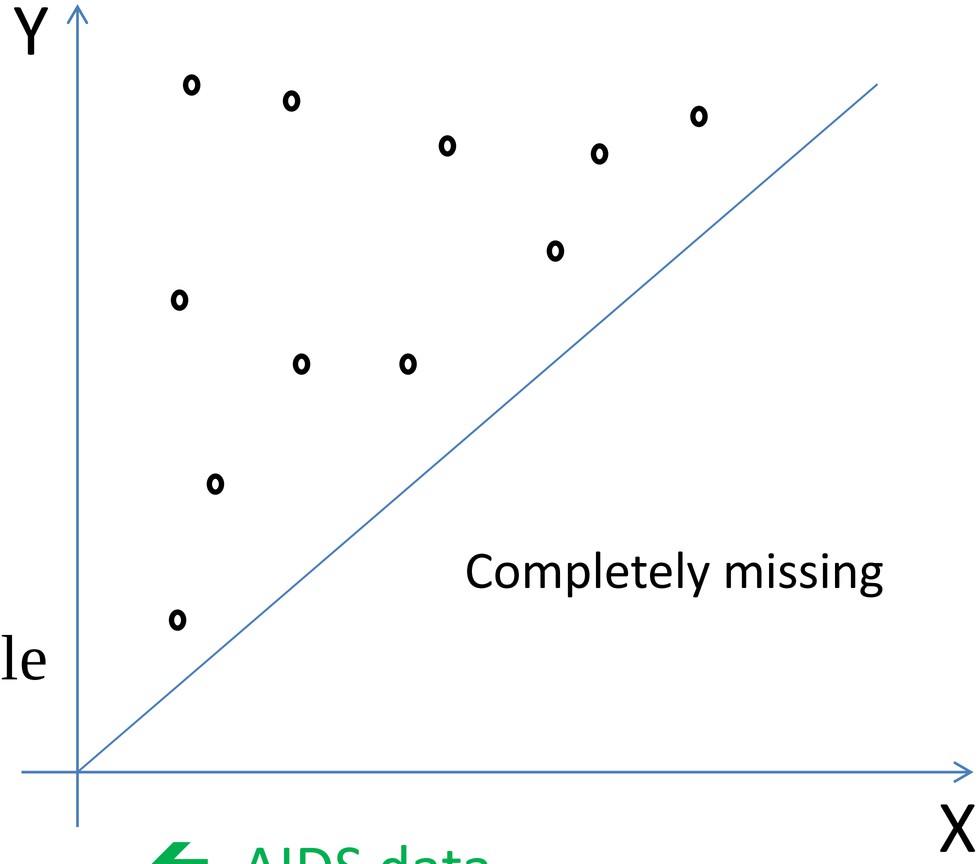
where (X, Y) is

the "*population*" random variable

Target : Estimation of

$F_X(x) = \Pr(X \leq x)$ ← AIDS data

$F_Y(y) = \Pr(Y \leq y)$ ← Power transformer data



Traditional analysis

- Nonparametric estimator

$$\hat{F}_X(x) = \prod_{u>x}\left\{1 - \frac{\sum_{j=1}^{n} I(X_j = u)}{\sum_{j=1}^{n} I(X_j \le u, Y_j \ge u)}\right\}, \quad \hat{F}_Y(y) = 1 - \prod_{u \le y}\left\{1 - \frac{\sum_{j=1}^{n} I(Y_j = u)}{\sum_{j=1}^{n} I(X_j \le u, Y_j \ge u)}\right\}$$

(Lynden-Bell, 1971; Lagakos et al., 1988)

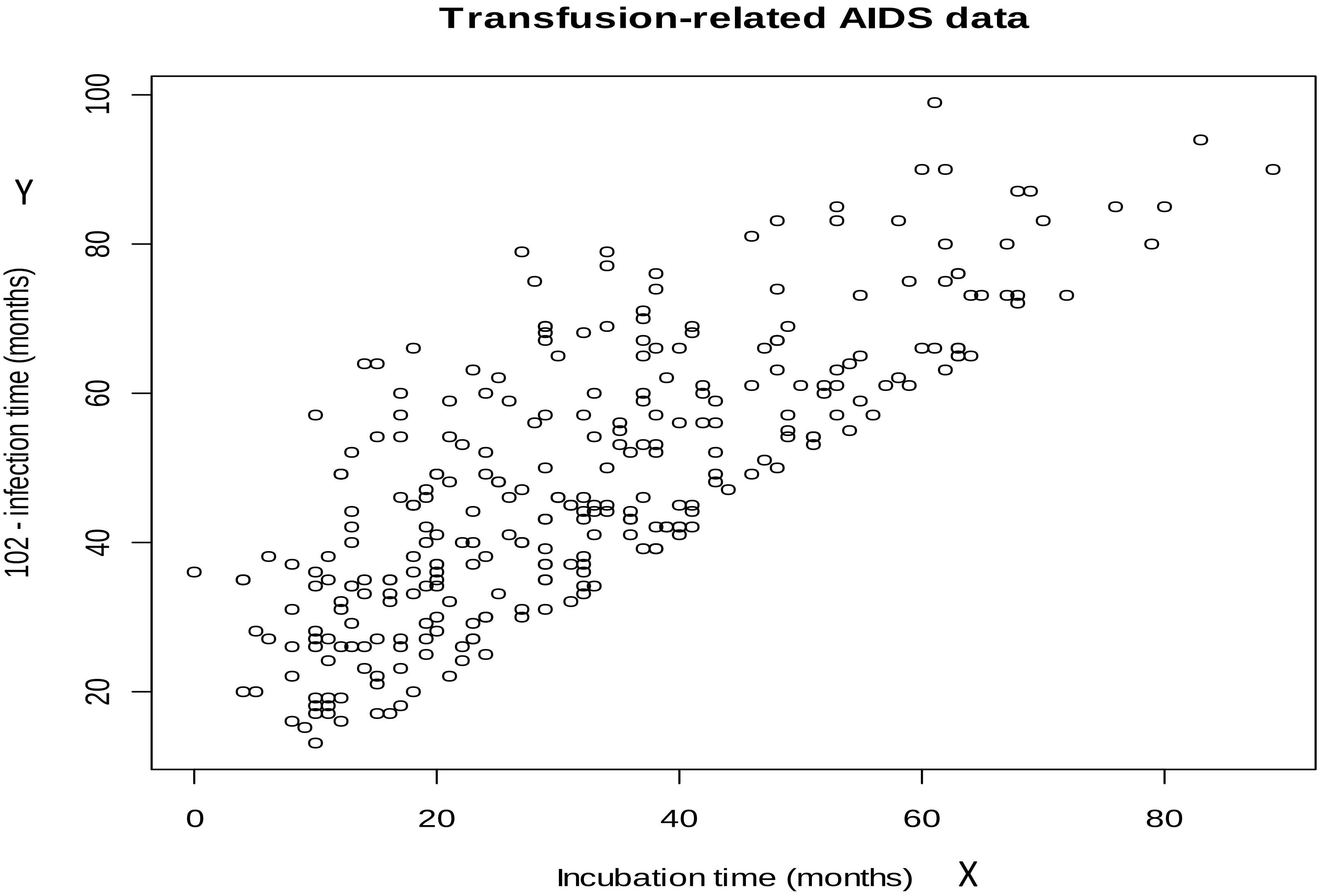
- Key assumption: Quasi-independence (Tsai, 1990):

$$X \perp_Q Y: \ \ \Pr(X \le x, Y \le y \mid X \le Y) = \frac{\iint_{\substack{\mathrm{u \le x, v \le y} \\ \mathrm{u \le v}}} dF_X(u)\,dF_Y(v)}{\iint_{\mathrm{u \le v}} dF_X(u)\,dF_Y(v)}$$

*Quasi-independence assumption is testable

(Tsai, 1990; Chen et al., 1996; Martin & Betensky, 2005; Emura & Wang; 2010)

Quasi-independence is rejected at
P-value = 0.040 (Marting & Betensky 2005 *JASA*)
= 0.048 (Emura & Wang 2010 *JMVA*)



- Chaieb et al. (2006 *Biometrika*) relax the quasi-independnece by using “Copulas”

Copula

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

- Example 1: Independence copula

$$C[u, v] = uv$$

- Example 2: Frank copula (Genest, 1986; Frank, 1979)

$$C_\alpha[u, v] = \log_{\alpha^{-1}}\left\{1 + \frac{(\alpha^{-u} - 1)(\alpha^{-v} - 1)}{(\alpha^{-1} - 1)}\right\}, \quad \alpha > 0$$

- Example 3: Normal copula

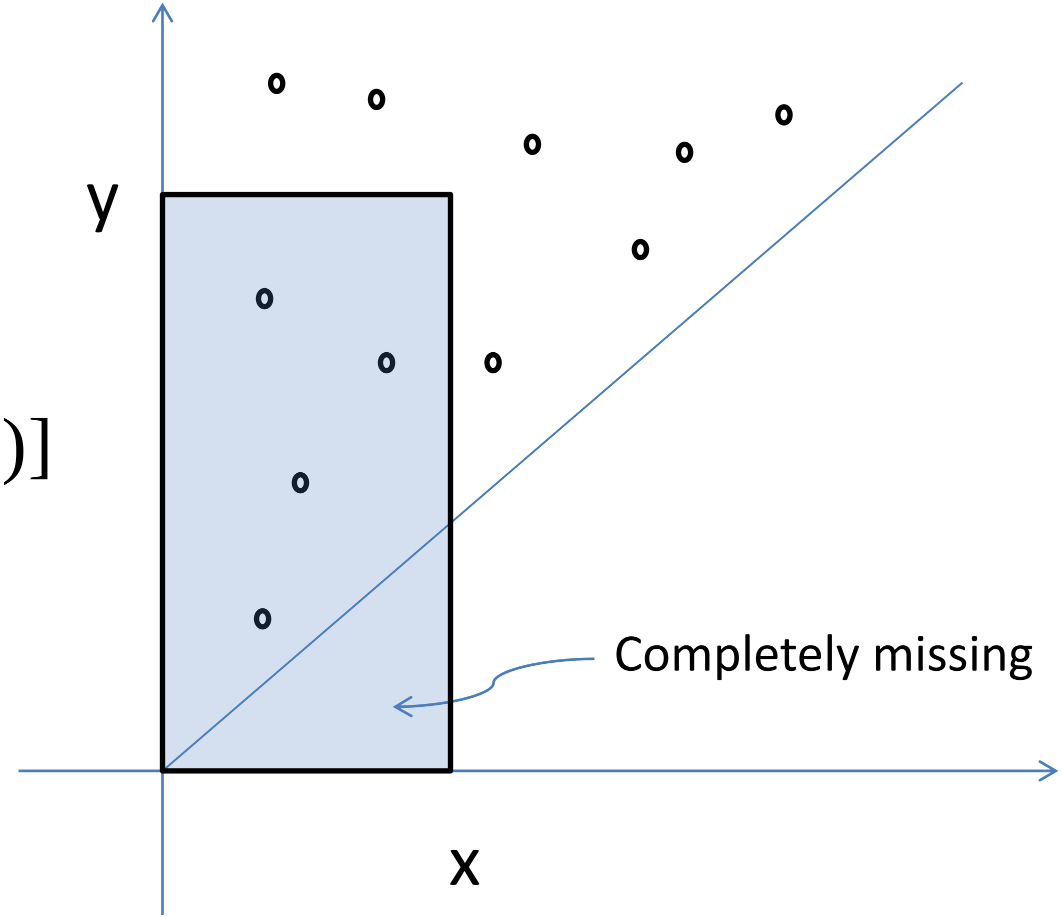
$$C_\rho[u, v] = \Phi_\rho[\Phi^{-1}(u), \Phi^{-1}(v)], \quad -1 < \rho < 1$$

Φ_ρ : Joint CDF of standard bivariate normal

Copula model

$$\Pr(X \leq x, Y \leq y)$$
$$= C[\Pr(X \leq x), \Pr(Y \leq y)]$$

The model is unidentifiable



Copula model

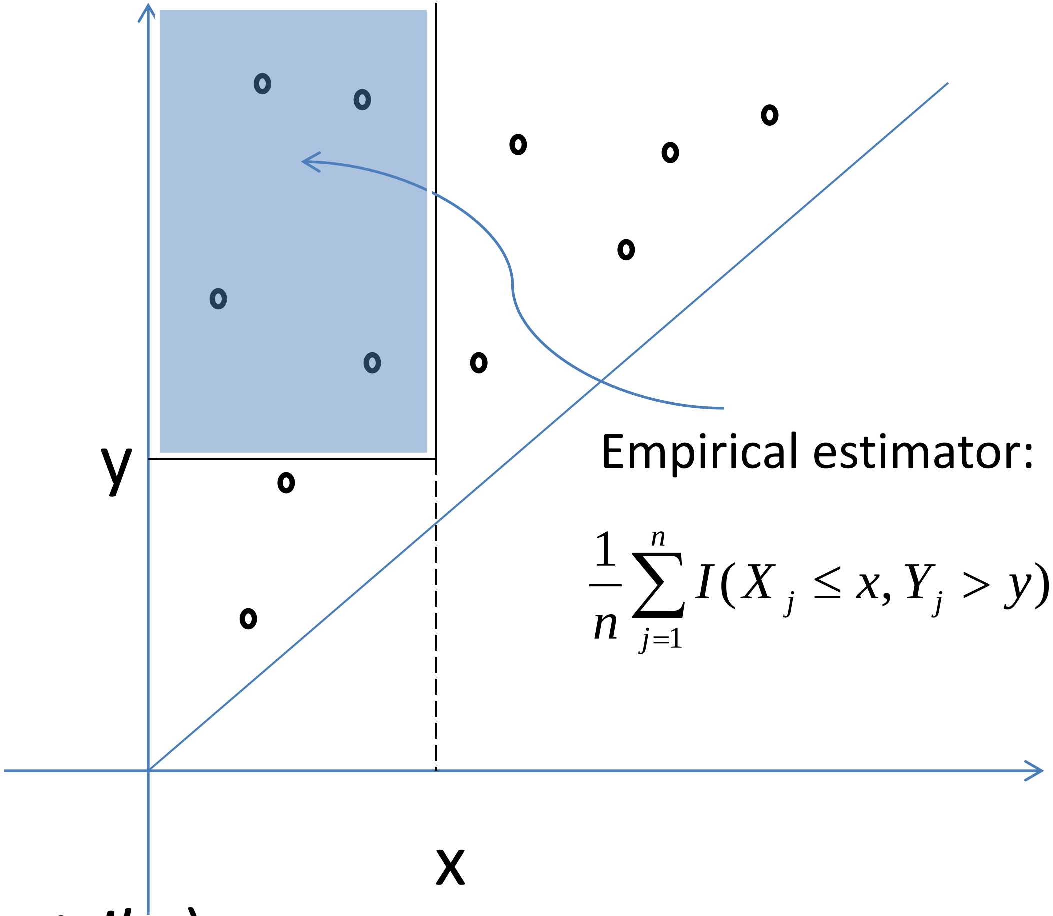
$$\Pr(X \le x, Y > y \mid X \le Y)$$

$$= \frac{C_\alpha[F_X(x), S_Y(y)]}{c(\alpha, F_X, S_Y)}$$

where

$$c(\alpha, F_X, S_Y) =$$

$$\iint_{x \le y} \frac{\partial^2}{\partial x \partial y} C_\alpha[F_X(x), S_Y(y)]dxdy$$



Empirical estimator:

$$\frac{1}{n}\sum_{j=1}^{n} I(X_j \le x, Y_j > y)$$

- Semi-survival copula

(Chaieb et al., 2006, *Biometrika*)

- Quasi-independence: $C_\alpha[u, v] = uv$

Existing procedures

- Archimedean family: $C_\alpha[u,v]=\phi_\alpha^{-1}\{\phi_\alpha(u)+\phi_\alpha(v)\}$

  Useful for solving moment equations:

  [Chaieb et al. (2006) and Emura et al. (2011 *Stat. Sinica*)]

$$\because)\frac{1}{n}\sum_{j=1}^{n}I(X_j\le t,Y_j>t)=\frac{C_\alpha[F_X(t),S_Y(t)]}{c(\alpha,F_X,S_Y)},$$

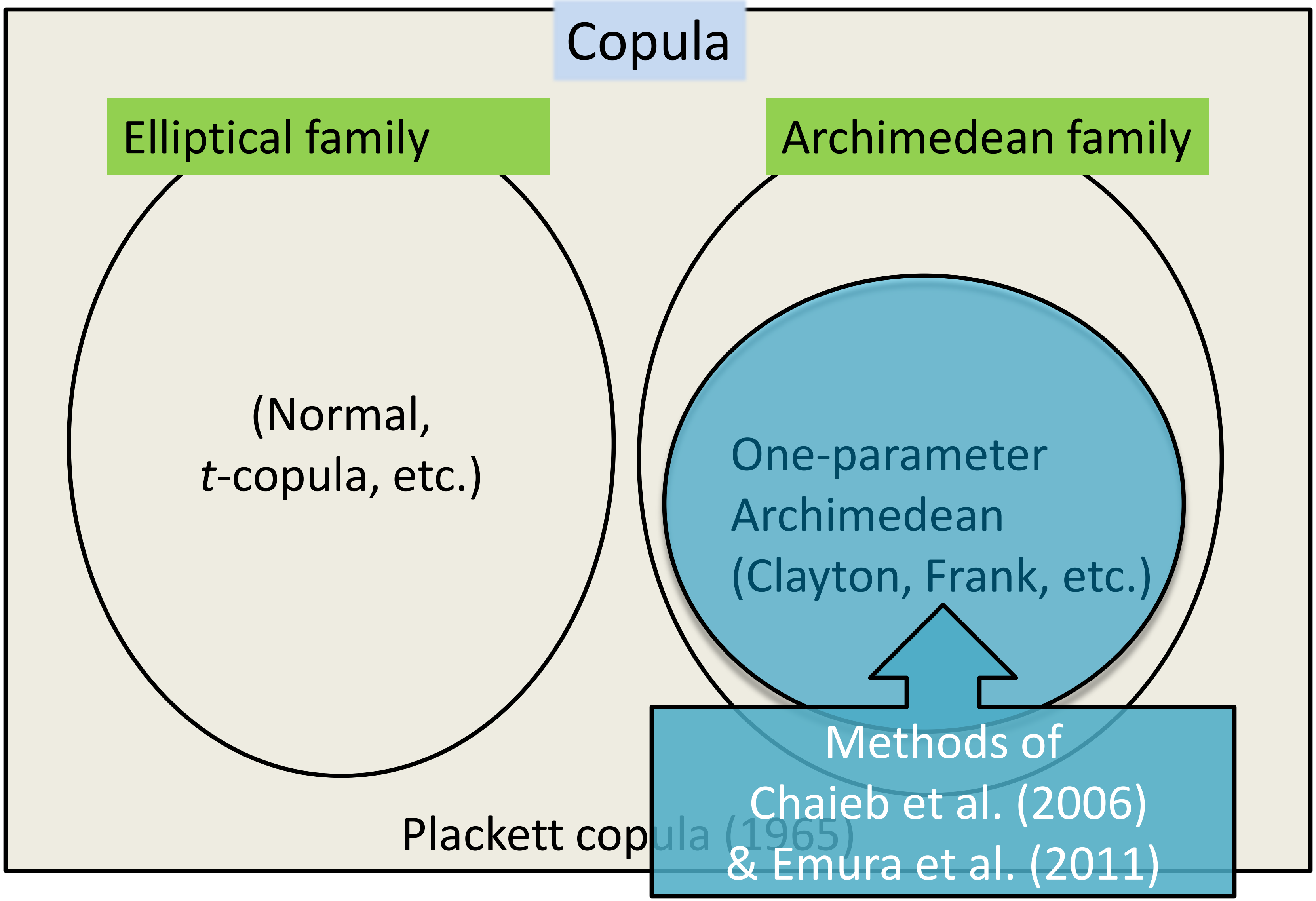
$$\Leftrightarrow \phi_\alpha\left(\frac{c(\alpha,F_X,S_Y)}{n}\sum_{j=1}^{n}I(X_j\le t,Y_j>t)\right)=\phi_\alpha(F_X(t))+\phi_\alpha(S_Y(t))$$

$$\Rightarrow F_X(t)=\phi_\alpha^{-1}\left\{\phi_\alpha\left(\frac{c(\alpha,F_X,S_Y)}{n}\sum_{j=1}^{n}I(X_j\le t,Y_j>t)\right)-\phi_\alpha(S_Y(t))\right\}$$

$$S_Y(t)=\phi_\alpha^{-1}\left\{\phi_\alpha\left(\frac{c(\alpha,F_X,S_Y)}{n}\sum_{j=1}^{n}I(X_j\le t,Y_j>t)\right)-\phi_\alpha(F_X(t))\right\}$$

Sequentially solve: $S_X(X_1)\equiv 1\rightarrow F_X(X_1)\rightarrow S_X(X_2)\rightarrow\cdots$

Existing procedures



Copula

Elliptical family

(Normal, *t*-copula, etc.)

Archimedean family

One-parameter Archimedean (Clayton, Frank, etc.)

Methods of Chaieb et al. (2006) & Emura et al. (2011)

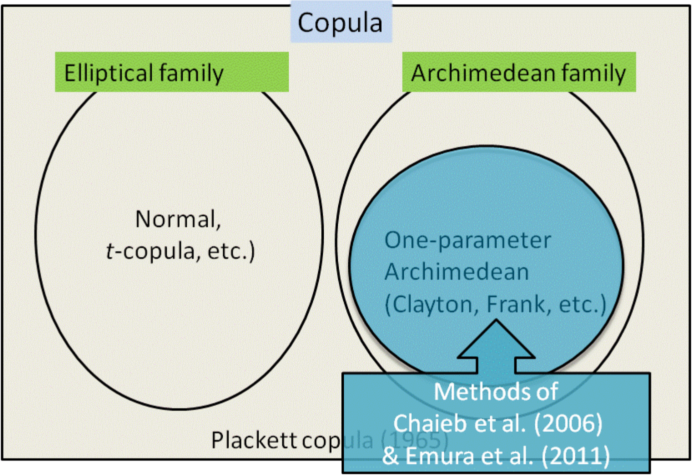
Plackett copula (1965)

Part II: Proposed method

Proposed method

- The preceding two methods use moment-based estimating equations for (F_X, S_Y) under **Archimedean copula family**

- In this talk, we propose to get $(\hat{F}_X, \hat{S}_Y)$ by the nonparametric maximum likelihood estimator (NPMLE; due to T



Copula

Elliptical family

Archimedean family

Normal,
t-copula, etc.)

One-parameter
Archimedean
(Clayton, Frank, etc.)

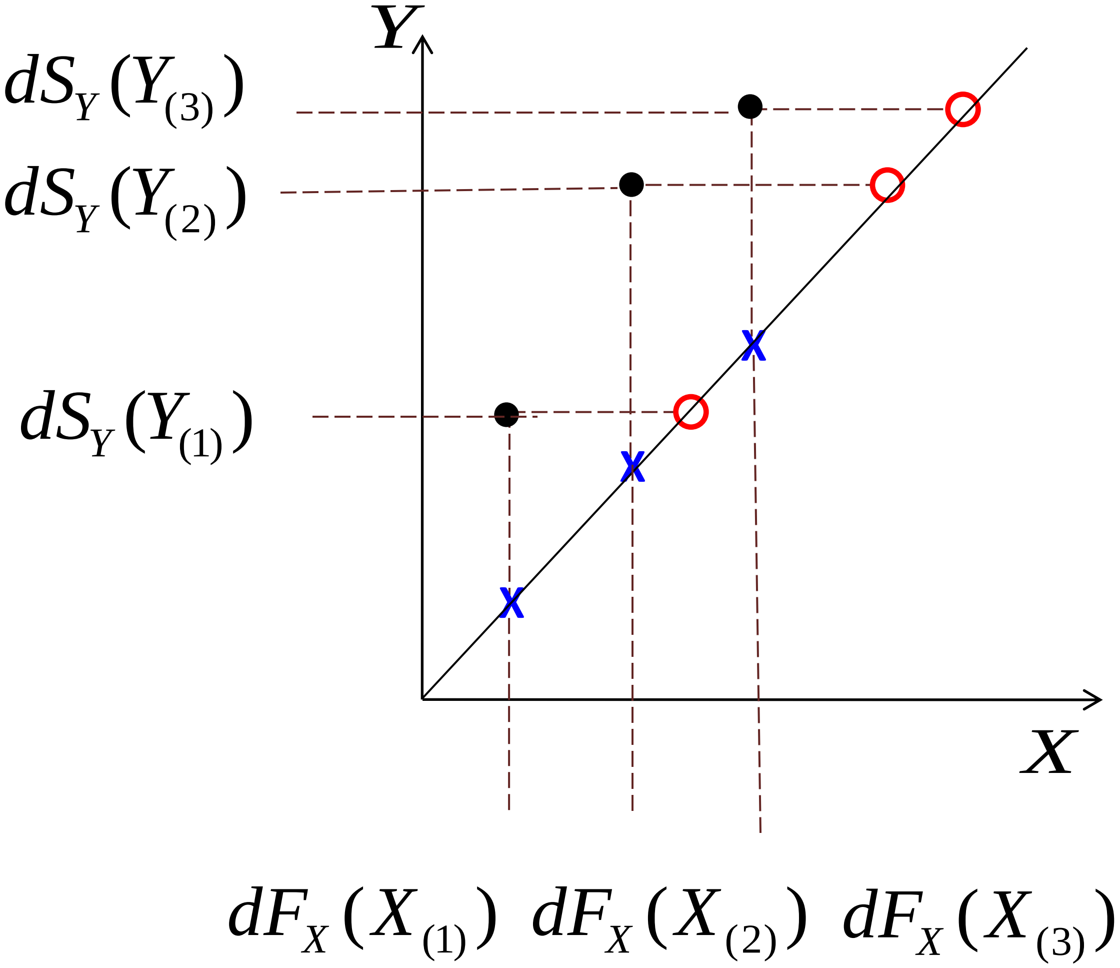
Methods of
Chaieb et al. (2006)
& Emura et al. (2011)

Plackett copula (1965)

Advantage:

Potentially applicable for
a broader class of copulas

NPMLE: due to Turnbull (1976)



Maximize the NPMLE

with constraints :

$0 \le dF_X(X_{(1)}) \le 1$

$0 \le dF_X(X_{(2)}) \le 1$

$0 \le dF_X(X_{(3)}) \le 1$

$$\sum_{i=1}^{3} dF_X(X_{(i)}) = 1$$

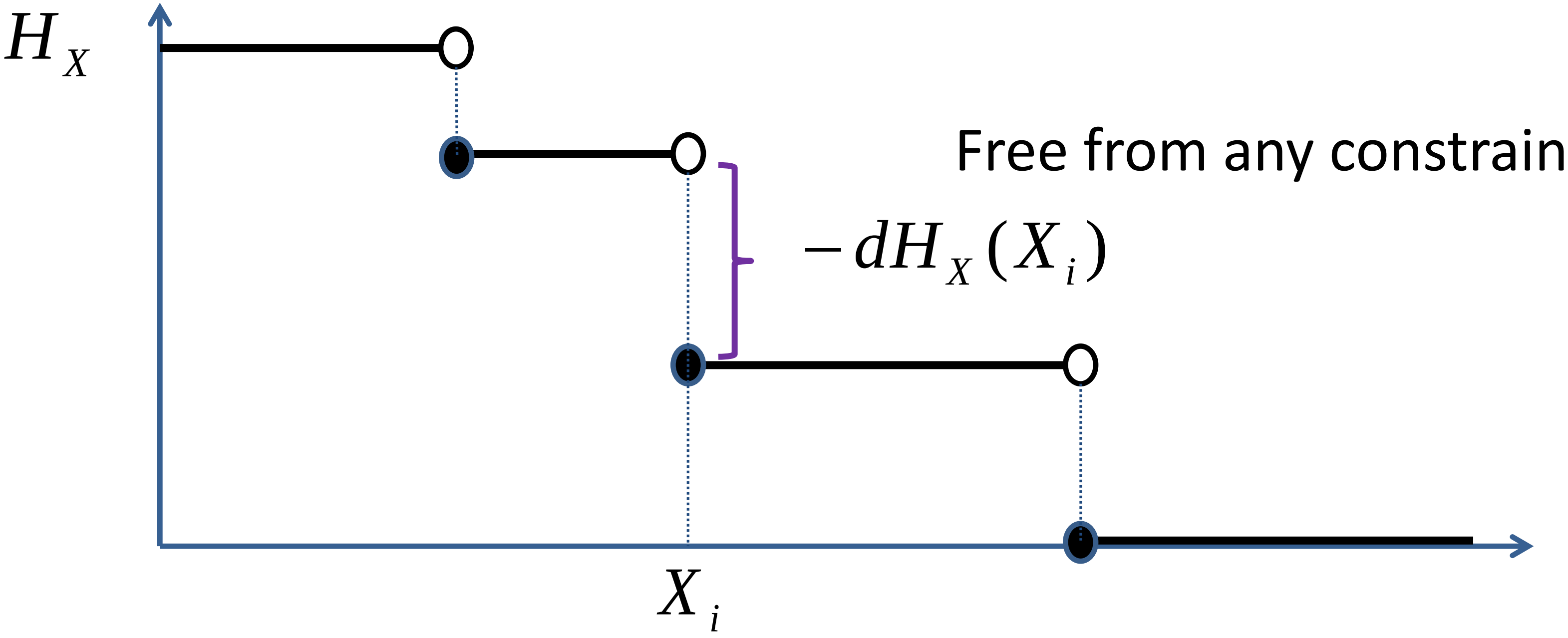
- Parameterize (F_X, S_Y) as follows:

$$F_X(x) = e^{-H_X(x)}, \quad S_Y(y) = e^{-\Lambda_Y(y)}$$

$* H_X(x)$: Reverse-time cumulative hazard

(Lagakos et al., 1988; Navaro & Ruiz, 1996)

$* \Lambda_Y(y)$: Cumulative hazard



Free from any constraint

Proposed method

- Semi-survival Copula (Chaieb et al., 2006)

$$\Pr(X \le x, Y > y \mid X \le Y) = \frac{C_\alpha[e^{-H_X(x)}, e^{-\Lambda_Y(y-)}]}{c(\alpha, H_X, \Lambda_Y)},$$

where $c(\alpha, H_X, \Lambda_Y) = \iint_{x \le y} -\frac{\partial^2}{\partial x \partial y} C_\alpha[e^{-H_X(x)}, e^{-\Lambda_Y(y-)}] dxdy$

leading to the density

$$\Pr(X = x, Y = y \mid X \le Y) = \frac{\eta_\alpha[H_X(x), \Lambda_Y(y-)]}{c(\alpha, H_X, \Lambda_Y)} \{-dH_X(x)\} d\Lambda_Y(y),$$

where $\eta_\alpha[x, y] = e^{-x} e^{-y} \frac{\partial^2}{\partial u \partial u} C_\alpha[u, u] \Big|_{u=e^{-x}, v=e^{-y}}$

Proposed method

- Log-likelihood

$$l_n(\alpha, H_X, \Lambda_Y) =$$

$$\sum_{j=1}^{n} \log \eta_\alpha[H_X(X_j), \Lambda_Y(Y_j-)] + \log\{-dH_X(X_j)\} + \log d\Lambda_Y(Y_j) - \log c(\alpha, H_X, \Lambda_Y)$$

➔ Maximize for (*2n+1*) parameters

$$(\alpha, -dH_X(X_1), \ldots, -dH_X(X_n), d\Lambda_Y(Y_1), \ldots, d\Lambda_Y(Y_n))$$

Proposed method

- Identifiability problem

  The maximum of $l_n(\alpha, H_X, \Lambda_Y)$ is not unique

  (# of parameters =2n+1 > # of observed points = 2n)

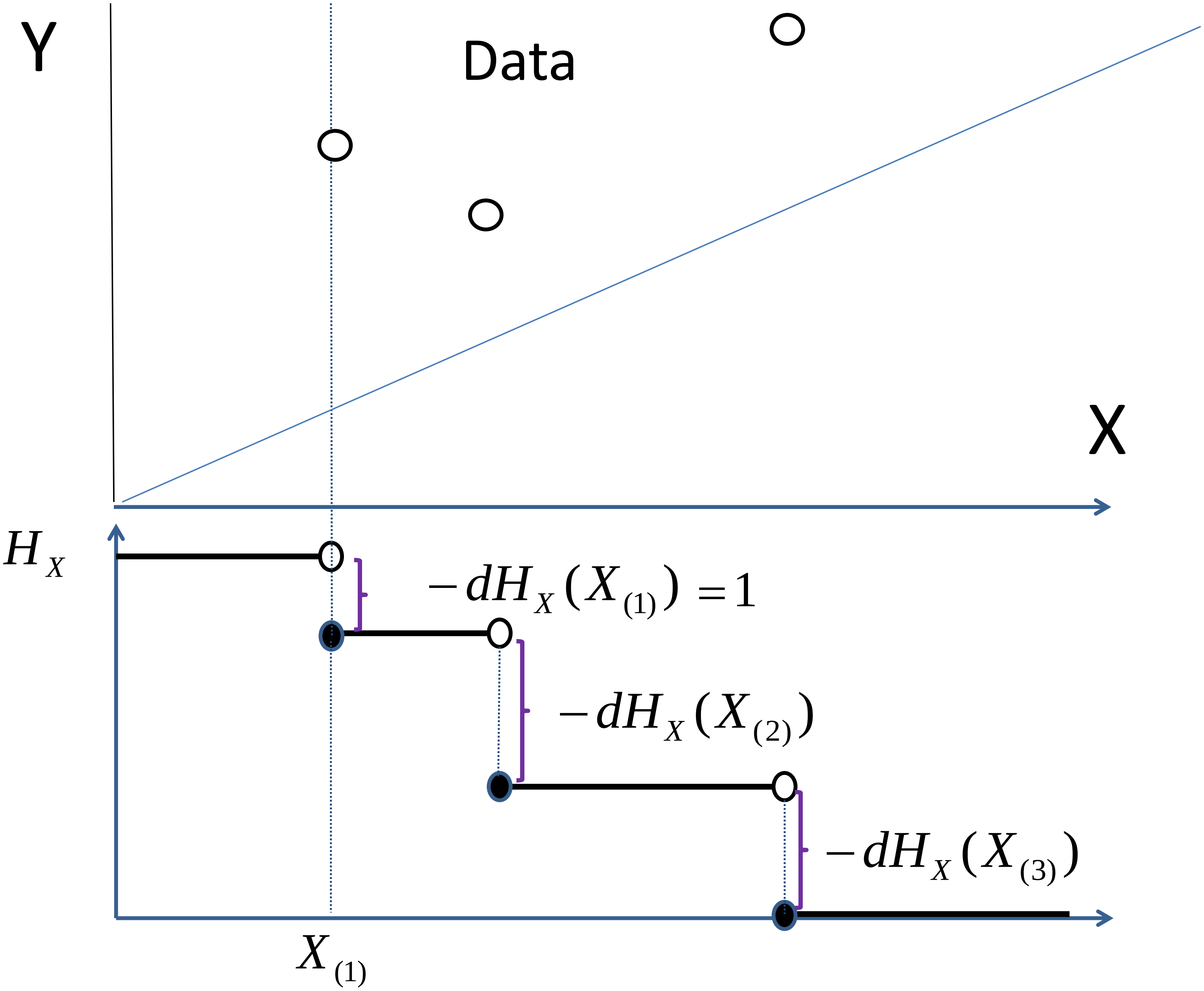
  *analogy with the linear regression with p > n

- Idea: reduces to 2n-1 parameters

$$(\alpha, -dH_X(X_1), \ldots, -dH_X(X_n), d\Lambda_Y(Y_1), \ldots, d\Lambda_Y(Y_n))$$

$$\Downarrow$$

$$(\alpha, \underbrace{-dH_X(X_{(1)})}_{\equiv 1}, \ldots, -dH_X(X_{(n)}), d\Lambda_Y(Y_{(1)}), \ldots, \underbrace{d\Lambda_Y(Y_{(n)})}_{\equiv 1})$$


Y
Data
X
H_X
$-dH_X(X_{(1)}) = 1$
$-dH_X(X_{(2)})$
$-dH_X(X_{(3)})$
$X_{(1)}$

Proposed method

- *2n-1* score equations

$$0 = \partial l_n(\alpha, H_X, \Lambda_Y) / \partial \alpha$$

$$0 = \partial l_n(\alpha, H_X, \Lambda_Y) / \partial \{-dH_X(X_{(j)})\}, \quad j = 2,.., n$$

$$0 = \partial l_n(\alpha, H_X, \Lambda_Y) / \partial d\Lambda_Y(Y_{(j)}), \quad j = 1,.., n-1$$

leading to a self-consistency (Turnbull, 1976) type equations

$$H_X(x) = \int_x^\infty \frac{\sum_{j=1}^n I(X_j = u)}{\sum_{j=1}^n \Psi_j^{(1,0)}(u; \alpha, H_X, \Lambda_Y)}$$

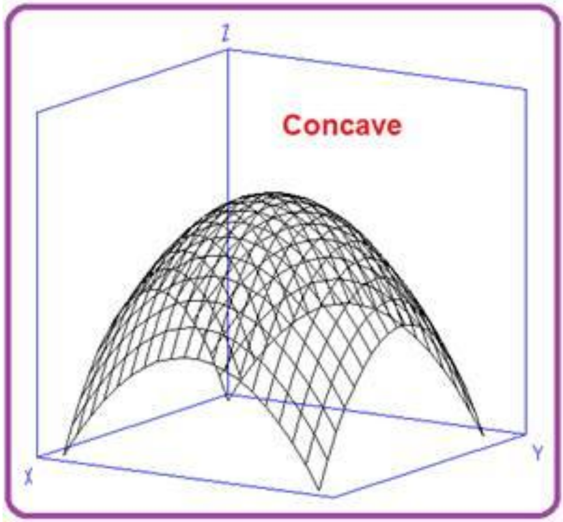
$$\Lambda_Y(x) = \int_0^y \frac{\sum_{j=1}^n I(Y_j = u)}{\sum_{j=1}^n \Psi_j^{(0,1)}(u; \alpha, H_X, \Lambda_Y)}$$

Proposed method

- Quasi-Newton algorithm to maximize

$$l_n(\alpha, H_X, \Lambda_Y) =$$

$$\sum_{j=1}^{n} \log \eta_\alpha [H_X(X_j), \Lambda_Y(Y_j -)] + \log\{-dH_X(X_j)\} + \log d\Lambda_Y(Y_j) - \log c(\alpha, H_X, \Lambda_Y)$$



<u>Requirements:</u>

•twice differentiable

•Unique maximum

•No constraint in parameter space

Especially, one can apply “nlm” routine in R.

Convergence criteriat:

1. *2n-1* scores are zero at $(\hat{\alpha}, \hat{H}_X, \hat{\Lambda}_Y)$

2. Hessian matrix at $(\hat{\alpha}, \hat{H}_X, \hat{\Lambda}_Y)$ is positive definite

(Hessian matrix are concave, it should be a global maxima)

Proposed method

- The NPMLE $(\hat{\alpha}, \hat{H}_X, \hat{\Lambda}_Y)$ is consistent & asymptotic normal (Emura & Wang 2012 *JMVA*)
- Observed Fisher information

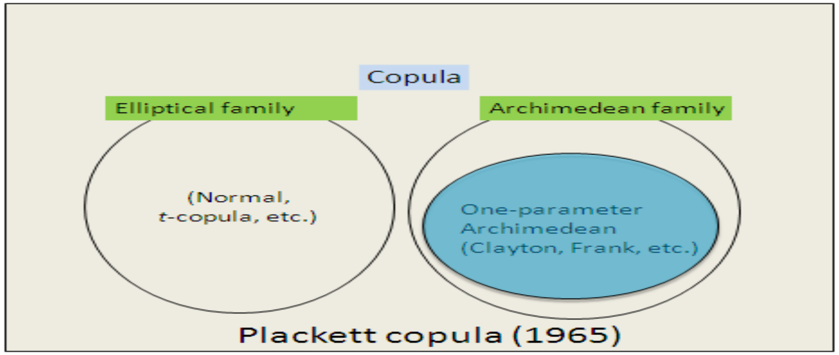
  = minus of the Hessian of $l_n(\alpha, H_X, \Lambda_Y)$

$$\hat{i}_n(\hat{\alpha}, \hat{H}_X, \hat{\Lambda}_Y) = \begin{bmatrix} \hat{i}_{n,11} & \hat{i}'_{n,12} \\ \hat{i}_{n,12} & \hat{i}_{n,22} \end{bmatrix}$$

- Consistent variance estimator, e.g.,

$$\hat{V}_n(\hat{\alpha}) \approx (\hat{i}_{n,11} - \hat{i}'_{n,12}\hat{i}_{n,22}^{-1}\hat{i}_{n,12})^{-1}$$

Simulation setting (I):



- Plackett copula (not Archimedean family)

$$C_\alpha[u,v] = \frac{1}{2(\alpha-1)} + \frac{u+v}{2} - \frac{[\{1+(\alpha-1)(u+v)\}^2 - 4uv\alpha(\alpha-1)]^{1/2}}{2(\alpha-1)}$$

$$\alpha = 1/2.51,\ 1/5.11,\ 2.51,\ 5.11$$

$$(\text{s.t. Spearmen's rho} = 0.25,\ 0.5, -0.25, -0.5)$$

- Exponential margins

$$H_X(x) = -\log(1-e^{-1.5x})$$

$$\Lambda_Y(y) = 0.5y$$

- Data generation: $\Pr(X \le x, Y > y) = C_\alpha[e^{-H_X(x)}, e^{-\Lambda_Y(y-)}]$

If $X_j \le Y_j$ then included in the sample. Otherwise truncated.

Repeat until we get *n* (=125 or 250) pair of (X_j, Y_j)

Positive dependence, 200 repetitions

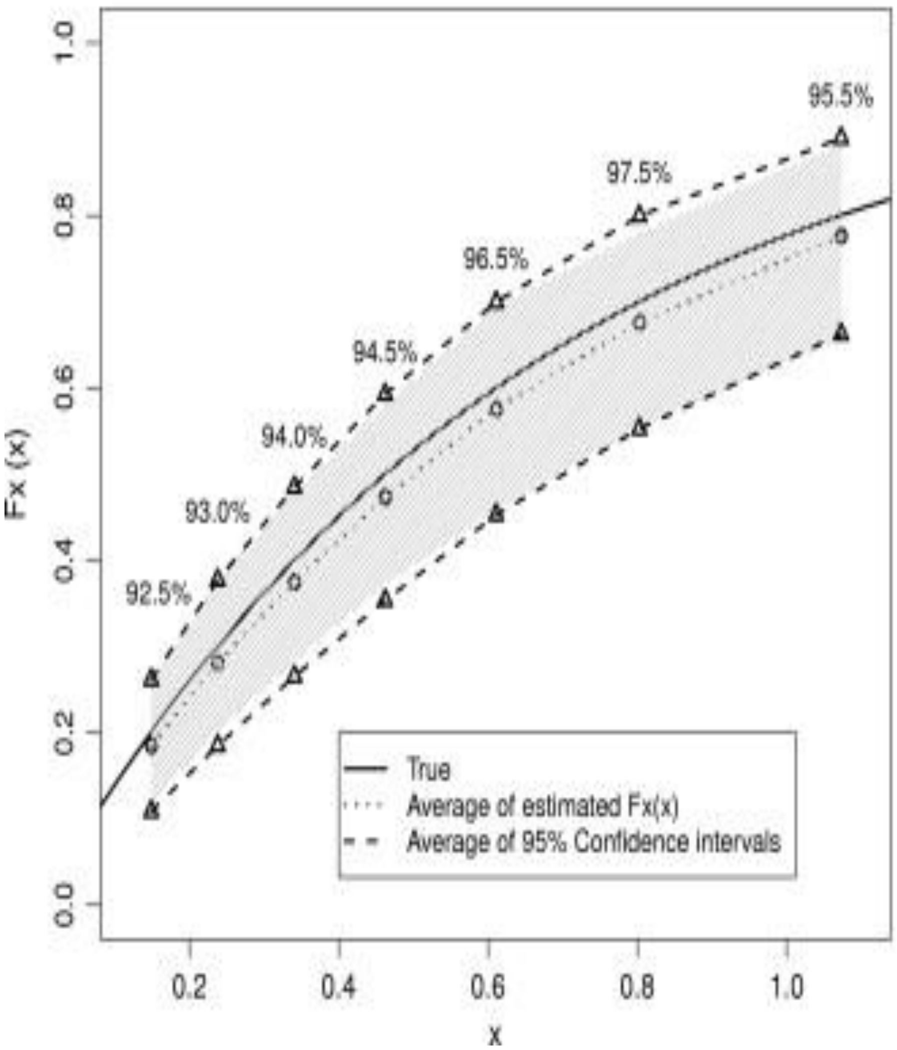
Parameter		*Mean*(*Bias*)	*SE*	*SEE*	95%*Cov*
Spearman's ρ = 0.25 (α =1/2.15, $\Pr(X \leq Y) = 0.79$)					
$\log(\alpha)$ =-0.765	n =125	-0.778 (-0.013)	0.407	0.407	0.945
	n =250	-0.697 (0.068)	0.311	0.296	0.965
$H_X(t) = 0.693$	n =125	0.736 (0.043)	0.123	0.121	0.955
	n =250	0.733 (0.040)	0.090	0.086	0.970
$\Lambda_Y(t) = 0.693$	n =125	0.710 (0.017)	0.144	0.139	0.960
	n =250	0.725 (0.032)	0.104	0.102	0.970
Spearman's ρ = 0.50 (α =1/5.11, $\Pr(X \leq Y) = 0.84$)					
$\log(\alpha)$ =-1.631	n =125	-1.642 (-0.011)	0.323	0.319	0.965
	n =250	-1.652 (-0.021)	0.231	0.222	0.940
$H_X(t) = 0.693$	n =125	0.726 (0.033)	0.101	0.092	0.910
	n =250	0.716 (0.023)	0.067	0.064	0.920
$\Lambda_Y(t) = 0.693$	n =125	0.704 (0.011)	0.110	0.102	0.960
	n =250	0.701 (0.008)	0.068	0.069	0.950

Negative dependence, 200 repetitions

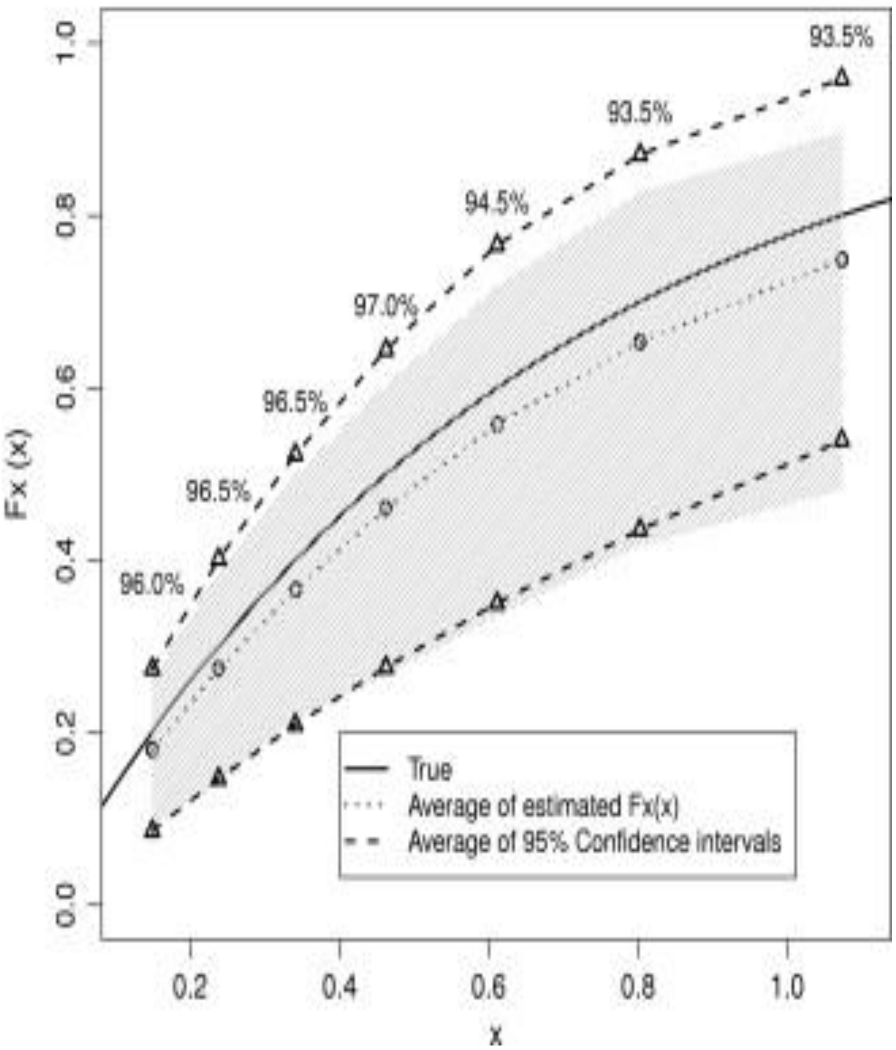
Parameter		*Mean*(*Bias*)	*SE*	*SEE*	95%*Cov*
Spearman’s ρ = -0.25 (α =2.15, $\Pr(X \le Y) = 0.72$)					
$\log(\alpha)$ =0.765	n =125	0.859 (0.094)	0.598	0.554	0.960
	n =250	0.717 (-0.048)	0.342	0.359	0.930
$H_X(t) = 0.693$	n =125	0.809 (0.116)	0.313	0.244	0.960
	n =250	0.717 (0.024)	0.139	0.138	0.935
$\Lambda_Y(t) = 0.693$	n =125	0.793 (0.100)	0.363	0.267	0.960
	n =250	0.699 (0.006)	0.139	0.137	0.930
Spearman’s ρ = -0.50 (α =5.11, $\Pr(X \le Y) = 0.70$)					
$\log(\alpha)$ = 1.631	n =125	1.758 (0.127)	0.818	0.598	0.915
	n =250	1.708 (0.077)	0.534	0.386	0.955
$H_X(t) = 0.693$	n =125	0.883 (0.190)	0.582	0.343	0.925
	n =250	0.787 (0.094)	0.374	0.196	0.960
$\Lambda_Y(t) = 0.693$	n =125	0.862 (0.169)	0.624	0.354	0.885
	n =250	0.775 (0.082)	0.404	0.207	0.955

Performance of $\hat{F}_X(x) = e^{-\hat{H}_X(x)}$

Positive Association: Kendall's tau = 0.25



Negative Association: Kendall's tau = -0.25



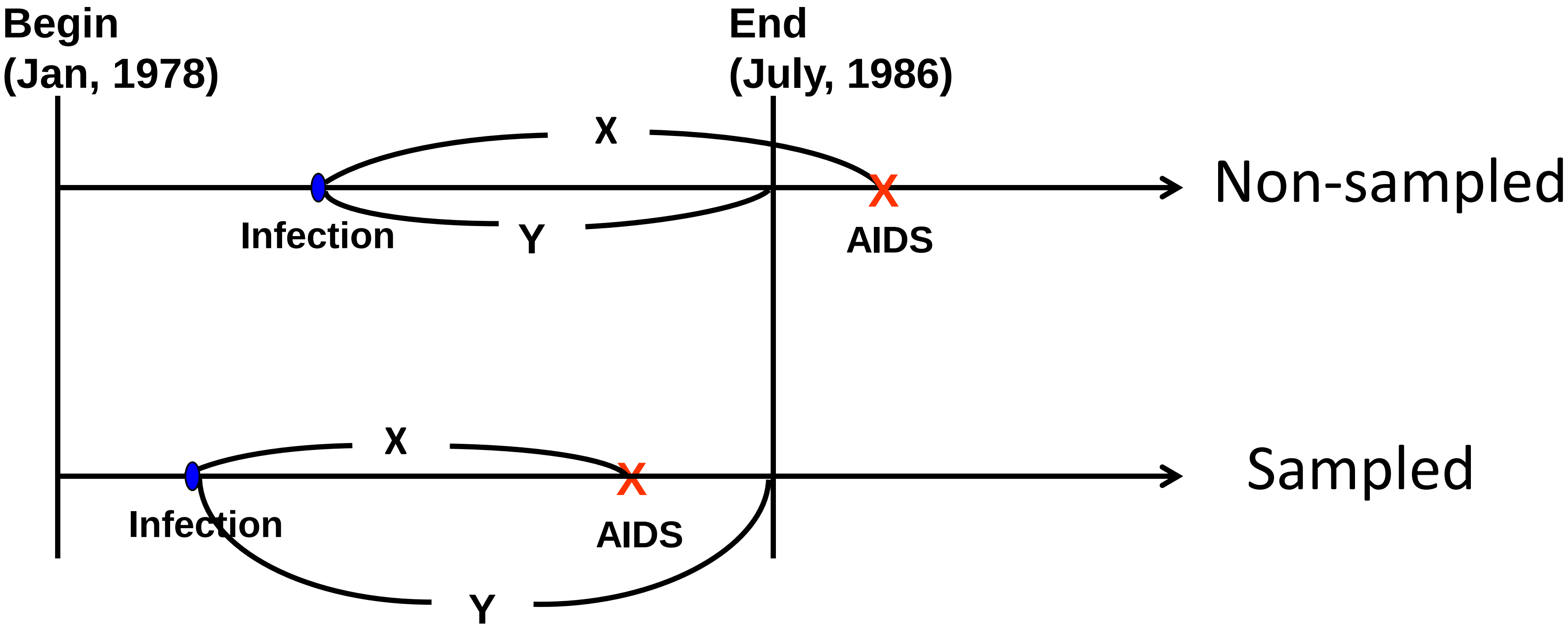
Data analysis

- Transfusion-related AIDS (Kalbfleisch & Lawless, 1989, JASA)

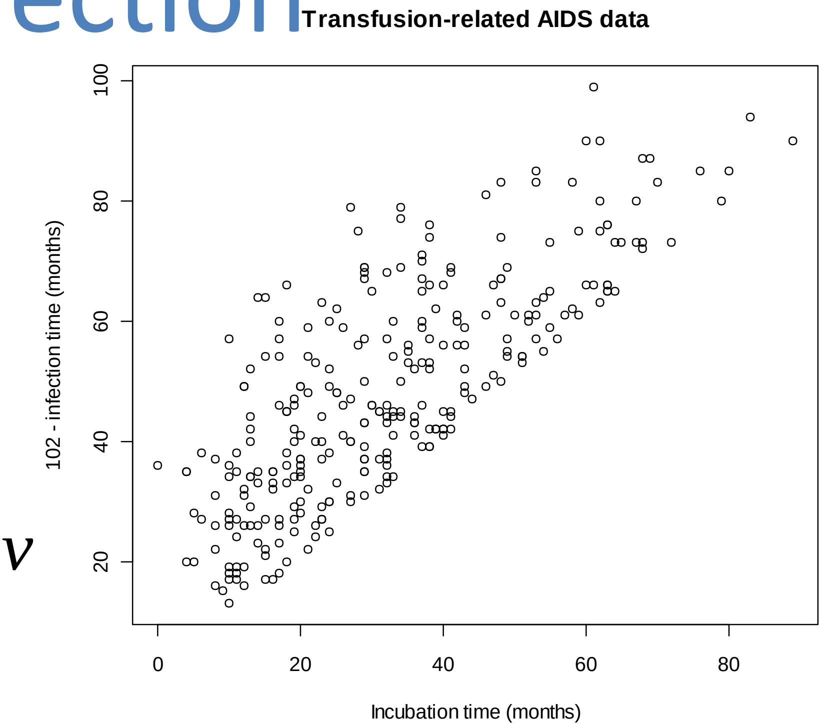
X : Time from infection to AIDS (month) ← Estimation

Y : 102 - time of infection (month)

n : sample size = 293



Model selection



- (*K+1*) candidate copulas

$$\begin{cases} C^{(0)}[u,v] = uv \\ C_\alpha^{(k)}[u,v], \quad k = 1,\dots,K \end{cases}$$

$$\text{where} \quad \lim_{\alpha \to 1} C_\alpha^{(k)}[u,v] = uv$$

- Deviance: $C_\alpha^{(k)}$ vs. $C^{(0)}$

$$2\{l_n^{(k)}(\hat{\alpha}, \hat{H}_X, \hat{\Lambda}_Y) - l_n(1, \hat{H}_X^{\alpha=1}, \hat{\Lambda}_Y^{\alpha=1})\} \quad \sim \quad \chi^2_{df=1}$$

Step 1: Calculate deviances for *K* copulas

Step 2: Choose the copula with smallest p-value (<0.05)

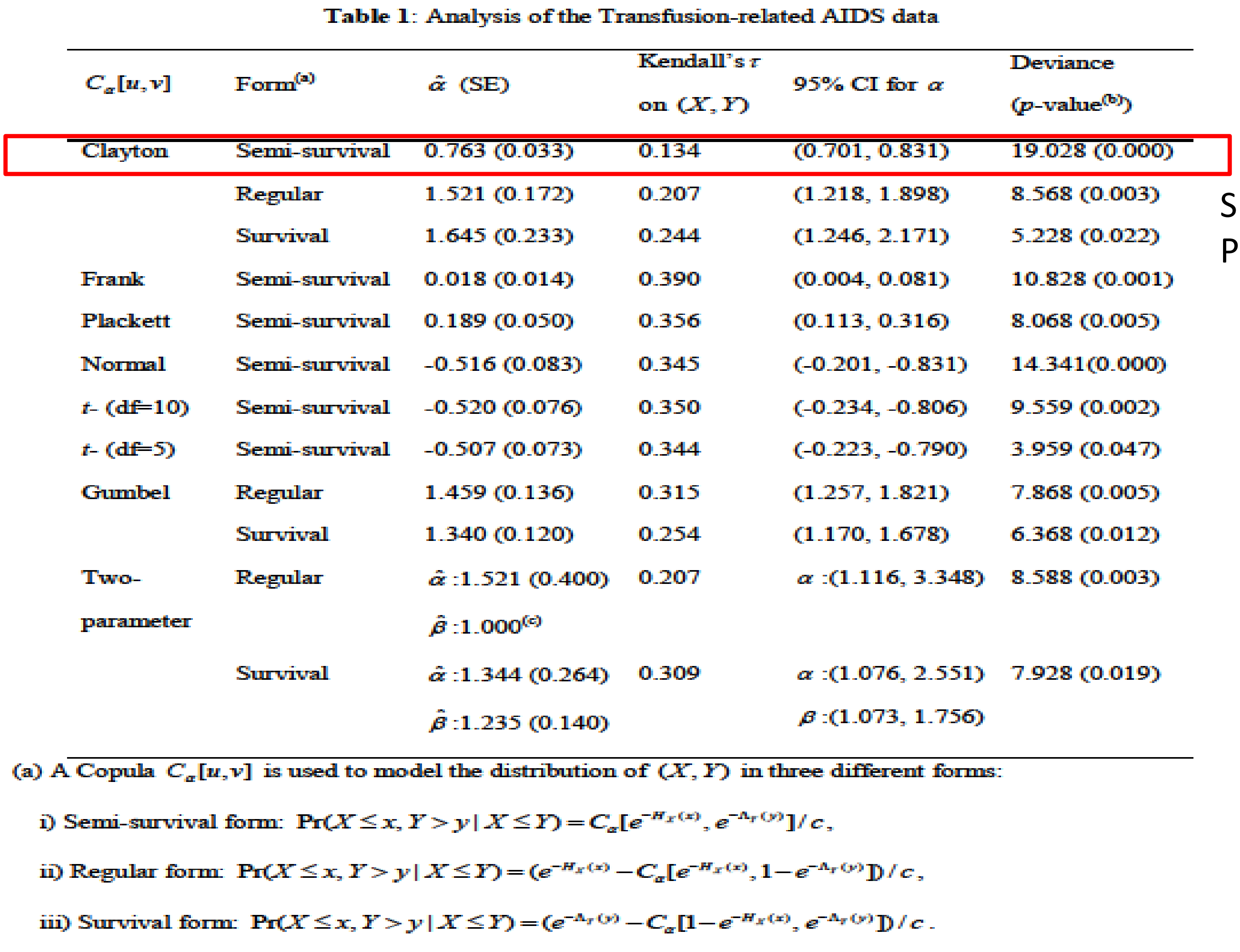


Table 1: Analysis of the Transfusion-related AIDS data

$C_\alpha[u,v]$	Form[a]	$\hat{\alpha}$ (SE)	Kendall's τ on (X,Y)	95% CI for α	Deviance (p-value[b])
Clayton	Semi-survival	0.763 (0.033)	0.134	(0.701, 0.831)	19.028 (0.000)
	Regular	1.521 (0.172)	0.207	(1.218, 1.898)	8.568 (0.003)
	Survival	1.645 (0.233)	0.244	(1.246, 2.171)	5.228 (0.022)
Frank	Semi-survival	0.018 (0.014)	0.390	(0.004, 0.081)	10.828 (0.001)
Plackett	Semi-survival	0.189 (0.050)	0.356	(0.113, 0.316)	8.068 (0.005)
Normal	Semi-survival	-0.516 (0.083)	0.345	(-0.201, -0.831)	14.341(0.000)
t- (df=10)	Semi-survival	-0.520 (0.076)	0.350	(-0.234, -0.806)	9.559 (0.002)
t- (df=5)	Semi-survival	-0.507 (0.073)	0.344	(-0.223, -0.790)	3.959 (0.047)
Gumbel	Regular	1.459 (0.136)	0.315	(1.257, 1.821)	7.868 (0.005)
	Survival	1.340 (0.120)	0.254	(1.170, 1.678)	6.368 (0.012)
Two-parameter	Regular	$\hat{\alpha}$:1.521 (0.400) $\hat{\beta}$:1.000[c]	0.207	α :(1.116, 3.348)	8.588 (0.003)
	Survival	$\hat{\alpha}$:1.344 (0.264) $\hat{\beta}$:1.235 (0.140)	0.309	α :(1.076, 2.551) β :(1.073, 1.756)	7.928 (0.019)

(a) A Copula $C_\alpha[u,v]$ is used to model the distribution of (X,Y) in three different forms:

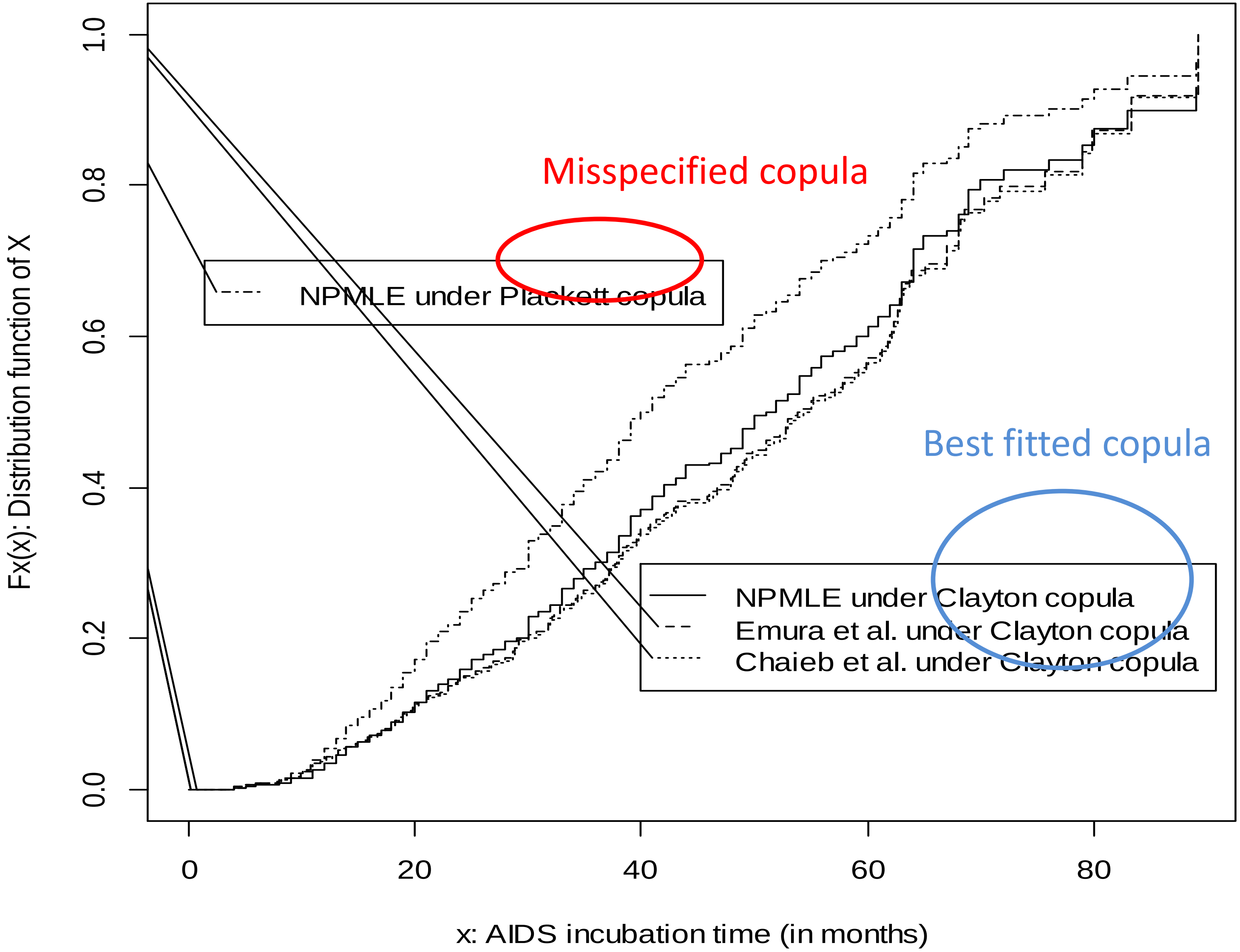
i) Semi-survival form: $\Pr(X \le x, Y > y \mid X \le Y) = C_\alpha[e^{-H_X(x)}, e^{-\Lambda_Y(y)}]/c$,

ii) Regular form: $\Pr(X \le x, Y > y \mid X \le Y) = (e^{-H_X(x)} - C_\alpha[e^{-H_X(x)}, 1-e^{-\Lambda_Y(y)}])/c$,

iii) Survival form: $\Pr(X \le x, Y > y \mid X \le Y) = (e^{-\Lambda_Y(y)} - C_\alpha[1-e^{-H_X(x)}, e^{-\Lambda_Y(y)}])/c$.

Smallest P-value

$\hat{F}_X(x) = e^{-\hat{H}_X(x)}$: Time from infection to AIDS (month)



Summary

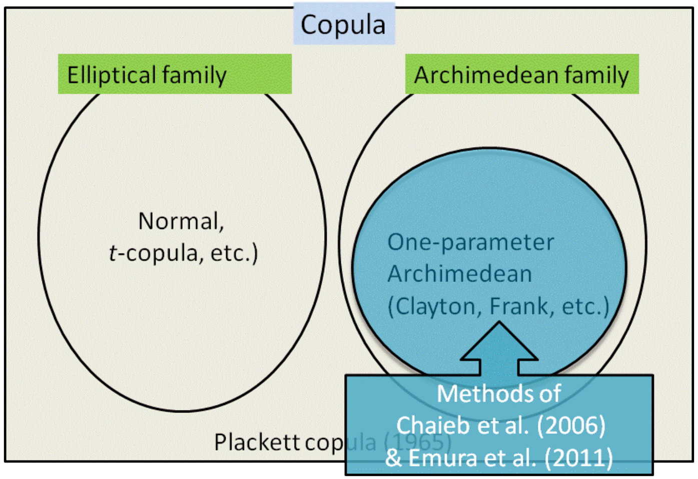
We proposed the NPMLE under semi-survival copula with reverse-time hazard modeling

Advantage of the proposed NPMLE

- Broader class of copula than existing methods

than Chaieb et al.(2006) & Emura et al. (2011)

- Equip Copula model selection via likelihood comparison



Disadvantage

- NPMLE is computationally very demanding

(numerical maximization in high-dimensional space)

Thank you for your kind attention

- Emura T & Wang W (2012)
  Nonparametric maximum likelihood estimation for dependent truncation data based on copulas.
  to appear in Journal of Multivariate Analysis.

- Proposed method is implemented in R depend.truncation package.
  available at CRAN http://cran.r-project.org/