A copula-based Markov chain model for attribute data

Takeshi Emura

Joint work with

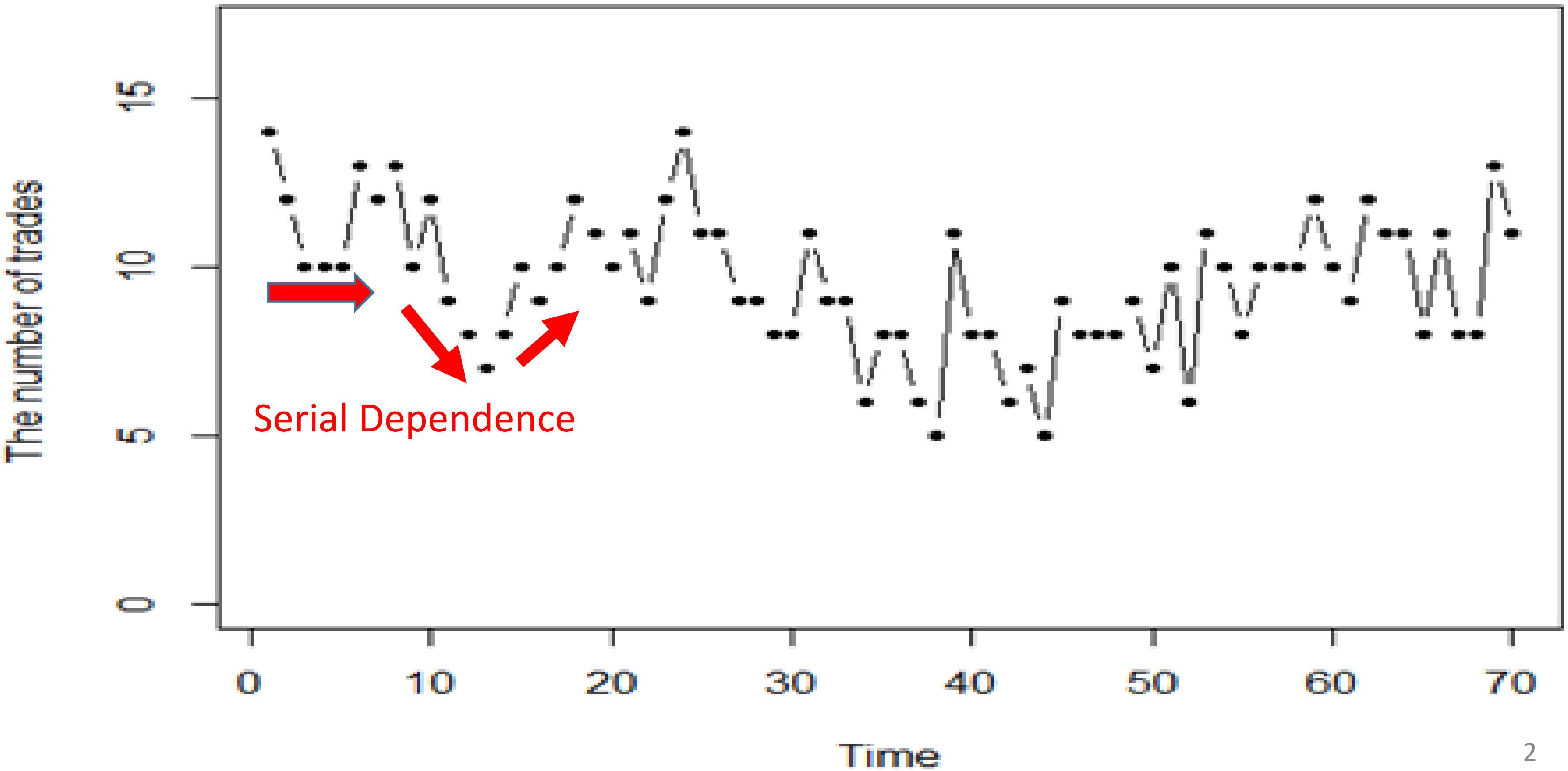
Xinwei Huang (NCTU) and Wei-Ru Chen (NCU)

Presented at 國立高雄大學統計學研究所

Korean stock market data

Weiß and Kim (2013 Stat Pap)

- **No. of trades** in n = 22 sectors, at $T = 70$ different time points



Data structure

- Time series $\{Y_t : t = 1, 2, \ldots, T\}$
  - discrete time, discrete valued

- Binomial margins $Y_t \sim Bin(n, p)$
  - p is unknown, n is known

- No. of trades in n = 22 sectors, at $T = 70$ different time points

Binomial AR(1) model

McKenzie (1985) proposed the binomial AR(1) model defined by

$$Y_t = \alpha \circ Y_{t-1} + \beta \circ (n - Y_{t-1}), \qquad t = 2, 3, ..., T,$$

where $\beta \equiv p(1-\rho)$, $\alpha \equiv \beta + \rho$, $p \in (0,1)$, $\rho \in (\max\{-p/(1-p), (-1+p)/p\}, 1)$, and

$\alpha \circ y := \sum_{i=1}^{y} X_i$, where $X_i \sim Bin(1, \alpha)$

$$Corr(Y_t, Y_{t-1}) = \frac{\text{Cov}(Y_t, Y_{t-1})}{\sqrt{V(Y_t)}\sqrt{V(Y_{t-1})}} = \frac{\rho np(1-p)}{np(1-p)} = \rho$$ ←Unknown

Binomial AR(1) model

Transition Probability

$$g_{p,\rho}(y_t \mid y_{t-1}) = P(Y_t = y_t \mid Y_{t-1} = y_{t-1}) = P(\alpha \circ Y_{t-1} + \beta \circ (n - Y_{t-1}) = y_t \mid Y_{t-1} = y_{t-1})$$

$$= \sum_{k=\max\{0, y_t + y_{t-1} - n\}}^{\min\{y_t, y_{t-1}\}} \binom{y_{t-1}}{k} \binom{n - y_{t-1}}{y_{t-1} - k} \alpha^k (1-\alpha)^{y_{t-1}-k} \beta^{y_t - k} (1-\beta)^{n - y_{t-1} + k - y_t} ,$$

The log-likelihood function based on the observations $\{Y_t : t = 1, 2, ..., T\}$ is given by

$$\ell_{\mathrm{AR1}}(p, \rho) = \log\left\{ \binom{n}{Y_1} p^{Y_1} (1-p)^{n-Y_1} \right\} + \sum_{t=2}^{T} \log g_{p,\rho}(Y_t \mid Y_{t-1}) .$$

MLE by $(\hat{p}^{\mathrm{AR1ML}}, \hat{\rho}^{\mathrm{AR1ML}}) = \operatorname{argmax} \ell_{\mathrm{AR1}}(p, \rho)$,

Weiß and Kim (2013 *Statistics*)

AR(1) vs. Copula model

- **AR(1) model**

  Only a linear dependence

- **Copula model (proposed)**

  Flexible dependence patterns

  -Clayton, Gumbel, Frank, FGM, Plackett, etc.

  Extreme dependence

  - Lower tail dependence (Clayton),
  - Upper tail dependence (Joe)
  - Both lower and upper tail (*t*-copula)

  Higher order models

  - 2^{nd} order Markov chain

Copula-based methods have not been developed for discrete margins

→ We develop for the first time !

Copulas

Copula in Latin: a link, a tie, a bond. (Sklar, 1959)

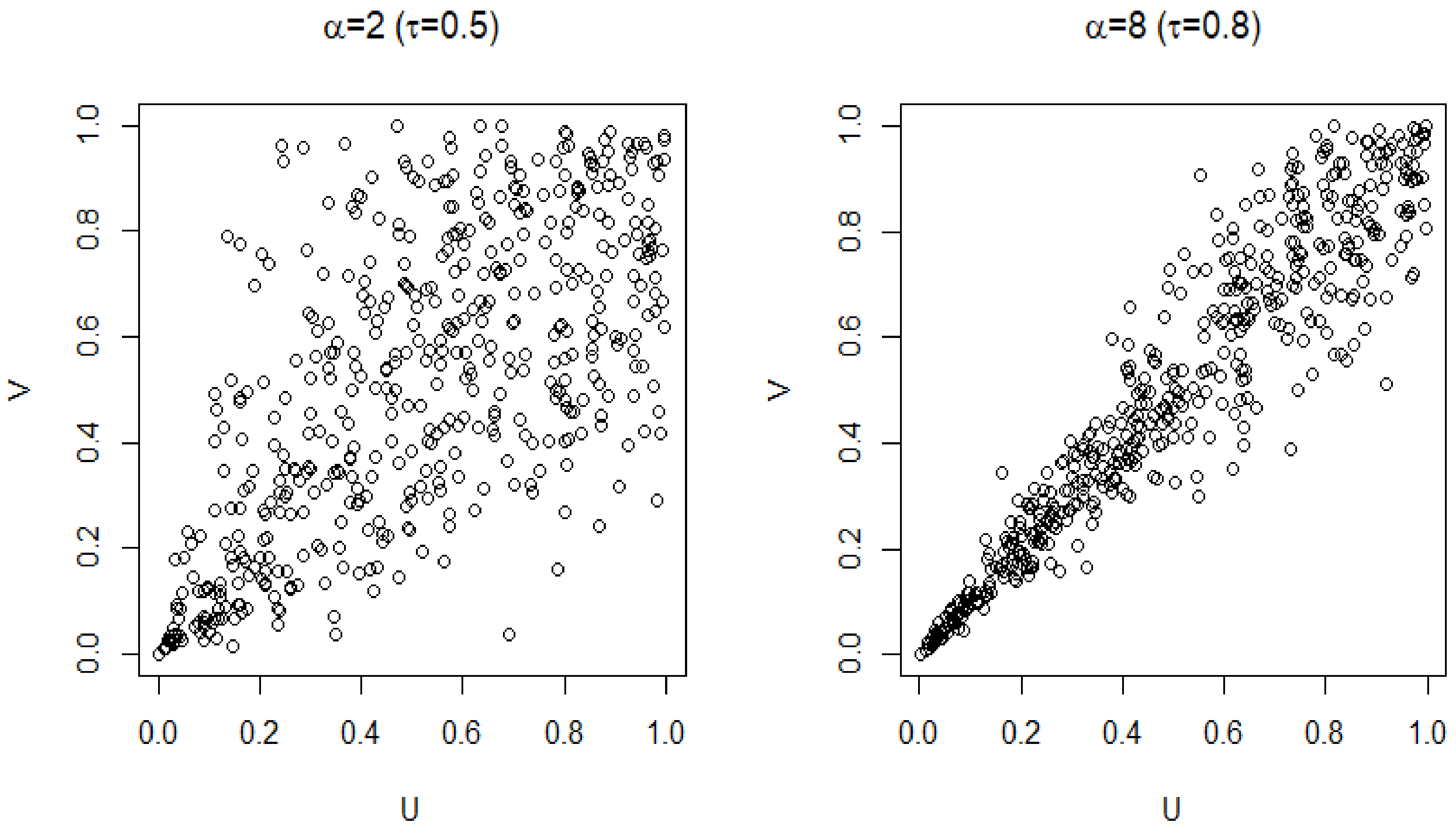
A bivariate copula is a function $C:[0,1]^2 \to [0,1]$ satisfying:

(1) $C(u,0)=C(0,v)=0$, $C(u,1)=u$ and $C(1,v)=v$

(2) For every $u_1 < u_2$ and $v_1 < v_2$,

$$C(u_2,v_2)-C(u_2,v_1)-C(u_1,v_2)+C(u_1,v_1)\geq 0$$

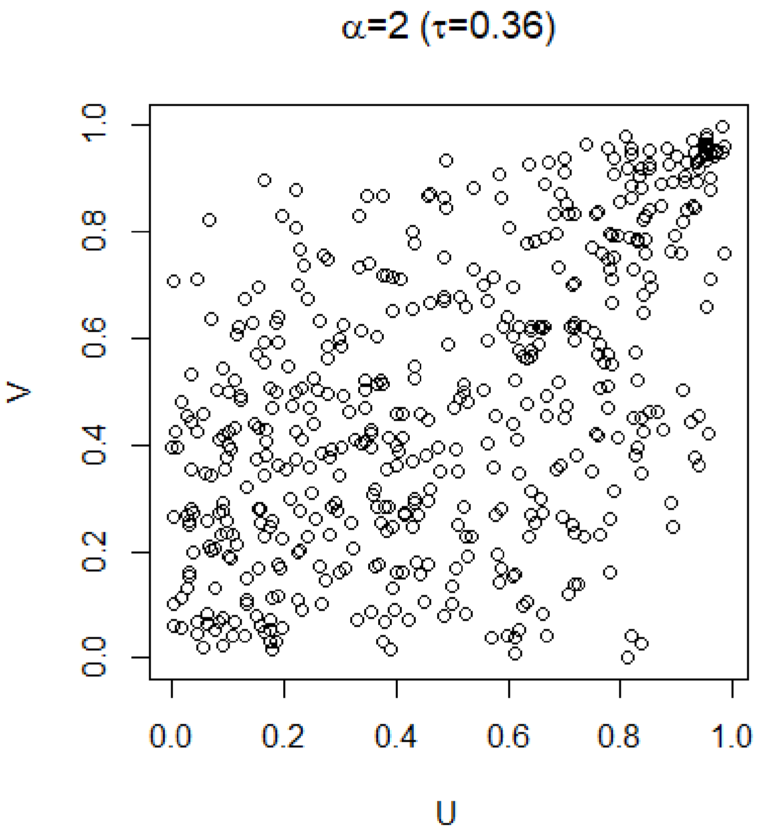
The Clayton copula; $C_\alpha(u,v) = \max(u^{-\alpha} + v^{-\alpha} - 1, 0)^{-1/\alpha}$

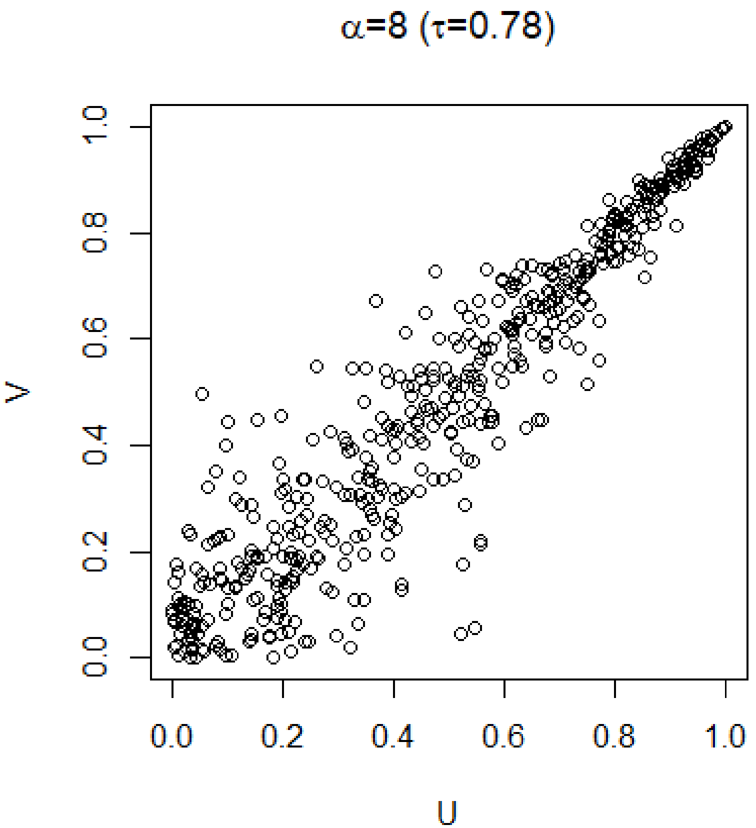


Kendall's tau: $\tau = 4\int_0^1\int_0^1 C(u,v)\,dC(u,v) - 1 = \alpha/(\alpha+2)$

The Joe copula;

$$C_\alpha(u,v) = 1 - \{(1-u)^\alpha + (1-v)^\alpha - (1-u)^\alpha(1-v)^\alpha\}^{1/\alpha}$$





Kendall's tau: $\tau = 1 - 4\int_0^\infty t(1-e^{-t})^{2/\alpha-2} e^{-2t} / \alpha^2 dt$

Copula-based Markov chain

Darsow et al. (1992 Illinois J of Math)

- Copula structure between *t-1* and *t*.

$$\Pr(Y_t \le y_t, Y_{t-1} \le y_{t-1}) = C\{G(y_t), G(y_{t-1})\}$$

- Markov assumption

$$\Pr(Y_t \le y_t \mid Y_{t-1} = y_{t-1}, Y_{t-1} = y_{t-2}, \ldots) = \Pr(Y_t \le y_t \mid Y_{t-1} = y_{t-1})$$

$$\{Y_t : t = 1, 2, \ldots, T\}$$

Stationary Markov process

(Joe 1997; Chen and Fan 2006)

Statistical Inference for copula-based Markov time-series models

- Chen and Fan (2006) → nonparametric margins
- Long and Emura (2014) → normal margins
- Emura, Long, Sun (2017) → R package "*Copula.Markov*"
- Sun, Lee and Emura (2018) → *t*-margins, Bayesian inference
- Huang and Emura (2019) → Model disgnoistc
- Lin, Emura, Sun (2019)→ Normal mixture margins
- Huang, Chen, Emura (20??)→ Binomial margins

Proposed model

(Assumption I) Markov property:

$$\Pr(Y_t \le y_t \mid Y_{t-1} = y_{t-1}, Y_{t-1} = y_{t-2}, \ldots) = \Pr(Y_t \le y_t \mid Y_{t-1} = y_{t-1})$$

(Assumption II) Marginal distribution:

$$G(y) = \sum_{x=0}^{y} \binom{n}{x} p^x (1-p)^{n-x}, \qquad y = 0, 1, \ldots, n$$

(Assumption III) Parametric copula

$$\Pr(Y_t \le y_t, Y_{t-1} \le y_{t-1}) = C_{\boldsymbol{\alpha}}\{G(y_t), G(y_{t-1})\}$$

Dependence parameter

Data generation (inverse method)

Algorithm 1: Data generation

Step 1: Draw $Y_1 \sim Bin(n, p)$.

Step 2: Given y_{t-1}, obtain Y_t as the solution to the equation

$$U_t \times g(y_{t-1}) = C_\alpha(G(y_t), G(y_{t-1})) - C_\alpha(G(y_t), G(y_{t-1}-1))$$

for y_t, $t = 2, 3, ..., T$, where $U_t \sim U(0,1)$.

- Proposed R function:

Clayton.Markov.DATA.binom(n, size, prob, alpha)

n = number of observations

size = number of binomial trials

prob = binomial probability; 0<p<1

alpha = copula parameter

Example; Clayton copula

We first consider the Clayton copula model

$$P(Y_t \le y_t \ , Y_{t-1} \le y_{t-1}) = A_\alpha(y_t \ , \ y_{t-1})^{-1/\alpha}, \qquad y_t = 0, 1, ..., n$$

where $A_\alpha(y_t \ , \ y_{t-1}) = G(y_t)^{-\alpha} + G(y_{t-1})^{-\alpha} - 1$. The transition distribution function is

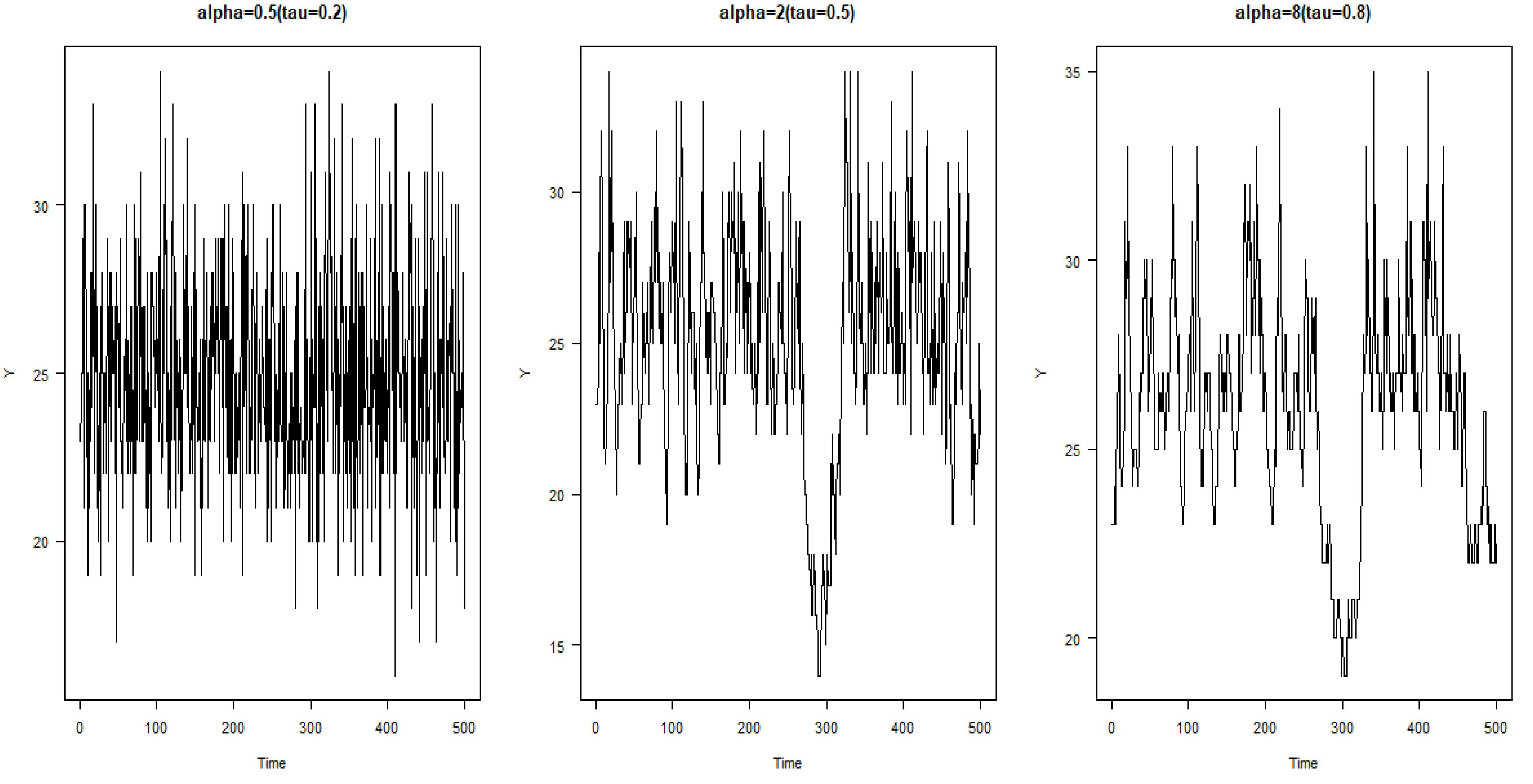
$$P(Y_t \le y_t \middle| Y_{t-1} = y_{t-1}) = \frac{A_\alpha(y_t \ , \ y_{t-1})^{-1/\alpha} - A_\alpha(y_t \ , \ y_{t-1} - 1)^{-1/\alpha}}{g(y_{t-1})}, \qquad y_t = 0, 1, ..., n\,.$$

The conditional density function of Y_t given $Y_{t-1} = y_{t-1}$ is

$$g(\, y_t \middle| y_{t-1}) = P(Y_t \le y_t \middle| Y_{t-1} = y_{t-1}) - P(Y_t \le y_t - 1 \middle| Y_{t-1} = y_{t-1})$$

$$= \frac{A_\alpha(y_t, \ y_{t-1})^{-1/\alpha} - A_\alpha(y_t, \ y_{t-1} - 1)^{-1/\alpha}}{g(y_{t-1})} - \frac{A_\alpha(y_t - 1, \ y_{t-1})^{-1/\alpha} - A_\alpha(y_t - 1, \ y_{t-1} - 1)^{-1/\alpha}}{g(y_{t-1})},$$

Generating time series from the Clayton: *T*=500 time points, *n*=50, and *p*=0.5



Likelihood (Clayton copula)

$$\ell(p,\alpha) = \log\{g(y_1)\} - \sum_{t=2}^{T} \log\{g(y_{t-1})\}$$

$$+\sum_{t=2}^{T} \log\{A_\alpha(y_t, y_{t-1})^{-1/\alpha} - A_\alpha(y_t, y_{t-1}-1)^{-1/\alpha}$$

$$-A_\alpha(y_t - 1, y_{t-1})^{-1/\alpha} + A_\alpha(y_t - 1, y_{t-1}-1)^{-1/\alpha}\}$$

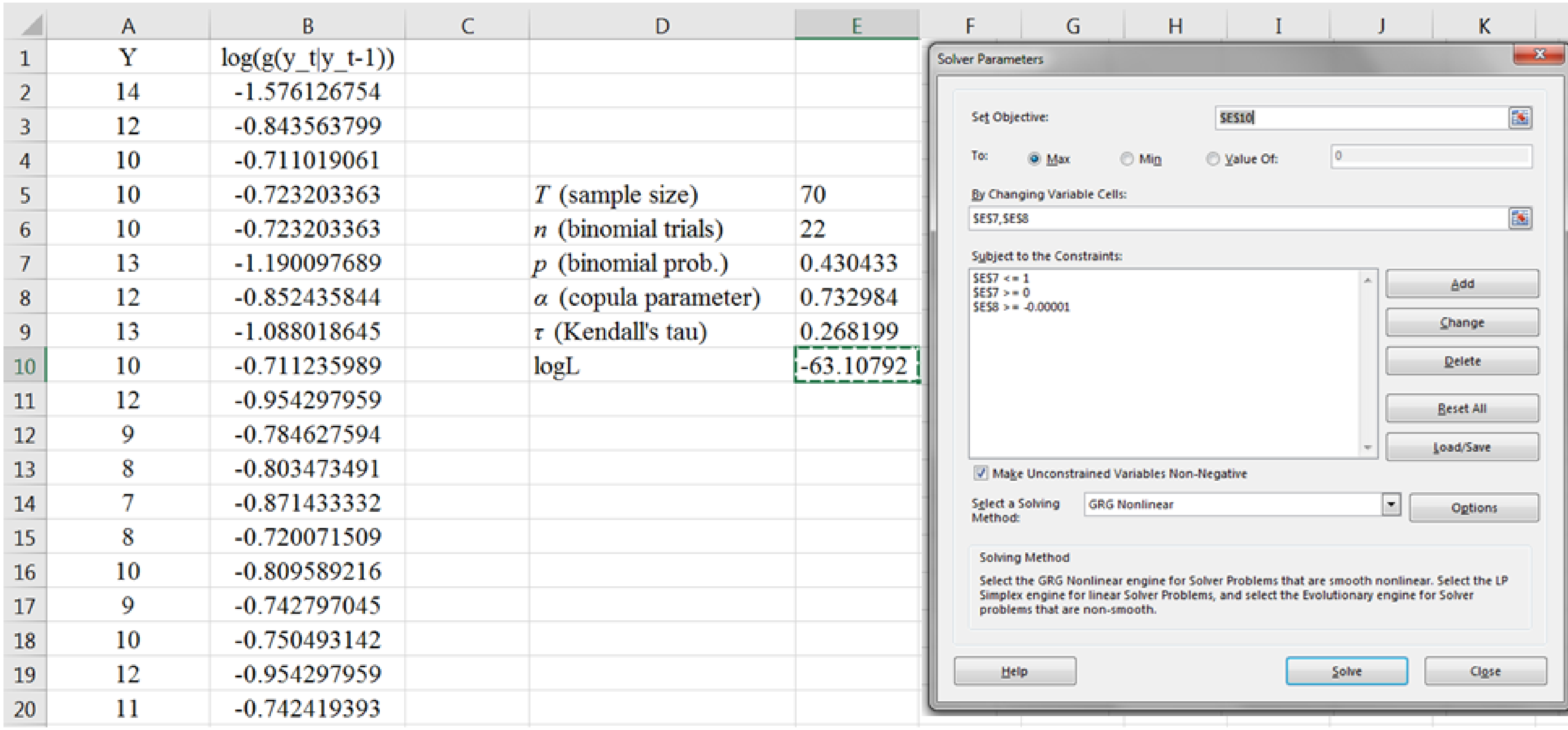
where $A_\alpha(y_t, y_{t-1}) = G(y_t)^{-\alpha} + G(y_{t-1})^{-\alpha} - 1$

Proposed R function for finding the MLE:

Clayton.Markov.MLE.binom(Y, size, k = 3, method="nlm", plot = TRUE, GOF=FALSE)

Y = vector of observations
size = numbe of binomial trials
method = nlm or Newton
k = determine the length between LCL and UCL (k=3 corresponds to 3-sigma limit)
GOF = show the model diagnostic plot if TRUE

Excel users can use “Solver”



	A	B	C	D	E
1	Y	log(g(y_t|y_t-1))			
2	14	-1.576126754			
3	12	-0.843563799			
4	10	-0.711019061			
5	10	-0.723203363		T (sample size)	70
6	10	-0.723203363		n (binomial trials)	22
7	13	-1.190097689		p (binomial prob.)	0.430433
8	12	-0.852435844		α (copula parameter)	0.732984
9	13	-1.088018645		τ (Kendall's tau)	0.268199
10	10	-0.711235989		logL	-63.10792
11	12	-0.954297959			
12	9	-0.784627594			
13	8	-0.803473491			
14	7	-0.871433332			
15	8	-0.720071509			
16	10	-0.809589216			
17	9	-0.742797045			
18	10	-0.750493142			
19	12	-0.954297959			
20	11	-0.742419393			

Fig. A. The use of Excel Solver under the Calyton copula.

Maximum likelihood estimator (MLE)

- Transformation:

$$\mathrm{P} = \log(1/p - 1) \quad , \quad \mathrm{A} = \log(\alpha + 1)$$

such that $\mathrm{P} \in \mathbb{R}$ and $\mathrm{A} \in \mathbb{R}$, where $\mathbb{R} \equiv (-\infty, \infty)$.

- The transformed log-likelihood function

$$\tilde{\ell}\,(\mathrm{P}, \mathrm{A}) = \ell\,(1/(\,\mathrm{e}^{\mathrm{P}} + 1\,),\ \mathrm{e}^{\mathrm{A}} - 1) = \ell\,(p, \alpha)\,.$$

The MLE

$$(\,\hat{\mathrm{P}}^{\mathrm{ML}}, \hat{\mathrm{A}}^{\mathrm{ML}}) = \underset{\mathrm{A}\in\mathbb{R},\ \mathrm{P}\in\mathbb{R}}{\operatorname{argmax}}\, \tilde{\ell}\,(\mathrm{P}, \mathrm{A})\,.$$

- $\sqrt{T}\,(\hat{\mathrm{P}}^{\mathrm{ML}} - \mathrm{P}, \hat{\mathrm{A}}^{\mathrm{ML}} - \mathrm{A})^{\mathrm{T}} \xrightarrow{\ d\ } N(\mathbf{0}, \mathbf{I}^{-1}(\mathrm{P}, \mathrm{A})) \quad \text{as} \quad T \to \infty\,.$

  Ref: Billingsley (1961)

Model diagnostic
- testing the binomial distribution

Goodness-of-fit

$$H_0 : \Pr(Y_t \le y) = G(y) = \sum_{x=0}^{y} \binom{n}{x} p^x (1-p)^{n-x}$$

$$H_1 : \Pr(Y_t \le y) \ne G(y)$$

Test statistics

$$K = \sup \left| G_n(y_j) - G(y_j;\hat{p}) \right| \quad \text{or} \quad C = \sum_j \left\{ G_n(y_j) - G(y_j;\hat{p}) \right\}^2$$

The goodness-of-fit test with parametric bootstrap

Step 1: Generate $\{ Y_t^{(b)} : t = 1, ..., T \}$ under H_0 and given $\hat{p}^{\mathrm{ML}}$ and $\hat{\alpha}^{\mathrm{ML}}$ for $b = 1, 2, \ldots, B$.

Step 2: Compute $\hat{p}^{\mathrm{ML(b)}}$, $\hat{\alpha}^{\mathrm{ML(b)}}$, $F^{ML}(y_t, y_{t-1}; \hat{\alpha}^{ML(b)}, \hat{p}^{ML(b)})$, and $F^{NP(\mathrm{b})}(y_t, y_{t-1})$ from the data $\{ Y_t^{(b)} : t = 1, ..., T \}$. Then, compute $C^{(b)}$ for each $b = 1, 2, \ldots, B$.

Step 3: The P-value of the test is calculated as $\sum_{b=1}^{B} \mathbf{I}(C^{(b)} \ge C) / B$.

MLE for 3-σ Control Limits

$$\mu = E[Y_t], \quad \sigma^2 = Var(Y_t),$$

- Lower Control Limit

  $$LCL = \mu - 3\sigma$$

- Upper Control Limit

  $$UCL = \mu + 3\sigma$$

Here we apply $\hat{\mu}^{ML} = n\hat{p}^{ML}$ and $\hat{\sigma}^{ML} = \sqrt{n\hat{p}^{ML}(1-\hat{p}^{ML})}$.

Proposed R functions in *Copula.Markov*

Clayton.Markov.DATA.binom

Clayton.Markov.GOF.binom

Clayton.Markov.MLE.binom

Joe.Markov.DATA.binom

Joe.Markov.MLE.binom

Benchmark: The standard (naïve) method

$$\hat{\mu}^{\text{STD}} = \frac{1}{T}\sum_{t=1}^{T} Y_t \,, \qquad \hat{\sigma}^{\text{STD}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} Y_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T} Y_t\right)^2} \,,$$

$$\text{LCL} = \hat{\mu}^{\text{STD}} - 3\hat{\sigma}^{\text{STD}}$$

$$\text{UCL} = \hat{\mu}^{\text{STD}} + 3\hat{\sigma}^{\text{STD}} .$$

🡐Consistent, but not efficient

$$\hat{p}^{\text{STD}} = \hat{\mu}^{\text{STD}} / n$$

$$\hat{\alpha}^{\text{STD}} = \tau^{-1}(\hat{\tau}_b^{\text{STD}})$$

←Inconsistent estimator

Sample Kendall's tau for

$$(Y_1, Y_2), \ (Y_2, Y_3), \ \ldots \ , (Y_{T-1}, Y_T) .$$

Simulation settings ; n=50

True model	T	p
Clayton $\tau = 0.5$ $\alpha = 2$	50	0.01
		0.05
		0.10
	100	0.01
		0.05
		0.10
	200	0.01
		0.05
		0.10
AR(1) $\rho = 0.5$	50	0.01
		0.05
		0.10
	100	0.01
		0.05
		0.10
	200	0.01
		0.05
		0.10

Evaluation criterion:
Mean squared error (MSE)

$$\mathrm{MSE}(\hat{p}) = \mathrm{E}[(\hat{p} - p)^2]$$

$$\mathrm{MSE}(\hat{\mu} + 3\hat{\sigma}) = \mathrm{E}[\{\hat{\mu} + 3\hat{\sigma} - (\mu + 3\sigma)\}^2]$$

Upper Control Limit

True model
= Clayton copula

Best performance
= Smallest MSE

True model
= AR(1)

Best performance
= Smallest MSE

e model	T	p	$E(\hat{p})$			$MSE(\hat{p})$		
			Proposed	AR(1)	Standard	Proposed	AR(1)	Standard
Clayton copula	50	0.01	0.009700	0.009925	0.009933	0.000009	0.000010	0.000010
		0.05	0.048473	0.049553	0.049477	0.000070	0.000099	0.000102
		0.10	0.097216	0.099436	0.099253	0.000157	0.000227	0.000234
	100	0.01	0.009864	0.010009	0.010012	0.000004	0.000005	0.000005
		0.05	0.049319	0.049872	0.049868	0.000033	0.000049	0.000049
		0.10	0.098715	0.099643	0.099632	0.000070	0.000116	0.000117
	200	0.01	0.009936	0.010007	0.010009	0.000002	0.000002	0.000002
		0.05	0.049659	0.049916	0.049923	0.000014	0.000023	0.000024
		0.10	0.099357	0.099758	0.099762	0.000030	0.000057	0.000058
AR(1)	50	0.01	0.010139	0.009985	0.009978	0.000016	0.000012	0.000011
		0.05	0.050845	0.050067	0.050070	0.000069	0.000054	0.000055
		0.10	0.100720	0.099770	0.099799	0.000128	0.000103	0.000104
	100	0.01	0.010291	0.009959	0.009957	0.000009	0.000006	0.000006
		0.05	0.051257	0.050061	0.050059	0.000038	0.000029	0.000029
		0.10	0.101609	0.099947	0.099951	0.000070	0.000055	0.000055
	200	0.01	0.010325	0.009970	0.009971	0.000004	0.000003	0.000003
		0.05	0.051387	0.050044	0.050049	0.000020	0.000014	0.000014
		0.10	0.101751	0.099842	0.099854	0.000037	0.000027	0.000028

True model
= Clayton copula

Best performance
= Smallest MSE

True model
= AR(1)

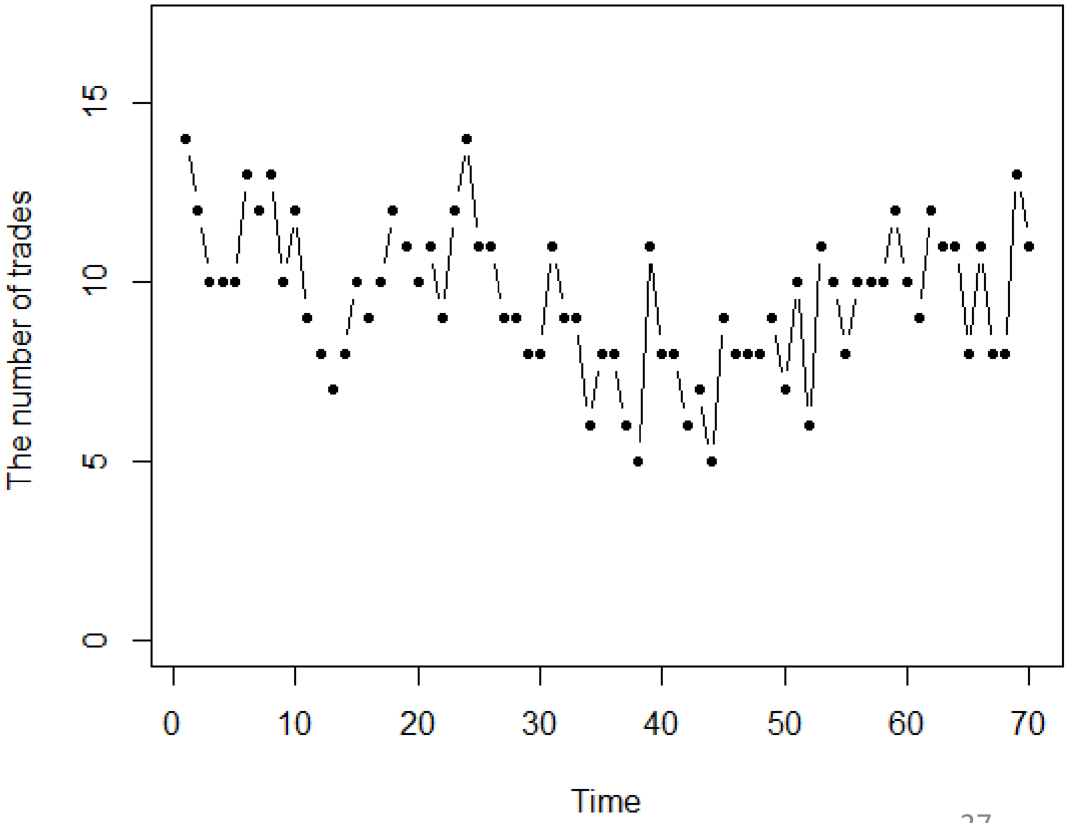
Best performance
= Smallest MSE

			$E(\hat{\mu}+3\hat{\sigma})$			$MSE(\hat{\mu}+3\hat{\sigma})$		
e model	T	$\mu+3\sigma$	Proposed	AR(1)	Standard	Proposed	AR(1)	Standard
ayton	50	2.61	2.537858	2.569410	2.571437	0.226859	0.256254	0.250403
= 2		7.12	6.960515	7.053953	7.045808	0.645722	0.907078	0.939309
= 0.5)		11.36	11.129399	11.296059	11.280818	1.001672	1.437919	1.490273
	100	2.61	2.577538	2.598669	2.599108	0.105581	0.118108	0.118370
		7.12	7.050526	7.098339	7.097803	0.297503	0.438198	0.443465
		11.36	11.256230	11.324290	11.323242	0.438674	0.720785	0.730584
	200	2.61	2.595106	2.605548	2.605831	0.050508	0.055207	0.055423
		7.12	7.087649	7.109375	7.109945	0.126124	0.206260	0.208934
		11.36	11.310664	11.339330	11.339532	0.183930	0.350892	0.356498
.R(1)	50	2.61	2.592335	2.576765	2.575684	0.358081	0.277372	0.276382
= 0.5)		7.12	7.185474	7.115688	7.115706	0.593400	0.477103	0.485885
		11.36	11.407847	11.335805	11.337937	0.778561	0.630644	0.639533
	100	2.61	2.634559	2.588611	2.588321	0.191788	0.138666	0.138389
		7.12	7.231848	7.121617	7.121403	0.325778	0.253768	0.255017
		11.36	11.483155	11.354482	11.354717	0.426060	0.334374	0.337536
	200	2.61	2.650236	2.598019	2.598199	0.096640	0.071038	0.071444
		7.12	7.248567	7.123830	7.124275	0.170016	0.123414	0.124562
		11.36	11.497498	11.348936	11.349806	0.222027	0.167233	0.168577

Data analysis - Korean stock market

Weiß and Kim (2013 *Stat Pap*)

- n = 22
- p = ?
- serial dependence



Model diagnostic

- Likelihood-based copula selection

$$\ell_{\text{Clayton}} = -145.3113 > \ell_{\text{Joe}} = -145.3636$$

➔ Select the Clayton copula

- Goodness-of-fit test for the binomial

The P-value = 0.76

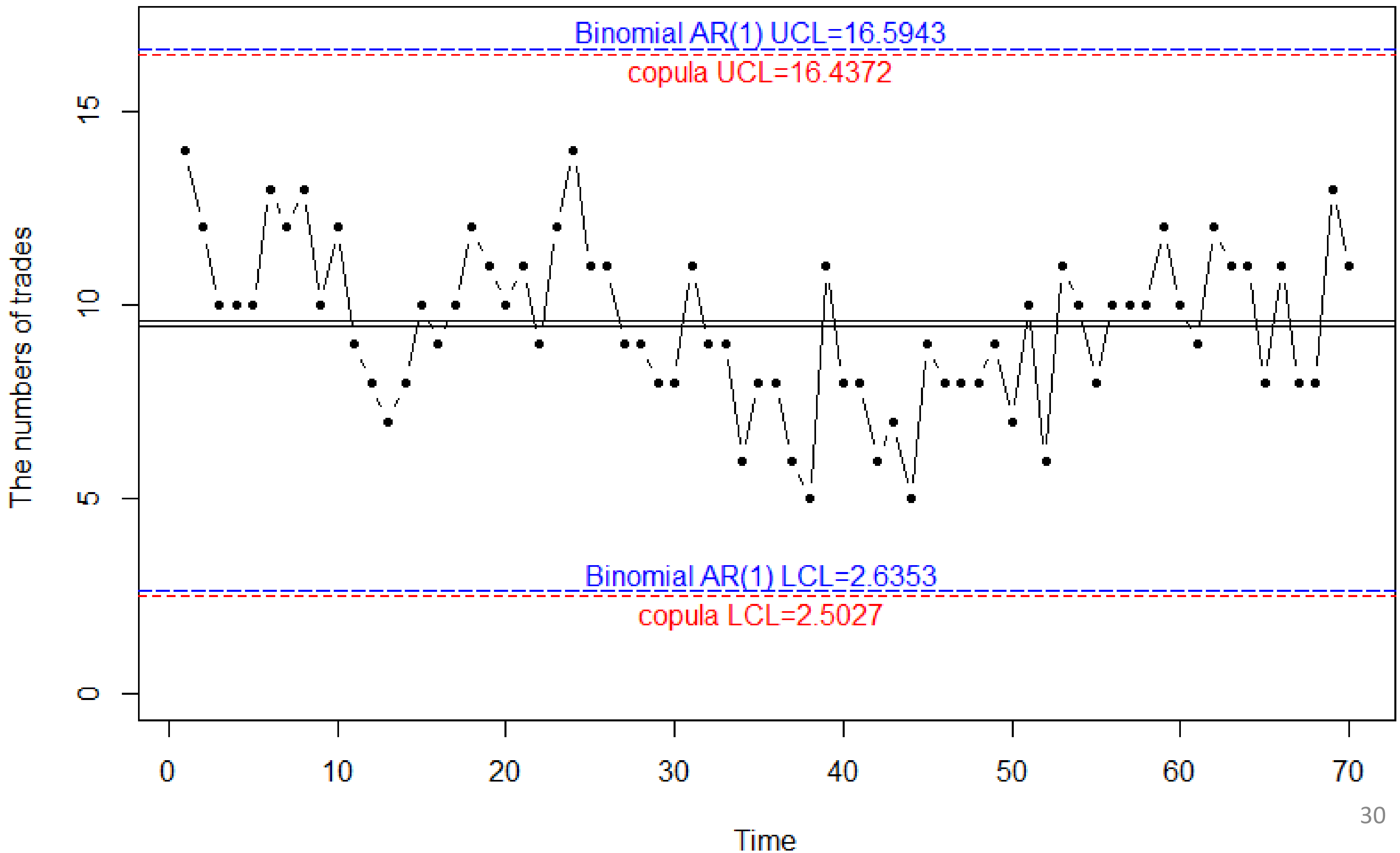
➔ No evidence against the binomial model

Fitted results

	Proposed	Chen & Fan	Standard	AR(1)
$\hat{p}$	0.4304	-	0.4331	0.4370
$\hat{\mu}$	9.4699	-	9.5285	9.6148
$\hat{\sigma}$	2.3224	-	2.3241	2.3265
$\hat{\alpha}$	0.7329	0.5604	0.9915	-
$\hat{\tau}$	0.2681	0.2188	0.3314	-
$\hat{\rho}$	-	-	-	0.5114
$SE(\hat{p})$	0.0185	-	-	0.0220
$SE(\hat{\alpha})$ or $SE(\hat{\rho})$	0.2646	-	-	0.0904

- All models show positive dependence.

MLEs for 3-σ control limits
➔ the stock market is in a stable condition



Summary

- **Proposed a copula-based Markov model for binomial data**

- MLE (Computation, Asymptotic)
- Control Limit (3-σ control limits)
- Copula selection (Clayton vs. Joe)
- R package (Data generation + MLE + Model diagnosis)
- Goodness-of-fit test (parametric bootstrap)

- **Future work:**

- More complex & general distribution
  (e.g., COM-Poisson distribution)
- Change Point model (ongoing with Lai Jay)
- Survival data (ongoing with Xinwei Huang)
  (event time series is censored by death)