Changepoint detection method with copula conditional distribution to Multivariate multistage case

Jong-Min Kim

Joint Work with Ning Wang and Yumin Liu, Zhengzhou University

SAMSI and University of Minnesota-Morris, MN, USA

December 14, 2018

The 3rd Pacific Rim Statistical Conference for Production Engineering, National Tsing Hua University

Hsinchu, Taiwan

Introduction.

- The Latin word **copula** means a linking verb, or anything that provides a link between two items.
- Copula is a new way of modeling the correlation structure between variables.
- The basic idea of copula is to separate the dependence and the marginal distributions in a multivariate distribution.

Brief historical background.

- 1940s: Hoeffding studied properties of multivariate distributions.
- The word copula appeared for the first time (Sklar).
- 1990s: Related books published Nelsen R.B. 1999 An introduction to copulas.
- Joe H. 1997, Multivariate models and dependence concepts 2000s: Applying copulas to various fields.

Why do we use copula?

- It is well known that multivariate financial data are typically nonlinear, non-normally distributed, and non-linear dependence.
- By using copula method, the assumptions like Gaussian distribution or linear dependence required for the correlation coefficient can be discarded.
- Copula functions are popular due to the following properties: copulas are invariant to strictly increasing transformations of the random variables.
- This property is particularly useful as transformations, such as log-transformation, are commonly used in financial analysis.

Copula Quality Control.

- Dokouhaki and Noorossana (2013), A Copula Markov CUSUM Chart for Monitoring the Bivariate Auto-correlated Binary Observations, *Quality and Reliability Engineering International*, 29(6), 911-919.
- Emura, T., Long, T.H. and Sun, L.H. (2017) R routines for performing estimation and statistical process control under copula-based time series models. *Communications in Statistics - Simulation and Computation* 2017; **46**(4): 3067-3087.
- Kim, J.-M, Baek, J. and Reller, M (2018), Control charts of mean and variance using copula Markov SPC and conditional distribution by copula. *Communications in Statistics - Simulation and Computation*, In Press.

Changepoints

- A changepoint is a time at which the structural patterns of a time series changes.
- Multiple changepoints are inherent in time series.
- Changpoint can affect means, variances, and extremes.
- Changepoint detection can be viewed as a model selection problem, hypothesis testing procedure, or classification analysis.
- More generally change detection also includes the detection of anomalous behavior: anomaly detection.

The Second Pacific Rim Statistical Conference for Production Engineering at SNU, Korea, 2016

- From Proceedings of the Pacific Rim Statistical Conference for Production Engineering Big Data, Production Engineering and Statistics (2017), we can find a meaningful and important papers related to this research as follows:
- Hsiang-Ling Hsu, Ching-Kang Ing, Tze Leung Lai and Shu-Hui Yu, Multistage Manufacturing Processes: Innovations in Statistical Modeling and Inference 67–84.
- Peihua Qiu and Lu You, Recent Research in Dynamic Screening System for Sequential Process Monitoring, 85–94.

Definition of Copula (1)

- Extraction of dependence structure of random variables from joint distribution.
- Joint distribution F_{XY} using a copula function C is defined by

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)) = C(u, v)$$

where F_X and F_Y are marginals.

For example,
1) Archimedean copula (Clayton, Frank) and Gaussian copula
2) Clayton copula: $C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$, for $\theta > 0$: Kendal's correlation$= \theta/(\theta + 2)$

Definition of Copula (2)

Definition

• A bivariate copula is a function $C : [0,1]^2 \to [0,1]$, with the following three properties:

(a) $C(u,0) = C(0,v) = 0, \quad \forall u, v \in [0,1]$

(b) $C(u,1) = C(1,u) = u, \quad \forall u \in [0,1]$

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

$\forall u_1, u_2, v_1, v_2 \in [0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$.

• Sklar (1973) showed that

joint F_{XY} can be represented as each marginal F_X and F_Y by using a two-dime'l copula $C(\cdot,\cdot)$.

Properties of Copula

More specifically, the copula may be written as

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)).$$

If F_X and F_Y are continuous, then copula, C, is unique and invariant under strictly monotone transformations of the random variables.
Let X and Y be continuously distributed random variables with copula C and let h_1 and h_2 be strictly increasing transformation functions, then, the copula of $h_1(x)$ and $h_2(y)$ is the same as copula of x and y, called the invariance property, (i.e., $C(h_1(x), h_2(y)) = C(x, y)$).

Sklar's Theorem

Sklar (1959): There exists a copula function C such that

$$F(x_1, x_2, \cdots, x_n) = C(F_1(x_1), F_2(x_2), \cdots, F_n(x_n))$$

where F_k is the k-th marginals, for $k = 1, 2, \cdots, n$.

Remarks: C need not be unique, but unique if the marginals are continuous.

For continuous marginals, the unique copula function is

$$C(u_1, \cdots, u_n) = F(F_1^{-1}(u_1), \cdots, F_n^{-1}(u_n))$$

where $F_1^{-1}, \cdots, F_n^{-1}$ denote the quantile functions of $F_1, \cdots, F_n$.

Normal Copula.

Denote distribution function of standard normal by:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{w^2}{2}\} dw.$$

Consider an n-variate normal random vector

$$Z = (Z_1, Z_2, \cdots, Z_n)$$

with Z_k is distributed as $N(0,1)$ for $k = 1, 2, \cdots, n$ and has positive definite, symmetric covariance matrix $\mathbf{V} = (v_{ij})$.

With elements

$$v_{ij} = \begin{cases} 1, & \text{if } i = j, \\ corr(Z_i, Z_j), & \text{otherwise.} \end{cases}$$

Normal Copula.

Joint density of Z

$$f(z_1, \cdots z_n) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp\{-\frac{1}{2} Z^T V^{-1} Z\} dw.$$

Joint distribution:

$$H(Z_1, \cdots, Z_n) = \int_{-\infty}^{z_n} \int_{-\infty}^{z_{n-1}} \cdots \int_{-\infty}^{z_1} f(x_1, x_2, \cdots, x_n) dx_1 \cdots dx_n.$$

The copula defined by

$$C(u_1, \cdots, u_n) = H(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_n))$$

is called the normal copula.

The multivariate Student-t distribution.

The n-dimensional random vector $X = (X_1, \cdots, X_n)$ is said to have a (non-singular) multivariate Student-t distribution with ν degrees of freedom, mean vector μ and positive-definite dispersion or scatter matrix Σ, denoted $X \sim t_n(\nu, \mu, \Sigma)$, if its density is given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)^n|\Sigma|}}\left(1 + \frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{\nu}\right)^{-\frac{\nu+n}{2}},$$

Note that in this standard parameterisation $cov(X) = \frac{\nu}{\nu-2}\Sigma$ so that the covariance matrix is not equal to Σ and is in fact only defined if $\nu > 2$.

Archimedean copula.

Let C be an associative, Archimedean copula. Then, there exists a strictly decreasing and convex (hence continuous) function (called the generator) $\varphi : [0,1] \to [0,+\infty)$ with $\varphi(1) = 0$ such that for every pair (u,v) in $[0,1] \times [0,1]$,

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

where $\varphi^{[-1]}$ is the "pseudo-inverse" of φ, given by

$$\varphi^{[-1]}(x) = \begin{cases} \varphi^{-1}(x), & \text{if } 0 \le x \le \varphi(0), \\ 0, & \text{if } \varphi(0) < x \le +\infty. \end{cases}$$

Archimedean copula functions.

Copula	Copula function
FGM	$C(u,v,\theta_1) = uv + \theta_1 uv(1-u)(1-v), \quad \theta_1 \in (-1,1]$
Clayton	$C(u,v,\theta_2) = \left(u^{-\theta_2} + v^{-\theta_2} - 1\right)^{-1/\theta_2}, \quad \theta_2 \in (0,\infty)$
Frank	$C(u,v,\theta_3) = -\frac{1}{\theta_3}\log\left[1 + \frac{(e^{-\theta_3 u}-1)(e^{-\theta_3 v}-1)}{e^{-\theta_3}-1}\right], \theta_3 \in \mathbb{R}\setminus\{0\}$
Gumbel	$C(u,v,\theta_4) = \exp\left[-\left((-\log u)^{\theta_4} + (-\log v)^{\theta_4}\right)^{1/\theta_4}\right], \theta_4 \geq 1$

Measurement of Dependence.

Bivariate measures of dependence for continuous variables are as follows:

- Spearman's rho:

$$\rho_C = 12 \int_0^1 \int_0^1 [C(u,v) - uv]\, dudv$$

- Kendall's tau:

$$\tau_C = 4 \int_0^1 \int_0^1 C(u,v) dC(u,v) - 1$$

Copula Conditional Function.

Corollary 1. Using Archimedean copulas introduced, we can derive the conditional distribution of X_1 given on X_2 for two random variables X_1 and X_2, $F_{1|2}(x_1|x_2;\theta_{12})$ is

$$F_{1|2}(x_1|x_2;\theta_{12}) = \frac{\partial C^C(u_1,u_2,\theta_{12})}{\partial u_2},$$

where $u_1 = F(x_1)$ and $u_2 = F(x_2)$.

NPMVCP by Holland and Hawkins (2014)

- A nonparmatric control chart using multivariate rank based test by Choi and Marden (1997) is proposed in Holland and Hawkins (2014). It gives the multivariate change point model (Hawkins et al. (2003)).
- It gives the multivariate change point model (Hawkins et al. (2003)) to identify changes in a sequence, $X_1, \ldots, X_t$ as follows

$$X_i \sim \begin{cases} F(\mu), & i \leq \tau \\ F(\mu + \sigma), & i > \tau \end{cases}$$

and $H_0 : \delta = 0$, versus $H_1 : \delta \neq 0$.

NPMVCP by Holland and Hawkins (2014)

- The test statistics and its asymptotic distribution are given for $k \in \{1, \ldots, t-1\}$ as:

$$\frac{tk}{t-k} \bar{r}_t^{(k)^T} \tilde{\sum}_{k,t}^{-1} \bar{r}_t^{(k)} \xrightarrow{d} \chi_d^2, \text{ if } t \to \infty,$$

where $\tilde{\sum}_{k,t}$ is the pooled sample covariance matrix of the centered rank vector $\bar{r}_t^{(k)}$ computed using kernel function.

NPMVCP by Holland and Hawkins (2014)

- Holland and Hawkins (2014) uses the test statistic

$$r_{k,t} = \bar{r}_t^{(k)^T} \hat{\sum}_{k,t}^{-1} \bar{r}_t^{(k)},$$

  where $\hat{\sum}_{k,t} = \frac{t-k}{tk} \hat{\sum}_t$ is the unpooled estimator of covariance matrix of centered ranks.
- Simulation study by Holland Hawkins (2014) shows that the power of using pooled or unpooled estimator of covariance matrix leads to similar performance. However for convenience of computation the unpooled covariance estimator $\hat{\sum}_{k,t}$ is employed.

Energy test based control chart (ETCC)

- Okhrin and Xu (2017) proposed a nonparametric change point model which can simultaneously monitor any change (mean, variance, dependence structure all together) in the multivariate distribution in online style.
- The control chart was integrated with the maximum energy divergence based permutation test to online detect the multiple change points for multivariate time series.
- The energy test, not similar to common rank test, uses discrepancy of empirical characteristic functions of two random vectors. And the empirical distribution of the test statistic can be obtained by permutation samples.

Energy test based control chart (ETCC)

- Then the sequential detection of change points can be conducted under the algorithm introduced by change point model (see Hawkins, Qiu and Kang (2003)) to form an online detection.
- For a change point detection problem, it is set that the change occurs at τ when the two random vectors $\{X_i \in R^d : X_i \sim F, i = 1, \ldots, \tau\}$ and $\{Y_j \in R^d : Y_j \sim G, j = \tau + 1, \ldots, T\}$ have distribution shift.
- In a multiple change point case, τ_i, $i = 1, 2, \ldots,$ the changes' detection can be formulated as

Energy test based control chart (ETCC)

- $$X_t \sim \begin{cases} F_0, & t \leq \tau_1 \\ F_1, & \tau_1 < t \leq \tau_2 \\ F_2, & \tau_2 < t \leq \tau_3 \\ \cdots & \\ F_j, & \tau_j < t \leq \tau_{j+1} \\ \cdots & \end{cases}$$

- Because the corresponding characteristic functions of X_i and Y_j, i.e. f_x and f_y, are uniquely determined, hence using the divergence between characteristic functions of the two random vectors to monitor the change is an applicable routine.

Energy test based control chart (ETCC)

- Székely and Rizzo (2005) used an integrated weighted distance between two characteristic functions, and showed that the larger the distance the more possible a change may occur between the two random vectors.
- Okhrin and Xu (2017) called a nonparametric change point model which is nonparametric control chart as energy test based control chart (ETCC).
- Based on ETCC, Xu (2017) made an R package 'EnergyOnlineCPM' which centers on the Phase II nonparametric change point model to online detect the multiple change points in context of high dimensional time series based on the maximum energy test statistic using permutation samples.

Empirical CDF

- From Definition 1, any bivariate distribution function of X and Y, $F_{X,Y}$, can be represented as a function of its marginals, F_X and F_Y, through a copula C.
- Any monotone strictly increasing transformation of X and Y will leave the joint dependence between them unchanged, and thus the copula associated with a pair (X, Y) will be the same as the copula with transformed variables (U, V).
- We can describe the dependence structure between X and Y by considering only the pairs of ranks of X and Y.

Empirical CDF

- Given n pairs of data points $\{(x_i, y_i); i = 1, \ldots, n\}$ from a pair of continuous random variables (X, Y), we can approximate the corresponding couple $(u_i, v_i) = (F_X(x_i), F_Y(y_i))$ using the pairs of ranks $\{(R_i, S_i); i = 1, \ldots, n\}$ where R_i is the rank of x_i among $x_1, \ldots, x_n$ and S_i is the rank of y_i among $y_1, \ldots, y_n$: $u_i = \frac{R_i}{n+1}$ and $v_i = \frac{S_i}{n+1}$.
- Using these normalized ranks, one can transform X and Y into U and V. Since the marginal distributions of U and V are uniform on $[0, 1]$ and thus parameter-free, this rank-based approach allows us to compute joint probabilities without knowing marginal distributions.

Proposed Method

The procedures to estimate parameters of the copula for the conditional cumulative distribution function can be summarized by:

- Step 1 Uses the empirical CDF to transform the observations to uniform distribution data in [0,1],
- Step 2 The parameters θ_{ij}, θ_{jk} of the joint CDF's $F(X_i, X_j)$ and $F(X_j, X_k)$ are estimated by the IFM method by Joe (1997),
- Step 3 The conditional CDFs $F(X_i|X_j; \hat{\theta}_{ij})$ and $F(X_k|X_j; \hat{\theta}_{jk})$ are computed with the estimates $\hat{\theta}_{ij}$ and $\hat{\theta}_{jk}$,
- Step 4 The parameter $\theta_{ik|j}$ of the CDF's $F(F(X_i|X_j), F(X_k|X_j))$ are estimated by the IFM method by Joe (1997),
- Step 5 Compute $u_{i|jk}$ with u_i, u_j and u_k by using Corollary 1.

Proposed Method

By using these procedures, we make the copula transformed conditional values which we plug into Energy test based control chart (ETCC) and NPMVCP by Holland and Hawkins (2014) for detecting change points.

Simulation Study

- For X data set, we simulate the highly correlated multivariate data by using 'copula' R package with 'normalCopula' function for the 5 variables with the correlation parameters $(0.9, 0.8, 0.8, 0.8, 0.7, 0.7, 0.7, 0.6, 0.5, 0.4)$ specifying the type of the symmetric positive definite matrix characterizing the elliptical copula is unstructured.
- For each marginal distribution, we set three gamma distributions (gamma(shape=5, scale=1), gamma (shape=5, scale=2), gamma (shape=5, scale=3)) and three exponential distributions (exp(rate=5), exp(rate =2).

Simulation Study

- For Y data set, we simulate the highly correlated multivariate data by using 'copula' R package with 'normalCopula' function for the 5 variables with the correlation parameters $(0.4, 0.5, 0.6, 0.7, 0.7, 0.7, 0.7, 0.6, 0.5, 0.4)$ specifying the type of the symmetric positive definite matrix characterizing the elliptical copula is unstructured.
- For each marginal distribution, we set three gamma distributions (gamma(shape=2, scale=1), gamma (shape=2, scale=2), gamma (shape=2, scale=3)) and three exponential distributions (exp(rate=2), exp(rate =5).

Simulation Study

- For Z data set, we simulate the highly correlated multivariate data by using 'copula' R package with 'normalCopula' function for the 5 variables with the correlation parameters $(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$ specifying the type of the symmetric positive definite matrix characterizing the elliptical copula is unstructured.
- For each marginal distribution, we set three gamma distributions (gamma(shape=3, scale=1), gamma (shape=3, scale=2), gamma (shape=3, scale=3)) and three exponential distributions (exp(rate=4), exp(rate =3).

Correlation matrix with the simulated data.

Jong-Min Kim
Joint Work with Ning Wang and Yumin Liu, Zhengzhou University

	X_1	X_2	X_3	X_4	X_5
X_1	1	0.9005	0.7873	0.7762	0.7778
X_2	0.9005	1	0.7081	0.6399	0.6988
X_3	0.7873	0.7081	1	0.5858	0.4561
X_4	0.7762	0.6399	0.5858	1	0.3780
X_5	0.7778	0.6988	0.4561	0.3780	1
	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	1	0.3501	0.4843	0.5644	0.6425
Y_2	0.3501	1	0.6626	0.6010	0.6890
Y_3	0.4843	0.6626	1	0.5559	0.4614
Y_4	0.5644	0.6010	0.5559	1	0.3409
Y_5	0.6425	0.6890	0.4614	0.3409	1
	Z_1	Z_2	Z_3	Z_4	Z_5
Z_1	1	0.5409	0.5325	0.5097	0.5043
Z_2	0.5409	1	0.5007	0.4843	0.4783
Z_3	0.5325	0.5007	1	0.4811	0.5467
Z_4	0.5097	0.4843	0.4811	1	0.4540
Z_5	0.5043	0.4783	0.5467	0.4540	1

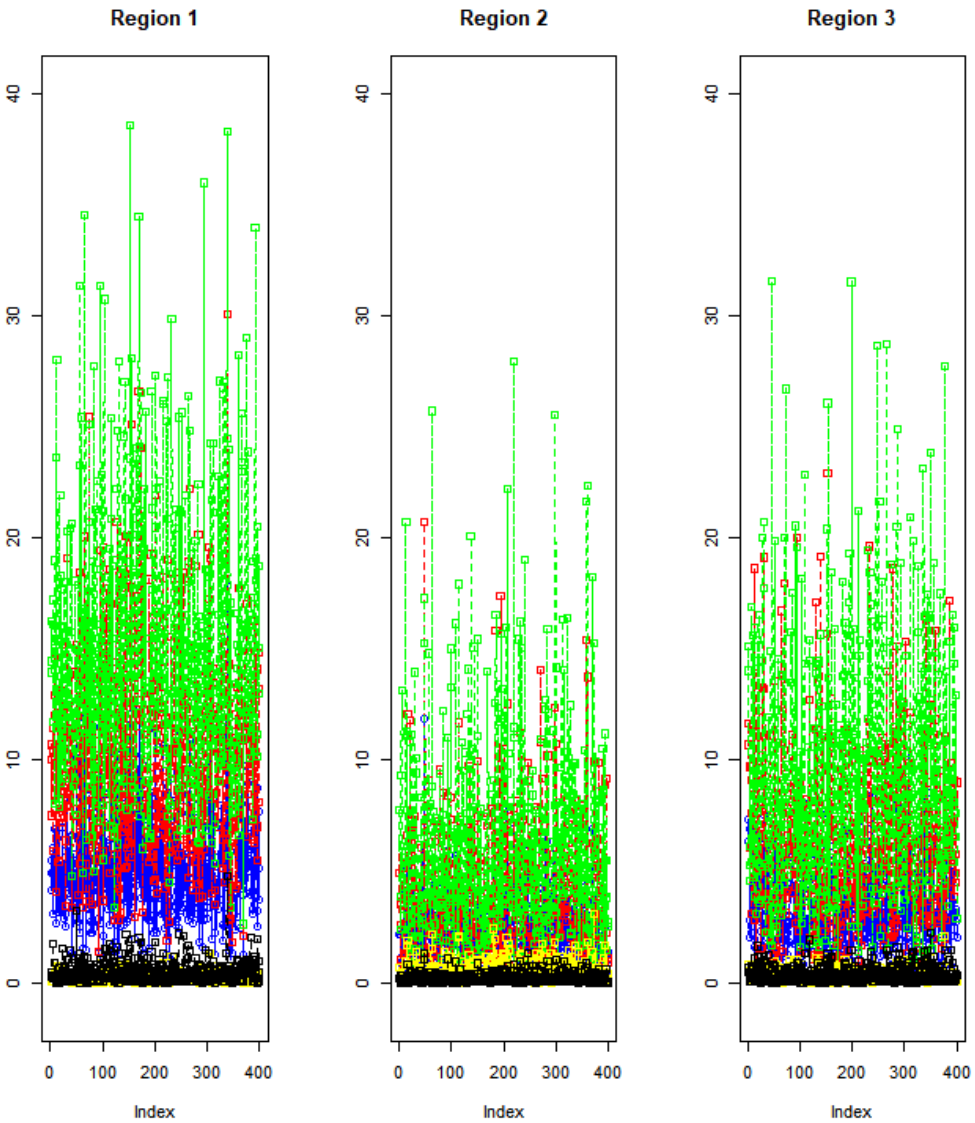
Change Point Detection.

Table: Change Point Detection with PCA with simulation data.

	$\mathbf{Y_1}$		$\mathbf{Y_2}$		$\mathbf{Y_3}$	
Method	ETCC	NPMVCP	ETCC	NPMVCP	ETCC	NPMVCP
X_1, X_2, X_3, X_4 , X_5, X_6	170 191	150 186	208 264 243 323 368 382	193 219 240 304 361 382	42 238	42 238
PCA Scores	174	174	206 218	193 218	238 296 321	238 294 321
	$\mathbf{Y_1}$		$\mathbf{Y_{2\|1}}$		$\mathbf{Y_{3\|21}}$	
Method	ETCC	NPMVCP	ETCC	NPMVCP	ETCC	NPMVCP
t-Copula	174	174	208 223	193 220	323	323
Gaussian Copula	174	174	208 223 371	193 220 371	48 238 322	42 238 321
Frank Copula	174	174	208 218 371	193 218 371	238 322	238 321

Plots with simulated data2.

Jong-Min Kim
Joint Work with Ning Wang and Yumin Liu, Zhengzhou University



Illustrated Study

Our data set contains daily foreign exchange rates for the twenty four most traded currencies (8 countries in Asia, 8 countries in Europe and 8 countries in America) against the euro from 3 January 2013 to 6 October 2014. The data set is obtained from the currency database retrieval system provided by Professor Werner Antweilers website at UBCs Sauder School of Business, . An increase in the nominal exchange rates considered reflects a depreciation of those currencies to the euro. Let S_t be an observed daily foreign exchange rate process in discrete time, $t = 1, \ldots, n$ and $r_t = \log(S_t/S_{t-1})$ is the rates of return of the exchange rates at time t.

Illustrated Study

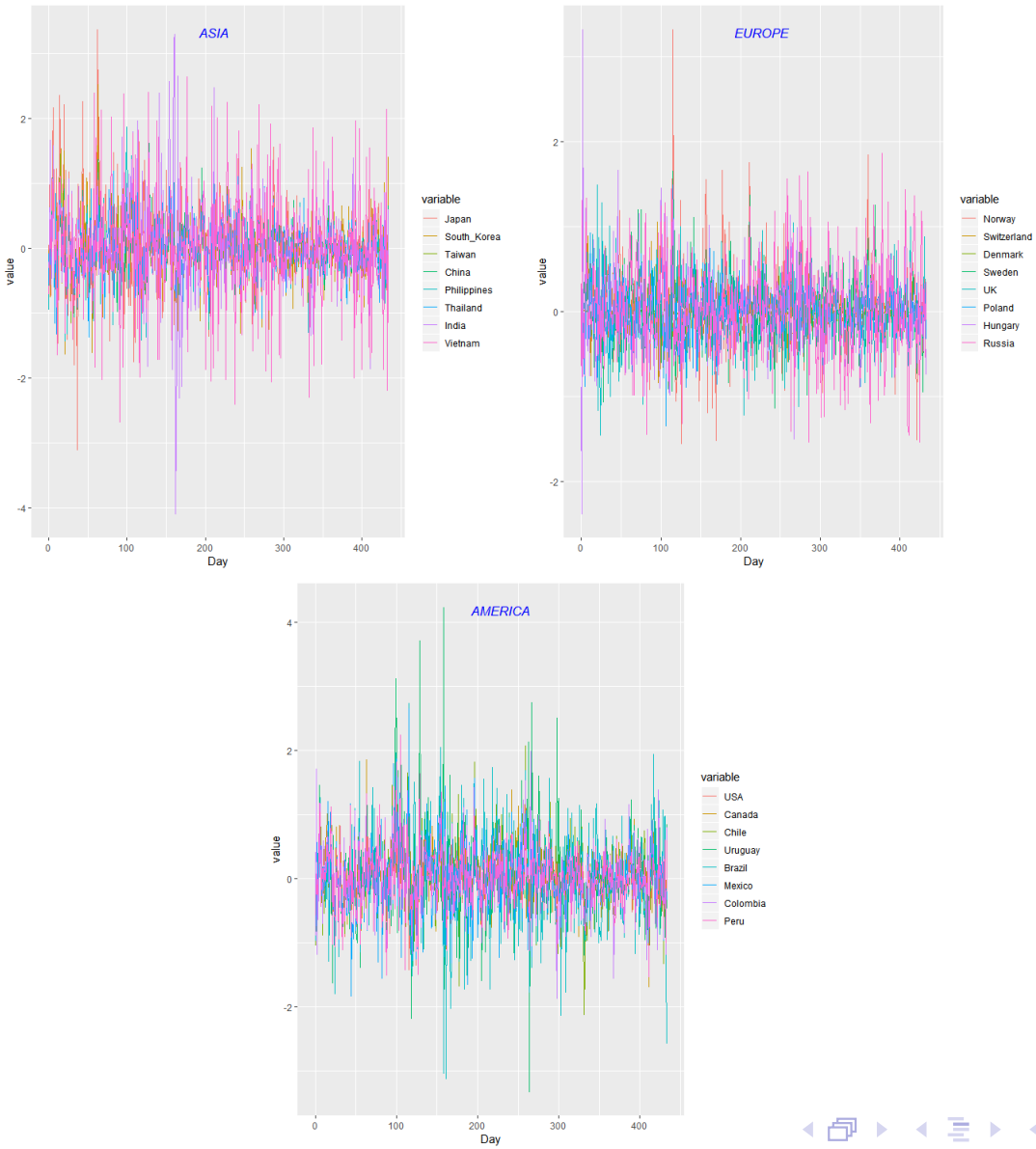
Table: 8 countries in Asia, 8 countries in Europe and 8 countries in America. Order by 2017 GDP.

Order	ASIA	EUROPE	AMERICA
1	JAPAN	NORWAY	USA
2	SOUTH KOREA	SWITZERLAND	CANADA
3	TAIWAN	DENMARK	CHILIE
4	CHINA	SWEDEN	URUGUAY
5	PHILIPPINES	UNITED KINGDOM	BRAZIL
6	THAILAND	POLAND	MEXICO
7	INDIA	HUNGARY	COLUMBIA
8	VIETNAM	RUSSIA	PERU

Twenty Four Countries in America, Europe, Asia.

Changepoint detection method with copula conditional distribution to Multivariate multistage case

Jong-Min Kim
Joint Work with Ning Wang and Yumin Liu, Zhengzhou University



Change Point Detection with Real Exchange Currency Data.

Table: Change Point Detection with Real Exchange Currency Data (1/3/2013 to 10/6/2014).

Whole Data	Asia		Europe		America	
Method	ETCC	NPMVCP	ETCC	NPMVCP	ETCC	NPMVCP
Whole Data	347	306	106 340	106 304	103 353	103 302
PCA Scores	344	175	109 341	106 304	103 127 342	103 126 302

Change Point Detection of Copula Conditional Distribution and PCA.

Jong-Min Kim
Joint Work with Ning Wang and Yumin Liu, Zhengzhou University

Table: Change Point Detection of Copula Conditional Distribution and PCA with Real Exchange Currency Data (1/3/2013 to 10/6/2014).

PCA Scores	Asia		Europe|Asia		America|(Europe,Asia)	
Method	ETCC	NPMVCP	ETCC	NPMVCP	ETCC	NPMVCP
t-Copula	382	307	0	0	103	103
Gaussian Copula	385	307	0	0	43 103 381	31 103 287
Frank Copula	382	307	0	0	42 103	31 103

Change Point Detection with Real Exchange Currency Data.

Table: Change Point Detection with Real Exchange Currency Data with America (1/3/2013 to 10/3/2014) and Asia (1/4/2013 to 10/6/2014).

Whole Data	America		Asia	
Method	ETCC	NPMVCP	ETCC	NPMVCP
Whole Data	103 127 346	103 126 302	344	306
PCA Scores	106 127 341	106 126 302	34 109 351	30 105 306

Change Point Detection of Copula Conditional Distribution and PCA.

Jong-Min Kim Joint Work with Ning Wang and Yumin Liu, Zhengzhou University

Table: Change Point Detection of Copula Conditional Distribution and PCA with Real Exchange Currency Data of America (1/3/2013 to 10/3/2014) and Asia (1/4/2013 to 10/6/2014).

PCA Scores	America		Asia|America	
Method	ETCC	NPMVCP	ETCC	NPMVCP
t-Copula	381	307	0	0
Gaussian Copula	383	302	378	306
Frank Copula	381	302	381	306

Conclusion

- We did multivariate multi-stage changepoint by copula conditional distribution and PCA.
- We will apply deep learning to multivariate multi-stage changepoint by copula conditional distribution.
- We will also apply functional data analysis to multivariate multi-stage changepoint by copula conditional distribution.

Acknowledgements

- Organizer committees for The 3rd Pacific Rim Statistical Conference for Production Engineering.
- National Tsing Hua University, Hsinchu, Taiwan.
- Professors and Students.
- All Great Today's Attendants.

Suggestions and Questions

Thank you for your attention today.

jongmink@morris.umn.edu

Jong-Min Kim