A Phase II Nonparametric Multivariate Multiple Change Points Model for High Dimensional Financial Time Series with Application in U.S. ETF Portfolio Surveillance

Yafei Xu*[1]

[2]*School of Business and Economics, Humboldt-Universität zu Berlin, Germany*

Abstract

In this article author constructed a nonparametric Phase II high dimensional multiple change points model for statistical process control based on energy statistic. Three highlights are included. Firstly, it is a nonparametric change point model which requires no pre-knowledge on the process. Secondly, it is oriented to Phase II detection which is central for financial stream surveillance. Thirdly the model is designed for high dimensional series. In simulation study the metrics for in-control and out-of-control average run length (IC-ARL, OC-ARL) of the model have been investigated in context of mean change, variance change in Gaussian and t, and asymmetric tail dependence change in Gaussian copula to R-Vine copulas with upper and lower tails separately. In empirical application we employed a high dimensional time series data set from U.S. ETF market to analyze the model's performance for portfolio surveillance especially in financial turmoil period. The results from simulation and empirical study showed that the model has great potential in sequential detection of multiple change points analysis for high dimensional time series. In this article author contributed an R package '`EnergyOnlineCPM`'.

Keywords: Phase II statistical process control; Change-point; Energy test; Vine copula; High dimensional time series; Nonparametric

*Corresponding author: `yafei.xu@hu-berlin.de`. The grant by the Chinese Government Scholarship through CSC-No.: 201508080128 is gratefully acknowledged.

1 Introduction

Change point model (CPM) plays a pivotal role in monitoring time series, e.g. in high dimensional portfolio surveillance. It is usually assumed that the series includes t random variables $\{X_t;\ t \in \mathbb{N}\},\ X_t \in \mathbb{R}^d$ which are identically independently distributed in $d \geq 1$ dimensions. In the series the number of change points $\{\tau_k;\ k \in \mathbb{N}\}$ is unknown which divide the series $\{X_t\}$ into diverse segments and each adjacent pair of segments follows different distributions, such that

$$X_i \sim \begin{cases} F_0, & t \leq \tau_1, \\ F_1, & \tau_1 < t \leq \tau_2, \\ \dots & \\ F_j, & \tau_j < t \leq \tau_{j+1}, \\ \dots . & \end{cases} \tag{1}$$

After decades' study, the change point model has arrived in the context of high dimension, i.e the multivariate change point model. Among them a part of researches are based on parametric assumptions, such as Crosier (1988) for multivariate CUSUM and Lowry, Woodall, Champ & Rigdon (1992) for multivariate EWMA and Zou & Tsung (2011), which assume multivariate Gaussian distribution. Qiu & Hawkins (2001), Qiu & Hawkins (2003), Hawkins & Deng (2010) have developed nonparametric change point models, however these models still need pre-knowledge in in-control distribution. Recent advance of Phase II nonparametric multivariate change point models can be found in Holland & Hawkins (2014) and Zhou, Zi, Geng & Li (2015).

Previous works focusing on energy divergence test theory and its applications in change point detection are briefly reviewed as follows. In Székely & Rizzo (2004), Zech & Aslan (2003), Székely & Rizzo (2013), the energy statistic and the related test and power analysis for distributional equality have been discussed. For application, Kim, Marzban, Percival & Stuetzle (2009) use the energy test in sliding window scheme with fixed window size to detect change points in image data. Matteson & James (2014) and James & Matteson (2015) employ energy test combined with two different clustering approaches in change point retrospective analysis or Phase I analysis.

In this paper authors developed an Phase II nonparametric change point model for detecting multiple change points in high dimensional time series. The essence of this model is to

integrate the maximum energy divergence based permutation test into Phase II context to sequentially detect the multiple change points. In simulation study, the model shows superior performance compared with a mainstream Phase II nonparametric change point models introduced in Holland & Hawkins (2014) in mean change, variance change and dependence structure change, which gives a promising application future for the model especially in monitoring evolution of high dimensional financial time series. For this purpose we have given an application of our model in surveillance of an ETF portfolio experiencing through financial quiescence and financial turmoil. For reproducible research and further study, an R package '`EnergyOnlineCPM`' is contributed.

The paper is structured as follows. In Section 2, the methodology is given including the preliminary of change point model in Phase I and Phase II, the energy divergence, the permutation test of maximum energy divergence. Simulation study is performed in Section 3 followed by an application of monitoring a high dimensional U.S. ETF portfolio. The last section concludes. Some supplemental materials about the copula function and the R package are affiliated in the appendix.

2 Methodology

In this section three parts are included. In Section 2.1, a brief review of the energy divergence and energy test is given. Section 2.2 introduce the Phase I change point model. The Phase II change point model is followed in Section 2.3, which is nonparametric for multiple change points detection in multivariate time series context.

2.1 Energy Divergence and Test

For a change point model used in high dimensional time series, the change happens at $\tau + 1$ when the two samples $X = \{X_i;\ i = 1, \ldots, \tau\} \overset{\mathcal{L}}{\sim} F$ and $Y = \{Y_j;\ j = \tau + 1, \ldots, T\} \overset{\mathcal{L}}{\sim} G$ have distribution shift in d dimension. Because the corresponding characteristic functions of the two samples, ϕ_x and ϕ_y are uniquely determined. Therefore using the divergence between characteristic functions of the two samples to monitor the change is an applicable routine. Székely & Rizzo (2005) uses an integrated weighted distance between two characteristic functions, hence the larger the distance the more possible a change may occur between the two samples.

Theorem 1. *Let* $X = \{X_i;\ i = 1, \ldots, \tau\} \overset{\mathcal{L}}{\sim} F$ *and* $Y = \{Y_j;\ j = \tau+1, \ldots, T\} \overset{\mathcal{L}}{\sim} G$ *be two samples in* d *dimension.* X', Y' *are copies of* X *and* Y. *The corresponding characteristic functions of the two samples are* ϕ_X *and* ϕ_Y. *If* $0 < \alpha < 2$ *with* $E||X||_2{}^{\alpha} < \infty$ *and* $E||Y||_2{}^{\alpha} < \infty$ *then*

$$\int_{\mathbb{R}^d} \frac{|\phi_X(\mathbf{t}) - \phi_Y(\mathbf{t})|^2}{||\mathbf{t}||_2^{d+\alpha}} d\mathbf{t} = C(d,\alpha)\mathcal{E}^{\alpha}(X,Y), \tag{2}$$

where

$$\begin{aligned}
\phi_X(\mathbf{t}) &= E\{\exp[i(X,\mathbf{t})]\} = E[\cos(X,\mathbf{t}) + i\,\sin(\mathrm{X},\mathbf{t})], \\
C(d,\alpha) &= \frac{2\Pi^{\frac{d}{2}}\Gamma(1-\frac{\alpha}{2})}{\alpha 2^{\alpha}\Gamma(\frac{\alpha+d}{2})}, \\
\Gamma(r) &= \int_0^{\infty} t^{r-1}e^{-t}dt,\ r \neq 0, -1, -2, \ldots, \\
\mathcal{E}^{\alpha}(X,Y) &= 2E||X-Y||_2^{\alpha} - E||X-X'||_2^{\alpha} - E||Y-Y'||_2^{\alpha}.
\end{aligned} \tag{3}$$

Proof. See Lemma 1 in Appendix of Székely & Rizzo (2005). □

Theorem 2. *Following set-up in Theorem 1,* $\mathcal{E}^{\alpha}(X,Y) = 0$ *iff* X *and* Y *are identically distributed.*

Proof. See Theorem 2 (ii) in Székely & Rizzo (2005). □

Therefore the metric $\mathcal{E}^{\alpha}(X,Y)$ can be used to measure the divergence between two distributions. The empirical conterpart of Equation (3) is derived as

$$\begin{aligned}
\hat{\mathcal{E}}^{\alpha}(X,Y) &= \frac{\tau(T-\tau)}{T}\Bigg(\frac{2}{\tau(T-\tau)}\sum_{i=1}^{\tau}\sum_{j=1}^{T-\tau}||x_i - y_j||_2 \\
&\quad -\frac{1}{\tau^2}\sum_{i=1}^{\tau}\sum_{j=1}^{\tau}||x_i - x_j||_2 - \frac{1}{(T-\tau)^2}\sum_{i=1}^{T-\tau}\sum_{j=1}^{T-\tau}||y_i - y_j||_2\Bigg),
\end{aligned} \tag{4}$$

where $\{x_i,\ i = 1, 2, \ldots, \tau\}, \{y_j,\ j = 1, 2, \ldots, T-\tau\}$ are observations of X and Y. Since X and Y are i.i.d. samples, and Equation (3) can be used as the metric of divergence between two unknown distributions, therefore Equation (4) can be employed as the test statistic, where the empirical distribution of $\hat{\mathcal{E}}^{\alpha}$ can be obtained by permutation samples.

2.2 Phase I Change Point Model

In statistical process control there are mainly two types of detection termed as Phase I and Phase II detection. Let $x = \{x_1, ..., x_T\}$ denote a sample of observations with length of T. In Phase I detection, the sample and its size T are fixed, i.e. no new in-comer. The detection is performed only based on data x as historical data. Hence this type change point analysis is retrospective and static. Phase I analysis has broad applications in bio-statistics and transportation statistics, see Székely & Rizzo (2005) and Matteson & James (2014).

Assume there is a change occurs at k, then the Equation 1 can be represented in the following test hypotheses,

$$
\begin{aligned}
H_0: \quad & X_t \overset{\mathcal{L}}{\sim} F_0(x|\theta),\ 1 \le t \le T, \\
H_1: \quad & X_t \overset{\mathcal{L}}{\sim} \begin{cases} F_0(x|\theta), & 1 \le t \le k, \\ F_1(x|\theta), & k < t \le T. \end{cases}
\end{aligned}
$$

A two-sample parametric or nonparametric test can be applied in this case under test statistic denoted with $Z_{k,T}$. If $Z_{k,T}$ is larger than a threshold $h_{k,T}$, i.e. $Z_{k,T} > h_{k,T}$, then the null hypothesis is rejected, that is the two samples are not identically distributed. Then a change is admitted at x_k. Since the change point location is unkown, hence the two-sample test will be performed at every point $1 < k < T$. The test statistic is changed to

$$Z_T = \max_{1<k<T} Z_{k,T}.$$

The null hypothesis is rejected if $Z_T > h_T$. The Type I error α in this context means that the model signals a change point when actually there is no change occurs. The distribution of the test statistic Z_T can be obtained either by its asymptotic distribution (if available) or by simulation methods. At the end, the change location can be estimated by

$$\hat{\tau} = \arg\max_{1<k<T} Z_{k,T}.$$

2.3 Phase II Change Point Model

Different from Phase I detection, Phase II detection considers the sample x with not fixed T size but an increasing size, i.e. T is increased with time proceeding. Phase II detection

is also termed as online detection and sequential detection. For instance, the stock price is updated with time proceeding, therefore the length of time series x is always increased. Hence the detection in Phase II concentrates on dynamic stream data not static data.

With the Phase I analysis in Section 2.2, Phase II can be extended from the Phase I with increasing sample size. That is whenever a new observation arrives then a new sample $x = \{x_1, \ldots, x_T, x_{T+1}, \ldots, x_t\}$ is constructed with size t. For every new observation the Phase I process will be performed. Hence the null hypothesis is rejected if $Z_t > h_t$. The Type I error α can be represented with

$$\begin{aligned} Pr(Z_1 > h_1) &= \alpha,\ t = 1, \\ Pr(Z_t > h_t | Z_{t-1} \leq h_{t-1}, \ldots, Z_1 \leq h_1) &= \alpha,\ t > 1. \end{aligned} \tag{5}$$

In statistical process control, the in-control average run length (IC-ARL) is the inverse of the Type I error, i.e. $1/\alpha$, which stands for the average run length of the model until the first erroneous alarm signals.

3 Simulation Study

In statistical process control, the assessment of change point model uses mainly two measures, the in-control average run length (IC-ARL) and the out-of-control average run length (OC-ARL). IC-ARL assumes the process follows a common distribution without change in order to calculate the length until the first error signal flags, therefore the larger the IC-ARL the better the model. OC-ARL assumes the process has a change point in order to compute the length until the model detects the change. Hence the model is expected to have a samll OC-ARL.

In this simulation study, the proposed model, P2MECPM (Phase II Maximum Energy Change Point Model), is tested in three scenarios including mean change, variance change and asymmetric dependence structure change. And the benchmark model for comparison is NPMVCP model introduced in Holland & Hawkins (2014). For fair comparability here the quanrantine optimization is not considered for both models. The warm-up is set as 32 consistent to the default set-up in R package `NPMVCP` and the significant level as 0.005 or IC-ARL= 200 in mean change and variance change categories. For asymmetric tail change it is set the significant level as 0.01. In mean change and variance change it is considered to implement the model in five and ten dimensions but in tail change part we

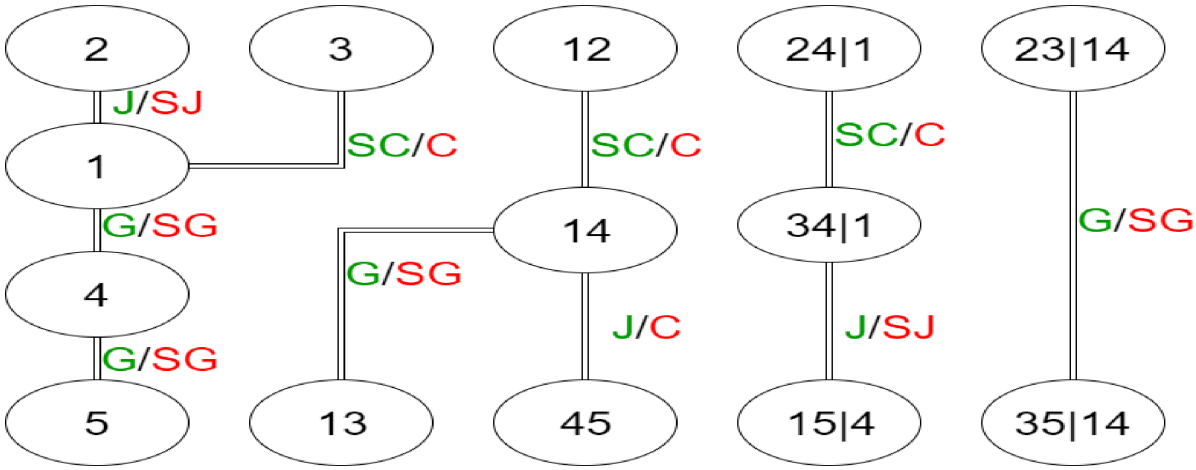

Figure 1: The structure of a five dimensional regular vine copula. The employed bivariate copula family for upper tail dependence structure is in green, lower tail dependence in red. Abbreviation: J: Joe, G: Gumbel, C: Clayton, S: Survival.

only compare the model in five dimensions. Since all measures are computed based on the i.i.d. sample, hence under the law of large numbers the mean will converge. Therefore all simulations in this part are performed based on 100 runs.

In mean change scenario, the standard normal distribution and student-t distribution with five degrees of freedom are employed. The range of mean change is set in four categories, i.e. $\mu_1 = \mu_0 + C,\ C = 0, 1, 3, 10$. If $C = 0$, then the model is tested for erroneous signal ratio (ESR), i.e. proportion of wrong detections through a homogeneous distributional sample. Here the ESR is tested in a length of 82 points series, if the model does flag in the next 50 steps then the error signal will be counted. In variance change part, the DGP will change from a standard Gaussian process to a more volatile process with σ^2 changing from 0 to 5 and 10.

Since copula is a practical tool in constructing flexible distributions, hence in asymmetric tail dependence change category, two different regular vine copulas, R-Vine-U and R-Vine-L, are used. A short introduction of copula function and pair copula constructions is given in appendix. The R-Vine-U copula has an upper tail dependence and the R-Vine-L the lower tail dependence. The structures of the two regular vine copulas used in this paper are illustrated in Figure 1, where the bivariate copula family for R-Vine-U is colored in green and R-Vine-L in red. The star notation denotes the better model in every category in Table 3. Every record in tail shift category is computed in a length of 82 points series.

The results for simulation study are given in Table 3. Some conclusions are as follows.

Firstly, in mean change category, P2MECPM performs better in OC-ARL in five dimensions context in both Gaussian and Student-t cases. In ten dimension category, NPMVCP has advantage in OC-ARL but P2MECPM is far better in IC-ARL.

Secondly, P2MECPM outperforms NPMVCP in Gaussian variance change no matter in low dimension or high dimension or in small change or large change. This has special meaning in monitoring time series in volatile circumstances, e.g. in financial crisis.

Thirdly, if the time series has change from zero tail dependence like Gaussian process to lower or upper tail dependence, then P2MECPM is much faster to alarm the change. Generally the market is tend to have heavy lower tail dependence structure in financial turmoil period, see Hu (2006), hence the sensitivity to lower tail dependence change has vital practical meaning. In parentheses it presents the result based on 50 warm-up setting in the P2MECPM, which shows a high improvement of the model in tail change detection.

4 Empirical Study

In this section the P2MECPM is applied to monitor a U.S. ETF portfolio. In this portfolio five assets are included with tickers DGT, EWD, GLD, IGV and IUSG. The data is obtained from the Wall Street Journal website. The length of close prices of each ETF is 2517 from 20060104 to 20151231, 10 years' data set.

Before we implement the change point model to this data set, we need to filter this five dimensional time series into independent series. In this study the GJR-GARCH$(1,1)$, see Glosten, Jagannathan & Runkle (1993), is employed to handle the data, such that,

$$\begin{aligned} y_t &= \mu + \phi(L)y_t + \gamma(L)\epsilon_t, \\ \epsilon_t &= \sigma_t z_t,\ z_t \overset{\mathcal{L}}{\sim} \mathcal{D}(0,1), \\ \sigma_t^2 &= K + \delta\sigma_{t-1}^2 + \alpha\epsilon_{t-1}^2 + \phi\epsilon_{t-1}^2 I_{t-1}, \end{aligned}$$

where $I_{t-1} = 0$ if $\epsilon_{t-1} \leq 0$, and $I_{t-1} = 1$ if $\epsilon_{t-1} < 0$.

After filtration the independent residuals are obtained. The correlations of Pearson, Kendall and Spearman of the residuals are shown in Table 2. It is clear that the GLD has almost no correlation with the other four ETFs, while the other four are strong correlated. The correlations of the five ETFs can also be observed from Figure 2 that DGT, EWD, IGV, IUSG have pairwise positive correlation.

			P2MECPM	NPMVCP
			detection location	detection location
Mean Shift	Gaussian	μ		
	5 dim.	0	0.04	0.02*
		1	10	11
		3	7	8
		10	6	8
	10 dim.	0	0.03	0.29
		1	9	6*
		3	6	6
		10	6	6
	Student-t_5	μ		
	5 dim	0	0.04	0.02*
		1	10	11
		3	7	9
		10	6	8
	10 dim	0	0.03	0.35
		1	9	6*
		3	6	5*
		10	6	5*
Variance Shift	Gaussian	σ^2		
	5 dim.	5	14	19
		10	10	19
	10 dim.	5	11	11
		10	8	13
Tail Shift	Gaussian to R-Vine-U	θ		
	5 dim.	5	40 (22)	40
		10	40 (20)	40
		30	36 (20)	39
	Gaussian to R-Vine-L	θ		
	5 dim.	5	39 (21)	40
		10	37 (20)	40
		30	33 (19)	38

Table 1: Simulation study in mean shift, variance shift and tail dependence shift. The ESR for Gaussian and Student-t with 5 degree of freedom are given in mean shift category, otherwise out-of-control ARLs are reported.

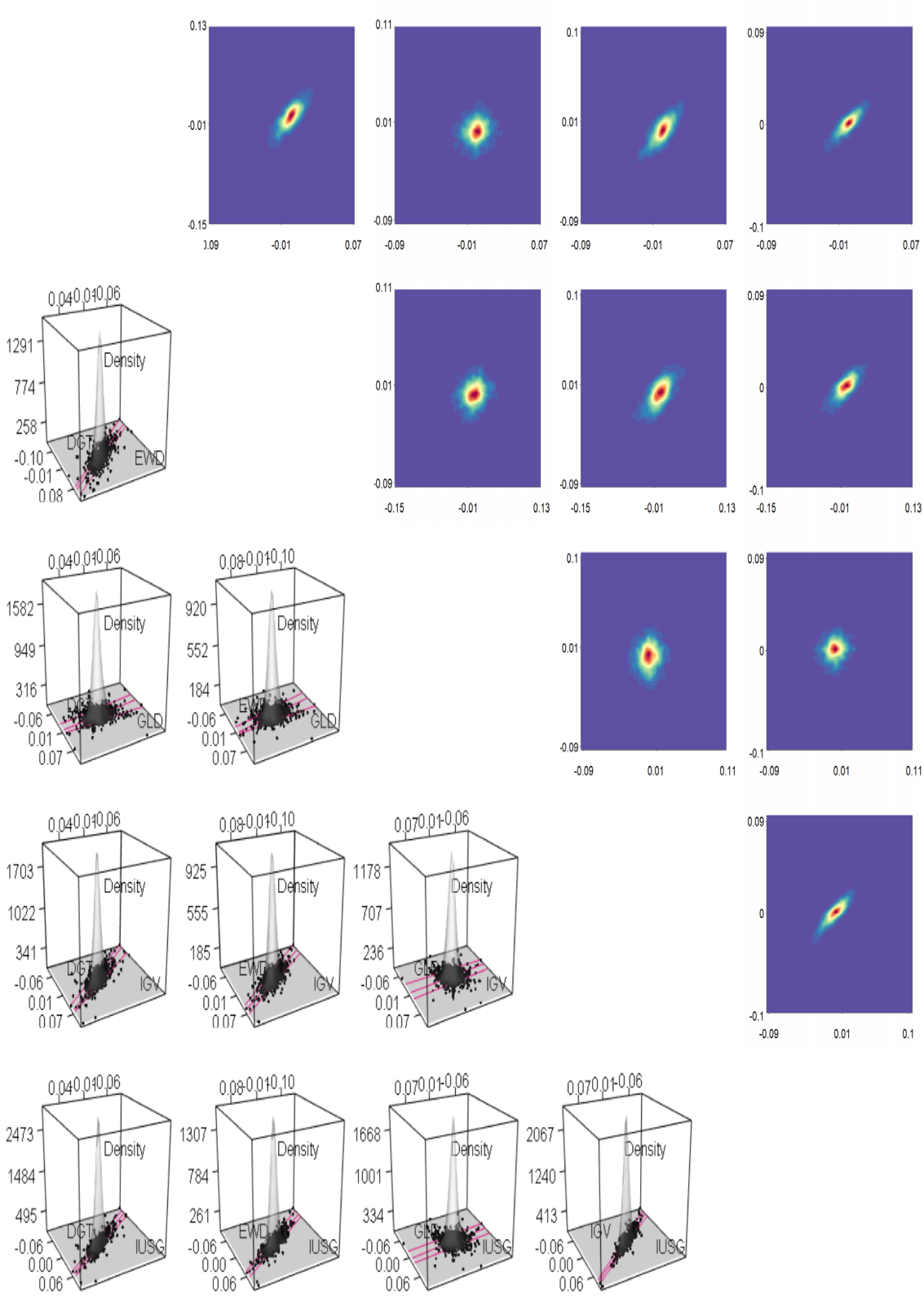

Figure 2: The lower triangular plots give 2-dimensional kernel density estimations containing scatter plots of pairwise GJR-GARCH(1, 1)-filtered log-returns with quantile regressions under 0.10, 0.5, 0.90 quantiles. The upper triangular plots give pairwise contours of residuals of five ETFs.

	DGT	EWD	GLD	IGV	DGT	EWD	GLD	IGV	DGT	EWD	GLD	IGV
EWD	0.802				0.567				0.745			
GLD	0.092	0.161			0.095	0.138			0.136	0.198		
IGV	0.768	0.739	0.037		0.516	0.461	0.038		0.687	0.631	0.055	
IUSG	0.847	0.820	0.075	0.905	0.611	0.544	0.069	0.686	0.786	0.727	0.099	0.860

Table 2: Pairwise dependence measures include Pearson's correlation (left), Kendall's correlation (middle) and Spearman's correlation (right).

The P2MECPM is set in this empirical study with 32 warm-ups, 0.005 significant level. At last 11 change points are signaled. Figure 3 illustrates the close price series and the residual series. The market benchmark S&P500 index is shown with its close and residual series in the bottom of the Figure 3. It can be observed that the DGT, EWD, IGV and IUSG are co-moved, especially IGV and IUSG have similar trajectories with the S&P500 index. It is clear that the P2MECPM has caught the main volatile periods, e.g. the financial crisis in 2008/09. The lowest point of S&P500 happened on 20090302, but the P2MECPM has detected the change on 20080929. An even earlier alarm was given at the beginning of 20070821, which denotes the preceeding of the sub-prime mortgage crisis. This means the model can give alarm when the market has evolute to the turmoil and risky side, which is crucial for financial institutions to make trading decisions.

5 Conclusion

In this paper, authors has proposed to use a nonparametric Phase II change point model, P2MECPM, to detect the multiple change points in high dimensional time series, whose main advantage lies in Phase II setting and powerful nonparametric permutation based maximum energy test. In simulation study the mean, variance and tail change have been investigated and shown superior performance compared to a benchmark model. In empirical application, the model is implemented for surveillance of an ETF portfolio which contains five ETF assets. The model show the ability to detect the change of the market from quiescent period to volatile period which provides reference to financial institutions to prevent from market risk. .

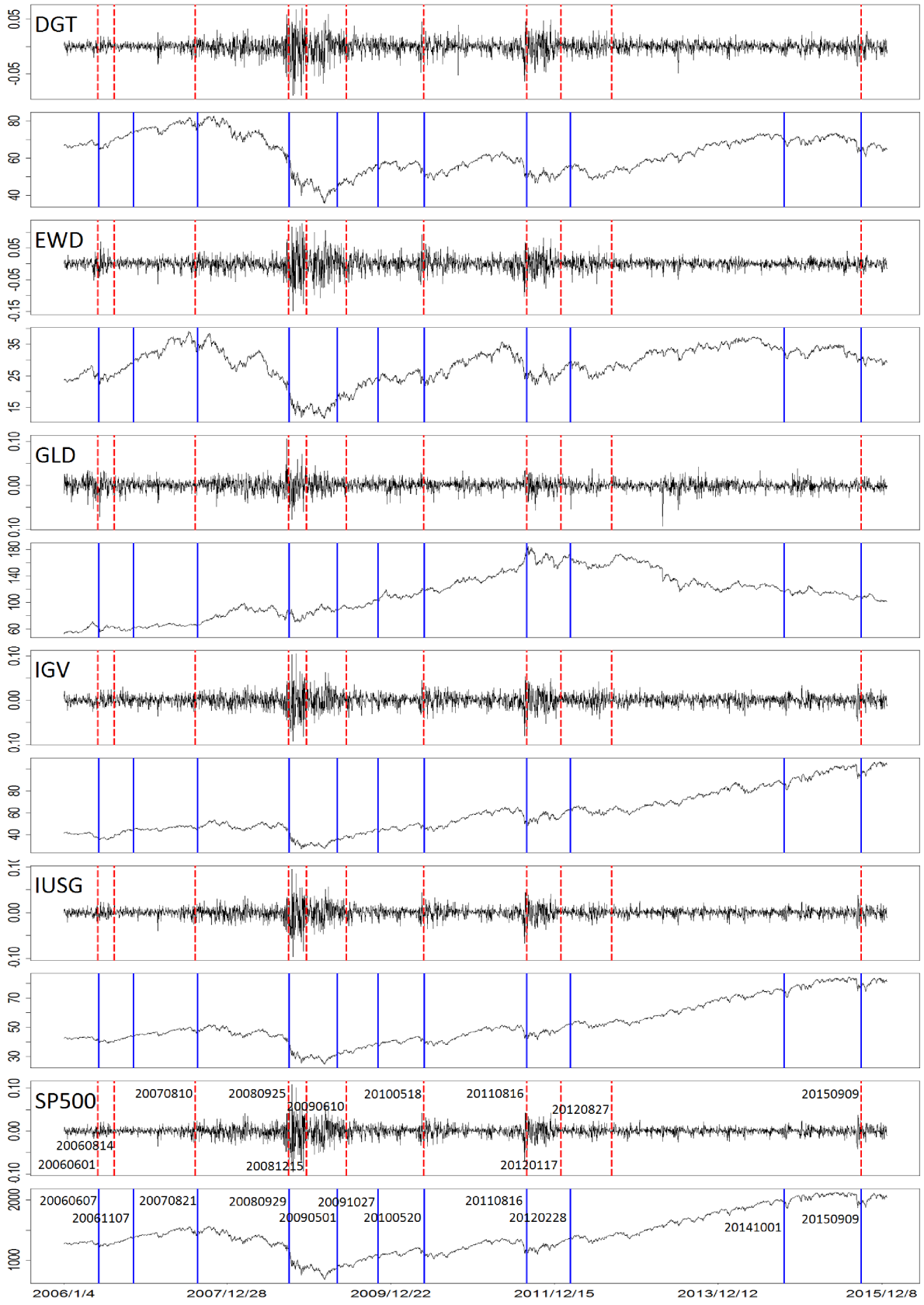

Figure 3: Change time points (red) and detection time points (blue) are denoted for five ETFs and S&P500 index. The change time points are illustrated on residuals plots and the detection time points on close price plots. The time window is from 20060104 to 20151231, totally 2516 trading days.

References

Aas, K., Czado, C., Frigessi, A. & Bakken, H. (2009). Pair-copula constructions of multiple dependence, *Insurance: Mathematics and Economics* **44**(2): 182–198.

Bedford, T. & Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines, *Annals of Mathematical and Artificial Intelligence* **32**: 245–268.

Bedford, T. & Cooke, R. M. (2002). Vines – a new graphical model for dependent random variables, *Annals of Statistics* **30**(4): 1031–1068.

Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes, *Technometrics* **30**(3): 291–303.

Glosten, L. R., Jagannathan, R. & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks, *The journal of finance* **48**(5): 1779–1801.

Hawkins, D. M. & Deng, Q. (2010). A nonparametric change-point control chart, *Journal of Quality Technology* pp. 165–173.

Holland, M. & Hawkins, D. (2014). A control chart based on a nonparametric multivariate change-point model, *Journal of Quality Technology* **46**: 1975â1987.

Hu, L. (2006). Dependence patterns across financial markets: A mixed copula approach, *Applied Financial Economics* **16**: 717–729.

James, N. & Matteson, D. (2015). ecp: An r package for nonparametric multiple change point analysis of multivariate data, *Journal of Statistical Software* **62**(1): 1–25.

Joe, H. (1996). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters, *in* L. Rüschendorf, B. Schweizer & M. Taylor (eds), *Distribution with fixed marginals and related topics*, IMS Lecture Notes – Monograph Series, Institute of Mathematical Statistics.

Joe, H. (2014a). *Dependence Modeling with Copulas*, Chapman and Hall/CRC, Boca Raton.

Joe, H. (2014b). *Dependence Modeling with Copulas*, Chapman and Hall/CRC, Boca Raton.

Kim, A. Y., Marzban, C., Percival, D. B. & Stuetzle, W. (2009). Using labeled data to evaluate change detectors in a multivariate streaming environment, *Signal Processing* **89**: 2529â2536.

Kurowicka, M. & Cooke, R. M. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*, John Wiley & Sons, New York.

Lowry, C. A., Woodall, W. H., Champ, C. W. & Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart, *Technometrics* **34**(1): 46–53.

Matteson, D. S. & James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data, *Journal of the American Statistical Association* **109**(505): 334–345.

Nelsen, R. (2006). *An Introduction to Copulas*, Springer- New York.

Qiu, P. & Hawkins, D. (2001). A rank-based multivariate cusum procedure, *Technometrics* **43**(2): 120–132.

Qiu, P. & Hawkins, D. (2003). A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions, *Journal of the Royal Statistical Society: Series D (The Statistician)* **52**(2): 151–164.

Rizzo, M. L. & Szekely, G. J. (2016). Package 'energy', *R User's Manual* .

Sklar, A. (1959). Fonctions dé repartition á n dimension et leurs marges, *Publ. Inst. Stat. Univ. Paris* **8**: 299–231.

Székely, G. J. & Rizzo, M. L. (2004). Testing for equal distributions in high dimension, *InterStat* **5**.

Székely, G. J. & Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method, *Journal of classification* **22**(2): 151–183.

Székely, G. J. & Rizzo, M. L. (2013). Energy statistics: statistics based on distances, *Signal Processing* **143**: 1249–1272.

Xu, Y. (2017). Reference manual: An r package 'EnergyOnlineCPM'.
URL: *https://sites.google.com/site/EnergyOnlineCPM/*

Zech, G. & Aslan, B. (2003). A multivariate two-sample test based on the concept of minimum energy, *Proc. Statistical Problems in Particle Physics, Astrophysics, and Cosmology* pp. 8–11.

Zhou, M., Zi, X., Geng, W. & Li, Z. (2015). A distribution-free multivariate change-point model for statistical process control, *Communications in Statistics: Simulation and Computation* **44**: 1975–1987.

Zou, C. & Tsung, F. (2011). A multivariate sign ewma control chart, *Technometrics* **53**: 84–97.

A Copula Fundamental

A.1 d-Dimensional Copula

A d-dimensional copula is a distribution function on $[0,1]^d$ having all marginal distributions uniform on $[0,1]$. Sklar's Theorem, c.f. Sklar (1959), introduces the relation between joint distribution function, copula function and marginal functions.

Theorem 3 (Sklar (1959))**.** *Let F be a multivariate distribution function with margins $F_1, \ldots, F_d$, then there exists the copula C such that*

$$F(x_1, \ldots, x_d) = C\{F_1(x_1), \ldots, F_k(x_d)\}, \quad x_1, \ldots, x_d \in \overline{\mathbb{R}}.$$

If F_i are continuous for $i = 1, \ldots, d$ then C is unique. Otherwise C is uniquely determined on $F_1(\overline{\mathbb{R}}) \times \cdots \times F_d(\overline{\mathbb{R}})$.

Conversely, if C is a copula and $F_1, \ldots, F_d$ are univariate distribution functions, then function F defined above is a multivariate distribution function with margins $F_1, \ldots, F_d$.

The representation in Sklar's Theorem can be used for constructing new multivariate distributions by changing either the copula function or marginal distributions. For detail of copula theory, it can be referred to monographs Nelsen (2006) and Joe (2014a).

A.2 Gaussian Copula

The elliptical copula used in this work is Gaussian copula given by,

$$C_{gs}(u_1, \ldots, u_d; G) = \Phi_d \left\{ \Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d); G \right\}, \; u_k \in [0,1], \; k = 1, \ldots, d, \qquad (6)$$

Archimedean Copula	Representation $C(u_1, \ldots, u_d; \theta)$	Generator Function $\varphi(t; \theta)$	Parameter θ
Clayton	$\left(\sum_{k=1}^{d} u_k^{-\theta} - d + 1\right)^{-\theta^{-1}}$	$\frac{1}{\theta}(t^{-\theta} - 1)$	$[-1/(d-1), \infty) \backslash \{0\}$
Gumbel	$\exp\left\{-\sum_{k=1}^{d}(-\log u_k)^{\theta}\right\}^{\theta^{-1}}$	$\{-\log(t)\}^{\theta}$	$[1, +\infty)$
Joe	$1 - \left\{\sum_{k=1}^{d}(1-uk)^{\theta} - \prod_{k=1}^{d}(1-u_k)^{\theta}\right\}^{\frac{1}{\theta}}$	$-\log\left\{1-(1-t)^{\theta}\right\}$	$[1, +\infty)$

Table 3: Structures of common Archimedean copulas.

where G is a $(d \times d)$ correlation matrix, Φ_d is a d-dimensional standard Gaussian CDF and Φ is a one dimensional standard Gaussian CDF. Gaussian copula is symmetric with zero tail dependence.

A.3 Archimedean Copula

Another important family of copula employed in this paper is the Archimedean copula family, whose elements can be constructed as

$$C_A(u_1, \ldots, u_d; \theta) = \begin{cases} \varphi^{-1}\left\{\varphi(u_1; \theta) + \cdots + \varphi(u_d; \theta); \theta\right\} & \text{if } \sum_{k=1}^{d} \varphi(u_k; \theta) \leq \varphi(0; \theta), \\ 0 & \text{else,} \end{cases}$$

where a decreasing function $\varphi \colon [0, 1] \to [0, +\infty)$ is the generator function with $\varphi(1) = 0$ and $\varphi(+\infty) = 1$. Here three most well-known Archimedean copulas are considered, i.e. Clayton, Gumbel and Joe. Table 3 lists the representations, generator functions and parameter spaces of these four common Archimedean copulas.

Clayton copula has lower tail dependence but no upper tail dependence and this is important for modeling losses. Gumbel copula is the only extreme value copula, and often used in modeling gains. Joe copula, having upper tail dependence. Therefore those copulas with such a specific attribute are used in our study.

A.4 Vine Copula

Vine copula or pair-copula constructions are originally proposed in Joe (1996) and developed in depth by Bedford & Cooke (2001), Bedford & Cooke (2002), Kurowicka & Cooke (2006) and Aas, Czado, Frigessi & Bakken (2009). The catchy name is due to similarities of the graphical representation of vine copulae and botanical vines. The fundamental idea of the vine copula is to construct a d-dimensional copula by decomposing the dependence structure into $d(d-1)/2$ bivariate copulas.

Let S be the index subset of $D = \{1, \ldots, d\}$ referring to the index set of conditioning variables and T be the index set of conditioned variables with $T \cup S = D$. Let $\sharp M$ denote the cardinality of set M. The cdf of variables with index in S is denoted by F_S, so that $F(x) = F_D(x)$. The conditional cdf of variables with index in T conditional on S is denoted $F_{T|S}$. A similar notation is used for the corresponding copulas. To derive a vine copula for a given $x = (x_1, \ldots, x_d)^\top$ in the spirit of Joe (2014b), we start from a d-dimensional distribution function, i.e.

$$F(x) = \int\limits_{(-\infty, x_S]} F_{T|S}(x_T|y_S) dF_S(y_S), \tag{7}$$

and replace the conditional distribution $F_{T|S}(x_T|x_S)$ by the corresponding $\sharp T$-dimensional copula $F_{T|S}(x_T|x_S) = C_{T;S}\{F_{j|S}(x_j|x_S) : j \in T\}$. The copula $C_{T;S}\{F_{j|S}(x_j|x_S) : j \in T\}$ is implied by Sklar's Theorem with margins $F_{j|S}(x_j|x_S),\ j \in T$. It is not a conditional distribution although with conditional distribution as margins. This yields a copula-based representation of the joint d-dimensional distribution function from (7), which is given by

$$F(x) = \int\limits_{(-\infty, x_S]} C_{T;S}\{F_{j|S}(x_j|y_S) : j \in T\} dF_S(y_S). \tag{8}$$

Note that the support of the integral in (7) and (8) is a cube $(-\infty, x_S] \in \mathbb{R}^{\sharp S}$. Converting all univariate margins to uniform distributed random variables allows rewriting $F(x)$ as a d-dimensional copula

$$C(u) = \int\limits_{[0, u_S]} C_{T;S}\{G_{j|S}(u_j|v_S) : j \in T\} dC_S(v_S), \tag{9}$$

where $G_{j|S}(u_j|v_S)$ is a conditional distribution from copula $C_{S\cup\{j\}}$. If $T = \{i_1, i_2\}$, then

$$C_{S\cup\{i_1,i_2\}}(u_{S\cup\{i_1,i_2\}}) = \int\limits_{[0, u_S]} C_{i_1,i_2;S}\{G_{i_1|S}(u_{i_1}|v_S), G_{i_2|S}(u_{i_2}|v_S)\} dC_S(v_S). \tag{10}$$

Since the essential idea of vine copula is based on building a joint dependence structure by $d(d-1)/2$ bivariate copulae, (10) is an important building block in the construction of vines referring to a $(\sharp S + 2)$-dimensional copula built from a bivariate copula $C_{i_1,i_2;S}$.

In case of continuous random variables, the d-dimensional distribution function admits a density function $f(x_1, \ldots, x_d)$, which can be decomposed and represented by bivariate copula densities in an analogue manner. An example of density decompositions for the 5-dimensional case related to so called R-vine (regular vine) copula is given as follows.

The density of the R-vine structure which is used in simulation study, is

$$
\begin{aligned}
& c\{F_1(x_1), \ldots, F_5(x_5)\} \qquad (11)\\
& = c_{12}\{F_1(x_1), F_2(x_2)\} \cdot c_{13}\{F_1(x_1), F_3(x_3)\} \cdot c_{14}\{F_1(x_1), F_4(x_4)\} \cdot c_{45}\{F_4(x_4), F_5(x_5)\}\\
& \cdot c_{15;4}\{F(x_1|x_4), F(x_5|x_4)\} \cdot c_{24;1}\{F(x_2|x_1), F(x_4|x_1)\} \cdot c_{34;1}\{F(x_3|x_1), F(x_4|x_1)\}\\
& \cdot c_{23;14}\{F(x_2|x_{14}), F(x_3|x_{14})\} \cdot c_{35;14}\{F(x_3|x_{14}), F(x_5|x_{14})\}\\
& \cdot c_{25;314}\{F(x_2|x_{314}), F(x_5|x_{314})\}.
\end{aligned}
$$

Since vine copula can provide flexible way to construct joint distribution hence the literature on pair-copula construction is spreading steadily, and most recent information about it can be found on vine copula homepage[1].

B An R Package 'EnergyOnlineCPM'

At the moment there are two R packages for energy statistic, James & Matteson (2015) and Rizzo & Szekely (2016). Rizzo & Szekely (2016) is focused on the energy tests and James & Matteson (2015) concentrates on the Phase I change point model used for retrospective analysis. The package 'EnergyOnlineCPM' is the first package which centers on the nonparametric Phase II change point model to onlinely detect multiple change points for high dimensional time series based on the maximum energy test statistic using permutation samples.

The package is at the moment hosted in Github and it can be installed on the R terminal with following lines. Please note the package requires R version >= 3.3.2.

```
install.packages("devtools")
library(devtools)
install_github("YafeiXu/EnergyOnlineCPM")
library(EnergyOnlineCPM)
```

An example of simulation study for P2MECPM to detect a five dimensional time series with two change points in mean shift by standard Gaussian can be found in the manual Xu (2017). Installation, user manual, examples and more information can be referred to the project homepage `https://sites.google.com/site/EnergyOnlineCPM/`.

[1] `http://www.statistics.ma.tum.de/en/research/vine-copula-models/`

Package 'EnergyOnlineCPM'

February 18, 2017

Type Package

Title EnergyOnlineCPM Package

Version 1.0

Date 2017-02-14

Author Yafei Xu

Maintainer Yafei Xu <yafei.xu@hu-berlin.de>

Depends parallel, energy, R (>= 3.3.2)

Description This R package provides users a new function for nonparametric Phase II multiple multivariate change points detection.

URL https://sites.google.com/site/EnergyOnlineCPM

Repository https://github.com/YafeiXu/EnergyOnlineCPM

License GPL (>= 2)

R topics documented:

`EnergyOnlineCPM` *Installation of the R package 'EnergyOnlineCPM' for nonparametric Phase II multiple change points detection for high dimensional time series.*

Description

This R package provides users a new function for nonparametric Phase II multiple multivariate change points detection. In example part of this section the installation is given.

Details

Package:	EnergyOnlineCPM
Type:	Package
Version:	1.3
Date:	2017-02-14
License:	GPL (>= 2)

Author(s)

Yafei Xu <yafei.xu@hu-berlin.de>

Examples

```
# Installation of the package from Github
install.packages("devtools")
library(devtools)
install_github("YafeiXu/EnergyOnlineCPM")
library(EnergyOnlineCPM)
```

`maxEnergyCPMv` *Phase II Multiple Change Points Model for High Dimensional Time Series*

Description

This R function centers on nonparametric Phase II multiple change points detection for high dimensional time series. Three highlights are included in the function. Firstly, the new model is nonparametric which does not require any distributional pre-knowledge about the process. The test is based on the maximum energy statistic (see Gabor J. Szekely and Maria L. Rizzo 2004, Testing for Equal Distributions in High Dimension) and permutation samples. Secondly, the model is a Phase II change point model which is used for onlinely detection of stream data not for batch data. Phase II set-up has practical meaning in time series change detection. Thirdly, it is concentrated on high dimensional data, i.e. multivariate context. An important remark is that the data used in this function must be independent, i.e. every row in the N*d matrix must be an independent observation. If your data set contains not-independent observations then you need to handle the data using some filter functions, e.g. GARCH family to obtain the residuals which are theoretically independent.

Usage

```
maxEnergyCPMv(data1, wNr, permNr, alpha)
```

Arguments

`data1`	an N*d matrix, N is the number of observations and d the dimensions.
`wNr`	a scalar of warm-up.
`permNr`	a scalar of times of permutation.
`alpha`	a scalar of significant level

Details

The function returns ONLY ONE vector containing even number components, where the first half stands for detection time vector and the rest half stands for the vector of change time locations.

Value

`result`	a vector of locations of detection time in the first half, locations of change time in the second half.

Author(s)

Yafei Xu <yafei.xu@hu-berlin.de>

Examples

```
library(MASS)

# simulate 300 length time series
simNr=300

# simulate 300 length 5 dimensonal standard Gaussian series
Sigma2 <- matrix(c(1,0,0,0,0,  0,1,0,0,0,   0,0,1,0,0,  0,0,0,1,0,  0,0,0,0,1),5,5)
Mean2=rep(1,5)
sim2=(mvrnorm(n = simNr, Mean2, Sigma2))

# simulate 300 length 5 dimensonal standard Gaussian series
Sigma3 <- matrix(c(1,0,0,0,0,  0,1,0,0,0,   0,0,1,0,0,  0,0,0,1,0,  0,0,0,0,1),5,5)
Mean3=rep(0,5)
sim3=(mvrnorm(n = simNr, Mean3, Sigma3))

# construct a data set of length equal to 90.
# first 20 points are from standard Gaussian.
# second 30 points from a Gaussian with a mean shift with 555.
# last 40 points are from standard Gaussian.
data1=sim6=rbind(sim2[1:20,],(sim3+555)[1:30,],sim2[1:40,])

# set warm-up number as 20, permutation 200 times, significant level 0.005
wNr=20
permNr=200
alpha=1/200
maxEnergyCPMv(data1,wNr,permNr,alpha)
```

Index