A Non-parametric Multivariate Control Chart for High-Dimensional Financial Surveillance

Ostap Okhrin[1] and Ya Fei Xu*[2]

[1]*Institute of Economics and Transport, Technische Universität Dresden, Germany*

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

Abstract

This article presents a non-parametric control chart based on the change point model, for multivariate statistical process control (MSPC). The main constituent of the chart is the energy test that focuses on the discrepancy between empirical characteristic functions of two random vectors. Simulation study discusses in-control and out-of-control measures in context of mean shift and covariance shift. In real application, three financial data sets (in 5, 29, 90 dimensions) were employed to analyze the charting performance for financial surveillance in 2008-2009 crisis. The results from both simulation and empirical studies, compared with benchmarks, strongly advocate the proposed chart. This new control chart highlights in four aspects. Firstly, it is non-parametric, requiring no pre-knowledge of the random processes. Secondly, this control chart can monitor mean and covariance simultaneously. Thirdly it is devised for multivariate time series which is more practical in real data application. Fourthly, it is designed for online detection (Phase II), which is central for real time surveillance of stream data. This paper also contributes an `R` package `'EnergyOnlineCPM'` in CRAN for further research and practice.

Keywords: Phase II statistical process control; multivariate statistical process monitoring; change point model; energy test; financial surveillance; R package

*Corresponding author: `yafei.xu.huberlin@foxmail.com`.

1 Introduction

Control chart plays a pivotal role in statistical process monitoring. It is not rare to assume that a d-dimensional sequence $X_1, \ldots, X_t$, are identically independently distributed. In the series the number of change points is typically unknown, hence the problem that the control chart will tackle is to identify these change points, i.e. separation of the series $X_1, \ldots, X_t$ into diverse segments, where each adjacent pair of segments follows different distributions.

In early stage, feature research on statistical process control chart can be referred to Shewhart (1931), Shewhart & Deming (1939), Page (1954b), Page (1954a) and Roberts (1959). Since multivariate process are more useful and common in practical quality engineering (Woodall & Montgomery (2014)), therefore in recent decades, numerous papers have contributed to forward statistical process control (SPC) to multivariate context. A part of research is based on parametric assumptions, such as Crosier (1988) for multivariate CUSUM and Lowry, Woodall, Champ & Rigdon (1992) for multivariate EWMA and Zou & Tsung (2011) also assumed multivariate Gaussian distribution. Qiu & Hawkins (2001), Qiu & Hawkins (2003), Hawkins & Deng (2010) developed change point models with assumed pre-knowledge in in-control distribution. Another part of research focusing on online non-parametric multivariate change point models can be found in Zou, Wang & Tsung (2012), Holland & Hawkins (2014) and Zhou, Zi, Geng & Li (2015). A special accumulation of recent papers on nonparametric control chart can be referred to Chakraborti, Qiu & Mukherjee (2015). The latest review of nonparametric SPC control chart can be found in Qiu (2017).

For a proper detection of the changes, different statistical tests with different pros and cons were used, e.g. Student-t test, Bartlett test and Generalized Likelihood Ratio test, see Hawkins, Qiu & Kang (2003), Hawkins & Zamba (2005a), and Hawkins & Zamba (2005b). This paper employs the energy test, which is non-parametric and simple in implementation (as only means are to be computed) and has good power. Székely & Rizzo (2004), Zech & Aslan (2003), Székely & Rizzo (2013) investigated the energy statistic and the related test and performed the power analysis for distributional equality. Further, Kim, Marzban, Percival & Stuetzle (2009) show the satisfactory performance of the test in sliding window scheme with fixed window size in detection of change points in image data. Matteson & James (2014) and James & Matteson (2015) employ energy test combined with two different clustering schemes in change point retrospective analysis, i.e. the batch analysis (Phase I).

This paper proposes a non-parametric control chart for online detection of multiple change points for multivariate time series. This control chart has four main features. Firstly, it is non-parametric, what implies no need of pre-knowledge on the process comparing with traditional parametric control charts. Second feature is *online* monitoring, which can be applied life in many areas using real-time data. Thirdly, this control chart monitors multivariate time series which is pervasive in practice, e.g. in financial portfolio management. Last but not least, this new control chart can surveillance more general changes in multivariate time series, i.e. simultaneous surveillance of mean and covariance.

To our best knowledge, this is the first non-parametric control chart which can simultaneously monitor mean and covariance changes in the multivariate distribution in online fashion.

From the methodological side, the new 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 uses discrepancy of empirical characteristic functions of two random vectors, what differs from the common rank test. And the empirical distribution of the test statistic is thus obtained by permutation samples. Afterwards the sequential detection of change points can be conducted under the algorithm introduced by change point model (see Hawkins et al. (2003)) to perform online detection.

The simulation study investigates the proposed control chart in detecting mean shift (in Gaussian, Student-t and Laplace distribution) and covariance shift (in Gaussian and Student-t and Laplace distribution). The performance of the proposed control chart was compared with the benchmark control charts including the spatial rank based EWMA (SREWMA) by Zou et al. (2012), the self-starting multivariate minimal spanning tree (SMMST) based control chart by Zhou et al. (2015) and the non-parametric multivariate change point (NPMVCP) model based control chart by Holland & Hawkins (2014). The result indicates the outstanding performance of the proposed control chart.

In real-data application, the proposed control chart was employed in financial surveillance, i.e. monitoring high dimensional financial portfolios. Three data sets were used, separately in 5, 29, and 90 dimensions. The time windows of all three data sets covered the 2008-2009 global financial crisis, with window width of more than 1000 observations. The result shows that the new control chart is capable to detect the abnormal distributional change in financial market. For the purpose of reproducible research and practice of non-parametric online MSPC, authors contributed an `R` package '`EnergyOnlineCPM`' in this paper. Among recent control chart researches this package is the first `R` package which can be used in online simultaneous monitoring of mean and covariance for multivariate data. More details on the package can be checked in user manual (Xu (2017)) and the homepage (https://sites.google.com/site/energyonlinecpm).

The paper is structured as follows. In Section 2, the methodology is given, introducing the energy test and the preliminary of change point model in two diverse phases (Phase I and II). Simulation study, application study and their corresponding results are presented in Section 3 and 4 respectively. Section 5 concludes. An introduction of the package 'EnergyOnlineCPM' and some supplementary materials about the data meta information are attached in appendix.

2 Methodology

2.1 Energy Test

It is known that the corresponding characteristic functions of d-dimensional random vectors X and Y, i.e. ϕ_X and ϕ_Y, are uniquely determined since $X \sim F_X$ and $Y \sim F_Y$, hence using the divergence between characteristic functions of the two random vectors to monitor the change is an applicable routine. Székely & Rizzo (2005) used an integrated weighted distance between two characteristic functions, and showed that the larger the distance the more possible that the two random vectors are not identically distributed.

Theorem 1. *Let $X \sim F_X$ and $Y \sim F_Y$ be two d-dimensional random vectors. X', Y' are independent copies of X and Y. The corresponding characteristic functions of the two random vectors are ϕ_X and ϕ_Y. If $0 < \alpha < 2$ with $\mathbb{E}||X||_2^\alpha < \infty$ and $\mathbb{E}||Y||_2^\alpha < \infty$ then*

$$\int\limits_{\mathbb{R}^d} \frac{|\phi_X(p) - \phi_Y(p)|^2}{||p||_2^{d+\alpha}} dp = W(d,\alpha)\mathcal{E}^\alpha(X,Y), \tag{1}$$

with

$$\begin{aligned} W(d,\alpha) &= \frac{2\Pi^{\frac{d}{2}}\Gamma(1-\frac{\alpha}{2})}{\alpha 2^\alpha \Gamma(\frac{\alpha+d}{2})}, \textit{where } \Gamma(\cdot) \textit{ being the Gamma function,} \\ \mathcal{E}^\alpha(X,Y) &= 2\mathbb{E}||X-Y||_2^\alpha - \mathbb{E}||X-X'||_2^\alpha - \mathbb{E}||Y-Y'||_2^\alpha. \end{aligned} \tag{2}$$

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

Theorem 2. *Under assumptions of 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. Let the random samples of random vectors X, Y be $S_X = \{X_1, \dots, X_m\}$ and $S_Y = \{Y_1, \dots, Y_n\}$ respectively. The empirical counterpart of (2) can be derived as

$$\begin{aligned}\hat{\mathcal{E}}^\alpha(S_X, S_Y) &= \frac{mn}{m+n}\Bigg(\frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}||X_i - Y_j||_2^\alpha \\ &\quad -\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}||X_i - X_j||_2^\alpha - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||Y_i - Y_j||_2^\alpha\Bigg). \qquad (3)\end{aligned}$$

With Theorem 2 it is clear that the larger the quantity of $\hat{\mathcal{E}}^\alpha(S_X, S_Y)$ the higher the likelihood that the components in S_X, S_Y are from diverse distributions. Hence (3) can be used as the distance between two unknown distributions of random-samples, therefore (3) can be employed as the test statistic, where the empirical distribution of $\hat{\mathcal{E}}^\alpha$ can be obtained by permutation samples with the following approach.

In order to conduct a permutation based statistical test, first the hypothesis is set that

H_0 : components in S_X and S_Y are identically distributed,

H_1 : components in S_X and S_Y have different distributions.

As mentioned above, the test statistic is set as $\hat{\mathcal{E}}^\alpha(S_X, S_Y)$. Next step is to construct the distribution of the test statistic. Since the theoretical distribution of the test statistic is intractable, hence here the permutation test is employed under the assumption of independent random vectors. In order to accomplish this work, P permutation samples can be generated by random shuffling of $\{x_1, \dots, x_m, y_1, \dots, y_n; x_i, y_j \in \mathbb{R}^d\}$. For the start sample $\{x_1, \dots, x_m, y_1, \dots, y_n; x_i, y_j \in \mathbb{R}^d\}$, since the sample size is $m+n$, therefore there are $(m+n)!$ permuted samples. For every shuffling sample the energy test statistic $\hat{\mathcal{E}}^\alpha(S_X, S_Y)$ is calculated, hence finally it obtains a P-vector of test statistics based on P different permutation samples, then the empirical distribution of $\hat{\mathcal{E}}^\alpha(S_X, S_Y)$ can be obtained by sorting the values in the P-vector and the critical value can be obtained by choosing a quantile following the given confidence level. Readers for more details about the permutation test and its related empirical distribution can be referred to Fisher (1937), Pitman (1937) and Pitman (1938).

2.2 Review of SREWMA, SMMST and NPMVCP

In this sub-section, three recent published non-parametric control charts are briefly reviewed, including the SREWMA by Zou et al. (2012), the SMMST by Zhou et al. (2015) and the NPMVCP by Holland & Hawkins (2014). These three control charts will appear in the simulation section as the benchmark control charts.

2.2.1 A Review of SREWMA

Zou et al. (2012) proposed a non-parametric multivariate EWMA control chart based on the spatial rank test to monitor the location parameter change. It assumes that for a sequence of random vectors $X_{-g+1}, \ldots, X_0, X_1, \ldots, X_t \in \mathbb{R}^d$, where $X_{-g+1}, \ldots, X_0$ are g vectors before the start point X_1, the multivariate change point problem is represented as

$$X_i \overset{\text{i.i.d.}}{\sim} \begin{cases} \mu_0 + \Omega\varepsilon_i & \text{if } i \leq \tau, \\ \mu_1 + \Omega\varepsilon_i & \text{if } i > \tau, \end{cases} \tag{4}$$

where τ stands for the change index, Ω for a full-rank $d \times d$ transformation matrix and $M := \Omega^{-1}$. It is set that $\varepsilon_i \in \mathbb{R}^d$ is i.i.d. with $\mathrm{Cov}(\varepsilon_i) = I_d$ and $\mathbb{E}(\varepsilon_i) = 0$. Then the test statistic is given as

$$Q_t^{R_E} = \frac{(2-\lambda)d}{\lambda\xi_t}||V_t||^2, \tag{5}$$

where

$$\begin{aligned} V_t &= (1-\lambda)V_{t-1} + \lambda R_E(\hat{M}_{t-1}X_t),\ V_0 = 0, \\ \xi_t &:= \hat{\mathbb{E}}\{||R_F(MX_t)||^2\}, \\ &\approx \frac{1}{g+t-1}\left\{\sum_{j=1}^{g}||\tilde{R}_E(\hat{M}_g X_j)||^2 + \sum_{j=1}^{t-1}||R_E(\hat{M}_{j-1}X_j)||^2\right\}, \\ \tilde{R}_E(\hat{M}_g X_j) &= \frac{1}{g}\sum_{k=1}^{n} U(\hat{M}_g(X_j - X_k)), \\ R_E(\hat{M}_{t-1}X_t) &= \frac{1}{g+t-1}\sum_{j=1}^{t-1} U\{\hat{M}_{t-1}(X_t - X_j)\}. \end{aligned}$$

Here $U(X)$ is called the spatial sign function that

$$U(X) = \begin{cases} ||X||^{-1}X & \text{if } X \neq 0, \\ 0 & \text{if } X = 0, \end{cases}$$

where $||X|| = (X^\top X)^{1/2}$ is the Euclidean norm of the d-vector X. And $R_E(X_t) = \frac{1}{g}\sum_{j=1}^{g} U(X_t - X_j)$ is the empirical version of the spatial rank for the $d-$vector X_t, and the theoretical counterpart is $R_F(X_t) = \mathbb{E}_{X_j}\{U(X_t - X_j)\}$. Under the regulation (see Proposition 2 in Zou et al. (2012)), the test statistic (5) has the asymptotic distribution following

$$Q_t^{R_E} = \frac{(2-\lambda)d}{\lambda \xi_t}||V_t||^2 \to \chi_d^2, \text{ if } \lambda \to 0, \lambda t \to \infty.$$

2.2.2 A Review of SMMST

Zhou et al. (2015) integrated the multivariate version Wald-Wolfowitz runs test (Friedman & Rafsky (1979)) into the change point model (Hawkins et al. (2003)) based control chart to perform non-parametric multivariate location surveillance. The main idea of the multivariate Wald-Wolfowitz runs test in Friedman & Rafsky (1979) is to use the minimal spanning tree (MST) approach to generalize the sorted list in uni-variate runs test to multivariate context. That is, in the d-dimensional data set with N points, every data point is seen as a node and all the nodes can be connected by $N(N-1)/2$ edges. And for every edge a quantity can be granted by using the Euclidean distance of two $d-$dimensional nodes. Then Friedman & Rafsky (1979) gives three steps to compute the test statistic.

1. Use the MST algorithm (see Appendix in Friedman & Rafsky (1979)) to construct the MST for all nodes in the data set.
2. Remove all edges, of which the two nodes are from diverse groups.
3. Compute the runs statistic R, i.e. the number of the disjoint sub-trees in the MST.

The test null hypothesis for a two sample problem (with $m, n,$ where $N := m + n$, as the sample sizes of the two groups), i.e. $\mathrm{H}_0 : F_X = F_Y$, will be rejected if R is smaller than a critical value.

It is defined that $Z_i,\ 1 \leq i \leq N-1$ is an indicator function such that

$$Z_i = \begin{cases} 1 & \text{if the } i\text{-th edge links nodes from diverse groups,} \\ 0 & \text{else.} \end{cases} \tag{6}$$

Then $R := \sum_{i=1}^{N-1} Z_i + 1$. The mean and conditional variance of R can be derived as follows,

$$\begin{aligned}
\mathbb{E}(R) &= \frac{2mn}{N} + 1, \\
\mathrm{Var}(R|C) &= \frac{2mn}{N(N-1)} \left\{ \frac{2mn - N}{N} + \frac{C - N + 2}{(N-2)(N-3)(N(N-1) - 4mn + 2)} \right\},
\end{aligned}$$

where C is determined by the node degrees. At last the test statistic W has the asymptotic distribution based on the permutation samples, such that

$$W := \frac{R - \mathbb{E}(R)}{\{\mathrm{Var}(R|C)\}^{1/2}} \rightarrow N(0,1), \text{if } m, n \rightarrow \infty.$$

2.2.3 A Review of NPMVCP

Holland & Hawkins (2014) devised a non-parametric control chart using multivariate rank based test by Choi & Marden (1997). It gives the multivariate change point model (Hawkins et al. (2003)) to identify changes in a sequence, $X_1, \ldots, X_t$, that

$$X_i \sim \begin{cases} F(\mu) & \text{if } i \leq \tau, \\ F(\mu + \delta) & \text{if } i > \tau, \end{cases} \tag{7}$$

and $\mathrm{H}_0 : \delta = 0, \;\; \mathrm{H}_1 : \delta \neq 0$.

Choi & Marden (1997) state that under the null hypothesis, i.e. there is no change, then the asymptotic distribution of the statistic $\frac{tk}{t-k} \bar{r}_t^{(k)^\top} \tilde{\Sigma}_{k,t}^{-1} \bar{r}_t^{(k)}$, $k \in \{1, \ldots, t-1\}$, can be represented as follows,

$$\frac{tk}{t-k} \bar{r}_t^{(k)^\top} \tilde{\Sigma}_{k,t}^{-1} \bar{r}_t^{(k)} \rightarrow \chi_d^2, \tag{8}$$

where

$$\begin{aligned}
\tilde{\Sigma}_{k,t} &= \frac{t^2}{t-2} \left\{ \frac{1}{k^2} \sum_{i=1}^{k} R_k(X_t) R_k(X_i)^\top + \frac{1}{(t-k)^2} \sum_{i=k+1}^{t} R_{t,k}^*(X_i) R_{t,k}^*(X_i)^\top \right\}, \\
R_k(X_i) &= \sum_{j=1}^{k} h(X_i, X_j), \\
R_{t,k}^*(X_i) &= \sum_{j=k+1}^{t} h(X_i, X_j), \\
h(X_i, X_j) &= \frac{X_i - X_j}{||X_i - X_j||}.
\end{aligned}$$

Here $\tilde{\Sigma}_{k,t}$ is the pooled sample covariance matrix of the centered rank vector, and $R_k(X_i)$ is the multivariate centered rank, and $h(X_i, X_j)$ is the kernel function that $h(X_i, X_j) = -h(X_j, X_i)$. At last Holland & Hawkins (2014) uses the test statistic

$$r_{k,t} \;=\; \bar{r}_t^{(k)\top} \hat{\Sigma}_{k,t}^{-1} \bar{r}_t^{(k)},$$

where $\hat{\Sigma}_{k,t} = (\frac{t-k}{tk})\hat{\Sigma}_t$ and $\hat{\Sigma}_t = \frac{1}{t-1}\sum_{i=1}^{t} R_t(X_i)R_t(X_i)^\top$ is the unpooled estimator of covariance matrix. It states that in the simulation study the power of using pooled or unpooled estimator of covariance matrix leads to similar performance. However for convenience of computation the unpooled covariance estimator $\hat{\Sigma}_t$ is employed.

2.3 Phase I Change Point Model

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

Assume there is only one change occurred at $\tau + 1$, then the change point detection problem can be represented in the following test hypotheses,

$$\begin{aligned} \mathrm{H}_0: &\quad X_i \sim F_0,\ 1 \le i \le T, \\ \mathrm{H}_1: &\quad X_i \sim \begin{cases} F_0, & 1 \le i \le \tau, \\ F_1, & \tau + 1 \le i \le T. \end{cases} \end{aligned}$$

A two-sample parametric or non-parametric test with test statistics $B_{i,T}$ is usually applied in this case. Before conducting the permutation test the significant level should be set. If $B_{i,T}$ is larger than a predefined critical value $h_{i,T}$, i.e. $B_{i,T} > h_{i,T}$, then the null hypothesis is rejected, meaning that the two sets of random vectors are not identically distributed. Then a detection point is admitted at i-th point. Since the change point location is unknown, hence the two-sample test will be performed at every point i, $1 \le i < T$, i.e. conducting $T - 1$ dichtomizations. According to the change point model (Hawkins et al.

(2003)), the test statistic is derived from $B_{i,T},\ i = 1, \dots, T-1$, as the largest value, such that

$$B_T = \max_{1 \leq i < T} B_{i,T}.$$

The null hypothesis is rejected if $B_T > l_T$, where l_T is the critical value derived from the distribution of B_T. Please note that $h_{i,T}$ is the critical value of the test statistic $B_{i,T}$, and l_T is the critical value of the test statistic B_T, and $B_T = \max_{1 \leq i < T} B_{i,T}$. The Type I error α in this context means that the model signals a change point when actually there is actually no change occurs. The distribution of the test statistic B_T can be obtained either by its asymptotic distribution (if available) or by simulation methods e.g. permutation test scheme. At the end, the change location can be estimated by

$$\hat{\tau} = \arg\max_{1 \leq i < T} B_{i,T}.$$

2.4 Phase II Change Point Model

In contrary to the Phase I detection based on the fix-sized sample $\{x_1, ..., x_T\}$, Phase II detection considers the dynamic-sized sample $\{x_1, ..., x_t\}$ with an increasing size, i.e. the sample size t always increases with time proceeding. For this reason Phase II detection is also termed as online detection and sequential detection, e.g. the stock price is updated with time, therefore the length of time series $\{x_1, ..., x_t\}$ is always increased, i.e. the t is not fixed or static but dynamic. Hence the detection in Phase II concentrates on the dynamic stream data.

With the Phase I analysis in Section 2.3, Phase II can be extended from the Phase I with increasing sample size to update the old sample size. That is whenever a new observation x_t arrives then a new sample $\{x_1, \dots, x_T, x_{T+1}, \dots, x_t\}$ is constructed and the new sample size is denoted here as t. For example, if the old sample is $\{x_1, \dots, x_T\}$ and the new arrival is x_{T+1}, then the new sample becomes $\{x_1, \dots, x_T, x_{T+1}\}$. In this case $t = T+1$. For every new arrival of observation the Phase I analysis will be performed based on the new sample $\{x_1, \dots, x_T, x_{T+1}, \dots, x_t\}$. For this sample, $t-1$ two-sample tests will be performed, therefore $\{B_{1,t}, \dots, B_{t-1,t}\}$ can be obtained, further $B_t = \max\{B_{1,t}, \dots, B_{t-1,t}\}$. Hence the null hypothesis is rejected if $B_t > l_t$. The Type I error α can be represented with

$$\begin{aligned} \mathbb{P}(B_1 > l_1) &= \alpha,\ t = 1, \\ \mathbb{P}(B_t > l_t | B_{t-1} \leq l_{t-1}, \dots, B_1 \leq l_1) &= \alpha,\ t > 1. \end{aligned} \tag{9}$$

In statistical process control, the in-control average run length (IC-ARL), ARL_0, is the inverse of the Type I error, i.e. $ARL_0 = 1/\alpha$, which stands for the average step length of the detection until the first erroneous alarm signals.

3 Simulation Study

3.1 Set-up in Simulation Study

In the study of statistical process monitoring, the assessment of change-point detection methods uses mainly two measures, the IC-ARL and the out-of-control average run length (OC-ARL). IC-ARL assumes that the time series follow a distribution without change in order to calculate the steps until the first erroneous signal flags, therefore the larger the IC-ARL the better the model. OC-ARL assumes that the process has a change point in a known point in order to compute the average step length until the model detects this pre-set change. Since commonly there is delay in detection, hence the detection method is expected to have a small OC-ARL.

In simulation study of this work, the proposed model is assessed in different scenarios including mean change and covariance change. Here the OC-steps are set separately as 100 and 200 for middle-term (OC 100 steps) and long-term detection (OC 200 steps).

In mean shift part for middle and long term detection assessment with $\tau = 32$ and OC-steps of 100 and 200, the DGPs are Gaussian, Student-t_5 and $Laplace(0, \Sigma_L)$, $\Sigma_L = (a_{ij})$, $a_{ii} = 11$, $a_{ij} = 10$. The shifts are set as $\delta = 0, 0.25, 0.5, 0.75, 1, \ldots, 9$. The result can be referred to Table 5. Here the benchmark is NPMVCP. Especially, the IC-ARLs in this scenario comparison are given in Table 1 and Figure 2.

In single component mean shift part. the detection assessment is conducted with $\tau = 32$ and OC-steps of 100 and 200, and the DGPs are the same as the first scenario including Gaussian, Student-t_5 and Laplace. The shifts are set the $\delta = 0, 0.25, 0.5, 0.75, 1, \ldots, 9$. The result can be referred to Table 7 and Figure 3. Here the benchmark is NPMVCP.

Additionally, as mentioned in Section 2.2, the proposed model is compared with the SMMST and the SREWMA in scenario of 200 OC-steps mean shift under Gaussian, t_5 and $Gamma_5$. τs are set as 40 and 90, and $\delta = 1, 1.5, 2, 3, 4$. Result is given in Table 8 and Figure 5.

In covariance shift part for middle and long term detection assessment with $\tau = 32$ and OC-steps of 100 and 200, the DGPs are set as the Gaussian $N(0, I)$ and Student-t_5 with $\sigma^2 = 0.25, 0.5, 0.75, 2, 3, \ldots, 11$. Refer to Table 6 and Figure 4 for the result.

The recent paper studying non-parametric multivariate control chart using the change point model (Hawkins et al. (2003)) is NPMVCP in Holland & Hawkins (2014). Therefore, the benchmark model for comparison in this paper is NPMVCP model, which is a mainstream non-parametric change point model for multivariate location shift detection. Since this paper used the code provided in `R` package `NPMVCP` in Holland & Hawkins (2014) without using optimal quarantine technique, therefore for fair comparability, here the quarantine was not considered for both models. The warm-up was set as 32 consistent to the default set-up in `NPMVCP`.

Since the test integrated in the proposed control chart is based on permutation samples, hence the choice of simulation runs is necessarily to be considered. Because all metrics were computed based on the i.i.d. samples, hence under the law of large numbers the mean of OC-ARLs will converge. In order to choose an appropriate size of simulation, a study was conducted, see Figure 1. It is clear that the simulation runs larger than 50 led the similar results and the mean of both models' OC-ARLs arrived closely to the run of 50. Hence in this paper, the simulation size was chosen as 50 runs for sufficiency.

3.2 Results and Analysis of Simulation Study

In mean shift of middle and long-term scenario with 100 and 200-OC steps (Table 1 and Figure 2), the proposed control chart outperformed NPMVCP in most cases in all three DGPs. More detailed, in all three distributions the proposed control chart performed better in moderate shift ($\delta \geq 2$) for three dimensional cases and in small shift ($\delta \geq 0.75$) in ten dimensional cases, see Gaussian and t_5. It is clear when the dimension of data set increases then the proposed control chart's performance is enhanced.

In single mean shift middle and long-term scenario with 100 and 200-OC steps (Table 7, Figure 3), it shows that NPMVCP performs well in small shift and the proposed control chart performs well in moderate shift ($\delta \geq 2$). However Table 1 clearly shows that in all categories, the proposed control chart outperforms the NPMVCP in IC-ARL. It is clear that the NPMVCP has only roughly 60 percent correct detection, which is far worse than the proposed control chart. According to the Table 1 and the above analysis, it can conclude that the proposed control chart in mean shift detection is capable and robust.

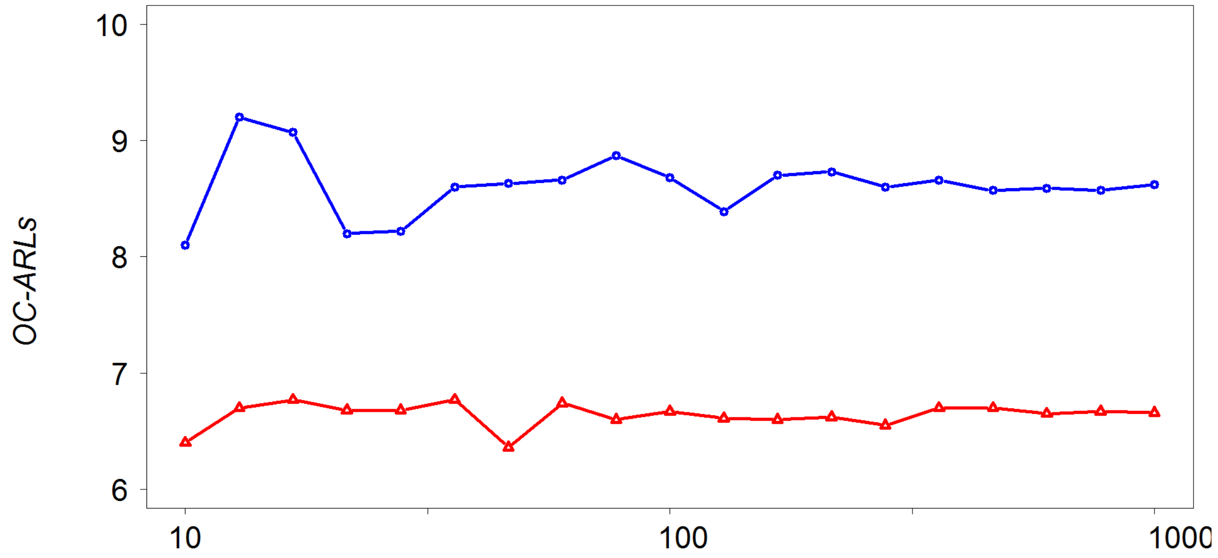

Figure 1: Comparison of OC-ARLs (Out-of-Control Average Run Length) under NPMVCP and proposed control chart through 10 to 1000 runs of simulation. In the simulation, the DGP is set from a five dimensional standard Gaussian distribution shifted with mean plus 3 and the warm-up is set to 32 identical to the setting in Holland and Hawkins' package NPMVCP. $\tau = 32$.

In order to further show the robustness and capacity of the proposed control chart, Table 8 provides another evidence. In this table the proposed control chart is compared with another two non-parametric control charts SMMST (Zhou et al. (2015)) and SREWMA (Zou et al. (2012)). In Figure 5, it is clear that the proposed control chart performs generally better than the other benchmarks, especially in Gaussian and t_5 cases.

In covariance shift part for middle and long-term detection assessment with $\tau = 32$ and OC-steps of 100 and 200, the proposed control chart outperformed the NPMVCP in most cases, while NPMVCP had ability to detect the small covariance shift, e.g. in scale of $\sigma^2 = 2$. In larger covariance shifts or larger dimension data sets, the proposed model gave better results. It is clear that NPMVCP has almost constant change no matter the change of dimensions or distributions, while the proposed control chart shows high sensitivity to the increase of dimension, see Table 6 and Figure 4.

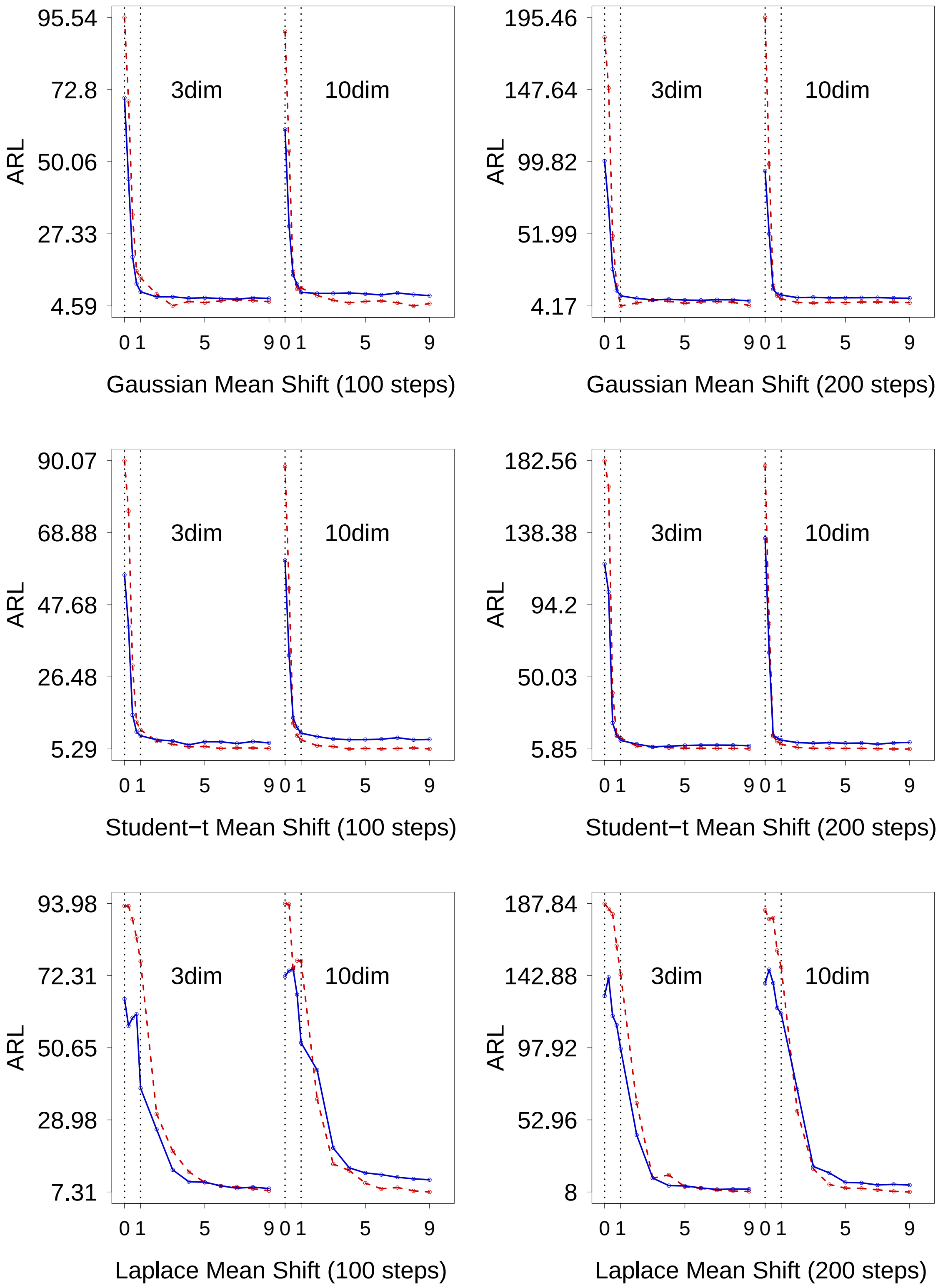

Figure 2: Simulation results (Table 5) for mean shift with DGPs of Gaussian, Student-t_5 and Laplace distributions. The blue line stands for NPMVCP and the red for proposed control chart.

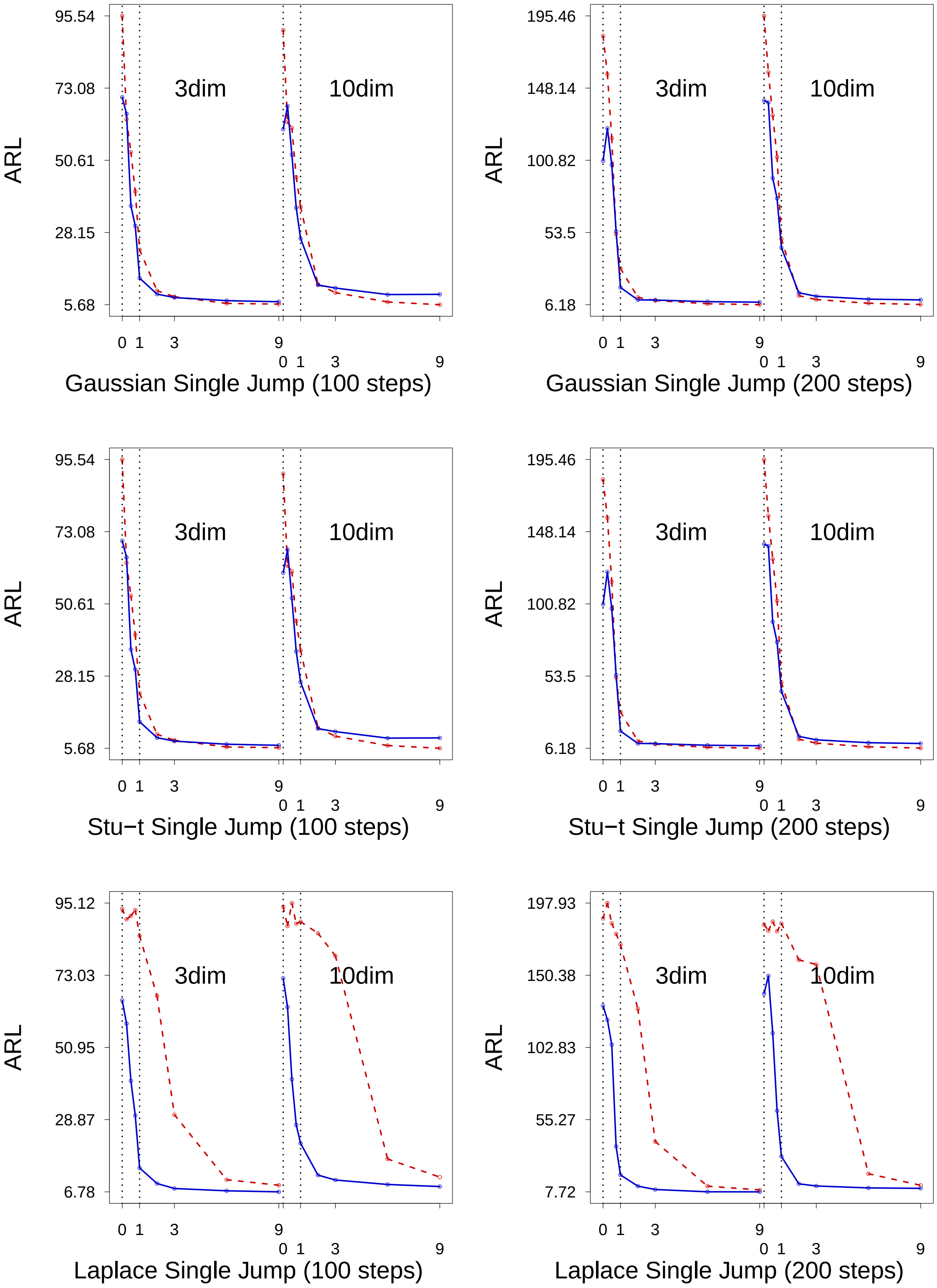

Figure 3: Single mean shift (see Table 7) for multivariate Gaussian, Student-t_5 and Laplace with mean $\mu_k + \delta$, $\delta \in \{0, 0.25, 0.50, 0.75, 1, 2, 3, 6, 9\}$. The red line stands for the proposed control chart and the blue line for the Holland & Hawkins (2014).

		Gaussian		t		Laplace	
ARL_0	Dim.	proposed	NPMVCP	proposed	NPMVCP	proposed	NPMVCP
200	3	182.36 (50.29)	124.82 (71.55)	182.56 (48.89)	118.66 (74.11)	187.84 (41.96)	135.62 (67.62)
	10	195.46 (19.71)	138.62 (70.29)	179.26 (52.06)	135.02 (69.45)	183.47 (45.77)	140.28 (67.84)
100	3	95.54 (16.91)	67.30 (40.25)	90.07 (27.13)	68.00 (34.34)	93.30 (20.45)	62.84 (35.35)
	10	91.13 (24.29)	58.24 (38.50)	88.38 (29.25)	74.12 (34.49)	93.98 (21.11)	69.34 (34.68)

Table 1: Comparison of proposed model against the NPMVCP model (Holland & Hawkins (2014)) in In-Control ARL for mean shift with 100 and 200 OC-steps. In parentheses the standard deviations are given.

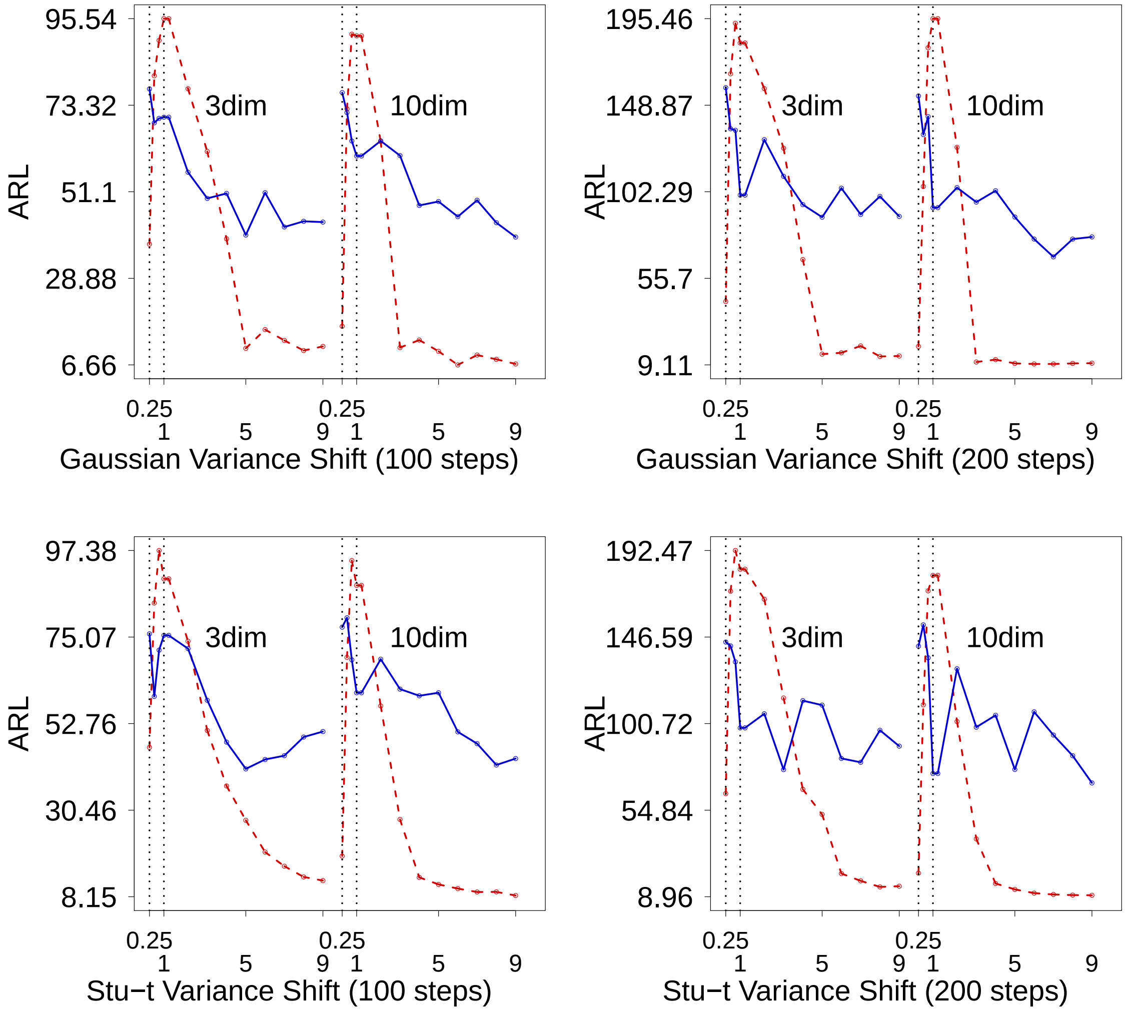

Figure 4: Simulation results for covariance shift (Table 6) with DGPs of Gaussian and Student-t_5. The blue line stands for NPMVCP and the red line for proposed model.

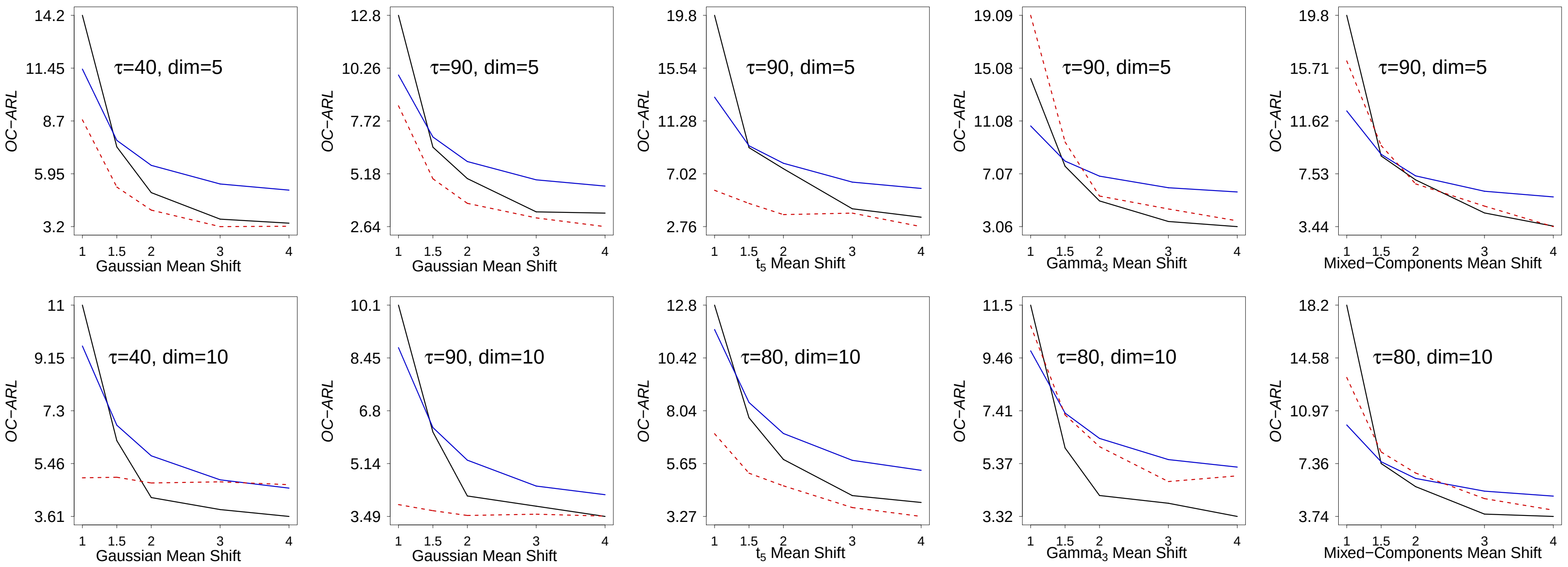

Figure 5: Comparison of simulation results (Table 8) of the proposed control chart (red) with SMMST (Zhou et al. (2015)) in black and SREWMA (Zou et al. (2012)) in blue appeared in Table 2, 3, 4, 5 in Zhou et al. (2015).

4 Real Data Application in Financial Surveillance

4.1 Data Sets

In this section, three data sets were employed. The first data set is a five dimensional close prices on the U.S. ETF (Exchange-Traded Fund) market, including five tickers of DGT, EWD, GLD, IGV and IUSG, see Table 2. The data set is obtained from the Wall Street Journal web site.

The second data set contains 29-dimensional close prices from DJIA (Dow Jones Industrial Average) component firms, see Table 3 in Appendix. The third data set contains 90 close prices from S&P 100 components, see Table 4 in Appendix. The both were obtained from Yahoo Finance. An illustration of the both data sets can be found in Figure 7

The window length spans from 20070103-20101231, in general 1007 observations for each data set. Therefore the global financial crisis occurred in 2008-2009 is covered by all three data sets, which is our interest to check if the proposed control chart is capable to detect the market shift.

In previous section of methodology, the energy test was introduced, where it is known that this test needs independent samples. Therefore before using the proposed control chart, all three data sets need to be handled. In this work the VAR (Vector AutoRegressive), see Sims (1980), was used to filter out the residuals from the raw data sets in the first step. VAR model generalizes the uni-variate auto-regressive model (AR model) by allowing for more than one endogenous variable to capture the linear inter-dependency among multiple time series.

A z-th order VAR, denoted VAR(z), is

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_z y_{t-z} + e_t,$$

where the z-th observation y_{t-z} is the z-th lag of y, c is a $d \times 1$ vector of constants, A_z is a time-invariant $d \times d$ matrix and e_t is a $d \times 1$ vector of error terms satisfying $\mathbb{E}(e_t) = 0$, $\mathbb{E}(e_t {e_t}^\top) = 0$ and $\mathbb{E}(e_t {e_{t-k}}^\top) = 0$.

After filtering out residuals by VAR model, the serial correlation between multivariate residuals should be checked. A multivariate portmanteau test, see Hosking (1980), was employed to test the independence of VAR model's multivariate residuals. The three data sets were fitted by VAR models separately in VAR(10) for ETF data set, VAR(5) for DJIA and S&P100 data sets. The out-filtered residuals of DJIA and S&P data sets are shown in Figure 8, and the Figure 6 show the residuals' relationships for the five dimensional ETF data set.

The proposed control chart is set in this application section with 32 warm-ups, 0.005 significant level, which are the same setting in package `NPMVCP`.

4.2 Results and Analysis of Application

In real data applications, actually the real data set can be seen as a complex scenario combining mean shift and covariance shift together, therefore a control chart with the capacity to simultaneously detect the above scenarios will have competitive edge. The findings in applications show similar to those in simulation study that the performance of the proposed control chart stands out.

First of all, the proposed control chart detected out the changes of the market regimes in five dimensional ETF data set. Figure 9 illustrate the detection points using the proposed control chart and the NPMVCP. It is clear that the control chart by NPMVCP has more detection points than the proposed which is similar to the result in simulation study. A possible reason is that the NPMVCP has more erroneous detection than the proposed model, see the IC-ARL performance in mean shift.

Secondly, the proposed control chart has strong detection power for covariance shift detection for high dimensional data, consistent with the results shown in simulation study. In Figure 10, it shows that NPMVCP has obvious large delay in detection of financial turmoil period (2008.09-2009.03). The first detection point for NPMVCP model on financial crisis is 20090205 (estimation of change point on 20081118), while the proposed control chart is on 20081007 (estimation of change point on 20081007). Therefore the proposed model can signal alarm for investors of the in-crisis, while the NPMVCP can not do this.

Thirdly, similar to the ETF data set, the proposed control chart detected out the change points of financial crisis in 29 and 90 dimensional data sets. In Figure 11 it shows the proposed control chart signaled detection points for in-crisis, separately on 20081005 for 29-dimensional DJIA data set (similar to result in James & Matteson (2015)) and on 20080917 for 90-dimensional S&P data set. Hence the proposed model can be used to serve as an alarm tool for the investors in financial market.

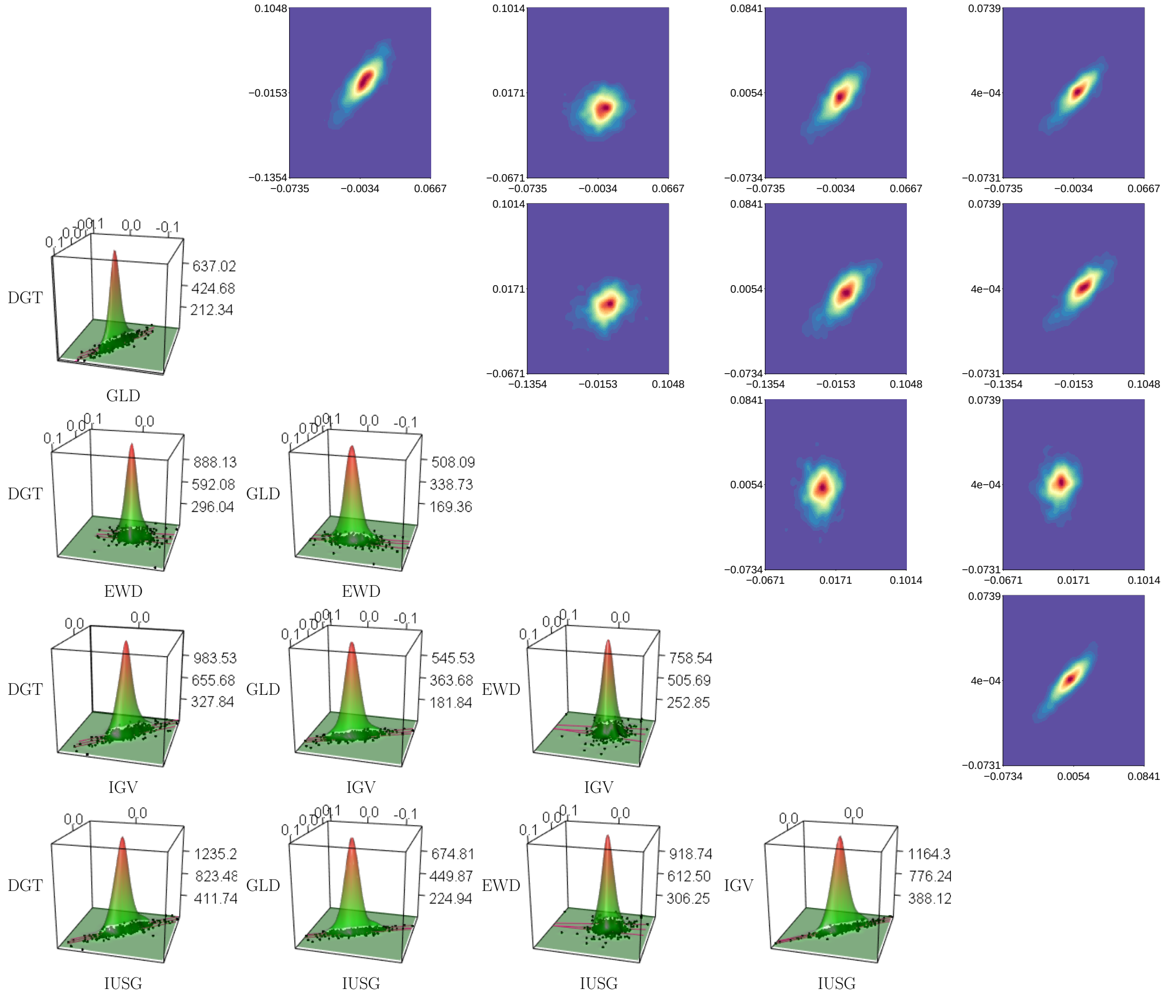

Figure 6: The lower triangular panels show the residuals scatter-points with quantile regressions in 0.05, 0.5, 0.95 quantiles, on which the estimated kernel density is illustrated. The upper triangular panels show the contours of the pairwise residuals from the data set of five dimensional ETF data set (Table 2).

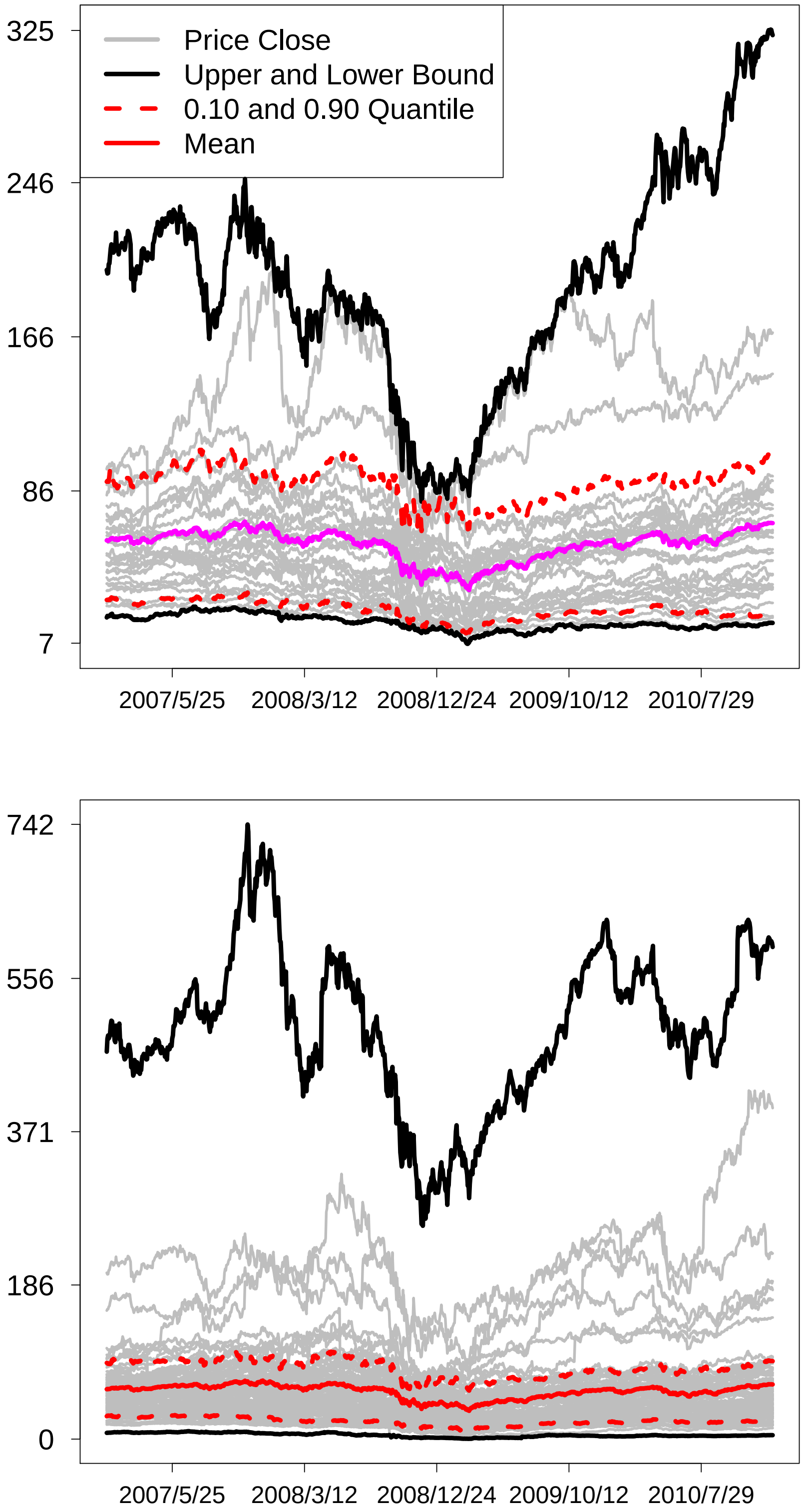

Figure 7: The upper panel presents the 29-dimensional DJIA data set. The lower panel illustrates the 90-dimensional SP100 data set (Table 3).

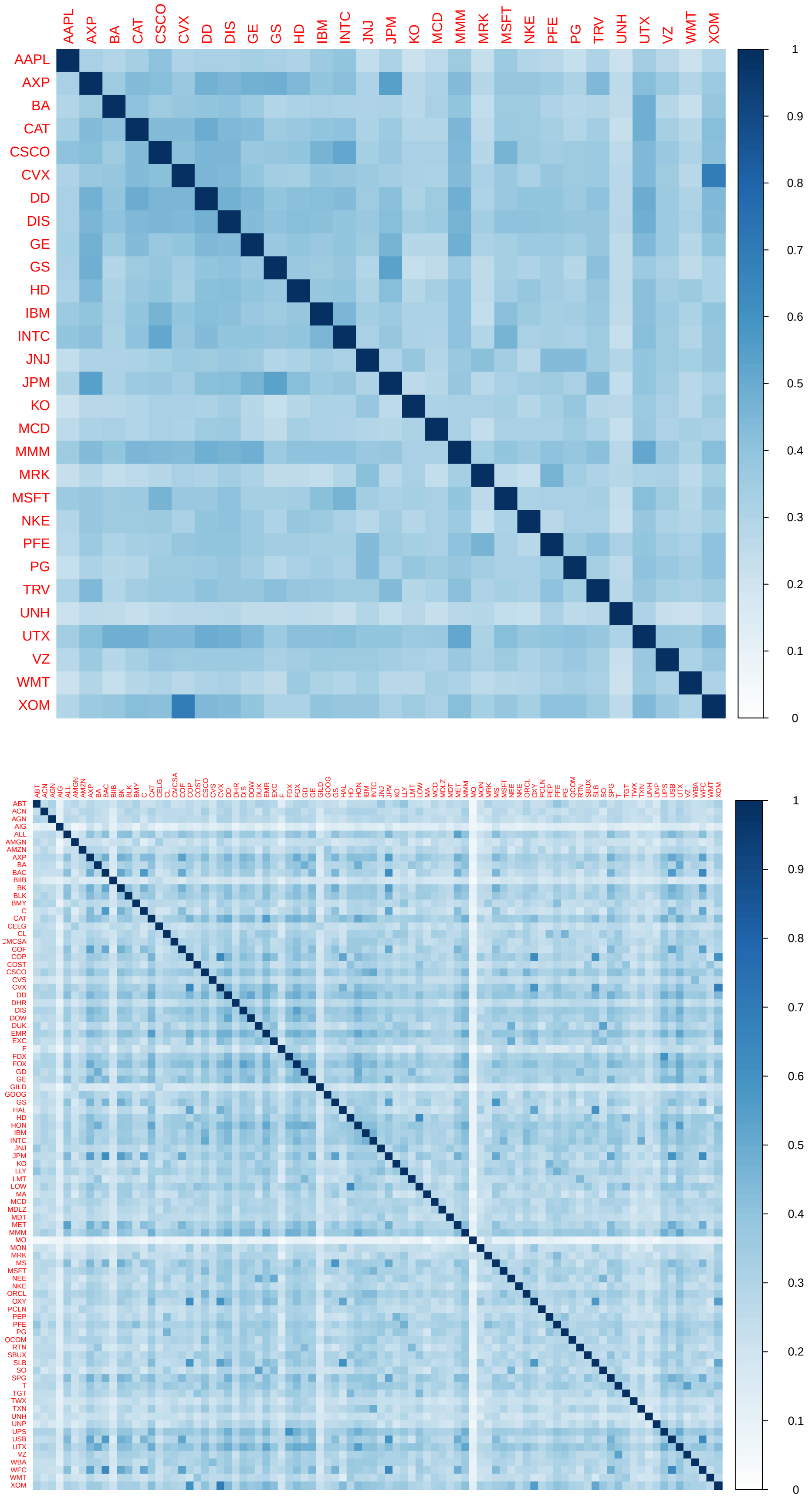

Figure 8: Pearson correlations between 29 dimensional DJIA residuals in upper panel. Pearson correlations between 90 dimensional SP100 residuals in lower panel.

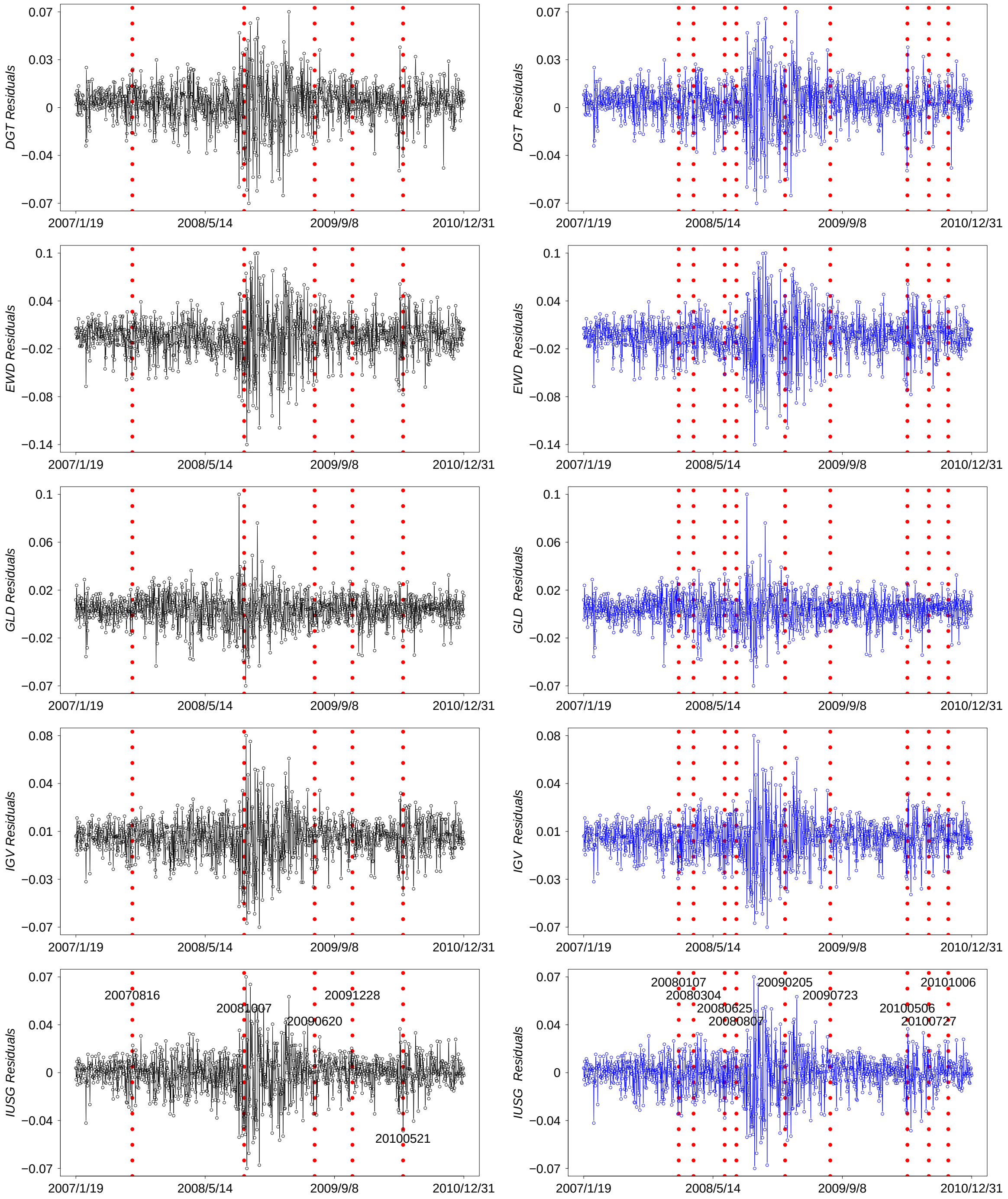

Figure 9: Change detection by the proposed (black) and the NPMVCP (blue) control charts. DGT: SPDR Global Dow ETF, EWD: iShares MSCI Sweden Capped ETF, GLD: SPDR Gold Trust, IUSG: iShares Core SP U.S. Growth ETF, IGV: iShares North American Tech-Software ETF.

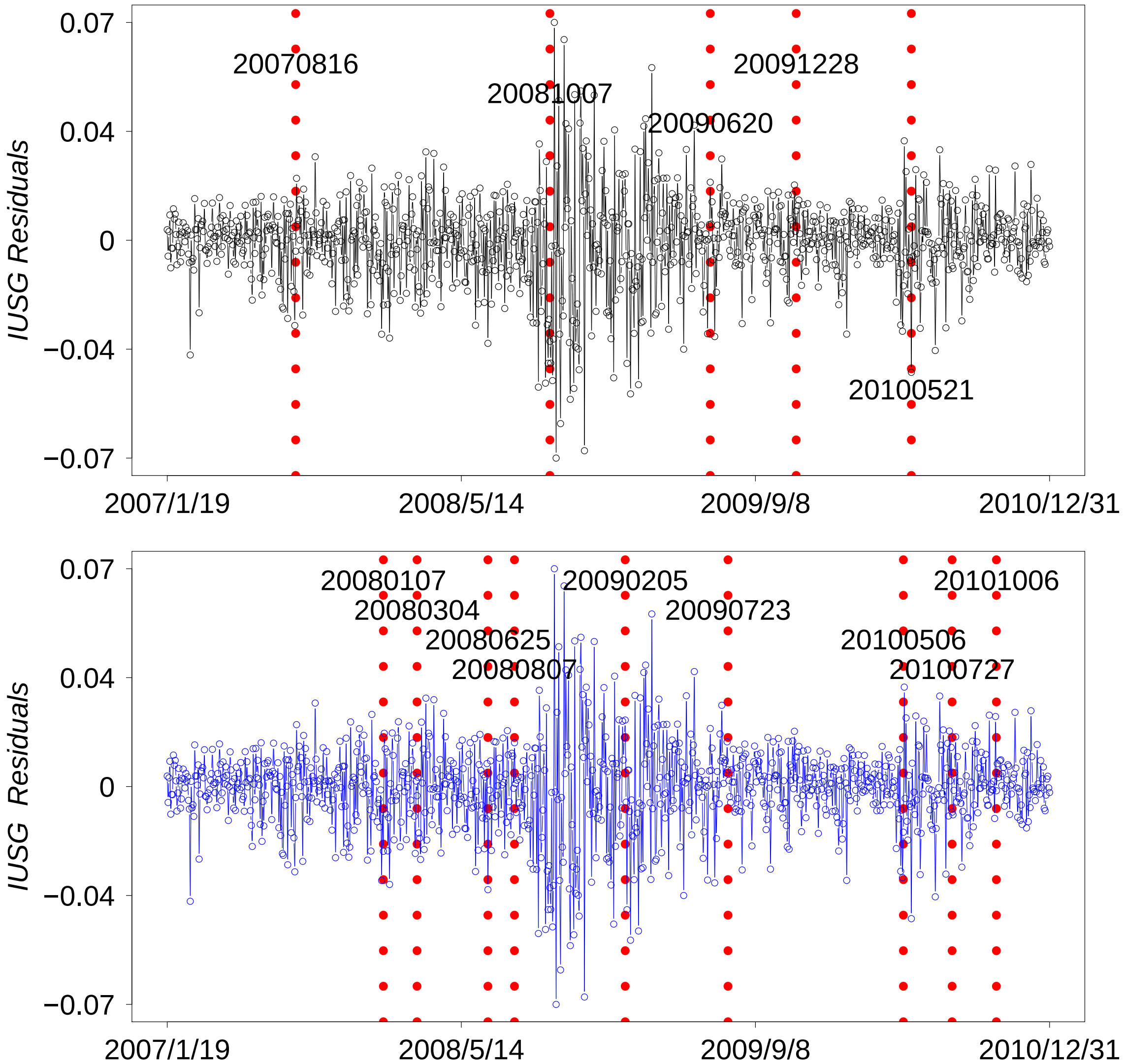

Figure 10: Change detection comparison between the proposed (upper panel) and the NPMVCP (lower panel) control charts for IUSG.

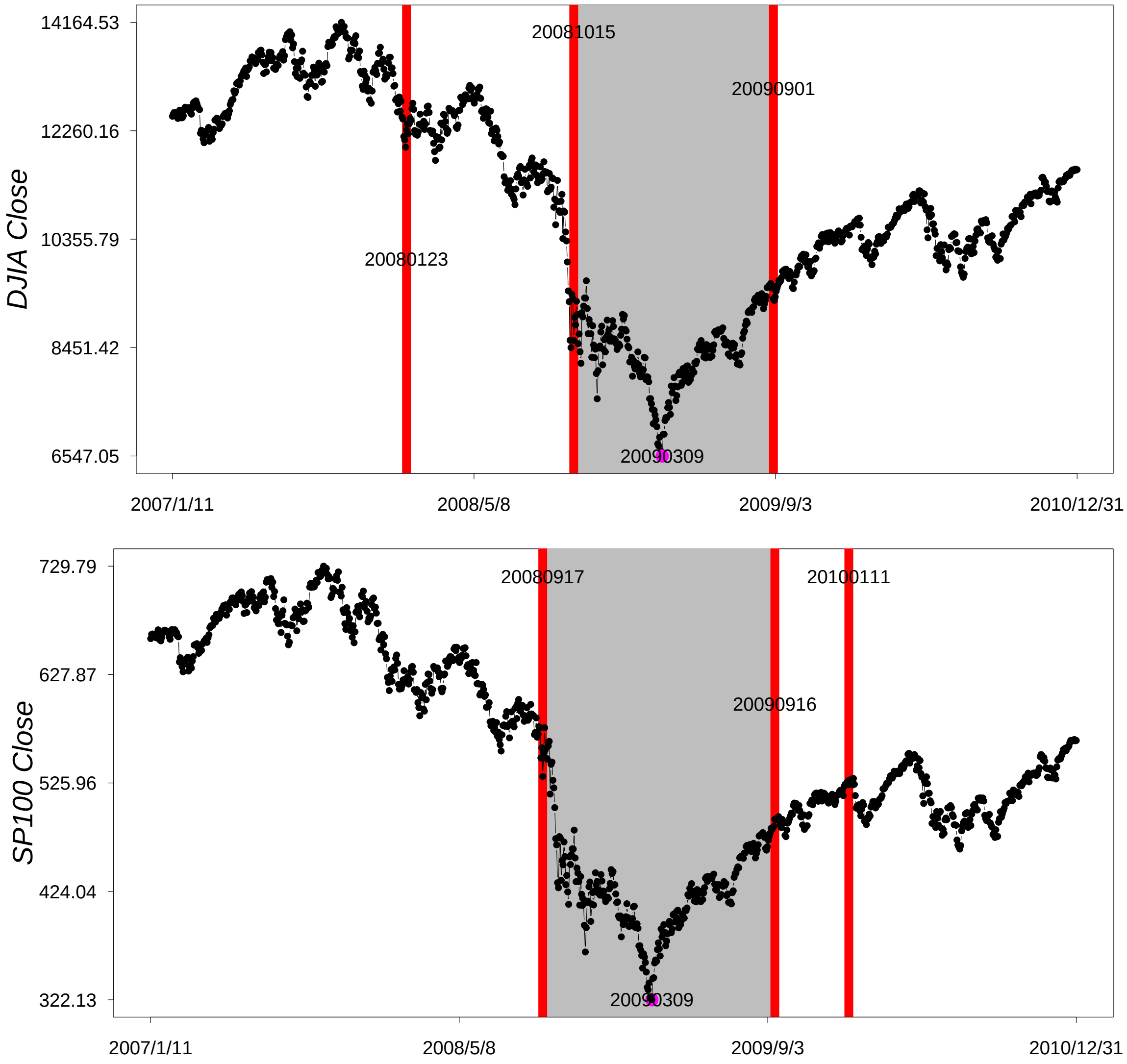

Figure 11: Proposed control chart for detection points of DJIA (upper) and SP100 (lower) data sets. The red line stands for the detection point. The pink point stands for the lowest point in each index.

5 Conclusion

This paper proposes a non-parametric multivariate control chart to detect the multiple change points in high-dimensional stream data (high dimensional financial time series). It has four features. Firstly, it is a non-parametric control chart requiring no assumption on the process, compared with the classical parametric control chart. Secondly, it is oriented to Phase II change point detection which is central for real time surveillance of stream data and can be applied extensively, e.g. in industrial quality control, finance, medical science, geology et al. Thirdly, the control charts is designed for multivariate time series, which is more practical and informative for catching the essence of data as a whole than uni-variate time series.

Last but the most important feature of the proposed control charts is that it monitors not only mean or only covariance, but monitors mean and covariance simultaneously, not separately.

In simulation study the mean and covariance shifts were investigated and the control chart has shown outstanding performance compared to the benchmark models. In real data application, the proposed control charts was implemented for surveillance of three high-dimensional portfolios in 5, 29 and 90 dimensions separately. The proposed control chart shows the capacity to detect changes of the market regimes from quiescent period to volatile period, which provides reference to financial investors to take measures for the in-crisis investment. An `R` package '`EnergyOnlineCPM`' for Phase II non-parametric multivariate statistical process control is contributed in this paper.

References

Chakraborti, S., Qiu, P. & Mukherjee, A. (2015). Editorial to the special issue: Nonparametric statistical process control charts, *Quality and Reliability Engineering International* **31**(1): 1–2.

Choi, K. & Marden, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance, *Journal of the American Statistical Association* **92**(440): 1581–1590.

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

Erdman, C., Emerson, J. W. et al. (2007). bcp: an R package for performing a Bayesian analysis of change point problems, *Journal of Statistical Software* **23**(3): 1–13.

Fisher, R. A. (1937). *The design of experiments*, Oliver And Boyd; Edinburgh; London.

Friedman, J. H. & Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests, *The Annals of Statistics* pp. 697–717.

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

Hawkins, D. M., Qiu, P. & Kang, C. W. (2003). The changepoint model for statistical process control, *Journal of Quality Technology* **35**(4): 355.

Hawkins, D. M. & Zamba, K. (2005a). A change-point model for a shift in variance, *Journal of Quality Technology* **37**(1): 21.

Hawkins, D. M. & Zamba, K. (2005b). Statistical process control for shifts in mean or variance using a changepoint formulation, *Technometrics* **47**(2): 164–173.

Holland, M. D. (2013). NPMVCP: Nonparametric multivariate change point model, *Reference manual* .
URL: *ftp://cran.r-project.org/pub/R/web/packages/NPMVCP/index.html*

Holland, M. & Hawkins, D. (2014). A control chart based on a nonparametric multivariate change-point model, *Journal of Quality Technology* **46**: 1975–1987.

Hosking, J. R. (1980). The multivariate portmanteau statistic, *Journal of the American Statistical Association* **75**(371): 602–608.

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.

Killick, R. & Eckley, I. (2011). Changepoint analysis with the changepoint package in R, *The R User Conference, useR! 2011 August 16-18 2011 University of Warwick, Coventry, UK*, p. 51.

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**: 2529–2536.

Knoth, S. (2016). spc: Statistical process control - collection of some useful functions, *Reference manual* .
URL: *https://cran.r-project.org/web/packages/spc/index.html*

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.

Page, E. (1954a). An improvement to wald's approximation for some properties of sequential tests, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 136–139.

Page, E. S. (1954b). Continuous inspection schemes, *Biometrika* **41**(1/2): 100–115.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations, *Supplement to the Journal of the Royal Statistical Society* **4**(1): 119–130.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: Iii. the analysis of variance test, *Biometrika* **29**(3/4): 322–335.

Qiu, P. (2017). Some perspectives on nonparametric statistical process control, **to appear in**, *Journal of Quality Technology* .

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. & Székely, G. J. (2016). Package 'energy', *R User's Manual* .

Roberts, S. (1959). Control chart tests based on geometric moving averages, *Technometrics* **1**(3): 239–250.

Ross, G. J. et al. (2013). Parametric and nonparametric sequential change detection in R: The cpm package, *Journal of Statistical Software* **78**.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*, ASQ Quality Press.

Shewhart, W. A. & Deming, W. E. (1939). *Statistical method from the viewpoint of quality control*, Courier Corporation.

Sims, C. A. (1980). Macroeconomics and reality, *Econometrica: Journal of the Econometric Society* pp. 1–48.

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.

Woodall, W. H. & Montgomery, D. C. (2014). Some current directions in the theory and application of statistical process monitoring, *Journal of Quality Technology* **46**(1): 78.

Xu, Y. F. (2017). *EnergyOnlineCPM: Distribution free multivariate control chart based on energy test.*
URL: *https://cran.r-project.org/web/packages/EnergyOnlineCPM/index.html*

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.

Zeileis, A., Leisch, F., Kleiber, C. & Hornik, K. (2005). Monitoring structural change in dynamic econometric models, *Journal of Applied Econometrics* **20**(1): 99–121.

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.

Zou, C., Wang, Z. & Tsung, F. (2012). A spatial rank-based multivariate ewma control chart, *Naval Research Logistics (NRL)* **59**(2): 91–110.

A Package ‘EnergyOnlineCPM’

For research of control chart, many R packages have been provided. In this section, an introduction of the proposed control chart based R package ‘EnergyOnlineCPM’ is presented. A review of mainstream control chart R packages, installation of package ‘EnergyOnlineCPM’ and an example of usage are given in the following.

Nowadays many packages are devised for control chart. We review some main packages based on R programming. Zeileis, Leisch, Kleiber & Hornik (2005) contributed ‘strucchange’ which is used for univariate change point analysis (Phase I) for mean monitoring. Erdman, Emerson et al. (2007) gave ‘bcp’ focused still on Phase I change point analysis for univariate data but used Bayesian method for mean surveillance. ‘changepoint’ in Killick & Eckley (2011) is used for mean or/and variance monitoring based on (non)parametric model in Phase I. ‘cpm’ in Ross et al. (2013) is used for Phase II analysis but only for univariate data set. ‘spc’ in Knoth (2016) collects some parametric control chart models using for Phase II monitoring of mean or/and variance. ‘NPMVCP’ in Holland (2013) is a package for multivariate data monitoring using a nonparametric change point model for surveillance of location changes. ‘ecp’ in James & Matteson (2015) is used for uni/multivariate Phase I data using a nonparametric model to surveillance distribution changes.

Energy statistic (Székely & Rizzo (2004)) is attracting attention for empirical discrepancy of characteristic functions. At the moment there are two R packages for energy statistic, James & Matteson (2015) and Rizzo & Székely (2016). Rizzo & Székely (2016) focused on the energy tests and James & Matteson (2015) concentrated on the Phase I change point model used for retrospective analysis. The package ‘EnergyOnlineCPM’ is the first package which centers on the online nonparametric change point model to monitor multiple change points for high dimensional time series based on the maximum energy test statistic using permutation samples.

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

```
install.packages("EnergyOnlineCPM")
library(EnergyOnlineCPM)
```

Next we show an example of using ‘EnergyOnlineCPM’ to detect a simulated data set with five dimensions. The data-driven-process is set as a process with three segments.

The first segment has 20 readings following $N(1_{5\times1}, I_{5\times5})$. The second segment has 30 observations following $N(2_{5\times1}, I_{5\times5})$. The third segment follows $N(1_{5\times1}, I_{5\times5})$, the same with the first segment, but has 50 observations. Therefore the 20-th and 50-th points are two theoretical change points. The task for 'EnergyOnlineCPM' is to detect these two points with least delayed steps. The script is given as follows.

```
library(MASS)
simNr = 300 # simulate 300 steps time series
# simulate 300 length 5 dimensional 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 steps 5 dimensional 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 2.
# last 40 points are from standard Gaussian.
data1 = rbind(sim2[1:20,], (sim3+2)[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)
```

After running the codes above, a plot (Figure 12) can be obtained, which shows the change points and detection points for the first univariate column in the five dimensional data set. The middle segment between blue lines shows a process with mean equal to 2, while the other two side-segments' means are all equal to 1. The two red lines give the detection points. The blue lines show the estimated change points. Installation, user manual, examples and more information can be referred to the user manual Xu (2017) and the project homepage: https://sites.google.com/site/energyonlinecpm.

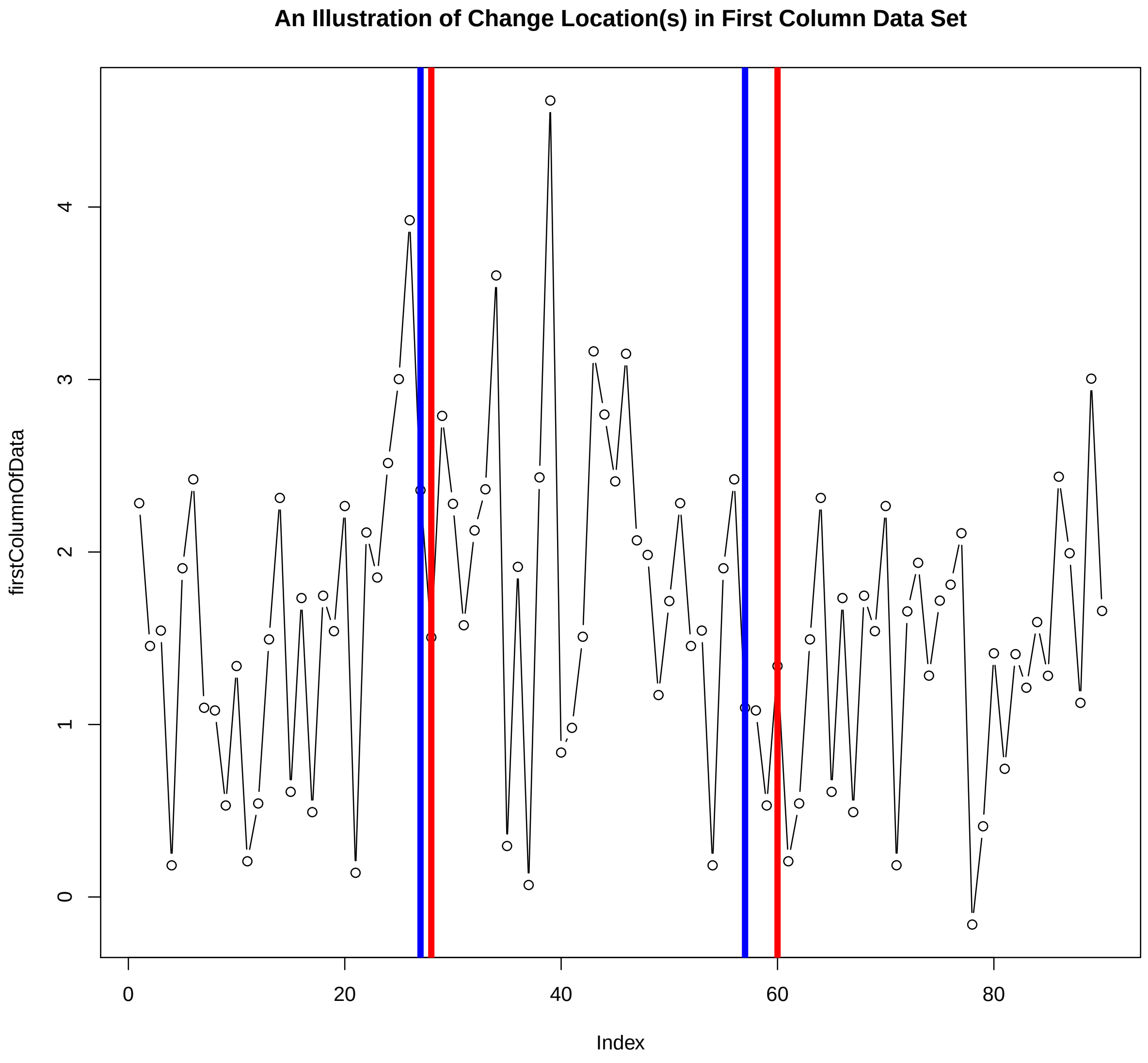

Figure 12: An example of change detection of a 5-dimensional data set. The blue line stands for the estimated change point and the red for the detection point.

B Information of Data Sets and Supplemental Tables

Symbol	Company
DGT	SPDR Global Dow ETF
EWD	iShares MSCI Sweden Capped ETF
GLD	SPDR Gold Trust
IGV	iShares Core S&P U.S. Growth ETF
IUSG	iShares North American Tech-Software ETF

Table 2: Related information of components of 5-dimensional data set of ETFs.

Company	Exchange	Symbol	Industry
Apple	NASDAQ	AAPL	Consumer electronics
American Express	NYSE	AXP	Consumer finance
Boeing	NYSE	BA	Aerospace anddefense
Caterpillar	NYSE	CAT	Construction andmining equipment
Cisco Systems	NASDAQ	CSCO	Computer networking
Chevron	NYSE	CVX	Oil & gas
DuPont	NYSE	DD	Chemical industry
Walt Disney	NYSE	DIS	Broadcasting andentertainment
General Electric	NYSE	GE	Conglomerate
Goldman Sachs	NYSE	GS	Banking,Financial services
The Home Depot	NYSE	HD	Home improvementretailer
IBM	NYSE	IBM	Computers andtechnology
Intel	NASDAQ	INTC	Semiconductors
Johnson & Johnson	NYSE	JNJ	Pharmaceuticals
JPMorgan Chase	NYSE	JPM	Banking
Coca-Cola	NYSE	KO	Beverages
McDonald's	NYSE	MCD	Fast food
3M	NYSE	MMM	Conglomerate
Merck	NYSE	MRK	Pharmaceuticals
Microsoft	NASDAQ	MSFT	Software
Nike	NYSE	NKE	Apparel
Pfizer	NYSE	PFE	Pharmaceuticals
Procter & Gamble	NYSE	PG	Consumer goods
Travelers	NYSE	TRV	Insurance
UnitedHealth Group	NYSE	UNH	Managed health care
United Technologies	NYSE	UTX	Conglomerate
Verizon	NYSE	VZ	Telecommunication
Walmart	NYSE	WMT	Retail
ExxonMobil	NYSE	XOM	Oil & gas

Table 3: Related information of components of 29-dimensional data set from DJIA.

Symbol	Company	Symbol	Company	Symbol	Company
ABT	Abbott Laboratories	EMR	Emerson Electric Co.	MS	Morgan Stanley
ACN	Accenture plc	EXC	Exelon	MSFT	Microsoft
AGN	Allergan plc	F	Ford Motor	NEE	NextEra Energy
AIG	American International Group Inc.	FDX	FedEx	NKE	Nike
ALL	Allstate Corp.	FOX	21st Century Fox	ORCL	Oracle Corporation
AMGN	Amgen Inc.	GD	General Dynamics	OXY	Occidental Petroleum Corp.
AMZN	Amazon.com	GE	General Electric Co.	PCLN	Priceline Group Inc/The
AXP	American Express Inc.	GILD	Gilead Sciences	PEP	Pepsico Inc.
BA	Boeing Co.	GOOG	Alphabet Inc	PFE	Pfizer Inc
BAC	Bank of America Corp	GS	Goldman Sachs	PG	Procter & Gamble Co
BIIB	Biogen Idec	HAL	Halliburton	QCOM	Qualcomm Inc.
BK	The Bank of New York Mellon	HD	Home Depot	RTN	Raytheon Company
BLK	BlackRock Inc	HON	Honeywell	SBUX	Starbucks Corporation
BMY	Bristol-Myers Squibb	IBM	International Business Machines	SLB	Schlumberger
C	Citigroup Inc	INTC	Intel Corporation	SO	Southern Company
CAT	Caterpillar Inc	JNJ	Johnson & Johnson Inc	SPG	Simon Property Group, Inc.
CELG	Celgene Corp	JPM	JP Morgan Chase & Co	T	AT&T Inc
CL	Colgate-Palmolive Co.	KO	The Coca-Cola Company	TGT	Target Corp.
CMCSA	Comcast Corporation	LLY	Eli Lilly and Company	TWX	Time Warner Inc.
COF	Capital One Financial Corp.	LMT	Lockheed-Martin	TXN	Texas Instruments
COP	ConocoPhillips	LOW	Lowe's	UNH	UnitedHealth Group Inc.
COST	Costco	MA	MasterCard Inc	UNP	Union Pacific Corp.
CSCO	Cisco Systems	MCD	McDonald's Corp	UPS	United Parcel Service Inc
CVS	CVS Health	MDLZ	Mondelez International	USB	US Bancorp
CVX	Chevron	MDT	Medtronic Inc.	UTX	United Technologies Corp
DD	DuPont	MET	Metlife Inc.	VZ	Verizon Communications Inc
DHR	Danaher	MMM	3M Company	WBA	Walgreens Boots Alliance
DIS	The Walt Disney Company	MO	Altria Group	WFC	Wells Fargo
DOW	Dow Chemical	MON	Monsanto	WMT	Wal-Mart
DUK	Duke Energy	MRK	Merck & Co.	XOM	Exxon Mobil Corp

Table 4: Related information of components of 90-dimensional data set from S&P100.

		Mean Gaussian Shift				Mean t Shift				Mean Laplace Shift			
		ARL_o=200		ARL_o=100		ARL_o=200		ARL_o=100		ARL_o=200		ARL_o=100	
Dimensions	δ	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP
3	0	182.36	100.48	95.54	70.24	182.56	98.46	90.07	75.50	187.84	114.12	93.30	65.36
	0.25	148.75	**70.12**	69.06	**44.54**	166.55	**102.04**	75.20	**41.26**	184.58	**141.88**	93.22	**57.22**
	0.50	50.69	**28.62**	33.40	**20.08**	40.31	**21.86**	29.82	**15.28**	181.26	**117.92**	89.22	**59.66**
	0.75	17.82	**14.42**	15.58	**11.64**	14.45	**14.02**	13.29	**10.36**	161.56	**112.22**	83.76	**60.74**
	1	4.17	10.86	13.53	**9.06**	12.63	**11.10**	10.90	**9.18**	143.72	**97.40**	76.62	**38.50**
	2	6.18	9.20	8.31	**7.46**	7.58	8.84	7.65	8.00	63.46	**43.54**	30.78	**26.10**
	3	7.97	8.20	4.67	7.48	7.00	7.20	6.66	7.64	16.62	16.66	19.58	**13.98**
	4	7.27	8.64	6.00	7.04	6.63	7.56	5.90	6.50	18.61	**11.98**	13.42	**10.40**
	5	5.95	8.06	5.65	7.18	6.28	7.98	6.03	7.42	11.23	11.78	10.17	10.26
	6	6.89	7.94	6.22	6.92	6.33	8.22	5.43	7.40	10.28	10.54	9.05	9.16
	7	7.08	8.28	6.51	6.72	6.15	8.22	5.60	6.90	9.12	9.60	8.81	**8.44**
	8	6.65	8.24	6.29	7.16	6.18	8.18	5.59	7.48	8.60	9.82	8.27	8.78
	9	4.48	7.56	5.95	6.96	5.98	7.78	5.46	7.08	8.27	9.78	7.71	8.34
10	0	195.46	93.74	91.13	60.28	179.26	74.28	88.38	60.72	183.77	99.64	93.98	65.86
	0.25	98.29	**52.02**	53.54	**29.76**	82.53	**64.86**	52.45	**32.82**	178.21	146.68	93.73	**73.82**
	0.50	17.74	**15.22**	15.51	**14.34**	13.68	14.22	12.86	14.44	178.88	138.26	73.94	74.44
	0.75	10.90	12.18	9.94	11.42	10.52	12.64	9.25	11.54	158.58	122.72	76.83	**66.64**
	1	8.97	11.46	10.27	**8.88**	8.68	11.26	7.97	9.94	148.20	119.28	76.71	**52.14**
	2	6.56	9.70	7.90	8.54	6.81	9.76	6.26	8.94	58.41	71.94	35.18	43.94
	3	6.09	9.94	6.38	8.56	6.27	9.40	6.01	8.24	22.38	23.90	15.70	20.54
	4	6.65	9.52	5.59	8.70	6.21	9.68	5.29	8.02	12.68	19.88	13.76	14.56
	5	6.35	9.58	6.01	8.42	6.17	9.36	5.44	8.06	10.41	13.98	9.97	13.04
	6	6.76	9.64	6.23	8.08	6.20	9.50	5.34	8.16	10.23	13.70	8.27	12.52
	7	6.75	9.72	5.61	8.68	6.05	8.78	5.43	8.60	9.37	12.38	8.64	11.74
	8	6.73	9.42	4.59	8.20	5.87	9.56	5.61	8.00	8.38	12.76	7.66	11.26
	9	6.35	9.30	5.34	7.84	5.85	9.92	5.31	8.10	8.00	12.30	7.31	10.98

Table 5: Mean shift in standard Gaussian, Student-t_5 and Laplace cases. The outperformed points of NPMVCP compared with the proposed control chart are in bold. The in-control length is set as 32 and out-of-control as 100 and 200, and change point $\tau = 32$.

		Gaussian Covariance Shift				t Covariance Shift			
		ARL_o=200		ARL_o=100		ARL_o=200		ARL_o=100	
Dimensions	σ^2	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP
3	0.25	43.24	158.30	37.66	77.50	63.62	143.94	46.64	75.90
	0.50	165.77	**136.22**	80.88	**68.80**	170.91	**141.86**	83.84	**59.78**
	0.75	192.96	**135.38**	89.96	**69.94**	192.47	**133.46**	97.38	**71.70**
	2	157.84	**130.32**	77.56	56.12	166.64	**105.92**	73.98	**72.10**
	3	125.76	**110.62**	61.44	49.40	114.27	**76.34**	50.98	58.78
	4	65.78	95.38	39.02	50.68	65.87	112.88	36.70	48.00
	5	14.92	88.60	10.91	40.00	52.58	110.50	27.85	41.10
	6	15.63	104.22	15.72	50.86	21.24	82.26	19.66	43.52
	7	19.32	90.08	12.94	42.06	17.38	80.22	16.02	44.50
	8	13.65	99.78	10.35	43.52	14.16	97.18	13.25	49.30
	9	13.88	89.04	11.40	43.32	14.54	88.78	12.31	50.72
	10	13.84	89.74	12.01	44.82	12.31	59.92	12.19	49.10
	11	8.95	82.02	16.02	41.70	12.14	89.34	10.96	38.24
10	0.25	19.06	153.84	16.58	76.56	21.48	141.76	18.64	77.62
	0.50	105.14	133.10	72.38	**71.44**	110.68	153.04	69.84	80.04
	0.75	179.82	**142.72**	91.56	**64.14**	171.08	**135.62**	94.78	**69.20**
	2	126.20	**104.52**	64.02	64.16	101.82	129.78	57.32	69.36
	3	10.66	96.72	11.10	60.38	39.60	98.82	28.06	61.66
	4	11.96	102.84	13.04	47.60	15.98	105.12	13.14	59.94
	5	9.91	88.72	10.15	48.60	12.86	76.44	11.32	60.72
	6	9.61	76.84	6.66	44.72	10.88	106.94	10.25	50.64
	7	9.61	67.26	9.19	48.92	10.15	94.66	9.37	47.60
	8	9.91	76.82	8.09	43.16	9.83	83.68	9.40	42.10
	9	10.02	77.98	6.90	39.48	9.73	69.28	8.46	43.76
	10	9.11	74.90	10.36	39.38	8.96	90.80	8.15	47.56
	11	6.49	85.68	6.71	40.78	8.94	77.72	7.78	45.74

Table 6: Covariance shift in Gaussian and Student-t_5 cases. The outperformed points of NPMVCP compared with the proposed control chart are in bold. The in-control length is set as 32 and out-of-control as 100 and 200, and change point $\tau = 32$.

		Gaussian Mean Shift				Student t Mean Shift				Laplace Mean Shift			
		$ARL_o = 200$		$ARL_o = 100$		$ARL_o = 200$		$ARL_o = 100$		$ARL_o = 200$		$ARL_o = 100$	
Dimensions	δ	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP	Proposed	NPMVCP
3	0.25	157.26	121.82	95.54	70.24	151.82	87.72	62.02	52.36	197.93	121.02	90.16	58.24
	0.5	115.36	97.72	63.36	65.08	116.18	76.62	51.64	34.96	184.61	104.70	91.22	40.72
	0.75	52.54	54.22	53.15	36.42	64.36	28.36	28.38	20.56	177.69	37.62	92.98	30.14
	1	29.98	17.46	40.98	30.20	24.68	17.60	19.50	15.46	170.64	18.86	85.08	14.08
	2	11.00	9.36	22.76	13.94	10.70	10.20	10.04	8.88	128.30	11.48	66.72	9.34
	3	8.92	9.22	10.04	8.98	8.35	8.80	7.67	8.06	40.85	9.26	30.38	7.80
	6	6.81	8.16	8.16	7.90	6.85	8.46	6.24	7.80	11.49	7.76	10.48	7.10
	9	6.18	7.82	6.08	6.94	6.20	7.98	5.88	6.88	8.96	7.72	8.80	6.78
10	0.25	158.56	138.72	62.52	67.50	147.69	132.30	64.00	71.06	179.80	149.96	88.09	63.32
	0.5	130.30	89.02	60.70	52.36	136.47	87.58	58.22	51.70	185.72	112.28	95.12	41.22
	0.75	102.92	75.80	45.36	35.82	77.60	59.78	53.24	37.64	179.38	61.16	88.80	27.26
	1	49.32	43.48	36.12	26.30	48.02	33.86	28.98	18.38	184.47	31.00	89.42	21.62
	2	12.10	13.98	12.04	11.80	11.56	13.34	11.39	11.56	160.52	13.00	85.88	11.90
	3	9.54	11.72	9.43	10.86	9.46	11.66	8.51	9.96	157.46	11.58	78.98	10.40
	6	7.12	9.82	6.53	8.84	6.96	10.06	6.33	8.90	19.66	10.32	16.84	9.04
	9	6.32	9.32	5.68	8.90	6.42	10.06	5.78	8.46	12.04	10.02	11.28	8.40

Table 7: Single mean shift in Gaussian, Student-t_5 and Laplace cases. The in-control length is set as 32 and out-of-control as 100 and 200, and change point $\tau = 32$.

		Gaussian						t_5			$Gamma_5$			Mix		
		SMMST		SREWMA		Proposed		SMMST	SREWMA	Proposed	SMMST	SREWMA	Proposed	SMMST	SREWMA	Proposed
dim	δ	$\tau = 40$	$\tau = 90$	$\tau = 40$	$\tau = 90$	$\tau = 40$	$\tau = 90$	$\tau = 90$	$\tau = 90$	$\tau = 90$	$\tau = 90$	$\tau = 90$	$\tau = 90$	$\tau = 90$	$\tau = 90$	$\tau = 90$
5	1	14.20	12.80	11.40	9.93	8.76	8.45	19.80	13.20	5.69	14.30	10.70	19.09	19.80	12.40	16.27
	1.5	7.35	6.46	7.69	6.95	5.26	4.94	9.13	9.28	4.63	7.64	8.03	9.48	8.91	9.04	9.72
	2	4.97	4.95	6.39	5.77	4.06	3.76	7.43	7.87	3.73	5.01	6.89	5.38	7.06	7.37	6.75
	3	3.59	3.35	5.42	4.89	3.20	3.06	4.20	6.35	3.85	3.45	6.01	4.40	4.50	6.18	5.03
	4	3.38	3.29	5.10	4.59	3.22	2.64	3.52	5.84	2.76	3.06	5.69	3.50	3.48	5.74	3.44
		$\tau = 40$	$\tau = 90$	$\tau = 40$	$\tau = 90$	$\tau = 40$	$\tau = 90$	$\tau = 80$	$\tau = 80$	$\tau = 80$	$\tau = 80$	$\tau = 80$	$\tau = 80$	$\tau = 80$	$\tau = 80$	$\tau = 80$
10	1	11.00	10.10	9.57	8.77	4.96	3.66	12.80	11.70	4.22	11.50	9.73	10.70	18.20	10.00	13.25
	1.5	6.26	6.13	6.80	6.27	4.98	3.80	7.72	8.41	5.58	5.97	7.32	7.25	7.36	7.48	8.15
	2	4.27	4.13	5.73	5.25	4.78	3.88	5.84	7.01	4.38	4.13	6.34	6.02	5.78	6.34	6.71
	3	3.85	3.81	4.89	4.44	4.82	3.76	4.21	5.80	3.60	3.83	5.52	4.67	3.90	5.47	4.96
	4	3.61	3.49	4.60	4.17	4.72	3.32	3.90	5.35	3.40	3.32	5.23	4.89	3.74	5.13	4.17

Table 8: Out-of-control ARLs' comparison between the proposed control chart and the SMMST and the SREWMA control charts in context of Gaussian, t_5, $Gamma_3$ and mix-component distribution mean shift. The performance of the SMMST and the SREWMA control charts is based on the Table 2, 3, 4, 5 in Zhou et al. (2015).