CHANGE-POINT DETECTION

Consider a time series data set consisting of n normally distributed observations. And suppose that the first segment of k observations has a $N(\mu_1, \sigma^2)$ distribution whereas the remaining segment of $n-k$ observations has a $N(\mu_2, \sigma^2)$ distribution where $\mu_1 \neq \mu_2$. That is, a change in mean occurs at some unknown step k. The **change-point detection** is a collection of methods to identify the value of k.

The change-point detection problem is also applicable to finding k when the mean doesn't change but the variance does, or when both mean and variance change. It also extends to the case of several segments with different means, variances, or both.

There are many well-developed methods to identify the point(s) of change. We will present the theory for the most basic approach.

The method of binary segmentation is often used to detect the change points. First, one change point is detected in the complete set of observations, then the series is split around this change point, and the algorithm is applied to the two resulting segments. The process continues until a pre-specified number of splits is detected.

To identify the value of k where the change occurs, the method of maximum likelihood estimation is employed. We assume that $y_1, \dots, y_k \sim N(\mu_1, \sigma^2)$, and $y_{k+1}, \dots, y_n \sim N(\mu_2, \sigma^2)$. The maximum likelihood estimators of μ_1, μ_2, and σ^2 are

$$\hat{\mu}_1 = \bar{y}_1 = \frac{1}{k}\sum_{i=1}^{k} y_i, \quad \hat{\mu}_2 = \bar{y}_2 = \frac{1}{n-k}\sum_{i=k+1}^{n} y_i,$$

and

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(y_i - \bar{y}\big)^2, \text{ where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The likelihood function for these data has the form:

$$L = L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2 \,|\, y_1, \dots, y_n) = (2\pi\,\hat{\sigma}^2)^{-n/2} \exp\Big\{ - \frac{\sum_{i=1}^{k}(y_i - \bar{y}_1)^2 + \sum_{i=k+1}^{n}(y_i - \bar{y}_2)^2}{2\,\hat{\sigma}^2}\Big\}.$$

The value of k that maximizes the likelihood function is the optimal one.

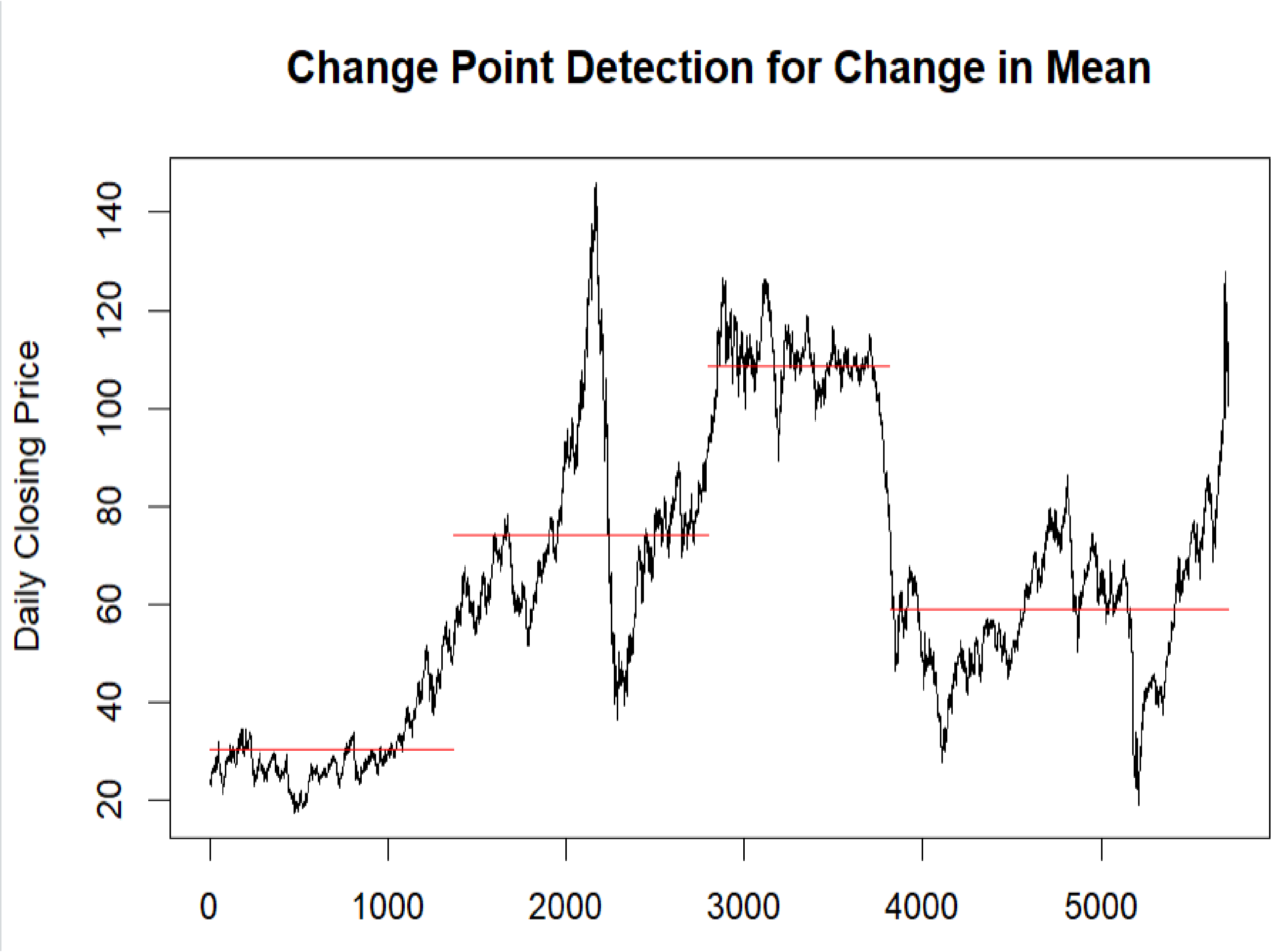
To apply the change-point detection to real-life data, we will use the library "changepoint" in R, and utilize functions cpt.mean(), cpt.var(), and cpt.meanvar() with options `method="BinSeg"` (binary segmentation), `Q=` (the number of splits k), and `penalty="AIC"`. Here AIC stands for Akaike Information Criterion which dictates choosing k that minimizes $AIC = -2\ln L + p\ln n$ where p is the number of parameters that have to be estimated from the data (in our example, we estimated

μ_1, μ_2 and σ^2, so $p = 3$).

Example. The file "crudeoil_data.csv" contains daily closing prices of (Brent) crude oil between 2000 and 2022. These data were extracted from the file “commodity 2000-2022.csv” downloaded from kaggle.com. We apply change-point methods to identify changes in mean, changes in variance, and simultaneous changes in mean and variance.

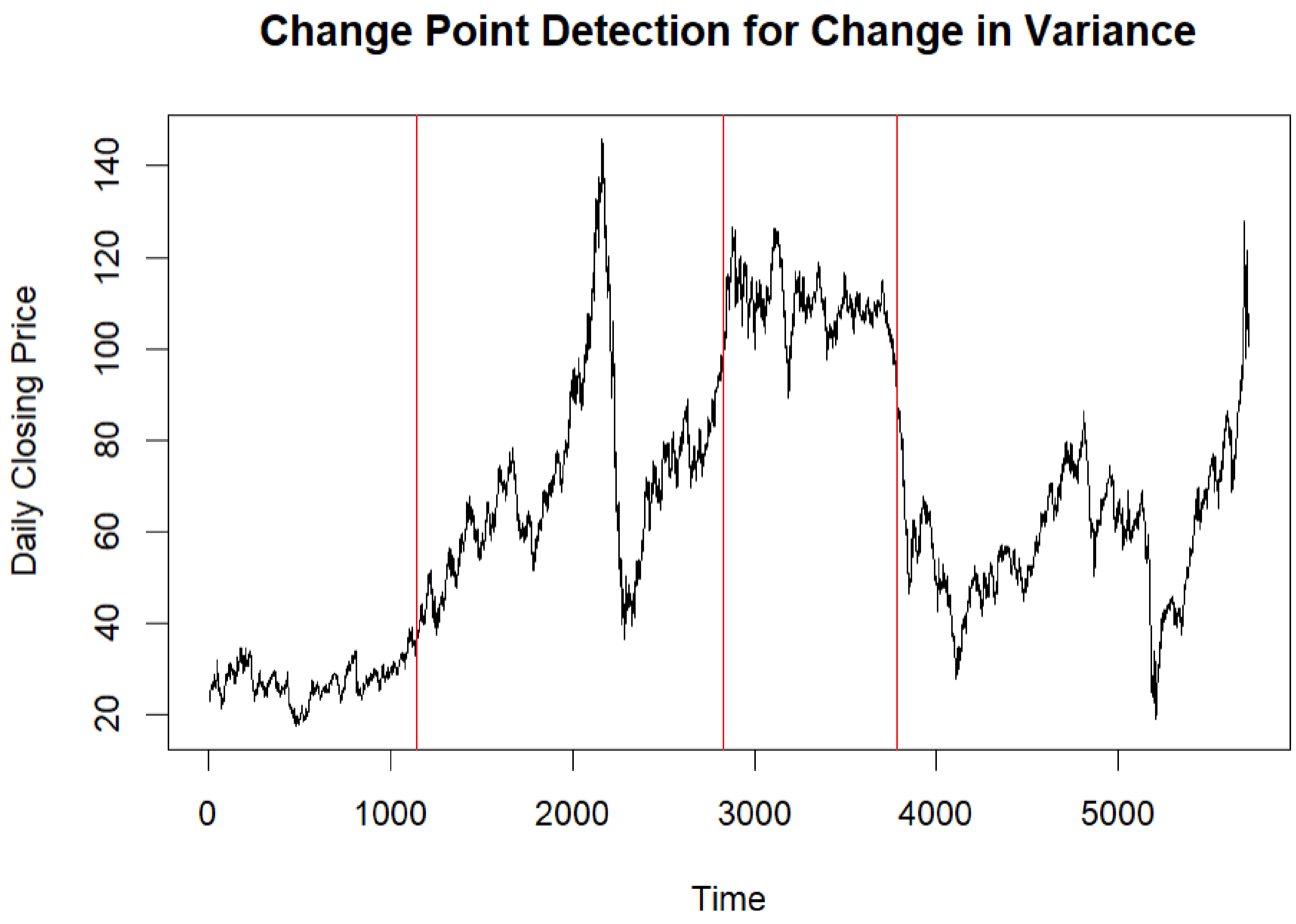
```
#application to crude oil’s closing price data
crudeoil.data<- read.csv(file="./crudeoil_data.csv", header=TRUE, sep=",")

library(changepoint)
ansmean=cpt.mean(crudeoil.data$Close, penalty="AIC", method="BinSeg", Q=3)
plot(ansmean, cpt.col="red", ylab="Daily Closing Price", main="Change Point Detection for Change in Mean")
print(ansmean)
```



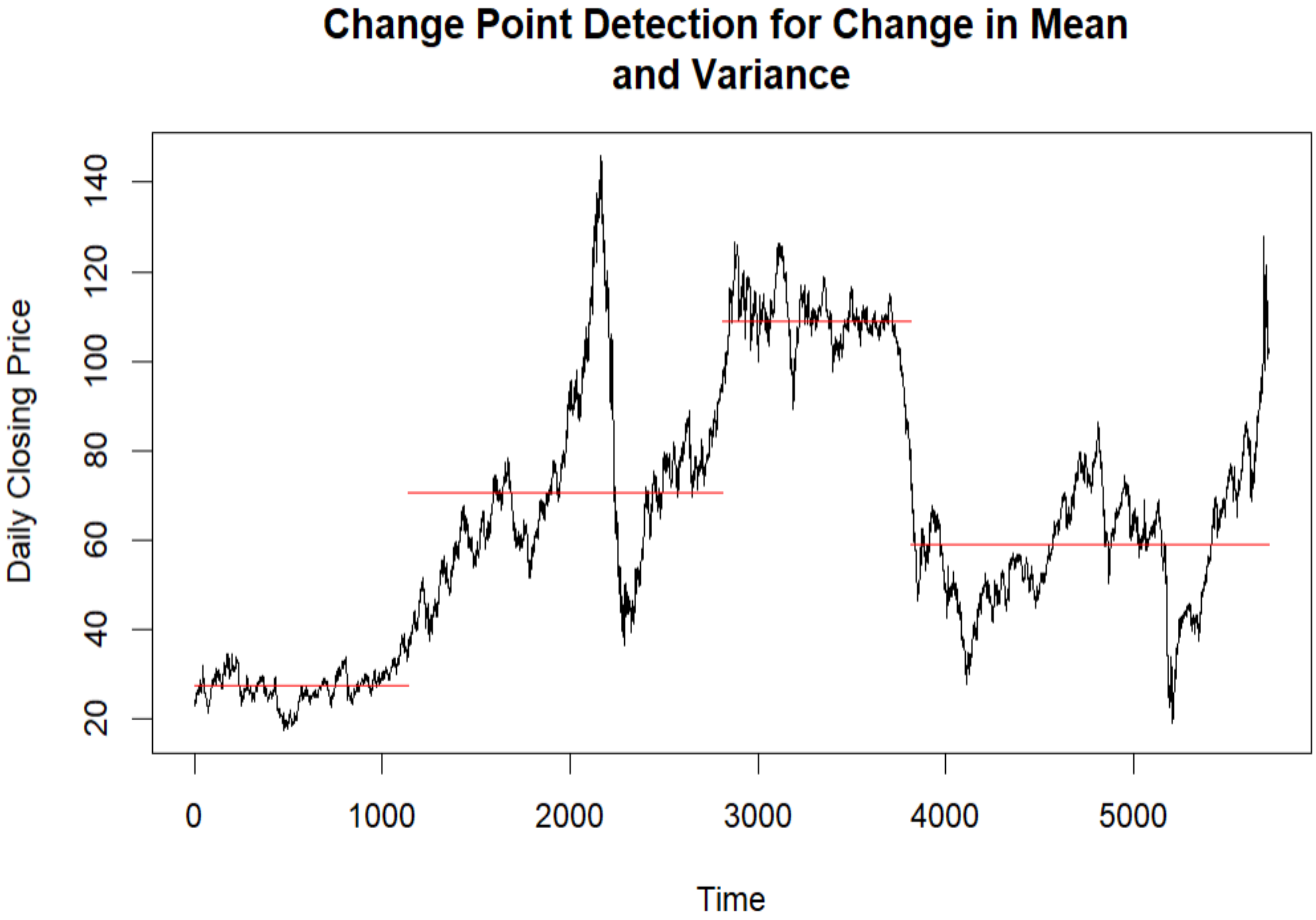
```
Changepoint Locations : 1368 2794 3816
```

```
ansvar=cpt.var(crudeoil.data$Close, penalty="AIC", method="BinSeg", Q=3)
plot(ansvar, cpt.col="red", ylab="Daily Closing Price", main="Change Point Detection for Change
in Variance")
print(ansvar)
```



```
Changepoint Locations : 1144 2827 3784
```

```
ansmeanvar=cpt.meanvar(crudeoil.data$Close, penalty="AIC", method="BinSeq", Q=3)
plot(ansmeanvar, cpt.col="red", ylab="Daily Closing Price", main="Change Point Detection for
Change in Mean and Variance")
print(ansmeanvar)
```



```
Changepoint Locations : 1140 2812 3816
```

□