

## 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  $\mu_1$ ,  $\mu_2$ , and  $\sigma^2$  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

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

The likelihood function for these data has the form:

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

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