In many real applications such as biology, economics, quality control, an important problem is to identify the change point of a time series. In the simplest case, the change point detection problem may be described in mathematics as the following. Suppose that $\{\epsilon_t: t = 1, \cdots n\}$ is a mean zero stationary time series. Suppose that $\{Y_t: t = 1, \cdots, n\}$ is another time series satisfying $Y_t = \mu_k + \epsilon_t$ for $t \in (\tau_k, \tau_{k+1}]$, with $0 = \tau_0 < \tau_1 < \cdots < \tau_{K+1} = n$ and $\mu_k \neq \mu_{k+1}$. In general, we can only observe Y_t. The change points τ_k , the changes μ_k $(k = 1, \cdots K)$ and even K are also generally unknown. The goal of this type of analysis is to develop statistical methods that can accurately identify the change points τ_k, precisely estimate μ_k and K (See Figure 1).

In a little bit more complex setting, Y_t may be written as $Y_t = \mu_k + \sigma_k \epsilon_t$, where the unknown parameters σ_k can be the same or different for different k. More generally, Y_t may also depend on some other observable covariates X_t, i.e. $Y_t = g(X_t) + \mu_k + \sigma_k \epsilon_t$, where the function g is unknown but usually can be assumed to belong to a function family (e.g. linear functions). In such models, in order to accurately detect the change points, we also need to estimate the unknown function g and the unknown constants σ_k.

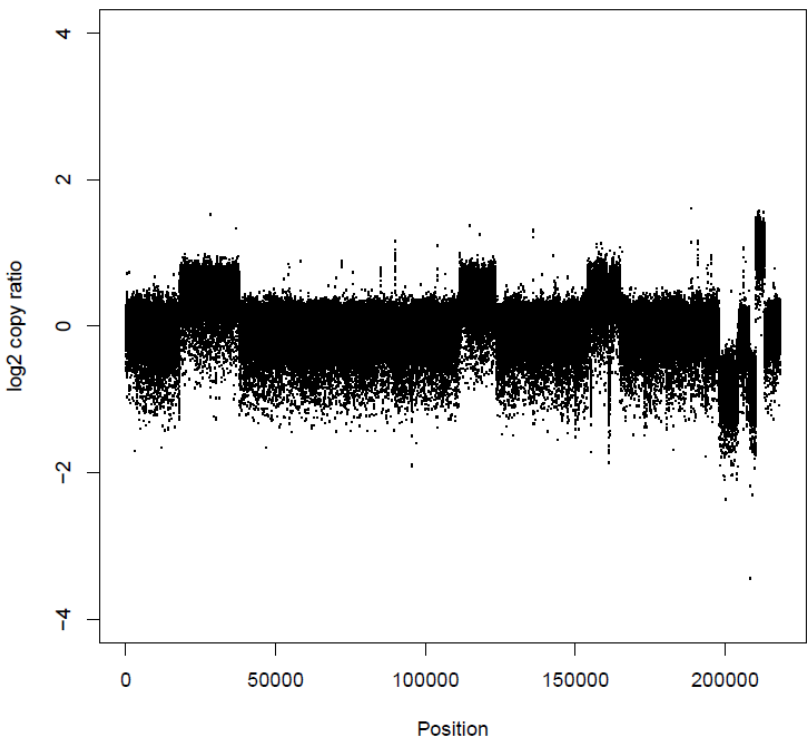

Figure 1: The scatter plot of the raw data of the project; We can clearly see some change point from the plot. This data is processed from the data of a genomic study

Aim of the project:

1. Develop a statistical method that can efficiently and accurately detect the change point in the model $Y_t = \mu_k + \epsilon_t$. You may assume ϵ_t being a Gaussian ARMA(p, q) time series (p and q are known), i.e. ϵ_t satisfies $A(\mathcal{B})\epsilon_t = B(\mathcal{B})\eta_t$ with η_t being the normal white noise and A, B as defined in the textbook . Use simulations to show that your algorithm works. In the simulation, you will need to consider a few settings that you think are typical in real applications and generate at least 100 simulated data for each setting. You also need to design a criterion to evaluate the performance of your method. Justify why you think this criterion is appropriate for the problem. Apply your method to the data as given in this project. Evaluate the performance of your method on this real data set. Plot your estimated changes μ_k and the raw data at the same plot. If possible (optional), prove that your method has some good asymptotic property.
2. Develop a statistical method for change point detection in the model $Y_t = \beta^T X_t + \mu_k + \sigma_k \epsilon_t$. Use simulation to show that your algorithm works.