Change Point Detection

Weilu Zhao(weilu.zhao@uth.tmc.edu (mailto:weilu.zhao@uth.tmc.edu)), Wang Qian(qian.wang@uth.tmc.edu (mailto:qian.wang@uth.tmc.edu))

4/28/2022

1.Introduction

In linear models, one of the major assumptions is that the model holds for the entire dataset. However, in practice, it is common to encounter such a scenario where the structure (coefficients) of the model changes when the data reaches a certain threshold, which may or may not be known beforehand. Such point model coefficients change drastically is called change point. Some representative examples of change points in linear models can be found in studies regarding Alzheimer disease. Cognitive ability is commonly used as the indicator for Alzheimer disease. Physicians noticed that when patients are in their youth and middle age, the cognitive ability changes slowly. However, when the age of patients passes a certain threshold, the speed of the degeneration of cognitive ability accelerates, indicating a drastic change in the coefficient of age with respect to cognitive ability in the linear model. In this case, that age threshold is considered as the change point of the linear model for that patient.
Change points are widely seen in the real world. Therefore, the development of techniques to detect change points in a linear model has been a popular topic. Studies regarding change point detection have been focusing on 3 major area: determining the number of change point; determining the location of the change point; and determining the coefficient of linear model for each segment. This study discusses the techniques for

In general, change point can be divided into two types: Discontinuous an continuous change point. Discontinuous change point refers to the change point such that both

In this literature review, we discussed the fundamental theories of change point detection in linear models. This literature review is organized in the following way: First, discuss the methods to detect change point in ordinary linear model. Then, we extend the change point detection method to generalized linear model. Finally, we discuss the change point detection in certain generalized linear model.

2. Change Point Detection in ordinary linear models

2.1 Simple linear regression

First, we consider a simple linear regression with one change point. We will discuss the discontinuous change point models first. Discontinuous change point models are models with no continuity constraints at the change points. Thus, the models of all segments are not restricted to common values at the change points. Consequently, for a known change point, the models of each segment are autonomous and all parameters in the linear predictor can be estimated separately. An unknown change point can be either estimated by a simple grid search over all feasible possibilities or by analyzing recursive residuals. The second method is appropriate if there is only one change point in the model, or the number of change points is small with respect to the sample size. This section considers change point models with one discontinuous change point. After introducing such a model for ordinary linear models(OLMs), it is then generalized for the wider class of GLMs. Finally, recursive residuals for OLMs and GLMs are introduced, which can be used to estimate the change point as well as to test the necessity of a change point.

Consider a simple linear regression model with a discontinuous change at a fixed but unknown change point. Let $(x_i, y_i), i = 1, \ldots, n$, denote pairs of observations, where y_i is the response and x_i some explanatory variable. Let us further assume that such n pairs (x_i, y_i) of observations can be arranged in some natural ordering. In this article, if not quoted otherwise, the index i describes this kind of order. Thus, the change point τ is given by any index i and determines the observation x_τ , after which the structural change in the relationship between x_i and y_i might occur. The change point τ partitions the data into two separate segments, in which the mean structure as well as the variance may be different. In fact, the first τ observations in a sample of size n follow one OLM and the last $n - \tau$ observations follow another OLM. The linear parameters of these two models are $\boldsymbol{\beta_1} = (\beta_{10}, \beta_{11})^T$ and $\boldsymbol{\beta_2} = (\beta_{20}, \beta_{21})^T$, respectively. Then, such an OLM can be written as

$$y_i = \begin{cases} \beta_{10} + x_i\beta_{11} + \varepsilon_{1i} & i = 1, \ldots, \tau \\ \beta_{20} + x_i\beta_{21} + \varepsilon_{2i} & i = \tau + 1, \ldots, n, \end{cases} \quad (2.1)$$

where the errors ε_{di} are independent random variables and follow a normal distribution with zero mean and variance σ_1^2 for $i \le \tau$ and variance σ_2^2 for $i > \tau$, i.e. $\varepsilon_{1i} \overset{iid}{\sim} N(0, \sigma_1^2)$ and $\varepsilon_{2i} \overset{iid}{\sim} N(0, \sigma_2^2)$, respectively. Such a model was first considered by Quandt (1958). He introduced a maximum likelihood (ML) method for estimating the unknown parameters $\boldsymbol{\beta} = (\boldsymbol{\beta_1}^T, \boldsymbol{\beta_2}^T)^T$, $\sigma^2 = \sigma_1^2 = \sigma_2^2$ and τ , which we describe in the following paragraph.

The parameters of interest are the linear parameter $\boldsymbol{\beta}$ and the change point τ . To guarantee the estimable of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2)^T$, possible values of τ are restricted to $\{3, 4, \ldots, n - 3\}$. To estimate these parameters with the ML method, we have to take a closer look to the log likelihood according to model (2.1). The log likelihood of a simple linear regression is

$$\ell(\alpha, \beta, \sigma^2 \,|\, \boldsymbol{y}) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2,$$

where α and β are the intercept and slope of the simple linear regression, respectively. Then, in the case where τ is known, the log likelihood under model (2.1) is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \,|\, \tau, \boldsymbol{y}) = -\frac{\tau}{2} log(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{\tau}(y_i - \beta_{10} - \beta_{11}x_i)^2 - \frac{n-\tau}{2} log(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}\sum_{i=\tau+1}^{n}(y_i - \beta_{20} - \beta_{21}x_i)^2,$$

or

$$\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \,|\, \tau, \boldsymbol{y}) = \ell(\beta_{10}, \beta_{11}, \sigma_1^2 \,|\, y_1, \ldots, y_\tau) + \ell(\beta_{20}, \beta_{21}, \sigma_2^2 \,|\, y_{\tau+1}, \ldots, y_n). \quad (2.2)$$

The first term on the right hand side of (2.2) is the log likelihood of the first τ observations and the second term is the log likelihood of the last $n - \tau$ observations. For τ known, both terms are mutually independent. Thus, the ML estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ are the ML estimates of the two separate models.

In the case of τ unknown, the change point has to be estimated. The main problem in estimating the change point is that there is no solution in closed form for estimating the parameter $\boldsymbol{\beta}$ and τ simultaneously. This is due to the fact that the ML estimate of the parameter $\boldsymbol{\beta}$ is a function of τ . It is only possible for a given value of τ to derive the ML estimate of $\boldsymbol{\beta}$. Therefore, the only feasible way to estimate the change point is to apply a grid search over a set of all possible values of τ. Now, for an arbitrary $\tau \in \{3, 4, \ldots, n-3\}$ the log likelihood given the ML estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$ is

$$\ell_u(\tau \mid \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2, \boldsymbol{y}) = -\frac{n}{2} log(2\pi) - \frac{\tau}{2} log\,\hat{\sigma}_1^2 - \frac{n-\tau}{2} log\,\hat{\sigma}_2^2 - \frac{n}{2} \qquad (2.3)$$

for $\sigma_1^2 \neq \sigma_2^2$ and

$$\ell_e(\tau \mid \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2, \boldsymbol{y}) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\tau\hat{\sigma}_1^2 + (n-\tau)\hat{\sigma}_2^2) - \frac{n}{2} \qquad (2.4)$$

for $\sigma_1^2 = \sigma_2^2 = \sigma^2$, where the subscripts u and e stand for unequal and equal variances, respectively. The ML estimate $\hat{\tau}$ is the value of τ that maximizes (2.3) respectively (2.4). Next we consider testing whether there is a change in the regression regime or not. A very common method for testing hypothesis is the likelihood ratio (LR) test. It is applicable for testing nested models and the test statistic is defined as

$$\lambda(\boldsymbol{y}) = \frac{sup_{\Theta_0} L(\boldsymbol{\theta} \mid \boldsymbol{y})}{sup_{\Theta} L(\boldsymbol{\theta} \mid \boldsymbol{y})},$$

where $L(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the likelihood function of the parameter vector $\boldsymbol{\theta}$ for the given data $\boldsymbol{y}$ and Θ is the entire parameter space. The set Θ_0 is the parameter space restricted under H_0 and is necessarily a subset of Θ, i.e. $\Theta_0 \subset \Theta$. Using the ML method for estimating the parameter $\boldsymbol{\theta}$, the LR test statistic can be written as

$$\lambda(\boldsymbol{y}) = \frac{L(\hat{\boldsymbol{\theta}}_0 \mid \boldsymbol{y})}{L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})},$$

where $\hat{\boldsymbol{\theta}}$ is the unrestricted ML estimate of $\boldsymbol{\theta}$ which can be realized in the entire parameter space Θ, and $\hat{\boldsymbol{\theta}}_0$ is the restricted ML estimate where the maximization is restricted to Θ_0. Under some regularity conditions, minus twice the LR test statistic, i.e.

$$\Lambda(\boldsymbol{y}) = -2\, log\, \lambda(\boldsymbol{y}),$$

follows asymptotically a χ^2-distribution with q degrees of freedom, where q is the difference of the number of parameters in the models under H_0 and H_1, respectively (see Casella & Berger, 2002, for a detailed discussion).

To test whether there is a change in the regression regime or not, and considering model (2.1), the hypothesis is

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$$

An assumption for applying the LR test is that the models under H_0 and H_1 are nested. For model (2.1) it is not obvious that a simple linear regression without a change point is nested in model (2.1). To see this let $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + \boldsymbol{\delta}$ with $\boldsymbol{\delta} = (\delta_0, \delta_1)^T$ and

$$z_i = \begin{cases} 0 & i = 1, \ldots, \tau \\ 1 & i = \tau + 1, \ldots, n. \end{cases}$$

Then (2.1) can be written as

$$y_i = \beta_{10} + \beta_{11} x_i + z_i (\delta_0 + \delta_1 x_i) + \varepsilon_i \quad i = 1, \ldots, n, \qquad (2.5)$$

and it can be clearly seen, that an OLM without a change point, given by

$$y_i = \beta_{10} + \beta_{11} x_i + \varepsilon_i \quad i = 1, \ldots, n,$$

is nested in (2.5). Thus, this assumption for the LR test is satisfied.

Recall that the maximized log likelihood according to a simple linear regression is

$$\ell_e(\hat{\boldsymbol{\beta}}, \tilde{\sigma}^2 | \boldsymbol{y}) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\tilde{\sigma}^2) - \frac{n}{2} \qquad (2.6)$$

where $\tilde{\sigma}^2$ is the usual ML estimate of σ^2 based on all observations. Then $\Lambda_u(\boldsymbol{y})$ is obtained by subtracting (2.4) from (2.6) as

$$\Lambda_u(\boldsymbol{y}) = n\, log(\tilde{\sigma}^2) - \tau\, log(\hat{\sigma}_1^2) - (n - \tau)\, log(\hat{\sigma}_2^2).$$

In case of equal variances this becomes

$$\Lambda_e(\boldsymbol{y}) = n\, log(\tilde{\sigma}^2) - n\, log(\tau \hat{\sigma}_1^2 - (n - \tau) \hat{\sigma}_2^2).$$

As mentioned above, under standard regularity conditions $\Lambda_e(\boldsymbol{y})$ is asymptotically χ^2-distributed. However, as Seber and Wild (1989) noted, standard asymptotical theory does not apply here because τ takes only discrete values and $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ is also true if the change point lies outside the range of the data. Moreover, Hawkins (1980) showed that the LR test statistic tends to infinity as n increases. Therefore, the LR test can only be used as an approximative device. Another test was introduced by

Chow (1960). He assumed that the change point is known and uses the usual F-test statistic for testing two nested models in linear regression. As usually the change point is unknown it is taken to be $\tau = n/2$. The problem that arises here is, that either the model on the left hand side or the model on the right hand side of the change point contains observations of the other regime. Thus, this test only provides satisfactory results if the true change point is $n/2$.

Farley and Hinich (1970) presented another test statistic for testing a change point in an OLM based on a Bayesian approach. They considered the model

$$y_i = \begin{cases} \alpha + \beta x_i + \varepsilon_i & i = 1, \ldots, \tau - 1 \\ \alpha - \delta x_\tau + (\beta + \delta) x_i + \varepsilon_i & i = \tau, \ldots, n, \end{cases} \qquad (2.7)$$

where δ determines the shift at the change point and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Using the notation from above and defining

$$z_i = \begin{cases} 0 & i = 1, \ldots, \tau - 1 \\ x_i - x_\tau & otherwise, \end{cases}$$

then (2.7) can be written as

$$y_i = \alpha + \beta x_i + \delta z_i + \varepsilon_i \qquad i = 1, \ldots, n,$$

and the hypothesis for testing a shift δ in the OLM at the change point is

$$H_0 : \delta = 0 \, H_1 : \delta \neq 0 \, .$$

Farley and Hinich (1970) suggested, that a priori every value of τ is equally likely, i.e.

$$P(\tau = i) = 1/n \qquad for \qquad i = 1, \ldots, n. \qquad (2.8)$$

Then under H_0 the marginal response mean is assumed to follow

$$E_0[y_i] = \alpha + \beta x_i. \qquad (2.9)$$

Under the alternative, i.e. if a shift of size δ occurs at the change point $\tau = i^*$, we have the conditional mean model

$$E_\delta[y_i \,|\, \tau = i^*] = \alpha + \beta x_i + \delta z_i.$$

which is

$$E_\delta[y_i \mid \tau = i^*] = \begin{cases} \alpha + \beta x_i & i = 1, \ldots, i^* \\ \alpha + \beta x_i + \delta(x_i - x_{i^*}) & otherwise. \end{cases}$$

Using (2.8) yields the marginal mean

$$E_\delta[y_i] = \frac{1}{n}\sum_{j=1}^{n} E_\delta[y_i \mid \tau = j]$$

which is

$$E_\delta[y_i] = \begin{cases} \alpha + \beta x_i & i = 1 \\ \alpha + \beta x_i + \delta\frac{1}{n}\sum_{j=1}^{i}(x_i - x_j) & otherwise. \end{cases} \quad (2.10)$$

Farley and Hinich (1970) substituted (2.9) and (2.10) in the likelihood function of the OLM with and without a change point respectively, and gave a first order approximation of the LR test statistic. Furthermore, they mentioned that for σ^2 known, this statistic follows a normal distribution.

2.2 Multiple linear regression

Next we consider an OLM with more than one explanatory variable, commonly known as multiple linear regression. Again, let $y_i, i = 1, 2, \ldots, n$ denote observations on the response variable. In contrast to section 2.1, let $x_i \in \mathrm{R}^{p \times 1}$ denote the column vector of p independent explanatory variables, i.e. $\boldsymbol{x}_i = (1, x_{i2}, \ldots, x_{ip})^T$, with $x_{i1} = 1$ for all i, to include an intercept in the model. Then an OLM with one discontinuous change point can be written as

$$y_i = \begin{cases} \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \varepsilon_{1i} & i = 1, \ldots, \tau \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \varepsilon_{2i} & i = \tau + 1, \ldots, n, \end{cases} \quad (2.11)$$

where $\boldsymbol{\beta}_d, \quad d = 1, 2$, are $p \times 1$ vectors of unknown parameters and ε_{di} are iid errors with $\varepsilon_{di} \overset{iid}{\sim} N(0, \sigma_d^2)$. To ensure valid estimates for $\boldsymbol{\beta}_d$ and σ_d^2 the possible values of τ are restricted to $\{p + 1, \ldots, n - p - 1\}$. Moreover, it is assumed that the first $p + 1$ and the last $n - p - 1$ vectors of $\boldsymbol{x}_i$ are linearly independent.

In matrix representation, model (2.11) can be written as two separate OLMs

$$\begin{cases} \boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \\ \boldsymbol{y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2, \end{cases} \quad (2.12)$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are both column vectors of the first τ and the last $n - \tau$ observations of the response variable, respectively. The matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are the first τ and the last $n - \tau$ row vectors of the design matrix respectively, and hence given by $\boldsymbol{X}_1 = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\tau)^T$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_{\tau+1}, \ldots, \boldsymbol{x}_n)^T$. Furthermore, the error vectors follow a Normal distribution, i.e. $\boldsymbol{\varepsilon}_d \sim N(\boldsymbol{0}, \sigma_d^2\boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ is the identity matrix with rank τ for $d = 1$ and rank $n - \tau$ for $d = 2$.

As there is no continuity constraint for the two models at the change point, the two models of (2.12) are autonomous and can be written as

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}. \quad (2.13)$$

Note that the design matrix in (2.13) is block diagonal, which indicates independence between the estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Thus, if the change point τ is known, the log likelihood can be partitioned into two terms, namely

$$\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \,|\, \tau, \boldsymbol{y}) = \ell(\boldsymbol{\beta}_1, \sigma_1^2 \,|\, y_1, \ldots, y_\tau) + \ell(\boldsymbol{\beta}_2, \sigma_2^2 \,|\, y_{\tau+1}, \ldots, y_n), \quad (2.14)$$

with $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2)^T$. These two terms correspond to the log likelihood of the first τ observations and the last $n - \tau$ observations and both terms are mutually independent. Thus, the ML estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ are the ML estimates of the two separate models and are given by

$$\hat{\boldsymbol{\beta}}_d = (\boldsymbol{X}_d^T\boldsymbol{X}_d)^{-1}\boldsymbol{X}_d^T\boldsymbol{y}_d, , \quad d = 1, 2$$

and

$$\hat{\sigma}_1^2 = \frac{1}{\tau}\hat{S}_1^2, \quad \hat{\sigma}_2^2 = \frac{1}{n - \tau}\hat{S}_2^2,$$

where

$$\hat{S}_d^2 = (\boldsymbol{y}_d - \boldsymbol{X}_d\hat{\boldsymbol{\beta}}_d)^T(\boldsymbol{y}_d - \boldsymbol{X}_d\hat{\boldsymbol{\beta}}_d)$$

is the residual sum of squares for the model in the dth segment.

In the case of τ unknown, however, the change point has to be estimated. The ML estimate of τ is again the value which maximizes the log likelihood (2.14) at the given ML estimates $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ and $\hat{\boldsymbol{\sigma}}^2$ of the two models. Note that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$ are again functions of τ and have to be estimated for each τ separately. The log likelihood (2.14) at these ML estimates is given by

$$\ell(\tau \mid \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2, \boldsymbol{y}) = -\frac{n}{2} log(2\pi) - \frac{\tau}{2} log\,\hat{\sigma}_1^2 - \frac{n-\tau}{2} log\,\hat{\sigma}_2^2 - \frac{1}{2\hat{\sigma}_1^2}\tau\hat{\sigma}_1^2 - \frac{1}{2\hat{\sigma}_2^2}(n-\tau)\hat{\sigma}_2^2.$$

Thus, $\hat{\tau}$ is obtained by maximizing

$$\ell(\tau \mid \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2, \boldsymbol{y}) = -\frac{n}{2} log(2\pi) - \frac{\tau}{2} log\,\hat{\sigma}_1^2 - \frac{n-\tau}{2} log\,\hat{\sigma}_2^2 - \frac{n}{2}. \qquad (2.15)$$

with respect to $\tau = p + 1, \ldots, n - p - 1$.

Under the assumption $\sigma_1^2 = \sigma_2^2 = \sigma^2$ the ML estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\left(\hat{S}_1^2 + \hat{S}_2^2\right)$$

and (2.15) reduces to

$$\ell(\tau \mid \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{y}) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log\,\hat{\sigma}^2 - \frac{n}{2}.$$

This means, that in the case of equal variances, $\hat{\tau}$ minimizes $\hat{S}_1^2 + \hat{S}_2^2$.

For testing the necessity of a change in an OLM, consider again the LR test statistic and Chow’s test. The quantity $\Lambda_u(\boldsymbol{y})$ based on the LR test statistic for model (2.12) is

$$\Lambda_u(\boldsymbol{y}) = \left[n\,log\frac{\tilde{S}^2}{n} - \tau\,log\frac{\hat{S}_1^2}{\tau} - (n-\tau)\,log\frac{\hat{S}_2^2}{n-\tau} \right]_{\tau=\hat{\tau}}, \qquad (2.16)$$

where

$$\tilde{S}^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

is the residual sum of squares for the model assumed under the null hypothesis. This statistic can be used to tests for a change in the variance as well as for a change in the regression coefficients (Worsley, 1983). In the case of equal variances, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we have

$$\Lambda_u(\boldsymbol{y}) = n\, log\left[\frac{\tilde{S}}{\hat{S}_1^2 + \hat{S}_2^2}\right]_{\tau = \hat{\tau}}.$$

For τ known and $\sigma_1^2 = \sigma_2^2$ the usual F-test statistic for $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ is

$$F_\tau = \frac{\left[\tilde{S}^2 - \left(\hat{S}_1^2 + \hat{S}_2^2\right)\right]/p}{\left(\hat{S}_1^2 + \hat{S}_2^2\right)/(n - 2p)},$$

which under H_0 follows an F-distribution with p and $n - 2p$ degrees of freedom. Worsley (1983) and Beckman and Cook (1979) suggested to use a generalized F-test statistic, namely

$$F_{max} = \ \max{}_{p < \tau < n-p} F_\tau$$

for testing $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. They gave an approximation to the distribution of F_{max} under the null hypothesis based on the Bonferroni inequality. As the distribution of the F-test statistic depends on the configuration of the design matrix, Beckman and Cook (1979) simulated four OLMs with different design matrices to investigate the influence of the design on the distribution of the F-test statistic. They showed that there is a non-negligible influence. Furthermore, they gave approximative upper bounds for the 90% percentiles of the F_{max} distribution based on these simulations. The bounds were conservative when testing for a change in the linear regression or if the variability of the explanatory variable is large. Therefore, if the variability of the explanatory variable is greater than in the considered design of Beckman and Cook (1979), they recommended to apply the usual Bonferroni inequality instead of the simulated values. Worsley (1983) introduced upper bounds for the percentiles of the F_{max} distribution, based on an improved Bonferroni inequality (Worsley, 1982). Furthermore, to avoid the integration for calculating these bounds, he approximated these bounds using the MacLaurin series. He showed that both, the exact and the approximated bounds are more accurate than the bounds calculated with the usual Bonferroni inequality.

Farley, Hinich, and McGuire (1975) introduced a simpler interpretation of the test presented by Farley and Hinich (1970). Furthermore, they compared the power of the three methods, the Chow test, the approach based on F_{max} and the method introduced by Farley and Hinich (1970). Their results, based on a few simulations were that Chow's test using $\tau = n/2$ is most powerful if the change point lies in the middle of the data. In this case, the method introduced by Farley and Hinich (1970) has less power than that of Chow, but performs better than the LR test. In contrast, if the change point lies near the left or right extremes of the data, the LR test is most powerful.

Esterby and El-Shaarawi (1981) considered a linear regression with one change point, where the explanatory variables are polynomials of unknown degree p_1 and p_2 for the first and second segment, respectively. They showed that the maximum likelihood for the assumed change point model is proportional to $\hat{\sigma}_1^{-\tau}\hat{\sigma}_2^{-(n-\tau)}$ assuming equal variances, and proportional to $(\tau - p_1 - 1)\hat{\sigma}_1^2 + (n - \tau - p_2 - 1)\hat{\sigma}_2^2$ assuming unequal variances, where $\hat{\sigma}_d^2$ are the ML estimates of σ_d^2. Thus, in the case of equal variances, maximizing the log likelihood corresponds to minimizing the residual sums of squares. Furthermore, they proposed an iterative method for estimating simultaneously the degrees of the polynomials and the change point.

Tests for general hypotheses, where the variance additionally changes at the change point, were first introduced by Brown, Durbin, and Evans (1975) using recursive residuals. These residuals will be considered Later in the article. A more detailed discussion on testing a change point in OLMs is given in Seber and Wild (1989).

3. Change Point Detection in Generalized linear model

In this section GLMs with one discontinuous change point are considered. GLMs are a generalization of OLMs.First, the response variable must be no longer normal distributed, but can follow any distribution from the linear exponential family. Second, in GLMs the mean structure is determined by a continuous link function $g(\cdot)$ and an unknown parameter vector $\boldsymbol{\beta}$, namely,

$$g(\mu) = \eta = \boldsymbol{x}^T\boldsymbol{\beta},$$

where η is the so called linear predictor. Third, it follows that the response variance is the product of a so-called dispersion parameter ϕ and the variance function $V(\cdot)$, which is allowed to depend on μ, i.e.

$$Var(y) = \phi V(\mu).$$

In general, a change in the mean structure as well as a change in the variance structure is imaginable. A different mean structure for both segments may be due to different link functions as well as different linear predictors, where the difference of the linear predictors may be due to different sets of explanatory variables or different values of the linear parameter $\boldsymbol{\beta}$. A change in the variance structure can be due to different probability models for each segment, which indicates different variance functions

$$V(\cdot)$$

or different dispersion parameters specific for each segment. However, in this work only a change in the linear parameters is considered. Moreover, the probability model is the same for all segments and we assume that the dispersion parameter is constant for all observations and segments. Thus, in the remainder of this work a common dispersion parameter ϕ is considered. It is important to note that the variance of the observation y is a function of the mean μ. Thus, a change in the mean indicates a change in the variance of y, as well, even if the dispersion parameter is constant for all observations and the variance function is the same for all segments.

As GLMs are generalizations of OLMs, the model (2.11) with one discontinuous change point is extended to GLMs as

$$g(\mu_i) = \begin{cases} \boldsymbol{x}_i^T \boldsymbol{\beta}_1 & i = 1, \ldots, \tau \\ \boldsymbol{x}_i^T \boldsymbol{\beta}_2 & i = \tau + 1, \ldots, n, \end{cases} \qquad (2.17)$$

where both parameter vectors $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^T$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p})^T$ are of the same dimension $(p \times 1)$. Again, the change point τ is an index i and determines the observation $\boldsymbol{x}_\tau$ after which the relationship between the response and the explanatory variable changes. Under the assumptions about a unique link function and a unique variance function for both segments, (2.17) can be partitioned into two autonomous GLMs, which can be written as

$$\begin{cases} g(\boldsymbol{\mu}_1) = \boldsymbol{X}_1 \boldsymbol{\beta}_1 \\ g(\boldsymbol{\mu}_2) = \boldsymbol{X}_2 \boldsymbol{\beta}_2, \end{cases} \qquad (2.18)$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are both column vectors containing the first τ and last $n - \tau$ values of the mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, and the matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are build up by the first τ and the last $n - \tau$ row vectors of the design matrix, respectively.

To derive the ML estimates of $\boldsymbol{\beta}_d$, $d = 1, 2, \phi$, and τ , again a closer look at the log likelihood is necessary. The log likelihood of GLMs without a change point of a sample $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is

$$\ell(\boldsymbol{\theta}, \phi \,|\, \boldsymbol{y}) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right], \qquad (2.19)$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ is the vector of the canonical parameter of the exponential family. Usually in GLMs, $\boldsymbol{\beta}$ is the parameter of interest. Thus, it is common to write the log likelihood in terms of $\boldsymbol{\beta}$, i.e. $\ell(\boldsymbol{\beta}, \phi \,|\, \boldsymbol{y})$.

First, consider the case where τ is known. The log likelihood for a GLM with one discontinuous change point τ and the parameter of interest $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T + \boldsymbol{\beta}_2^T)^T$ is given by

$$\ell(\boldsymbol{\beta}, \phi \,|\, \tau, \boldsymbol{y}) = \sum_{i=1}^{\tau} \left[\frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right] + \sum_{i=\tau+1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right]. \qquad (2.20)$$

Note that $b^{'}(\theta_i) = \mu_i$, and (2.17) holds. As the y_i's are independent, again, both terms on the right hand side are autonomous (see Subsection 2.1). Consequently the ML estimates $\hat{\boldsymbol{\beta}}_d$ are the ML estimates of the two models of (2.18) corresponding to the first τ and last $n - \tau$ observations, respectively. The dispersion parameter ϕ is estimated by the usual Pearson statistic based on all observations.

In the case where τ is unknown, the change point has to be estimated. As the estimates of the parameters $\hat{\boldsymbol{\beta}}_d$ and ϕ depend on the change point τ, the same problem arises as in OLMs. That is, there is no closed form solution of the estimates $\hat{\tau}$, $\hat{\boldsymbol{\beta}}_d$, and $\hat{\phi}$. Hence, a grid search over all reasonable change points is applied to find the global maximum of the log likelihood

$$\ell(\boldsymbol{\beta}, \phi, \tau \,|\, \boldsymbol{y}) = \ell(\boldsymbol{\beta}_1, \phi \,|\, y_1, \ldots, y_\tau) + \ell(\boldsymbol{\beta}_2, \phi \,|\, y_{\tau+1}, \ldots, y_n).$$

To guarantee the estimable of the parameters $\hat{\boldsymbol{\beta}}_d$ and ϕ, the reasonable values of τ are restricted to $\{p+1, \ldots, n-p-1\}$.

A common quantity to evaluate the goodness-of-fit of a GLM is the deviance. As the fitted value $ {_i}$ of a GLM is a function of the explanatory variables and the estimated linear parameter $\hat{\boldsymbol{\beta}}$, we reparameterize the log likelihood. Thus, in what follows, we denote the log likelihood in terms of $\hat{\boldsymbol{\mu}}$ instead of $\hat{\boldsymbol{\beta}}$. In a GLM without a change point, the deviance is defined as

$$D = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \phi) = 2\phi[\ell(\boldsymbol{y}, \phi \,|\, \boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}}, \phi \,|\, \boldsymbol{y})],$$

where $\ell(\boldsymbol{y}, \phi \,|\, \boldsymbol{y})$ is the log likelihood of the saturated model with $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$. As for a given ϕ and given data set, $\ell(\boldsymbol{y}, \phi \,|\, \boldsymbol{y})$ is a constant, maximizing the log likelihood is equivalent to minimizing the deviance. Besides applying the deviance to evaluate the goodness-of-fit of GLMs, it is widely used to compare nested models. This is done by considering the difference between the deviances of the two models under consideration. In particular, differences between the deviances are used to decide if some additional explanatory variables improve the fit of the model. In general, the difference of the deviance of two nested GLMs equals the LR test statistic. Therefore, under certain regularity conditions, it follows asymptotically a χ^2-distribution with q degrees of freedom, where q is the difference of the number of parameters of these two models.

As mentioned in Subsection 2.1, an OLM without a change point can be considered as nested in a OLM with a change point. This holds for GLMs if the structure of the variance is the same over the entire model and because the design matrix for a GLM with a change point is the same as for an OLM with a change point. Hence, an intuitive and obvious method to compare a GLM with a change point to a GLM without a change point is to analyze the difference between the deviances of these two models. The deviances of the two submodels of (2.18) are

$$D(\boldsymbol{y}_1, \hat{\boldsymbol{\mu}}_1, \phi) = 2\phi\left[\sum_{i=1}^{\tau} \ell(y_i, \phi \,|\, y_i) - \sum_{i=1}^{\tau} \ell(\hat{\mu}_i, \phi \,|\, y_i)\right] D(\boldsymbol{y}_2, \hat{\boldsymbol{\mu}}_2, \phi) = 2\phi\left[\sum_{i=\tau+1}^{n} \ell(y_i, \phi \,|\, y_i) - \sum_{i=\tau+1}^{n} \ell(\hat{\mu}_i, \phi \,|\, y_i)\right],$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are column vectors of the first τ and last $n - \tau$ observations, respectively. As the deviance of two autonomous models are additive, the deviance of a GLM with a change point is

$$D^{cp} = D^{cp}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \phi) = D(\boldsymbol{y}_1, \hat{\boldsymbol{\mu}}_1, \phi) + D(\boldsymbol{y}_2, \hat{\boldsymbol{\mu}}_2, \phi),$$

where the superscript denotes that this is the deviance corresponding to a GLM with a change point. Then the difference between the deviance D of a GLM without change point and D^{cp} is

$$D - D^{cp} = -2\phi[\ell(\hat{\boldsymbol{\mu}}, \phi \,|\, \boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}}_1, \phi \,|\, y_1, \ldots, y_\tau) - \ell(\hat{\boldsymbol{\mu}}_2, \phi \,|\, y_{\tau+1}, \ldots, y_n)],$$

By definition, this is minus twice the LR test statistic of a GLM with and without a change point. For normal errors, identity link function and equal variances, this difference is

$$D - D^{cp} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) - \left[(\boldsymbol{y}_1 - \hat{\boldsymbol{\mu}}_1)^T(\boldsymbol{y}_1 - \hat{\boldsymbol{\mu}}_1) + (\boldsymbol{y}_2 - \hat{\boldsymbol{\mu}}_2)^T(\boldsymbol{y}_2 - \hat{\boldsymbol{\mu}}_2)\right] = \tilde{S}^2 - \left[\hat{S}_1^2 + \hat{S}_2^2\right],$$

which is the difference between the residual sum of squares of the two models. As the same deviations from the regularity conditions mentioned in Subsection 2.1 arises, the difference $D - D_{cp}$ does not follow a χ^2-distribution, even in the case of a common dispersion parameter for all observations. Thus, this difference can only be used as an approximative test.

4. Bootstrap (recent development)

To find such structural breaks as soon as possible. Such problem arises across many scientific areas: quality control Lai (1995), cybersecurity Blazek and Kim (2001), Wang et al. (2004), econometrics Spokoiny (2009), Mikosch and Starica (2004), geodesy e.t.c. Article Shiryaev (1963) describes classical results in change point detection theory. Overview of the state-of-art methods are presented in Polunchenko and Tartakovsky (2011) and Shiryaev (2010).

This section considers sequential hypothesis testing, in which each hypothesis $(\mathrm{P}_1 = \mathrm{P}_2)$ monitors the presence of change point through Likelihood Ratio Test (LRT) using sliding window. At each time step the procedure extracts a data slice, splits it in two parts of equal size and executes LRT on it. High values of LRT indicate possible distribution difference in the window parts $(\mathrm{P}_1 \neq \mathrm{P}_2)$. Procedures with LRT have demonstrated above. The work Quandt (1960) proposes application of LRT for detection of breaks in linear regression model. It was further developed by many authors, e.g. Haccou et al. (1987), Srivastava and Worsley (1986). Papers Liu et al. (2008), Zou et al. (2007) investigate LRT for change point detection for nonparametric case. Nonparametric approaches are easily adaptable for complex data but in general they need more information for model building than their parametric alternatives. Introduction of parametric assumption: $\mathrm{P}_1, \mathrm{P}_2 \in \{\mathrm{P}(\theta) : \theta \in \mathrm{R}^p\}$ allows to reduce the sufficient number of observations as soon as $\mathrm{P}(\theta)$ has less degrees of freedom than nontapametric model. The state-of-the-art review of parametric models based on LRT and its application to economics and bio-informatics are presented by Chen and Gupta (2012). The paper Gombay (2000) explores how LRT can be used for sequential change point detection in case $\mathrm{P}(\theta)$ is exponential family.

This section provides description of the Change Point Detection algorithm which employs Likelihood Ratio Test (LRT). Let $(\mathrm{P}(\theta), \theta \in \mathrm{R}^p, L(\theta) = log(\partial^n \mathrm{P}(\theta)/\partial Y))$ be a parametric assumption about the nature of data inside the window $(Y_{t-h}, \ldots, Y_{t+h-1})$ with central point t and size $2h$. Here and further we assume, that the observations $\{Y_i\}_{i=1}^n$ are independent, so

$$L(\theta, \mathrm{Y}) = \sum_i l_i(\theta) \qquad (L)$$

.

Denote argmax of the Likelihood function and the “real” model parameter value as follows

$$\hat{\theta} = argmax\theta L(\theta, \mathrm{Y}), \quad \theta^* = argmax\theta \mathrm{E} L(\theta, \mathrm{Y}).$$

The algorithm sequentially computes LRT statistic ($T_h(t)$) for each t in the sliding window procedure. The LRT statistic itself corresponds to the gain from window split into two parts ($\mathrm{Y}_l, \mathrm{Y}_r$):

$$T_h(t) = L(\hat{\theta}_l, \mathrm{Y}_l) + L(\hat{\theta}_r; \mathrm{Y}_r) - L(\hat{\theta}, \mathrm{Y}), \qquad (T)\mathrm{Y}_l = (Y_{t-h}, \ldots, Y_{t-1}), \quad \mathrm{Y}_r = (Y_t, \ldots, Y_{t+h-1}), \hat{\theta}_l = argmax\theta L(\theta, \mathrm{Y}_l), \quad \hat{\theta}_r = argmax\theta L(\theta, \mathrm{Y}_r)$$

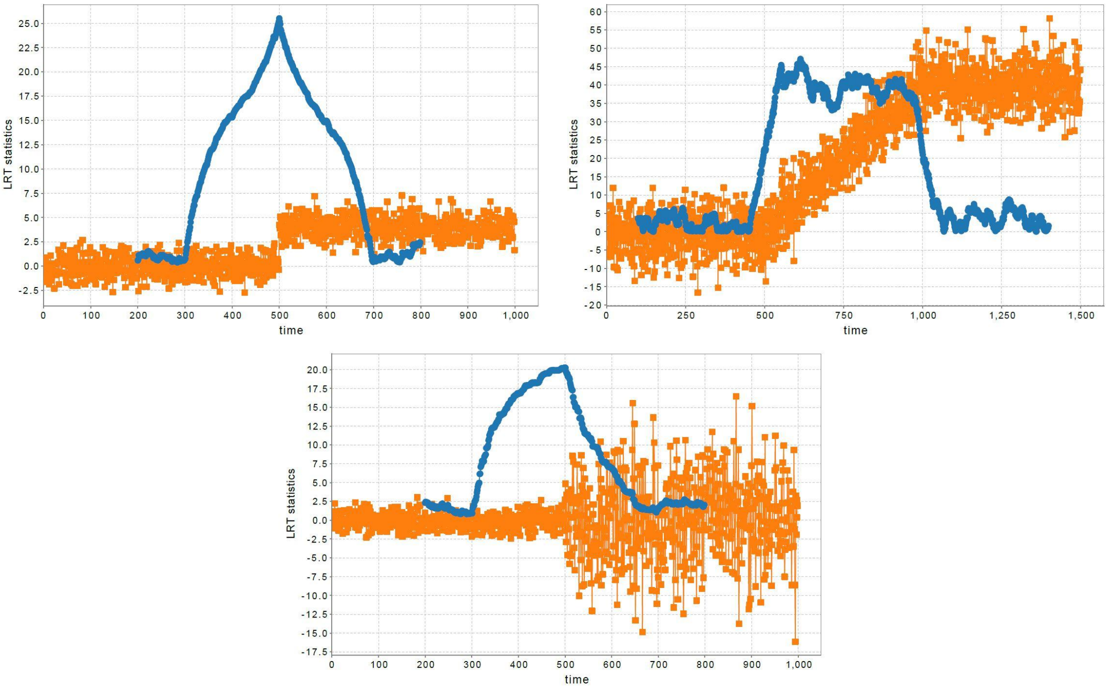

Figure 1: Types of change point and the geometry of change-point patterns: triangle pattern – abrupt mean transition, trapezium pattern – smooth mean transition, horn pattern – abrupt variance transition.

Assume that MLE parameters belong to the local region $\hat{\theta}, \hat{\theta}_l, \hat{\theta}_r \in \Theta(r)$ and the likelihood has a fit quadratic expansion (A), then with probability $1 - 4e^{-x}$ for each t

$$\left|\sqrt{2T_h(t)} - \parallel D_{lr}(\theta_r^* - \theta_l^*)(t) + \xi_{lr}(t)\parallel\right| \le 7\Diamond(\sqrt{2}r, \mathbf{x}),$$

where

$$\xi_{lr}(t) = D_{lr}\{D_l^{-2}\nabla L(\theta_l^*, \mathrm{Y}_l) + D_r^{-2}\nabla L(\theta_r^*, \mathrm{Y}_r)\}, \quad D_{lr} = D_l D^{-1} D_r.$$

According to above, encountering change point, statistic $2T_h(t) \approx \| \xi(t) + \Delta(t) \|^2$ starts growing according to change point pattern type (for example spades, trapezium, horn, ref. the Figure 1). In order to match pattern positions, the procedure monitors $2h$ values of the LRT simultaneously and convolves them with each of the predefined pattern functions $P_\tau(t)$:

$$TP_h(\tau) = \sum_t P_\tau(t)\sqrt{2T_h(t)}. \qquad (TP)$$

High values of $TP_h(\tau)$ correspond to a suffcient correlation of $\sqrt{2T_h}$ and $P\tau$ (similar to the dependence on t). The algorithm marks a time moment τ at a scale h as a change point, if the test statistic $TP_h(\tau)$ exceeds a calibrated (by bootstrap procedure) critical value z_h:

$$\{\tau \text{ is a change point}\} \Leftrightarrow \{\exists h : TP_h(\tau) > z_h\}.$$

The greater window size h is chosen, the more probably the algorithm will mark τ as a change point. Again, small windows may mark τ faster.

$ {Weighted,,,, bootstrap,,,, procedure}$ enables resampling of the statistic $max_{1 \le \tau \le n} TP_h(\tau)$ and thus calculation of the critical value z_h for the window size $2h$. It generates a sequence of weighted likelihood functions, where each element is a convolution of independent likelihood components and weight vector $(u_1^b, \ldots u_n^b)$:

$$L^b(\theta, \mathrm{Y}) = \sum_i u_i^b l_i(\theta), \qquad (Lb)$$

where $\{u_i^b\}_{i=1}^n$ are i.i.d. and $u_i^b \in \mathcal{N}(1, 1)$. At each weights generation one gets a new value of $L^b(\theta)$ and its optimal parameter θ^b and thus bootstrap procedure enables to estimate $L(\hat{\theta})$ fluctuations. The corresponding bootstrap LRT statistic is

$$T_h^b(t) = L^b(\theta_l^b, \mathrm{Y}_l) + L^b(\theta_r^b, \mathrm{Y}_r) - \sup\theta\{L^b(\theta, \mathrm{Y}_l) + L^b(\theta + \hat{\theta}_r - \hat{\theta}_l, \mathrm{Y}_r)\}, \qquad (Tb)\theta^b = \mathrm{argmax}\theta L^b(\theta, \mathrm{Y}).$$

Parameter $(\hat{\theta}_r - \hat{\theta}_l)$ is required for condition $T_h^b \approx \| \xi^b \|$. In this case one can estimate $max_{1 \le \tau \le n}\ TP_h^b(\tau)$ quantiles under the null hypothesis $(\Delta^b(t) \propto \hat{\theta}_r(t) - \hat{\theta}_l(t))$ instead of the false assumption $(\Delta^b(t) = 0)$.

Empirical bootstrap version generates subsamples of data $\{Y_k\}$ from the complete dataset with random independent indexes of size n. In this case

$$L^{\epsilon}(\theta, \mathrm{Y}) = \sum_{i} l_{k(i)}(\theta), \qquad \text{(Le)}$$

where $\{k(i)\}_{i=1}^{n}$ are i.i.d. and $k(i) \in \{1, \ldots, n\}$. For all window positions $\hat{\theta}_r = \hat{\theta}_l = \hat{\theta}$ and here bias correction is not required. So the corresponding LRT statistic is like (T):

$$T_h^{\epsilon}(t) = L^{\epsilon}(\theta_l^{\epsilon}, \mathrm{Y}_l) + L^{\epsilon}(\theta_r^{\epsilon}, \mathrm{Y}_r) - L^{\epsilon}(\theta^{\epsilon}, \mathrm{Y}), \qquad \text{(Te)}\theta^{\epsilon} = argmax\theta L^{\epsilon}(\theta, \mathrm{Y}).$$

Empirical bootstrap works better in the application but less suitable for theoretical investigations (the distribution is discontinuous).

5. Detecting multiple change points: the PULSE criterion

In this section, we mainly focus on detecting mean changes, and as an adoption of the method, detecting variance changes. The following brief review stimulates us to consider a new way to investigate this issue, which has potential to handle with more complex data structures. In the literature, some objective function-based criteria with optimization algorithms for exhaustive search have been proposed for the problems with fixed numbers of change points. Yao (1988) suggested a BIC type criterion. Fricket al. (2014) suggested a simultaneous multiscale change-point estimator(SMUCE) by solving an optimization problem, Yao and Au (1989) proposed a penalized least squares-based approach for mean changes. A weighted least squares function-based method was suggested by Gao et al. (2018). Harchaoui and Levy-Leduc (2010) proposed a LASSO-based approach. The estimation consistency can be ensured under certain regularity conditions. One of the main concerns about these methods are about their computational complexities. See the comments by Niu et al. (2016). When the number of change points goes to infinity as the sample size tends to infinity, the methods require more computational costs. In contrast, cumulative sum-based approaches (CUSUM) are very popular with less computational cost. The relevant methods are based on hypothesis testing, which in many cases are efficient in detection. The seminal paper by Page (1954) has great influence for later developments. Vostrikova (1981) designed some tests for multiple changes through binary segmentation procedures. To alleviate the difficulty caused by short spacings between change points or small jump magnitudes, Fryzlewicz (2014, 2020) introduced an additional randomization step in the algorithms called WBS and WBS2, respectively, where WBS2 can be computationally more efficient. Using moving sum (MOSUM) or “scan” statistic to construct test statistic is also a popularly used technique such as Bauer and Hackl (1980) and Chu et al. (1995). Wu and Zhao (2007) and Cao and Wu (2015) discussed the limiting distributions of the maxima of MOSUM. Hao et al. (2014) considered a MOSUM-based test statistic, called screening and ranking algorithm (SaRa) to simultaneously detect multiple change points. A further development is by Fang et al. (2020) who also used hypothesis testing-based method to detect multiple changes and gave a good way to control false positives in terms of the study on the large deviation theory. To handle the case with diverging number of change points, for the i.i.d. normal errors, Baranowski et al (2019) extended CUSUM-based procedure. Wang et al. (2020) extended the WBS procedure. Eichinger and Kirch (2018) also suggested a MOSUM-based statistic, to simultaneously determine changes when the number of change points goes to infinity as the sample size tends to infinity. They used the maximum of local MOSUM’s over all possible local intervals such that all local changes that have sufficient large magnitudes, similarly as Hao et al. (2014) assumed, can be detected. This is also a computational efficient approach.

5.1 Notations

Let $X_1, \ldots, X_n$ be independent one-dimensional random variables decomposed as

$$X_i = \mu_i + \varepsilon_i, 1 \le i \le n,$$

where $\mu_i = E(X_i)$ are the means. Assume that there are K change points $1 < z_1 < z_2 < \cdots < z_K < n$ such that $\mu_{z_{k-1}+j} = \mu^{(k)}, fork = 1, \ldots, K+1$ and $0 \le j \le z_k - z_{k-1} - 1$ where $z_0 = 0$ and $z_{K+1} = n$. For $k = 1, \ldots, K$, write $\beta_k = |\mu^{(k+1)} - \mu^{(k)}|$ for the (non-zero) difference in means between consecutive segments. The number K can go to infinity as the sample size goes to infinity.

Write the minimum length of segments as α_n^*:

$$\alpha_n^* := \min 0 \le k \le K \{z_{k+1} - z_k\} \qquad (5.1)$$

and the minimum magnitudes of mean changes as v:

$$v := \min 1 \le k \le K \beta_k. \qquad (5.2)$$

Denoted by $1 < \hat{z}_1 \le \hat{z}_2 \le \cdots \le \hat{z}_k < n - 1$ as the estimated locations.

5.2 Criterion Construction

Construct a signal statistic by the following steps. Consider the mean changes detection problem first.

Difference of Moving Averages: To character the mean information, let $S(i)$ be the moving sum with the window size α_n for every location i as:

$$S(i) = \sum_{j=i}^{i+\alpha_n-1} \mu_j \qquad (5.3)$$

As the difference between two successive moving sums at the population level can show the mean change at its location z_k, we define $D(i)$ as: $(for\ 1 \le i \le n - 2\alpha_n,\ if\ 2\alpha_n < \alpha_n^*)$

$$D(i) := \frac{1}{\alpha_n}(S(i) - S(i - \alpha_n)) = \frac{1}{\alpha_n}\left(\sum_{j=i}^{i+\alpha_n-1} \mu_j - \sum_{j=i-\alpha_n}^{i-1} \mu_j\right) \qquad (5.4)$$

For any fixed k, we have:

$$D(i) = \begin{cases} \frac{i-(z_k-\alpha_n)}{\alpha_n}(\mu_{k+1}-\mu_k) & z_k-\alpha_n \le i \le z_k \\ \frac{z_k+\alpha_n-i}{\alpha_n}(\mu_{k+1}-\mu_k) & z_k \le i \le z_k+\alpha_n \\ 0, & z_{k-1}+\alpha_n \le i \le z_k-\alpha_n. \end{cases} \qquad (5.5)$$

This is because, when $z_{k-1}+\alpha_n \le i \le z_k-\alpha_n$, $S(i) = S(i+\alpha_n)$. $D(i)$ attains a local maximum/minimum at $i = z_k$ for any k with $1 \le k \le K$ within the segment of length $2\alpha_n$. This is not a new idea while just the idea of MOSUM. Identifying local minima would be a way to identifying changes. As it can be expected to have too many local maxima/minima due to the randomness oscillation, we may have difficulty to accurately determine the number of change points and their locations. To make the differences more smoothly at the sample level, we consider a smoothing step by doubly averaging below. It is worth pointing out that the second averaging step in theory is not a necessary step, but in practice, we found it is useful for a better detection.

Doubly Averaging: The second round of averaging is to repeatedly use datum points in every average. It is worth pointing out that at the population level, this step is not necessary, but at the sample level, this step is designed to alleviate the oscillation of the sequence. Denote $\tilde{D}(i)$ by the averages of $D(i)$ within the window of size α_n:

$$\tilde{D}(i) = \frac{1}{\alpha_n}\sum_{j=i}^{i+\alpha_n-1} D(j). \qquad (5.6)$$

As the result, we have that

$$\tilde{D}(i) = \begin{cases} >0, & z_k-2\alpha_n \le i \le z_k+\alpha_n, \\ 0, & otherwise, \end{cases}$$

with the following detail:

$$
|\tilde{D}(i)| = \begin{cases}
0, & z_{k-1}+\alpha_n \le i \le z_k - 2\alpha_n; \\
\dfrac{i-(z_k+2\alpha_n+1)\cdot(i-z_k+2\alpha_n)}{\alpha_n^2}\beta_k, & z_k - 2\alpha_n < i \le z_k - \alpha_n; \\
\dfrac{-i^2-\alpha_n i+2iz_k-i+z_k-z_k^2+\alpha_n z_k+\frac{1}{2}(\alpha_n^2-\alpha_n)}{\alpha_n^2}\beta_k & z_k-\alpha_n < i < z_k - \dfrac{\alpha_n}{2} - \sqrt{\alpha_n}; \\
\left(\dfrac{3}{4} - \dfrac{\alpha_n - \sqrt{\alpha_n}}{\alpha_n^2}\right)\beta_k & i = z_k - \dfrac{\alpha_n}{2} - \sqrt{\alpha_n}; \\
\dfrac{-i^2-\alpha_n i+2iz_k-i+z_k-z_k^2+\alpha_n z_k+\frac{1}{2}(\alpha_n^2-\alpha_n)}{\alpha_n^2}\beta_k & z_k - \dfrac{\alpha_n}{2} - \sqrt{\alpha_n} < i < z_k - \dfrac{\alpha_n}{2}; \\
\dfrac{3}{4}\beta_k, & i = z_k - \dfrac{\alpha_n}{2}; \\
\dfrac{-i^2-\alpha_n i+2iz_k-i+z_k-z_k^2+\alpha_n z_k+\frac{1}{2}(\alpha_n^2-\alpha_n)}{\alpha_n^2}\beta_k & z_k - \dfrac{\alpha_n}{2} < i < z_k - \dfrac{\alpha_n}{2} + \sqrt{\alpha_n}; \\
\left(\dfrac{3}{4} - \dfrac{\alpha_n - \sqrt{\alpha_n}}{\alpha_n^2}\right)\beta_k & i = z_k - \dfrac{\alpha_n}{2} + \sqrt{\alpha_n}; \\
\dfrac{-i^2-\alpha_n i+2iz_k-i+z_k-z_k^2+\alpha_n z_k+\frac{1}{2}(\alpha_n^2-\alpha_n)}{\alpha_n^2}\beta_k & z_k - \dfrac{\alpha_n}{2} + \sqrt{\alpha_n} < i \le z_k; \\
\dfrac{(-i+z_k+\alpha_n+2)\cdot(-i+1+\alpha_n+z_k)}{\alpha_n^2}\beta_k, & z_k < i \le z_k + \alpha_n; \\
0, & z_k + \alpha_n \le i \le z_{k+1} - 2\alpha_n.
\end{cases}
$$

where $\beta_k = |\mu_{k+1} - \mu_k|$. Clearly, $\tilde{D}(i)$ attains local maxima at $z_k - \frac{1}{2}\alpha_n$ for each k with $1 \le k \le K$. The local maximizers of $\tilde{D}(i)$ plus $\frac{1}{2}\alpha_n$ are the locations of change points. Similarly as $D(i)$, the sequence $\tilde{D}(i)$ cannot be directly used to be a signal statistic either. Now we construct a sequence of ridge ratios as a signal statistic that is of a “pulse” pattern such that change points can be well identified.

Signal function (we will call it the signal statistic at the sample level). Consider the ratios between $\tilde{D}(i)$ and $\tilde{D}(i + \frac{3}{2}\alpha_n)$. Define the ridge ratios $T(i)$ at the population level as

$$T(i) = \frac{|\tilde{D}(i)| + c_n}{|\tilde{D}(i + \frac{3}{2}\alpha_n)| + c_n}, \qquad (5.7)$$

where $c_n \to 0$ as $n \to \infty$, to be selected later, to avoid the undefined terms $0/0$. In addition, for $i \in (z_{k-1} + \alpha_n, z_k - 2\alpha_n)$, $|\tilde{D}(i)| = 0$ and $|\tilde{D}(i + \frac{3}{2}\alpha_n)|$ monotonically increases. For $i \in (z_k - 2\alpha_n, z_k - \frac{1}{2}\alpha_n)$, $|\tilde{D}(i)|$ monotonically increases, and $|\tilde{D}(i + \frac{3}{2}\alpha_n)|$ monotonically decreases. For $i \in (z_k - \frac{\alpha_n}{2}, z_k + \alpha_n)$, $|\tilde{D}(i + \frac{3}{2}\alpha_n)| = 0$ and $|\tilde{D}(i)|$ monotonically decreases. Then c_n could also play a role of making $T(i)$ monotonic, to avoid the scenario where there are too many points tending to 0. In summary, the following property could be easily justified: letting ↘ and ↗ mean decreasing and increasing with respect to the index i; $\to 0$ and $\to \infty$ mean going to zero and infinity as $n \to \infty$,

$$T(i) = \begin{cases} 1, & z_{k-1} + \alpha_n \le i \le z_k - \frac{7}{2}\alpha_n; \\ \frac{c_n}{|\tilde{D}(i+\frac{3}{2}\alpha_n)| + c_n} \searrow, & z_k - \frac{7}{2}\alpha_n < i < z_k - 2\alpha_n; \\ \frac{c_n}{|\tilde{D}(i+\frac{3}{2}\alpha_n)| + c_n} \to 0, & i = z_k - 2\alpha_n; \\ \frac{|\tilde{D}(i)| + c_n}{|\tilde{D}(i+\frac{3}{2}\alpha_n)| + c_n} \nearrow, & z_k - 2\alpha_n < i < z_k - \frac{\alpha_n}{2}; \\ \frac{|\tilde{D}(i)| + c_n}{c_n} \to \infty, & i = z_k - \frac{\alpha_n}{2}; \\ \frac{|\tilde{D}(i)| + c_n}{c_n} \searrow, & z_k - \frac{\alpha_n}{2} < i < z_k + \alpha_n; \\ 1, & z_k + \alpha_n \le i < z_{k+1} - \frac{7}{2}\alpha_n. \end{cases}$$

Any true change point is just the local minimizer plus $2\alpha_n$. Based on the signal function, using the local minimizers to identify change points is convenient to implement.

Sample Version. To define the signal statistic at the sample level, which is called the signal function at the population level above, we can use the sample averages to estimate $D(i)$ and $\tilde{D}(i)$. Let $\hat{S}(i) = \sum_{j=i}^{i+\alpha_n-1} X_j$ to estimate $S(i)$, $D_n(i) = \frac{1}{\alpha_n}(\hat{S}(i) - \hat{S}(i + \alpha_n))$ and $\tilde{D}_n(i) = \frac{1}{\alpha_n}\sum_{j=i}^{i+\alpha_n-1} D_n(j)$

The signal statistic is then defined as: for $i = 1, \ldots, n - \frac{7}{2}\alpha_n$,

$$T_n(i) = \frac{|\tilde{D}_n(i)| + c_n}{|\tilde{D}_n(i + \frac{3}{2}\alpha_n)| + c_n}, \qquad (5.8)$$

and the ridge value c_n tends to 0 at a certain rate specified later. We can see that $E\tilde{D}_n(i) = \tilde{D}(i)$.

Criterion: As we discussed above, the signal statistic should be highly oscillating and there would be too many local minima. Thus, we restrict our search, separately within each chosen interval, to find a local minimum of $T_n(i)$. We do this through a threshold τ with $0 < \tau < 1$. That is,

$$\{i, \alpha_n + 1 \leq i \leq n - \frac{5}{2}\alpha_n : T_n(i) < \tau\}.$$

From the properties of $T_n(i)$, all these indices can be separated into disjoint subsets each containing only one change point asymptotically. Therefore, we can search, separately within the disjoint subsets, for local minima. To make the search easily implemented, we simply recommend $\tau = 0.5$ as a compromised value to avoid possible overestimation with large τ close to 1 and underestimation with small τ close to 0. From the definition of $T_n(i)$'s and its pulse pattern, we can also search for the changes through identifying local maxima that, at the population level, tend to infinity. But it is equivalent to using $1/T_n(i)$. Thus, we do not discuss its use in this article. Further, from the definition of $T(i)$'s at the population level, the gap between two local minimizers must be larger than $2\alpha_n$. Due to the consistency of the involved estimators, there are $\hat{K}$ pairs $\{m_k, M_k\}$ where m_k and M_k with $m_k < M_k$ are determined by $T_n(i) < 0.5$ and m_k satisfies that $T_n(m_k - 1) \geq 0.5$ and $T_n(m_k) < 0.5$, and $T_n(M_k) < 0.5$ and $T_n(M_k + 1) \geq 0.5$. Write $\hat{z}_k - 2\alpha_n$ as the minimizer in each interval (m_k, M_k).

5.3. Change Points in Variances

In this section, we adopt the criterion for detecting change points in variances. Consider second moments of X_i's that are generated from the following model:

$$X_i = \mu + \varepsilon_i, 1 \leq i \leq n, \qquad (5.10)$$

where μ is an unknown mean and $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2_{(i)}$. Similarly, we assume that $\sigma^2_{(i)}$'s follow a piecewise constant structure with $K + 1$ segments. In other words, there are K change points $1 < z_1 < z_2 < \cdots < z_K < n - 1$ such that, for any k with $0 \leq k \leq K$,

$$\sigma^2_{z_k + 1} = \cdots = \sigma^2_{z_{k+1}} = \sigma^2_k, \qquad (5.11)$$

As before, define $z_0 = 0$ and $z_{K+1} = n$. At the population level, we can similarly define $D(i)$ and $\tilde{D}(i)$:

$$D(i) = log\sigma_{(i)} - log\sigma_{(i-\alpha_n)} \quad and \quad \tilde{D}(i) = \frac{1}{\alpha_n}\sum_{j=i}^{i+\alpha_n-1} D(j).$$

We can estimate μ by the sample mean and the variance by

$$\hat{\sigma}^2_{(i)} = \frac{1}{\alpha_n}\sum_{t=i}^{i+\alpha_n-1}(X_t - \frac{1}{n}\sum_{j=1}^{n}X_j)^2, \qquad (5.12)$$

$D_n(i)$ and $\tilde{D}_n(i)$ are defined as the difference of moving averages and the average of $D_n(j)$'s:

$$D_n(i) = log\hat{\sigma}_{(i)} - log\hat{\sigma}_{(i+\alpha_n)} \quad and \quad \tilde{D}_n(i) = \frac{1}{\alpha_n}\sum_{j=i}^{i+\alpha_n} D_n(j). \qquad (5.13)$$

Finally, we take the ratios of $\tilde{D}(i)$ to acquire the required estimator of $T(i)$:

$$T_n(i) = \frac{|\tilde{D}_n(i)| + c_n}{|\tilde{D}_n(i + \frac{3}{2}\alpha_n)| + c_n}, \qquad (5.14)$$

The criterion is exactly the same as that before by using

$$\{i, 1 \le i \le n - \frac{7}{2}\alpha_n : T_n(i) < \tau\}.$$

This approach can achieve estimation consistency with less computational complexity. In the construction of $\tilde{D}$ with the segment length α_n would be seen from a nonparametric estimation perspective for a function with fixed designed points $t = 1, \ldots, n$. This is because moving averages could be regarded as a local smoothing procedure. Thus, the optimal selection of α_n, if we want to study the estimation efficiency of T_n, could be regarded as the optimal selection of a tuning parameter. Note that the optimal selection of tuning parameter in nonparametric estimation which tries to balance between the estimation bias and variance.Yet, the optimality, if it exists, is related to the rate of convergence of the signal statistic. Thus, this is an essential difference in methodology. It deserves a further study to see in what sense we need an optimal selection and whether the optimal α_n exists.

In addition, this approach could be extended to handle more general models than mean or variance changes. For example, this approach might be used to detect change points in distribution or regression functions. Besides, our approach might also be applied to multivariate data which was considered in Matteson and James (2014) or, under certain regularity conditions, to high dimensional data as Wang and Samworth (2018) considered. A rough idea is to define a criterion that is the minimum of the signal statistics over all components. Note that such a minimum of component-based signal statistics will no longer have a pulse pattern because we can check that the maximum value of this minimum signal statistic is one. But the minima near change points are still zero, which can be used to find out changes. The research is ongoing. More general, when coping with change points detection in functional data mentioned in Berkes et al. (2009), this approach might also work. Other than the component-based method mentioned above, another possible way is to use projected variables.

References