An Analytical Method for Detecting the Change-Point in Simple Linear Regression Model. Application at Weibull Distribution
Dariush Ghorbanzadeh, Philippe Durand, Luan Jaupi

To cite this version:

HAL Id: hal-02464929
https://hal-cnam.archives-ouvertes.fr/hal-02464929
Submitted on 7 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
An analytical method for detecting the change-point in simple linear regression model. Application at Weibull distribution.

Dariush GHORBANZADEH(1)  Philippe DURAND(2)  Luan JAUPI(3)
dariush.ghorbanzadeh@cnam.fr(1)  philippe.durand@cnam.fr(2)  jaupi@cnam.fr(3)
CANAM- département IMATH, 292, Rue Saint-Martin, 75141-Paris CEDEX 03, France

Abstract

In this paper we study an analytical method to detect the change-point in the model of simple linear regression. The study method is used to estimate the parameters of a Weibull model representative a change-point. The procedure proposed in this paper is illustrated through a classical change-point data. For the accuracy of the method a simulation study is performed.

Keywords: Change-point; simple linear regression model; Weibull distribution.

1 Introduction

Change-point models have originally been developed in connection with applications in quality control, where a change from the in-control to the out-of-control state has to be detected based on the available random observations. Up to now various change-point models have been suggested for a broad spectrum of applications like quality control, reliability, econometrics, medicine, signal processing, meteorology, etc.

The general change-point problem can be described as follows: A random process indexed by time is observed and we want to investigate whether a change in the distribution of the random elements occurs. Formally, let $X_1, \ldots, X_n$ denote a sequence of independent random variables, where
the elements $X_1, \ldots, X_k$ have an identical distribution function $f_1$ and $X_{k+1}, \ldots, X_n$ are distributed according to $f_2$ and the change-point $k$ is unknown.

The change point problem has been considered and studied by several authors. Change-point analysis concerns with the detection and estimation of the point at which the distribution changes. One change point problem or multiple change points problem have been studied in the literature, depending on whether one or more change points are observed in a sequence of random variables. Several methods, parametric or non-parametric, have been developed to approach the solution of this problem while the range of applications of change point analysis is broad. There is an extensive bibliography on the subject and several methods to search for the change-point problem have appeared in the literature. TheCUSUM (cumulative sum) approach: Basseville & Nikiforov [1], Lucas & Crosier [12], Ritov [15] and Yashchin [16]. The maximum-logarithm of the likelihood ratio approach: Guralnik & Srivastava [10], Gustafsson [11] and Ghorbanzadeh [7]. The Bayesian approach: Bradley & all [5], Barry & Hartigan [2] and Ghorbanzadeh & Lounes [9]. The Non-Parametric approach: Pettitt [13], Dehling & all [6] and Ghorbanzadeh & Picard [8].

In this work we consider the change-point model for a simple linear regression with one change point. Consider $n$ pairs of observations $(X_i, Y_i)$ and we suppose that the relationship between $X$ and $Y$ can be described by a simple linear regression, where the structure changes after a change point $k \in \{4, \ldots, n-4\}$. This restriction on $k$ is needed to ensure that the parameters in the model are estimable. Thus, the observations $(X_i, Y_i)$ follow a linear model for $i \leq k$ and another linear model for $i > k$. Therefore, the model is given by

\[
\begin{align*}
Y_i &= B_1 + A_1 h(X_i) + \varepsilon_{1,i} & \text{if } i = 1, \ldots, k \\
Y_i &= B_2 + A_2 h(X_i) + \varepsilon_{2,i} & \text{if } i = k+1, \ldots, n
\end{align*}
\]  

(1)

where $A_j$ and $B_j$ $(j = 1, 2)$, are the unknown parameters, $\varepsilon_{j,i}$ are independent errors and $h$ is a known function.

The method proposed in this paper is illustrated through a classical change-point data from Quandt [14]. We use the model (1) to estimate the parameters of a Weibull model representative a change-point. For the accuracy of the method a simulation study is performed.

2 Analytical method for the change-point estimate

For $k_0 \in \{4, \ldots, n-4\}$, we construct $n - 7$ two subsamples as follows:
Table 1: Distribution of data into two subsamples.

<table>
<thead>
<tr>
<th>$k_0$</th>
<th>Sample$_1(k_0)$</th>
<th>Sample$_2(k_0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>$X_1, \ldots, X_4$</td>
<td>$X_5, \ldots, X_n$</td>
</tr>
<tr>
<td>5</td>
<td>$X_1, \ldots, X_5$</td>
<td>$X_6, \ldots, X_n$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$n - 4$</td>
<td>$X_1, \ldots, X_{n-4}$</td>
<td>$X_{n-3}, \ldots, X_n$</td>
</tr>
</tbody>
</table>

For each $k_0 \in \{4, \ldots, n - 4\}$, we consider the following models

$$
\begin{align*}
Y_i &= B_1(k_0) + A_1(k_0) h(X_i) + \varepsilon_{1,i}(k_0) \quad \text{if } i = 1, \ldots, k_0 \\
Y_i &= B_2(k_0) + A_2(k_0) h(X_i) + \varepsilon_{2,i}(k_0) \quad \text{if } i = k_0 + 1, \ldots, n
\end{align*}
$$

(2)

For each $k_0 \in \{4, \ldots, n - 4\}$, $A_1(k_0)$, $B_1(k_0)$, $A_2(k_0)$ and $B_2(k_0)$ solve the following minimization problem:

$$
\min_{(A_1(k_0), B_1(k_0), A_2(k_0), B_2(k_0))} D(k_0, A_1(k_0), B_1(k_0), A_2(k_0), B_2(k_0))
$$

where

$$
D(k_0, A_1(k_0), B_1(k_0), A_2(k_0), B_2(k_0)) = \sum_{i=1}^{k_0} \varepsilon_{1,i}^2(k_0) + \sum_{i=k_0+1}^{n} \varepsilon_{2,i}^2(k_0)
$$

(3)

By classics calculations, we obtain the estimator of $A_1(k_0)$, $B_1(k_0)$, $A_2(k_0)$ and $B_2(k_0)$

$$
\begin{align*}
\hat{A}_1(k_0) &= \frac{\sum_{i=1}^{k_0} (h(X_i) - \bar{h}_{k_0}(X))(Y_i - \bar{Y}_{k_0})}{\sum_{i=1}^{k_0} (h(X_i) - \bar{h}_{k_0}(X))^2}, \quad \hat{B}_1(k_0) = \bar{Y}_{k_0} - \hat{A}_1(k_0) \bar{h}_{k_0}(X) \\
\hat{A}_2(k_0) &= \frac{\sum_{i=k_0+1}^{n} (h(X_i) - \bar{h}_{k_0}(X))(Y_i - \bar{Y}_{k_0}^*)}{\sum_{i=k_0+1}^{n} (h(X_i) - \bar{h}_{k_0}(X))^2}, \quad \hat{B}_2(k_0) = \bar{Y}_{k_0}^* - \hat{A}_2(k_0) \bar{h}_{k_0}(X)
\end{align*}
$$

(4)
where

\[
\begin{align*}
\tau_{k_0}(X) &= \frac{1}{k_0} \sum_{i=1}^{k_0} h(X_i), \\
\tau_{k_0}^*(X) &= \frac{1}{n - k_0} \sum_{i=k_0 + 1}^{n} h(X_i), \\
\bar{Y}_{k_0} &= \frac{1}{k_0} \sum_{i=1}^{k_0} Y_i, \\
\bar{Y}_{k_0}^* &= \frac{1}{n - k_0} \sum_{i=k_0 + 1}^{n} Y_i.
\end{align*}
\]

Let \( D(k_0) = D(k_0, \hat{A}_1(k_0), \hat{B}_1(k_0), \hat{A}_2(k_0), \hat{B}_2(k_0)) \) and

\[
k^* = \arg\min_{k_0} D(k_0)
\]

By equation (4), we deduce the estimators of \( A_1, B_1, A_2 \) and \( B_2 \)

\[
\hat{A}_1 = \hat{A}_1(k^*), \quad \hat{B}_1 = \hat{B}_1(k^*), \quad \hat{A}_2 = \hat{A}_2(k^*), \quad \hat{B}_2 = \hat{B}_2(k^*)
\]

and the change-point time is estimated by

\[
\hat{k} = \text{length of Sample}_1(k^*)
\]

### 3 Application to Quandt’s data

This data was illustrated by Quandt [14]. He considered a simple linear regression model with one point-change. The data, listed in Table 2.

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_i)</td>
<td>4</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>12</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>(Y_i)</td>
<td>3.473</td>
<td>11.555</td>
<td>5.714</td>
<td>5.710</td>
<td>6.046</td>
<td>7.650</td>
<td>3.140</td>
<td>10.312</td>
<td>13.353</td>
<td>17.197</td>
</tr>
<tr>
<td>i</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>(X_i)</td>
<td>15</td>
<td>11</td>
<td>3</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>10</td>
<td>7</td>
<td>19</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 2: Quandt’s data.

The results obtained by the model (1) show a change after the first \( \hat{k} = 12 \) observations, giving

\[
\begin{align*}
Y_i &= 2.2215 + 0.6912 X_i \quad \text{if } i = 1, \ldots 12 \\
Y_i &= 5.9141 + 0.4787 X_i \quad \text{if } i = 13, \ldots 20
\end{align*}
\]

The following graph shows the estimation results.
The change-point detection model for the Weibull distribution

In the following, we note $W(a, b)$ the Weibull distribution with the cumulative distribution function $F(x) = 1 - \exp\left(-\left(\frac{x}{a}\right)^b\right)$. In this section we assume that a sequence of observations $X_1, \ldots, X_n$ represents a change point with:

$$
\begin{cases}
X_i \sim W(a_1, b_1) & \text{if } i = 1, \ldots, k \\
X_i \sim W(a_2, b_2) & \text{if } i = k + 1, \ldots, n
\end{cases}
$$

(9)

For $k_0 \in \{4, \ldots, n-4\}$, we build $n-7$ two subsamples and we order them as the Table 3.

For each subsample, we use the Benard’s approximation (Bernard & Bosi-Levenbach [4]) for median ranks, given by:

$$
\begin{cases}
MR_1(i) = \frac{i - 0.3}{k_0 + 0.4} & i \in \{1, \ldots, k_0\} \\
MR_2(i) = \frac{i - 0.3}{n - k_0 + 0.4} & i \in \{k_0 + 1, \ldots, n\}
\end{cases}
$$

(10)

The cumulative distribution function of Weibull distribution will be transformed to a
linear function:

\[ \ln \left( - \ln \left( 1 - F(x) \right) \right) = b \ln x - b \ln a \]

Let \( Y = \ln \left( - \ln \left( 1 - F(x) \right) \right) \), \( A = b \) and \( B = -b \ln a \).

To estimate the values of the cumulative distribution function, we use the median rank. For each subsample, we have

\[
\begin{aligned}
\left\{
Y_i &= \ln \left( - \ln \left( 1 - MR_1(i) \right) \right) \quad \text{if } i = 1, \ldots, k_0 \\
Y_i &= \ln \left( - \ln \left( 1 - MR_2(i) \right) \right) \quad \text{if } i = k_0 + 1, \ldots, n
\end{aligned}
\]

Then the model (9) is written:

\[
\begin{aligned}
\left\{
Y_i &= B_1(k_0) + A_1(k_0) \ln(X_i) \quad \text{if } i = 1, \ldots, k_0 \\
Y_i &= B_2(k_0) + A_2(k_0) \ln(X_i) \quad \text{if } i = k_0 + 1, \ldots, n
\end{aligned}
\]

By the equations (6), (7) and (8), we deduce the estimators of \( a_1, b_1, a_2 \) and \( b_2 \):

\[
\hat{a}_j = \exp \left( - \frac{\hat{B}_j(k^*)}{\hat{b}_j} \right), \quad \hat{b}_j = \hat{A}_j(k^*) \quad (j = 1, 2)
\]

<table>
<thead>
<tr>
<th>( k_0 )</th>
<th>Sample(_1(k_0) ) ordered</th>
<th>Sample(_2(k_0) ) ordered</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>( X_1^{(1)}, \ldots, X_1^{(4)} )</td>
<td>( X_2^{(1)}, \ldots, X_2^{(n-4)} )</td>
</tr>
<tr>
<td>5</td>
<td>( X_1^{(1)}, \ldots, X_1^{(5)} )</td>
<td>( X_2^{(1)}, \ldots, X_2^{(n-5)} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( n-4 )</td>
<td>( X_1^{(1)}, \ldots, X_1^{(n-4)} )</td>
<td>( X_2^{(1)}, \ldots, X_2^{(4)} )</td>
</tr>
</tbody>
</table>

Table 3: Distribution of data into two subsamples ordered. \( X_j^{(i)} \) denotes the \( i-th \) order statistic of sample\(_j(k_0) \) (\( j = 1, 2 \)).

5 Illustrative data, simulations and application

5.1 Illustrative data.

To illustrate all the steps of the method studied in this paper, we propose the data presented in the Table 4. These data have been simulated from Python 3.3. This is a sample with size 30 representing a point-change. The first 13 data are simulated
according to the Weibull distribution \( W(6, 3) \) and the remains are simulated according to the Weibull distribution \( W(10, 9) \).

\[
\begin{array}{ccccccccccc}
\hat{i} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
X_1 & 5.66 & 4.78 & 5.49 & 6.30 & 4.69 & 7.29 & 4.02 & 5.01 & 5.59 & 3.79 \\
\hat{i} & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 \\
\hat{i} & 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 & 29 & 30 \\
\end{array}
\]

Table 4: Illustrative data.

The steps of the study method are illustrated in the Table 5.
Table 5: The steps of the calculations for the data in the Table 4.

The following figure represents the sum of squared errors defined in equation (3).

Figure 2: Sum of squared errors defined in equation (3).

The following figure shows the weibull probability plot.
Figure 3: The weibull probability plot for the illustrative data. The estimators are the values: \( \hat{a}_1 = 5.78 \), \( \hat{b}_1 = 6.15 \), \( \hat{a}_2 = 10.16 \), \( \hat{b}_2 = 9.83 \) and \( k = 13 \).

### 5.2 Simulations

In order to study the performance of the method, we simulated 1000-samples of sizes \( n = 30 \) and 100 with a change-points \( k = 13 \) and 41. We considered three cases: the first relates to the change in the second parameter of the Weibull distribution, the second case, the change in the first parameter and the third case, the change in both parameters. For each sample we calculated the parameter estimators, the following table summarizes the results obtained for different values of \( a_1, a_2, b_1 \) and \( b_2 \).
Table 6: Statistics of estimators of $a_1, a_2, b_1, b_2$ for size $n = 30$ and change-point $k = 13$.

<table>
<thead>
<tr>
<th></th>
<th>$a_1 = a_2 = 6, b_1 = 2, b_2 = 5$</th>
<th></th>
<th>$a_1 = a_2 = 6, b_1 = b_2 = 4$</th>
<th></th>
<th>$a_1 = 6, a_2 = 10, b_1 = 3, b_2 = 9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean of $\hat{a}_1$</td>
<td>6.1633</td>
<td>mean of $\hat{a}_2$</td>
<td>6.0181</td>
<td>mean of $\hat{b}_1$</td>
<td>2.0747</td>
</tr>
<tr>
<td>std of $\hat{a}_1$</td>
<td>1.0793</td>
<td>std of $\hat{a}_2$</td>
<td>0.4221</td>
<td>std of $\hat{b}_1$</td>
<td>0.7989</td>
</tr>
<tr>
<td>mean of $\hat{a}_1$</td>
<td>6.5194</td>
<td>mean of $\hat{a}_2$</td>
<td>9.3962</td>
<td>mean of $\hat{b}_1$</td>
<td>4.1095</td>
</tr>
<tr>
<td>std of $\hat{a}_1$</td>
<td>0.9955</td>
<td>std of $\hat{a}_2$</td>
<td>0.8481</td>
<td>std of $\hat{b}_1$</td>
<td>2.1639</td>
</tr>
</tbody>
</table>

Table 7: Statistics of estimators of $a_1, a_2, b_1, b_2$ for size $n = 100$ and change-point $k = 41$.

<table>
<thead>
<tr>
<th></th>
<th>$a_1 = a_2 = 6, b_1 = 2, b_2 = 5$</th>
<th></th>
<th>$a_1 = a_2 = 6, b_1 = b_2 = 4$</th>
<th></th>
<th>$a_1 = 6, a_2 = 10, b_1 = 3, b_2 = 9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean of $\hat{a}_1$</td>
<td>6.1116</td>
<td>mean of $\hat{a}_2$</td>
<td>6.0298</td>
<td>mean of $\hat{b}_1$</td>
<td>1.9759</td>
</tr>
<tr>
<td>std of $\hat{a}_1$</td>
<td>0.6597</td>
<td>std of $\hat{a}_2$</td>
<td>0.2304</td>
<td>std of $\hat{b}_1$</td>
<td>0.5492</td>
</tr>
<tr>
<td>mean of $\hat{a}_1$</td>
<td>6.2783</td>
<td>mean of $\hat{a}_2$</td>
<td>9.3905</td>
<td>mean of $\hat{b}_1$</td>
<td>4.2397</td>
</tr>
<tr>
<td>std of $\hat{a}_1$</td>
<td>0.8557</td>
<td>std of $\hat{a}_2$</td>
<td>0.7253</td>
<td>std of $\hat{b}_1$</td>
<td>1.5346</td>
</tr>
<tr>
<td>mean of $\hat{a}_1$</td>
<td>6.8058</td>
<td>mean of $\hat{a}_2$</td>
<td>9.9860</td>
<td>mean of $\hat{b}_1$</td>
<td>2.7517</td>
</tr>
<tr>
<td>std of $\hat{a}_1$</td>
<td>0.8215</td>
<td>std of $\hat{a}_2$</td>
<td>0.2530</td>
<td>std of $\hat{b}_1$</td>
<td>0.4662</td>
</tr>
</tbody>
</table>
5.3 Application

We apply the method to study data used by Bhattacharya & Bhattacharjee \[\text{[3]}\] which represents the Average Monthly Wind Speed (m/s) at kolkata (from 1\textsuperscript{st} March, 2009 to 31\textsuperscript{st} March, 2009).

The following figure represents the sum of squared errors for the Average Monthly Wind Speed (m/s) at kolkata data.

![Figure 4: Sum of squared errors defined in equation \([3]\).](image)

The following figure shows the weibull probability plot for the Average Monthly Wind Speed (m/s) at kolkata data.

![Figure 5: The weibull probability plot. The estimators are the values: \(\hat{a}_1 = 1.13\), \(b_1 = 1.88\), \(\hat{a}_2 = 1.27\), \(b_2 = 1.65\), and \(k = 15\).](image)
6 conclusion

In this paper, we presented an analytical method of estimating change-point parameters. The results obtained for the Weibull distribution are satisfying. The proposed method is very simple to program that could be easily adapted to other distributions.

References


