Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

Een algemene analyse van change-point methoden

(Engelse titel: A general analysis of change-point methods)

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

Mohammed el Ouasgiri

Delft, Nederland
Maart 2019

BSc verslag TECHNISCHE WISKUNDE

"Een algemene analyse van change-point methoden"
(Engelse titel: "A general analysis of change-point methods)"

Mohammed el Ouasgiri

Technische Universiteit Delft

Begeleider

Dr.ir. J. Bierkens

Overige commissieleden

Drs. E. van Elderen

Dr. J.J. Cai

...

...

Maart, 2019

Delft

Abstract

The primary goal of this report is to provide a general overview of offline change-point literature as it is known today. Change-point methods are important statistical problems, where we are interested in determining whenever a certain data-set changes in *structure*. Furthermore, the term *off-line* is meant to indicate that the data itself is already known, whereas on the other hand we have *on-line* methods which deal with situations where new data is yet being received during localisation of the change-points. In this report we mainly consider off-line methods, as we feel off-line methods provide a more friendly introduction into change-point analysis and on-line methods are in principle just an extension of their off-line counterparts.

First off, these off-line change-point methods are considered under different assumptions (parametric, non-parametric). In each case, we treat a solution to the change-point problem under different models(normal and gamma model, mean or varianche change etc.). Eventually we shall also treat some widely used algorithms, meant to extend the problem into the localisation of multiple change-points.

Aside from theoretical considerations, an equally important part of this report will be focused on empirical results. Both for the statistics as algorithms will there be a performance study where the different methods will be empirically assessed and compared under different models, namely the robustness against *outliers*(extreme data-values) will be investigated. So to complement the primary goal, we will also focus on the following two subgoals:

1. Assessment and comparison of different change-point models
2. Evaluating and improving robustness against outliers

Table of Contents

1 Introduction

The subject that we will consider is that of change-point analysis. Briefly speaking, a change-point problem consists of determining when a given data-set *structurally* changes. What is meant specifically with a structural change depends mainly on the scenario, and will become more clear as we consider the problem under different assumptions.

As for the problem itself, change-point analysis has already been studied for quite a few decades(see [2], [4], [9]). Its first occurance was in the late thirties, where it was used as a diagnostic tool within quality control to assess whenever production levels were drastically changing. Adequately detecting these changes within a reasonable amount of time, meant that potential disasters or other unfortunate consequences could be timely averted. Since that period, change-point analysis has gradually grown in importance within both the statistical and computer science disciplines. As a result, the problem has been studied under a wide variety of different assumptions over the years.

The problem is separated into two main branches : *off-line* and *online* change-point analysis. In the former case, we are dealing with the situation when the entirety of the data has already been given, so we are mainly conducting a retrospective analysis of the data. In contrast, the *online* method(also called the sequential method) is devised to be able to conduct a change-point analysis while new data is being received. In this report, we will exclusively focus on *off-line* change-point models, as *online* models can be easily regarded as a repeated occurence of an offline model by fixing the available data at every timepoint.

We have described in general terms what change-point problems are, and now we shall give a more formal problem formulation. Intuitively speaking, when performing off-line change-point analysis, we wish to estimate two things: the amount of change-points and their locations. To this end we are assuming that the data's mean value is piecewise constant, so in this case a change-point location is wherever the data displays a noticeable *break*. In other words, a set of estimated change-points corresponds with a certain segmentation of the data, where the segments represent homogeneous sections of the data(i.e. sections where the data's mean remains roughly constant from a structural view).
Before we can state the problem more formally, we have to introduce some terminology:

• The data is represented as $y = \{y_t\}_{t=1}^{T}$, where t is the index for time and T is the end time(i.e. T is the length of the data)

• A subset of the data, from point a to point b, is either denoted as $\{y_t\}_{t=a}^{b}$ or $y_{a..b}$

• The total amount of change-points is denoted with K

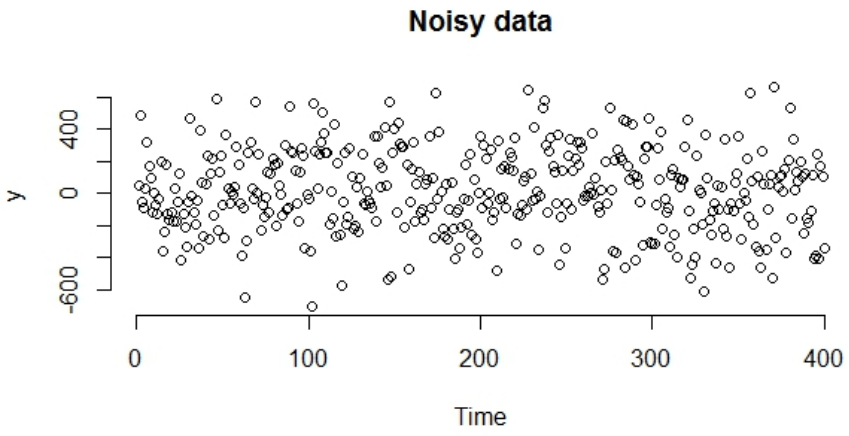


Figure 1: A plot of noisy data with a break at every 100 observations

• A set of estimated change-points will be denoted as $\tau = \{\tau_1, \tau_2, ..., \tau_K\}$

• Whenever the data is segmented, we denote with θ_i the relevant density parameter for the i-th segment

• The set of all possible segmentations of the data is denoted with Ω

In the course of the next two sections, we will consider several different models for y_t, and in general we assume that y_t will be real-valued and independently distributed with density function $f(\theta)$(we will also consider the case when f is unknown). The vector τ will then describe where y_t undergoes a change in its parameter θ.

We also make the additional assumption that the changes are abrupt, instead of gradually occuring over a period of time. In the end, the question remains how τ and its cardinality should be estimated.

An obvious first choice is visual inspection. This is especially effective when the changes relate to the expectation of the data and these differences are quite big. Obviously, this is quite an ideal situation, and certainly not something we see often with real life data. Indeed, it is quite common for the data itself to be *concealed* with noise. Moreover, the differences in location paramaters could be a lot smaller, to the point where it is very difficult to discern any differences with the naked eye. An example of such data is displayed above.

The data contains relatively small *breaks* every 100 observations within the large noise-levels, so any change in the data becomes almost impossible to spot. This is of course an extreme example, but it is not difficult to imagine a more nuanced example where the noise still remains an issue. We also had the simplifying assumption that the changes referred to the expectations, which are in general easier to discern than other parameters such as the variance.

We have shown that the change-point problem can not always be easily solved in a direct manner, so we must turn to more formal means for our purposes.
From a computational perspective, the change-point problem can be formulated as an optimization problem:

$$\min_{|\tau|=K} V(\tau, y) = C = \sum_{k=0}^{K} c(y_{\tau_k..\tau_{k+1}}) \tag{1}$$

Here we assume that $\tau_0 = y_{t=1}$ and $\tau_{K+1} = T$, and this may also be a maximization problem. Moreover, the function c is what we call a cost function, it uses the assumed model for y_t to produce a quantitative measure that allows us to measure the goodness-of-fit of all subsets of the data as homogeneous segments. The more extreme the value of $V(\tau, y)$, the less/more well approximated the data is by the segmentation corresponding to τ(whichever it is depends on how c is defined). Therefore, our goal is to minimize(or maximize) $V(\tau, y)$, by considering all possible segmentations τ of the data. However, as we will later discuss, here we face another challenge since the set of possible segmentations becomes very large even when the length of the data has a modest size. So large in fact, that from a practical point of view it becomes simply impossible to find a solution (at least within a reasonable time frame). With this problem in mind, we will treat a search method that employs a non-traditional programming-technique called *dynamic programming*. So, like cost-functions, search methods are another crucial part of change-point problems.

Finally, before we can start solving the problem, we have to keep in mind that the above problem has two forms depending on whether the amount of changepoints K is known. The situation treated above corresponds to when K is known.

However, when K is unknown the problem becomes slightly more complicated as now we also have to concern ourselves with estimating K. We can not formulate the problem in the same manner as before, as this does not prevent *overfitting*: the amount of fitted change-points is far too large for the data considered. So, we have to make a slight adjustment to our optimization problem:

$$\min_{\tau\in\Omega} V(\tau, y) = \sum_{k=0}^{K} c(y_{\tau_k..\tau_{k+1}}) + \text{pen}(\tau) \tag{2}$$

This time, we have added an artificial penalty term, that as the name implies penalizes a certain segmentation based on the amount of change-points it

has. In this way, we try to strike a balance between overfitting and minimizing the total cost $V(\tau, y)$. Of course, the problems do not stop here. There are many extensions possible to the above problem, such as making the method robust against outlying values and dealing with non-parametric or dependent data.

Before we conclude this section, we shall summarize the change-point procedure as described above:

1) Firstly, an appropiate choice of the cost function c$(\cdot)$ needs to be made. The choice depends mainly on the underlying model of the data(parametric/non-parametric)

2) Secondly, we need to assume whether the amount of change-points K is known or unknown

3) Finally, depending on the choice we made in step 2, we solve optimization problem (1) or (2)

We will spend section 2 and 3 treating cost-functions/test-statistics for both the parametric as non-parametric models under the assumption of a single change-point. After that, we will start looking into search methods in section 4 to deal with steps 2 and 3.

2 Parametric model

As stated before, the cost-functions used in the optimization problem depend on the underlying model. There are many different models we can consider, but they can be roughly categorized into parametric and non-parametric models. In other words, we can either assume that the underlying distribution of the data is known(parametric) or not(non-parametric). Thus, we start with the parametric model, as this corresponds to the most basic model. More specifically, we will look at various distributions from the parametric family, and use different methods to derive cost-functions for the following problem:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_T = \theta$$

$$H_1 : \theta_1 = \cdots = \theta_k \neq \theta_{k+1} = \cdots = \theta_T$$

As mentioned, θ is the changing parameter that depends on the underlying model of y_t, and k is the hypothesized change-point location. Also, we only consider the case of a single change-point, but we extend this to multiple change-points when we consider search methods in section 4.

Thus, the question remains how we should apply the cost-functions to test the above hypotheses. In other words, how should we solve the change-point problem from a statistical perspective(unlike problems (1) and (2), which are only valid from a computational perspective)? First, we note that we will not directly apply the cost-function itself as a test-statistic, but rather combine them into a so-called *discrepancy*-function d:

$$d(y_{1..k}, y_{k+1..T}) = c(y_{1..T}) - [c(y_{1..k}) + c(y_{k+1..T})] \tag{3}$$

It is this discrepancy that we will use as a test-statistic for every case that we consider, since it directly compares the fit of the null and alternative model to the data(instead of substraction we may also use division). We then repeat this procedure for every possible change-point location, and eventually consider for which case the discrepancy is maximized, thus obtaining a change-point location estimate. Also, depending on the significance of the test, the critical values may be determined by considering the upper quantiles of the empirical distribution, which in turn is derived from a large number of simulations under H_0. The specific details will follow in section 2.3, where we conduct a simulation study to assess and compare the parametric cost-functions.

2.1 Normal distribution

We start with the normal distribution, as this is the most prevalent distribution in parametric change-point analysis. Let us assume that the data is normal and independently distributed, i.e. $y_t \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for t = 1,...,T and i = 1,..,K. So we need to consider two different parameters: the expectation μ and the variance σ^2.

2.1.1 Mean change

To familiarize the reader with the problem, we have displayed a figure below where a mean change is showcased. The approach that we consider here is termed the *likelihood* based approach, and it starts with some hypotheses about the change-point location. More specifically:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_T = \mu$$

$$H_1 : \mu_1 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_T$$

The change-point location is denoted with k, and note that we are testing for a single change-point only. However, as we will see later in our treatment of search methods such as binary segmentation, multiple change-point problems can easily be seen as a repeated occurence of a single change-point problem, so it suffices to consider the above hypothesis problem. We consider both the cases

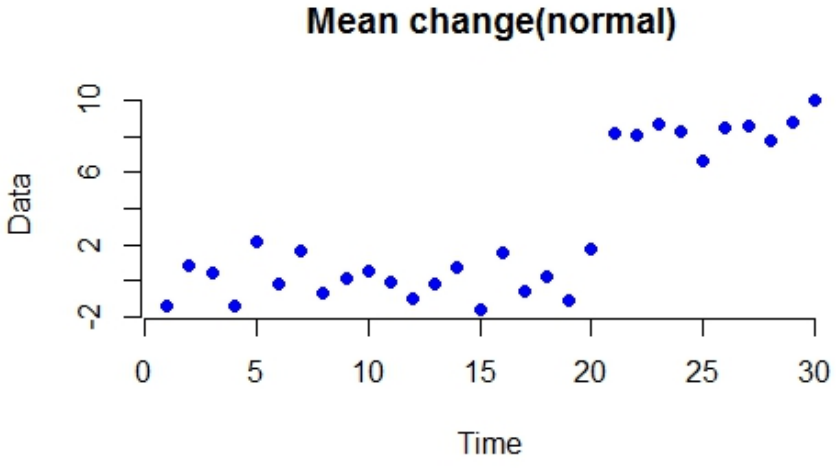


Figure 2: A plot of data with a change in mean at 20th observation

when the variance is known and not.

2.1.1.1 Known variance

In order to test the hypotheses that we have defined above, we will use the likelihood of the data under both hypotheses to derive a MLE-estimate for the mean, and correspondingly an appropiate statistic for detecting a mean-change(see [2], [3]). The technical details can be found in Appendix A1.1.

Here we used the assumptions that the data is normal and independently distributed, and for the alternative hypothesis the two different means are denoted with μ_1 and μ_T. The likelihood functions that were derived are the probabilities of the data occuring for the given values of μ and σ^2(null hypothesis) or μ_1, μ_T and σ^2(alternative hypothesis).

Using the MLE-estimate of the mean, we can derive a test statistic for testing a change. We define the following quantities:

$$S = \sum_{i=1}^{T} (y_i - \bar{y})^2 \tag{4a}$$

$$\hat{\mu}_1 = \overline{y}_k = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{4b}$$

$$\hat{\mu}_n = \overline{y}_{T-k} = \frac{1}{T-k} \sum_{i=k+1}^{T} y_i \tag{4c}$$

$$S_k = \sum_{i=1}^{k} (y_i - \overline{y}_k)^2 + \sum_{i=k+1}^{T} (y_i - \overline{y}_{T-k})^2 \tag{4d}$$

We let S and S_k represent the null and alternative hypotheses respectively. We can test for a change in mean by considering the difference between S and S_k :

$$M_k = S - S_k \tag{5}$$

Obviously, M_k provides a quantitative measure for the discrepancy between H_0 and H_1, i.e. no change-points or 1 change-point. The higher the value of M_k, the more likely it is that a change-point has occured at time k. Therefore, we wish to calculate M_k for all possible values of k and consider its highest value as a potential candidate for a change-point, i.e.: $M = \max_{1 \leq k \leq T-1} M_k$.

2.1.1.2 Unknown variance

A more practically viable approach can be derived when we assume that the variance is unknown. Here, we can for the most part use the likelihood like we did before, except now we also need a MLE estimate for the variance(see A1.2). The relevant likelihood functions are:

$$L_0(\mu, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma})^T} e^{-\sum_{t=1}^{T}(y_t-\mu)^2/2\sigma^2} \tag{6a}$$

$$L_1(\mu_1, \mu_T, \sigma_1^2) = \frac{1}{(\sqrt{2\pi\sigma_1})^T} e^{-\sum_{t=1}^{k}(y_t-\mu_1)^2/2\sigma_1^2 - \sum_{t=k+1}^{T}(y_t-\mu_T)^2/2\sigma_1^2} \tag{6b}$$

As before, when we assume that the change-point is at location k, then the corresponding MLEs are:

$$\hat{\mu}_1 = \overline{y}_k = \frac{1}{k}\sum_{t=1}^{k} y_t, \qquad \hat{\mu}_T = \overline{y}_{T-k} = \frac{1}{T-k}\sum_{i=k+1}^{T} y_t$$

and

$$\hat{\sigma}_1^2 = \frac{1}{T}\left[\sum_{t=1}^{k}(y_t - \overline{y}_k)^2 + \sum_{t=k+1}^{T}(y_t - \overline{y}_{T-k})^2\right].$$

Now, we could fill in these MLEs into equations (6) and consider their discrepancy, but to make it a bit more manageable we instead adopt a method from [21], where again as quantitative measures for the hypotheses we have:

$$S = \sum_{t=1}^{T}(y_t - \overline{y})^2 \ , \ S_k = \sum_{t=1}^{k}(y_t - \overline{y}_k)^2 + \sum_{t=k+1}^{T}(y_t - \overline{y}_{T-k})^2.$$

We will need to define one more quantity before we can form the test statistic from these cost-functions:

$$E_k = k(\overline{y}_k - \overline{y}_T)^2 + (T-k)(\overline{y}_{T-k} - \overline{y}_T)^2$$

It is easily verifiable through some simple algebra that $S = S_k + E_k$. Finally, using these quantities we can define our likelihood based test statistic as follows:

$$C = \max_{1\le k\le T-1} \frac{S}{S_k} = \max_{1\le k\le T-1} 1 + \frac{E_k}{S_k} = \max_{1\le k\le T-1} 1 + N_k \tag{7}$$

This is equivalent to: $N = \max_{1\le k\le T-1} N_k = \max_{1\le k\le T-1} \frac{E_k}{S_k} = \max_{1\le k\le T-1} \frac{E_k}{(S-E_k)}$.

2.1.2 Variance change

Now that we have shown how to treat a change in mean, we can similarly derive how to detect a change in variance. We will treat two different methods for this: first we will look at the likelihood-based approach, which will be quite similar as before. Secondly, we will derive an alternative, model selection-based method called the model-selection based approach. Again, we show a figure below where a variance change is showcased.

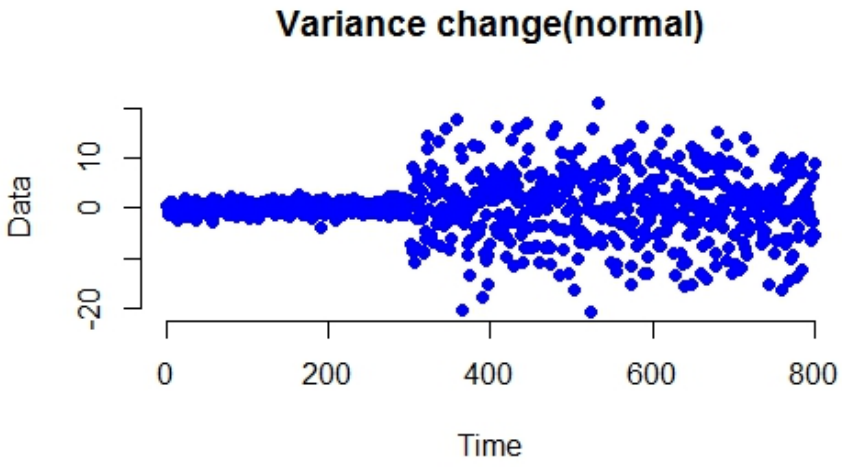


Figure 3: A plot of data with a change in variance at 300th observation

2.1.2.1 Likelihood approach

Assume as before that the data is normally and independently distributed, i.e. $y_t \sim \mathcal{N}(\mu, \sigma_i^2)$ for $t = 1, ..., T$ and $i = 1, ..., K$. Here, μ is assumed to be known. As with the change in mean, we start with some hypotheses :

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_T^2 = \sigma^2$$

$$H_1 : \sigma_1^2 = \cdots = \sigma_{\tau_1}^2 \neq \sigma_{\tau_1+1}^2 = \cdots = \sigma_{\tau_2}^2 \neq \cdots \neq \sigma_{\tau_K+1}^2 = \cdots = \sigma_T^2$$

Here we test for multiple change-points, where $\tau = \{\tau_1, \tau_2, ..., \tau_K\}$ denotes the change-point locations and K the (unknown) amount of change-points. However, as we explained before, single change-point problems are a better starting

point and can be easily extended to multiple change-points:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_T^2 = \sigma^2$$
$$H_1 : \sigma_1^2 = \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_T^2$$

We will use the log-likelihood function, $l(\mu, \sigma^2; y_1, ..., y_T)$, and the MLE of the variance to derive a cost-function, and with it a test-statistic. Indeed, we can plug in the MLE into the log-likelihood and consider its difference under H_0 and H_1 to measure the discrepancy, just as we did when testing for a change in mean. More specifically, under H_0 we have:

$$l(\sigma^2) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i - \mu)^2$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{T}(y_i - \mu)^2}{T}$$

This leads to a maximum log-likelihood:

$$l(\hat{\sigma}^2) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln \hat{\sigma}^2 - \frac{T}{2} \tag{8}$$

Likewise, we can compute the maximum log-likelihood under the alternative hypothesis H_1:

$$l(\sigma_1^2, \sigma_T^2) = -\frac{T}{2}\ln 2\pi - \frac{k}{2}\ln \sigma_1^2 - \frac{T-k}{2}\ln \sigma_T^2 - \frac{\sum_{i=1}^{k}(y_i - \mu)^2}{2\sigma_1^2} - \frac{\sum_{i=k+1}^{T}(y_i - \mu)^2}{2\sigma_T^2},$$

where σ_1^2 and σ_T^2 are the variances before and after the change respectively. Let $\hat{\sigma}_1^2$, $\hat{\sigma}_T^2$ denote their respective MLEs, then:

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{k}(y_i - \mu)^2}{k}, \quad \hat{\sigma}_T^2 = \frac{\sum_{i=k+1}^{T}(y_i - \mu)^2}{T-k}$$

This eventually gives us a maximum log-likelihood of:

$$l(\hat{\sigma}_1^2, \hat{\sigma}_T^2) = -\frac{T}{2}\ln 2\pi - \frac{k}{2}\ln \hat{\sigma}_1^2 - \frac{T-k}{2}\ln \hat{\sigma}_T^2 - \frac{T}{2} \tag{9}$$

Subtracting these maximum likelihoods/cost-functions from each other, and discarding all constants that do not depend on the observations, we obtain at last the following test statistic for a change in variance:

$$\lambda_{var} = \max_{1<k<T-1} [T\ln \hat{\sigma}^2 - k\ln \hat{\sigma}_1^2 - (T-k)\ln \hat{\sigma}_T^2] \tag{10}$$

Again, our best candidate for a change-point is located at the timepoint k such that the above statistic is maximized. For multiple change-points, say k, we look at the k-highest values of λ_{var}. In this sense, the approach is identical to that of detecting a change in mean.

2.1.2.2 Model-selection based approach

We will now look at another method for change-point inference, the so-called *model-selection based approach*(see [2], [6]).

To put it shortly, the model-selection based approach is based on selecting the best fitting model out of a set of possible models. Here the models conform to either no or 1 change-point. To be more specific, we have 1 model for the case of no changpoints and a set of other models for 1 change-point, which all correspond to a specific change-point location. By using model evaluating quantities, we can search for the model that maximizes/minimizes these quantities. A natural, first choice for such a measure is the *Akaike Information Criterion*[7] or AIC for short:

$$AIC(m) = -2\ln L(\hat{\Theta}_m) + 2p_m, \;\; m = 1, 2, ..., M \tag{11}$$

Here, $L(\hat{\Theta}_m)$ denotes the maximum likelihood of model m, which obviously serves as a model evaluation. Furthermore, p_m is the amount of parameters in model m. The parameters depend on the type of change-point problem, so for detecting a single change in the variance while the mean is known, there is either one parameter(variance) under H_0 or there are two relevant parameters: σ_1^2 and σ_T^2(the variances before and after the change-point) under H_1. We are interested in the model for which AIC(m) is minimized, or MAICE (Minimum AIC Estimate).

However, as Schwarz(1978)[8] has shown, the MAICE may not always give accurate results, especially when the data becomes large. Therefore, we will use an alternative measure that Schwarz devised, called the Schwartz Information Criterion or SIC :

$$SIC_m = -2\ln L(\hat{\Theta}_m) + p_m \ln T, \;\; m = 1, 2, ..., M \tag{12}$$

It is identical to AIC, except the second term has been altered. So how

exactly can we find the change-point in this setup? First off, we need to calculate the corresponding SIC-values under all the possible models, which we will denote by SIC_{var}(T)[no change-points] and SIC_{var}(k)[1 change-point] respectively:

$$\begin{aligned}\mathrm{SIC}_{var}(T) &= -2\ln L_0(\mu, \hat{\sigma}^2) + \ln T \\ &= T\ln 2\pi + T\ln \hat{\sigma}^2 + T + \ln T\end{aligned}$$

$$\begin{aligned}\mathrm{SIC}_{var}(k) &= -2\ln L_1(\mu, \hat{\sigma}^2) + 2\ln T \\ &= T\ln 2\pi + k\ln \hat{\sigma}_1^2 + (T-k)\ln \hat{\sigma}_T^2 + T + 2\ln T\end{aligned}$$

For the above formulae we used the maximum log-likelihoods that we derived before. Note that SIC_{var}(k) will have to be recalculated for every possible change-point location k, and that due to our MLEs the possible locations for k are constrained between 2 and T - 2. Once all these values are calculated, we calculate the minimum value and make one of the following conclusions:

• If $\mathrm{SIC}_{var}(T) < \min_{2\leq k\leq T-2} \mathrm{SIC}_{var}(k)$, then the model describing no change-points is the best fit, so we may conclude that there are no change-points.

• If there is a k such that $\mathrm{SIC}_{var}(T) > \mathrm{SIC}_{var}(k)$, then we reject this first model and estimate the change-point location by $\hat{k}$ such that $SIC_{var}(\hat{k}) = \min_{2\leq k\leq T-2} \mathrm{SIC}_{var}(k)$

In order to bring this method more in line with our standard hypothesis testing procedure, we also require proper critical values c_α(section 2.3) which will serve to gauge significance. After all, if the SIC-values are closely located, then the conclusions above may not be valid anymore as the results are likely to be caused by random fluctuations of the data instead of an actual change. Therefore, we adopt a method from [2], where it was shown that the critical values may be estimated as:

$$c_\alpha \simeq \left\{ -\frac{1}{a(\log T)} \log\log[1-\alpha+\exp(-2e^{b(\log T)})]^{-\frac{1}{2}} + \frac{b(\log T)}{a(\log T)} \right\}^2 - \log T \tag{13}$$

where:

$$a(\log T) = (2\log\log T)^{\frac{1}{2}} \tag{14a}$$

$$b(\log T) = 2\log\log T + \frac{1}{2}\log\log\log T - \log\Gamma(\frac{1}{2}) \tag{14b}$$

Here, Γ is the so-called gamma-function(see (15)). Using these critical values, we now accept the model of no change-points when $\mathrm{SIC}_{var}(T) < \min_{2\leq k\leq T-2} \mathrm{SIC}_{var}(k) + c_\alpha$, so it is harder to reject this model now. An important

reason for defining this alternative model, is that it is more capable of detecting change-points under the presence of outliers(see the simulation study in section 2.3).

2.2 Gamma distribution

Another well-known distribution from the parametric family that we will look at is the gamma model. The underlying assumptions that we will make about the data are identical to the normal model that we considered before, namely that the data are independently and identically distributed. More formally:

$$y_t \sim \frac{1}{\theta_t^{\xi}\Gamma(\xi)} y_t^{\xi-1} e^{-(\frac{y_t}{\theta_t})}, \quad \xi, \theta_t > 0, y_t > 0, t = 1, ..., T$$

Here, ξ and θ are the two parameters, which we will call the shape and scale respectively. Furthermore, $\Gamma(\xi)$ is the so-called gamma function:

$$\Gamma(\xi) = \int_0^{\infty} x^{\xi-1} e^{-x} dx \tag{15}$$

For our change-point problem we will consider both the scale and the shape, while simultaneously assuming that the other parameter is known. As for practical uses, the gamma distribution itself encompasses some well-known distributions, for example the exponential and chi-square distributions are special cases of the general gamma family. As a result, the gamma distribution is prominently featured in areas such as finance, reliability studies, survival analysis and other relevant disciplines(see [2]). Therefore, it has become an important topic in the last decades to sufficiently detect any changes that might occur in gamma-distributed variables. For a specific example, the shape parameter within a reliability problem directly controls the failure rate(it is decreasing, increasing or constant whenever ξ - 1 is negative, positive or zero respectively).

The first change-point problem that we will look at is the general case, so we will consider the scale and shape parameters separately. Our procedure here is almost identical as for the normal case: we will still derive a solution using the likelihood approach, but with a Bayesian element.

2.2.1 Scale

Assume that ξ is known, we will try to determine when a change in θ takes place. Obviously, the first step to accomplishing this is by defining appropiate hypothesis tests as before:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_T = \theta_0$$

$$H_1 : \theta_1 = \cdots = \theta_k = \theta_0 \neq \theta_{k+1} = \cdots = \theta_T = \theta_0 + \delta$$

Here, k is the unknown change-point position, θ_0 is unknown and δ is chosen such that $\mid \delta \mid > 0$ and $\theta_0 + \delta > 0$.

In order to become more acquainted with the scale-change problem, we have plotted 2 figures below.

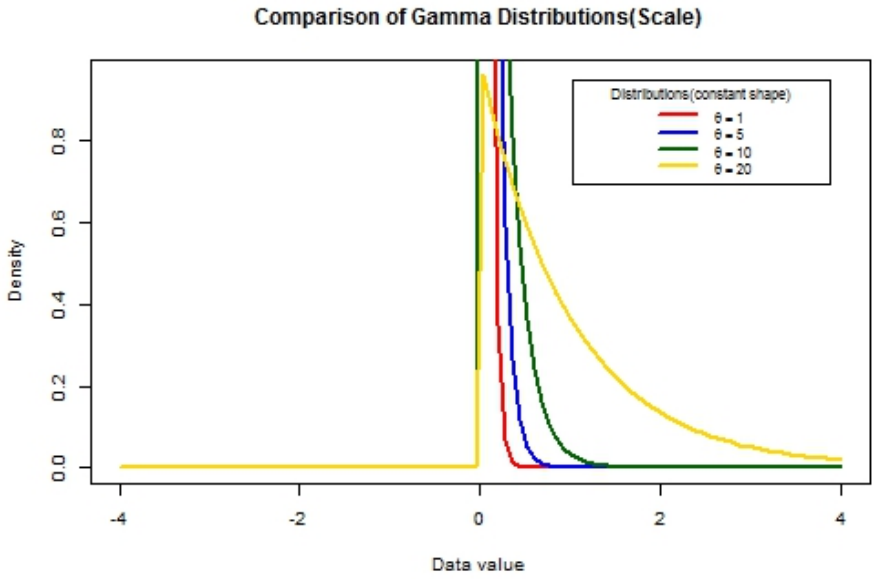


Figure 4: A plot of various gamma-densities with differing scales but constant shape

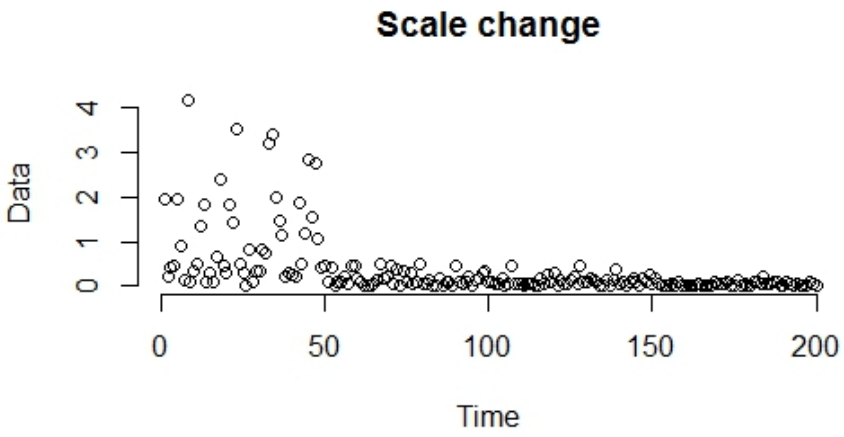


Figure 5: A plot of data with a decrease in scale at every 50th observation

As is apparent, the data becomes more "condensed" as the scale decreases. The problem generally arises when the differences in "density" are more subtle, to the point where directly trying to spot the change-points becomes an impossible task. So we will treat a systematic approach, namely the likelihood/Bayes-based approach[4].

2.2.1.1 Likelihood/Bayes based approach(Scale)

We mostly repeat the same procedure as for the normal model, as in we will calculate the likelihood under both hypotheses, and use these to define an appropiate test statistic(the precise technical details can be found in A2.1). However, before we can do so, we need to make an additional assumption about the *a priori* distribution of the change-point. We will assume that every possible change-point location is equally likely to occur, i.e. for $k = 1, 2, ..., T-1$ the probability of a change-point at k is $\frac{1}{T-1}$. This *a priori* assumption will be denoted with $\pi_T(j)$:

$$\pi_T(j) = \begin{cases} \frac{1}{T-1}, & j = 1, 2, ..., T-1 \\ 0, & \text{otherwise} \end{cases}$$

For the null hypothesis, it is quite straightforward to calculate the likelihood(the

precise derivations may be found in appendix A2):

$$L_0(\theta_0) = \frac{\prod_{t=1}^{T} y_t^{\xi-1}}{(\Gamma(\xi))^T} \exp\left[\sum_{t=1}^{T}\left(-\frac{y_t}{\theta_0} - \ln\theta_0^{\xi}\right)\right] \tag{16}$$

Likewise, under H_1 we have:

$$L_1(\theta_0,\delta) = \frac{1}{T-1}\frac{\prod_{t=1}^{T} y_t^{\xi-1}}{\Gamma^T(\xi)}\sum_{j=1}^{T-1}\exp\left[\sum_{t=1}^{j}\left(-\frac{y_t}{\theta_0} - \ln\theta_0^{\xi}\right)\right]\cdot\exp\left[\sum_{t=j+1}^{T}\left(-\frac{y_t}{\theta_0+\delta} - \ln(\theta_0+\delta)^{\xi}\right)\right]$$

The test statistic is based on the likelihood-ratio of the above hypotheses, so in order to simplifiy our derivation a bit we will approximate the second term of $L_1(\theta_0,\delta)$ with a Taylor expansion as $(\delta/\theta_0) \to 0$:

$$-\frac{y_t}{\theta_0+\delta} - \ln(\theta_0+\delta)^{\xi} = -\frac{y_t}{\theta_0} - \ln\theta_0^{\xi} + \delta\left(\frac{y_t}{\theta_0^2} - \frac{\xi}{\theta_0}\right) + o\left(\frac{\delta}{\theta_0}\right)$$

Thus, our likelihood under H_1 becomes:

$$L_1(\theta_0,\delta) = \frac{1}{T-1}\frac{\prod_{t=1}^{T} y_t^{\xi-1}}{\Gamma^T(\xi)}\sum_{j=1}^{T-1}\left\{\exp\left[\sum_{t=1}^{j}\left(-\frac{y_t}{\theta_0} - \ln\theta_0^{\xi}\right)\right]\cdot\exp\left[\sum_{t=j+1}^{T}\left(-\frac{y_t}{\theta_0} - \ln\theta_0^{\xi} + \frac{y_t\delta}{\theta_0^2} - \frac{\xi\delta}{\theta_0} + o(\delta)\right)\right]\right\} \tag{17}$$

As stated, in order to conduct a change-point analysis we will use a likelihood-ratio:

$$\Lambda = \frac{L_1(\theta_0,\delta)}{L_0(\theta_0)} = 1 + \frac{\delta}{\theta_0}\left[\frac{1}{(T-1)\theta_0}\sum_{j=1}^{T-1}\sum_{t=j+1}^{T} y_t - \frac{T\xi}{2}\right] + o(\delta) \tag{18}$$

However, we can simplify our calculations a bit by only partly considering the above expression. Since we are dealing with a likelihood ratio, it suffices to only consider the terms depending on the observations, namely $\frac{1}{\theta_0}\sum_{j=1}^{T-1}\sum_{t=j+1}^{T} y_t$. Our new likelihood-ratio based test statistic is thus:

$$\lambda = \frac{1}{\theta_0}\sum_{j=1}^{T-1}\sum_{t=j+1}^{T} y_t$$

$$
\begin{aligned}
&= \frac{1}{\theta_0} \sum_{t=2}^{T} (t-1) y_t \\
&= \frac{1}{\theta_0} \sum_{t=1}^{T-1} t y_{t+1} \\
&= \frac{1}{\theta_0} \sum_{t=1}^{T} (t-1) y_t
\end{aligned}
$$

Filling in the MLE of the scale, finally gives us the following statistic:

$$
\lambda_{sc} = \frac{T\xi}{\sum_{t=1}^{T} y_t} \sum_{t=1}^{T} (t-1) y_t \tag{19}
$$

2.2.2 Shape

We consider the reversed situation, where a change has taken place in ξ and θ remains constant. This leads to the following hypotheses:

$$
\begin{aligned}
H_0 &: \xi_1 = \xi_2 = \cdots = \xi_T = \xi_0 \\
H_1 &: \xi_1 = \cdots = \xi_k = \xi_0 \neq \xi_{k+1} = \cdots = \xi_T = \xi_0 + \delta
\end{aligned}
$$

Again, we assume that the value of the parameter is unknown. The terminology that we use here is also identical to the previous problem. In order to clarify what a change in shape entails, we have displayed two figures below:

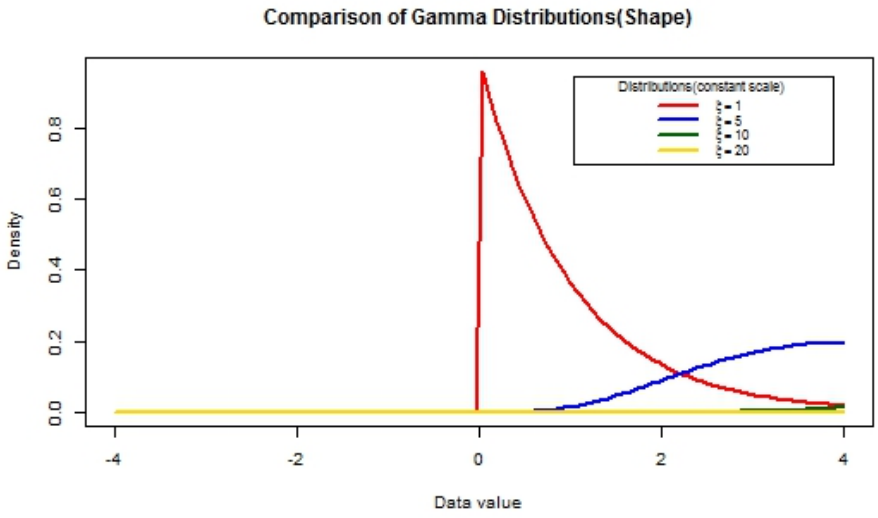


Figure 6: A plot of various gamma-densities with differing shapes but constant scale

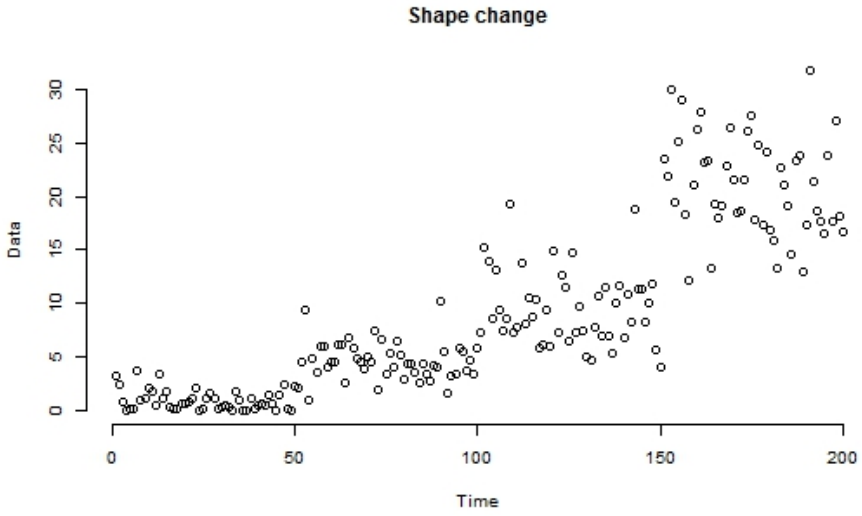


Figure 7: A plot of data with an increase in shape at every 50th observation

It is clear that increasing the scale both disperses the data as increases its absolute value, though the example considered here is an obvious one. The problem of course remains the same as we considered it before, so in the same

way we will again treat a likelihood/Bayes-based approach[5].

2.2.2.1 Likelihood/Bayes-based approach(Shape)

The procedure remains the same, but as we shall see the details will vary a bit. Again, we first calculate the likelihood under the above alternative hypothesis, which turns out can be written as(see A.2.2):

$$L_1(y_t \mid \xi_0, \delta, \theta, k) = \prod_{t=1}^{T} \frac{1}{\theta^{\xi_t}\Gamma(\xi_t)} y_t^{\xi_t - 1} e^{-(\frac{y_t}{\theta})}$$

$$= (T-1)^{-1}\exp\left\{-\sum_{t=1}^{T}\frac{y_t}{\theta} + \eta(y_t;\xi,\theta)\right\}\cdot\sum_{k=1}^{T-1}\left\{1+\delta\sum_{t=k+1}^{T}[\ln y_t - \ln\theta - \Psi(\xi) + o(\delta)]\right\},$$

as $\delta \to 0$.

We are now in the position to derive a convenient expression for our likelihood-ratio(when $\delta \to 0$):

$$\Lambda = \frac{L_1(y_t \mid \xi_0, \delta, \theta)}{L_0(\xi_0, \delta)}$$

$$= \frac{(T-1)^{-1}\exp\left\{-\sum_{t=1}^{T}\frac{y_t}{\theta} + \eta(y_t;\xi,\theta)\right\}\cdot\sum_{k=1}^{T-1}\left\{1+\delta\sum_{t=k+1}^{T}[\ln y_t - \ln\theta - \Psi(\xi) + o(\delta)]\right\}}{\frac{\prod_{t=1}^{T} y_t^{\xi-1}}{\Gamma^T(\xi)}\exp\left[\sum_{t=1}^{T}\left(-\frac{y_t}{\theta_0} - \ln\theta_0^{\xi}\right)\right]}$$

$$= 1 + \frac{\delta}{T-1}\sum_{k=1}^{T-1}\sum_{t=k+1}^{T}[\ln y_t - \ln\theta - \Psi(\xi)] + o(\delta)$$

$$= 1 + \frac{\delta}{T-1}\sum_{t=1}^{T-1} t\ln y_{t+1} - \frac{\delta T}{2}\{\Psi(\xi) + \ln\theta\} + o(\delta)$$

In order to further simplify this expression, we will fill in the MLE-estimate under H_0 for the term $\Psi(\xi)+\ln\hat{\theta}$ where $\hat{\theta}$ is the previously derived MLE-estimate for the scale θ. The peculiarity here is that while an analytic MLE can be derived for the specified term, this is not possible for the shape itself.
In the end, we obtain the following MLE-estimator:

$$\Psi(\xi_0) + \ln\hat{\theta} = \frac{1}{T}\sum_{t=1}^{T}\ln y_t \tag{20}$$

Finally, returning to the likelihood-ratio from before we can now fill in (17) to obtain:

$$
\begin{aligned}
\Lambda &= \frac{L_1(y_t \mid \xi_0, \delta, \theta)}{L_0(\xi_0, \delta)} \\
&= 1 + \frac{\delta}{T-1} \sum_{t=1}^{T-1} t \ln y_{t+1} - \frac{\delta T}{2} \{\Psi(\xi) + \ln\theta\} + o(\delta) \\
&= 1 + \frac{\delta}{2(T-1)} \sum_{t=1}^{T} 2(t-1) \ln y_t - \frac{\delta}{2(T-1)} \sum_{t=1}^{T} (T-1) \ln y_t + o(\delta) \\
&= 1 + \frac{\delta}{T-1} \sum_{t=1}^{T-1} t \ln y_{t+1} - \frac{\delta}{2} \sum_{t=1}^{T} \ln y_t + o(\delta) \\
&= 1 + \frac{\delta}{2(T-1)} \sum_{t=1}^{T} (2t - T - 1) \ln y_t + o(\delta)
\end{aligned}
$$

Hence, we take the monotonic part and use it as our test statistic for detecting changes in shape:

$$
\lambda_{sh} = \left| \sum_{t=1}^{T} (2t - T - 1) \ln y_t \right| \tag{21}
$$

2.3 Simulation study(Parametric)

Now that we have defined the relevant cost-functions, we need to assess their performance for a broad range of practical possibilities. As we have explained before, we will be assessing the discrepancy function formed from the cost-functions, and use this as a test-statistic instead. The simulations will generally occur in two phases: in the first phase, we will empirically derive the critical values under the null-hypothesis, by repeatedly simulating data with no change-points and assessing the upper quantiles of the empirical distribution formed from the calculated discrepancy values. These critical values will in turn be used for the second phase where we estimate the power and change-point location $\hat{\tau}$ for data that does actually have a change-point(alternative hypothesis). In essence, we will consider the effect from the following parameters on the power and $\hat{\tau}$:

- T, the length of the data.
- δ, the size of the change itself.
- τ , the actual location of the change-point.

All the simulations that we conduct from now on, will for both phases have a size of 10000 with a significance level of $\alpha = 0.05$. Aside from these parameters, we will also study the effect of some more general changes. First we study

the performance under the ideal situation of normally distributed data, after which we slightly deviate to the student-t distribution. Not only does this allow us to study the sensitivity towards a slight change in distributional form, but also the robustness of the test-statistics against outlying values(the student-t distribution has heavier tails). The R code for all subsequent simulations can be found at the link displayed below[1].

2.3.1 Normal distribution

First off, we start with the ideal situation where the data is normally distributed. In order to study the effect of the mentioned parameters, we only alternate in one given parameter while keeping the others constant. The relevant table is shown below.

Parameters			Mean change				Variance change			
T	δ	τ	Power(M)	$\hat{\tau}_M$	Power(N)	$\hat{\tau}_N$	Power(λ_{var})	$\hat{\tau}_{\lambda_{var}}$	Power(SIC$_{var}$)	$\hat{\tau}_{SIC_{var}}$
30	0.1	5	0.045	13.39(10.14)	0.05	13.99(10.21)	0.771	5.28(3.25)	0.6	5.03(2.41)
		10	0.056	14.62(9.93)	0.041	14.5(9.61)	0.92	9.73(2.72)	0.79	9.5(2.24)
		15	0.0572	15.31(9.54)	0.061	15.7(9.22)	0.94	14.37(2.61)	0.7985	14.28(2.42)
	1.1	5	0.398	6.71(4.89)	0.33	6.58(4.74)	0.051	14.4(10.26)	0.007	15.94(8.6)
		10	0.59	10.77(4.32)	0.51	10.65(3.94)	0.06	16.34(10.32)	0.008	15.9(8.7)
		15	0.67	15.11(3.81)	0.62	15.07(3.62)	0.04	15.13(10.37)	0.005	13.9(8.5)
	2	5	0.92	5.38(2.14)	0.87	5.33(1.92)	0.078	13.1(9.57)	0.02	11.28(7.9)
		10	0.99	10.11(1.69)	0.98	10.09(1.61)	0.13	12.96(7.99)	0.02	12.41(6.73)
		15	0.998	15.002(1.55)	0.99	15(1.55)	0.13	14.78(7.93)	0.03	15.18(6.45)
100	0.1	5	0.0436	46.62(35.64)	0.048	43.93(35.67)	0.8388	5.97(8.66)	0.7311	5.51(5.65)
		25	0.047	50.56(36.1)	0.043	49.53(35.3)	1	24.4(2.4)	1	24.3(2.5)
		50	0.061	49.69(33.28)	0.0656	48.7(33.68)	1	49.2(2.48)	1	49.3(2.46)
	1.1	5	0.38	11.42(18.4)	0.36	11.36(18.49)	0.049	42.34(36.71)	0.01	44.97(37.17)
		25	0.98	25.71(6.44)	0.97	25.69(6.44)	0.072	46.99(38.6)	0.012	48.86(35.51)
		50	0.9976	50.04(5.53)	0.9964	50.054(5.52)	0.066	53.9(37.51)	0.01	52.62(34.2)
	2	5	0.94	5.73(5.24)	0.93	5.69(5.15)	0.079	35.22(35.09)	0.02	32.72(31.6)
		25	1	25.04(1.45)	1	25.04(1.45)	0.269	33.23(22.05)	0.1039	31.81(17.8)
		50	1	49.97(1.37)	1	49.97(1.37)	0.4188	51.85(17.7)	0.2055	52.1(14.15)

Table 1: A showcase of the power and change-point estimation, when the data changes from $\mathcal{N}(0,1)$ to $\mathcal{N}(\delta,1)$[mean change] or $\mathcal{N}(0,\delta)$[variance change] at change-point location τ. The values in parentheses indicate standard deviations.

[1] https://github.com/StudBch96/Bachelor-project-R-code

The first conclusion that we can draw, is that all the statistics considered seem to work better when we move τ towards the middlepoint of the data. Whenever τ is on the extreme end, we see that the performance may significantly worsen if the change is not big enough, regardless of T. Thus under the condition that τ is not close enough to the extreme ends, we may also make the obvious observation that whenever T is bigger or δ differs more, the performance will improve. As for the statistics themselves, we see that M and N provide quite similar performance when $T = 100$. However, when $T = 30$ the differences become more pronounced at a reasonable change size($\delta = 1.1$). Here, M provides a somewhat better power, which is to be expected as N was derived under the additional assumption that the variance was unknown. Furthermore, it can clearly be observed that SIC_{var} in general has a lower power than λ_{var}, though when the change is more drastic($\delta = 0.1$ or 2) it does tend to provide better τ -approximations.

2.3.2 Student-t distribution

Now we have simulated data with the student-t distribution, which under H_0 has a mean of 0 and 3 degrees of freedom(df). We also introduce two separate parameters, δ_1 and δ_2, to denote the change-sizes for the mean and degrees of freedom(df) respectively. For the variance change, we have to keep in mind that $\delta_2 = -2$ is the most drastic change, followed by the values 100 and 20. In every possible situation we consider here, the data will have more outlying values than the normal distribution we considered first.

We can immediately see from the table below, that the statistics in general have worse performance. In particular, looking at M and N, the only case for which they have good performance is when $T = 100$, $\delta_1 = 2$ and τ is far enough from the edge. Otherwise they have subpar performance at best, though it would appear that N does have somewhat better performance when δ_1 is not drastically small.

As before, the variance statistics do have more glaring differences. If we consider the power, then SIC_{var} is always superior in that respect, which is the opposite of when the data was normally distributed. Still, even SIC_{var} only performs well when the change is very drastic(δ_2 = -2). Furthermore, it can easily be seen that change-point approximation is very bad in all situations, so neither statistic is suited for precisely locating where a change has occured. All in all, an easy conclusion that we can draw is that SIC_{var} is more robust against outliers than λ_{var}.

Parameters				Mean change				Variance change			
T	δ_1	δ_2	τ	Power(M)	$\hat{\tau}_M$	Power(N)	$\hat{\tau}_N$	Power(λ_{var})	$\hat{\tau}_{\lambda_{var}}$	Power(SIC$_{var}$)	$\hat{\tau}_{SIC_{var}}$
30	0.1	-2	5	0.049	15.7(11.8)	0.051	15.6(12.6)	0.55	16.4(6.5)	0.76	15.8(6.72)
			10	0.045	15.7(11.9)	0.048	15.2(12.7)	0.57	17.6(5.7)	0.76	16.9(5.8)
			15	0.05	15.1(11.8)	0.047	15.4(12.8)	0.54	19.95(4.7)	0.71	19.3(4.9)
	1.1	20	5	0.079	11.95(10.4)	0.1	9.5(9.4)	0.032	3.5(2.4)	0.08	5.4(5.4)
			10	0.11	12.5(8.9)	0.19	11.6(7.2)	0.04	5.6(3.2)	0.13	7.4(4.7)
			15	0.12	14.9(8.2)	0.18	14.98(6.9)	0.049	7.8(4.5)	0.15	9.53(5.4)
	2	100	5	0.19	7.96(7.4)	0.33	6.3(5.2)	0.033	3.4(2.3)	0.09	4.9(4.5)
			10	0.44	10.9(4.6)	0.57	10.6(3.8)	0.05	5.9(3.3)	0.14	7.2(4.68)
			15	0.53	14.95(3.99)	0.63	14.96(3.4)	0.054	7.7(4.7)	0.16	9.31(5.3)
100	0.1	-2	5	0.044	48.7(43.7)	0.05	51.08(45.4)	0.74	51.2(24.6)	0.98	49.4(25.9)
			25	0.054	52.2(42.99)	0.052	51.5(45.8)	0.85	49.7(20.97)	0.9937	48.1(21.2)
			50	0.048	48.6(42.9)	0.054	51.3(45.2)	0.86	63.02(13.8)	0.9923	62.2(14.94)
	1.1	20	5	0.069	37.4(41.1)	0.082	30.8(40.2)	0.017	3.3(1.8)	0.12	9.9(18.7)
			25	0.3	30.4(20.9)	0.41	28.6(17.9)	0.059	14.6(7.7)	0.35	19.4(13.3)
			50	0.47	50.04(15.4)	0.61	49.99(14.01)	0.065	26.4(15.7)	0.47	34.9(17.2)
	2	100	5	0.18	17.7(29.2)	0.28	11.95(22.6)	0.019	3.6(1.6)	0.11	8.35(16.5)
			25	0.95	25.8(7.8)	0.92	25.6(7.02)	0.064	15.1(7.7)	0.39	18.8(11.9)
			50	0.99	49.98(6.2)	0.97	49.96(5.8)	0.076	27.5(15.4)	0.51	35.7(16.6)

Table 2: A showcase of the power and change-point estimation when the data follows a student-t distribution. The data initially has a mean of zero and 3 df(degrees of freedom), and after the change has a mean of δ_1 or $3 + \delta_2$ df.

2.3.3 Robustness

As we saw, the parametric statistics performed significantly worse when the data was heavier-tailed. In other words, they are not robust against outlying values, which is to be expected as they were derived under the assumption of normality. It is of practical importance that change-points, even under these more extreme circumstances, can be reliably detected. Though it is always possible to screen the data before-hand and manually pick out the outliers, this might be unpreferable due to time constraints. We will therefore discuss two adjustments to the cost-functions in order to make them more robust.

The first adjustment we could make is to simply replace the estimators of the relevant parameters with more robust versions. For example, while deriving M and N we used the MLE-estimator of the mean:

$$\hat{\mu} = \frac{\sum y_t}{T}$$

But erroneously big values of the data could easily have a likewise big influence on the numerator of the MLE-estimator, thus any cost-functions using the estimator could similarly be biased towards the outlier. So we would like to replace the MLE-estimator with a more robust one. Fortunately, such an estimator is easy to define: the *median*. It simply orders the data based

Parameters			Mean change			
T	δ_1	τ	Power(M)	$\hat{\tau}_M$	Power(N)	$\hat{\tau}_N$
100	0.1	5	0.06	55.6(39.74)	0.058	54.88(39.67)
		25	0.054	56.06(40.19)	0.06	50.07(41.71)
		50	0.059	57.06(39.76)	0.054	53.14(40.79)
	1.1	5	0.11	34.95(38.99)	0.13	25.37(33.52)
		25	0.5	29.02(16.24)	0.53	29.4(16.05)
		50	0.74	50.52(13.36)	0.65	50.41(12.49)
	2	5	0.38	12.37(21.38)	0.42	10.65(16.32)
		25	0.99	25.65(6.57)	0.96	27.53(6.9)
		50	1	49.97(4.97)	0.98	50.04(5.11)

Table 3: Power and change-point approximation when using robust estimators and bounded cost-functions(with $P = 5$). The data follows a student-t distribution as before.

on size(see (23)), and picks out the point in the middle(or the average of the middle-points when the data-length is even) as an estimator for the mean. In this way, large values will be assigned the highest rank and their effect will be severely mitigated.

Aside from robust estimators, an additional improvement follows from the observation that the cost-functions are unbounded. Even though a robust estimator such as the median might perform well in the presence of one outlier, it generally breaks down when the amount of outliers becomes too large. Thus, additional robustness could be obtained by somehow making the cost-functions bounded. This is also not hard to do, as it simply requires us to impose additional conditions(denote the cost function as c):

$$c(y_t, \theta) = \begin{cases} c(y_t, \theta), & \text{for } \mid y_t - \theta \mid < P \\ C, & \text{for } \mid y_t - \theta \mid \geq P \end{cases}$$

Here, the maximum admissable outlier is bounded at P, if it exceeds this then c gets assigned the fixed constant C. So no matter the amount of outliers, if a true change causes the cost-function to exceed the value C, then the method should in theory become completely robust against outliers. However, the tricky part lies in correctly choosing the value P. If on one hand P is chosen too large, then we are being too lenient and if P is chosen too small, then we might needlessly disregard useful data. With both of these adjustments applied, we show the performance for the mean-change statistics in table 3.

We see that unless $\delta_1 = 0.1$, the results have indeed improved both in the form of a higher power and lower standard deviations. Especially M shows a strong improvement, since it was particularly afflicted by the outliers due to its quadratic nature.

2.3.4 Bonferroni-correction

Aside from robustness, there is also an inherent problem with the simulations themselves. Basically, we try to find a change-point by consequtively conducting hypothesis tests for every point between 2 and T - 2. If these tests have a significance level of α, then the odds of wrongfully rejecting the null hypothesis is also α(the so-called Type-I error). So, if one is conducting a series of similar tests, then this chance of a false positive quickly adds up. For our example, when we conduct T - 3 tests then the chance of at least detecting one false positive is: $1 - (1 - \alpha)^{T-3}$. For $T = 100$ and $\alpha = 0.05$, this value is approximately 0.99.

To counter-act this, we will apply the so-called Bonferroni-correction to our critical values. This is a rather conservative method where we simply reduce the original α to $\frac{\alpha}{T-3}$, so in a sense we are trying to preserve the overall significance of our testing procedure. Below, we again show our results for the normally distributed data when $T = 100$.

A commonly directed criticism at the Bonferroni-correction is that in the pursuit of eliminating false positives, it also drags down the power in the process, and this is of course also very obvious from our results(unless the change itself is big enough). But we can also see that in difficult situations where false positives are more likely to occur, i.e. when the change is small or it occurs close to the edges, the change-point approximations may improve. For example, when $\delta = 0.1$ and $\tau = 25$, M provides a better τ-estimate.

Parameters			Mean change				Variance change			
T	δ	τ	Power(M)	$\hat{\tau}_M$	Power(N)	$\hat{\tau}_N$	Power(λ_{var})	$\hat{\tau}_{\lambda_{var}}$	Power(SIC$_{var}$)	$\hat{\tau}_{SIC_{var}}$
100	0.1	5	0	/	0	/	0.68	5.16(3.2)	0.22	4.8(1.38)
		25	0.002	22.5(6.36)	0.002	26(1.41)	0.995	24.34(2.7)	0.91	24.27(2.2)
		50	0.004	56.5(42.3)	0	/	1	49.17(2.56)	0.94	49.24(2.33)
	1.1	5	0.026	8.31(12.65)	0.014	11.57(8.92)	0.003	75(29.46)	0	/
		25	0.72	25.46(5.11)	0.29	26.46(6.13)	0.001	10(/)	0	/
		50	0.86	49.85(5.012)	0.64	49.88(7.5)	0.001	98(/)	0	/
	2	5	0.595	5.44(3.69)	0.27	7.02(6.14)	0.003	61.67(50.5)	0	/
		25	1	25.06(1.33)	0.97	27.01(4.22)	0.014	30.93(19.99)	0	/
		50	1	49.99(1.41)	0.999	50.1(3.59)	0.054	52.07(12.7)	0.0001	61(/)

Table 4: Power and change-point approximation with Bonferonni-correction.

2.3.5 Gamma-distribution

Finally, we consider the gamma-statistics. Using the same simulation setup for the normal and student-t distribution, we have displayed both the power and change-point approximations in table 5.

What is readily apparent from the results, is that the change-point location τ affects the results much more negatively than it did in previous simulations. In general, unless τ is around the mid-point of the data and the changes are big enough, the statistics perform very poorly. This is especially the case for λ_{sh}, though this is unsurprising as before we have shown that shape changes entail both the mean and variance, whereas scale changes mainly correspond with variance(at least for the change sizes we considered here). Therefore, the locations of shape changes in general will be harder to discern due to a relatively higher amount of noise, and this is clearly demonstrated here.

It should be noted though that while λ_{sh} always provides a poor chang-point approximation, its power is less influenced by τ when it moves away from the mid-point, at least for $\delta_2 = 2$. So for big shape changes it may be harder to discern their precise locations, but due to their more extreme overall change the existence of a change-point may be more easily evidenced, even for less ideal positions of τ.

Parameters				Scale change		Shape change	
T	δ_1	δ_2	τ	Power(λ_{sc})	$\hat{\tau}_{\lambda_{sc}}$	Power(λ_{sh})	$\hat{\tau}_{\lambda_{sh}}$
30	4	2	5	0.0081	13.37(3.26)	0.0335	25.34(5.85)
			10	0.0915	12.55(1.97)	0.28	19.13(11.75)
			15	0.8161	13.17(1.073)	0.5273	12.45(12.57)
	2	0.5	5	0.005	13.74(3.16)	0.0071	18.35(12.008)
			10	0.069	13.11(1.91)	0.0368	15.61(12.39)
			15	0.356	13.49(1.35)	0.0463	13.17(12.33)
	0.5	0.1	5	0.005	15.02(1.73)	0.0094	14.34(12.24)
			10	0.0105	13.96(2.496)	0.0079	10.77(12.04)
.			15	0.0278	14.31(1.61)	0.017	11.54(11.89)
100	4	2	5	0.0022	48.77(3.52)	0	/
			25	0.1945	44.46(5.27)	0.6138	76.68(38.38)
			50	1	47.94(1.605)	0.9728	41.73(46.68)
	2	0.5	5	0.0007	51.86(6.34)	0.0002	49(67.88)
			25	0.0607	46.14(5.02)	0.0171	65.08(43.56)
			50	0.9636	48.31(2.36)	0.0487	40.78(45.53)
	0.5	0.1	5	0.001	48.1(4.48)	0.0001	4(/)
			25	0.0043	47.86(3.83)	0.0008	72.13(43.62)
.			50	0.0785	49.49(3.65)	0.0033	45.42(46.38)

Table 5: Power and change-point approximations for the parametric gamma-statistics, where the data changes from $\Gamma(1,1)$ to either $\Gamma(1+\delta_1,1)$ [scale change] or $\Gamma(1,1+\delta_2)$ [shape change].

3 Non-parametric model

Now we will consider a different class of change-point problems, namely the non-parametric model. Unlike the parametric case, we may not make any assumptions about the distributional structure of the data. In a practical sense, working under these new conditions is obviously more realistic than making presumptive guesses about the underlying distribution. Furthermore, if the assumptions about alleged distributions happen to be completely misfounded, then it will tend to have an adverse effect in the form of additional false positives(as we clearly saw for example when the parametric variance statistics were evaluated under the student-t distribution).

This relaxation of the assumptions may make it seem like that the test-statistics we derive under the new setting may perform worse than their parametric counterparts, since we have less information to work with. However, as we will later show in a simulation study, the non-parametric test statistics will generally perform better(in a non-Gaussian setting that is).

As before, we will consider both a single change in mean and variance. However, since the distributions are now unknown we can also make a more general inference: the change-points of distributions themselves. As for the reasoning of only considering a single change-point, we again refer to our treatment of search methods in section 4, where it is shown how we can easily consider multiple change-points from the framework that we set up here.

In order to set up a statistical framework for change-point detection, we use the following hypothesis testing procedure:

$$
\begin{aligned}
H_0 &: y_t \sim F_0 \ \forall t \\
H_1 &: y_t \sim \begin{cases} F_0, & \text{if } t < \tau \\ F_1, & \text{otherwise} \end{cases}
\end{aligned}
$$

Note that now we can also consider general distributions, and not just distribution-specific parameters like we did previously. Also, the way that we define test-statistics will be more straight-forward. Instead of defining cost-functions and using the discrepancy function (3) to define test-statistics, we now immeadiately treat some well-established two-sample tests which themselves can serve as test-statistics. These will be denoted as $D_{\tau,T}$.

From here on the idea remains the same as for the likelihood-approach. In order to estimate a change-point location we will calculate the absolute value of $D_{\tau,T}$ for all possible τ-locations and take the maximum value. Formally:

$$
D_T = \mid \max_{\tau} \frac{D_{\tau,T} - \mu_{D_{\tau,T}}}{\sigma_{D_{\tau,T}}} \mid, \qquad 1 < \tau < T \tag{22}
$$

As shown, we have made another slight modification to our standard procedure. For large samples we may invoke the central limit theorem to conclude that $D_{\tau,T}$ is approximately normally distributed, and thus its standardized version standard normally distributed. This is preferable of course, as critical values and the like are readily available in that case. Also, standardizing the test-statistic will make it less skewed by the change-point locations, as the variances of these statistics always seem to depend on τ(the statistics will primarily use only the part of the data prior to τ). In other words, if a certain change-point location causes a high variance, then obviously the absolute value of the statistic will be biased towards that change-point location.

Finally, through simulations under H_0 we may determine the upper(lower) α quantile c_α of the empirical null-distribution, and use this as a threshold for D_T so that H_0 is rejected whenever $D_T > c_\alpha (D_T < c_\alpha)$. We then simply extract a change-point location $\hat{\tau}$ for which the maximization(minimization) holds and which also exceeds the defined threshold.

3.1 Ranks

Instead of densities and the corresponding likelihood, we need to find an alternative way to garner information about the data. Thankfully, at least when it comes to detecting any changes in the mean or variance, the *ranks* of the data-points seem to fit our purposes quite nicely.

Simply speaking, the ranks denote the relative positions of the data-points when it comes to size. So the lowest rank denotes the smallest value, and vice-versa. Formally, we may define the rank of a data-point y_t as follows:

$$r(y_t) = \sum_{t \neq s}^{T} I(y_t \geq y_s) \tag{23}$$

Here, $I(\cdot)$ is the indicator-function. For now, we will use ranks to give appropiate cost-functions and their correponding test-statistics $D_{\tau,T}$ as defined previously. Their performance will be evaluated later in a simulation study, where we shall also compare them with their parametric counterparts in both a Gaussian as a non-Gaussian setting.

3.1.1 Mean change

In order to detect a change in mean, we will make use of the so-called *Mann-Whitney* two-sample test[9]. Obviously the two considered samples are the data-sets prior to and after the change-point, which we will respectively denote with V and W(with sizes T_1 and T_2).

Furthermore, under the assumptions of no changes and tied ranks, it can

easily be shown that the sum of the first T ranks is:

$$\frac{T(T+1)}{2}$$

After all, it is nothing more than the sum of the first T positive integers. With this sum, we can also easily obtain the expected rank under H_0 by simply taking the empirical mean, i.e. dividing by T. We can then measure the discrepancy from H_0 by summarizing over the differences between the actual ranks in V and the expected rank under H_0:

$$U_{\tau,T} = \sum_{t=1}^{T_1} (r(y_t) - \frac{T+1}{2}) \tag{24}$$

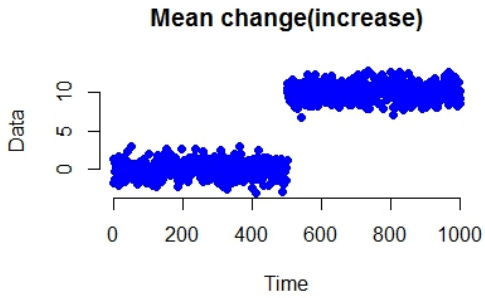
In order to standardize $U_{\tau,T}$, we use the following formulas for the mean and variance(see [9]):

$$m_{U_{\tau,T}} = \frac{T_1 T_2}{2} \tag{25a}$$

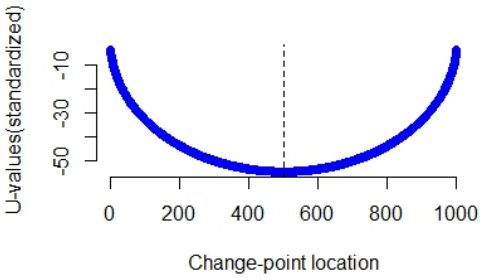
$$\sigma_{U_{\tau,T}} = \sqrt{\frac{T_1 T_2 (T_1 + T_2 + 1)}{12}} \tag{25b}$$

Finally, we repeat this procedure for every possible change-point location and H_0 is rejected whenever the maximum absolute value exceeds the defined threshold c_α.

Below we show a few figures of the Mann-Whitney statistic in action, where alongside the actual data we have plotted all standardized values of $U_{\tau,T}$ against the possible change-point locations.



An increase in mean



Corresponding standardized U-values

Figure 8: The effectiveness of Mann-Whitney displayed for an increase in mean. The vertical dashed line marks the most likely candidate for a change-point

3.1.2 Variance change

We now consider a change in variance, by considering multiple distinct test-statistics. As usual, we assume that the data is independent within each sample, and that the samples themselves are also independent of each other. We start with the Mood-statistic[10], which is essentially a direct extension of the Mann-Whitney statistic.

3.1.2.1 Mood-statistic

As with the Mann-Whitney statistic, we again look directly at the discrepancy between the ranks and the expected rank. But now we must summarize over the squared differences in V:

$$MD_{\tau,T} = \sum_{t=1}^{T_1}(r(y_t) - \frac{T+1}{2})^2 \tag{26}$$

In order to standardize it, we use the following(see [10]):

$$m_{MD_{\tau,T}} = \frac{T_1(T^2-1)}{12} \tag{27a}$$

$$\sigma_{MD_{\tau,T}} = \sqrt{\frac{T_1T_2(T+1)(T^2-4)}{180}} \tag{27b}$$

The basic idea is that the sum is large whenever the corresponding ranks of the sample are large or small, which of course also coincides with a higher variance(or a smaller variance when the sum is small). However, even though the method is intuitively simple, it requires us to assume that the means of the samples are equal. Any deviation from this assumption may lead to worse performance(see table 7). The next statistic that we consider is a modification of the Mood-statistic, meant to mitigate this problem.

3.1.2.2 Mood's test(differing means)

Instead of measuring the discrepancy relative to the overall mean of the ranks, we instead only consider the corresponding mean of the first sample V:

$$\overline{r} = \sum_{t=1}^{T_1} \frac{r(y_t)}{T_1} \tag{28}$$

We also divide it by $T_1 - 1$ to make it a bit more robust, finally leading us to:

$$MT_{\tau,T} = \frac{1}{T_1 - 1} \sum_{t=1}^{T_1} (r(y_t) - \overline{r})^2 \tag{29a}$$

$$m_{MT_{\tau,T}} = \frac{T(T+1)}{12} \tag{29b}$$

$$\sigma_{MT_{\tau,T}} = \sqrt{\frac{TT_2(T+1)(3(T+1)(T_1+1) - TT_1)}{360T_1(T_1-1)}} \tag{29c}$$

So we lessen its dependence on the second sample, and thus its dependence on the assumptions about the means, by only considering the expected rank of the first sample. However, in those cases where the means are still (roughly) identical we would be needlessly limiting the amount of information, so we would expect the normal Mood test to still perform better then. We now look at a more drastic change of the test setup, termed the Ansari-Bradley method, which will hopefully not suffer from the same drawbacks.

3.1.2.3 Ansari-Bradley

We now treat a novel method proposed in [11], where the use of ranks still remains integral, but the test setup is different. Instead of mainly working with one set of ranks, we now center the ranks around a mid-point. Therefore, we have to make a distinction between two cases depending on whether the data-length $T_1+T_2 = T$ is even or odd(we shall denote the ranks as $\{r_t^*\}_{t=1}^T$):

$$\{r_t^*\}_{t=1}^T = \begin{cases} 1, 2, ..., \frac{T}{2}, \frac{T}{2}, ..., 2, 1 & \text{if T is even} \\ 1, 2, ..., \frac{T-1}{2}, \frac{T+1}{2}, \frac{T-1}{2}, ..., 2, 1 & \text{if T is odd} \end{cases}$$

For example, the smallest and largest values are both given a rank of 1. Then the second-smallest and second-largest values are given a rank of 2, and the same process is repeated until we assign a rank to the data-points in the middle between the two extremes. This may seem like a peculiar setup, but for detecting a change in variance between two samples it works well: the data-points with the higher variance will more likely be assigned the smaller ranks. Thus if we were to sum up the ranks from $\{r_t^*\}_{t=1}^T$ that belong to a certain sample, let us say V, then if it is significantly small we may be inclined to conclude that V has a higher variance. Formally we write it down as follows:

$$AB_{\tau,T} = \sum_{t=1}^{T_1} R_t^*$$

$$m_{AB_{\tau,T}} = \begin{cases} \frac{T_1(T+2)}{4} & \text{if T is even} \\ \frac{T_1(T+1)}{4T} & \text{if T is odd} \end{cases}$$

$$\sigma_{AB_{\tau,T}} = \begin{cases} \sqrt{\frac{T_1T_2(T+2)(T-2)}{48(T-1)}} & \text{if T is even} \\ \sqrt{\frac{T_1T_2(T+1)[3+T^2]}{48T^2}} & \text{if T is odd} \end{cases}$$

As mentioned, the R_t^* denote the ranks of V from $\{r_t^*\}_{t=1}^T$. It should also be more obvious now why we had to assume that the difference in means could not be too big, for if the variance-change is relatively small compared to the mean-change, then the highly variant sample may not attain the extreme values anymore, thus compromising the effectiveness of our test-statistic. In spite of that, we still expect this statistic to be more robust than the Mood-statistic

against different means, since $AB_{\tau,T}$ does not have a quadratic nature.

3.2 General distributional changes

Since we do not make any assumptions about the distributions anymore, we can now also consider the very general problem of detecting a change in the distribution itself. However, this also means that if we do detect a change, it usually becomes difficult to trace back the cause of the change. Indeed, in the most general setting we can of course not assume anything about the type of change. The causes may be included to: a mean and/or variance change, a change in kurtosis of the underlying distributions, and in general any well-defined characteristic of a distribution that we can think of.

Thus the use of ranks becomes impractible, and we have to likewise consider a more general solution. Practically any method that deals with this problem, employs so-called *empirical* cumulative distribution functions(ECDF):

$$\hat{F}(s) = \frac{1}{T}\sum_{t=1}^{T} I(y_t \leq s) \tag{30}$$

Using these approximations for the CDFs, we can measure the discrepancy by considering the maximum distance between different ECDFs. As usual, we assume that the data is continous, independent and contains 1 change-point. We begin with the two-sample variant of a very popular statistic, namely *Kolmogorov-Smirnov*([12], [13]).

3.2.1 Kolmogorov-Smirnov

We define a separate ECDF for each of the two samples, denoted as $\hat{F}_V(s)$ and $\hat{F}_W(s)$. We then simply try to find the maximum absolute difference between the two:

$$KS_{\tau,T} = \sup_{\mathrm{s}} \mid \hat{F}_V(s) - \hat{F}_W(s) \mid \tag{31}$$

There is however a problem concerning its standardization, as to the best of our knowledge there still is not an analytical expression for its mean and variance. So since we can not adopt the KS-statistic into our existing framework, we have to use an alternative method of testing its significance. Namely

we will use *p-values*, denoted as $p_{\tau,T}$, which are defined as the probabilities of observing a more extreme value than $KS_{\tau,T}$ under H_0. Clearly, the smaller this value is, the less likely it is that we can attribute an unusual value of $KS_{\tau,T}$ to chance(thus providing stronger evidence against H_0).

The question then remains of how to estimate $p_{\tau,T}$. We will adopt a method outlined in [16], which will be briefly summarized here:

$$p_{\tau,T} = Q(KS_{\tau,T}\sqrt{\frac{T_1T_2}{T_1+T_2}} + \beta) \tag{32}$$

, where

$$\begin{aligned} Q(z) &= 2\sum_{i=1}^{\infty}(-1)^{i-1}\exp(-2i^2z^2) \\ &\approx 2(\exp(-2z^2) - \exp(-8z^2)) \text{ and} \end{aligned}$$

$$\beta = \begin{cases} \frac{1}{2\sqrt{T_1}}, & \text{if } T_1 > 2T_2 \\ \frac{2}{3\sqrt{T_1}}, & \text{if } T_2 \leq T_1 \leq 2T_2 \text{ and } T_1 \text{ a multiple of } T_2 \\ \frac{2}{5\sqrt{T_1}}, & \text{otherwise} \end{cases}$$

Finally, in order to bring it more in line with the existing framework, we define $q_{\tau,T} = 1 - p_{\tau,T}$. Maximizing $q_{\tau,T}$ is equivalent to minimizing $p_{\tau,T}$, so for the change-point estimation we solve:

$$q_T = \max_{1\leq\tau\leq T-1} q_{\tau,T} \tag{33}$$

So a change-point is detected at $\hat{\tau}$ when it maximizes $q_{\tau,T}$ and it exceeds a threshold c_α.

3.2.2 Cramer-Von-Mises

An alternative test-statistic that we could use is the Cramer-Von-Mises test([14], [15]), which similarly utilizes a distance-based statistic from the ECDFs:

$$CVM_{\tau,T} = \int_{-\infty}^{\infty} \mid \hat{F}_V(s) - \hat{F}_W(s) \mid^2 dF(s) \tag{34}$$

However, in our case the independent variable s denotes the time, which is discrete and finite. So instead of evaluating a complicated integral, we can simply summarize over all y_t:

$$CVM_{\tau,T} = \sum_{t=1}^{T} \mid \hat{F}_V(y_t) - \hat{F}_W(y_t) \mid^2 \tag{35}$$

We use the following corrected form of the statistic, which should help standardize it against variant data:

$$CVM_{\text{corr}} = \frac{T_1 \cdot T_2}{T_1 + T_2} \cdot CVM_{\tau,T} \tag{36}$$

Here, T_1 and T_2 are the lengths of the subsegments resulting from the (potential) change-point. The resulting manner in which we detect a change-point, is identical as before.

3.3 Simulation study(Non-parametric)

In this simulation study, we do not only wish to assess the performance of the relevant statistics, but also compare them with the parametric model. Namely, we are interested if the rank-based statistics are indeed more robust against outliers, so we will be conducting the same simulations as we did for the student-t distribution. Afterwards, we will also be investigating whether David's test performes better when the mean and variance change simultaneously, which is what we hypothesized earlier. Finally we consider the general distributional change-statistics, which we will assess under the student-t and gamma distributions.

The simulation setup remains identical, except now we also outright apply the Bonferroni-correction discussed during the previous study.

3.3.1 Rank-based statistics

On the next page we have displayed the results for the rank-based statistics, which we have denoted as follows:

- MW: Mann-Whitney
- Mood: Mood's test
- MT: Mood's test(differing means)
- AB: Ansari-Bradley

Due to the fact that we mainly use ranks, which are generally more robust againts outliers than likelihood-based statistics, we see that the results have generally improved. Even accounting for the robustness-modifications that we discussed for the parametric model(table 3), Mann-Whitney still performs

better given the right circumstances. As for the variance-change statistics, their change-point approximations still leave some things to be desired, but aside from the marginal power decrease compared to SIC_{var} their performance has also increased somewhat. We can furthermore conclude that the Mood test and Ansari-Bradley provide very similar performance, with Mood's differing means variant performing significantly worse.

We also wish to compare the latter against the other methods when a change in mean has occured. To that end, we showcase the results when an additional mean-change of 2 has occured in table 7. We can conclude that while MT does provide higher power in some cases, in general it still performs worse than the other statistics.

Parameters				Mean change		Variance change					
T	δ_1	δ_2	τ	Power(MW)	$\hat{\tau}_{MW}$	Power(Mood)	$\hat{\tau}_{Mood}$	Power(MT)	$\hat{\tau}_{MT}$	Power(AB)	$\hat{\tau}_{AB}$
30	0.1	-2	5	0.051	12.6(5.2)	0.98	10.1(5.5)	0.06	8.2(6.7)	0.98	11.1(5.4)
			10	0.058	14.1(4.4)	0.97	13.7(5.9)	0.056	9.3(7.4)	0.95	14.7(5.9)
			15	0.041	13.3(4.3)	0.92	16.9(6.2)	0.059	8.5(7.8)	0.91	17.8(6.2)
	1.1	20	5	0.13	12.4(3.7)	0.44	10.4(8.6)	0.086	5.2(5.5)	0.43	12.2(9.1)
			10	0.47	11.2(2.8)	0.47	10.7(8.5)	0.084	5.9(5.5)	0.46	12.7(8.9)
			15	0.55	13.8(2.1)	0.5	9.7(7)	0.036	4.6(4.9)	0.5	11.3(7.8)
	2	100	5	0.38	10.3(3.7)	0.44	10.6(8.7)	0.066	4.8(5.3)	0.43	12.7(9.2)
			10	0.81	10.2(2)	0.52	10.1(7.8)	0.048	5.7(6.5)	0.48	11.2(7.9)
			15	0.92	14.02(1.3)	0.5	10.03(7.2)	0.041	3.7(3.4)	0.49	11.6(7.9)
100	0.1	-2	5	0.053	48.6(9.5)	1	11.6(7.7)	0.06	22.5(16.5)	1	12.6(7.7)
			25	0.053	47.4(11.2)	1	25.6(11.8)	0.069	28.2(15.02)	1	26.6(11.8)
			50	0.073	49.8(9.6)	1	38.4(20.9)	0.08	20.6(12.7)	1	39.4(20.9)
	1.1	20	5	0.084	45.98(8.96)	0.59	39.6(35.7)	0.07	16.8(12.5)	0.59	43.3(35.5)
			25	0.77	34.7(9.7)	0.73	29.2(28.8)	0.086	19.7(12.4)	0.71	30.9(29.9)
			50	0.98	49.3(3.97)	0.83	28.7(22.7)	0.058	21.8(13.4)	0.82	29.8(23.4)
	2	100	5	0.097	46.1(8.4)	0.59	37.94(34.9)	0.056	20.5(14.5)	0.6	42.9(35.3)
			25	0.995	29.2(7)	0.74	25.6(26.3)	0.06	18.7(11.7)	0.73	29.6(28.95)
			50	1	48.96(1.8)	0.79	28.9(23.2)	0.055	19.1(11.8)	0.79	30.7(24.02)

Table 6: A showcase of the power and change-point estimation for the non-parametric statistics when the data follows a student-t distribution. The data initially has a mean of zero and 3 df(degrees of freedom), and after the change has a mean of δ_1 or $3 + \delta_2$ df.

Parameters			Variance change					
T	δ_2	τ	Power(Mood)	$\hat{\tau}_{Mood}$	Power(MT)	$\hat{\tau}_{MT}$	Power(AB)	$\hat{\tau}_{AB}$
100	-2	5	1	19.38(14.82)	0.16	32.2(13.68)	1	20.38(14.82)
		25	1	37.81(14.84)	0.92	41.6(20.41)	1	38.75(14.84)
		50	0.99	58.7(15.97)	0.89	28.88(21.64)	0.99	59.43(16.21)
	20	5	0.95	9.88(18.02)	0.11	33.84(12.12)	0.93	14.83(22.67)
		25	1	24.14(10.05)	0.89	43.92(19.92)	0.99	26.42(11.41)
		50	0.56	44.35(30.94)	0.86	32.21(21.45)	0.6	45.81(31.42)
	100	5	0.96	9.31(16.49)	0.14	34.36(14.16)	0.94	13.88(20.95)
		25	1	23.81(9.17)	0.89	43.44(19.96)	1	26.1(11)
		50	0.62	42.06(31.89)	0.89	31.94(21.67)	0.64	45.9(32.46)

Table 7: The variance statistics evaluated when there is also a mean-change of size 2.

3.3.2 General distributional changes

For the final simulation study, we consider the general distribution change-statistics Kolmogorov-Smirnoff and Cramer-von-Mises. We wish to compare their performance to the parametric and rank-based statistics, and to that end we have displayed the results below for both the student-t and gamma distributions.

(**a**)

Parameters				Mean change				Variance change			
T	δ_1	δ_2	τ	Power(KS)	$\hat{\tau}_{KS}$	Power(CvM)	$\hat{\tau}_{CvM}$	Power(KS)	$\hat{\tau}_{KS}$	Power(CvM)	$\hat{\tau}_{CvM}$
			5	0	/	0	/	0.002	15.5(13.44)	0	/
	0.1	-2	10	0.003	26(2)	0.008	17.13(6.35)	0.009	20.78(7.87)	0.004	18.5(2.08)
			15	0.001	28(/)	0.001	15(/)	0.004	27.5(0.58)	0.002	19(2.83)
			5	0	/	0	/	0	/	0.001	19(/)
30	1.1	20	10	0.028	13.18(3.95)	0.049	11.71(3.2)	0	/	0.006	12.67(9.03)
			15	0.078	16.68(3.31)	0.14	14.85(2.25)	0.001	4(/)	0.005	14.6(7.09)
			5	0.34	5.87(2.34)	0.013	8(2.31)	0.006	6.17(3.66)	0.001	11(/)
	2	100	10	0.51	10.3(2.13)	0.34	10.35(1.6)	0.005	10.2(9.04)	0	/
.			15	0.73	15.31(1.93)	0.53	14.94(1.71)	0.002	5.5(2.12)	0.002	13(5.66)

(**b**)

Parameters				Scale change				Shape change			
T	δ_1	δ_2	τ	Power(KS)	$\hat{\tau}_{KS}$	Power(CvM)	$\hat{\tau}_{CvM}$	Power(KS)	$\hat{\tau}_{KS}$	Power(CvM)	$\hat{\tau}_{CvM}$
			5	0.054	7.78(3.97)	0.022	10.68(3.76)	0.086	5.29(1.43)	0.038	8.39(3.32)
	4	2	10	0.12	11.05(2.53)	0.026	11.15(1.78)	0.17	10.41(1.58)	0.083	10.52(1.07)
			15	0.5	15.81(2.44)	0.43	15.43(1.95)	0.53	15.01(2.17)	0.61	14.98(1.76)
			5	0.021	9.33(5.17)	0	/	0.002	13.5(7.78)	0	/
30	2	0.5	10	0.051	12.2(4.12)	0.014	12(3.21)	0.005	14(3.32)	0.003	12.67(4.62)
			15	0.17	16.27(3.26)	0.075	15.67(2.35)	0.022	15.77(2.76)	0.009	13.78(2.54)
			5	0.001	20(/)	0	/	0.001	18(/)	0.001	12(/)
	0.5	0.1	10	0.014	14.21(5.16)	0.002	19(11.31)	0.007	16.71(4.23)	0.001	24(/)
.			15	0.022	18(4.25)	0.007	17(3.27)	0.004	19.75(1.5)	0.004	19(7.07)

Table 8: Power and change-point approximation for Kolmogorov-Smirnoff(KS) and Cramer von Mises(CvM). Table (a) displays the results for the student-t distribution, while (b) shows them for the gamma-distribution.

We only consider $T = 30$, since computing the ECDFs requires subtantive computational work. Beginning with the student-t distribution, we see that the only reasonable performance is obtained for the biggest mean change, whereas for every other situation the power tends to be very low. Surprisingly though, we may get some relatively accurate change-point approximations even when

the power is very low and the changes are not drastic. For example, when $\delta_2 = 20$ we see that CvM gives some pretty accurate approximations for $\tau = 10$ and $\tau = 15$. This is of course the exception rather than the norm, as the same method again performs poorly for the more drastic case of $\delta_2 = -2$. Aside from these exceptions, we can still discern a general pattern when the performance is not too low: the power of KS tends to be higher than CvM, while the latter has lower standard deviations. Furthermore, comparing these to our previous statistics, we can clearly see from table 6 that the rank-based statistics are vastly superior. However, table 2 still shows that KS and CvM provide better performance for a mean-change than the parametric statistics, even though the Bonferroni-correction was not applied there. This shows us that KS and CvM are still relatively more robust against outliers than the parametric statistics.

The comparison with the gamma-statistics in table 5 is more interesting. While the power is still lower(especially if we consider the scale change), we do see more accurate change-point approximations here. This is quite relieving, since we clearly saw that the gamma-statistics were not well-suited for discerning a precise change-point location(particularly the shape-change). So, whichever statistic should one use? It depends on the original goal: if one simply seeks to confirm the existence of a change-point, then the parametric statistics are preferable, considering their superior power and computation time. If the precise location also needs to be determined, then we would advocate the use of the general distributional statistics.

4 Search methods

As we mentioned during the beginning of this report, change point detection procedures generally consist of three parts: cost functions, search methods and penalty terms. We have dedicated the last two sections to treating various cost functions, all valid under their own set of assumptions. However, we were restricted to only consider the detection of a single change point, seeing as we did not have dedicated search methods that could quickly skim over the data and find the optimal segmentation to either solve problem (1) or (2). Also, depending on which specific optimization problem we solve, we also need to have a penalty term. The problem of finding an optimal penalty value is quite difficult and we instead refer the interested reader to [20], where it is discussed how an optimal penalty may be efficiently found from a range of values. With a properly chosen penalty value, it is possible to strike a good balance between goodness-of-fit and complexity.

For now, we will treat several search methods, each designed to find the optimal segmentation in an unique way. We start with approximate algorithms, as these are generally easier to implement than exact ones. *R*-code implementing all treated search methods may be found below[2].

4.1 Approximate algorithms

While approximate algorithms may not offer the same accuracy as exact algorithms, they are still widely used in practice for their speed and ease of implementation. Therefore, we will treat two widely known search methods belonging to this category:

- Window sliding
- Binary segmentation

4.1.1 Window sliding

As the name implies, this search method offers a very intuitive way of looking for multiple change-points. First, two windows are formed around the change-point candidate, their widths specified by the user beforehand. Then we simply *slide* these windows across the data, while using the discrepancy function (3) to continuously compute the discrepancy between the segmented data at change-point candidate t and the unsegmented data:

$$d(y_{v..t}, y_{t..w}) = c(y_{v..w}) - [c(y_{v..t}) + c(y_{t..w})]$$

So the left side of the first window is at time-point v, and the right side

[2] https://github.com/StudBch96/Bachelor-project-R-code

of the second window at time-point w. Afterwards, we will have obtained a discrepancy curve, in which we will search for local maximum values as these are the strongest indications of a structural change incurring. This procedure is summarized in the figures below:

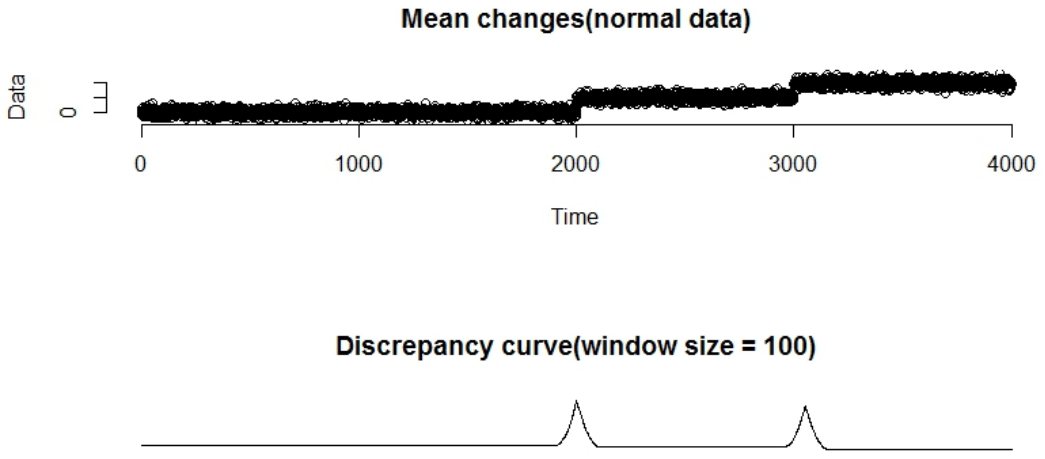


Figure 9: A plot of normally distributed data undergoing mean changes, alongside a discrepancy curve generated by the window-sliding algorithm.

While intuitively simple, it is also computationally fast as the total amount of calculations it needs to perform scales linearly with the length of the data(namely, calculating the discrepancy curve). However, it also has a few serious disadvantages. First off, we can not use it to detect any changes that are less than a window-size located to the edges of the data, as we only consider the midpoint of the windows as potential change-points. Therefore, especially when we are working with big datasets and windows, we will be more prone to missing out on actual change-points occuring closely to the edges.

Also, this algorithm will have trouble locating change-points close to each other(as in, both are in the span of one window). After all, the cost function was derived under the assumption of 1 change-point, so it will work poorly when there are multiple change-points present.

4.1.2 Binary Segmentation

As the second algorithm, we consider perhaps one of the most widely used search methods in practice, namely the binary segmentation[17]. This algorithm allows us to naturally expand on the single change-point problem by sequentially trying to test for a single change for a given dataset, and repeating the very same test for the two newly formed subsegments when a change has been identified. This process is repeated until a certain stopping criterion is fulfilled. A schematic overview of this procedure is displayed below:

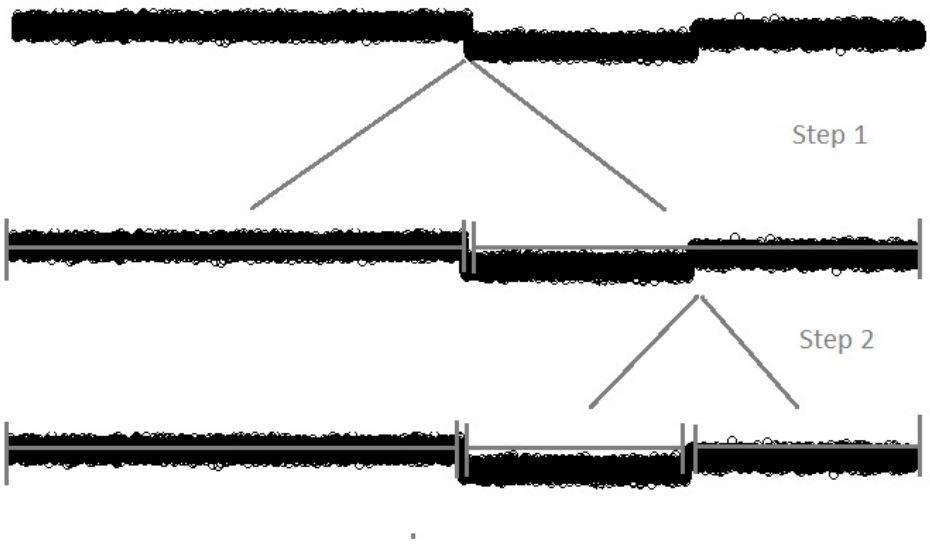


Figure 10: A schematic overview of Binary Segmentation.

Just like the window-sliding algorithm, the binary segmentation is also computationally fast(with a complexity of $\mathcal{O}(T \log T)$, see [13]). Unlike the previous algorithm though, it is sequential and above all else *greedy*: it is only interested in the change-points that give the highest possible likelihood.

Although it might be preferable to have an intuitive algorithm that is not as graphically dependent as the window-sliding algorithm, we may still run into the same issues that we had before, namely a poor performance whenever the minimum-distance between change-points is too small. To counter-act this we discuss a relatively recent development, termed the *wild binary segmentation*[18], which naturally expands the standard method to specifically deal with this problem.

4.1.2.1 Wild Binary Segmentation

While still functionally identical, there is a key difference when selecting the relevant subsegments for inferring a change. In contrast to the standard method, where we chose fixed subsegments controlled entirely by the change-points, we now generate random intervals whose starting/end points are uniformly distributed(within the range of the adjacent estimated change-points).

Why does this particular setup give superior performance when the change-points are closely clustered? Basically, by adding a random element such as this, we want to forcibly create a scenario where the cost function can perform better(i.e. at most 1 change-point). In other words, what we are hoping for is that in at least one of the simulated intervals around the potential change-point location, there are no other (potential) change-points present so that the cost function may give a more pronounced value(unaffected by any neighbouring change-points). As we will show shortly, the wild binary segmentation provides similar performance in the case of closely located change-points, though it will come with an increased computational cost due to all the additional simulations it has to perform.

4.2 Exact algorithms

When designing and implementing exact search methods, there are additional problems that can arise. Namely, in order to gain an exact solution it is basically required to search through the entire solution space. In the specific case of change-point estimation, the corresponding computational cost can be enormous:

• For the unconstrained optimization problem (1), there are 2^{T-1} possible solutions(every location between 1 and T - 1 can be a change-point)

• For the constrained version (2) the amount decreases to $\binom{T-1}{K-1}$

In both cases, even for moderate sizes of the data-length T and amount of change-points K, the total cost can increase very quickly. For example, for $T = 500$ and $K = 10$, the amount of possible segmentations is $\binom{499}{9} \approx 4.916 \cdot 10^{18}$. Therefore, it is simply imfeasible to iterize over all possible solutions, and it is for this reason that traditional programming techniques are of no interest. Thankfully, there is a separate branch within programming literature that deals with problems like these: *dynamic programming*. In essence, it is possible to turn difficult problems into a set of linked sub-problems, and by solving each of these sub-problems we may eventually combine the solutions to solve the original problem. The trick here is that for every consequent sub-problem, any previously obtained information should be used to efficiently obtain the new solutions. We discuss a commonly used exact algorithm, the Segment-Neighbourhood[19],

which will clearly demonstrate the principles of dynamic programming.

4.2.1 Segment Neighbourhood

A good way of efficiently searching through a solution space is with the use of recursions, and change-point problems are no exception. If we denote the optimal segmentation(let us assume for now that optimal is equivalent to maximum likelihood) with K change-points for $y_{1..t}$ as $\{\tau_0, \tau_1, ..., \tau_{K+1}\}$ (where $\tau_0 = 1$ and $\tau_{K+1} = t$), and denote its corresponding cost as $c_{K,t}$ then it is possible to derive a recursion for the cost:

$$\begin{aligned} c_{K,t} &= \max_{\tau} \sum_{i=0}^{K} c(y_{\tau_i+1:\tau_{i+1}}) \\ &= \max_{\tau_K} \left\langle \max_{\tau_{1:(K-1)}} \sum_{i=0}^{K-1} c(y_{(\tau_i+1):\tau_{i+1}}) + c(y_{(\tau_K+1):\tau_{K+1}}) \right\rangle \\ &= \max_{\tau_K} \left\langle c_{K-1,\tau_K} + c(y_{(\tau_K+1):\tau_{K+1}}) \right\rangle \\ &= \max_{\tau_K \in K,\ldots,t-1} \left\langle c_{K-1,\tau_K} + c(y_{(\tau_K+1):t}) \right\rangle \end{aligned}$$

So in every iteration, we could consider 1 additional change-point and use the above recursion to quickly find the optimal solution. Note that within each iteration, this process needs to be repeated for t up until T, and for each t the minimization/maximization also needs to performed(in order to locate the change-point). Furthermore, we need to perform this whole process for every possible amount of change-points, which for simplicity is constrained to K. Taking all these computations together, means that this algorithm will have a complexity of $\mathcal{O}(KT^2)$.

We see that as a trade-off for obtaining an exact algorithm, the computational cost has been increased quite a bit. Still, it is quite impressive that the original seemingly impossible problem of considering every possible change-point configuration, turns out to be pretty feasible due to dynamic progamming.

4.3 A quick performance review

With the cost-functions and search methods now appropiately defined, we may finally assess their performance for multiple-changepoints problems. To this end, we consider two aspects of the algorithms:

- Speed
- Accuracy

With speed, we simply refer to the so-called *run-time* of the algorithms. As for accuracy, we will primarily look at the discrepancy values that the algorithms return for every changepoint-location. These values, as we explained in sections 2 and 3, are a quantitative measure for the significance of the change-point. So if the algorithm provides a higher discrepancy value for a certain changepoint-location, then it is more confident in that result.

We consider a relatively simple example, displayed in the figure below. We now have multiple mean changes, and the first changes occuring are very big in size and do not occur too closely to each other, so they should be easy to detect. The last few change-points(occuring after time 300) have in contrast smaller change sizes and are also more densely located, so these should be harder to detect. For this particular data set, we have tabulated the performance of the search methods in table 9(the cost-function used was the parametric mean change from section 2).

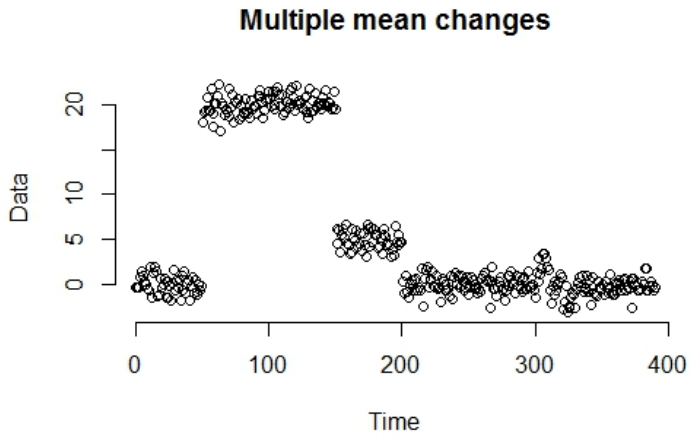


Figure 11: A plot of data with multiple changes in its mean.

Algorithm	τ-location and discrepancy values							Runtime(sec.)
Window-sliding	50	150	200	300	310	320	330	0.026
	293.55	139.04	25.05	4.004	9.09	0.37	2.4	
Bin.Segmentation	50	150	200	300	310	320	330	0.058
	6627.22	6781.5	513.62	20.62	8.23	7.72	5.58	
Wild Bin.Segmentation	50	150	200	300	310	320	330	9.56
	6704.47	12412.99	495.63	19.93	27.84	8.11	11.33	
Segm.Neighbourhood	50	150	200	300	310	320	330	0.52
	14487.84	14415.85	14370.42	14363.5	14352.98	14343.54	14323.95	

Table 9: A table displaying the accuracy and speed for the treated search methods. For the wild binary segmentation, 1000 intervals were simulated at every iteration.

First, we see that in terms of speed the window-sliding algorithm performs best, followed by the binary segmentation, the segment neighbourhood and the wild binary segmentation. In particular the latter seems to be much slower, which is mainly due to all the additional intervals that had to be simulated at every iteration.

On the other hand, the window-sliding algorithms provides lower discrepancy values across the board. It is still able to detect the first few changes with relative ease, but it performs significantly worse for the closely located changes. If we were to try and solve problem (2) with a penalty value of 10, then this algorithm would not even be able to detect any of the latter changes, as none of the discrepancy values would exceed the penalty threshold. The cause of course lies in the fact that we only consider a small part of the data at once, so we are disregarding a lot of information.

In this regard, the binary segmentation does offer better performance. But it again does not provide good enough performance for the densely located changes, which is what we hypothesized earlier. It is for this reason that we had also treated the wild binary segmentation, and judging from the table it does indeed provide better performance for these kind of changes(especially for points 310 and 330), though for point 320 it still only has a marginal improvement. But as we just discussed, this comes at the cost of a big increase in run-time.

So if we ignore wild binary segmentation, then the approximate algorithms are indeed fast, but offer poor performance for densely located change-points and/or small change sizes. The wild binary segmentation mitigated this problem somewhat, but at the expense of a big increase in computational cost. If we now finally consider the Segment Neighbourhood, then not only does it provide vastly superiour performance for all the change-points, but also at only a modest increase of run-time. The advantages of dynamic programming are clearly demonstrated here: finding exact, accurate solutions in an efficient manner such that the increase in computational cost is not too big. We conclude this performance study by recommending binary segmentation when the changes are big and sparsely located, and the segment neighbourhood when harder to discern

changes causes the former to perform badly.

5 Discussion

In this report we have given a general analysis of change-point models under a variety of different models and assumptions. We first defined change-point models in a general sense, where we defined these models as being constituted of the following three parts:

1) Cost functions
2) Search methods
3) Penalty terms

We then spent sections 2 and 3 on treating several different cost-functions. Namely, in section 2 we focused on parametric models, where we could assume a known form of the underlying distribution of the data. Specifically, the normal and gamma models were treated, as these are some of the most used models in practice([2], [4]). We also leaned into the practicality of these cost-functions, by conducting a simulation study where the performance of the derived statistics where evaluated under different situations. The main conclusion that we could draw was that the parametric statistics were not robust against outliers, even after we tried mitigating this with robust estimators and bounded cost-functions.

As a means of extending the practical use of change-point models, and to perhaps also find a more robust method, we then treated non-parametric models. There were no specific distribution functions/densities to work with now, so we had to resort to rank-based statistics and empirical distribution functions. As it turned out when we performed the same simulation studies, these non-parametric statistics were a lot more robust against outliers(especially the rank-based ones).

Finally, we treated some widely-used search methods, which we defined as computational means of extending the single-changepoint problem to multiple change-points. We saw that the approximate algorithms were fast, but did not perform well when the changes were not so drastic and also densely located. On the other hand, the exact algorithm provided vastly superior change-point approximations in every scenario, albeit at the cost of a modest decrease in speed.

Appendix A

A1: Normal statistics

A1.1: Mean change(known variance)

Maximum likelihood mean

The likelihood functions for the null and alternative hypotheses respectively are:

$$\begin{aligned} L_0(\mu,\sigma^2;y_1,...,y_T) &= \prod_{i=1}^{T} f_y(y_i;\mu,\sigma^2) \\ &= \prod_{i=1}^{T} \frac{\exp(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2})}{(2\pi\sigma^2)^{\frac{1}{2}}} \\ &= \frac{\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i-\mu)^2)}{(2\pi\sigma^2)^{\frac{T}{2}}} \end{aligned}$$

$$\begin{aligned} L_1(\mu_1,\mu_T,\sigma^2;y_1,...,y_T) &= \cdots \\ &= \frac{\exp\left(\frac{-(\sum_{i=1}^{k}(y_i-\mu_1)^2+\sum_{i=k+1}^{T}(y_i-\mu_T)^2)}{2}\right)}{(2\pi\sigma^2)^{\frac{T}{2}}} \end{aligned}$$

Since σ^2 is already given, we can obtain unbiased estimators for the mean by calculating the log-variant of the likelihood, maximize it with respect to μ and obtain the MLE for μ :

$$\begin{aligned} l(\mu,\sigma^2;y_1,...,y_T) &= \ln(L(\mu,\sigma^2;y_1,...,y_T)) \\ &= \ln\left((2\pi\sigma^2)^{\frac{-T}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i-\mu)^2\right)\right) \\ &= \ln\left((2\pi\sigma^2)^{\frac{-T}{2}}\right)+\ln\left(\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i-\mu)^2\right)\right) \\ &= -\frac{T}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i-\mu)^2 \\ &= -\frac{T}{2}\ln(2\pi)-\frac{T}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i-\mu)^2 \end{aligned}$$

In order to maximize this expression for μ, we set its derivative with respect to μ equal to zero(this will guarantee us a maximum as the logarithmic function is increasing):

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2; y_1, ..., y_T) = 0 \Rightarrow$$
$$\frac{\partial}{\partial \mu}\left(-\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i-\mu)^2\right) = 0 \Rightarrow$$
$$\frac{1}{\sigma^2}\sum_{i=1}^{T}(y_i-\mu) = 0 \Rightarrow$$
$$\frac{1}{\sigma^2}\left(\sum_{i=1}^{T} y_i - n\mu\right) = 0 \Rightarrow$$
$$\sum_{i=1}^{T} y_i - n\mu = 0 \Rightarrow$$
$$\hat{\mu} = \overline{y} = \frac{1}{T}\sum_{i=1}^{T} y_i$$

A1.2: Mean change(unknown variance)

Maximum likelihood variance

By taking the derivative w.r.t the variance of the log-likelihood and setting it equal to zero:

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2; y_1, ..., y_T) = 0 \Rightarrow$$
$$\frac{\partial}{\partial \sigma^2}\left(-\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_i-\mu)^2\right) = 0 \Rightarrow$$
$$-\frac{T}{2\sigma^2} - \left[\frac{1}{2}\sum_{i=1}^{T}(y_i-\mu)^2\right]\frac{d}{d\sigma^2}\left(\frac{1}{\sigma^2}\right) = 0 \Rightarrow$$
$$-\frac{T}{2\sigma^2} - \left[\frac{1}{2}\sum_{i=1}^{T}(y_i-\mu)^2\right]\left(-\frac{1}{(\sigma^2)^2}\right) = 0 \Rightarrow$$
$$-\frac{T}{2\sigma^2} + \left[\frac{1}{2}\sum_{i=1}^{T}(y_i-\mu)^2\right]\frac{1}{(\sigma^2)^2} = 0 \Rightarrow$$
$$\frac{1}{2\sigma^2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{T}(y_i-\mu)^2 - T\right] = 0$$

Since $\sigma^2 \neq 0$, the above equation is only satisfied when $\hat{\sigma}^2 = \frac{1}{T}\sum_{i=1}^{T}(y_i-\overline{y})^2$.

$S = S_k + E_k$

First, we rewrite S_k as follows:

$$\begin{aligned}
S_k &= \sum_{t=1}^{k}(y_t - \overline{y}_k)^2 + \sum_{t=k+1}^{T}(y_t - \overline{y}_{T-k})^2 \\
&= \sum_{t=1}^{k}(y_t^2 - 2y_t\overline{y}_k + \overline{y}_k^2) + \sum_{t=k+1}^{T}(y_t^2 - 2y_t\overline{y}_{T-k} + \overline{y}_{T-k}^2) \\
&= \sum_{t=1}^{T} y_t^2 - 2\overline{y}_k \sum_{t=1}^{k} y_t + \frac{(\sum_{t=1}^{k} y_t)^2}{k} - 2\overline{y}_{T-k} \sum_{t=k+1}^{T} y_t + \frac{(\sum_{t=k+1}^{T} y_t)^2}{T-k} \\
&= \sum_{t=1}^{T} y_t^2 - \frac{(\sum_{t=1}^{k} y_t)^2}{k} - \frac{(\sum_{t=k+1}^{T} y_t)^2}{T-k}
\end{aligned}$$

Furthermore, we may rewrite E_k as:

$$\begin{aligned}
E_k &= k(\overline{y}_k - \overline{y}_T)^2 + (T-k)(\overline{y}_{T-k} - \overline{y}_T)^2 \\
&= k(\overline{y}_k^2 - 2\overline{y}_k\overline{y}_T + \overline{y}_T^2) + (T-k)(\overline{y}_{T-k}^2 - 2\overline{y}_T\overline{y}_{T-k} + \overline{y}_T^2) \\
&= k\overline{y}_k^2 + (T-k)\overline{y}_{T-k}^2 + k(-2\overline{y}_k\overline{y}_T + \overline{y}_T^2) + (T-k)(-2\overline{y}_T\overline{y}_{T-k} + \overline{y}_T^2)
\end{aligned}$$

Finally, we can write:

$$\begin{aligned}
S_k + E_k &= \sum_{t=1}^{T} y_t^2 + k(-2\overline{y}_k\overline{y}_T + \overline{y}_T^2) + (T-k)(-2\overline{y}_T\overline{y}_{T-k} + \overline{y}_T^2) \\
&= \sum_{t=1}^{T} y_t^2 - 2\overline{y}_T \sum_{t=1}^{k} y_t + k\overline{y}_T^2 - 2\overline{y}_T \sum_{t=k+1}^{T} y_t + (T-k)\overline{y}_T^2 \\
&= \sum_{t=1}^{T} y_t^2 - 2\overline{y}_T \sum_{t=1}^{T} y_t + T\overline{y}_T^2 \\
&= \sum_{t=1}^{T}(y_t^2 - 2y_t\overline{y}_T + \overline{y}_T^2) = S
\end{aligned}$$

A2: Gamma statistics

A2.1: Scale change

Likelihood functions

$$\begin{aligned}
L_0(\theta_0) &= f(y_1, ..., y_T; \theta_0) \\
&= \prod_{t=1}^{T} \frac{1}{\theta_t^{\xi}\Gamma(\xi)} y_t^{\xi-1} e^{-(\frac{y_t}{\theta_t})} \\
&= \frac{\prod_{t=1}^{T} y_t^{\xi-1}}{\Gamma^T(\xi)} \prod_{t=1}^{T} \frac{1}{\theta_t^{\xi}} e^{-(\frac{y_t}{\theta_t})} \\
&= \frac{\prod_{t=1}^{T} y_t^{\xi-1}}{\Gamma^T(\xi)} \exp\left[\sum_{t=1}^{T}\left(-\frac{y_t}{\theta_0} - \ln\theta_0^{\xi}\right)\right]
\end{aligned}$$

$$\begin{aligned}
L_1(\theta_0, \delta) &= f(y_1, ..., y_T; \theta_0, \delta) \\
&= \sum_{j=1}^{T-1} \pi_T(j) f(y_1, ..., y_T; \theta_0, \delta) \mid j)
\end{aligned}$$

$$= \frac{1}{T-1} \sum_{j=1}^{T-1} \left[\left(\prod_{t=1}^{j} \frac{1}{\theta_t^{\xi}\Gamma(\xi)} y_t^{\xi-1} e^{-(\frac{y_t}{\theta_t})}\right)\left(\prod_{t=j+1}^{T} \frac{1}{\theta_t^{\xi}\Gamma(\xi)} y_t^{\xi-1} e^{-(\frac{y_t}{\theta_t})}\right)\right]$$

$$= \frac{1}{T-1} \sum_{j=1}^{T-1} \left[\left(\prod_{t=1}^{j} \frac{1}{\theta_t^{\xi}\Gamma(\xi)} y_t^{\xi-1} e^{-(\frac{y_t}{\theta_t})}\right)\left(\prod_{t=j+1}^{T} \frac{1}{(\theta_0+\delta)^{\xi}\Gamma(\xi)} y_t^{\xi-1} e^{-(\frac{y_t}{(\theta_0+\delta)})}\right)\right]$$

$$= \frac{1}{T-1} \frac{\prod_{t=1}^{T} y_t^{\xi-1}}{\Gamma^T(\xi)} \sum_{j=1}^{T-1} \exp\left[\sum_{t=1}^{j}\left(-\frac{y_t}{\theta_0} - \ln\theta_0^{\xi}\right)\right]$$

$$\cdot \exp\left[\sum_{t=j+1}^{T}\left(-\frac{y_t}{\theta_0+\delta} - \ln(\theta_0+\delta)^{\xi}\right)\right]$$

Maximum likelihood scale

The likelihood function is:

$$\prod_{t=1}^{T} \frac{1}{\theta^{\xi}\Gamma(\xi)} y_t^{\xi-1} e^{-(\frac{y_t}{\theta})} = \frac{1}{\Gamma(\xi)^T} \frac{1}{\theta^{T\xi}} e^{-\frac{K}{\theta}} \prod_{t=1}^{T} y_t^{\xi-1}$$

Here we defined K as $\sum_{t=1}^{T} y_t$. By applying the logarithm and deriving the

resulting expression with respect to θ and equating it to zero, we obtain the scale's MLE estimate:

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\log(\frac{1}{\Gamma(\xi)^T}\frac{1}{\theta^{T\xi}}e^{-\frac{K}{\theta}}\prod_{t=1}^{T}y_t^{\xi-1}) = 0 \Rightarrow \\
&\frac{\partial}{\partial\theta}((\xi-1)\sum_{t=1}^{T}\log y_t - \frac{K}{\theta} + (T\xi)\log(\frac{1}{\theta}) - T\log(\Gamma(\xi))) = 0 \Rightarrow \\
&\frac{K}{\theta^2} - T\xi\frac{1}{\theta} = 0 \Rightarrow \\
&\frac{1}{\theta}(\frac{K}{\theta} - T\xi) = 0 \Rightarrow \\
&\hat{\theta} = \frac{K}{T\xi} = \frac{\sum_{t=1}^{T}y_t}{T\xi}
\end{aligned}$$

Likelihood-ratio

$$\begin{aligned}
\Lambda = \frac{L_1(\theta_0,\delta)}{L_0(\theta_0)} &= \frac{1}{T-1}\sum_{j=1}^{T-1}\exp\left[\sum_{t=j+1}^{T}\frac{\delta y_t}{\theta_0^2} - \frac{\xi\delta}{\theta_0} + o(\delta)\right] \\
&= \sum_{j=1}^{T-1}\frac{1}{T-1}\left\{1 + \sum_{t=j+1}^{T}\left[\frac{\delta y_t}{\theta_0^2} - \frac{\xi\delta}{\theta_0}\right] + (T-j)o(\delta)\right\} \\
&= \sum_{j=1}^{T-1}\frac{1}{T-1}\left\{1 + \sum_{t=j+1}^{T}\left[\frac{\delta y_t}{\theta_0^2} - \frac{\xi\delta}{\theta_0}\right] + o(\delta)\right\} \\
&= 1 + \frac{1}{T-1}\sum_{j=1}^{T-1}\sum_{t=j+1}^{T}\left[\frac{\delta y_t}{\theta_0^2} - \frac{\xi\delta}{\theta_0}\right] + o(\delta) \\
&= 1 + \frac{\delta}{\theta_0}\left[\frac{1}{(T-1)\theta_0}\sum_{j=1}^{T-1}\sum_{t=j+1}^{T}y_t - \frac{T\xi}{2}\right] + o(\delta)
\end{aligned}$$

A2.2: Shape change

Likelihood function

$$\begin{aligned}
L_1(y_t \mid \xi_0,\delta,\theta,k) &= \prod_{t=1}^{T}\frac{1}{\theta^{\xi_t}\Gamma(\xi_t)}y_t^{\xi_t-1}e^{-(\frac{y_t}{\theta})} \\
&= \exp\left\{-\sum_{t=1}^{T}\frac{y_t}{\theta}\right\}\prod_{t=1}^{T}y_t^{\xi_t-1}\prod_{t=1}^{T}\frac{1}{\theta^{\xi_t}}\prod_{t=1}^{T}\frac{1}{\Gamma(\xi_t)}
\end{aligned}$$

$$= \exp\left\{-\sum_{t=1}^{T}\frac{y_t}{\theta} + \sum_{t=1}^{k}[(\xi_0-1)\ln y_t - \ln\Gamma(\xi_0) - \xi_0\ln\theta] + \sum_{t=k+1}^{T}[(\xi_0+\delta-1)\ln y_t - \ln\Gamma(\xi_0+\delta) - (\xi_0+\delta)\ln\theta]\right\}$$

Now we define $\eta(y_t;\xi,\theta) = (\xi-1)\ln y_t - \ln\Gamma(\xi) - \xi\ln\theta$, for t = 1, ..., T. As before, we will derive a semi-bayesian approach where we make an a priori assumption about the change-point location, i.e. that it is equally likely to occur on any of its possible locations. By summing over all the possible locations k between 1 and T - 1, we now obtain for our likelihood under H_1 :

$$L_1(y_t \mid \xi_0,\delta,\theta) = \sum_{k=1}^{T-1}(T-1)^{-1}L_1(y_t \mid \xi_0,\delta,\theta,k)$$

$$= (T-1)^{-1}\exp\left\{-\sum_{t=1}^{T}\frac{y_t}{\theta}\right\}\cdot\sum_{k=1}^{T-1}\exp\left\{\sum_{t=1}^{k}\eta(y_t;\xi,\theta) + \sum_{t=k+1}^{T}\eta(y_t;\xi+\delta,\theta)\right\}$$

Under the additional assumption that the jump δ is not too big, we can obtain a more convenient expression for the likelihood. First, we note that in a close neighbourhood of ξ (i.e. a small value of δ), we can approximate $\eta(y_t;\xi+\delta,\theta)$ as follows:

$$\begin{aligned}\eta(y_t;\xi+\delta,\theta) &= \eta(y_t;\xi,\theta) + \delta\eta'(y_t;\xi,\theta) + o(\delta)\\ &= \eta(y_t;\xi,\theta) + \delta[\ln y_t - \ln\theta - \Psi(\xi)] + o(\delta)\end{aligned}$$

Here, we define $\Psi(\xi)$ to be the first derivative of $\ln\Gamma(\xi)$. This allows us to write the likelihood as follows:

$$L_1(y_t \mid \xi_0,\delta,\theta) = (T-1)^{-1}\exp\left\{-\sum_{t=1}^{T}\frac{y_t}{\theta} + \eta(y_t;\xi,\theta)\right\}\cdot$$

$$\sum_{k=1}^{T-1}\exp\left\{\delta\sum_{t=k+1}^{T}[\ln y_t - \ln\theta - \Psi(\xi) + o(\delta)]\right\}$$

$$= (T-1)^{-1}\exp\left\{-\sum_{t=1}^{T}\frac{y_t}{\theta} + \eta(y_t;\xi,\theta)\right\}\cdot$$

$$\sum_{k=1}^{T-1}\left\{1+\delta\sum_{t=k+1}^{T}[\ln y_t - \ln\theta - \Psi(\xi) + o(\delta)]\right\}, \text{ as } \delta\to 0.$$

Maximum likelihood $\Psi(\xi) + \ln\hat{\theta}$

We start from the log-likelihood assuming H_0 holds:

$$
\begin{aligned}
l(\xi_0) = \ln L_0(\xi_0, \hat{\theta}) &= \ln \prod_{t=1}^{T} f(y_t; \xi_0, \hat{\theta}) \\
&= \ln \prod_{t=1}^{T} \frac{\prod_{t=1}^{T} y_t^{\xi_0 - 1}}{\Gamma^T(\xi_0)} \exp\left[\sum_{t=1}^{T}\left(-\frac{y_t}{\hat{\theta}_0} - \ln \hat{\theta}_0^{\xi_0}\right)\right] \\
&= (\xi_0 - 1)\sum_{t=1}^{T} \ln(y_t) - \sum_{t=1}^{T} \frac{y_t}{\hat{\theta}} - Tk\ln(\hat{\theta}) - T\ln(\Gamma(\xi_0)) \\
&= (\xi_0 - 1)\sum_{t=1}^{T} \ln y_t - T\xi_0 - T\xi_0 \ln\left(\frac{\sum y_t}{\xi_0 T}\right) - T\ln(\Gamma(\xi))
\end{aligned}
$$

Taking the derivative with respect to ξ then yields:

$$
\begin{aligned}
\partial_\xi l(\xi_0) &= \sum_{t=1}^{T} \ln y_t - T - T\ln\left(\frac{\sum y_t}{\xi_0 T}\right) - \frac{T}{\Gamma(\xi_0)} \cdot \Gamma\prime(\xi_0) + T\xi_0 \frac{\xi_0 T}{\sum y_t} \frac{\sum y_t}{T} \frac{1}{\xi_0^2} \\
&= \sum_{t=1}^{T} \ln y_t - T\ln\left(\frac{\sum y_t}{\xi_0 T}\right) - T\frac{\Gamma\prime(\xi_0)}{\Gamma(\xi_0)}
\end{aligned}
$$

Denoting $\dfrac{\Gamma\prime(\xi_0)}{\Gamma(\xi_0)}$ as $\Psi(\xi_0)$ and equating the equation to zero finally leads us to:

$$
\sum_{t=1}^{T} \ln y_t - T\ln \sum_{t=1}^{T} y_t + T\ln\xi_0 + T\ln T - T\Psi(\xi_0) = 0
$$

The equation can not be simplified anymore in regards to ξ_0, meaning that it is not possible to derive an analytic expression for the MLE of ξ_0. However, using some clever mathematical reasoning we can derive a MLE-estimator for $\Psi(\xi_0) + \ln\hat{\theta}$. First note that from the above equation we can write:

$$
\begin{aligned}
T(\ln\xi_0 - \Psi(\xi_0)) = T\ln \sum_{t=1}^{T} y_t - \sum_{t=1}^{T} \ln y_t - T\ln T \Rightarrow \\
\ln\xi_0 - \Psi(\xi_0) = \ln \sum_{t=1}^{T} y_t - \frac{1}{T}\sum_{t=1}^{T} \ln y_t - \ln T \\
= \ln \frac{1}{T}\sum_{t=1}^{T} y_t - \frac{1}{T}\sum_{t=1}^{T} \ln y_t
\end{aligned}
$$

We now cleverly rewrite $\Psi(\xi_0) + \ln\hat{\theta}$:

$$
\Psi(\xi_0) - \ln\xi_0 + \ln\xi_0 + \ln\hat{\theta}
$$

$$
\begin{aligned}
&= \frac{1}{T}\sum_{t=1}^{T}\ln y_t - \ln\frac{1}{T}\sum_{t=1}^{T} y_t + \ln\xi_0 + \ln\hat{\theta} \\
&= \frac{1}{T}\sum_{t=1}^{T}\ln y_t - \ln\frac{1}{T}\sum_{t=1}^{T} y_t + \ln\xi_0 + \ln\frac{1}{T}\sum_{t=1}^{T} y_t - \ln\xi_0 \\
&= \frac{1}{T}\sum_{t=1}^{T}\ln y_t
\end{aligned}
$$

Bibliography

References

[1] G.A. Young, R.L.Smith. *Essentials of Statistical Inference*. Cambridge University Press, 2005.

[2] J. Chen, Arjun K. Gupta. *Parametric Statistical Change Point Analysis, with applications to Genetics, Medicine and Finance*. Birkhäuser Basel, 2012.

[3] P. Fearnhead, G. Rigaill. *Changepoint detection in the presence of outliers*. *arXiv:1609.07363*, 2017.

[4] D.A. Hsu. *Detecting Shifts of Parameter in Gamma Sequences, with Applications to Stock Price and Air Traffic Flow Analysis*. Journal of the American Statistical Association, 74:365, 31-40, 1979.

[5] A. Ramanayake. *Tests for a Change Point in the Shape Parameter of Gamma Random Variables*. Communications in Statistics - Theory and Methods, 33:4, 821-833, 2005.

[6] Y.C. Yao. *Estimating the number of change-points via Schwarz' criterion*. Statistics and Probability Letters, 6:181-189, 189.

[7] H. Akaike. *Information theory and an extension of the maximum likelihood principle*. 2nd International Symposium on Information Theory, 267-281, 1973.

[8] G. Schwarz. *Estimating the Dimension of a Model*. The Annals of Statistics, 6(2):461-464, 1978.

[9] A. Sen, M.S. Srivastava. *On tests for detecting change in mean*. Annals of Statistics, 3:98–108, 1975.

[10] A.M. Mood. *On the asymptotic efficiency of certain nonparametric two-sample tests*. The Annals of Mathematical Statistics, 25(3):514-522, 1954.

[11] A.R. Ansari, R.A. Bradley. *Rank-sum Tests for Dispersions*. Virginia Polytechnic Institute, 1959.

[12] N.V. Smirnov. *Estimate of deviation between empirical distribution functions in two independent samples*. Bulletin Moscow University, 2:3-16, 1933.

[13] A.N. Kolmogorov. *Sulla determinazione empirica di una legge di distribuzione*. Giornale dell' Istituto Italiano degli Attuari, 4: 83–91, 1933.

[14] H. Cramér. *On the Composition of Elementary Errors*. Scandinavian Actuarial Journal, 1:13-74, 1928.

[15] R.E. von Mises. *Wahrscheinlichkeit*. Statistik und Wahrheit(Julius Springer), 1928.

[16] G.J. Ross, N.M. Adams. *Two non-parametric control charts for detecting arbitrary distribution changes*. Journal of Quality Technology, 44(2), 1959.

[17] L. Vostrikova. *Detecting 'disorder' in multidimensional random processes*. Soviet Mathematics Doklady, 24:55-59, 1981.

[18] P. Fryzlewicz. *Wild Binary Segmentation for multiple change-point detection*. The London School of Economics and Political Science, 2014.

[19] I.E. Auger, C.E. Lawrence. *Algorithms for the optimal identification of segment neighborhoods*. Bulletin of Mathematical Biology, 51(1):39-54, 1989.

[20] K. Haynes, I.A. Eckley, P.Fearnhead. *Efficient penalty search for multiple changepoint problems. arXiv:1412.3617*, 2014.

[21] D.M. Hawkins *Testing a sequence of observations for a shift in location. Journal of the American Statistical Association*, 72:180-186, 1977.