Nonparametric Changepoint Analysis for Bernoulli Random Variables Based on Neural Networks

Gichuhi, Anthony Waititu

Vom Fachbereich Mathematik
der Technische Universität Kaiserslautern
zur Erlangung des Akademischen Grades
Doktor der Naturwissenschaften
(Doctor rerum naturalium, Dr. rer. nat.)
Genehmigte Dissertation

1. **Gutachter: Prof. Dr. Jürgen Franke**
2. **Gutachter: Prof. Dr. Heinrich von Weizsäcker**

To my Family: Rose, Joy and Victor

Acknowledgment

I thank the almighty God for His grace and protection throughout my studies.

My special regards go to Prof. Dr. Jürgen Franke, Kaiserslautern University, Germany, for supporting, guiding and encouraging me throughout the study period. I still remember when he wrote to DAAD Nairobi Kenya, recommending me for a Ph.D. position in Kaiserslautern University. Without his assistance and understanding, I would not have succeeded in my Ph.D. studies.

My gratitude also goes to Dr. Marlene Müller, Fraunhofer Institute, Germany, for accepting to be the second referee to my work.

The efforts put across by Dr. Mwita, Jomo Kenyatta University, Kenya, can not go unmentioned. Thanks a lot, Dr. Mwita, for introducing me first to quantile regression and then to Prof. Dr. Jürgen Franke.

I also thank the Kaiserslautern Statistics group for all the seminars we did together and for all the academic and non-academic discussions we held. Special regards go to the secretary of statistics department, Ms. Beate Siegler, for her kindness and willingness to help.

I know that my family, to whom I have dedicated this work, sacrificed a lot to see me through my studies. I sincerely thank my wife Rose and children Joy and Victor for their understanding. I know I have been away for such a long time but life is back to normal now that I have finished my studies.

My gratitude also goes to my close friends Fr. Joachim Lieberich, Dr. Stephane Lieberich's family and HansPeter's family. I wish these people God's blessings.

Last but not the least, I thank all my friends and all those people not mentioned above but contributed to my success in one way or another.

Abstract

In many medical, financial, industrial, e.t.c. applications of statistics, the model parameters may undergo changes at unknown moment of time.

In this thesis, we consider change point analysis in a regression setting for dichotomous responses, i.e. they can be modeled as Bernoulli or 0-1 variables. Applications are widespread including credit scoring in financial statistics and dose-response relations in biometry.

The model parameters are estimated using neural network method. We show that the parameter estimates are identifiable up to a given family of transformations and derive the consistency and asymptotic normality of the network parameter estimates using the results in Franke and Neumann [24].

We use a neural network based likelihood ratio test statistic to detect a change point in a given set of data and derive the limit distribution of the estimator using the results in Gombay and Horvath ([28], [30] under the assumption that the model is properly specified. For the misspecified case, we develop a scaled test statistic for the case of one-dimensional parameter. Through simulation, we show that the sample size, change point location and the size of change influence change point detection.

In this work, the maximum likelihood estimation method is used to estimate a change point when it has been detected. Through simulation, we show that change point estimation is influenced by the sample size, change point location and the size of change.

We present two methods for determining the change point confidence intervals: Profile log-likelihood ratio and Percentile bootstrap methods. Through simulation, the Percentile bootstrap method is shown to be superior to profile log-likelihood ratio method.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

In many medical, financial, industrial, etc applications of statistics, it is important to consider that the model parameters may undergo changes at unknown moment of time.

The time moment when the model has changed is called *change point*. Other synonyms are *probabilistic diagnostics* and *disorder problems*.

As an example, the annual discharges of the Nile River at Aswan from 1871 to 1970 has been studied by many authors including Cobb [9]. The objective of the Nile river study was first to detect whether there was a change in flow discharge and secondly to estimate the change if it was detected. The above authors found out that there was a change in discharge flow in the year 1899.

The change point problem is two fold: Change point detection and change point estimation.

Let $x_1, \ldots, x_n$ be a sequence of independent and identically distributed random variables having distribution function $F(x; \theta_{H0})$ under normal operating conditions and $F(x; \theta_{H1})$ when the normal operating conditions change. θ is the mean parameter.

To set up the problem of change point detection, one first specifies the acceptable region ω_0 in which θ should reside under normal operating conditions and the unacceptable region ω_1. This is accomplished by formulating two hypothesis as shown below:

$$\begin{aligned} H_0 \quad & : \theta_1 \quad = \ldots = \theta_n = \theta_{H0} \\ & \textit{Versus} \\ H_1 \quad & : \theta_1 \quad = \ldots = \theta_K = \theta_{H0}; \theta_{K+1} = \ldots = \theta_n = \theta_{H1} \end{aligned}$$

where $K = 1, \ldots, n-1$ is an unknown index of the shift point.

One then develops the rejection criterion of the above hypothesis as discussed in chapter two. A change point is detected when H_0 is rejected. The

next step of change point estimation is carried out only when the null hypothesis of no change point is rejected.

In this thesis we consider change point analysis in a regression setting in case that the responses are dichotomous, i.e. they can be modeled as Bernoulli or 0-1 variables. Applications are widespread, e.g. credit scoring in financial statistics or dose-response relations in biometry, to mention only two of them.

The parameter of interest is the probability for observing 1 which depends on various predictor variables. We consider a nonparametric setting where this parameter is a rather arbitrary function. We study the change point problem i.e. we consider tests if this function changes somewhere in the sample, and we discuss how to estimate the location of the change.

Before we proceed to formulate an appropriate mathematical framework, we discuss various aspects of change point problems and the relevant literature.

Completeness of a priori statistical information

Depending on whether the probabilistic model of data is known or not, one can distinguish between parametric, semi-parametric and non-parametric methods of change point detection and estimation.

Initial change point studies were based on a sequence of random variables without considering regression models. Worsley [70] used the likelihood ratio method to test for a change in probability of a sequence of independent binomial variables. In this paper, the exact iterative procedure for the exact null and alternative distribution of likelihood ratio statistics were found.

Non-parametric detection of a change point in a sequence of random variables was studied by many authors. Page [53] used the cumulative sum (CUSUM) technique to test for a possible change point. Worsley [70] used the cumulative sum statistics to test for a change in probability of a sequence of independent binomial random variables.

Maximum likelihood estimate (MLE) method has been used to estimate a change point when the probabilistic data model is known. Hinkley [36] applied the MLE method to estimate a change point in a sequence of normally distributed random variables whereby he derived the asymptotic distribution of the estimator using random walk theory. Hinkley and Hinkley [37] used the MLE method to estimate the change point in a sequence of zero-one variables.

Pettitt [58] used a Mann-Whitney type statistic to non-parametrically estimate a change point when it is known that a change has taken place at an unknown point in a sequence of random variables. In this work, the estimate is compared with MLE using Monte Carlo techniques and found to

be fairly constant over various distributions like normal distribution.

Regression based approaches to change point analysis have been considered frequently. The parametric testing of a change point in simple linear regression is discussed in Kim and Siegmund [46] where a likelihood ratio test method is used. In their work, Kim and Siegmund [46] dealt with two cases. The first case is when the alternative hypothesis specifies that only the intercept changes while the second case is when the alternative allows both the intercept and the slope to change.

Non-parametric testing of a change point in linear regression has been studied by authors like Aue *et al* [4]. Aue *et al* [4] studies two schemes for change point detection in generalized linear models. The first scheme is essentially the CUSUM of residuals test while the second scheme is based on the work of Clark and McCracken [8]. The latter scheme is based on squared prediction errors.

Hinkley [35] dealt with the estimator of a change point in the linear regression setting. The emphasis in this paper was in estimating and making inference about the change point estimator. This work employed MLE techniques to estimate the two-phase regression change point.

For non-parametric estimation of a change point in a linear case, Hsu [39] uses the linear least squares method. Hsu [39] showed that the least squares change point estimator remains consistent when there is a one-time break but it may identify a spurious change when there is none.

Change point problems occur frequently in medical research. An example can be found in MacNeil and Mao [50] where it was found that cancer incidence rates remain relatively stable for people at a younger age but change drastically after a certain threshold. Also, Gallant [26] found out that the weight/height ratio of preschool boys relates to their age in one way before a certain age but that the functional relation of the two changes after-wards.

Nonlinear regression has been used to model biomedical/epidemiological change point problems. The modeling is presented within the framework of logistic regression models which have been used by authors like Pastor-Barriuso *et al* [57] and Vexler and Gurevich [66] to analyze the relationship between some explanatory variables and a dependent Bernoulli variable.

Pastor-Barriuso *et al* [57] use a logistic regression technique to model dose-response relationship which is believed to follow two different regressions. To test and estimate a change point, they use a modified iterative reweighted least squares algorithm.

Vexler and Gurevich [66] use a nonparametric logistic regression method to model threshold problems. In particular, they use polynomial approximation where they estimate the parameters using local maximum likelihood method.

In this study, we focus on nonparametric regression with an epidemiological application in mind. The model functions are estimated using neural network techniques.

As indicated above, parametric test statistics for a change point are based on likelihood ratio statistics and the estimation on maximum likelihood method. More general results can be found in Csörgő and Horváth [11].

As noted in Guan [32], most nonparametric change point models assume no relationship between the response and explanatory variables. See Csörgő and Horváth [11] for more details.

Semi-parametric methods bridge the gap between parametric and nonparametric methods. In many applications, it is common to assume a link between the response and the explanatory variables. This means that an assumption about the existence of such a link function is made. The logistic function is such a function that has been used to link a dichotomous outcome to some regressors. Guan [32] use a semi-parametric approach to test for and estimate a change point. This is achieved by using the empirical likelihood method to efficiently use auxiliary information about the relationship between the two population distributions.

Method of Data Acquisition

Change point problems can be classified as either fixed sample (also called off-line) or sequential setting (also called on-line).

In on-line change point problems, a sequence of independent observations $Y_1, Y_2, \ldots$ is observed sequentially from a given process. At first, the $Y_i's$ have the same distribution g_o. The process is therefore said to be *in control*. However, the process may go out of control at some unknown time K and the $Y_i's$ have another distribution g_1.

Various procedures for the on-line change point problem exist depending on whether g_o and g_1 (or one of them) are assumed known or unknown.

Page's CUSUM and Shewhart's control chart are some of the popular procedures used when both the pre-change distribution g_o and post-change distribution g_1 are completely specified. Yashchin [72] uses the likelihood ratio strategy.

However, in line with statistical quality control, standard procedure assumes that the pre-change distribution g_o is known but the post-change distribution g_1 is unknown and therefore has to be estimated. Such a study has been done in Siegmund and Venkatraman [61].

Some real life problems arise where both g_o and g_1 are unknown. Gordon and Pollak [31] deal with the case when g_o is not completely specified. Mei [52] deals with the case when both g_o and g_1 are completely unspecified.

In on-line change point problems, one has to fine tune the sliding window size because the procedures involved here do not take into account the whole data at once for change point detection and estimation.

Off-line change point problems deal with a fixed sample $Y_1, \ldots, Y_n$ which is first observed and then the detection and estimation of the change point is carried out.

Page [53] introduced the off-line change point problem by assuming that for an unknown change point K, $Y_1, \ldots, Y_K$ has a distribution function $f(Y; \theta_0)$ and $Y_{K+1}, \ldots, Y_n$ has a distribution function $f(Y; \theta_1)$.

When both θ_0 and θ_1 are known, Page ([53],[54]) use the likelihood ratio method to detect the change point and the maximum likelihood method to estimate it.

Hinkley and Hinkley [37] dealt with the estimation of a change point in a sequence of binomial variables for the case when both θ_0 and θ_1 (or one of them) are known or unknown.

A number of authors have studied off-line change point models with covariates, see Kim and Siegmund [46] and Csörgő and Horváth [11] for more details.

A lot of literature focuses on change point problems for continuous outcomes. As noted in Pastor-Barriuso *et al* [57], few models are available for dichotomous outcomes.

Our motivation comes from credit scoring and epidemiological perspectives where the response variable is dichotomous and regresses on various covariates. In particular, we deal with an off-line nonlinear regression model.

Data Characteristics

A random process is either in the discrete or continuous domain. One can therefore distinguish between change point problems for discrete and continuous random processes.

Change point problems for discrete random processes have been studied by many authors. Hinkley and Hinkley [37] studied the inference about the change point in a sequence of binomial variables where the asymptotic distribution of the MLE is derived using random walk results. In this study, the asymptotic distribution of the likelihood ratio statistic for testing hypothesis about the change point is obtained.

The random walk distributions in Hinkley and Hinkley [37] are computer intensive especially when the sample size is large. Worsley [70] developed an iterative procedure for the exact null and alternative distributions of likelihood ratio for testing a change in probability of a sequence of independent binomial random variables.

Due to recursion, as noted in Horváth [38], the iterative procedure in

Worsley [70] is computationally very difficult and time consuming especially if the sample size is large. Horváth [38] developed the limit theorems for likelihood ratio and cumulative sum tests for a sequence of binomial random variables.

More recent studies concentrate on regression based discrete change point problems. Such a study has been done in Pastor and Guallar [56], Pastor-Barriuso *et al* [57] and Vexler and Gurevich [66]. All the three studies above deal with a binomial response variable which is regressed on some covariates.

In this study, we model a discrete (Bernoulli) response variable regressed on a given number of explanatory variables.

Continuous change point problems have also attracted attention from many authors. Inference about the change point in a sequence of normally distributed random variables can be found in Hinkley [36]. Pettitt [58] uses a Mann-Whitney type statistic to estimate a change point in a sequence of continuous random variables.

Jandhyala and Fotopoulos [43] developed a change point methodology for identifying changes in the scale and shape parameters of a Weibull distribution. Weibull distribution is widely applied to model data on climatological factors such as maximum/minimum temperatures.

Regression based continuous change point models exist in the literature. Loader [47] considered a normally distributed regression model in which the mean function may have a discontinuity at an unknown point.

Lastly, a random process may or may not exhibit statistical dependence. Change point problems can therefore be formulated for random sequences with independent observations and random sequences with dependent observations.

For the dependent case, Davis *et al* [12] tested for a change in the parameter values and order of an autoregressive model and showed that if the white noise in the AR model is weakly stationary with finite fourth moments, then under the null hypothesis of no change point, the normalized Gaussian likelihood ratio test statistic converges in distribution to the Gumbel extreme value distribution.

Earlier work on regression based change point models in time series setting was done by Solow [62] and later modified by Easterling and Peterson [14], Elsner *et al* [16] and Lund and Reeves [49].

The independent observations case has also attracted a lot of research. All the literature quoted in the above section *Completeness of a priori statistical information* deal with the independent case.

This study deals with an independent case with the motivation that iid assumption is more suitable for cross-sectional samples than time series, see White [67] for more details. Moreover, the iid assumption greatly simplifies

the necessary assumptions and proofs as well as allowing a very clear statement of the essential results. When the appropriate structure to replace the iid assumption is available (eg strictly stationary ergodic vectors), one can relax the iid assumption and reach the same kind of conclusions, White [67].

Type of Change Point

The change in probabilistic characteristics of observations can be abrupt or gradual.

All the literature quoted in the above three sections, *Completeness of a priori statistical information*, *Method of Data Acquisition*, and *Data Characteristics*, deal with abrupt change point. Abrupt change point problems are problems with a sudden break of model parameter(s).

Changes in a sequence of random process can be more than one. Pan and Chen [55] applied the modified information criterion to detect multiple change points in an off-line sequence of independent random variables. Abrupt multiple change point problems in on-line setting have been studied by authors like Braun *et al* [6] and Fearnhead and Liu [19].

In application areas like engineering and ecology, one often observes a sequence of variables that at some unknown point starts changing its behavior gradually. Such a problem has been studied in Jaruskova [44].

This study deals with the abrupt case with one change point which has wide application in areas like econometrics and biomedicine.

Chapter 2

CONCEPTS AND RESULTS NEEDED

2.1 Neural Networks and Logistic Regression

In this section, we discuss the neural network used. We then discuss some principles of logistic regression. Lastly, we discuss the link between neural network and logistic regression.

2.1.1 Neural Network

An artificial neural network (ANN) is a parallel connection of a set of nodes called neurons. From the statistical viewpoint, it represents a function of explanatory variables which is composed of simple building blocks and which may be used to provide an approximation of conditional expectations or, in particular, probabilities in regression. Below we define an ANN with an input layer, hidden layer and an output layer.

Definition of the ANN

In this study, we only consider a feed-forward net with $d+1$ input nodes, one layer of H hidden nodes, one output node and an activation function $\psi(x)$.The input and hidden layer nodes are connected by weights W_{hj} for $h \in \{1, \ldots, H\}$ and $j \in \{0, \ldots, d\}$.
The hidden and output layers are connected by weights α_h for $h \in \{0, \ldots, H\}$.

Considering an input vector $\mathbf{x} = (x_1, \ldots, x_d) \in \Re^d$,then the input $\upsilon_h(\mathbf{x})$

to the h^{th} hidden node is the value

$$\upsilon_h(\mathbf{x};\boldsymbol{\theta}) = W_{h0} + \sum_{j=1}^{d} W_{hj}x_j \tag{2.1}$$

The output $\phi_h(\mathbf{x};\boldsymbol{\theta})$ of the h^{th} hidden node is the value

$$\phi_h(\mathbf{x};\boldsymbol{\theta}) = \psi(\upsilon_h(\mathbf{x};\boldsymbol{\theta})) \tag{2.2}$$

The net input to the output node is the value

$$O_H(\mathbf{x};\boldsymbol{\theta}) = \alpha_0 + \sum_{h=1}^{H} \alpha_h\phi_h(\mathbf{x},\boldsymbol{\theta}) \tag{2.3}$$

Finally, the output $Z(\mathbf{x};\boldsymbol{\theta})$ of the net is the value

$$Z(\mathbf{x};\boldsymbol{\theta}) = \psi(O_H(\mathbf{x};\boldsymbol{\theta})) \tag{2.4}$$

$\boldsymbol{\theta}$ stands for all the parameters $\alpha_0,\ldots,\alpha_H$ and $W_{hj},\ h=1,\ldots,H,\ j=0,\ldots,d$, of the network. We also write $\boldsymbol{\alpha} = (\alpha_0,\ldots,\alpha_H)^T$ and $\mathbf{W} = (W_{hj},\ h=1,\ldots,H,\ j=0,\ldots,d)$

We note that one does not necessarily need the same function $\psi(.)$ in (2.2) and (2.4). In unbounded regression, $\psi(.)$ in (2.4) is frequently just the identity function, i.e. $Z(\mathbf{x};\boldsymbol{\theta}) = O_H(\mathbf{x};\boldsymbol{\theta})$. However, as we are interested in conditional probabilities, it is convenient to use a function $\psi(.)$ with values in [0,1] in (2.4). As such bounded and typically sigmoid functions are also appropriate for (2.2), we do not distinguish between different functions for ease of notation.

As stated earlier, $\psi(.)$ is an activation function. There exists two main activation functions:

(i)**The Unipolar or Logistic Activation Function**
It takes the form

$$\begin{aligned}\psi(x) &= \frac{exp(a(x-b))}{1+exp(a(x-b))} \\ &= \frac{1}{1+exp(-a(x-b))}\end{aligned} \tag{2.5}$$

where

$$\psi(x) = \begin{cases} 1, & x\to\infty \\ 0, & x\to-\infty \end{cases} \tag{2.6}$$

a is the learning rate while b is called the bias. Since the unipolar activation function maps $\Re \rightarrow [0,1]$, it is very practical to Bernoulli or binomial data.

(ii)The Bipolar or Hyperbolic Tangent Activation Function
It takes the form

$$\begin{aligned}
\psi(x) &= 2\{\frac{1}{1+exp(-ax)}\} - 1 \\
&= \frac{1-exp(-ax)}{1+exp(-ax)} \\
&= \frac{exp(ax/2)\{1-exp(-ax)\}}{exp(ax/2)\{1+exp(-ax)\}} \\
&= \frac{\{exp(ax/2)-exp(-ax/2)\}/2}{\{exp(ax/2)+exp(-ax/2)\}/2} \\
&= \frac{sinh(ax/2)}{cosh(ax/2)} \\
&= tanh(ax/2) \qquad (2.7)
\end{aligned}$$

where

$$\begin{cases}
\psi(x) \rightarrow 1 & \text{as } x \rightarrow \infty \\
\psi(x) \rightarrow -1 & \text{as } x \rightarrow -\infty \\
\psi(x) + \psi(-x) = 0
\end{cases} \qquad (2.8)$$

Both the unipolar and bipolar sigmoids are continuously differentiable.

The connection weights are adjusted through training. There exists two training paradigms: Non supervised and supervised learning. We discuss and later apply supervised learning.

The supervised training of a neural net requires the following:

1. A sample of n input vectors, $\mathbf{X} = \mathbf{X}_1, \ldots, \mathbf{X}_n \ \in \Re^d$ of size d each and an associated output vector, $\mathbf{Y} = Y_1, \ldots, Y_n \ \in \Re$.

2. The selection of an initial weight set.

3. A repetitive method to update the current weights to optimize the input-output map.

4. A stopping rule.

We discuss the third requirement above because it is more challenging.
The maximum likelihood method is used to determine the error function which is then used to train a given network. We now discuss error functions.

Error Functions

The error function chosen depends on the conditional distribution of the target/training data.

Assuming that the model errors are Gaussian with mean zero and variance σ^2 which is independent of $\mathbf{x}$, one has as the conditional density of Y_i given $\mathbf{X}_i = \mathbf{x}$

$$f(\mathbf{Y}|\mathbf{X}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-(\mathbf{Y} - Z(\mathbf{X};\boldsymbol{\theta}))^2}{2\sigma^2} \right\} \tag{2.9}$$

so that the log-likelihood function is given by

$$L = \frac{-\sum_{i=1}^{n}(Y_i - Z(\mathbf{X}_i;\boldsymbol{\theta}))^2}{2\sigma^2} - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(2\sigma^2) \tag{2.10}$$

For maximization purposes, the second and third term on the right-hand side of equation (2.10) are independent of the weights $\boldsymbol{\theta}$ and can therefore be omitted so that maximizing equation (2.10) is equivalent to minimizing

$$S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - Z(\mathbf{X}_i;\boldsymbol{\theta}))^2 \tag{2.11}$$

The weights are then adjusted in such a way that the error function in equation (2.11) is minimized.

However, the error function in equation (2.11) is based on the assumption that the target data Y_i were generated from a smooth deterministic function with added Gaussian noise.

Our work deals with a classification problem and therefore the targets are binary variables. The Gaussian noise model does not provide a good description of binary variables.

The error function for binary variables with mean $Z(\mathbf{X};\boldsymbol{\theta})$ is got by first realizing that the probability weights of Y_i given $\mathbf{X}_i = \mathbf{x}$ are

$$\pi(Y|\mathbf{X}) = Z(\mathbf{X};\boldsymbol{\theta})^Y (1 - Z(\mathbf{X};\boldsymbol{\theta}))^{(1-Y)}, Y = 0, 1 \tag{2.12}$$

so that the likelihood function is given by

$$L = \Pi_{i=1}^{n} Z(\mathbf{X}_i;\boldsymbol{\theta})^{Y_i} (1 - Z(\mathbf{X}_i;\boldsymbol{\theta}))^{(1-Y_i)} \tag{2.13}$$

$\hat{\boldsymbol{\theta}}$ is then the value of $\boldsymbol{\theta}$ that maximizes the likelihood equation (2.13).

Since it is more convenient to minimize the negative of the log likelihood, one has

$$S(\mathbf{Y},\mathbf{X};\boldsymbol{\theta}) = -\sum_{i=1}^{n}\{Y_i \ln(Z(\mathbf{X}_i;\boldsymbol{\theta})) + (1-Y_i)\ln(1-Z(\mathbf{X}_i;\boldsymbol{\theta}))\} \tag{2.14}$$

In line with equations (2.11) and (2.14), the weights are adjusted in such a way that the error between the targets $\mathbf{Y}$ and the actual outputs, $Z(\mathbf{X};\boldsymbol{\theta})$, is minimized.

We now turn to the very important step of training the network. This step involves updating the weights until the error function $S(\mathbf{Y},\mathbf{X};\boldsymbol{\theta})$ is minimized. There are various methods of minimizing $S(\mathbf{Y},\mathbf{X};\boldsymbol{\theta})$. We discuss the major three.

(a)Backpropagation (BP)

Backpropagation is kind of a coordinate wise gradient descent method. Taking a unipolar $\psi(x)$, the weights are adjusted as follows:

$$\begin{aligned}\mathbf{W}^{r+1} &= \mathbf{W}^{r} + \Delta\mathbf{W} \\ \boldsymbol{\alpha}^{r+1} &= \boldsymbol{\alpha}^{r} + \Delta\boldsymbol{\alpha}\end{aligned} \tag{2.15}$$

Taking individual weights, we have the r^{th} iteration weights as

$$\begin{aligned}\alpha_h^{(r+1)} &= \alpha_h^{(r)} - \lambda_1\{\frac{\partial S(\mathbf{Y},\mathbf{X};\boldsymbol{\theta}^{(r)})}{\partial\alpha_h}\} \\ &\quad \text{for } i=1,\ldots,n \text{ and } h=1,\ldots,H\end{aligned} \tag{2.16}$$

Similarly,

$$\begin{aligned}W_{hj}^{(r+1)} &= W_{hj}^{(r)} - \lambda_2\{\frac{\partial S(\mathbf{Y},\mathbf{X};\boldsymbol{\theta}^{(r)})}{\partial W_{hj}}\} \\ &\quad \text{for } i=1,\ldots,n,\ h=1,\ldots,H \text{ and } j=0,\ldots,d\end{aligned} \tag{2.17}$$

λ_1 and λ_2 represent the step gain.

The weights are adjusted until the stopping criterion is met. Under this method, each weight is adjusted n times at each iteration. This means that for I iterations, each weight is adjusted In times. The method is therefore slow especially because I is normally large. The method is also not very stable and leads to asymptotically sub-efficient estimates (White, [69]).

(b)The Quasi-Newton method

This method was independently developed by the authors: Broyden [7],

Fletcher [22] and Goldfarb [27]. It is therefore commonly referred to as the BFGS method.

The training starts by first inputing an initial set of weights $\boldsymbol{\theta}^{(0)}$. From $\boldsymbol{\theta}^{(0)}$, $S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0)})$ is determined. From the principles of second-order Taylor expansion, $S(y_i, \mathbf{x}_i; \boldsymbol{\theta}^{(1)})$ can be found:

$$S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(1)}) = S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0)}) + A_0(\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}) + \frac{1}{2}(\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)})B_0(\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}) \tag{2.18}$$

where $\mathbf{A}_0$ is the first order derivative vector of $S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})$ and $\mathbf{B}_0$ is the Hessian matrix of $S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})$ both at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. For a neural net with a total of M parameters in $\boldsymbol{\theta}$, both $\mathbf{A}_0$ and $\mathbf{B}_0$ are evaluated numerically as

$$\begin{aligned}
\mathbf{A}_0 \;=\; & \frac{S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)} + \iota_1, \ldots, \boldsymbol{\theta}^{(0,M)}) - S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,M)})}{\iota_1} \\
& \vdots \\
& \frac{S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,m)} + \iota_m, \ldots, \boldsymbol{\theta}^{(0,M)}) - S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,M)})}{\iota_m} \\
& \vdots \\
& \frac{S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,M)} + \iota_M) - S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,M)})}{\iota_M}
\end{aligned} \tag{2.19}$$

where $\iota_m = \max(\epsilon, \epsilon\boldsymbol{\theta}^{(0,m)})$ with e.g. $\epsilon = 10^{-6}$ for $m = 1, \ldots, M$.

The direct off-diagonal elements of the matrix $\mathbf{B}_0$ are evaluated as:

$$\begin{aligned}
\frac{\partial^2 S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0)})}{\partial \boldsymbol{\theta}^{(0,j)} \partial \boldsymbol{\theta}^{(0,k)}} \;=\; & \frac{1}{\iota_j \iota_k} \Big\{ S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,j)} + \iota_j, \ldots, \boldsymbol{\theta}^{(0,k)} + \iota_k, \ldots, \boldsymbol{\theta}^{(0,M)}) \\
& - S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,j)}, \ldots, \boldsymbol{\theta}^{(0,k)} + \iota_k, \ldots, \boldsymbol{\theta}^{(0,M)}) \\
& - S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,j)} + \iota_j, \ldots, \boldsymbol{\theta}^{(0,k)}, \ldots, \boldsymbol{\theta}^{(0,M)}) \\
& - S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,M)}) \Big\} \\
& \text{for } j, k = 1, \ldots, M
\end{aligned} \tag{2.20}$$

The direct diagonal elements of $\mathbf{B}_0$ are evaluated as:

$$\begin{aligned}
\frac{\partial^2 S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0)})}{\partial (\boldsymbol{\theta}^{(0,j)})^2} \;=\; & \frac{1}{\iota_j^2} \Big\{ S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,j)} - \iota_j, \ldots, \boldsymbol{\theta}^{(0,M)}) \\
& - 2S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,M)}) \\
& + S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0,1)}, \ldots, \boldsymbol{\theta}^{(0,j)} - \iota_j, \ldots, \boldsymbol{\theta}^{(0,M)}) \Big\} \\
& \text{for } j = 1, \ldots, M.
\end{aligned} \tag{2.21}$$

$\boldsymbol{\theta}^{(1)}$ is obtained by differentiating the right-hand side of equation (2.18) with respect to $\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}$ and equating the result to zero to get:

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} - \mathbf{B}_0^{-1}\mathbf{A}_0 \tag{2.22}$$

The quantity $\mathbf{B}_0^{-1}\mathbf{A}_0$ is called the direction of a change. It is a vector describing a segment of a path from iteration 0 to iteration 1. $\mathbf{B}_0$ represents the "angle" of the direction while $\mathbf{A}_0$ represents the "size" of the direction.

The minimization then continues from iteration 1, to 2,...until the stopping criterion is met. In general, the r^{th} iteration weights are given by:

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} - \mathbf{B}_r^{-1}\mathbf{A}_r \tag{2.23}$$

This method is very accurate and fast. However, the Hessian matrix $\mathbf{B}_r$ may become singular. This means that $\mathbf{B}_r^{-1}$ would be undefined. The BFGS method solves this problem by numerically approximating $\mathbf{B}_r$.

The method first re-defines equation (2.23) as follows:

$$\boldsymbol{\theta}^{(r+1)} = \omega_r\boldsymbol{\theta}^{(r)} - \mathbf{B}_r^{-1}\mathbf{A}_r \tag{2.24}$$

where ω_r is called the step length and it is found such that $S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(r)} - \omega_r\mathbf{B}_r^{-1}\mathbf{A}_r) < S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(r)})$. This is done using various methods like step halving and golden section search. In step halving, ω_r is first set at 1 and the function $S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(r)} - \omega_r\mathbf{B}_r^{-1}\mathbf{A}_r)$ tested for a decrease. If it fails, ω_r is decreased by 1/2 and the test carried out again. This process continues until a decrease in the function occurs. The final value of ω_r is the required step length.

By letting

$$\mathbf{a}_r = \boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)} = -\omega_r\mathbf{B}_r^{-1}\mathbf{A}_r \tag{2.25}$$

represent the change in parameters in the r^{th} iteration and

$$\mathbf{b}_r = \mathbf{A}_{r+1} - \mathbf{A}_r \tag{2.26}$$

represent the change in gradients in the current r^{th} iteration, the BFGS method has:

$$\mathbf{B}_{r+1}\mathbf{a}_r = \mathbf{b}_r \tag{2.27}$$

Therefore, $\mathbf{B}_{r+1}$ is the ratio of the change in the gradient to the change in the parameters. This is what is called the Quasi-Newton condition.

The BFGS method solves equation (2.27) for $\mathbf{B}_{r+1}$ as:

$$\mathbf{B}_{r+1} = \mathbf{B}_r + \frac{\mathbf{b}_r \mathbf{b}_r^t}{\mathbf{b}_r^t \mathbf{a}_r} - \frac{\mathbf{B}_r \mathbf{a}_r \mathbf{a}_r^t \mathbf{B}_r}{\mathbf{a}_r^t \mathbf{B}_r \mathbf{a}_r} \tag{2.28}$$

where $\mathbf{c}^t$ represents the transpose of vector $\mathbf{c}$.
The BFGS update matrix $\mathbf{B}_r$ remains positive definite as long as $\mathbf{b}_r^t \mathbf{a}_r > 0$ and holds automatically since $S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})$ is strictly convex.

Having defined all the necessary equations, we now summarize the BFGS algorithm:

1. Input the initial weights $\boldsymbol{\theta}^{(0)}$ and $\mathbf{B}_0$, an identity matrix whose size is equal to the length of vector $\boldsymbol{\theta}^{(0)}$.
2. Set $\boldsymbol{\sigma}_r = -\mathbf{B}_r^{-1} \mathbf{A}_r$.
3. Compute the step length ω_r and determine $\boldsymbol{\theta}^{(r+1)}$ as $\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r+1)} + \omega_r \boldsymbol{\sigma}_r$
4. Compute the values $\mathbf{a}_r = \boldsymbol{\theta}^{r+1} - \boldsymbol{\theta}^{(r)}$ and $\mathbf{b}_r = \mathbf{A}^{r+1} - \mathbf{A}^r$
5. Compute $\mathbf{B}_{r+1}$ as $\mathbf{B}_{r+1} = \mathbf{B}_r + \frac{\mathbf{b}_r \mathbf{b}_r^t}{\mathbf{b}_r^t \mathbf{a}_r} - \frac{\mathbf{B}_r \mathbf{a}_r \mathbf{a}_r^t \mathbf{B}_r}{\mathbf{a}_r^t \mathbf{B}_r \mathbf{a}_r}$.
6. Continue with the next r until termination criterion are satisfied.

Later in this work, we apply the BFGS method to minimize our functions.

(c)The Simulated Annealing method

This method differs from Quasi-Newton method in that it does not consider the first- or second-order derivatives. The optimization is through a stochastic search method. Simulated annealing method originated from statistical mechanics. The method is summarized below:

1. Initialize $\boldsymbol{\theta}^{(0)}$ and hence determine $S(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}^{(0)})$.
2. Compute the r^{th} iteration temperature as $T(r) = \frac{\overline{T}}{1+ln(r)}$
3. Perturbate the solution vector randomly to obtain the r^{th} solution vector $\hat{\boldsymbol{\theta}}^{(r)}$ and hence determine $S(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}}^{(r)})$.
4. Generate probability $P(r)$ from the uniform distribution; $0 \leq P(r) \leq 1$.
5. Compute the Metropolis ratio $M(r)$ as $M(r) = exp\{\frac{-[S(\mathbf{Y},\mathbf{X};\hat{\boldsymbol{\theta}}^{(r)}) - S(\mathbf{Y},\mathbf{X};\hat{\boldsymbol{\theta}}^{(r-1)})]}{T(r)}\}$

6. If $S(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}}^{(r)}) - S(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}}^{(r-1)}) < 0$ then $\boldsymbol{\theta}^{(r)} = \hat{\boldsymbol{\theta}}^{(r)}$
   else
   if $P(r) \leq M(r)$ then $\boldsymbol{\theta}^{(r)} = \hat{\boldsymbol{\theta}}^{(r)}$.

7. Repeat steps 2-6 till $r = \overline{T}$ where $r = 1, \ldots, \overline{T}$.

$\overline{T}$ should be large enough to make sure that all proposed transitions are accepted by the algorithm.

We conclude this section by discussing two important issues:

Number of Layers

A neural net can have more than one hidden layer, see Looney [48] for more details. However, it is shown in White [69] that one hidden layer with sufficient number of neurons is enough to approximate any function of interest. In practice, however, a network with more than one layer may provide a more parsimonious model for the data.

The number of neurodes, H, in the hidden layer can be determined as in Looney [48] by a rule of thumb:

$$H = \|1.7 * \log_2(n)\| + 1 \tag{2.29}$$

Alternatively, one can use the Black Information Criterion (BIC) as proposed in Swanson and White [64] to sequentially determine H. We discuss the BIC procedure later in this work.

The Stopping Rule

There are four common stopping rules:

1. $\|\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}\| < \epsilon$, for $\epsilon > 0$.
2. $\left|S(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}}^{(r+1)}) - S(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}}^{(r)})\right| < \epsilon$ for $\epsilon > 0$ but small.
3. $S(\mathbf{Y}, \mathbf{X}; \hat{\boldsymbol{\theta}}^{(r)}) < E_{min}$, a pre-specified lower bound for the training error
4. $r > MAX$, where MAX is the pre-specified number of iterations.

We note that Rule(4) can be used together with Rule(1), Rule(2) or Rule(3).

2.1.2 Logistic regression

In this section, we give the motivation for the use of logistic regression for modeling binary response data.

Logistic regression is a sub-category of generalized linear models. This broad category includes simple linear regression, log-linear regression, etc. Logistic regression is used to model a discrete outcome (e.g good/bad credit risk, dead/alive, presence/absence or success/failure), depending on a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these.

Logistic regression can be used to model probabilities of group membership as in Joanes [45] among many others. Discriminant analysis is also used to predict group membership with only two groups. The motivation for using logistic regression emanates from the fact that whereas discriminant analysis can only be used with continuous covariates, logistic regression puts no condition on the distribution of the independent or dependent variables. In logistic regression, both the independent and dependent variables can take any form. The variables do not have to be normally distributed, linearly related or of equal variance within each group.

The Logistic Function

This function has been discussed by many researchers like Müller *et al* [34]. The logistic function is given as:

$$\begin{aligned} f(g) &= \frac{exp(g)}{1+exp(g)} \\ &= \frac{1}{1+exp(-g)} \end{aligned} \tag{2.30}$$

When modelling a Bernoulli response variable with multiple covariates, one directly models the probabilities of group membership, Joanes [45], as follows ;

$$P(Y=1|\mathbf{x}) = \frac{1}{1+exp(-(\beta_0+\sum_{j=1}^{d}\beta_j x_j))} \tag{2.31}$$

where now g in equation (2.30) is given by;

$$g = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d \tag{2.32}$$

To illustrate the applicability of the logistic function, we present in figure (2.1) the logistic function of data on age of an individual and whether or not there were signs of Coronary heart disease.

The following are some properties of the logistic function:

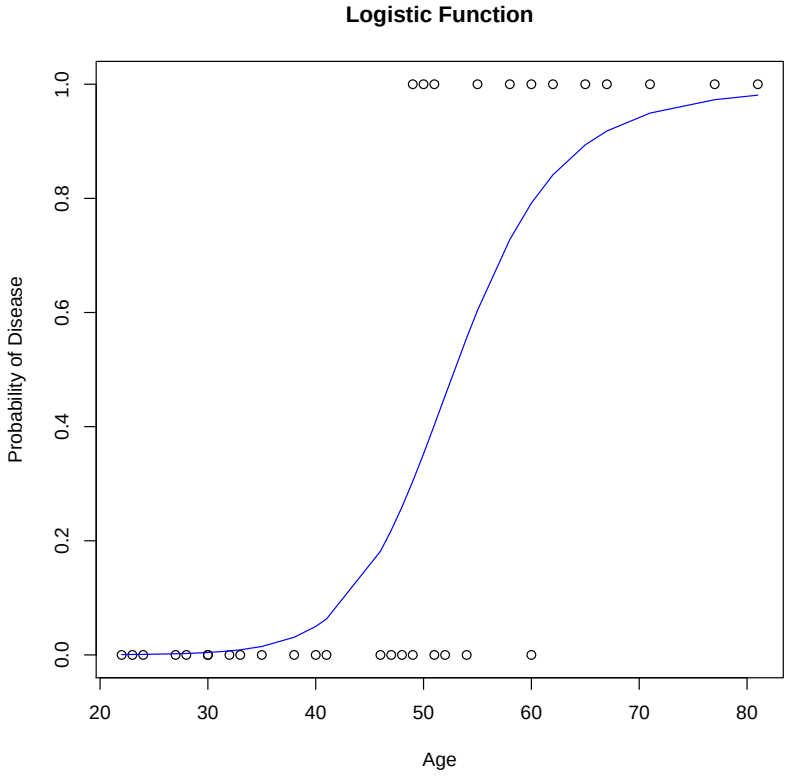

Figure 2.1: *The Logistic Function: The Blue line represents the effect of age on the risk of coronary heart disease.*

1. It is bounded between zero and one. This property eliminates the possibility of getting estimated/predicted probabilities outside this range which would not make sense. In linear regression, it is possible to get predicted values outside this range which is statistically inadmissible.

2. With a proper transformation, one can get a linear model from the logistic function. Fan *et al* [17] use the logit function to transform the logistic mean function of a Bernoulli distributed response variable.

Transforming equation (2.31) as in Fan *et al* [17], we have

$$\begin{aligned}
\text{logit}(P(Y=1|\mathbf{x})) &= \log_e[\frac{P(Y=1|\mathbf{x})}{1-P(Y=1|\mathbf{x})}] \\
&= \log_e\{\frac{1+exp((\beta_0+\sum_{j=1}^{d}\beta_j x_j))}{1+exp(-(\beta_0+\sum_{j=1}^{d}\beta_j x_j))}\} \\
&= \log_e\{exp((\beta_0+\sum_{j=1}^{d}\beta_j x_j))\} \\
&= \beta_0+\sum_{j=1}^{d}\beta_j x_j \qquad (2.33)
\end{aligned}$$

2.1.3 Fitting the Logistic Regression Model

As pointed out in Fan *et al* [17], estimating the function $P(Y=1|\mathbf{x})$ in equation (2.31) is equivalent to estimating the function $g(\mathbf{X};\boldsymbol{\beta})=\beta_0+\beta_1 x_1+\ldots+\beta_d x_d$ in equation (2.32).

There exist two approaches of estimating $g(\mathbf{X};\boldsymbol{\beta})$ in the literature: Parametric and non-parametric approaches.
Parametric estimation of $g(\mathbf{X};\boldsymbol{\beta})$ can be found in Joanes [45], Pastor and Gualler [56] and Pastor-Barriuso *et al* [57] among others. This approach uses the MLE method. As pointed out by the above authors, one first defines the likelihood function. For a Bernoulli case we have

$$L(\mathbf{Y},\mathbf{X};\boldsymbol{\beta})=\Pi_{i=1}^{n}[P(Y_i=1|\mathbf{X}_i=\mathbf{x})]^{Y_i}[1-P(Y_i=1|\mathbf{X}_i=\mathbf{x})]^{1-Y_i} \quad (2.34)$$

so that after taking logs and upon simplifying, one has

$$L^*(\mathbf{Y},\mathbf{X};\boldsymbol{\beta})=\sum_{i=1}^{n}\{Y_i g(\mathbf{X}_i;\boldsymbol{\beta})-\ln(1+exp(g(\mathbf{X}_i;\boldsymbol{\beta})))\} \quad (2.35)$$

Clearly, $L^*(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta})$ depends entirely on the unknown parameters $\boldsymbol{\beta} = \beta_0, \beta_1, \ldots, \beta_d$. Then, the MLE's $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ are the values that minimize the likelihood function given in equation (2.35). In essence, one has $d+1$ likelihood equations obtained by differentiating $L^*(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta})$ with respect to each of $\beta_0, \beta_1, \ldots, \beta_d$.

However, in many cases, a parametric form of $g(\mathbf{X}; \boldsymbol{\beta})$ is not known by the modeler. In such cases, misspecification of the parametric model may lead to serious errors in the subsequent data analysis.

It is for this reason the recent work has concentrated on nonparametric estimation of $g(\mathbf{X}; \boldsymbol{\beta})$. Fan *et al* [17] use Kernel methods to estimate $g(\mathbf{X}; \boldsymbol{\beta}) = \ln(P(Y = 1|\mathbf{x}))$. In this work, the local maximum likelihood estimation (LMLE)technique is introduced. The LMLE technique is the nonparametric counterpart of the parametric MLE.

Another nonparametric method that can be used to estimate $g(\mathbf{X}; \boldsymbol{\beta})$ is the polynomial method. This method is used by Vexler and Gurevich [66]. The method emanated from Weierstrass approximation theorem which states that every continuous function defined on a set [a,b] can be uniformly approximated as closely as desired by a polynomial function. The accuracy increases with the increase in the polynomial power/degree.

To illustrate the polynomial method, assume that $\mathbf{x} = (x_1, x_1)$. Then, $g(\mathbf{X})$ can be approximated polynomially as;

$$g(\mathbf{X}) \approx g(\mathbf{X}; \boldsymbol{\beta}) = \sum_{i=0}^{D} \beta_{1i} x_1^i + \sum_{i=1}^{D} \beta_{2i} x_2^i + \sum_{i=2}^{D} \sum_{j=2}^{i} \rho_{i-j+1\ j-1} x_1^{i-j+1} x_2^{j-1} \tag{2.36}$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\rho}$ are vectors of parameters to be estimated. D represents the polynomial degree.
The polynomial parameters can be estimated using the LMLE method as in Fan *et al* [17] and Vexler and Gurevich [66].

As noted in McNelis ([51],pp 18), the number of parameters in polynomial estimation rises exponentially with the degree of expansion and the dimension of $\mathbf{x}$. Moreover, polynomials of high degree compared to the sample size typically tend to fluctuate if fitted to data.

ANN provide an alternative which frequently provides a parsimonious approximation to the underlying nonlinear function. For sigmoid transfer function, the reason may be that functions to be approximated in practice are locally linear which is true for a sigmoid function around its center of symmetry. Combining several sigmoid functions with different centers of symmentry results in a superposition of locally linear functions.

It is stated in McNelis ([51],pp 20) that ANN performs well in forecasting data generated by unknown and highly nonlinear processes. There, it is also concluded that ANN approximates $g(\mathbf{X})$ as accurately as the polynomial method and with fewer parameters.
When the two methods have the same number of parameters, McNelis ([51],pp 20) concludes that ANN performs better. It is for this reason that later in this work we use ANN to estimate a given function.

Both the polynomial and ANN methods are special cases of general sieve estimates. There, the functional form of the estimator is given up to finitely many parameters, but the number of parameters is determined by the polynomial degree or the number of neurons. This is because in both methods, the functional forms are given but the degree of the polynomial or the number of neurons are not. Therefore, the parameters are neither limited in number nor do they have a straightforward interpretation as the parameters do in linear models.

There is a relationship between logistic regression and ANN, McNelis([51],pp 52). Using the unipolar activation function defined in (2.1.1), the logistic function represents an ANN with one hidden neuron.

As in Franke *et al* ([23],pp 390) and McNelis([51],pp 52), the predicting probability $Z(\mathbf{x};\boldsymbol{\theta}) = P(Y_i = 1|\mathbf{X}_i = \mathbf{x})$ for a network with H hidden neurons and $j+1$ input nodes can be represented as

$$\begin{aligned} Z(\mathbf{x};\boldsymbol{\theta}) &= \psi(\alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j)) \\ &= \frac{1}{1 + exp\{-(\alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j))\}} \end{aligned} \tag{2.37}$$

where now $g(\mathbf{x};\boldsymbol{\theta}) = \alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j)$

We note that the predicting probability $Z(\mathbf{x};\boldsymbol{\theta})$ can also be obtained by using the bipolar activation function

$$\begin{aligned} Z^*(\mathbf{x};\boldsymbol{\theta}) &= \frac{1 - exp\{-(\alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j))\}}{1 + exp\{(\alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j))\}} \\ &= \tanh(\alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j)) \end{aligned} \tag{2.38}$$

where now ψ is a hyperbolic tangent activation function as described in equation (2.7). This implies that $Z^*(\mathbf{x};\boldsymbol{\theta}) \in (-1, 1)$. To transform $Z^*(\mathbf{x};\boldsymbol{\theta})$

into the range (0,1), we use the transformation function:

$$Z(\mathbf{x};\boldsymbol{\theta}) = \frac{Z^*(\mathbf{x};\boldsymbol{\theta}) + 1}{2} \tag{2.39}$$

The network parameters are estimated using any of the methods discussed in section (2.1.1).
The ANN method is very appealing but the parameter vector $\boldsymbol{\theta}$ is unidentifiable, see Ding and Hwang [41] for more details. However, it can be shown that $\boldsymbol{\theta}$ is identifiable up to a given family of transformations. We thoroughly deal with this problem in chapter 3.

2.2 Change Point Detection

Testing for possible structural changes in a model has become one of the principal objectives of econometric analysis. This is due to the fact that if there is a change in the generating process, then there will be an induced structural instability in the original model.

Definition Let $\mathbf{Y} = Y_1, \ldots, Y_n$ be a sequence of independently distributed random variables. Cobb [9] defines a change point as the point $K \in (1, n)$ at which the data generating mechanism of $\mathbf{Y}$ changes.
This implies that when there is one change point at K, $\mathbf{Y}$ can be segmented into two parts with different densities as follows.

$Y_1, \ldots, Y_K$ are i.i.d. with density $f(Y;\boldsymbol{\theta}_0)$ and $Y_{K+1}, \ldots, Y_n$ are i.i.d. with density $f(Y;\boldsymbol{\theta}_1)$. Analogously, for discrete random variables, $f(Y;\boldsymbol{\theta}_i), i = 0, 1$ would be probability weights characterizing the distribution.

Frequently, it is assumed that the form of the density remains unchanged, and a change in parameter only influences the mean of the data. Thus, the change point problem can be easily extended to cover regression models. Feder [21], for example, considers a setting where

$$Y_i = Z(X_i;\boldsymbol{\theta}) + \epsilon_i \quad , \quad \text{for } 1 \leq i \leq n \tag{2.40}$$

where ϵ_i's are the zero mean model errors and the function $Z(\mathbf{X}_i;\boldsymbol{\theta}) = E[Y_i = 1|\mathbf{X}_i = \mathbf{x}]$ is known only up to the parameter $\boldsymbol{\theta}$.

The problem of change in the mean of variables has attracted a lot of research. See for example Worsley [70], Yao and Davis [71], Gombay and Horvath [28], Feder ([21],[20]) and Hinkley [36] among many others.

Change point detection entails testing the null hypothesis H_0 against the alternative hypothesis H_1. H_0 postulates that the model distribution does not change throughout the whole period. H_1 postulates that the model

distribution remains unchanged as in H_0 up to a certain unknown time point when the model distribution changes.
The objective of change-point detection is to test for a possible change-point in a given set of random variables.

Assuming that the change point K is unknown and taking $\mu(\mathbf{X}_i)$ as the mean function, the hypotheses can be represented as follows for model (2.40).

$$\begin{aligned} H_0 \quad &: \quad \mu(\mathbf{X}_i) = Z(\mathbf{X}_i; \boldsymbol{\theta}_0), i = 1, \ldots, n \\ &\textit{versus} \\ H_1 \quad &: \quad \exists K \in 2, \ldots, n-1 \text{ such that} \\ &\qquad \mu(\mathbf{X}_i) = Z(\mathbf{X}_i; \boldsymbol{\theta}_0), i = 1, \ldots, K \\ &\qquad \mu(\mathbf{X}_i) = Z(\mathbf{X}_i; \boldsymbol{\theta}_1), i = K+1, \ldots, n \end{aligned} \tag{2.41}$$

There exist two testing procedures for a possible change point namely retrospective testing and sequential testing. We now discuss each of them.

2.2.1 Sequential testing

This procedure is also called on-line testing. It is also referred to as prospective testing.

Suppose that there is a process of independent observations $Y_1, Y_2, \ldots$. The process is considered to be 'in control' up to an unknown time K with the distribution of $Y_1, \ldots, Y_K$ given by $f(Y; \boldsymbol{\theta}_0)$. At the unknown time K, the process goes 'out of control' so that the distribution of $Y_{K+1}, Y_{K+2} \ldots$ is given by $f(Y; \boldsymbol{\theta}_1)$.

The aim of sequential testing is to raise an alarm as soon as the process is out of control so that an appropriate action can be taken. Sequential testing has three quantities of interest namely;

1. False alarm rate, $\alpha = P(\text{Accept } H_1 | H_0 \text{ is true})$
2. Mis-detection rate, $\beta = P(\text{Accept } H_0 | H_1 \text{ is true})$
3. Expected stopping time also called the decision delay time given as $E[N]$ where N is the number of samples up to the change point until the change is detected.

The frequentist and Bayesian formulation methods are used to balance the trade off between the above three quantities effectively. In the frequentist method, α and β are fixed appropriately. One then minimizes $E[N]$ with respect to $f(Y; \boldsymbol{\theta}_0)$ and $f(Y; \boldsymbol{\theta}_1)$. Page's CUSUM method is an example of

the frequentist formulation.

In the Bayesian formulation, one minimizes the following weighted expression

$$\begin{aligned} &\text{Minimize } C_1\alpha + C_2\beta + C_3 E[N] \\ &\quad \text{for some weights } C_1, C_2, C_3 \end{aligned} \tag{2.42}$$

The procedure has been researched by many authors like Antoch *et al* [2] and Berkes *et al* [5] among others.

The procedure is done on-line as new data become available and the goal is to stop the process immediately once a change point is detected.

Applications of sequential testing include statistical quality control for detection of changes in quality operations, monitoring the number of cases of a disease for potential outbreak, global warming and detection of business cycles. This list is not exhaustive by far.

2.2.2 Retrospective Testing

Testing of the change point under various forms of $Z(\mathbf{x};\boldsymbol{\theta})$, the mean function, has attracted a lot of research.

Hinkley [36] , Hinkley and Hinkley [37] and many others studied the case when $Z(\mathbf{x};\boldsymbol{\theta})$ assumes a constant value say μ_0 before a change and after the change it assumes another constant say μ_1. The hypotheses under this case are of the form:

$$\begin{aligned} H_0 \quad &: \quad Y_i = \mu_0 + \epsilon_i \ , i = 1, \ldots, n \\ &\textit{versus} \\ H_1 \quad &: \quad \exists K \in 2, \ldots, n-1 \text{ such that} \\ &\qquad Y_i = \mu_0 + \epsilon_i \ , i = 1, \ldots, K \\ &\qquad Y_i = \mu_1 + \epsilon_i \ , i = K+1, \ldots, n \end{aligned} \tag{2.43}$$

where μ_0, μ_1 and K are unknown.

Various procedures like R-test procedures, CUSUM test procedures and M-test procedures can be used to test the hypotheses in equation (2.43) above. We discuss the M-test and CUSUM procedures.

Method	$\phi(x)$	
Huber	x	$\lvert x\rvert \leq A$
	A sign(x)	$\lvert x\rvert > A$
Welsh	$x\ \exp\{-(\frac{x}{B})^2\}$	$\lvert x\rvert \in \Re_1$
Fair	$\frac{x}{1+\lvert x\rvert/C}$	$\lvert x\rvert \in \Re_1$
Tukey	$x(1-(\frac{x}{D})^2)^2$	$\lvert x\rvert \leq D$
	0	$\lvert x\rvert > D$

Table 2.1:

M-test Procedures

For the theory of M-procedures, we refer to Huber [40] . This method is based on the partial sums defined as

$$S_K(\phi) = \sum_{i=1}^{K} \phi(Y_i - \hat{\mu}_n(\phi)) \ , K = 1, \ldots, n-1 \tag{2.44}$$

where $\phi(x)$ is a monotone and skew symmetric function satisfying $\phi(-x) = -\phi(x)\ \forall x \in \Re$. $\hat{\mu}_n(\phi)$ is the M-estimator of the mean and it is any solution of the equation

$$\sum_{i=1}^{n} \phi(Y_i - a) = 0 \tag{2.45}$$

Then, the maximum M-test statistic is defined as

$$Q_n = \max_{1 \leq K \leq n-1} \left\{ \frac{1}{\hat{\sigma}_n(\phi)} \left| \sqrt{\frac{n}{K(n-K)}} \sum_{i=1}^{K} \phi(Y_i - \hat{\mu}_n(\phi)) \right| \right\} \tag{2.46}$$

where

$$\hat{\sigma}_n^2(\phi) = \min_{1 \leq K \leq n-1} \left\{ \frac{1}{n} \left(\sum_{i=1}^{K} \phi^2(Y_i - \hat{\mu}_K(\phi)) + \sum_{i=K+1}^{n} \phi^2(Y_i - \hat{\mu}_{n-K}(\phi)) \right) \right\} \tag{2.47}$$

$\hat{\mu}_K(\phi)$ is calculated from $Y_1, \ldots, Y_K$ while $\hat{\mu}_{n-K}(\phi)$ is calculated from $Y_{K+1}, \ldots, Y_n$.

Table (2.1) shows various types of the score function $\phi(x)$, see Huber [40] and Antoch and Visek [3] for more details.

The Huber function is mostly used in practice. We note that, by letting $A \to \infty$ in the Huber function , one has $\phi(x) = x$ which in effect makes the

M-test statistics a classical least squares statistics. Under this case, equation (2.46) reduces to

$$Q_n = \max_{1\leq K\leq n-1}\left\{\frac{1}{\hat{\sigma}_n(\phi)}\left|\sqrt{\frac{n}{K(n-K)}}\sum_{i=1}^{K}(Y_i-\hat{\mu}_n(\phi))\right|\right\} \tag{2.48}$$

where now

$$\hat{\sigma}_n^2(\phi) = \min_{1\leq K\leq n-1}\left\{\frac{1}{n}\left(\sum_{i=1}^{K}(Y_i-\hat{\mu}_K(\phi))^2+\sum_{i=K+1}^{n}(Y_i-\hat{\mu}_{n-K}(\phi))^2\right)\right\} \tag{2.49}$$

When $A\to 0$, the M-test statistic reduces to the L_1-norm test statistic. Lastly, when $\phi(x) = sign(x), x\in\Re$, the procedure reduces to the L_1 test procedure where now the test statistic is given by

$$Q_n = \max_{1\leq K\leq n-1}\left\{\sqrt{\frac{n}{K(n-K)}}\left|\sum_{i=1}^{K}\text{sign}(Y_i-\tilde{M}_n)\right|\right\} \tag{2.50}$$

where $\tilde{M}_n$ is the sample median based on all the observations.

CUSUM Procedures

This method has many versions, see Page ([53],[54]) for example. We briefly describe the CUSUM procedure in Taylor [65]. The test is based on the CUSUM statistic;

$$Q_n = S_i \tag{2.51}$$

where $S_i = S_{i-1}+(Y_i-\bar{Y})\, i=1,\ldots,n$, $S_0=0$ and ,$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$

An upward slope of the CUSUM chart indicates a period where the observed values tend to be above $\bar{Y}$. Similarly, a downward slope of the CUSUM indicates a period where the observed values tend to be below $\bar{Y}$. A change point is detected when there is a sudden change in direction of the CUSUM.

We note that many times the mean function is often conditional, meaning that the mean of the response variable depends on the covariates. More research has been done along this line, see for example Pastor and Guallar [56].

A simple linear regression case is one where Y depends on only one covariate and where the hypotheses are of the form

$$\begin{aligned} H_0 \quad &: \quad Y_i = a_0+b_0x_i+\epsilon_i\ , i=1,\ldots,n \\ &\textit{versus} \\ H_1 \quad &: \quad \exists K\in 2,\ldots,n-1 \text{ such that} \\ &\qquad Y_i = a_0+b_0x_i+\epsilon_i\ , i=1,\ldots,K \\ &\qquad Y_i = a_1+b_1x_i+\epsilon_i\ , i=K+1,\ldots,n \end{aligned} \tag{2.52}$$

where $(a_0, b_0) \neq (a_1, b_1)$

The testing can also be done through the M-test procedures where the M-residuals are estimated as

$$\hat{\epsilon}_i = \phi\left(Y_i - \hat{a}_{0n} - \hat{b}_{0n}x_i\right) \ , i = 1, \ldots, n \tag{2.53}$$

where $\hat{a}_{0n}$ and $\hat{b}_{0n}$ are the M-estimators estimated from all the n observations by solving the following simultaneous equations

$$\sum_{i=1}^{n} \phi(Y_i - a_0 - b_0 x_i) = 0 \text{ and } \sum_{i=1}^{n} x_i \phi(Y_i - a_0 - b_0 x_i) = 0 \tag{2.54}$$

Then, the M-test statistic is given by

$$Q_n = \max_{1 \leq K \leq n-1} \Lambda_K \tag{2.55}$$

with

$$\begin{aligned} \Lambda_K = \quad & \frac{1}{\hat{\sigma}^2_{n,M}} \left\{ \frac{n}{K(n-K)} \left(\sum_{i=1}^{K} \hat{\epsilon}_i \right)^2 + \right. \\ & \left. \frac{\left[\sum_{i=1}^{n} (x_i - \hat{x}_n)^2\right] . \left[\sum_{i=1}^{K} (x_i - \hat{x}_n)^2 \hat{\epsilon}_i\right]^2}{\left[\sum_{i=1}^{K} (x_i - \hat{x}_n)^2\right] . \left[\sum_{i=K+1}^{n} (x_i - \hat{x}_n)^2\right]} \right\} \end{aligned} \tag{2.56}$$

where

$$\begin{aligned} \hat{\sigma}^2_{n,M} = \quad & \min_{1 \leq K \leq n-1} \frac{1}{n} \left\{ \sum_{i=1}^{K} \phi^2 \left(Y_i - \hat{a}_n - \hat{b}_n x_i \right) + \right. \\ & \left. + \sum_{i=K+1}^{n} \phi^2 \left(Y_i - \hat{a}^*_n - \hat{b}^*_n x_i \right) \right\} \end{aligned} \tag{2.57}$$

$\hat{a}^*_n$ and $\hat{b}^*_n$ are the M-estimators of a_1 and b_1 based on $Y_{K+1}, \ldots, Y_n$. However, many times the dependent variable Y depends on various independent variables.

Recent papers, see for example Feder [21], Zhan *et al* [73], Pastor and Gurevich [56], Pastor-Barriuso *et al* [57] and Vexler and Gurevich [66], have dealt with the case when $Z(\mathbf{x}; \boldsymbol{\theta})$ is considered to be polynomial.

Zhan *et al* [73] propose the following regression model when a change point K exists:

$$Y_i = \pi(\mathbf{X}_i) + \epsilon_i \quad , \quad \text{for } 1 \leq i \leq n \tag{2.58}$$

with ϵ_i's i.i.d. with mean 0 and finite variance and $\boldsymbol{\theta}$ is a vector of unknown parameters. The polynomial function $\pi(\mathbf{X}_i)$ has the following form

$$\pi(\mathbf{X}_i) = \begin{cases} Z(\mathbf{X}_i; \boldsymbol{\theta}_0) & \text{for } i = 1, \ldots, K \\ Z(\mathbf{X}_i; \boldsymbol{\theta}_1) & \text{for } i = K+1, \ldots, n \end{cases} \tag{2.59}$$

The change point testing hypotheses under this case are of the form:

$$\begin{aligned} H_0 \quad &: \quad Y_i = Z(\mathbf{X}_i; \boldsymbol{\theta}_0) + \epsilon_i \ , i = 1, \ldots, n \\ &\textit{versus} \\ H_1 \quad &: \quad \exists K \in 2, \ldots, n-1 \text{ such that} \\ &\qquad Y_i = Z(\mathbf{X}_i; \boldsymbol{\theta}_0) + \epsilon_i \ , i = 1, \ldots, K \\ &\qquad Y_i = Z(\mathbf{X}_i; \boldsymbol{\theta}_1) + \epsilon_i \ , i = K+1, \ldots, n \end{aligned} \tag{2.60}$$

In the general, not necessarily regression ,case, when the distribution of Y_i is known up to the parameter $\boldsymbol{\theta}$, one can use the likelihood ratio test statistic to test for a change, see for example Gombay and Horvath [28]. By letting the density, probability weights in the discrete case, of Y_i under H_0 be denoted by $f_i(Y_i; \boldsymbol{\theta}_0)$ and that under H_1 be denoted by $f_i(Y_i; \boldsymbol{\theta}_1)$ and assuming that $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$ and K are unknown, then the log likelihood ratio statistic is given by

$$\Lambda_K = \arg_{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \boldsymbol{\Theta}} \max_{1 \leq K \leq n-1} \log \left[\frac{\Pi_{i=1}^{K} f_i(Y_i; \boldsymbol{\theta}_0) \Pi_{i=K+1}^{n} f_i(Y_i; \boldsymbol{\theta}_1)}{\Pi_{i=1}^{n} f_i(Y_i; \boldsymbol{\theta}_0)} \right] \tag{2.61}$$

so that the test statistic is given by

$$Q_n = \max_{1 \leq K \leq n-1} (2\Lambda_K) \tag{2.62}$$

In the regression setting (equation (2.60)), we would, for example, have

$$f_i(Y_i; \boldsymbol{\theta}) = g(Y_i - Z(\mathbf{X}_i; \boldsymbol{\theta})) \tag{2.63}$$

where $g(.)$ is the density of the ϵ_i.

In all these tests, the null hypothesis H_0 is rejected when $Q_n > C_\alpha$, where C_α is the α-level critical value.

2.3 Rejection Criteria

The critical values of a given test statistic for rejecting H_0 can be derived using asymptotic theory or carrying out simulations on Q_n under H_0.

2.3.1 Asymptotic Method

In the asymptotic method, the asymptotic distribution of the test statistic under H_0 is derived and from it the critical values determined. This approach has been taken by many authors like Gombay and Horvath ([28], [30],[29]), Horvath [38] and Worsley [70].

Gombay and Horvath ([28], [30]) applied the likelihood ratio test to detect a change point. In these papers, they determined the asymptotic distribution of Q_n in equation (2.62) under the null hypothesis. In this work, the distribution of Y_i is assumed known but general in nature.

Horvath [38] investigated the limit distribution of the likelihood ratio test and cumulative sum (CUSUM) test for a change in binomial probability. In this work, he shows that the asymptotic distribution of the likelihood ratio test is the double exponential distribution. The test is shown to be very powerful at the tails.

Similar work to that of Horvath [38] has been done by Worsley [70]. However, due to the recursive nature of the method, it is computationally difficult and time consuming especially if the sample is large. Horvath [38] and Worsley [70] dealt with the unconditional expectation function.

Conditional change point analysis of the expectation function $Z(\mathbf{x};\boldsymbol{\theta})$ under a Bernoulli setting is discussed in Pastor and Gualler [56], Pastor-Barriuso *et al* [57], Vexler and Gurevich [66] among others. Pastor and Gualler [56], Pastor-Barriuso *et al* [57] use a fully defined (parameterized) $Z(\mathbf{x};\boldsymbol{\theta})$.

Vexler and Gurevich [66] use polynomial approximation to estimate $Z(\mathbf{x};\boldsymbol{\theta})$. In our study, we use ANN method to estimate $Z(\mathbf{x};\boldsymbol{\theta})$.

2.3.2 Simulation Method

The asymptotic distribution of Q_n may not be known or may be less tractable to derive. The Monte Carlo method is usually used to simulate the critical values of Q_n in equation (2.62)). This is done by doing B repetitions of Q_n under the null hypothesis for a fixed sample size n. The advantage of Monte Carlo simulations is that it gives good approximations even when n is small. By denoting the simulated repetitions of Q_n by $Q^*_{n,b}, b = 1, \ldots, B$, the estimated critical value of Q_n at $1-\alpha$ level is then given by $Q^*_{n,((1-\alpha)(B+1))}$. B is usually chosen to be large, see Hall [33] and Efron and Tibshirani [15] on Monte Carlo quantiles. This method is used when the test statistic Q_n is built from known distribution(s) of Y_i. That is, Q_n is parametric.

When the test statistic Q_n is non-parametric in nature, the bootstrap or other re-sample methods can be used to derive the critical region. The re-sample method is used in Taylor [65] to develop re-sample critical values for the CUSUM Q_n. For the CUSUM method, given a sequence of random variables $Y_1, \ldots, Y_n$, one first defines an estimator of the magnitude of change

as:

$$\begin{aligned}
\Delta &= S_{max} - S_{min} \text{ , where} \\
S_{max} &= \max_{i=0,\ldots,n} S_i \\
S_{min} &= \min_{i=0,\ldots,n} S_i \\
S_i &= S_{i-1} + (Y_i - \bar{Y}), \text{for } i = 1, \ldots, n \text{ , } S_0 = 0 \text{ and} \\
\bar{Y} &= \frac{Y_1 + \ldots + Y_n}{n}
\end{aligned}$$

The procedure then continues as follows;

1. Generate a re-sample of n units by re-ordering the original n observations. This translates to sampling without replacement.

2. From the re-sample in (1) above, determine the CUSUM.

3. Calculate the maximum and minimum of the re-sample CUSUM, S^b_{max} and S^b_{min}. Then, the difference of the bootstrap CUSUM is given by $\Delta^b = S^b_{max} - S^b_{min}$.

4. Determine whether the re-sample difference Δ^b is greater than Δ.

5. Repeat steps (1)-(4) B times where B is large, see Efron and Tibshirani [15] for the choice of B.

By letting b be the number of times $\Delta^b < \Delta$ out of B samples, the re-sample confidence level (RCL) is then defined as

$$RCL = \frac{100 * b}{B}$$

One then rejects H_0 when $RCL \geq 90\%$. The re-sample confidence level method reduces false detections.

2.4 Change Point Estimation

Once the change point has been detected, its estimation can then be carried out.
Various methods for estimating the change point K in Bernoulli random variables exist in the literature. We discuss a few of them.

Assuming that μ_0 and μ_1 in equation (2.43) are known and that Y_i is a Bernoulli random variable, Hinkley and Hinkley [37] and Rukhin [59] use the MLE method to estimate the unknown change point as shown below:

$$\begin{aligned}\hat{K} &= \arg\max_K \left\{ \sum_{i=1}^{K} Y_i \log \mu_0 + (1-Y_i)\log(1-\mu_0) + \right. \\ &\quad \left. + \sum_{i=K+1}^{n} Y_i \log \mu_1 + (1-Y_i)\log(1-\mu_1) \right\} \\ &= \arg\max_K \left\{ S_K \log \frac{\mu_0(1-\mu_1)}{\mu_1(1-\mu_0)} + K \log \frac{(1-\mu_0)}{(1-\mu_1)} \right\} \end{aligned} \tag{2.64}$$

where $\mu_0 = P(Y_i = 1)$ for $i = 1, \ldots, K$, $\mu_1 = P(Y_i = 1)$ for $i = K+1, \ldots, n$ and $S_K = \sum_{i=1}^{K} Y_i$ for $K = 1, \ldots, n-1$. In practice however, μ_0 and μ_1 are often not known in advance.

When μ_0 and μ_1 (or one of them) are unknown, Hinkley and Hinkley [37] and Worsley [70] replace them with their MLE estimators so that equation (2.64) can be written as

$$\begin{aligned}\hat{K} &= \arg\max_K \left\{ \sum_{i=1}^{K} Y_i \log \hat{\mu}_0^K + (1-Y_i)\log(1-\hat{\mu}_0^K) + \right. \\ &\quad \left. + \sum_{i=K+1}^{n} Y_i \log \hat{\mu}_1^K + (1-Y_i)\log(1-\hat{\mu}_1^K) \right\} \end{aligned} \tag{2.65}$$

where

$$\hat{\mu}_0^K = \frac{1}{K} \sum_{i=1}^{K} Y_i$$

and

$$\hat{\mu}_1^K = \frac{1}{n-K} \sum_{i=K+1}^{n} Y_i$$

Worsley [70] and Horvath [38] estimate the change point in model (2.43) using the CUSUM method when both μ_0 and μ_1 are unknown. The change point estimator is given as

$$\hat{K} = \max_{1 \le K \le n} |Q_K| \tag{2.66}$$

where Q_K is the cumulative sum of all successes minus the proportion of all successes up to and including period K, divided by the sample standard deviation. Mathematically, we have;

$$
\begin{aligned}
Q_K &= \left(\sum_{i=1}^{K} Y_i - \frac{K}{n}\sum_{i=1}^{n} Y_i\right)/\sqrt{n\hat{\sigma}^2} \\
&= K\left(\bar{Y}_K - \bar{Y}\right)/\sqrt{n\hat{\sigma}^2} \qquad (2.67)
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{Y}_K &= \frac{\sum_{i=1}^{K} Y_i}{K} \\
\bar{Y} &= \frac{\sum_{i=1}^{n} Y_i}{n} \\
\hat{\sigma}^2 &= \bar{Y}(1-\bar{Y})
\end{aligned}
$$

The above work was based on the assumption that both μ_0 and μ_1 are unconditional.

A lot of research has been done on the case when both μ_0 and μ_1 depend on covariates. For the Bernoulli case, they are of the form;

$$
\begin{aligned}
\mu_{0i} &= P\left(Y_i = 1|\mathbf{X}_i = \mathbf{x}\right) \ , i = 1, \ldots, K \\
\mu_{1i} &= P\left(Y_i = 1|\mathbf{X}_i = \mathbf{x}\right) \ , i = K+1, \ldots, n \qquad (2.68)
\end{aligned}
$$

The logistic function defined in section (2.1.2) as

$$
P(Y_i = 1|\mathbf{X}_i = \mathbf{x}) = \frac{1}{1 + exp(-(\beta_0 + \sum_{j=1}^{d}\beta_j x_j))}
$$

is used to model binary outcomes when the mean function is known to be dependent on the covariates. This approach has been taken by Seber and Wild [60], Pastor and Gualler [56], Pastor-Barriuso *et al* [57] and Vexler and Gurevich [66] among others.

Two methods can be used to estimate K in the conditional probability case.
The first method is the likelihood ratio method given as

$$
\hat{K} = \arg\min\left\{Q_n = 2\log\Lambda_K : 1 \leq K < n\right\} \qquad (2.69)
$$

where Λ_K is the likelihood ratio given by

$$
\begin{aligned}
\Lambda_K &= \left\{\Pi_{i=1}^{K}\{P_{H_1}(Y_i = 1|X_i = \mathbf{x})\}^{Y_i}\{1 - P_{H_1}(Y_i = 1|X_i = \mathbf{x})\}^{1-Y_i}\right\} * \\
&\quad * \frac{\left\{\Pi_{i=K+1}^{n}\{P_{H_1}(Y_i = 1|X_i = \mathbf{x})\}^{Y_i}\{1 - P_{H_1}(Y_i = 1|X_i = \mathbf{x})\}^{1-Y_i}\right\}}{\Pi_{i=1}^{n}\{P_{H_o}(Y_i = 1|X_i = \mathbf{x})\}^{Y_i}\{1 - P_{H_o}(Y_i = 1|X_i = \mathbf{x})\}^{1-Y_i}} \qquad (2.70)
\end{aligned}
$$

where $P_{H_o}(Y_i = 1|X_i = \mathbf{x})$ is the mean function of Y_i before the change and $P_{H_1}(Y_i = 1|X_i = \mathbf{x})$ is the mean function of Y_i after the change. This approach has been used by Gombay and Horvath [30] among others.

The second method estimates K as the value of K which maximizes the log-likelihood function. That is

$$\begin{aligned}\hat{K} &= \arg\max_K \left\{ \sum_{i=1}^{K} [Y_i \log P_{H_o}(Y_i = 1|X_i = \mathbf{x}) + (1 - Y_i) \log(1 - P_{H_o}(Y_i = 1|X_i = \mathbf{x}))] \right. \\ &\quad \left. + \sum_{i=K+1}^{n} [Y_i \log P_{H_1}(Y_i = 1|X_i = \mathbf{x}) + (1 - Y_i) \log(1 - P_{H_1}(Y_i = 1|X_i = \mathbf{x}))] \right\} \end{aligned} \tag{2.71}$$

This approach has been taken by Pastor and Gualler [56], Pastor-Barriuso *et al* [57] and Vexler and Gurevich [66] among others.

Chapter 3

CHANGE-POINT DETECTION

We assume, as in Gurevich and Vexler [66], that a sample of n independent observations $(Y_i, X_i), 1 \leq i \leq n$, is available where $\mathbf{X}_i = [x_{1i}, \ldots, x_{di}]^T \in \Re^d$ are independent identically distributed (iid) random predictors with joint distribution function F on Ω. Y_i are independent Bernoulli variates whose exact distribution depends on the predictor $\mathbf{X}_i$. Further, we assume that $x_{1i}, \ldots, x_{di}$ are scalar values.

We first consider the following standard change point problem. Assume that the data $(Y_i, \mathbf{X}_i)$ are independent and the conditional distribution of Y_i given $\mathbf{X}_i = \mathbf{x}$ is $\mathcal{B}(1, p(\mathbf{x}))$. Then we consider the change point testing problem

$$\begin{aligned} H_0 \quad &: \quad p_i(\mathbf{x}) = p_0(\mathbf{x}),\ i = 1, \ldots, n \\ vs & \\ H_1 \quad &: \quad \text{for some } i \leq K,\ p_i(\mathbf{x}) = p_0(\mathbf{x}), \text{ and for } i \geq K+1,\ p_i(\mathbf{x}) = p(\mathbf{x}) \\ &\quad\ \ \text{where } 1 \leq K \leq n-1 \text{ is unknown, and } p_0 \neq p \end{aligned} \tag{3.1}$$

The general form of the likelihood function is, as $Y_i \in \{0, 1\}$

$$L = \Pi_{i=1}^{n} [p_i(\mathbf{x})]^{Y_i} [1 - p_i(\mathbf{x})]^{1-Y_i}$$

We approximate our model by a parametric one by replacing $p_o(\mathbf{x})$ by $Z(\mathbf{x}; \boldsymbol{\theta})$ and $p(\mathbf{x})$ by $Z(\mathbf{x}; \boldsymbol{\theta}^*)$ where

$$Z(\mathbf{x}; \boldsymbol{\theta}) = \psi(O_H(\mathbf{x}; \boldsymbol{\theta})),\ Z(\mathbf{x}; \boldsymbol{\theta}^*) = \psi(O_H(\mathbf{x}; \boldsymbol{\theta}^*))$$

ψ being the unipolar or logistic function as before, and

$$\begin{aligned} O_H(\mathbf{x};\boldsymbol{\theta}) &= \alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j), \\ O_H(\mathbf{x};\boldsymbol{\theta}^*) &= \alpha_0^* + \sum_{h=1}^{H} \alpha_h^* \psi(W_{h0}^* + \sum_{j=1}^{d} W_{hj}^* x_j) \end{aligned} \tag{3.2}$$

where $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, a suitably close compact subset of $\Re^D, D = 1 + (d+2)H$.

3.1 Model Definition

As mentioned above, $(Y_i, \mathbf{X}_i)$ are independent with the following conditional distribution, say

$$\mathcal{L}(Y_i|\mathbf{X}_i = \mathbf{x}) = B(1, p(\mathbf{x})) \tag{3.3}$$

We approximate this general model by a parametric one where $p(\mathbf{x})$ is replaced by the output function $Z(\mathbf{x};\boldsymbol{\theta})$ of a neural network:

$$\begin{aligned} Z(\mathbf{x};\boldsymbol{\theta}) &= \psi(O_H(\mathbf{x};\boldsymbol{\theta})), \\ O_H(\mathbf{x};\boldsymbol{\theta}) &= \alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j) \\ \text{and} & \\ \boldsymbol{\theta} &= \alpha_0, \ldots, \alpha_H, W_{10}, \ldots, W_{Hd} \in \boldsymbol{\Theta}, \text{ assumed compact} \end{aligned} \tag{3.4}$$

Therefore, in the following, we pretend that $\{(Y_i, \mathbf{X}_i), i = 1, \ldots, n\}$ are i.i.d. with

$$\mathcal{L}(Y_i|\mathbf{X}_i = \mathbf{x}) = B(1, Z(\mathbf{x};\boldsymbol{\theta})) \tag{3.5}$$

but we are aware that this model may be misspecified.

In line with Gombay and Horvath [30], $(Y_i, \mathbf{X}_i)$ has a density function $\pi(Y, \mathbf{x};\boldsymbol{\theta})$ with respect to a $\sigma-$finite measure ν.

By choosing $\mu = \mathcal{L}(\mathbf{X}_i)$, we have $\nu = [\delta_0 + \delta_1] \otimes \mu$ where $\delta_{\mathcal{L}}$ is point mass in $\mathcal{L}$. Since Y_i is a bernoulli random variable, $\delta_0 + \delta_1$ is a counting measure on $\{0, 1\}$.

We then write the following

$$\begin{aligned} P(Y_i \in A, X_i \in C) &= \int_{AXC} \pi(Y, \mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\left[\delta_0(Y) + \delta_1(Y)\right] \otimes \mu(\mathbf{x}) \\ &= \int_C P(Y_i = Y | \mathbf{X}_i = \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \end{aligned} \tag{3.6}$$

$$\Rightarrow \pi(Y, \mathbf{x}; \boldsymbol{\theta}) = P(Y_i = Y | \mathbf{X}_i = \mathbf{x}) = \begin{cases} Z(\mathbf{x}; \boldsymbol{\theta}) & , Y = 1 \\ 1 - Z(\mathbf{x}; \boldsymbol{\theta}) & , Y = 0 \end{cases} \tag{3.7}$$

so that

$$\pi(Y, \mathbf{x}; \boldsymbol{\theta}) = \delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta})) \tag{3.8}$$

$\pi(Y, \mathbf{x}; \boldsymbol{\theta})$ is a smooth function of $\boldsymbol{\theta}$ since $Z(\mathbf{x}; \boldsymbol{\theta})$ is smooth by equation (3.4).

3.1.1 Change Point Model Definition

In line with Rukhin [59], we let Q and R be two different Bernoulli distributions with probabilities $Z(\mathbf{x}; \boldsymbol{\theta}_0)$ and $Z(\mathbf{x}; \boldsymbol{\theta}_1)$ respectively. We then assume that the observed data $(Y_1, \mathbf{X}_1), \ldots, (Y_K, \mathbf{X}_K), \ldots, (Y_n, \mathbf{X}_n)$ consist of two independent segments.

The first segment $(Y_1, \mathbf{X}_1), \ldots, (Y_K, \mathbf{X}_K)$ is a random sample from Q while the second sample $(Y_{K+1}, \mathbf{X}_{K+1}), \ldots, (Y_n, \mathbf{X}_n)$ is a random sample from distribution R.

Specifically, we model the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ as follows:

$$\pi(Y, \mathbf{x}; \boldsymbol{\theta}) = \delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta})), Y = 0, 1 \tag{3.9}$$

We model a possible change point by assuming that

$$P(Y_i = 1 | \mathbf{X}_i = \mathbf{x}) = \begin{cases} Z(\mathbf{x}; \boldsymbol{\theta}_0) & , \text{ for } 1 \leq i \leq K \\ Z(\mathbf{x}; \boldsymbol{\theta}_1) & , \text{ for } K+1 \leq i \leq n \end{cases} \tag{3.10}$$

We note that, when there is no change point $Z(\mathbf{x}; \boldsymbol{\theta}_0) = Z(\mathbf{x}; \boldsymbol{\theta}_1)$ so that

$$P(Y_i = 1 | \mathbf{X}_i = \mathbf{x}) = Z(\mathbf{x}; \boldsymbol{\theta}_0) \text{ for } 1 \leq i \leq n \tag{3.11}$$

From equation 3.10, the statistical model that relates Y and $Z(\mathbf{x}; \boldsymbol{\theta})$ is;

$$Y_i = \begin{cases} Z(\mathbf{x}; \boldsymbol{\theta}_0) + \epsilon_i & , \text{ for } 1 \leq i \leq K \\ Z(\mathbf{x}; \boldsymbol{\theta}_1) + \epsilon_i & , \text{ for } K+1 \leq i \leq n \end{cases} \tag{3.12}$$

where ϵ_i's are the model errors and K is the change point.
Further, we make the following assumptions:

Assumption 1

(a) ϵ_i's are independently and under H_0 identically distributed with zero mean and finite variance.
(b) X_i's are i.i.d. random vectors.
(c) X_i's are independent of the ϵ_i's for $1 \leq i \leq n$

Some remarks are in order:
The function $Z(\mathbf{x}; \boldsymbol{\theta}) = E[Y_i | X_i = \mathbf{x}]$ and the change point K are unknown and are to be estimated. This essentially entails estimating $\boldsymbol{\theta}$ from the data for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. The i.i.d. assumption is chosen here because it is central to the classical theory of regression, see White [67] for more details. The i.i.d. assumption can be relaxed to accommodate time series data.

3.2 Parameter Estimation

As mentioned in chapter two of this work, the parameters are estimated using the artificial neural network (ANN) techniques. In particular, a feed-forward net is used in this study. We note that there are many training paradigms. Among the most common in literature are feed-forward and back propagation. Here, we use the feed-forward algorithm which White [69] shows to be more statistically efficient than the back propagation algorithm. In this study, we use the unipolar activation function, ψ , as defined in chapter two.

From equation (3.8), the negative log-likelihood is given by

$$S(Y, \mathbf{X}; \boldsymbol{\theta}) = -\sum_{i=1}^{n} \{\delta_1(Y_i) \ln(Z(\mathbf{X}_i; \boldsymbol{\theta})) + \delta_0(Y_i) \ln(1 - Z(\mathbf{X}_i; \boldsymbol{\theta}))\} \tag{3.13}$$

where as before

$$Z(\mathbf{X}; \boldsymbol{\theta})) = \psi(O_H(\mathbf{X}; \boldsymbol{\theta})) \tag{3.14}$$

and $\psi(.)$ is as defined in chapter two.

Then the estimator of $\boldsymbol{\theta}$ is given as,

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_o} S(Y, \mathbf{X}; \boldsymbol{\theta}) \tag{3.15}$$

Next, we discuss the conditions that guarantee the existence, consistency and asymptotic normality of (3.15).

3.3 Existence of the Estimator

The following lemma guarantees the existence of a solution to (3.15) if we assume that $\boldsymbol{\Theta}_o$ is compact which is a common assumption if dealing with artificial neural networks.

Lemma 3.1. *Assume (3.13) and (3.14) for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where the latter is compact. Then, there exists a solution of the maximum likelihood equation (3.15) a.s.*

Proof. By our choice of $\psi(.)$ and $O_H(.)$, $Z(\mathbf{x};\boldsymbol{\theta})$ given by (3.14) is continuous in $\mathbf{x}$ and $\boldsymbol{\theta}$, and $0 < Z(\mathbf{x};\boldsymbol{\theta}) < 1$ for all $\mathbf{x}, \boldsymbol{\theta}$. Therefore, $S(Y, \mathbf{X};\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for all $Y, \mathbf{X}$, and it assumes its minimum on compact sets. □

3.4 Model Irreducibility

Recall the definition of a neural network by (2.1) - (2.4) where we mainly consider the unipolar activation function $\psi(u)$ in this thesis as it assumes values in [0,1]. To refer to the existing literature more easily, we consider in the next two sections the case of the bipolar activation function $\widetilde{\psi}(u)$ for the hidden neurons and of the identity for the output neuron. The network is now given by

$$\begin{aligned} \upsilon_h(\mathbf{x};\boldsymbol{\theta}) &= W_{h0} + \sum_{j=1}^{d} W_{hj} x_j \\ \widetilde{\phi}_h(\mathbf{x};\boldsymbol{\theta}) &= \widetilde{\psi}(\upsilon_h(\mathbf{x};\boldsymbol{\theta})) \\ \widetilde{Z}(\mathbf{x};\boldsymbol{\theta}) &= \widetilde{O}_H(\mathbf{x};\boldsymbol{\theta}) = \widetilde{\alpha}_0 + \sum_{h=1}^{H} \widetilde{\alpha}_h \widetilde{\phi}_h(\mathbf{x},\boldsymbol{\theta}) \end{aligned} \tag{3.16}$$

We discuss identifiability issues, i.e. up to which respect the mapping $(x_1, \ldots, x_d) \longmapsto \widetilde{Z}(\mathbf{x};\boldsymbol{\theta})$ respectively $\longmapsto Z(\mathbf{x};\boldsymbol{\theta})$ determines the parameters.

From the definition of ϕ and $\widetilde{\phi}$, we immediately have

$$\widetilde{\psi}(u) = 2\psi(u) - 1, \;\; \psi(u) = \frac{1}{2}(1 + \widetilde{\psi}(u)) \tag{3.17}$$

Therefore, we have

$$\begin{aligned}\widetilde{O}_H(\mathbf{x};\widetilde{\boldsymbol{\theta}}) &= \widetilde{\alpha}_0 + \sum_{h=1}^{H} \widetilde{\alpha}_h(2\psi(\upsilon_h(\mathbf{x};\boldsymbol{\theta})) - 1) \\ &= (\widetilde{\alpha}_0 - \sum_{h=1}^{H} \widetilde{\alpha}_h) + \sum_{h=1}^{H} (2\widetilde{\alpha}_h)\psi(\upsilon_h(\mathbf{x};\boldsymbol{\theta})) = O_H(\mathbf{x};\boldsymbol{\theta})\end{aligned} \tag{3.18}$$

if we relate $\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}$ by

$$\begin{aligned}\alpha_0 &= \widetilde{\alpha}_0 - \sum_{h=1}^{H} \widetilde{\alpha}_h, \\ \alpha_h &= 2\widetilde{\alpha}_h, h = 1, \ldots, H\end{aligned} \tag{3.19}$$

and the W_{hj} are the same in both parameter vectors $\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}$. Therefore, identifiability of the parameter vector from the mapping holds for $O_H(\mathbf{x};\boldsymbol{\theta})$ iff it holds for $\widetilde{O}_H(\mathbf{x};\boldsymbol{\theta})$. Finally, we remark that by continuity and strict monotonicity of $\psi(u)$, $Z(\mathbf{x};\boldsymbol{\theta})$ and $O_H(\mathbf{x};\boldsymbol{\theta})$ determine each other uniquely. Therefore, we can study identifiability issues for $\widetilde{Z}(\mathbf{x};\widetilde{\boldsymbol{\theta}})$ instead of $Z(\mathbf{x};\boldsymbol{\theta})$.

To simplify notation, we now write ψ for the bipolar activation function and $\boldsymbol{\theta}$ for the corresponding parameter.

Before we deal with model irreducibility, we first define redundancy. A neural network with a given $\boldsymbol{\theta}$ is redundant if there exists another network with fewer neurons that gives the same input-output map.

A net with $\psi(\mathbf{x})$ satisfying (2.8) and $\boldsymbol{\theta}$ as in (3.2) is reducible if one of the following conditions hold:

(a) $\alpha_i = 0$ for some $i = 1, \ldots, H$.

(b) One of the functions $\upsilon_i(\mathbf{x};\boldsymbol{\theta})$ is a constant.

(c) There exist two different indexes $i_1, i_2 \in (1, \ldots, H)$ such that the functions $\upsilon_{i_1}(\mathbf{x})$ and $\upsilon_{i_2}(\mathbf{x})$ are sign equivalent. That is $\upsilon_{i_1}(\mathbf{x};\boldsymbol{\theta}) = \pm\upsilon_{i_2}(\mathbf{x};\boldsymbol{\theta})$.

A neural network that meets any of the 3 conditions above is redundant (see Sussmann [63]) because its input-output function can be achieved by another net with fewer hidden neurons. This is achieved after deleting one neuron. We note that if condition (a) holds, then the i-th hidden node makes no contribution to the net input $Z(\mathbf{x};\boldsymbol{\theta})$. The input-output map will therefore be unchanged if we remove the i-th node.

If a net is reducible because (b) holds, then $\upsilon_i = c$, c being a constant. One can therefore remove the i-th node and replace α_o by $\alpha_o + \alpha_i\psi(c)$. This condition can only arise if for a certain fixed i, $W_{ij} = 0$ for all $j = 1, \ldots, d$. In this case then, $\upsilon_i = W_{io}$.

Lastly, if a net is reducible because (c) holds, we can write $\upsilon_{i_1}(\mathbf{x}; \boldsymbol{\theta}) = \tau\upsilon_{i_2}(\mathbf{x}; \boldsymbol{\theta})$ where $\tau = 1$ or $\tau = -1$. Then, the nodes i_1 and i_2 contribute to $Z(\mathbf{x}; \boldsymbol{\theta})$ a combined value of

$$\begin{aligned} \alpha_{i_1}\psi(\upsilon_{i_1}(\mathbf{x}; \boldsymbol{\theta})) + \alpha_{i_2}\psi(\upsilon_{i_2}(\mathbf{x}; \boldsymbol{\theta})) &= \alpha_{i_1}\psi(\tau\upsilon_{i_2}(\mathbf{x}; \boldsymbol{\theta})) + \alpha_{i_2}\psi(\upsilon_{i_2}(\mathbf{x}; \boldsymbol{\theta})) \\ &= \tau\alpha_{i_1}\psi(\upsilon_{i_2}(\mathbf{x}; \boldsymbol{\theta})) + \alpha_{i_2}\psi(\upsilon_{i_2}(\mathbf{x}; \boldsymbol{\theta})) \\ &= (\tau\alpha_{i_1} + \alpha_{i_2})\upsilon_{i_2}(\mathbf{x}; \boldsymbol{\theta}) \qquad (3.20) \end{aligned}$$

This relation is due to the fact that $\psi(x)$ in (2.8) is an odd function, that is, $\psi(\tau x) = \tau\psi(x)$. In this type of reducibility, we can then remove node i_1 and replace α_{i_2} by $\tau\alpha_{i_1} + \alpha_{i_2}$.

Conditions (a) and (b) are brought about by the presence of irrelevant neurons in the hidden layer. To control these conditions, we use the Schwarz Information Criterion (SIC) proposed in Swanson and White [64] as our model selection criterion given as:

$$SIC(h) = \ln(\hat{\sigma}^2) + (h(2 + d) + 1)\frac{\ln(n)}{n} \qquad (3.21)$$

The first term is the goodness-of-fit measure while the second term is the complexity penalty.

Using the SIC criterion, we start with a single hidden neuron and determine SIC(1). Then a second hidden neuron is added and SIC(2) determined. The process is continued until when an extra hidden neuron does not improve the SIC. We therefore estimate $h + 1$ models in order to choose a model with h neurons.

This procedure ensures that $\alpha_i \neq 0\ \forall i$ and $W_{ij} \neq 0$ for $j = 1, \ldots, d\ \forall i$.

We therefore only have to make the following assumption to ensure that $\boldsymbol{\theta}$ is irreducible hence non-redundant:

Assumption 2

There exist no two different indexes $i_1, i_2 \in 1, \ldots, H$ such that the functions $\upsilon_{i_1}(\mathbf{x}; \boldsymbol{\theta})$ and $\upsilon_{i_2}(\mathbf{x}; \boldsymbol{\theta})$ are sign equivalent.

This assumption solves the irreducibility caused by condition (c) above. The result translates immediately to the case of a unipolar activation function by equation (3.19) and the discussion above. However even though $\boldsymbol{\theta}$ is now irreducible, it is unidentifiable as discussed below.

3.5 Model Identifiability

A fundamental problem of the ANN is the un-identifiability of the parameters. We therefore have different sets of parameters but the corresponding distributions of $(Y, \mathbf{X})$ are identical. The parameters are therefore not unique. This problem of unidentifiability has been studied by Sussmann [63] and Hwang and Ding [41] among others.

In order to explain this clearly, we represent all the weights as follows:

$$\begin{aligned} \alpha_o \text{ and } \boldsymbol{\beta}_i = (\alpha_i, \mathbf{W}_i) \text{ for } i = 1, \ldots, H \\ \text{where } \mathbf{W}_i = (W_{io}, W_{i1}, \ldots, W_{id}) \end{aligned} \tag{3.22}$$

We now discuss the sources of unidentifiability. As noted by Hwang and Ding [41], every ANN is unidentifiable.
The following two transformations leave a neural network input-output map unchanged:

(i) The permutation of $\boldsymbol{\beta}_i's$. That is, if we interchange two hidden nodes, say h_s and h_t where s and t denote the node's position, and relabel them as h_t and h_s and of course also relabel the corresponding weights as α_t and α_s, and $\mathbf{W}_t$ and $\mathbf{W}_s$, $Z(\mathbf{X}_i; \boldsymbol{\theta})$ remains unchanged. This transformation alone yields $H!$ different models that have the same input-output map.

(ii) The other invariant transformation is due to the symmetry of $\psi(x)$.That is $\psi(x) = -\psi(-x)$. This means that if we choose a hidden node h_t and negate $\mathbf{W}_t$ as well as α_t , the input-output map remains unchanged. This in effect means that $(\alpha_o, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_H)$ and $(\alpha_o, \boldsymbol{\beta}_1, \ldots, -\boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_H)$ have the same input-output map. This transformation alone yields 2^H different models with the same input-output map.

We note that similar results are found in Hwang and Ding [41] but the activation function in their case was unipolar. In particular, they show that $(\alpha_o, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_H)$ and $(\alpha_o + \alpha_i, \boldsymbol{\beta}_1, \ldots, -\boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_H)$ have the same input-output map.

As pointed out in Sussmann [63], these two transformations generate a family of $2^H H!$ elements. Call all of these transformations η. Similar to Hwang and Ding [41], we characterize each of these transformations as being a composite function of $(\eta_1, \ldots, \eta_H)$ where

$$\begin{aligned}
\eta_1([\alpha_o, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_H)]) &= (\alpha_o, -\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_H) \\
&\text{and} \\
\eta_i([\alpha_o, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_i, \ldots, \boldsymbol{\beta}_H)]) &= (\alpha_o, \boldsymbol{\beta}_i, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{i-1}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_{i+1} \ldots, \boldsymbol{\beta}_H) \\
&\text{for } i = 2, \ldots, H
\end{aligned} \tag{3.23}$$

The following theorem corresponds to theorem (2.3) of Hwang and Ding [41].

Theorem 3.1. *Assume models (2.8),(3.2) and (3.12) and further that model (2.8) is a continuous function satisfying condition A of Hwang and Ding [41]. Suppose that $\boldsymbol{\theta}$ in (3.2) is irreducible. Assume also that the distribution of $\boldsymbol{X}$ has support in $\Re^d$. Then, $\boldsymbol{\theta}$ is identifiable up to the family of transformations generated by (3.23).*

Proof. The proof of this theorem can be found in Hwang and Ding [41]. The task is therefore to prove that $\psi(\mathbf{x})$ satisfies condition A of Hwang and Ding [41] and that it is continuous.
In particular, this condition requires that for any $h > 0$, any scalars α_o , $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_i > 0$ for $i = 1, \ldots, h$ where $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j$ for every $i \neq j$, the condition $\zeta(\mathbf{x}; \boldsymbol{\theta}) = \alpha_0 + \sum_{i=1}^{h} \alpha_i \phi_i(\mathbf{x}; \boldsymbol{\theta}) = 0 \ \forall \mathbf{x} \in \Re^d$ implies that $\alpha_o = \alpha_1 = \ldots = \alpha_h = 0$. That is, the class of functions $\{\psi(\mathbf{W}\mathbf{x} + W_o), W_i > 0\} \cup \{\psi \equiv 1\}$ is linearly independent. This condition is proved in lemma 1 of Sussmann [63]. □

Remarks
1. A network with a tansig activation function, or any other activation function satisfying condition A of Hwang and Ding [41], has $2^h h!$ transformations which are the only ways to modify $\boldsymbol{\theta}$ without changing the input-output map.
2. Theorem 3.2 above implies that if there exists another $\boldsymbol{\theta}^+$ such that $Z(\mathbf{x}; \boldsymbol{\theta}^+) = Z(\mathbf{x}; \boldsymbol{\theta})$, then we can find a transformation generated by (3.23) that transforms $\boldsymbol{\theta}^+$ to $\boldsymbol{\theta}$.
3. The above work has dealt extensively with the irreducibility and identifiability of the tansig (Bipolar) sigmoid. We note that the tansig sigmoid can easily be transformed to the unipolar (Logistic) sigmoid, refer to equation 2.39, which models Bernoulli experiments well. It is for this reason that later in this work we apply the unipolar sigmoid.

3.6 Consistency and Asymptotic Normality of Network Parameter Estimates

In this section, we assume that $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$, are i.i.d. with

$$\mathcal{L}(Y_i|\mathbf{X}_i = \mathbf{x}) = B(1, p(\mathbf{x})) \tag{3.24}$$

We fit a neural network output function $Z(\mathbf{x}; \boldsymbol{\theta})$ to $p(\mathbf{x})$ by minimizing the negative log likelihood multiplied by $1/n$.

$$S(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} [Y_i \ln Z(\mathbf{X}_i; \boldsymbol{\theta}) + (1 - Y_i) \ln(1 - Z(\mathbf{X}_i; \boldsymbol{\theta}))] \tag{3.25}$$

Let $S_0(\boldsymbol{\theta}) = E[S(\boldsymbol{\theta})]$ denote the expectation of the target function $S(\boldsymbol{\theta})$. As $(Y_i, \mathbf{X}_i)$ are i.i.d. , we have

$$\begin{aligned} S_0(\boldsymbol{\theta}) &= -E[Y_1 \ln Z(\mathbf{X}_1; \boldsymbol{\theta}) + (1 - Y_1) \ln(1 - Z(\mathbf{X}_1; \boldsymbol{\theta}))] \\ &= -E[p(X_1) \ln Z(\mathbf{X}_1; \boldsymbol{\theta}) + (1 - p(\mathbf{X}_1)) \ln(1 - Z(\mathbf{X}_1; \boldsymbol{\theta}))] \end{aligned} \tag{3.26}$$

Assume that $S_0(\boldsymbol{\theta})$ has a unique minimum if $\boldsymbol{\theta}$ ranges over a given compact set $\boldsymbol{\Theta}$. Then, this minimum is characterized by

$$0 = \nabla S_0(\boldsymbol{\theta}) = -E\left[\frac{p(\mathbf{X}_1)}{Z(\mathbf{X}_1; \boldsymbol{\theta})} - \frac{1 - p(\mathbf{X}_1)}{1 - Z(\mathbf{X}_1; \boldsymbol{\theta})}\right] \nabla Z(\mathbf{X}_1; \boldsymbol{\theta}) \tag{3.27}$$

where we have used the fact that neural network output functions of the form (3.14) are continuous in $\mathbf{x}$ and $\boldsymbol{\theta}$ and continuously differentiable in $\boldsymbol{\theta}$ such that we may interchange expectation and differentiation.

For the correctly specified case where $p(\mathbf{x}) = Z(\mathbf{x}; \boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, equation (3.27) is obviously solved for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, i.e. $S_0(\boldsymbol{\theta})$ is minimized at the true parameter value $\boldsymbol{\theta}_0$. In the general case, where there is no true parameter value, we define $\boldsymbol{\theta}_0$ as

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} S_0(\boldsymbol{\theta}) \tag{3.28}$$

Consistency of the estimator $\hat{\boldsymbol{\theta}}$ which we get by minimizing equation (3.25) then means that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ for $n \to \infty$ in probability.

We want to rely as much as possible on previous work which mainly is in the context of classical regression models. Therefore, we rewrite our model as

$$Y_i = p(\mathbf{X}_i) + \epsilon_i, \;\; i = 1, \ldots, n \tag{3.29}$$

where the residuals are defined as

$$\epsilon_i = Y_i - p(\mathbf{X}_i) \tag{3.30}$$

As $(Y_i, \mathbf{X}_i)$ are i.i.d. and $P(Y_i = 1|\mathbf{X}_i) = E(Y_i|\mathbf{X}_i) = p(\mathbf{X}_i)$, we immediately have that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d., $E[\epsilon_i] = 0$ and

$$\begin{aligned} \mathrm{Var}(\epsilon_i) &= E(Y_i - p(\mathbf{X}_i))^2 \\ &= E\{E[(Y_i - p(\mathbf{X}_i))^2|\mathbf{X}_i]\} \\ &= E[p(\mathbf{X}_i)(1 - p(\mathbf{X}_i))^2 + (1 - p(\mathbf{X}_i))p^2(\mathbf{X}_i)] \\ &= E[p(\mathbf{X}_i)(1 - p(\mathbf{X}_i))] = \sigma_\epsilon^2 < \infty \end{aligned} \tag{3.31}$$

Moreover,

$$\mathrm{Var}(\epsilon_i|\mathbf{X}_i = \mathbf{x}) = \sigma_\epsilon^2(\mathbf{x}) = p(\mathbf{x})(1 - p(\mathbf{x}))$$

. We comment here that in this formulation, $\mathrm{Var}(\epsilon_i)$ does not depend on $\boldsymbol{\theta}$ which is good.

We need the following uniform law of large numbers (ULLN) whose proof can be found in Andrews [1].

Theorem 3.2. *Let $\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots$ be i.i.d. random vectors in $\Re^d$, $\boldsymbol{\Theta} \subseteq \Re^M$ compact, $\gamma : \Re^d \times \boldsymbol{\Theta} \to \Re$ measurable such that*

(i) $E|\gamma(\boldsymbol{U}_1; \boldsymbol{\theta})| < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

(ii) $\gamma(\boldsymbol{u}; \boldsymbol{\theta})$ is Lipschitz continuous in $\boldsymbol{\theta}$, i.e. for some $L(\boldsymbol{u}) > 0$

$$|\gamma(\boldsymbol{u}; \boldsymbol{\theta}) - \gamma(\boldsymbol{u}; \boldsymbol{\theta}')| \leq L(\boldsymbol{u}) \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|$$

for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$

(iii) $E[L(\boldsymbol{U}_1)] < \infty$

Then,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma(\boldsymbol{U}_i; \boldsymbol{\theta}) - E\gamma(\boldsymbol{U}_1; \boldsymbol{\theta}) \right| \to 0$$

in probability.

We want to apply the same kind of argument as Franke and Neumann [24] who discussed nonlinear least square estimates for neural network parameters in a setting allowing for misspecification. As the residuals ϵ_i are not only i.i.d. but also bounded in absolute value by 1, their assumptions reduce to

(A_1) The activation function ψ is bounded and twice continuously differentiable with bounded derivatives.

(A_2) $S_0(\boldsymbol{\theta})$ has a unique global minimum at $\boldsymbol{\theta}_0$ lying in the interior of $\boldsymbol{\Theta}$, and with the Hessian

$$A(\boldsymbol{\theta}_0) = \left(\frac{\partial^2}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_l} S_0(\boldsymbol{\theta}) \right) = \nabla^2 S_0(\boldsymbol{\theta})$$

where $A(\boldsymbol{\theta}_0)$ is positive definite.

(A_3) Let $\boldsymbol{\Theta}$ be chosen such that for some $\Delta > 0$, we have

$$\Delta \leq Z(\mathbf{x}; \boldsymbol{\theta}) \leq 1 - \Delta$$

for all $\mathbf{x} \in \Re^d, \boldsymbol{\theta} \in \boldsymbol{\Theta}$.

(A_4) $(Y_i, \mathbf{X}_i), i = 1, \ldots, n$ be i.i.d. with density $\xi(\mathbf{x})$ and $E \left\| X_1 \right\|^2 < \infty$.

(A_5) $p(\mathbf{x})$ is continuous in $\mathbf{x}$ and $0 < \delta \leq p(\mathbf{x}) \leq 1 - \delta < 1$ for some $\delta > 0$

Theorem 3.3. *Let $(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n$ be i.i.d. with $\mathcal{L}(Y_i | \boldsymbol{X}_i) = B(1, p(\boldsymbol{X}_i))$. Suppose that $(A_1) - (A_5)$ are satisfied. Then, for $n \to \infty$, with $\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0$ as above*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} N\left(0, \Sigma_1 + \Sigma_2\right)$$

where

$$\Sigma_1 = A^{-1}(\boldsymbol{\theta}_0) B_1(\boldsymbol{\theta}_0) A^{-1}(\boldsymbol{\theta}_0)$$

,

$$\Sigma_2 = A^{-1}(\boldsymbol{\theta}_0) B_2(\boldsymbol{\theta}_0) A^{-1}(\boldsymbol{\theta}_0)$$

with

$$B_1(\boldsymbol{\theta}_0) = E \frac{(p(\boldsymbol{X}_1) - Z(\boldsymbol{X}_1; \boldsymbol{\theta}_0))^2}{Z^2(\boldsymbol{X}_1; \boldsymbol{\theta}_0)(1 - Z(\boldsymbol{X}_1; \boldsymbol{\theta}_0))^2} \nabla Z(\boldsymbol{X}_1; \boldsymbol{\theta}_0) . \nabla^T Z(\boldsymbol{X}_1; \boldsymbol{\theta}_0)$$

$$B_2(\boldsymbol{\theta}_0) = E \frac{p(\boldsymbol{X}_1)(1 - p(\boldsymbol{X}_1))}{Z^2(\boldsymbol{X}_1; \boldsymbol{\theta}_0)(1 - Z(\boldsymbol{X}_1; \boldsymbol{\theta}_0))^2} \nabla Z(\boldsymbol{X}_1; \boldsymbol{\theta}_0) . \nabla^T Z(\boldsymbol{X}_1; \boldsymbol{\theta}_0)$$

and $A(\boldsymbol{\theta}_0)$ as above.

Before doing the proof, let us remark that the form $\Sigma_1 + \Sigma_2$ of the asymptotic covariance matrix reflects the two sources of error. $B_1(\boldsymbol{\theta}_0)$ contains the squared modeling bias $(p(\mathbf{x}) - Z(\mathbf{x}; \boldsymbol{\theta}_0))^2$ which vanishes in the correctly specified case. $B_2(\boldsymbol{\theta}_0)$ contains $p(\mathbf{X}_1)(1 - p(\mathbf{X}_1)) = \mathrm{Var}(Y_1 | X_1)$ which reflects the randomness in the response variable Y_1.

Proof. We introduce the auxiliary quantity

$$\tilde{S}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\left[p(\mathbf{X}_i)\ln Z(\mathbf{X}_i;\boldsymbol{\theta}) + (1-p(\mathbf{X}_i))\ln(1-Z(\mathbf{X}_i;\boldsymbol{\theta}))\right] \quad (3.32)$$

which is generated by replacing Y_i by its expectation given X_i in the definition of $S(\boldsymbol{\theta})$. We define $\tilde{\boldsymbol{\theta}}$ as the minimum of $\tilde{S}(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{S}(\boldsymbol{\theta}) \quad (3.33)$$

We prove

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \to N(0, \Sigma_1)$$
$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N(0, \Sigma_2)$$

and asymptotic independence of $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$. This implies the assertion of the theorem as

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

(a) By the ULLN stated above, we have for $\mathbf{U}_j = \mathbf{X}_j$,

$$\gamma(\mathbf{x};\boldsymbol{\theta}) = -p(\mathbf{x})\ln Z(\mathbf{x};\boldsymbol{\theta}) - (1-p(\mathbf{x}))\ln(1-Z(\mathbf{x};\boldsymbol{\theta})) \quad (3.34)$$

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\tilde{S}(\boldsymbol{\theta}) - S_0(\boldsymbol{\theta})\right| = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{n}\sum_{j=1}^{n}\gamma(\mathbf{X}_j;\boldsymbol{\theta}) - E\gamma(\mathbf{X}_1;\boldsymbol{\theta})\right| = o_p(1) \quad (3.35)$$

correspondingly, we get with

$$\mathbf{U}_j = (Y_j, \mathbf{X}_j), \gamma(Y, \mathbf{x};\boldsymbol{\theta}) = -(Y - p(\mathbf{x}))\ln\frac{Z(\mathbf{x};\boldsymbol{\theta})}{1-Z(\mathbf{x};\boldsymbol{\theta})}$$

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|S(\boldsymbol{\theta}) - \tilde{S}(\boldsymbol{\theta})\right| = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{n}\sum_{j=1}^{n}\gamma(Y_j, \mathbf{X}_j;\boldsymbol{\theta}))\right| = o_p(1) \quad (3.36)$$

as

$$E\gamma(Y_1, \mathbf{X}_1;\boldsymbol{\theta}) = -E(Y_1 - p(\mathbf{X}_1))\ln\frac{Z(\mathbf{X}_1;\boldsymbol{\theta})}{1-Z(\mathbf{X}_1;\boldsymbol{\theta})} = 0$$

using

$$E(Y_1|\mathbf{X}_1) = p(\mathbf{X}_1)$$

We only have to check if the conditions $(i) - (iii)$ of the ULLN are satisfied in both cases. From assumption (A_1), we immediately have that $Z(\mathbf{x};\boldsymbol{\theta})$ is continuous and, therefore, measurable in $\mathbf{x}$ and twice continuously differentiable in $\boldsymbol{\theta}$. In particular, $Z(\mathbf{x};\boldsymbol{\theta})$ is uniformly bounded in $\mathbf{x} \in \Re^d, \boldsymbol{\theta} \in \boldsymbol{\Theta}$. We consider the derivatives of $Z(\mathbf{x};\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}_k$ in detail in section (3.8). From the calculations there and assumption (A_1), we have that for some constant c and all $\mathbf{x} \in \Re^d, \boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\left|\frac{\partial}{\partial \boldsymbol{\theta}_l} Z(\mathbf{x};\boldsymbol{\theta})\right| \leq c \text{ if } \boldsymbol{\theta}_l = \alpha_0, \ldots, \alpha_H,\ W_{10}, \ldots, W_{H0},$$

$$\left|\frac{\partial}{\partial \boldsymbol{\theta}_l} Z(\mathbf{x};\boldsymbol{\theta})\right| \leq c|\mathbf{x}_i| \text{ if } \boldsymbol{\theta}_l = W_{1i}, \ldots, W_{Hi},\ i = 1, \ldots, d.$$

It follows for all $\mathbf{x} \in \Re^d, \boldsymbol{\theta} \in \boldsymbol{\Theta}$ and a suitable constant $c' > 0$

$$\|\nabla Z(\mathbf{x};\boldsymbol{\theta})\| \leq c' \|\mathbf{x}\|$$

Correspondingly, we get for some constant $c'' > 0$, using (A_3)

$$\|\nabla \ln Z(\mathbf{x};\boldsymbol{\theta})\| = \frac{\|\nabla Z(\mathbf{x};\boldsymbol{\theta})\|}{Z(\mathbf{x};\boldsymbol{\theta})} \leq c'' \|\mathbf{x}\|$$

$$\|\nabla \ln(1 - Z(\mathbf{x};\boldsymbol{\theta}))\| = \frac{\|\nabla Z(\mathbf{x};\boldsymbol{\theta})\|}{1 - Z(\mathbf{x};\boldsymbol{\theta})} \leq c'' \|\mathbf{x}\|$$

So, we have for $\gamma(\mathbf{x};\boldsymbol{\theta})$ of equation (3.35)

$$\begin{aligned} \left|\gamma(\mathbf{u};\boldsymbol{\theta}) - \gamma(\mathbf{u};\boldsymbol{\theta}')\right| &\leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\nabla\gamma(\mathbf{u};\boldsymbol{\theta})\| \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\| \\ &\leq (p(\mathbf{x}) + 1 - p(\mathbf{x}))c'' \|\mathbf{x}\| \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\| \\ &= c'' \|\mathbf{x}\| \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\| \end{aligned} \tag{3.37}$$

Using (A_4), we get (ii) and (iii) of the ULLN with $L(\mathbf{u}) = c'' \|\mathbf{u}\|$. (i) is satisfied as, from (A_3) and $0 \leq p(\mathbf{x} \leq 1)$, we immediately have that $\gamma(\mathbf{x};\boldsymbol{\theta})$ is uniformly bounded in $\mathbf{x} \in \Re^d, \boldsymbol{\theta} \in \boldsymbol{\Theta}$.

The argument for $\gamma(\mathbf{x};\boldsymbol{\theta})$ of (3.36) goes analogously where we additionally use that Y_j are bounded random variables.

From (3.35) and (3.36), we immediately have

$$|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}| = o_p(1),\ |\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| = o_p(1)$$

Hence by (A_2) with increasing probability, $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$ are interior points of $\boldsymbol{\Theta}$, and we have in particular

$$\nabla S(\hat{\boldsymbol{\theta}}) = \nabla \tilde{S}(\tilde{\boldsymbol{\theta}}) = \nabla S_0(\boldsymbol{\theta}_0) = 0$$

with probability going to 1 for $n \to \infty$.

(b) Hence, with probability going to 1,

$$\begin{aligned} 0 &= \nabla \tilde{S}(\tilde{\boldsymbol{\theta}}) - \nabla \tilde{S}(\boldsymbol{\theta}_0) + \nabla \tilde{S}(\boldsymbol{\theta}_0) \\ &= \nabla^2 S_0(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \frac{1}{n}\sum_{t=1}^{n}\left[\frac{p(\mathbf{X}_t)}{Z(\mathbf{X}_t;\boldsymbol{\theta}_0)} - \frac{1-p(\mathbf{X}_t)}{1-Z(\mathbf{X}_t;\boldsymbol{\theta}_0)}\right]\nabla Z(\mathbf{X}_t;\boldsymbol{\theta}_0) + R_1 \end{aligned} \tag{3.38}$$

where

$$\begin{aligned} R_1 &= \nabla \tilde{S}(\tilde{\boldsymbol{\theta}}) - \nabla \tilde{S}(\boldsymbol{\theta}_0) - \nabla^2 \tilde{S}(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \\ &\quad [\nabla^2 \tilde{S}(\boldsymbol{\theta}_0) - \nabla^2 S_0(\boldsymbol{\theta}_0)](\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &= o_p\left(\left\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|\right) \end{aligned} \tag{3.39}$$

By (A_1) and (A_4) analogously to the argument in the proof of theorem of Franke and Neumann [24],

$$\begin{aligned} 0 = \nabla S_0(\boldsymbol{\theta}_0) &= -E\left[Y_1 \nabla \ln Z(\mathbf{X}_1;\boldsymbol{\theta}_0) + (1-Y_1)\nabla \ln(1 - Z(\mathbf{X}_1;\boldsymbol{\theta}_0))\right] \\ &= -E\left[\frac{p(\mathbf{X}_1)}{Z(\mathbf{X}_1;\boldsymbol{\theta}_0)} - \frac{1-p(\mathbf{X}_1)}{1-Z(\mathbf{X}_1;\boldsymbol{\theta}_0)}\right]\nabla Z(\mathbf{X}_1;\boldsymbol{\theta}_0) \end{aligned} \tag{3.40}$$

we get from the central limit theorem that the middle term in equation (3.38) is of order $O_p(n^{-1/2})$. Here as well as in exchanging expectation and differentiation, we have used that $Z(\mathbf{x};\boldsymbol{\theta})$ is bounded and bounded away from 0 uniformly in $\mathbf{x} \in \Re^d, \boldsymbol{\theta} \in \boldsymbol{\Theta}$ by (A_3). Therefore, the logarithms in $S(\boldsymbol{\theta}), \tilde{S}(\boldsymbol{\theta})$ and $S_0(\boldsymbol{\theta})$ do not cause problems. So, we have

$$\nabla^2 S_0(\boldsymbol{\theta}_0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p\left(\left\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|\right) = O_p(n^{-1/2}),$$

which implies, as $\nabla^2 S_0(\boldsymbol{\theta}_0)$ is positive definite by (A_2)

$$\left\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\| = O_p(n^{-1/2})$$

inserting this into equation (3.38), we get, writing $A(\boldsymbol{\theta}_0)$ for $\nabla^2 S_0(\boldsymbol{\theta}_0)$

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = A^{-1}(\boldsymbol{\theta}_0)\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left[\frac{p(\mathbf{X}_t)}{Z(\mathbf{X}_t;\boldsymbol{\theta}_0)} - \frac{1-p(\mathbf{X}_t)}{1-Z(\mathbf{X}_t;\boldsymbol{\theta}_0)}\right]\nabla Z(\mathbf{X}_t;\boldsymbol{\theta}_0) + o_p(1) \tag{3.41}$$

(c) Now, recall that we may write our data as

$$Y_t = p(\mathbf{X}_t) + \epsilon_t,\ E[\epsilon_t|\mathbf{X}_t] = 0$$

We have as in (b) with probability tending to 1, noting that in (3.36),

$$\gamma(Y_j, \mathbf{X}_j; \boldsymbol{\theta}_0) = -\epsilon_j \ln \frac{Z(\mathbf{X}_j; \boldsymbol{\theta})}{1 - Z(\mathbf{X}_j; \boldsymbol{\theta})},$$

$$\begin{aligned} 0 = \nabla S(\hat{\boldsymbol{\theta}}) &= \nabla \tilde{S}(\hat{\boldsymbol{\theta}}) + \nabla[S(\hat{\boldsymbol{\theta}}) - \tilde{S}(\hat{\boldsymbol{\theta}})] \\ &= \nabla \tilde{S}(\hat{\boldsymbol{\theta}}) + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \nabla \gamma(Y_t, \mathbf{X}_t; \hat{\boldsymbol{\theta}}) \\ &= \nabla \tilde{S}(\hat{\boldsymbol{\theta}}) - \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \epsilon_t \frac{1}{Z(\mathbf{X}_t; \hat{\boldsymbol{\theta}})(1 - Z(\mathbf{X}_t; \hat{\boldsymbol{\theta}}))} \nabla Z(\mathbf{X}_t; \hat{\boldsymbol{\theta}}) \end{aligned} \tag{3.42}$$

As in part (b) above, compare also the similar argument in the proof of Theorem 1 of Franke and Neumann [24], part (iii), we get

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) = A^{-1}(\boldsymbol{\theta}_0) \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left[\frac{\nabla Z(\mathbf{X}_t; \boldsymbol{\theta}_0)}{Z(\mathbf{X}_t; \boldsymbol{\theta}_0)(1 - Z(\mathbf{X}_t; \boldsymbol{\theta}_0))} \epsilon_t \right] + o_p(1) \tag{3.43}$$

We get from equations (3.41) and (3.43) for suitable functions ς_1, ς_2 satisfying $E\varsigma_1(\mathbf{X}_t) = 0,\ E\varsigma_2(\mathbf{X}_t)\epsilon_t = 0$

$$\begin{aligned} \sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varsigma_1(\mathbf{X}_t) + o_p(1) \\ \sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varsigma_2(\mathbf{X}_t)\epsilon_t + o_p(1) \end{aligned} \tag{3.44}$$

where, by (A_3) and by the bound already used in (a), for $\nabla \|Z(\mathbf{x}; \boldsymbol{\theta})\|$,

$$\|\varsigma_1(\mathbf{x})\| \leq c_1 \|\mathbf{x}\|,\ \|\varsigma_2(\mathbf{x})\| \leq c_2 \|\mathbf{x}\|$$

for some constants $c_1, c_2 > 0$ and all $\mathbf{x} \in \Re^d$. As $E\|\mathbf{X}_t\|^2 < \infty$ and ϵ_t is bounded, and as $(\mathbf{X}_t, \epsilon_t)$ are i.i.d., we conclude by a multivariate central limit theorem that

$$\sqrt{n} \begin{pmatrix} \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} \end{pmatrix} \to N\left(0, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\right) \tag{3.45}$$

as for all k, l,

$$\begin{aligned} n \text{ cov}(\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0k},\ \hat{\boldsymbol{\theta}}_l - \tilde{\boldsymbol{\theta}}_l) &= \frac{1}{n}\sum_{t,\tau=1}^{n} E\left[\varsigma_{1k}(\mathbf{X}_t)\varsigma_{2l}(\mathbf{X}_\tau)\epsilon_\tau\right] + o_p(1) \\ &= \frac{1}{n}\sum_{t\neq\tau}^{n} E\left[\varsigma_{1k}(\mathbf{X}_t)\varsigma_{2l}(\mathbf{X}_\tau)\epsilon_\tau\right] + \\ &\quad \frac{1}{n}\sum_{t=1}^{n} E\left[\varsigma_{1k}(\mathbf{X}_t)\varsigma_{2l}(\mathbf{X}_t)\epsilon_t\right] + o_p(1) \\ &= 0 + o_p(1) \end{aligned} \tag{3.46}$$

using $E(\epsilon_\tau|\mathbf{X}_\tau) = 0$

(d) It remains to derive the form of Σ_1, Σ_2. We have, as $\varsigma_1(\mathbf{X}_t), t = 1, \ldots, n$, are i.i.d. and $E[\varsigma_1(\mathbf{X}_t)] = 0$

$$\begin{aligned} \Sigma_1 &= E\varsigma_1(\mathbf{X}_1)\varsigma_1^T(\mathbf{X}_1) \\ &= A^{-1}(\boldsymbol{\theta}_0)B_1(\boldsymbol{\theta}_0)A^{-1}(\boldsymbol{\theta}_0) \end{aligned} \tag{3.47}$$

where

$$B_1(\boldsymbol{\theta}_0) = E\left[\frac{p(\mathbf{X}_1)}{Z(\mathbf{X}_1;\boldsymbol{\theta}_0)} - \frac{1-p(\mathbf{X}_1)}{1-Z(\mathbf{X}_1;\boldsymbol{\theta}_0)}\right]^2 \nabla Z(\mathbf{X}_1;\boldsymbol{\theta}_0)\nabla^T Z(\mathbf{X}_1;\boldsymbol{\theta}_0)$$

and, correspondingly, using equation (3.31) ,that $E[\epsilon_t^2|\mathbf{X}_1] = p(\mathbf{X}_1)(1-p(\mathbf{X}_1))$

$$\begin{aligned} \Sigma_2 &= E\varsigma_2(\mathbf{X}_1)\varsigma_2^T(\mathbf{X}_1)\epsilon_1^2 \\ &= E\varsigma_2(\mathbf{X}_1)\varsigma_2^T(\mathbf{X}_1)p(\mathbf{X}_1)(1-p(\mathbf{X}_1)) \\ &= A^{-1}(\boldsymbol{\theta}_0)B_2(\boldsymbol{\theta}_0)A^{-1}(\boldsymbol{\theta}_0) \end{aligned} \tag{3.48}$$

where

$$B_2(\boldsymbol{\theta}_0) = E\left[\frac{p(\mathbf{X}_1)(1-p(\mathbf{X}_1))}{Z^2(\mathbf{X}_1;\boldsymbol{\theta}_0)(1-Z(\mathbf{X}_1;\boldsymbol{\theta}_0))^2}\nabla Z(\mathbf{X}_1;\boldsymbol{\theta}_0)\nabla^T Z(\mathbf{X}_1;\boldsymbol{\theta}_0)\right]$$

We finish by discussing the assumptions. (A_1) is usually satisfied, in particular for the unipolar and bipolar ψ. (A_2) is a standard assumption if dealing with nonlinear regression settings. (A_3) is a rather weak assumption on $\boldsymbol{\Theta}$ as we always have for $Z(\mathbf{x};\boldsymbol{\theta}) = \psi(O_H(\mathbf{x};\boldsymbol{\theta}))$ and unipolar ψ that $0 < Z(\mathbf{x};\boldsymbol{\theta}) < 1$ for all $\mathbf{x} \in \Re^d,\ \boldsymbol{\theta} \in \Re^M$. Condition (A_4) is standard, and condition (A_5) guarantees that the Bernoulli experiments do not become degenerate.

□

Having discussed the necessary theory, we now address the problem of change point detection which is based on testing the hypotheses in equation (3.1).

3.7 Testing for Change-Points

We first consider the following standard change point problem. Assume that the data $(Y_i, \mathbf{X}_i)$ are independent and the conditional distribution of Y_i given $\mathbf{X}_i = \mathbf{x}$ is $\mathcal{B}(1, p(\mathbf{x}))$. Then we consider the change point testing problem

$$\begin{aligned} H_0 &: p_i(\mathbf{x}) = p_0(\mathbf{x}),\ i = 1, \ldots, n \\ vs & \\ H_1 &: \text{for some } i \leq K,\ p_i(\mathbf{x}) = p_0(\mathbf{x}),\ \text{and for } i \geq K+1,\ p_i(\mathbf{x}) = p(\mathbf{x}) \\ &\quad \text{where } 1 \leq K \leq n-1 \text{ is unknown, and } p_0 \neq p \end{aligned}$$

The general form of the likelihood function is, as $Y_i \in \{0, 1\}$

$$L = \Pi_{i=1}^{n} [p_i(\mathbf{x})]^{Y_i} [1 - p_i(\mathbf{x})]^{1-Y_i}$$

As before, we approximate our model by a parametric one by replacing $p_o(\mathbf{x})$ by $Z(\mathbf{x}; \boldsymbol{\theta})$ and $p(\mathbf{x})$ by $Z(\mathbf{x}; \boldsymbol{\theta}^*)$ where

$$Z(\mathbf{x}; \boldsymbol{\theta}) = \psi(O_H(\mathbf{x}; \boldsymbol{\theta})),\ Z(\mathbf{x}; \boldsymbol{\theta}^*) = \psi(O_H(\mathbf{x}; \boldsymbol{\theta}^*))$$

ψ being the unipolar or logistic function as before, and

$$O_H(\mathbf{x}; \boldsymbol{\theta}) = \alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j),$$

$$O_H(\mathbf{x}; \boldsymbol{\theta}^*) = \alpha_0^* + \sum_{h=1}^{H} \alpha_h^* \psi(W_{h0}^* + \sum_{j=1}^{d} W_{hj}^* x_j),$$

where $\boldsymbol{\theta}, \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, a suitably close compact subset of $\Re^D$, $D = 1 + (d+2)H$.

For the moment, we assume that a change can happen only after time K, where $1 \leq K \leq n-1$. Then, the likelihood functions under H_0 and H_1 are given, using $1 - \psi(u) = \psi(-u)$, by

$$\begin{aligned} L_0(\boldsymbol{\theta}) &= \Pi_{i=1}^{n} [\psi(O_H(\mathbf{X}_i; \boldsymbol{\theta}))]^{Y_i} [1 - \psi(O_H(\mathbf{X}_i; \boldsymbol{\theta}))]^{1-Y_i} \\ &= \Pi_{i=1}^{n} [\psi(O_H(\mathbf{X}_i; \boldsymbol{\theta}))]^{Y_i} [\psi(-O_H(\mathbf{X}_i; \boldsymbol{\theta}))]^{1-Y_i} \end{aligned}$$

$$
\begin{aligned}
L_{1,K}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \Pi_{i=1}^{K}[\psi(O_H(\mathbf{X}_i;\boldsymbol{\theta}))]^{Y_i}[\psi(-O_H(\mathbf{X}_i;\boldsymbol{\theta}))]^{1-Y_i} \\
&\quad *\Pi_{i=K+1}^{n}[\psi(O_H(\mathbf{X}_i;\boldsymbol{\theta}^*))]^{Y_i}[\psi(-O_H(\mathbf{X}_i;\boldsymbol{\theta}^*))]^{1-Y_i}
\end{aligned}
$$

Let

$$\hat{\boldsymbol{\theta}}_0 = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_0(\boldsymbol{\theta}),$$

$$(\hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}_K^*) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\theta}^*\in\boldsymbol{\Theta}} L_{1,K}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$$

denote the maximum likelihood estimates under H_0, which does not depend on K, and under H_1. We use the following notation for the network weights corresponding to $\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}_K^*$:

$$\hat{\boldsymbol{\theta}}_0 = (\hat{\alpha}_0^0, \ldots, \hat{\alpha}_H^0, \hat{W}_{10}^0, \ldots, \hat{W}_{Hd}^0),$$

$$\hat{\boldsymbol{\theta}}_K = (\hat{\alpha}_0^K, \ldots, \hat{\alpha}_H^K, \hat{W}_{10}^K, \ldots, \hat{W}_{Hd}^K),$$

$$\hat{\boldsymbol{\theta}}_K^* = (\hat{\alpha}_0^{*K}, \ldots, \hat{\alpha}_H^{*K}, \hat{W}_{10}^{*K}, \ldots, \hat{W}_{Hd}^{*K}),$$

For testing for a fixed changepoint at location K, the likelihood ratio statistic is

$$\Lambda_K^n = \frac{L_0(\hat{\boldsymbol{\theta}}_0)}{L_{1,K}(\hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}_K^*)} \tag{3.49}$$

If K is not fixed and is unknown, we follow the approach of Gombay and Horvarth [30], and reject H_o iff

$$Q_n = \max_{1\le K\le n-1}(-2\ \log\Lambda_K^n) \tag{3.50}$$

is large.

3.8 Limit Distribution of the Change-Point Test Statistic

The main objective of this section is to determine the asymptotic null distribution of Q_n in equation (3.50) above. We want to use the approach of Gombay and Horvath [30] as far as possible. They consider independent data, so we have to consider the pairs $S_j = (Y_j, \mathbf{X}_j),\ j = 1, \ldots, n$, as our

original sample. We need densities w.r.t. some δ-finite measure. For that purpose, we assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. with distribution μ which is a δ-finite measure. Let $\nu = \delta_0 + \delta_1$, δ_0, δ_1 denoting point masses in 0,1, be the counting measure in $\{0, 1\}$. Then, $(Y_i, \mathbf{X}_i)$ has a density π_i w.r.t. the product measure $\nu \otimes \mu$. For $p_i(\mathbf{x}) = P(Y_i = 1 | \mathbf{X}_i = \mathbf{x})$ as usual, we have

$$\pi_i(Y, \mathbf{x}) = \begin{cases} p_i(\mathbf{x}) & , \text{for } Y = 1 \\ 1 - p_i(\mathbf{x}) & , \text{for } Y = 0 \end{cases}$$

such that we have for $Y \in \{0, 1\}$, $B \in \Re^d$ (B an arbitrary Borel set),

$$P(Y_i = Y, \mathbf{X}_i \in B) \;\; = \;\; \int_{\{Y\}} \int_B \pi_i(\mu, \nu) \mathrm{d}\nu(\mu) \otimes \mathrm{d}\mu(\nu)$$

First, we neglect the possibility of misspecification and we assume

$$p_i(\mathbf{x}) = Z(\mathbf{x}; \boldsymbol{\theta}_i), \;\; \boldsymbol{\theta}_i \in \boldsymbol{\Theta}, \;\; i = 1, \ldots, n,$$

for a neural network output function as in the previous section. Then, the changepoint problem has the form of testing

$$\begin{array}{rcl} H_0 & : & \boldsymbol{\theta}_1 = \ldots = \boldsymbol{\theta}_n \\ \text{vs} & & \\ H_1 & : & \boldsymbol{\theta}_1 = \ldots = \boldsymbol{\theta}_K \neq \boldsymbol{\theta}_{K+1} = \ldots = \boldsymbol{\theta}_n \\ & & \text{for some } 1 \leq K \leq n-1 \end{array}$$

as in Gombay and Horvath [30]. So, we may apply their results if we can show that their conditions are satisfied. We follow their enumeration. Mark that the density of $(Y_i, \mathbf{X}_i)$ is now of the form

$$\pi_i(Y, \mathbf{x}; \boldsymbol{\theta}_i) = \begin{cases} Z(\mathbf{x}; \boldsymbol{\theta}_i) & , \text{for } Y = 1 \\ 1 - Z(\mathbf{x}; \boldsymbol{\theta}_i) & , \text{for } Y = 0 \end{cases}$$

(C.1.) If $\boldsymbol{\theta} \neq \boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, the densities $\pi(Y, \mathbf{x}; \boldsymbol{\theta})$, $\pi(Y, \mathbf{x}; \boldsymbol{\theta}^*)$ do not coincide.

Proof. It is obvious that $\boldsymbol{\theta}$ is identifiable from the function $\pi(Y, \mathbf{x}; \boldsymbol{\theta})$ if it is identifiable from $Z(\mathbf{x}; \boldsymbol{\theta})$. We have given conditions for that identifiability in section (3.5). It suffices to assume that $\theta = (\alpha_1, \ldots, \alpha_H, W_{10}, \ldots, W_{Hd})$ satisfies, with $\mathbf{W}_h = (W_{h1}, \ldots, W_{hd}), h = 1, \ldots, H$,

(A.I) $\alpha_h > 0, h = 1, \ldots, H$.

(A.II) $\mathbf{W}_h > \mathbf{0}, h = 1, \ldots, H.$

(A.III) $(\mathbf{W}_h, W_{h0}) \neq (\mathbf{W}_{h'}, W_{h'0})$ for some $h \neq h'$.

□

To state and proof the other conditions, we introduce the following notations;

$$\begin{aligned}
\lambda(Y, \mathbf{x}; \boldsymbol{\theta}) &= \log \pi(Y, \mathbf{x}; \boldsymbol{\theta}) = \ln \left[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y) \left(1 - Z(\mathbf{x}; \boldsymbol{\theta})\right)\right] \\
\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_i} \log \pi(Y, \mathbf{x}; \boldsymbol{\theta}) \\
\lambda_{ij}(Y, \mathbf{x}; \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \log \pi(Y, \mathbf{x}; \boldsymbol{\theta}) \\
\lambda_{ijk}(Y, \mathbf{x}; \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_k} \log \pi(Y, \mathbf{x}; \boldsymbol{\theta}) \\
Z_i(\mathbf{x}; \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_i} Z(\mathbf{x}; \boldsymbol{\theta})
\end{aligned} \tag{3.51}$$

So that

$$\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta}) = \frac{\left(\delta_1(Y) - \delta_0(Y)\right) Z_i(\mathbf{x}; \boldsymbol{\theta})}{\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y) \left(1 - Z(\mathbf{x}; \boldsymbol{\theta})\right)} \tag{3.52}$$

(C.2.) Uniqueness of solutions $\hat{\boldsymbol{\theta}}_K$ and $\hat{\boldsymbol{\theta}}_K^*$,$K = 1, \ldots, n$, of

$$\sum_{1 \leq j \leq K} \lambda_i(Y_j, \mathbf{X}_j; \hat{\boldsymbol{\theta}}_K) = 0, \ i = 1, \ldots, D \tag{3.53}$$

and

$$\sum_{K < j \leq n} \lambda_i(Y_j, \mathbf{X}_j; \hat{\boldsymbol{\theta}}_K^*) = 0, \ i = 1, \ldots, D \tag{3.54}$$

where $\hat{\boldsymbol{\theta}}_K$ and $\hat{\boldsymbol{\theta}}_K^*$ maximize the log-likelihood function

$$\ln L_{1,K}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \sum_{i=1}^{K} \log \pi(Y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_K) + \sum_{i=K+1}^{n} \log \pi(Y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_K^*) \tag{3.55}$$

Proof. As usual for estimating parameters of neural networks, we have to assume that the parameter set $\boldsymbol{\Theta}$ is chosen appropriatly such that there are unique $\hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}_K^*$ solving (3.53), (3.54) or equivalently:

(A.IV) There are unique $\hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}_K^* \in \boldsymbol{\Theta}$ maximizing $\ln L_{1,K}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ given by (3.55).

We remark that $K = n$ corresponds to the case of no change point, i.e., then $\hat{\boldsymbol{\theta}}_n$ is the estimate of the parameter under H_0, and $\hat{\boldsymbol{\theta}}_n^*$ is not defined and not used at all. □

Let $\hat{\boldsymbol{\theta}}_0$ denote the true parameter value under H_0, i.e. $\hat{\boldsymbol{\theta}}_0$ maximizes

$$E\lambda_i(Y_1, \mathbf{X}_1; \boldsymbol{\theta}) = \frac{1}{n} E \ln L_{1,K}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \frac{1}{n} E L_0(\boldsymbol{\theta})$$

(C.3.) There is an open subset $\boldsymbol{\Theta} \subseteq \Re^D$ containing $\boldsymbol{\theta}_o$ such that $\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta})$, $\lambda_{ij}(Y, \mathbf{x}; \boldsymbol{\theta})$ and $\lambda_{ijk}(Y, \mathbf{x}; \boldsymbol{\theta})$, $1 \leq i, j, k \leq D$ exist and are continuous in $\boldsymbol{\theta}$ for all $Y \in \{0, 1\}$,$\mathbf{x} \in \Re^d$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Proof. This condition is satisfied by the fact that $\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta})$, $\lambda_{ij}(Y, \mathbf{x}; \boldsymbol{\theta})$ and $\lambda_{ijk}(Y, \mathbf{x}; \boldsymbol{\theta})$ are depending on $\boldsymbol{\theta}$ only through $Z(\mathbf{x}; \boldsymbol{\theta})$ which is continuously differentiable with respect to $\boldsymbol{\theta}$ infinitely often. □

(C.4.) There are functions $M_1(Y, \mathbf{x})$ and $M_2(Y, \mathbf{x})$ such that $|\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta})| \leq M_1(Y, \mathbf{x})$, $|\lambda_{ij}(Y, \mathbf{x}; \boldsymbol{\theta})| \leq M_2(Y, \mathbf{x})$ and $|\lambda_{ijk}(Y, \mathbf{x}; \boldsymbol{\theta})| \leq M_2(Y, \mathbf{x})$ for all $\mathbf{x} \in \Re^d$, $Y \in \{0, 1\}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $1 \leq i, j, k \leq D$. where M_1, M_2 satisfy

$$\sum_{Y=0}^{1} \int M_1(Y, \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) < \infty,$$

$$E_{\boldsymbol{\theta}_0} M_2(Y_1, \mathbf{X}_1) = \int [M_2(1, \mathbf{x}) Z(\mathbf{x}; \boldsymbol{\theta}_0) + M_2(0, \mathbf{x})(1 - Z(\mathbf{x}; \boldsymbol{\theta}_0))] \mathrm{d}\mu(\mathbf{x}) < \infty,$$

As $0 \leq Z(\mathbf{x}; \boldsymbol{\theta}_0) \leq 1$, the latter condition is satisfied if

$$\sum_{Y=0}^{1} \int M_2(Y, \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) < \infty.$$

Proof. **Part I** Recall

$$\pi(Y, \mathbf{x}; \boldsymbol{\theta}) = \delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta})) \tag{3.56}$$

where

$$Z(\mathbf{x};\boldsymbol{\theta}) = \psi(O_H(\mathbf{x};\boldsymbol{\theta}))$$

and

$$O_H(\mathbf{x};\boldsymbol{\theta}) = \alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(W_{h0} + \sum_{j=1}^{d} W_{hj} x_j) \tag{3.57}$$

Then,

$$\begin{aligned} |\lambda_i(Y,\mathbf{x};\boldsymbol{\theta})| &= \frac{\partial}{\partial \boldsymbol{\theta}_i} \log \pi(Y,\mathbf{x};\boldsymbol{\theta}) \\ &= \frac{|\delta_1(Y) Z_i(\mathbf{x};\boldsymbol{\theta})| + |\delta_0(Y) Z_i(\mathbf{x};\boldsymbol{\theta})|}{\delta_1(Y) Z(\mathbf{x};\boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x};\boldsymbol{\theta}))} \\ &= \frac{|Z_i(\mathbf{x};\boldsymbol{\theta})|}{\delta_1(Y) Z(\mathbf{x};\boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x};\boldsymbol{\theta}))} \\ &= \begin{cases} \frac{|Z_i(\mathbf{x};\boldsymbol{\theta})|}{Z(\mathbf{x};\boldsymbol{\theta})}, Y = 1 \\ \frac{|Z_i(\mathbf{x};\boldsymbol{\theta})|}{1 - Z(\mathbf{x};\boldsymbol{\theta})}, Y = 0 \end{cases} \end{aligned} \tag{3.58}$$

Since $Z(\mathbf{x};\boldsymbol{\theta}) = \psi(O_H(\mathbf{x};\boldsymbol{\theta}))$, then

$$\begin{aligned} Z_i(\mathbf{x};\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_i} \psi(O_H(\mathbf{x};\boldsymbol{\theta})) \\ &= \psi'(O_H(\mathbf{x};\boldsymbol{\theta})) \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x};\boldsymbol{\theta}) \end{aligned} \tag{3.59}$$

where

$$\begin{aligned} \psi(u) &= \frac{1}{1 + e^{-u}} \\ \text{so that} & \\ \psi'(u) &= \frac{e^{-u}}{(1 + e^{-u})^2} \\ &= \frac{(1 + e^{u}) e^{-u}}{(1 + e^{u})(1 + e^{-u})^2} \\ &= \frac{1 + e^{-u}}{(1 + e^{u})(1 + e^{-u})^2} \\ &= \frac{1}{(1 + e^{u})(1 + e^{-u})} \end{aligned} \tag{3.60}$$

From equation 3.60 above, we have

$$\frac{\left|\psi^{'}(u)\right|}{\psi(u)} = \frac{1}{1+e^{u}} \in [0,1] \tag{3.61}$$

and

$$\frac{\left|\psi^{'}(u)\right|}{1-\psi(u)} = \frac{1}{1+e^{-u}} \in [0,1] \tag{3.62}$$

For $Y = 1$ and using equation (3.61) above, we have

$$\begin{aligned} |\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta})| &= \frac{\left|\psi^{'}(O_H(\mathbf{x}; \boldsymbol{\theta}))\right|}{\psi(O_H(\mathbf{x}; \boldsymbol{\theta}))} \left| \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| \\ &\leq \left| \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| \end{aligned} \tag{3.63}$$

For $Y = 0$ and using equation (3.62) above, we have

$$\begin{aligned} |\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta})| &= \frac{\left|\psi^{'}(O_H(\mathbf{x}; \boldsymbol{\theta}))\right|}{1-\psi(O_H(\mathbf{x}; \boldsymbol{\theta}))} \left| \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| \\ &\leq \left| \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| \end{aligned} \tag{3.64}$$

Equations (3.63) and (3.64) lead to the conclusion that

$$|\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta})| \leq \left| \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| \tag{3.65}$$

For $\boldsymbol{\theta}_i$ corresponding to α_0;

$$\left| \frac{\partial}{\partial \alpha_0} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| = 1 \tag{3.66}$$

For $\boldsymbol{\theta}_i$ corresponding to $\alpha_i, i = 1, \ldots, H$;

$$\left| \frac{\partial}{\partial \alpha_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| = \left| \psi(W_{i0} + \sum_{j=1}^{d} W_{ij} x_j) \right| \leq 1 \tag{3.67}$$

For $\boldsymbol{\theta}_i$ corresponding to W_{h0};

$$\begin{aligned}
\left|\frac{\partial}{\partial W_{h0}} O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{h0}}\left[\alpha_0 + \sum_{h=1}^{H}\alpha_h\psi(W_{h0} + \sum_{j=1}^{d} W_{hj}x_j)\right]\right| \\
&= \left|\alpha_h\psi^{'}(W_{h0} + \sum_{j=1}^{d} W_{hj}x_j)\right| \\
&= \left|\frac{\alpha_h\psi(W_{h0} + \sum_{j=1}^{d} W_{hj}x_j)}{1 + e^{(W_{h0}+\sum_{j=1}^{d} W_{hj}x_j)}}\right| \\
&\leq \left|\alpha_h\psi(W_{h0} + \sum_{j=1}^{d} W_{hj}x_j)\right|, \text{since}\,|1-\psi(u)| \leq 1 \\
&\leq |\alpha_h|\,, \text{since}\,|\psi(u)| \leq 1 \qquad (3.68)
\end{aligned}$$

For $\boldsymbol{\theta}_i$ corresponding to W_{hl} for some l, calculations similar to equation (3.68) lead to

$$\left|\frac{\partial}{\partial W_{hl}} O_H(\mathbf{x};\boldsymbol{\theta})\right| = |\alpha_h|\,|x_l| \qquad (3.69)$$

Therefore, to get (C.4.), we use

$$M_1(0,\mathbf{x}) = M_1(1,\mathbf{x}) = M_1(\mathbf{x}) = \max(1, C.\sum_{l=1}^{d}|x_l|) \qquad (3.70)$$

with the assumption that

$$|\alpha_h| \leq \text{Const.}, h = 1,\ldots,H, \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \qquad (3.71)$$

and

$$E\,|\mathbf{X}_{l1}| < \infty, l = 1,\ldots,d. \text{ i.e } E\,\|\mathbf{X}_1\| < \infty \qquad (3.72)$$

Part II
We now deal with the second derivatives of $\lambda_{ij}(y,\mathbf{x};\boldsymbol{\theta})$

$$\begin{aligned}
\lambda_{ij}(Y,\mathbf{x};\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_j}\lambda_i(Y,\mathbf{x};\boldsymbol{\theta}) \\
&= \frac{\partial}{\partial \boldsymbol{\theta}_j}\frac{(\delta_0(Y)-\delta_1(Y))Z_i(\mathbf{x};\boldsymbol{\theta})}{\delta_1(Y)Z(\mathbf{x};\boldsymbol{\theta})+\delta_0(Y)(1-Z(\mathbf{x};\boldsymbol{\theta}))} \\
&= \frac{(\delta_0(Y)-\delta_1(Y))Z_{ij}(\mathbf{x};\boldsymbol{\theta})}{\delta_1(Y)Z(\mathbf{x};\boldsymbol{\theta})+\delta_0(Y)(1-Z(\mathbf{x};\boldsymbol{\theta}))} - \\
&\quad -\frac{(\delta_0(Y)-\delta_1(Y))^2 Z_i(\mathbf{x};\boldsymbol{\theta})Z_j(\mathbf{x};\boldsymbol{\theta})}{[\delta_1(Y)Z(\mathbf{x};\boldsymbol{\theta})+\delta_0(Y)(1-Z(\mathbf{x};\boldsymbol{\theta}))]^2} \qquad (3.73)
\end{aligned}$$

But

$$\begin{aligned}
\psi^{''}(u) &= \frac{d}{du}\psi^{'}(u) \\
&= \frac{-e^u}{(1+e^{-u})(1+e^u)^2}+\frac{e^{-u}}{(1+e^u)(1+e^{-u})^2} \\
&= \frac{e^{-u}-e^u}{(1+e^u)^2(1+e^{-u})^2} \\
&= \frac{(1+e^{-u})-(1+e^u)}{(1+e^u)^2(1+e^{-u})^2} \\
&= \frac{1}{(1+e^{-u})(1+e^u)^2}-\frac{1}{(1+e^u)(1+e^{-u})^2} \qquad (3.74)
\end{aligned}$$

From equation (3.74) above, we have

$$\begin{aligned}
\frac{\psi^{''}(u)}{\psi(u)} &= \frac{1}{(1+e^u)^2}-\frac{1}{(1+e^u)(1+e^{-u})} \\
&= (1-\psi(u))^2-\frac{\psi^{'}(u)}{\psi(u)} \in [-1,1] \qquad (3.75)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{\psi^{''}(u)}{1-\psi(u)} &= \frac{1}{(1+e^u)(1+e^{-u})}-\frac{1}{(1+e^{-u})^2}- \\
&= \frac{\psi^{'}(u)}{\psi(u)}-(1-\psi(u))^2 \in [-1,1] \qquad (3.76)
\end{aligned}$$

From equation (3.73);

$$\begin{aligned}
|\lambda_{ij}(Y,\mathbf{x};\boldsymbol{\theta})| &\leq \frac{|Z_{ij}(\mathbf{x};\boldsymbol{\theta})|}{\delta_1(Y)Z(\mathbf{x};\boldsymbol{\theta})+\delta_0(Y)(1-Z(\mathbf{x};\boldsymbol{\theta}))} \\
&\quad +\frac{|Z_i(\mathbf{x};\boldsymbol{\theta})|\,|Z_j(\mathbf{x};\boldsymbol{\theta})|}{[\delta_1(Y)Z(\mathbf{x};\boldsymbol{\theta})+\delta_0(Y)(1-Z(\mathbf{x};\boldsymbol{\theta}))]^2} \qquad (3.77)
\end{aligned}$$

since $|\delta_0(Y) - \delta_1(Y)| = 1$

For $Y = 1$,

$$|\lambda_{ij}(Y, \mathbf{x}; \boldsymbol{\theta})| \leq \frac{|Z_{ij}(\mathbf{x}; \boldsymbol{\theta})|}{Z(\mathbf{x}; \boldsymbol{\theta})} + |\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta})|\,|\lambda_j(Y, \mathbf{x}; \boldsymbol{\theta})| \quad (3.78)$$

where, from part I of the proof,

$$|\lambda_i(Y, \mathbf{x}; \boldsymbol{\theta})|\,|\lambda_j(Y, \mathbf{x}; \boldsymbol{\theta})| \leq \begin{cases} \text{Const.}\,|x_i|\,|x_j|\,, \text{for } \boldsymbol{\theta}_i = W_{hi}, \boldsymbol{\theta}_j = W_{hj}; i, j \geq 1 \\ \text{Const.}\,|x_i|\,, \text{for } \boldsymbol{\theta}_i = \alpha_i, \boldsymbol{\theta}_j = W_{hj} \\ \text{Const., for } \boldsymbol{\theta}_i = \alpha_i, \boldsymbol{\theta}_j = \alpha_j \end{cases} \quad (3.79)$$

so that for $E[M_2(Y_1, \mathbf{X}_1)] < \infty$, we need at least $E[\mathbf{X}_{1l}^2] < \infty, l = 1, \ldots, d$, or $E\left\|\mathbf{X}_1\right\|^2 < \infty$

We now determine $Z_{ij}(\mathbf{x}; \boldsymbol{\theta})$.

$$\begin{aligned} Z_{ij}(\mathbf{x}; \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_j} Z_i(\mathbf{x}; \boldsymbol{\theta}) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}_j} \left\{ \psi^{'}(O_H(\mathbf{x}; \boldsymbol{\theta})) \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \right\} \\ &= \psi^{''}(O_H(\mathbf{x}; \boldsymbol{\theta})) \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}_j} O_H(\mathbf{x}; \boldsymbol{\theta}) \\ &\quad + \psi^{'}(O_H(\mathbf{x}; \boldsymbol{\theta})) \frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} O_H(\mathbf{x}; \boldsymbol{\theta}) \end{aligned} \quad (3.80)$$

so that

$$\frac{|Z_{ij}(\mathbf{x}; \boldsymbol{\theta})|}{Z(\mathbf{x}; \boldsymbol{\theta})} \leq 1 * \left| \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}_j} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| + 1 * \left| \frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| \quad (3.81)$$

since from (3.61) and (3.75), $\frac{\left|\psi^{'}(u)\right|}{\psi(u)} \leq 1$ and $\frac{\left|\psi^{''}(u)\right|}{\psi(u)} \leq 1$ respectively.

Analogous results in (3.81) follow for the case $Y = 0$. That is,

$$\frac{|Z_{ij}(\mathbf{x}; \boldsymbol{\theta})|}{1 - Z(\mathbf{x}; \boldsymbol{\theta})} \leq 1 * \left| \frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x}; \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}_j} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| + 1 * \left| \frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} O_H(\mathbf{x}; \boldsymbol{\theta}) \right| \quad (3.82)$$

where $\frac{\left|\psi^{'}(u)\right|}{1-\psi(u)} \leq 1$ from equation (3.62) and $\frac{\left|\psi^{''}(u)\right|}{1-\psi(u)} \leq 1$ from (3.76).

Then,

$$|\lambda_{ij}(Y,\mathbf{x};\boldsymbol{\theta})| \leq \left|\frac{\partial}{\partial\boldsymbol{\theta}_i}O_H(\mathbf{x};\boldsymbol{\theta})\right|\left|\frac{\partial}{\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| + \left|\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| \tag{3.83}$$

where

$$\left|\frac{\partial}{\partial\boldsymbol{\theta}_i}O_H(\mathbf{x};\boldsymbol{\theta})\right|\left|\frac{\partial}{\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| \leq M_1^2(\mathbf{x}) \tag{3.84}$$

by (C.4.) part (I). We now deal with the second derivatives of $O_H(\mathbf{x};\boldsymbol{\theta})$. For $\boldsymbol{\theta}_i = \alpha_i$, $\boldsymbol{\theta}_j = \alpha_j$;

$$\left|\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| = 0 \tag{3.85}$$

For $\boldsymbol{\theta}_i = \alpha_i$, $\boldsymbol{\theta}_j = W_{hl}, i \neq h$;

$$\left|\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| = 0 \tag{3.86}$$

For $\boldsymbol{\theta}_i = \alpha_h$, $\boldsymbol{\theta}_j = W_{h0}$;

$$\begin{aligned}\left|\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{h0}}\psi(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \\ &= \left|\psi^{'}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \leq 1 \text{ , by equation 3.61}\end{aligned} \tag{3.87}$$

For $\boldsymbol{\theta}_i = \alpha_h$, $\boldsymbol{\theta}_j = W_{hl}, l > 0$;

$$\begin{aligned}\left|\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{hl}}\psi(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \\ &= \left|\psi^{'}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right||x_l| \\ &\leq |x_l| \text{ , by equation 3.87 above}\end{aligned} \tag{3.88}$$

For $\boldsymbol{\theta}_i = W_{hl}$, $\boldsymbol{\theta}_j = W_{h'l'}, h \neq h'$;

$$\begin{aligned}\left|\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{h'l'}}\alpha_h\psi(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \\ &= 0 \end{aligned} \tag{3.89}$$

For $\boldsymbol{\theta}_i = W_{h0}$, $\boldsymbol{\theta}_j = W_{h0}$;

$$\begin{aligned}\left|\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{h0}}\alpha_h\psi^{'}(W_{h0}+\sum_{j=1}^{d}W_{hj}x_j)\right| \\ &= \left|\alpha_h\psi^{''}(W_{h0}+\sum_{j=1}^{d}W_{hj}x_j)\right| \\ &\leq |\alpha_h| \end{aligned} \tag{3.90}$$

from equation (3.75).
For $\boldsymbol{\theta}_i = W_{h0}$, $\boldsymbol{\theta}_j = W_{hl}, l \geq 1$;

$$\begin{aligned}\left|\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{hl}}\alpha_h\psi^{'}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \\ &= \left|\alpha_h\psi^{''}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)x_l\right| \\ &\leq |\alpha_h|\,|x_l| \end{aligned} \tag{3.91}$$

For $\boldsymbol{\theta}_i = W_{hl}$, $\boldsymbol{\theta}_j = W_{hm}, l \geq 1$;

$$\begin{aligned}\left|\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{hm}}\alpha_h\psi^{'}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)x_l\right| \\ &= \left|\alpha_h\psi^{''}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)x_lx_m\right| \\ &\leq |\alpha_h|\,|x_lx_m| \end{aligned} \tag{3.92}$$

Therefore, for the existence of an appropriate bound, we need $|\alpha_1|, \ldots, |\alpha_H|$ bounded and $E\left\|\mathbf{X}_i\right\|^2 < \infty$

Part III
Finally, we determine the bound for $\lambda_{ijk}(Y, \mathbf{x}; \boldsymbol{\theta})$

$$\begin{aligned}
\lambda_{ijk}(Y, \mathbf{x}; \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_k} \lambda_{ij}(Y, \mathbf{x}; \boldsymbol{\theta}) \\
&= \frac{\partial}{\partial \boldsymbol{\theta}_k} \left\{ \frac{(\delta_0(Y) - \delta_1(Y)) Z_{ij}(\mathbf{x}; \boldsymbol{\theta})}{\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))} \right. \\
&\quad \left. - \frac{(\delta_0(Y) - \delta_1(Y))^2 Z_i(\mathbf{x}; \boldsymbol{\theta}) Z_j(\mathbf{x}; \boldsymbol{\theta})}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]^2} \right\} \\
&= \frac{(\delta_0(Y) - \delta_1(Y)) Z_{ijk}(\mathbf{x}; \boldsymbol{\theta})}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]} \\
&\quad - \frac{(\delta_0(Y) - \delta_1(y)) Z_{ij}(\mathbf{x}; \boldsymbol{\theta})(\delta_1(Y) Z_k(\mathbf{x}; \boldsymbol{\theta}) - \delta_0(Y) Z_k(\mathbf{x}; \boldsymbol{\theta}))}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]^2} \\
&\quad - \frac{(\delta_0(Y) - \delta_1(Y))^2 [Z_{ik}(\mathbf{x}; \boldsymbol{\theta}) Z_j(\mathbf{x}; \boldsymbol{\theta}) + Z_{jk}(\mathbf{x}, \boldsymbol{\theta}) Z_i(\mathbf{x}; \boldsymbol{\theta})]}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]^2} \\
&\quad - \frac{2(\delta_0(Y) - \delta_1(Y))^2 Z_i(\mathbf{x}; \boldsymbol{\theta}) Z_j(\mathbf{x}; \boldsymbol{\theta}) [\delta_1(Y) Z_k(\mathbf{x}; \boldsymbol{\theta}) - \delta_0(Y) Z_k(\mathbf{x}; \boldsymbol{\theta})]}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]^3}
\end{aligned} \tag{3.93}$$

Then, from equation (3.93) and the fact that $|\delta_0(Y) - \delta_1(Y)| = 1$, it follows that

$$\begin{aligned}
|\lambda_{ijk}(Y, \mathbf{x}; \boldsymbol{\theta})| &= \left| \frac{Z_{ijk}(\mathbf{x}; \boldsymbol{\theta})}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]} \right| \\
&\quad + \left| \frac{Z_{ij}(\mathbf{x}; \boldsymbol{\theta}) Z_k(\mathbf{x}; \boldsymbol{\theta})}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]^2} \right| \\
&\quad + \left| \frac{Z_{ik}(\mathbf{x}; \boldsymbol{\theta}) Z_j(\mathbf{x}; \boldsymbol{\theta})}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]^2} \right| \\
&\quad + \left| \frac{Z_{jk}(\mathbf{x}; \boldsymbol{\theta}) Z_i(\mathbf{x}; \boldsymbol{\theta})}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]^2} \right| \\
&\quad + \left| \frac{2 Z_i(\mathbf{x}; \boldsymbol{\theta}) Z_j(\mathbf{x}; \boldsymbol{\theta}) Z_k(\mathbf{x}; \boldsymbol{\theta})}{[\delta_1(Y) Z(\mathbf{x}; \boldsymbol{\theta}) + \delta_0(Y)(1 - Z(\mathbf{x}; \boldsymbol{\theta}))]^3} \right|
\end{aligned} \tag{3.94}$$

where each of the last four terms in equation (3.94) is bounded by $const. + const.|x_\alpha x_\beta x_\tau|$ for appropriate indices α, β, τ using (C.4) part (II).

We now determine $Z_{ijk}(\mathbf{x};\boldsymbol{\theta})$.

$$\begin{aligned}
Z_{ijk}(\mathbf{x};\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_k} Z_{ij}(\mathbf{x};\boldsymbol{\theta}) \\
&= \frac{\partial}{\partial \boldsymbol{\theta}_k}\left\{ \psi^{''}(O_H(\mathbf{x};\boldsymbol{\theta}))\frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x};\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\theta}_j} O_H(\mathbf{x};\boldsymbol{\theta}) \right. \\
&\quad \left. + \psi^{'}(O_H(\mathbf{x};\boldsymbol{\theta}))\frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} O_H(\mathbf{x};\boldsymbol{\theta}) \right\} \\
&= \psi^{'''}(O_H(\mathbf{x};\boldsymbol{\theta}))\frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x};\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\theta}_j} O_H(\mathbf{x};\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\theta}_k} O_H(\mathbf{x};\boldsymbol{\theta}) \\
&\quad + \psi^{''}(O_H(\mathbf{x};\boldsymbol{\theta}))\frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_k} O_H(\mathbf{x};\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\theta}_j} O_H(\mathbf{x};\boldsymbol{\theta}) \\
&\quad + \psi^{''}(O_H(\mathbf{x};\boldsymbol{\theta}))\frac{\partial^2}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_k} O_H(\mathbf{x};\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\theta}_i} O_H(\mathbf{x};\boldsymbol{\theta}) \\
&\quad + \psi^{''}(O_H(\mathbf{x};\boldsymbol{\theta}))\frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} O_H(\mathbf{x};\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\theta}_k} O_H(\mathbf{x};\boldsymbol{\theta}) \\
&\quad + \psi^{'}(O_H(\mathbf{x};\boldsymbol{\theta}))\frac{\partial^3}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_k} O_H(\mathbf{x};\boldsymbol{\theta})
\end{aligned} \tag{3.95}$$

But

$$\begin{aligned}
\psi^{'''}(u) &= \frac{d}{du}\psi^{''}(u) \\
&= \frac{d}{du}\left[\frac{1}{(1+e^{-u})(1+e^{u})^2} - \frac{1}{(1+e^{u})(1+e^{-u})^2} \right] \\
&= 2\left[\frac{1}{(1+e^{-u})^2(1+e^{u})} - \frac{1}{(1+e^{u})(1+e^{-u})^3} - \frac{1}{(1+e^{u})^3(1+e^{-u})} \right] \\
&= 2\left[\psi(u)\psi^{'}(u) - \psi^2(u)\psi^{'}(u) - (1-\psi(u))^2\psi^{'}(u) \right] \\
&= 2\psi^{'}(u)\left[-2\psi^2(u) + 3\psi(u) - 1 \right] \in [-2,2]
\end{aligned} \tag{3.96}$$

So that

$$\frac{\psi^{'''}(u)}{\psi(u)} = 2\frac{\psi^{'}(u)}{\psi(u)}\left[-2\psi^2(u) + 3\psi(u) - 1 \right] \in [-2,2] \tag{3.97}$$

and

$$\frac{\psi^{'''}(u)}{1-\psi(u)} = 2\frac{\psi^{'}(u)}{(1-\psi(u))}\left[-2\psi^2(u) + 3\psi(u) - 1 \right] \in [-2,2] \tag{3.98}$$

Then, for $Y = 1$ and using equation (3.95) and part II of the proof,,

$$\frac{|Z_{ijk}(\mathbf{x};\boldsymbol{\theta})|}{Z(\mathbf{x};\boldsymbol{\theta})} \leq const. + const.|x_\alpha x_\beta x_\tau| + \left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| \tag{3.99}$$

for suitable α, β, τ. Analogous result follows for $Y = 0$.
Therefore;

$$|\lambda_{ijk}(\mathbf{x},\boldsymbol{\theta})| \leq const. + const.|x_\alpha x_\beta x_\tau| + \left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| \tag{3.100}$$

It therefore remains to determine the bound of $\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})$
From the second derivatives of $O_H(\mathbf{x};\boldsymbol{\theta})$ in (C.4) part (II) above;
For $\boldsymbol{\theta}_i = \alpha_i$, $\boldsymbol{\theta}_j = \alpha_j$, $\boldsymbol{\theta}_k = \alpha_k$;

$$\left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| = 0 \tag{3.101}$$

For $\boldsymbol{\theta}_i = \alpha_i$, $\boldsymbol{\theta}_j = W_{hl}$, $\boldsymbol{\theta}_k = W_{hl}, i \neq h$;

$$\left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| = 0 \tag{3.102}$$

For $\boldsymbol{\theta}_i = \alpha_h$,$\boldsymbol{\theta}_j = W_{h0}$, $\boldsymbol{\theta}_k = W_{h0}$;

$$\begin{aligned}\left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{h0}}\psi^{'}(W_{h0} + \sum_{l=1}^{d} W_{hl}x_l)\right| \\ &= \left|\psi^{''}(W_{h0} + \sum_{l=1}^{d} W_{hl}x_l)\right| \\ &\leq 1 \end{aligned} \tag{3.103}$$

For $\boldsymbol{\theta}_i = \alpha_h$, $\boldsymbol{\theta}_j = W_{h0}$, $\boldsymbol{\theta}_j = W_{hl}, l > 0$;

$$\begin{aligned}\left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{hl}}\psi^{'}(W_{h0} + \sum_{l=1}^{d} W_{hl}x_l)\right| \\ &= \left|\psi^{''}(W_{h0} + \sum_{l=1}^{d} W_{hl}x_l)x_l\right| \\ &\leq |x_l| \end{aligned} \tag{3.104}$$

For $\boldsymbol{\theta}_i = \alpha_h$, $\boldsymbol{\theta}_j = W_{hl}$, $\boldsymbol{\theta}_k = W_{hl}$;

$$\begin{aligned}\left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{hl}}\psi^{'}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \\ &= \left|\psi^{''}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)x_lx_l\right| \\ &\leq |x_lx_l| \end{aligned} \tag{3.105}$$

For $\boldsymbol{\theta}_i = W_{h0}$, $\boldsymbol{\theta}_j = W_{h0}$, $\boldsymbol{\theta}_k = W_{h0}$;

$$\begin{aligned}\left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{h0}}\psi^{''}\alpha_h(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \\ &= \left|\alpha_h\psi^{'''}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \\ &\leq 2\,|\alpha_h| \text{ , by equation 3.96.} \end{aligned} \tag{3.106}$$

For $\boldsymbol{\theta}_i = W_{h0}$, $\boldsymbol{\theta}_j = W_{h0}$, $\boldsymbol{\theta}_k = W_{hl}, l \geq 1$;

$$\begin{aligned}\left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{hl}}\psi^{''}\alpha_h(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \\ &= \left|\alpha_h\psi^{'''}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)x_l\right| \\ &\leq 2\,|\alpha_h|\,|x_l| \end{aligned} \tag{3.107}$$

Lastly, for $\boldsymbol{\theta}_i = W_{hl}$, $\boldsymbol{\theta}_j = W_{ht}$,$\boldsymbol{\theta}_j = W_{hr}$, for $l, t, r \geq 1$;

$$\begin{aligned}\left|\frac{\partial^3}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_k}O_H(\mathbf{x};\boldsymbol{\theta})\right| &= \left|\frac{\partial}{\partial W_{hr}}\psi^{''}\alpha_h(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)\right| \\ &= \left|\alpha_h\psi^{'''}(W_{h0}+\sum_{l=1}^{d}W_{hl}x_l)x_lx_tx_r\right| \\ &\leq 2\,|\alpha_h|\,|x_lx_tx_r| \end{aligned} \tag{3.108}$$

Therefore, we have (C.4) with

$$M_2(1,\mathbf{x}) = M_2(0,\mathbf{x}) = M_2(\mathbf{x}) = C_1 + C_2\sum_{l,t,r}^{d}|x_lx_tx_r| \tag{3.109}$$

for large enough constants C_1, C_2.

To get the integrability conditions on M_1 and M_2, we remark that $|\alpha_1|, \ldots, |\alpha_h|$ are uniformly bounded in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ by the compactness of $\boldsymbol{\Theta}$. So, we only have to assume, as $M_1(\mathbf{x}) \leq const. + const. \|\mathbf{x}\|$, that $M_2(\mathbf{x})$ is integrable w.r.t. the distribution of $\mathbf{X}_t$ which follows from $E\|\mathbf{X}_1\| < \infty$ For a later condition, we even have to assume

(A.V) $E\|\mathbf{X}_1\|^4 < \infty$

□

(C.5.) $E_{\boldsymbol{\theta}}\lambda_i(Y_1, \mathbf{X}_1; \boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, for all i.

Proof.

$$\begin{aligned} E[\lambda_i(Y_1, \mathbf{X}_1; \boldsymbol{\theta})] &= E[E[\lambda_i(Y_1, \mathbf{X}_1; \boldsymbol{\theta})|\mathbf{X}_1]] \\ &= E[Z(\mathbf{X}_1; \boldsymbol{\theta})\lambda_i(1, \mathbf{X}_1, \boldsymbol{\theta}) + (1 - Z(\mathbf{X}_1; \boldsymbol{\theta}))\lambda_i(0, \mathbf{X}_1, \boldsymbol{\theta})] \\ &= E\left[Z(\mathbf{X}_1; \boldsymbol{\theta})\frac{Z_i(\mathbf{X}_1; \boldsymbol{\theta})}{Z(\mathbf{X}_1; \boldsymbol{\theta})} - (1 - Z(\mathbf{X}_1; \boldsymbol{\theta}))\frac{Z_i(\mathbf{X}_1; \boldsymbol{\theta})}{1 - Z(\mathbf{X}_1; \boldsymbol{\theta})}\right] \\ &= 0 \end{aligned} \tag{3.110}$$

□

(C.6.) $E_{\boldsymbol{\theta}}[\lambda_i(Y_1, \mathbf{X}_1; \boldsymbol{\theta})\lambda_j(Y_1, \mathbf{X}_1; \boldsymbol{\theta})] = -E_{\boldsymbol{\theta}}[\lambda_{ij}(Y_1, \mathbf{X}_1; \boldsymbol{\theta})] = I_{ij}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $I^{-1}(\boldsymbol{\theta})$ exists and $I(\boldsymbol{\theta}), I^{-1}(\boldsymbol{\theta})$ both are continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Proof.

$$\begin{aligned} E[\lambda_i(Y_1, \mathbf{X}_1; \boldsymbol{\theta})\lambda_j(Y_1, \mathbf{X}_1; \boldsymbol{\theta})] &= E[E[\lambda_i(Y_1, \mathbf{X}_1; \boldsymbol{\theta})|\mathbf{X}_1\lambda_j(Y_1, \mathbf{X}_1; \boldsymbol{\theta})|\mathbf{X}_1]] \\ &= E[Z(\mathbf{X}_1; \boldsymbol{\theta})\lambda_i(1, \mathbf{X}_1; \boldsymbol{\theta})\lambda_j(1, \mathbf{X}_1; \boldsymbol{\theta}) + \\ &\quad + (1 - Z(\mathbf{X}_1; \boldsymbol{\theta}))\lambda_i(0, \mathbf{X}_1; \boldsymbol{\theta})\lambda_j(0, \mathbf{X}_1; \boldsymbol{\theta})] \\ &= E\left[Z(\mathbf{X}_1; \boldsymbol{\theta})\frac{Z_i(\mathbf{X}_1; \boldsymbol{\theta})}{Z(\mathbf{X}_1; \boldsymbol{\theta})}\frac{Z_j(\mathbf{X}_1, \boldsymbol{\theta})}{Z(\mathbf{X}_1; \boldsymbol{\theta})} + \right. \\ &\quad \left. + \frac{(1 - Z(\mathbf{X}_1; \boldsymbol{\theta}))(-Z_i(\mathbf{X}_1; \boldsymbol{\theta})) - Z_j(\mathbf{X}_1; \boldsymbol{\theta})}{(1 - Z(\mathbf{X}_1; \boldsymbol{\theta}))^2}\right] \\ &= E\left[\frac{Z_i(\mathbf{X}_1; \boldsymbol{\theta})Z_j(\mathbf{X}_1; \boldsymbol{\theta})}{Z(\mathbf{X}_1; \boldsymbol{\theta})} + \right. \\ &\quad \left. + \frac{(Z_i(\mathbf{X}_1; \boldsymbol{\theta}))Z_j(\mathbf{X}_1; \boldsymbol{\theta})}{1 - Z(\mathbf{X}_1; \boldsymbol{\theta})}\right] \end{aligned} \tag{3.111}$$

Similarly,

$$\begin{aligned} E\left[\lambda_{ij}(Y_1, \mathbf{X}_1; \boldsymbol{\theta})\right] &= E\left[E\left[\lambda_{ij}(Y_1, \mathbf{X}_1; \boldsymbol{\theta})|\mathbf{X}_1\right]\right] \\ &= E\left[Z(\mathbf{X}_1; \boldsymbol{\theta})\left\{\frac{Z_{ij}(\mathbf{X}_1; \boldsymbol{\theta})}{Z(\mathbf{X}_1; \boldsymbol{\theta})} - \frac{Z_i(\mathbf{X}_1; \boldsymbol{\theta})Z_j(\mathbf{X}_1; \boldsymbol{\theta})}{Z^2(\mathbf{X}_1; \boldsymbol{\theta})}\right\} + \right. \\ &\quad \left. +(1 - Z(\mathbf{X}_1; \boldsymbol{\theta}))\left\{\frac{-Z_{ij}(\mathbf{X}_1; \boldsymbol{\theta})}{1 - Z(\mathbf{X}_1; \boldsymbol{\theta})} - \frac{Z_i(\mathbf{X}_1; \boldsymbol{\theta})Z_j(\mathbf{X}_1; \boldsymbol{\theta})}{(1 - Z(\mathbf{X}_1; \boldsymbol{\theta}))^2}\right\}\right] \\ &= -E\left[\frac{Z_i(\mathbf{X}_1; \boldsymbol{\theta})Z_j(\mathbf{X}_1; \boldsymbol{\theta})}{Z(\mathbf{X}_1; \boldsymbol{\theta})} + \right. \\ &\quad \left. + \frac{(Z_i(\mathbf{X}_1; \boldsymbol{\theta}))Z_j(\mathbf{X}_1; \boldsymbol{\theta})}{1 - Z(\mathbf{X}_1; \boldsymbol{\theta})}\right] \end{aligned} \tag{3.112}$$

So, to get $(C.6.)$, we only have to assume invertibility of $I(\boldsymbol{\theta})$.

(A.VI) If $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ is the true parameter value, then $I(\boldsymbol{\theta})$ is invertible.

□

(C.7.) $\mathrm{Var}\lambda_{ij}(Y_1, \mathbf{X}_1; \boldsymbol{\theta}_0) < \infty$ for all i, j

Proof. It suffices to show that $E[(\lambda_{ij}(Y_1, \mathbf{X}_1; \boldsymbol{\theta}))^2] < \infty$.
From (C.4.), we have $|\lambda_{ij}(Y_1, \mathbf{X}_1; \boldsymbol{\theta})| \leq C_1 + C_2|x_l x_m|$ depending on i, j.
Therefore, it follows from $(A.V)$ that we have $\mathrm{Var}\lambda_{ij}(Y_1, \mathbf{X}_1; \boldsymbol{\theta}) < \infty$.
□

(C.8.) $E|\lambda_i(Y_1, \mathbf{X}_1; \boldsymbol{\theta})|^\mu < \infty$, for all i, for some $\mu > 2$.

Proof. Since $|\lambda_i(Y_1, \mathbf{X}_1; \boldsymbol{\theta})| \leq C_1 + C_2|x_l|$, condition (C.8.) is satisfied with the assumption that $E\left\|\mathbf{X}_1\right\|^{1+\mu} < \infty$. □

Theorem 3.4. *If conditions $(A.I) - (A.VI)$ hold, then we have for*

$$Q_n = \max_{1 \leq K \leq n-1}(-2 \log \Lambda_K^n)$$

under the hypothesis H_0

$$\lim_{n \to \infty} P(a(log\ n)Q_n^{0.5} \leq x + b(log\ n)) = exp(-2\exp(-x))\ \forall x\ and\ x \in \Re$$

$$where\ a(s) = (2\log s)^{0.5}\ and\ b(s) = 2\log s + \frac{D}{2}\log(\log s) - \log(\Gamma(\frac{D}{2})) \tag{3.113}$$

D being the dimension of $\boldsymbol{\theta}$.

This result follows immediately from the considerations above and Theorem (2.1) of Gombay and Horvath [30].

We present the asymptotic critical values from (3.113), denoted as R_1, in table (3.1).

The right hand side of (3.113) is the square of a Gumbel distribution which is an extreme value distribution. As pointed out in Gombay and Horvath [30], the rate of convergence to extreme value distributions is usually slow. Therefore, (3.113) works only for large sample sizes. As we show later through simulation, (3.113) gives conservative rejection regions in case of small and moderate sample sizes.

It is for this reason that Gombay and Horvath [30] derived further approximations for $Q_n^{1/2}$ that yields good results for smaller sample sizes as described below:

As conditions (C.1) to (C.8) follow from $(A.I)-(A.IV)$, we also have the following result implied by Theorem 2.2 of Gombay and Horvath [30]:

Theorem 3.5. *If conditions* $(A.I)-(A.VI)$ *hold, then we have under* H_0

$$\left| Q_n^{1/2} - \sup_{\frac{1}{n}\le t\le 1-\frac{1}{n}} \left(\frac{B_n^{(D)}(t)}{t(1-t)} \right)^{1/2} \right| = O_p\left(exp(-\log n)^{1-\epsilon}\right)$$

for all $0 < \epsilon < 1$ *where* $\{B_n^{(D)},\ 0 \le t \le 1\}$ *is a sequence of stochastic processes distributed as*

$$B^{(D)}(t) = \sum_{1\le i\le D} B_i^2(t), 0 \le t \le 1$$

and

$$B_i(t),\ i = 1, \ldots, D$$

are independent Brownian bridges.

In particular, define for $0 < \alpha < 1$,

$$q_n = q_n(1-\alpha) = \sup\left\{x : P\left(Q_n^{1/2} \le x\right) \le 1-\alpha\right\} \tag{3.114}$$

and

$$\begin{aligned} V(a,b) &= V(a,b;1-\alpha) \\ &= \sup\left\{ x : P\left(\sup_{a\le t\le 1-b} \left(\frac{B^{(D)}(t)}{t(1-t)} \right)^{1/2} \le x \right) = 1-\alpha \right\} \end{aligned} \tag{3.115}$$

It is then shown that $V(a,b)$ is an asymptotically correct critical value of size α.

Theorem 3.6. *If conditions* $(A.I)-(A.VI)$ *hold, then under* H_0 *with*

$$a(n), b(n) \geq \frac{1}{n}$$

and with

$$\limsup_{n\to\infty} n[a(n)+b(n)]\exp\{-(\log n)^{1-\epsilon^*}\} < \infty$$

for some $0 \leq \epsilon^* \leq 1$*, then we have that*

$$\lim_{n\to} P\left\{Q_n^{1/2} > V(a(n), b(n))\right\} = \alpha \tag{3.116}$$

and that

$$|q_n - V(a(n), b(n))| = o\left((\log(\log n))^{1/2}\right) \tag{3.117}$$

Analogous to Gombay and Horvath [30], we choose

$$a(n) = b(n) = \frac{(\log\, n)^{3/2}}{n}$$

in (3.117) which makes $V(a,b)$ a good approximation for $q_n = q_n(1-\alpha)$.

However, since there is no known simple formula for the distribution function of $\sup_{a\leq t\leq 1-b}\left(\frac{B^{(D)}(t)}{t(1-t)}\right)^{1/2}$, we use itss inverted Laplace transform;

$$P\left(\sup_{a\leq t\leq 1-b}\left(\frac{B^{(D)}(t)}{t(1-t)}\right)^{1/2} \geq u\right) = \frac{u^D exp(-u^2/2)}{2^{D/2}\Gamma(D/2)}\left\{M - \frac{D}{u^2}M + \frac{4}{u^2} + O(\frac{1}{u^4})\right\} \tag{3.118}$$

where

$$M = \log\frac{(1-a)(1-b)}{ab}$$

,see Gombay and Horvath [30].

We represent the asymptotic critical values from (3.118), denoted as R_2, in table (3.1).

Sample Size	$1-\alpha$	R_1	R_2
50	0.90	3.7314	4.0634
	0.95	4.1672	4.3060
	0.99	5.1540	4.7870
100	0.90	3.8747	4.1518
	0.95	4.2866	4.3858
	0.99	5.2192	4.8545
150	0.90	3.9408	4.1946
	0.95	4.3418	4.4246
	0.99	5.2497	4.8874
200	0.90	3.9820	4.2220
	0.95	4.3762	4.4495
	0.99	5.2688	4.9085
500	0.90	4.0906	4.2966
	0.95	4.4672	4.5175
	0.99	5.3199	4.9666

Table 3.1: *Asymptotic critical values from equation (3.113), denoted as R_1 and from equation (3.118), denoted as R_2. Both critical values were evaluated at D=5 in line with our neural network structure.*

3.9 Simulation Study

For simulation purposes under H_1, we used the following model:

$$P(Y_i = 1|X_i = \mathbf{x}) = \begin{cases} (1 + exp(-(-1.5 + 2 * x_{1i} + 1 * x_{2i})))^{-}1, & i \leq K \\ (1 + exp(-(-1.5 + 3 * x_{1i} + 3 * x_{2i})))^{-}1, & K < i \leq n \end{cases} \tag{3.119}$$

The change point K was fixed at $n/2$ for $n = 50$ and $n = 500$. We then generated x_{1i} and x_{2i} as $uniform[0, 1]$. We then generated the Bernoulli random variables Y_i in line with equation (3.119).

Using equations (3.12), (3.15) and (3.49) the likelihood ratio was estimated and the results represented below. From table (3.1) and figure (3.1) of Hypothesis testing graph of simulated data, $Q_{50}^{1/2} = 4.8977$ rejected H_o under (3.113) at only 90% and 95% confidence interval but rejected it under (3.118) at 90% ,95% and 99% confidence intervals. This is due to slow convergence under (3.113).

However, figure (3.2) has $Q_{500}^{1/2} = 6.8690$ which rejected H_o under (3.113) and (3.118) at 90% and 95% and 99% confidence intervals. This is because

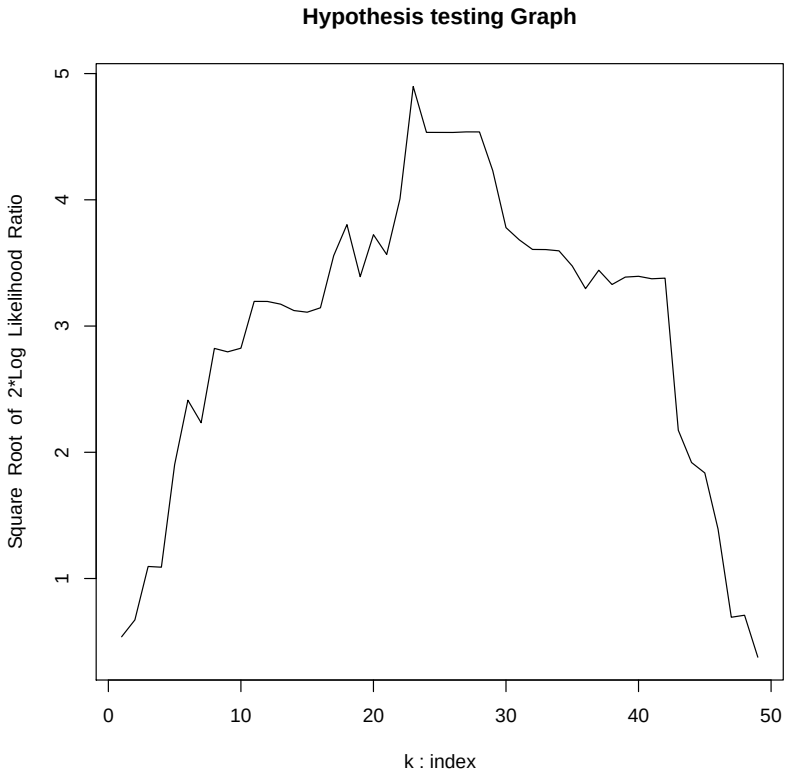

Figure 3.1: *Change Point testing Graph for n=50*

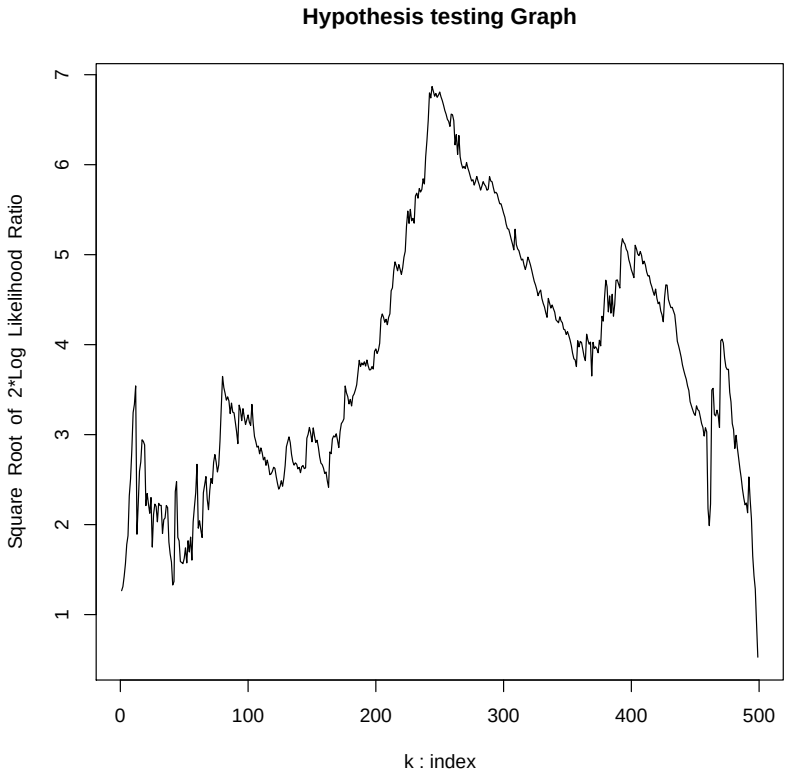

Figure 3.2: *Change Point Testing for n=500*

$n = 500$ is large enough for proper convergence. We did further simulations to test for a change when it was actually not present with the following model:

$$P(Y_i = 1|X_i = \mathbf{x}) = \begin{cases} (1 + exp(-(-1.5 + 3.5 * x_{1i} + 3 * x_{2i})))^{-}1, & i \leq 500 \end{cases} \tag{3.120}$$

where $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ were generated as in equation (3.119) and Y_i generated in line with equation (3.120). The results are presented in figure (3.3). From

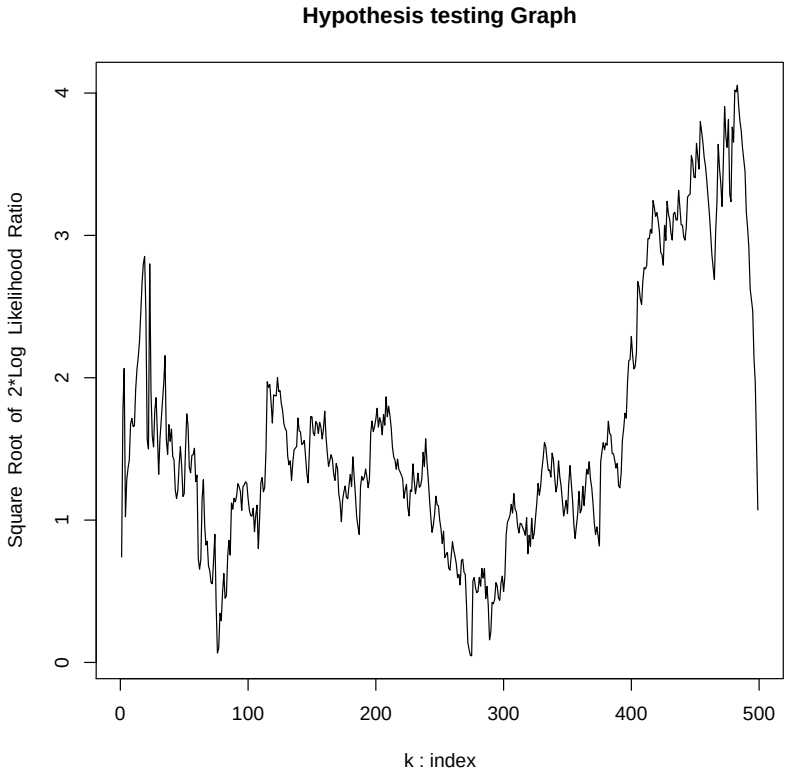

Figure 3.3: *Change Point Testing for* $n = 500$ *when actually there is no change*

table (3.1) and figure (3.3), $Q_{500}^{1/2} = 4.0555$ accepted H_o under (3.113) and (3.118) at 90% confidence interval.

3.10 Power of the Test

In this section, we have a look at the asymptotic power of the change point test discussed in section (3.8), i.e., for sake of simplicity, we restrict ourselves to the case of a correctly specified model. Recall that the main test statistic is

$$Q_n \;=\; \max_{1\le K\le n-1}(-2\ \ln\Lambda_K^n) = \max_{1\le K\le n-1}(2\ \ln(\Lambda_K^n)^{-1})$$

where

$$\Lambda_K^n \;=\; \frac{L_0(\hat{\boldsymbol{\theta}}_0)}{L_{1,K}(\hat{\boldsymbol{\theta}}_K,\hat{\boldsymbol{\theta}}_K^*)}$$

with $\hat{\boldsymbol{\theta}}_0,\hat{\boldsymbol{\theta}}_K,\hat{\boldsymbol{\theta}}_K^*$ being the maximum likelihood estimates under H_o and under the alternative of a change at K respectively. Writing

$$P_{\boldsymbol{\theta}}(Y|\mathbf{x}) = \begin{cases} \psi(O_H(\mathbf{x};\boldsymbol{\theta})), \text{ if } Y=1 \\ 1-\psi(O_H(\mathbf{x};\boldsymbol{\theta})), \text{ if } Y=0 \end{cases}$$

For the conditional probabilities of $Y_i = Y$ given $\mathbf{X}_j = \mathbf{x}$ provided that $\boldsymbol{\theta}$ is the true parameter, we have

$$\Lambda_K^n = \Pi_{i=1}^K \frac{P_{\hat{\boldsymbol{\theta}}_0}(Y_i|\mathbf{X}_i)}{P_{\hat{\boldsymbol{\theta}}_K}(Y_i|\mathbf{X}_i)} \Pi_{i=K+1}^n \frac{P_{\hat{\boldsymbol{\theta}}_0}(Y_i|\mathbf{X}_i)}{P_{\hat{\boldsymbol{\theta}}_K^*}(Y_i|\mathbf{X}_i)}$$

As Q_n is an increasing function of $\max_{1\le K\le n-1}((\Lambda_K^n)^{-1})$, the test for a specified level α rejects the hypothesis if

$$\max_{1\le K\le n-1}((\Lambda_K^n)^{-1}) > R$$

for some bound R. From theorem 3.3, we have that R grows asymptotically like n, as for given t depending on the level of the test

$$\frac{(t+b(\ln n))^2}{a^2(\ln n)} \approx 2\ln n$$

We now want to give some heuristic arguments for the test being consistent in the sense that, for given level α, the power converges to 1.

Let the alternative hold, i.e. there is some change point $1 \le K \le n-1$. Assume that for $n\to\infty$ we have $K, n-K \to \infty$ such that

$$\frac{K}{n} \to \tau \in (0,1)$$

i.e. the change happens after a fraction τ of the data. Let $\boldsymbol{\theta}_\tau, \boldsymbol{\theta}_\tau^*$ denote the true parameter values before and after the change point. Let, moreover, $\boldsymbol{\theta}_0$ denote the parameter values under H_0 which approximate the true distribution of the data best in the sense of section (3.6), i.e

$$\begin{aligned}
\boldsymbol{\theta}_0 &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[\frac{1}{n}\sum_{i=1}^{n} Y_i \ln Z(\mathbf{X}_i;\boldsymbol{\theta}) + (1-Y_i)\ln(1-Z(\mathbf{X}_i;\boldsymbol{\theta}))] \\
&= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\frac{K}{n} E\left\{Z(\mathbf{X}_i;\boldsymbol{\theta}_\tau)\ln Z(\mathbf{X}_i;\boldsymbol{\theta}) + (1-Z(\mathbf{X}_i;\boldsymbol{\theta}_\tau))\ln(1-Z(\mathbf{X}_i;\boldsymbol{\theta}))\right\}\right. \\
&\quad \left. + \frac{n-K}{n} E\left\{Z(\mathbf{X}_i;\boldsymbol{\theta}_\tau^*)\ln Z(\mathbf{X}_i;\boldsymbol{\theta}) + (1-Z(\mathbf{X}_i;\boldsymbol{\theta}_\tau^*))\ln(1-Z(\mathbf{X}_i;\boldsymbol{\theta}))\right\}\right]
\end{aligned}$$

From the consistency results of theorem (3.3), we have for $n \to \infty$

$$\hat{\boldsymbol{\theta}}_0 \to \boldsymbol{\theta}_0,\ \hat{\boldsymbol{\theta}}_K \to \boldsymbol{\theta}_\tau,\ \hat{\boldsymbol{\theta}}_K^* \to \boldsymbol{\theta}_\tau^*.$$

So, we asymptotically have by the law of large numbers

$$\frac{1}{n}\log\Lambda_K^n \sim \tau E_{\boldsymbol{\theta}_\tau} \log\frac{P_{\boldsymbol{\theta}_0}(Y_i|\mathbf{X}_i)}{P_{\boldsymbol{\theta}_\tau}(Y_i|\mathbf{X}_i)} + (1-\tau)E_{\boldsymbol{\theta}_\tau^*}\log\frac{P_{\boldsymbol{\theta}_0}(Y_i|\mathbf{X}_i)}{P_{\boldsymbol{\theta}_\tau^*}(Y_i|\mathbf{X}_i)}$$

We assume that the alternative holds, i.e. $\boldsymbol{\theta}_\tau^* \neq \boldsymbol{\theta}_\tau^*$, from which we also get $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_\tau,\ \boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_\tau^*$ by the definition of $\boldsymbol{\theta}_0$ as the parameters closest to a mixture of the distributions with parameters $\boldsymbol{\theta}_\tau$ and $\boldsymbol{\theta}_\tau^*$ respectively. In the correctly specified case, we have from our identifiability assumptions that, then, also $P_{\boldsymbol{\theta}_0} \neq P_{\boldsymbol{\theta}_\tau},\ P_{\boldsymbol{\theta}_0} \neq P_{\boldsymbol{\theta}_\tau^*}$. As log is strictly concave, we have from Jensen's inequality

$$\begin{aligned}
E_{\boldsymbol{\theta}_\tau}\log\frac{P_{\boldsymbol{\theta}_0}(Y_i|\mathbf{X}_i)}{P_{\boldsymbol{\theta}_\tau}(Y_i|\mathbf{X}_i)} &< \log E_{\boldsymbol{\theta}_\tau}\frac{P_{\boldsymbol{\theta}_0}(Y_i|\mathbf{X}_i)}{P_{\boldsymbol{\theta}_\tau}(Y_i|\mathbf{X}_i)} \\
&= \log\int\int \frac{P_{\boldsymbol{\theta}_0}(Y|\mathbf{x})}{P_{\boldsymbol{\theta}_\tau}(Y|\mathbf{x})} P_{\boldsymbol{\theta}_\tau}(Y|\mathbf{x})\mathrm{d}\nu(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\
&= \log\int\int P_{\boldsymbol{\theta}_0}(Y|\mathbf{x})\mathrm{d}\nu(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \log 1 = 0
\end{aligned}$$

and, analogously, for $\boldsymbol{\theta}_\tau^+$ replacing $\boldsymbol{\theta}_\tau$. So, we get for some constant $\zeta > 0$,

$$\frac{1}{n}\log\Lambda_K^n \sim -\zeta \text{ and } \log(\Lambda_K^n)^{-1} \sim n\zeta$$

This implies that the probability of an error of type II vanishes asymptotically, as

$$P\left(\max_{1\leq K\leq n-1}(\Lambda_K^n)^{-1} \leq R|H_1\right) \leq P\left((\Lambda_K^n)^{-1} \leq R|H_1\right) \to 0 \text{ for } n\to\infty$$

as $(\Lambda_K^n)^{-1}$ grows like $e^{n\zeta}$ and R only grows like n. Therefore, the asymptotic power of the change point test is 1.

We now investigate the power of the change point test for finite sample size by Monte Carlo simulation for specific alternatives of one change point. Recall that the test rejects the hypothesis if $Q_n^{\frac{1}{n}} > R$ where we get the critical bound R from the asymptotics of either Theorem 3.5 or Theorem 3.4. R depends on the level of α of the test and on the sample size n. The two possible values R_1, R_2 are tabulated in table (3.1).

The power of the test of level α against a particular alternative H_1 is then defined as the probability of rejecting H_0 correctly. i.e.

$$\tau(\alpha) = P\left(Q_n^{1/2} > R|H_1\right)$$

To get a more detailed asymptotic analysis of $\tau(\alpha)$, we need an approximation of the distribution of $Q_n^{1/2}$ under the alternative. This is beyond the scope of this thesis, and, therefore, we rely on simulation which were performed as follows:

Using the simulated data model in equation (3.119), $Q_n^{1/2}$ in equation (3.50) was estimated for every replicate where $B = 1000$ replicates and a defined sample size n. The power function for a given level $1 - \alpha$, was estimated by

$$\hat{\tau}(\alpha) = \left(1 + \#\left(Q_n^{1/2} > R_n(\alpha)\right)\right)/(1 + N) \tag{3.121}$$

where $\#(a > b)$ denotes the number of times a is greater than b. The procedure was carried out for the sample sizes $n = 150$ and $n = 200$.
The results are presented in table (3.2), table (3.3), table (3.5), figure (3.4) and figure (3.5).

We analyze the behavior of the test as the change point approaches data edges for a sample size of $n = 200$ using the critical bound R_1 in table (3.1). In figure 3.4 we plot the 95% power function under R_1 from 3.2 above. Below we analyse the behaviour of the test as the change point approaches data edges for a sample size of $n = 200$ using the critical bound R_2 in table (3.1). In figure (3.5) we plot the 95% power function using the critical bound R_2 from table (3.3) above. Results in table (3.2), table (3.3), figure (3.4) and figure (3.5) indicate that the change point test is less powerful when the change point is closer to the edges of the data. This finding is in agreement with the findings in James *et al* [42]. As noted in Jaruskova [44], this behavior at the data edges is due to (say K is near 0) comparing an estimate calculated from a relatively small number of observations - the first K observations -

	$\hat{\tau}(\alpha)$						
	Change Points						
α	25	50	75	100	125	150	175
0.10	0.7273	0.9501	0.9660	0.9760	0.9660	0.9131	0.6653
0.05	0.5285	0.8472	0.9141	0.9291	0.8931	0.7992	0.4535
0.01	0.1738	0.4825	0.6354	0.6523	0.5624	0.3606	0.0679

Table 3.2: *Change Point Power function of the likelihood Ratio test of a sample size* $n = 200$*. 1,000 simulations were done to determine each estimate.*

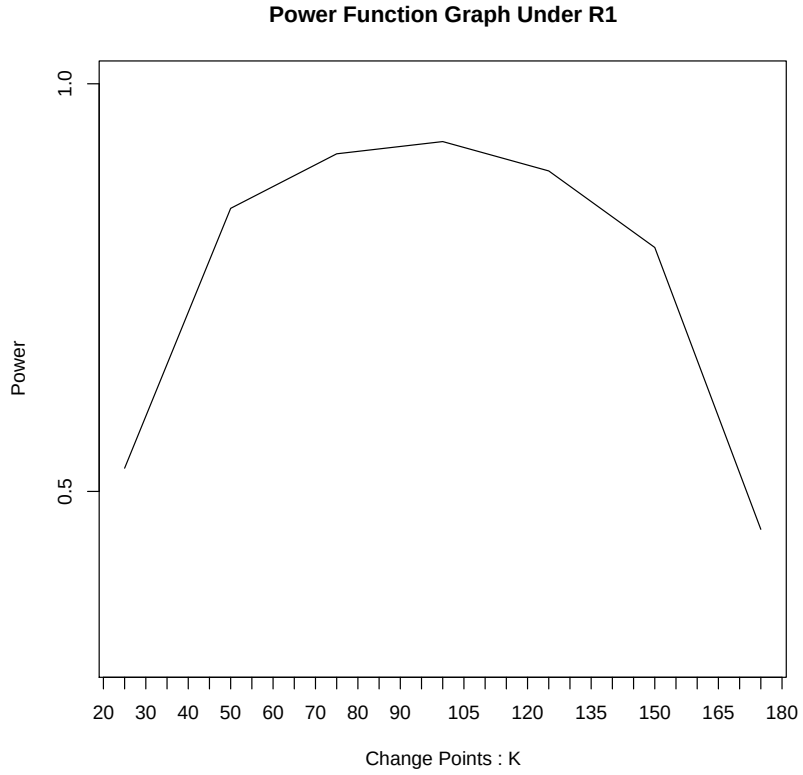

Figure 3.4: *The* 95% *power function using the critical bound* R_1 *for* $n = 200$

	$\hat{\tau}(\alpha)$						
	Change Points						
α	25	50	75	100	125	150	175
0.10	0.5944	0.8981	0.9371	0.9491	0.9341	0.8541	0.5315
0.05	0.4965	0.8262	0.9061	0.9151	0.8741	0.7632	0.4026
0.01	0.3057	0.6494	0.7572	0.7912	0.7293	0.5455	0.1828

Table 3.3: *Change Point Power function of the likelihood Ratio test of a sample size* $n = 200$. *1,000 simulations were done to determine each estimate.*

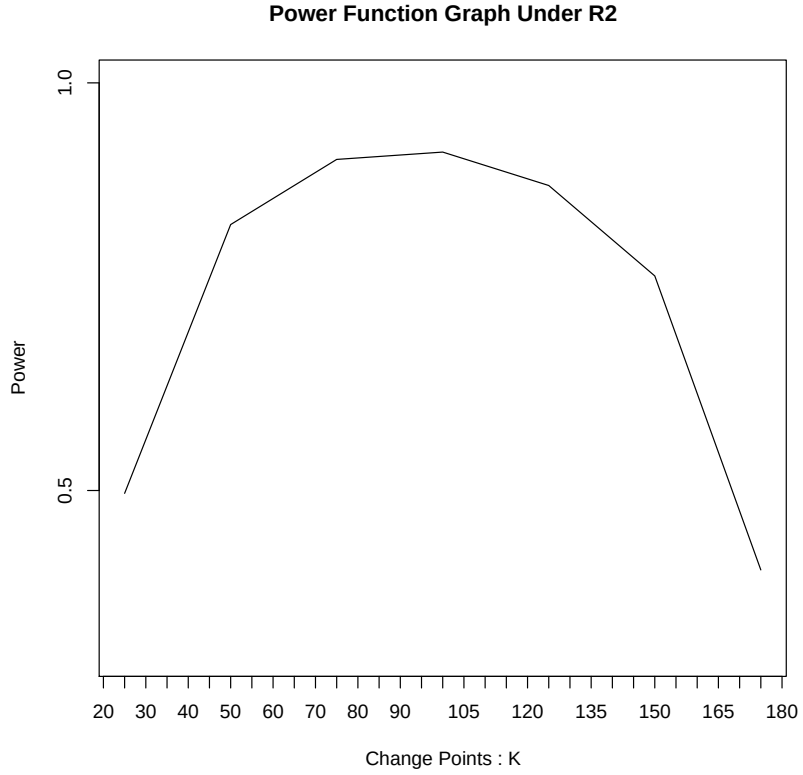

Figure 3.5: *The* 95% *power function under* R_2 *for* $n = 200$

	$\hat{\tau}(\alpha)$under R_1		
		Size of Change	
K	$1-\alpha$	$\Delta = 25$	$\Delta = 3.5$
25	0.90	0.4256	0.7273
	0.95	0.2128	0.5285
	0.99	0.0220	0.1738
50	0.90	0.5265	0.9501
	0.95	0.3127	0.8472
	0.99	0.0529	0.4825
75	0.90	0.6254	0.9660
	0.95	0.4016	0.9141
	0.99	0.0959	0.6354
100	0.90	0.6464	0.9760
	0.95	0.4456	0.9291
	0.99	0.1099	0.6523
125	0.90	0.6254	0.9660
	0.95	0.4016	0.8931
	0.99	0.0959	0.5624
150	0.90	0.5444	0.9131
	0.95	0.3017	0.7992
	0.99	0.0370	0.3606
175	0.90	0.3686	0.6653
	0.95	0.1668	0.4535
	0.99	0.0140	0.0679

Table 3.4: *Change Point Power function values of the likelihood Ratio test for different sizes of change and locations K. Sample size $n = 200$ and 1,000 simulations were done for each corresponding case.*

with an estimate calculated from a large number of observations - the last $n - K$ observations. This implies that the test is more likely to reject a change point at the edge of the data than when the change is relatively far away from the data edges.

Simulations were also carried out to investigate the power of the test in relation to the size of the change, denoted as Δ, and change point location under R_1. The results are presented in table (3.4) and figure (3.6). These results show that the power of the test increases with an increase in the size of change as expected. From table (3.4) and figure (3.6), the loss of power seems to be more due to the size of change than to the change point location. This result is relevant to application since it is more important to detect a change early once it has taken place irrespective of its location.

Lastly, we carried out simulations to study the power of the test as n increases. Since, as shown in table (3.2) and table (3.3), change point detection power depends on the location of the change point, we fixed $K = n/2$ for $n = \{150, 200\}$ for proper analysis. The results in table (3.5) confirm that the power of the test increases with the increase in n.

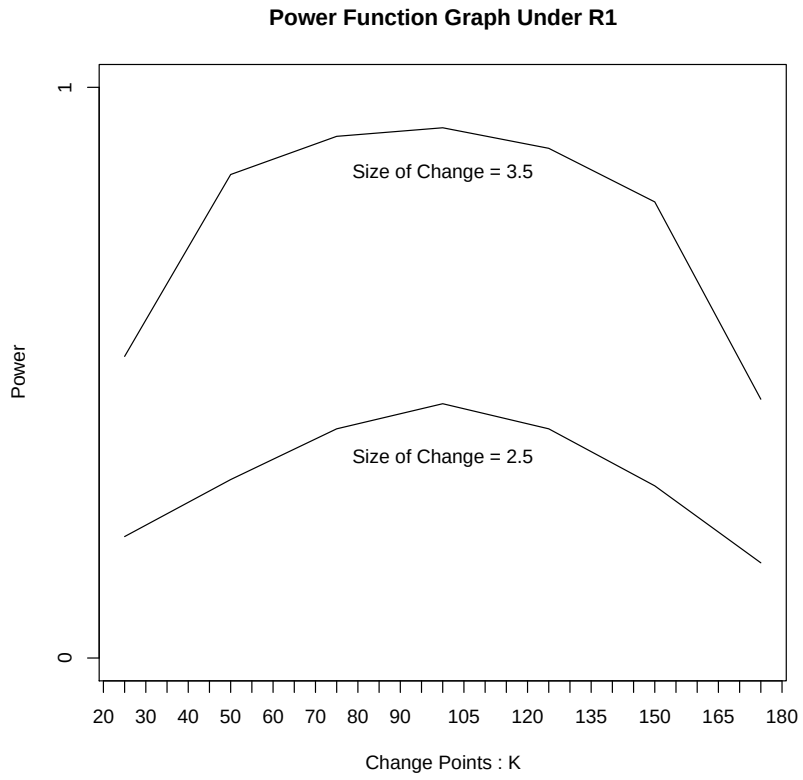

Figure 3.6: *The* 95% *power function for different sizes of change and locations* K *under* R_1 *for* $n = 200$

$\hat{\tau}(\alpha)$					
n=150			n=200		
K=75			K=100		
α	R_1	R_2	α	R_1	R_2
0.1	0.9321	0.8731	0.1	0.9760	0.9491
0.05	0.8242	0.7962	0.05	0.9291	0.9151
0.01	0.4505	0.5964	0.01	0.6523	0.7912

Table 3.5: *Change point Power values of the likelihood Ratio test of two sample sizes 150 and 200. 1,000 simulations were done to determine each estimate.*

3.11 Testing for Change Points under Misspecification

We now come back to the original testing problem of section (3.7) where $p_i = P(Y_i = 1|\mathbf{X}_i = \mathbf{x})$ is not necessarily of the form of an output function $Z(\mathbf{x};\boldsymbol{\theta})$ of a neural network. Nevertheless, we still apply the test of section (3.8) which now is based on a misspecified model.

We first discuss how this influences the conditions of Gombay and Horvath [30]. $(Y_i, \mathbf{X}_i),\ i = 1, \ldots, n$ are still independent with density

$$\pi_i(Y, \mathbf{x}) = \begin{cases} p_i(\mathbf{x}) & \text{, for } Y = 1 \\ 1 - p_i(\mathbf{x}) & \text{, for } Y = 0 \end{cases}$$

w.r.t. $\nu \otimes \mu$, where $\nu = \delta_0 + \delta_1$ is again the counting measure on $\{0, 1\}$ and where we still assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. with distribution μ which is a δ-finite measure.

Under the hypothesis H_0, we have $p_i(\mathbf{x}) = p_0(\mathbf{x}),\ i = 1, \ldots, n$. We define $\boldsymbol{\theta}_0$ as the parameter for which $Z(\mathbf{x};\boldsymbol{\theta}_0)$ approximates $p_0(\mathbf{x})$ as good as possible in the sense that, compare section (3.8),

$$\begin{aligned} \boldsymbol{\theta}_0 &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E\frac{1}{n}L_0(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[Y_1 \ln Z(\mathbf{X}_1;\boldsymbol{\theta}) + (1 - Y_1)(1 - Z(\mathbf{X}_1;\boldsymbol{\theta}))] \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} EE[\{Y_1 \ln Z(\mathbf{X}_1;\boldsymbol{\theta}) + (1 - Y_1)(1 - Z(\mathbf{X}_1;\boldsymbol{\theta}))\}|\mathbf{X}_1] \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[p_0(\mathbf{X}_1) \ln Z(\mathbf{X}_1;\boldsymbol{\theta}) + (1 - p_0(\mathbf{X}_1))(1 - Z(\mathbf{X}_1;\boldsymbol{\theta}))] \end{aligned} \tag{3.122}$$

We remark that under the misspecified model, the density of $(Y_i, \mathbf{X}_i)$ w.r.t. the dominating measure $\nu \otimes \mu$ is under H_0

$$\pi_i(Y, \mathbf{x};\boldsymbol{\theta}) = \begin{cases} Z(\mathbf{x};\boldsymbol{\theta}) & \text{, for } Y = 1 \\ 1 - Z(\mathbf{x};\boldsymbol{\theta}) & \text{, for } Y = 0 \end{cases}$$

Therefore, $\boldsymbol{\theta}_0$ can also be written as

$$\begin{aligned} \boldsymbol{\theta}_0 &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int\int \ln \pi(Y, \mathbf{x};\boldsymbol{\theta})\pi_0(Y, \mathbf{x})\mathrm{d}\nu(Y)\mathrm{d}\mu(\mathbf{x}) \\ &= \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[-E \ln \frac{\pi(Y, \mathbf{x};\boldsymbol{\theta})}{\pi_0(Y, \mathbf{x})}\right] \end{aligned} \tag{3.123}$$

i.e. $\boldsymbol{\theta}_0$ minimizes the Kullback-Leibler distance between the approximating parametric density $\pi(Y, \mathbf{x};\boldsymbol{\theta})$ and the true density $\pi_0(Y, \mathbf{x})$, compare White [68].

Now, let us discuss the conditions $(C.1.)-(C.8.)$ of Gombay and Horvath [30] under misspecification.

(C.1.) Under the conditions of section (3.8), we have that $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_0^*$ implies $Z(\mathbf{x};\boldsymbol{\theta}_0) \neq Z(\mathbf{x};\boldsymbol{\theta}_0^*)$. If $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_0^*$ solve (3.123) for $\pi_0(Y,\mathbf{x})$ and $\pi_0^*(Y,\mathbf{x})$ respectively , then if $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_0^*$, $Z(\mathbf{x};\boldsymbol{\theta}_0) \neq Z(\mathbf{x};\boldsymbol{\theta}_0^*)$ and, then, obviously $\pi_0 \neq \pi_0^*$. So, there are no different parameter values corresponding to the same distributions of the data under H_0.

For constructing the estimates, we work with the approximating parametric model, i.e. we have as in section (3.8)

$$\begin{aligned}\lambda(Y,\mathbf{x};\boldsymbol{\theta}) &= \ln[\delta_1(Y)Z(\mathbf{x};\boldsymbol{\theta}) + \delta_0(Y)(1-Z(\mathbf{x};\boldsymbol{\theta}))] \\ &= Y\ln Z(\mathbf{x};\boldsymbol{\theta}) + (1-Y)\ln(1-Z(\mathbf{x};\boldsymbol{\theta}))\end{aligned}$$

Therefore, using this condition, we have from (3.122)

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E\lambda(Y,\mathbf{x};\boldsymbol{\theta})$$

(C.2.) This condition follows immediatly if we adapt condition $(A.IV)$ to hold for the approximating parametric likelihood. Again, this is just a typical identifiability condition on $\boldsymbol{\Theta}$.

(C.3.)-(C.4.) These are just regularity conditions on $\lambda(Y,\mathbf{x};\boldsymbol{\theta})$ which continue to hold. The misspecification has only to be taken into account in calculating $E[M_2(Y_1,\mathbf{X}_1)]$ where the expectation is w.r.t. to the distribution $\pi_0(Y,\mathbf{x})$, but we still have

$$E[M_2(Y_1,\mathbf{X}_1)] = \int \{M_2(1,\mathbf{x})p_0(\mathbf{x}) + M_2(0,\mathbf{x})(1-p_0(\mathbf{x}))\}\mathrm{d}\mu(\mathbf{x}) < \infty$$

if $E\|\mathbf{X}\|^3 < \infty$

(C.5.) By definition of the parameter $\boldsymbol{\theta}_0$, it automatically satisfies, using the regularity of λ as a function of $\boldsymbol{\theta}$,

$$\nabla E\lambda(Y_1,\mathbf{X}_1;\boldsymbol{\theta}_0) = E\nabla\lambda(Y_1,\mathbf{X}_1;\boldsymbol{\theta}_0) = 0,$$

i.e. condition $(C.5.)$ is automatically satisfied by definition of $\boldsymbol{\theta}_0$.

(C.6.) This condition is the critical one for the misspecified case. It states the equivalence of the Fisher information matrix in the correctly specified model. It does not hold in the misspecified model and the asymptotic

covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}_0)$ under H_0 has the more complicated form

$$I(\boldsymbol{\theta}_0) = A^{-1}(\boldsymbol{\theta}_0)B(\boldsymbol{\theta}_0)A^{-1}(\boldsymbol{\theta}_0)$$

where $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta}) = B_1(\boldsymbol{\theta}) + B_2(\boldsymbol{\theta})$ are as in Theorem 3.3. We have to replace $(A.VI)$ by the condition

(A.VII) $I(\boldsymbol{\theta})$ is invertible (for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$)

(C.7.)-(C.8.) Both conditions continue to hold for $\lambda(Y, \mathbf{x}; \boldsymbol{\theta})$.

In the next section, we have a closer look at the maximum likelihood change point test under misspecification. There, we have to replace $(C.6.)$ by the assumption that $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$ each satisfy the corresponding conditions and, in particular, $A(\boldsymbol{\theta}_0)$ is positive definite which is automatically satisfied in the correctly specified case. For the case of one-dimensional parameter, we are able to show that the scaled test statistic

$$\hat{Q}_n = \max_{1 \le K \le n-1} (-2 \ \log \Lambda_K^n) \frac{A_n(\hat{\boldsymbol{\theta}}_n)}{B_n(\hat{\boldsymbol{\theta}}_n)}$$

has to be considered instead. Otherwise, the asymptotic behaviour of the change point test does not change. Theorem 3.3 continues to hold with $\hat{Q}_n$ replacing Q_n. However, for sieve estimates, we have to consider the multiparameter case $D > 1$ which remains to be done.

3.12 Testing for Change Points under Misspecification - the general case

In this section, we leave the context of Bernoulli regression models and consider the following more general situation. The data $\mathbf{X}_j$ correspond to the pairs $(Y_j, \mathbf{X}_j)$ in the previous sections.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \Re^m$ be independent random vectors with densities $f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})$ w.r.t. some σ-finite measure ν on $(\Re^m, \mathcal{B}^m)$. We want to test for a changepoint, i.e.

$$\begin{aligned} \bar{H}_0 &: f_1(\mathbf{x}) = \ldots = f_n(\mathbf{x}) \\ vs & \\ \bar{H}_1 &: f_1(\mathbf{x}) = \ldots = f_K(\mathbf{x}) \neq f_{K+1}(\mathbf{x}) = \ldots = f_n(\mathbf{x}) \text{ for some } 1 \le K \le n \end{aligned} \tag{3.124}$$

The form of the densities is not known but we still want to apply maximum likelihood ratio tests as in Gombay and Horvath ([28],[30]). As in the previous section, we approximate the unknown $f_j(\mathbf{x})$ by some parametric density $f(\mathbf{x};\boldsymbol{\theta}^j)$, and then we consider the maximum likelihood ratio test of the parametric changepoint problem:

$$\begin{aligned} \bar{H}_0 &: \boldsymbol{\theta}^1 = \ldots = \boldsymbol{\theta}^n \\ vs & \\ \bar{H}_1 &: \boldsymbol{\theta}^1 = \ldots = \boldsymbol{\theta}^K \neq \boldsymbol{\theta}^{K+1} = \ldots = \boldsymbol{\theta}^n \text{ for some } 1 \leq K \leq n \end{aligned} \tag{3.125}$$

$\boldsymbol{\theta}^j$ is, of course, a parameter of the distribution of $\mathbf{X}_j$, but in contrast to the setting of Gombay and Horvath, it does not completely specify the density $f_j(\mathbf{x})$. Nevertheless, the parametric setting provides a legitimate test of the original changepoint problem if we impose the following identifiability condition:

(A_1) For $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ implies that the densities $f(\mathbf{x};\boldsymbol{\theta}), f(\mathbf{x};\boldsymbol{\theta}^*)$ do not coincide.

Then, if the hypothesis $\bar{H}_0 : f_1(\mathbf{x}) = \ldots = f_n(\mathbf{x})$ holds, we immediately have $f_1(\mathbf{x};\theta^{(1)}) = \ldots = f_n(\mathbf{x};\theta^{(n)})$ and, by (A_1), $\theta^{(1)} = \ldots = \theta^{(n)}$, i.e. the parametric hypothesis H_0 holds too. So, if we reject H_0, we automatically reject $\bar{H}_0$.

The basic idea for constructing the test is to adopt the misspecified parametric model that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent with densities $f(\mathbf{x};\boldsymbol{\theta}^{(j)}), \boldsymbol{\theta}^{(j)} \in \boldsymbol{\Theta}, j = 1, \ldots, n$, and, then, apply the well-known parametric maximum likelihood tests to decide between H_0 and H_1. We have to investigate how the test statistic behaves asymptotically in this misspecified situation.

We define the relation between $f_j(\mathbf{x})$ and its parametric approximation in the following way:

We choose $\boldsymbol{\theta}^{(j)}$ such that the Kullback-Leibler distance (w.r.t. the dominating measure ν) between $f(\mathbf{x};\boldsymbol{\theta}^{(j)})$ and $f_j(\mathbf{x})$ is minimized

$$\boldsymbol{\theta}^{(j)} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \; -\int f_j(\mathbf{x}) \log \frac{f(\mathbf{x};\boldsymbol{\theta})}{f_j(\mathbf{x})} \nu(\mathrm{d}\mathbf{x}) \tag{3.126}$$

Equivalently, we have as only the numerator of the logarithmic term depends on $\boldsymbol{\theta}$,

$$\begin{aligned} \boldsymbol{\theta}^{(j)} &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int f_j(\mathbf{x}) \log f(\mathbf{x};\boldsymbol{\theta}) \nu(\mathrm{d}\mathbf{x}) \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \; E \log f(\mathbf{X}_j;\boldsymbol{\theta}) \end{aligned} \tag{3.127}$$

If the hypothesis H_0 holds, we denote the common value of $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ by $\boldsymbol{\theta}_0$

Our arguments follow closely those of Gombay and Horvath ([28],[30]), and we introduce a similar notation. Let

$$g(\mathbf{x}; \boldsymbol{\theta}) = \log f(\mathbf{x}; \boldsymbol{\theta}),\ g_i(\mathbf{x}; \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}_i} g(\mathbf{x}; \boldsymbol{\theta}), i = 1, \ldots, D$$

We assume that for any given $K = 1, \ldots, n$, there are unique solutions of the quasi likelihood equation, where $K = n$ corresponds to the case where H_0 holds.

(A_2) For all $K = 1, \ldots, n$, there are unique $\hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}^*_{n-K} \in \boldsymbol{\Theta}$ such that

$$\sum_{j=1}^{K} g_i(\mathbf{X}_j; \hat{\boldsymbol{\theta}}_K) = 0,\ \sum_{j=K+1}^{n} g_i(\mathbf{X}_j; \hat{\boldsymbol{\theta}}^*_{n-K}) = 0$$

We denote the quasi log likelihood functions before and after K by

$$L_K(\boldsymbol{\theta}) = \sum_{j=1}^{K} g(\mathbf{X}_j; \hat{\boldsymbol{\theta}}_K),\ L^*_{n-K}(\boldsymbol{\theta}) = \sum_{j=K+1}^{n} g(\mathbf{X}_j; \hat{\boldsymbol{\theta}}^*_{n-K})$$

Then, the quasi likelihood ratio statistic for testing for a changepoint in the approximate parametric model at a given K is

$$\begin{aligned} \Lambda_K &= \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \Pi_{j=1}^{n} f(\mathbf{X}_j; \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \boldsymbol{\Theta}} \Pi_{j=1}^{K} f(\mathbf{X}_j; \boldsymbol{\theta}) \Pi_{j=K+1}^{n} f(\mathbf{X}_j; \boldsymbol{\theta}^*)} \\ &= \frac{\Pi_{j=1}^{n} f(\mathbf{X}_j; \hat{\boldsymbol{\theta}}_n)}{\Pi_{j=1}^{K} f(\mathbf{X}_j; \hat{\boldsymbol{\theta}}_K) \Pi_{j=K+1}^{n} f(\mathbf{X}_j; \hat{\boldsymbol{\theta}}^*_{n-K})} \end{aligned} \tag{3.128}$$

As a test statistic for testing H_0 against H_1, we finally consider

$$Z_n = \max_{1 \le K < n} (-2 \log \Lambda_K)$$

Additionally to (A_1),(A_2), we need the following smoothness and moment conditions. Here, $\boldsymbol{\Theta}_0 \subseteq \boldsymbol{\Theta}$ denotes a suitably chosen compact subset of $\boldsymbol{\Theta}$ such that $\boldsymbol{\theta}_0$ lies in the interior of $\boldsymbol{\Theta}_0$. E_0 denotes the expectation under the hypothesis $\bar{H}_0 : f_j(\mathbf{x}) = f_0(\mathbf{x}),\ j = 1, \ldots, n$, i.e.

$$E_0 M(\mathbf{X}_j) = \int M(\mathbf{x}) f_0(\mathbf{x}) \nu(\mathrm{d}\mathbf{x}).$$

$\nabla,\ \nabla^2$ denote the gradient and the Hessian w.r.t. $\boldsymbol{\theta}$.

(A_3) The derivatives of $g(\mathbf{x};\boldsymbol{\theta}) = \log f(\mathbf{x};\boldsymbol{\theta})$w.r.t. $\boldsymbol{\theta}$

$$\begin{aligned} g_i(\mathbf{x};\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}_i} g(\mathbf{x};\boldsymbol{\theta}) \\ g_{ij}(\mathbf{x};\boldsymbol{\theta}) &= \frac{\partial^2}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} g(\mathbf{x};\boldsymbol{\theta}) \\ g_{ijl}(\mathbf{x};\boldsymbol{\theta}) &= \frac{\partial^3}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_l} g(\mathbf{x};\boldsymbol{\theta}) \end{aligned} \tag{3.129}$$

exist and are continous in $\boldsymbol{\theta}$ for all $\mathbf{x}$ and for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, $i, j, l = 1, \ldots, D$

(A_4) There are functions $M_1(\mathbf{x}),\ M_2(\mathbf{x})$ satisfying

$$\int M_1(\mathbf{x})\nu(\mathrm{d}\mathbf{x}) < \infty, E_0 M_2(\mathbf{X}_1) < \infty$$

such that for all $\mathbf{x}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$$\begin{aligned} |g_i(\mathbf{x};\boldsymbol{\theta})| &\leq M_1(\mathbf{x}) \\ |g_{ij}(\mathbf{x};\boldsymbol{\theta})| &\leq M_2(\mathbf{x}) \\ |g_{ijl}(\mathbf{x};\boldsymbol{\theta})| &\leq M_2(\mathbf{x}),\ i, j, l = 1, \ldots, D \end{aligned} \tag{3.130}$$

(A_5) $\boldsymbol{\theta}_0$ is the unique zero of $E_0 \nabla g(\mathbf{X}_1;\boldsymbol{\theta})$ in $\boldsymbol{\Theta}_0$.

(A_6) $A(\boldsymbol{\theta}) = -E_0 \nabla^2 g(\mathbf{X}_1;\boldsymbol{\theta})$ and $A^{-1}(\boldsymbol{\theta})$ exist and are continous for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and $A(\boldsymbol{\theta}_0)$ is positive definite. $B(\boldsymbol{\theta}) = E_0 \nabla g(\mathbf{X}_1;\boldsymbol{\theta}) \nabla^T g(\mathbf{X}_1;\boldsymbol{\theta})$ and $B^{-1}(\boldsymbol{\theta})$ exist and are continous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$.

(A_7) var $g_{ij}(\mathbf{X}_1;\boldsymbol{\theta}_0) < \infty,\ i, j = 1, \ldots, D$

(A_8) $E_0 |g_i(\mathbf{X}_1;\boldsymbol{\theta})|^{2+\sigma} < \infty,\ i = 1, \ldots, D$ for some $\sigma > 0$.

In deriving the asymptotics of Z_n under the hypothesis $\bar{H}_0$, we follow closely the arguments of Gombay and Horvath [28] and go into the details only where there are differences between the correctly and the misspecified case. For $\mathbf{Y} \in \Re^m$, $|\mathbf{Y}| = \max\{|Y_1|, \ldots, |Y_n|\}$ denotes the maximum norm. As an abbreviation, we use

$$\begin{aligned} Q_K &= \sum_{j=1}^{K} \nabla g(\mathbf{X}_j;\boldsymbol{\theta}_0), \\ Q^*_{n-K} &= \sum_{j=K+1}^{n} \nabla g(\mathbf{X}_j;\boldsymbol{\theta}_0) \end{aligned} \tag{3.131}$$

Then, we have

Lemma 3.2. *If $\bar{H}_0$ and (A_1)-(A_7) hold, we have for $n \to \infty$*

$$\max_{1\le K\le n} \frac{K}{\log\log K}\left|\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}_0 - \frac{1}{K}A^{-1}(\boldsymbol{\theta}_0)Q_K\right| = O_p(1),$$
$$\max_{1\le K\le n} \frac{K}{\log\log n - K}\left|\hat{\boldsymbol{\theta}}^*_{n-K} - \boldsymbol{\theta}_0 - \frac{1}{n-K}A^{-1}(\boldsymbol{\theta}_0)Q^*_{n-K}\right| = O_p(1) \quad (3.132)$$

Proof. The proof proceeds as the proof of the corresponding Lemma 2.1 of Gombay and Horvath ([28]). We only have to remark that

$$\lim_{K\to\infty} \hat{\boldsymbol{\theta}}_K = \boldsymbol{\theta}_0 a.s.$$

in the misspecified case too, as $\hat{\boldsymbol{\theta}}_K$ is an M-estimate of $\boldsymbol{\theta}_0$. Assumptions (A_1)-(A_5) are strong enough to apply Theorem 6.2.2 of Huber [40] which implies strong consistency of $\hat{\boldsymbol{\theta}}_K$. □

Lemma 3.3. *If $\bar{H}_0$ and (A_1)-(A_7) hold, we have for $n \to \infty$*

$$\max_{1\le K\le 1} \frac{K^{1/2}}{(\log\log K)^{3/2}} \quad \Big|L_K(\hat{\boldsymbol{\theta}}_K) - L_K(\boldsymbol{\theta}_0) - \frac{K}{2}(\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}_0)^T A(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta}_0)\Big| = O_p(1),$$
$$\max_{1\le K\le 1} \frac{(n-K)^{1/2}}{(\log\log(n-K))^{3/2}} \quad \Big|L^*_{n-K}(\hat{\boldsymbol{\theta}}^*_{n-K}) - L^*_{n-K}(\boldsymbol{\theta}_0) - \frac{n-K}{2}(\hat{\boldsymbol{\theta}}^*_{n-K} - \boldsymbol{\theta}_0)^T A(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}^*_{n-K} - \boldsymbol{\theta}_0)\Big| = O_p(1) \quad (3.133)$$

The proof proceeds exactly as that one of Lemma 2.2 of Gombay and Horvath ([28]), using Taylor expansions, a uniform law of large numbers and the law of the iterated logarithm. Lemma 1 and 2 together imply

Lemma 3.4. *If $\bar{H}_0$ and (A_1)-(A_7) hold, we have for $n \to \infty$*

$$\begin{aligned} \max_{1\le K\le 1} \frac{K^{1/2}}{(\log\log K)^{3/2}} \quad & \Big| L_K(\hat{\boldsymbol{\theta}}_K) - L_K(\boldsymbol{\theta}_0) \\ & -\frac{1}{2K} Q_K^T A^{-1}(\boldsymbol{\theta}_0) Q_K \Big| = O_p(1), \\ \max_{1\le K\le 1} \frac{(n-K)^{1/2}}{(\log\log(n-K))^{3/2}} \quad & \Big| L^*_{n-K}(\hat{\boldsymbol{\theta}}^*_{n-K}) - L^*_{n-K}(\boldsymbol{\theta}_0) \\ & -\frac{n-K}{2} Q^{*T}_{n-K} A^{-1}(\boldsymbol{\theta}_0) Q^*_{n-K} \Big| = O_p(1) \end{aligned} \tag{3.134}$$

The proof of theorem 1 also uses the following technical Lemma of Gombay and Horvath ([28])

Lemma 3.5. *Let $\boldsymbol{\eta}_m = (\eta_{m1}, \ldots, \eta_{md})^T, m = 1, 2, \ldots,$ be a sequence of i.i.d. random vectors with $E[\boldsymbol{\eta}_m] = 0$ and covariance matrix $E[\boldsymbol{\eta}_m \boldsymbol{\eta}_m^T] = I_d$, the $d \times d$ unit matrix, and*

$$\max_{1\le i\le d} E|\eta_{1i}|^{2+\sigma} < \infty \textit{for some } \sigma > 0.$$

Then, for all t,

$$\lim_{n\to\infty} P\{a(\log n) \max_{1\le K\le n} [\frac{1}{K} \sum_{i=1}^{d} (\sum_{j=1}^{K} \eta_{ji})^2]^{1/2} - b_d(\log n) \le t\} = \exp(-2\exp^{-t}), \tag{3.135}$$

where $a(u) = (2\log u)^{1/2}, b_d(u) = 2\log u + \frac{1}{2} d \log\log u - \log\Gamma(\frac{d}{2})$ with Γ denoting the Gamma function.

The main argument for deriving the asymptotic distribution of Z_n which

does not work in the misspecified case is the following, with $\boldsymbol{\eta}_j = A^{-1/2}(\boldsymbol{\theta}_0)\nabla g(\mathbf{X}_j;\boldsymbol{\theta}_0)$

$$\begin{aligned}\frac{1}{2K}Q_K^T A^{-1}(\boldsymbol{\theta}_0)Q_K &= \frac{1}{2K}(\sum_{j=1}^{K}\boldsymbol{\eta}_j)^T(\sum_{j=1}^{K}\boldsymbol{\eta}_j)\\ &= \frac{1}{2K}(\sum_{i,j=1}^{K}\boldsymbol{\eta}_j^T\boldsymbol{\eta}_i)\\ &= \frac{1}{2K}(\sum_{i,j=1}^{K}\sum_{l=1}^{d}\boldsymbol{\eta}_{jl}^T\boldsymbol{\eta}_{il})\\ &= \frac{1}{2K}\sum_{l=1}^{d}(\sum_{j=1}^{K}\boldsymbol{\eta}_{jl})^2\end{aligned} \tag{3.136}$$

$\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_n$ are independent with, by definition of $\boldsymbol{\theta}_0$, $E_0\boldsymbol{\eta}_j = A^{-1/2}(\boldsymbol{\theta}_0)E_0\nabla g(\mathbf{X}_j;\boldsymbol{\theta}_0) = 0$. In the correctly specified case, we have also for the covariance matrix of $\boldsymbol{\eta}_j$

$$E_0\boldsymbol{\eta}_j\boldsymbol{\eta}_j^T = A^{-1/2}(\boldsymbol{\theta}_0)E_0[\nabla g(\mathbf{X}_j;\boldsymbol{\theta}_0)\nabla^T g(\mathbf{X}_j;\boldsymbol{\theta}_0)]A^{-1/2}(\boldsymbol{\theta}_0) = I_{d.} \tag{3.137}$$

Then, we can apply Lemma 3.5, to derive the asymptotic distribution of

$$\max_{1\le K\le n}\{\frac{1}{2K}Q_K^T A^{-1}(\boldsymbol{\theta}_0)Q_K\}^{1/2}$$

which, by Lemma 3.4, also gives the asymptotic distribution of

$$\max_{1\le K\le n}\{L_K(\hat{\boldsymbol{\theta}}_K) - L_K(\boldsymbol{\theta}_0)\}^{1/2}$$

and, analogously for $L^*_{n-K}(\hat{\boldsymbol{\theta}}^*_{n-K})$ and $L_n\hat{\boldsymbol{\theta}}_n$, which then, by the arguments detailed in Gombay and Horvath ([28]) provides the asymptotic distribution of $Z_n^{1/2}$.

In the misspecified case, however, $A(\boldsymbol{\theta}_K)$ and $B(\boldsymbol{\theta}_K)$ usually do not coincide, and we have to transform the test statistic accordingly. We restrict ourselves to the case of a one-dimensional parameter, i.e. we now have $d = 1, \boldsymbol{\theta} \in \Re$, and $Q_K, A(\boldsymbol{\theta}_0), B(\boldsymbol{\theta}_0)$ e.t.c. are all scalar. Then, we immediately have from Lemma 3.4

Lemma 3.6. *If $\bar{H}_0$ and (A_1)-(A_7) hold and d=1, we have for $n \to \infty$*

$$\begin{aligned} \max_{1\le K\le 1} \frac{K^{1/2}}{(\log\log K)^{3/2}} \quad & \left| \frac{A(\boldsymbol{\theta}_0)}{B(\boldsymbol{\theta}_0)} \{L_K(\hat{\boldsymbol{\theta}}_K) - L_K(\boldsymbol{\theta}_0)\} \right. \\ & \left. -\frac{1}{2K}\frac{Q_K^2}{B(\boldsymbol{\theta}_0)} \right| = O_p(1), \\ \max_{1\le K\le 1} \frac{(n-K)^{1/2}}{(\log\log(n-K))^{3/2}} \quad & \left| \frac{A(\boldsymbol{\theta}_0)}{B(\boldsymbol{\theta}_0)} \{L^*_{n-K}(\hat{\boldsymbol{\theta}}^*_{n-K}) - L^*_{n-K}(\boldsymbol{\theta}_0)\} \right. \\ & \left. -\frac{1}{2K}\frac{(Q^*_{n-K})^2}{B(\boldsymbol{\theta}_0)} \right| = O_p(1) \end{aligned} \tag{3.138}$$

Now, as we have $A(\boldsymbol{\theta}_0)$ replaced by $B(\boldsymbol{\theta}_0)$,

$$\frac{1}{2K}\frac{Q_K^2}{B(\boldsymbol{\theta}_0)} = \frac{1}{2K}(\sum_{j=1}^{K} \boldsymbol{\eta}_j)^2$$

where $\boldsymbol{\eta}_j = B^{-1/2}(\boldsymbol{\theta}_0)\nabla g(\mathbf{X}_j;\boldsymbol{\theta}_0)$ are i.i.d. with mean 0 and variance 1 by definition of $B(\boldsymbol{\theta}_0)$, i.e. we may apply Lemma 3.5.

Lemma 3.6 together with Lemma 3.5 imply immediately the following result by replacing $-2\log\Lambda_K$ in the proof of the Theorem of Gombay and Horvath ([28]) everywhere by $-2\{\frac{A(\boldsymbol{\theta}_0)}{B(\boldsymbol{\theta}_0)}\log\Lambda_K\}$.

Theorem 3.7. *If $\bar{H}_0$ and (A_1)-(A_8) hold and d=1, then we have for all t*

$$\lim_{n\to\infty} P\{a(\log n)\left[\frac{A(\boldsymbol{\theta}_0)}{B(\boldsymbol{\theta}_0)}Z_n\right]^{1/2} \le t + b_1(\log n)\} = \exp(-2\exp^{-t})$$

where $a(u) = (2\log u)^{1/2}, b_1(u) = 2\log u + \frac{1}{2}\log\log u - \log\Gamma(\frac{1}{2})$ with Γ denoting the Gamma function.

In practice, $A(\boldsymbol{\theta}_0), B(\boldsymbol{\theta}_0)$ are not known and have to be replaced by their estimates. We consider the quasi maximum likelihood estimate $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$, and we replace the expectations by sample means to get, where $'$ denotes differentiation w.r.t. $\boldsymbol{\theta}$,

$$\hat{A}_n(\hat{\boldsymbol{\theta}}_n) = -\frac{1}{n}\sum_{j=1}^{n} g''(\mathbf{X}_j;\hat{\boldsymbol{\theta}}_n),\ \hat{B}_n(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{j=1}^{n}[g'(\mathbf{X}_j;\hat{\boldsymbol{\theta}}_n)]^2$$

Theorem 3.8. *If $\bar{H}_0$ and (A_1)-(A_8) hold and d=1, then we have for all t*

$$\lim_{n\to\infty} P\{a(\log n)\left[\frac{\hat{A}_n(\hat{\boldsymbol{\theta}}_n)}{\hat{B}_n(\hat{\boldsymbol{\theta}}_n)}Z_n\right]^{1/2} \le t + b_1(\log n)\} = \exp(-\exp^{-t})$$

Proof. We use the abbreviations $a_n = a(\log n),\ b_n = b_1(\log n), R_0 = \frac{A(\boldsymbol{\theta}_0)}{B(\boldsymbol{\theta}_0)}$ and $R_n = \frac{\hat{A}_n(\hat{\boldsymbol{\theta}}_n)}{\hat{B}_n(\hat{\boldsymbol{\theta}}_n)}$. We have

$$a_n[Z_n R_n]^{1/2} - b_n = (a_n[Z_n R_0]^{1/2} - b_n) + a_n Z_n^{1/2}[R_n^{1/2} - R_0^{1/2}]^{1/2}$$

The first term on the right-hand side has the correct asymptotic distribution by Theorem 3.7. So, we have to show that the second term is $o_p(1)$ for $n \to \infty$. As $b_n \sim 2 \log\log n$ for $n \to \infty$, $a_n Z_n^{1/2} = O_p(\log\log n)$ again by Theorem 3.7. So, we have to show that $R_n^{1/2} - R_o^{1/2}$ is $o_p([\log\log n]^{-1})$. By Theorem 3.3, we have $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ under $\bar{H}_0$. Together with the law of large numbers, we get, applying assumption (A_4),

$$R_n \xrightarrow{p} R_0 \neq 0 \text{ for } n \to \infty.$$

Then, we also have

$$R_n^{1/2} + R_0^{1/2} \xrightarrow{p} 2R_0^{1/2} \neq 0$$

, and, therefore,

$$R_n^{1/2} - R_0^{1/2} = \frac{R_n - R_0}{R_n^{1/2} + R_0^{1/2}}$$

is of the same order as $R_n - R_0$. Now,

$$\hat{B}_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{p} E_0[g'(\mathbf{X}_1; \boldsymbol{\theta}_0)]^2 = B(\boldsymbol{\theta}_0) > 0$$

Using again the law of large numbers and (A_4), it suffices to show that

$$A_n(\hat{\boldsymbol{\theta}}_n) - A(\boldsymbol{\theta}) = o_p(\log n \log n)^{-1}$$

Using a Taylor expansion of order 1 for $g''(\mathbf{x}; .)$ and the boundedness assumption on the third derivatives of g from (A_4), we get

$$\begin{aligned}
\left|A_n(\hat{\boldsymbol{\theta}}_n) - A(\boldsymbol{\theta}_0)\right| &= \left|\frac{1}{n}\sum_{j=1}^{n}\left\{g''(\mathbf{X}_j; \hat{\boldsymbol{\theta}}_n) - E_0 g''(\mathbf{X}_1; \boldsymbol{\theta}_0)\right\}\right| \\
&\leq \frac{1}{n}\sum_{j=1}^{n}\left|g''(\mathbf{X}_j; \hat{\boldsymbol{\theta}}_n) - g''(\mathbf{X}_j; \boldsymbol{\theta}_0)\right| \\
&\quad + \left|\frac{1}{n}\sum_{j=1}^{n}\left\{g''(\mathbf{X}_j; \boldsymbol{\theta}_0) - E_0 g''(\mathbf{X}_1; \boldsymbol{\theta}_0)\right\}\right| \\
&\leq \frac{1}{n}\sum_{j=1}^{n} M_2(\mathbf{X}_j)|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0| \\
&\quad + \left|\frac{1}{n}\sum_{j=1}^{n}\left\{g''(\mathbf{X}_j; \boldsymbol{\theta}_0) - E_0 g''(\mathbf{X}_1; \boldsymbol{\theta}_0)\right\}\right|
\end{aligned}$$

The first term asymptotically coincides with $E_0 M_2(\mathbf{X}_1)|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|$ by the law of large numbers, and $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0$ is of order $O_p(\frac{1}{\sqrt{n}})$ by Theorem 3.3. The second term is also of order $O_p(\frac{1}{\sqrt{n}})$ by the central limit theorem, using (A_7). So, we have

$$A_n(\hat{\boldsymbol{\theta}}_n) - A(\boldsymbol{\theta}_0) = O_p(\frac{1}{\sqrt{n}}) = o_p\left(\frac{1}{\log\log n}\right)$$

and the assertion follows. □

3.13 Some modifications of the changepoint test

In this section, we discuss two variants of the changepoint problem discussed in the previous sections. We restrict ourselves here to the correctly specified case where $p_i = P(Y = 1|\mathbf{X}_i = \mathbf{x}) = Z(\mathbf{x}; \boldsymbol{\theta})$ can be represented by a neural network. The misspecified situation can be handled in the same way as in sections (3.11) and (3.12).

To keep the notation simple, we assume in this section that $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}) \in \Re^2$, $\mathbf{X}_{i1}, \mathbf{X}_{i2} \in \Re$. The generalization to dimension where $\mathbf{X}_{i1} \in \Re^m, \mathbf{X}_{i2} \in \Re^{d-m}$ for some $1 \leq m < d$, is straight forward.

We are interested in a situation where $p_i(\mathbf{x}) = p_i(\mathbf{x}) = p_i(x_1, x_2)$ depends on x_1 only before the potential change, but it is allowed to depend on x_2 too after the change. Recall that $p_i(\mathbf{x}) = Z(\mathbf{x}; \boldsymbol{\theta}_i) = \psi(O_H(\mathbf{x}; \boldsymbol{\theta}_i))$ with O_H given by equation (2.3). We split the weight vector $\boldsymbol{\theta}_i$ into two parts: $\boldsymbol{\theta}_i = (\boldsymbol{\vartheta}_i, \boldsymbol{\tau}_i)$ where $\boldsymbol{\tau}_i \in \Re^H$ consists of all factors of x_2 in the representation of O_H, and $\boldsymbol{\vartheta}_i$ consists of the remaining network parameters. Then, the specific change point testing problem can be written as

$$\begin{aligned}
H_0 \;&:\; \boldsymbol{\vartheta}_1 = \ldots = \boldsymbol{\vartheta}_n,\ \boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_n = \mathbf{0} \\
vs \\
H_1 \;&:\; \boldsymbol{\vartheta}_1 = \ldots = \boldsymbol{\vartheta}_K \neq \boldsymbol{\vartheta}_{K+1} = \ldots = \boldsymbol{\vartheta}_n,\ \mathbf{0} = \boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_K \neq \boldsymbol{\tau}_{K+1} = \ldots = \boldsymbol{\tau}_n \\
&\quad \text{for some } 1 \leq K < n.
\end{aligned}$$

Now, under H_0, we estimate $\boldsymbol{\theta}$ by maximizing $L_0(\boldsymbol{\theta})$ only over those $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ which are of the form $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \mathbf{0})$, i.e. using a similar notation as in section (3.7), the maximum likelihood estimate of $\boldsymbol{\theta}$ under H_0 is of the form

$$\hat{\boldsymbol{\theta}}_0 = (\hat{\boldsymbol{\vartheta}}_0, \mathbf{0}) = (\hat{\alpha}_0^0, \ldots, \hat{\alpha}_H^0,\ \hat{W}_{10}^0, \ldots, \hat{W}_{H0}^0,\ \hat{W}_{11}^0, \ldots, \hat{W}_{H1}^0,\ 0, \ldots, 0)$$

Under the alternative, we maximize $L_{1,K}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$, where again $\boldsymbol{\theta}$ is restricted to the form $(\boldsymbol{\vartheta}, \mathbf{0})$, i.e. the parameter estimates before and after the potential change point are

$$\hat{\boldsymbol{\theta}}_K = (\hat{\boldsymbol{\vartheta}}_K, \mathbf{0}) = (\hat{\alpha}_0^K, \ldots, \hat{\alpha}_H^K,\ \hat{W}_{10}^K, \ldots, \hat{W}_{H0}^K,\ \hat{W}_{11}^K, \ldots, \hat{W}_{H1}^K,\ 0, \ldots, 0),$$

$$\hat{\boldsymbol{\theta}}_K^* = (\hat{\boldsymbol{\vartheta}}_K^*, \hat{\boldsymbol{\tau}}_K^*) = (\hat{\alpha}_0^{*K}, \ldots, \hat{\alpha}_H^{*K},\ \hat{W}_{10}^{*K}, \ldots, \hat{W}_{H0}^{*K},\ \hat{W}_{11}^{*K}, \ldots, \hat{W}_{H1}^{*K},\ \hat{W}_{12}^{*K}, \ldots, \hat{W}_{H2}^{*K}),$$

As for performing the test, we are only interested in the asymptotics of $Q_n = \max_{1 \le K \le n-1}(-2\ \log \Lambda_K^n)$ under the hypothesis H_0, looking at the proofs of Gombay and Horvath [30], remain unchanged if we consider the case $\boldsymbol{\theta} = (\boldsymbol{\vartheta}, \mathbf{0})$. We only have to reparametrize, and use $\boldsymbol{\vartheta}$ as the new parameter, and $\boldsymbol{\Theta}_{\boldsymbol{\vartheta}} = \{\boldsymbol{\vartheta}; \boldsymbol{\theta} = (\boldsymbol{\vartheta}, \mathbf{0}) \in \boldsymbol{\Theta}\}$ as the new parameter set. Thereoms 3.3, 3.4 and 3.5 continue to hold. We only have to recall that, then, D is the dimension of $\boldsymbol{\vartheta}$, not of the full parameter vector $\boldsymbol{\theta}$.

So, we may apply our test also to the specific situation where before a potential change point, certain parameters are shown to be 0. We shall remark that the specific situation does not change the asymptotic distribution of the test statistic under the hypothesis H_0, apart from the different value of D, but, of course, the power will be quite different.

Up to now, we have considered the classical change point problem, where the change happens, if at all, after some fixed time point K in the given sequence of observations. Vexler and Gurevich [66] consider a different change point problem where the change happens once a particular coordinate of the predictor variable or some real-valued function of the predictor variable exceeds a certain bound. Let us consider the most simple case of a change determined by the rise of the first coordinate x_1, i.e. $(Y_i, \mathbf{X}_i),\ i = 1, \ldots, n$, are satisfying $\mathbf{X}_{11} < \mathbf{X}_{21} < \ldots < \mathbf{X}_{n1}$. Then, a usual change point test is applied and, if the hypothesis is rejected, the change is attributed to the size of the $\mathbf{X}_{i1}$'s.

Ordering the data w.r.t. a coordinate of the predictor variable destroys. of course, the independence of the data. Vexler and Gurevich [66] nevertheless consider likelihood ratio statistics Λ_K^n which are calculated pretending that the rearranged data are independent. Then, they prove that the test based on $\max_{1 \le K \le n-1} \Lambda_K^n$ is still working as far as the level is concerned. However, the test is conservative as they only show upper bounds for the level, and, moreover, in calculating the Λ_K^n, they do not use the full maximum likelihood estimates of the parameters, but only suboptimal estimates using part of the data only. This device is necessary for the proofs.

In the real life example, we consider a similar situation, but we do not try to generalize the method of proof of Vexler and Gurevich [66], who only considered logistic regression, to neural networks. Apart from the sub-optimality of the test, there is a more fundamental reason: A change at a certain level of a predictor coordinate is more frequently described by threshold models in the literature than by change point techniques. Essentially, one is looking for a jump of $p(\mathbf{x}) = P(Y_j|\mathbf{X}_j = \mathbf{x})$ at a certain point $x_1 = \zeta$ if the first coordinate is considered. As we are interested in nonparametric function estimates based on neural networks, such jumps should be automatically approximated well if the complexity of the network, i.e. the number of neurons H, is large enough. So, a much more natural test for a change at $x_1 = \zeta$ would be to look at the maximal absolute value of the derivative w.r.t x_1 of the neural network estimate of $p(\mathbf{x})$ which should be large if there is a jump. For investigating the asymptotic behaviour of such a test, reordering the data and destroying the independence would not be necessary. We postpone such considerations to future work, as they would digress too far from the change point methods discussed in this thesis elsewhere.

3.14 Real Data Analysis

In this section, we deal with real data of size $n = 194$ from a lung cancer clinical trial conducted by the Eastern Cooperative Oncology Group, USA. The data can be found at *http://courses.washington.edu/b537/data/ECOGdata.txt* and in the appendix.

The data has four covariates namely:

1. Treatment
   This is a binary covariate taking 0 for treatment A and 1 for treatment B.

2. Age
   The covariate represents the patient's age in years.

3. Time
   This is the follow-up time in weeks.

4. Status
   This is a categorical variable taking the value 0 for Censored, 1 for local spread of the disease and 2 for distant spread of the disease.

The response variable is called Performance Status which is also categorical taking 0 for ambulatory and 1 for non-ambulatory.

Sample Size	$1-\alpha$	R_1	R_2
194	0.90	3.8931	4.4392
	0.95	4.2879	4.6640
	0.99	5.1821	5.1177

Table 3.6: *Asymptotic critical values from equation (3.113), denoted as* R_1 *and from equation (3.118), denoted as* R_2

Our study aims at establishing whether each of the above covariates had a change effect on the patient's performance. However, we note that the effects are mainly artefacts from using the changepoint test to the resorted data which destroys the independence. Table (3.6) gives the critical values for $n = 194$ and $D = 6$. The critical values are based on three covariates and one hidden neuron.

3.14.1 Change Point Detection due to Status

The observations ($\mathbf{Y}$,$\mathbf{X}$) were sorted in ascending order of the status variable, following the approach of Vexler and Gurevich [66]. The predictor $\mathbf{X}$ was partitioned into two sets so that

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$$

where

$$\mathbf{X}_1 = \{Treatment, Age, Time\}$$

and

$$\mathbf{X}_2 = \{Status\}$$

.

Using equations (3.12), (3.15) and (3.49) the likelihood ratio was estimated and the results presented in figure (3.7). From figure (3.7), there is a strong indication that a change point exists. The maximum value is $Q_{194} = 41.0364$ which gives $Q_{194}^{\frac{1}{2}} = 6.4060$. We are therefore able to reject H_0 in equation (3.1) at $90\%, 95\%$ and 99% confidence. There is a strong indication that patient's status affected the performance.

3.14.2 Change Point Detection due to Time

The observations ($\mathbf{Y}$,$\mathbf{X}$) were sorted in ascending order of the time variable and the predictor $\mathbf{X}$ partitioned into two sets so that

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$$

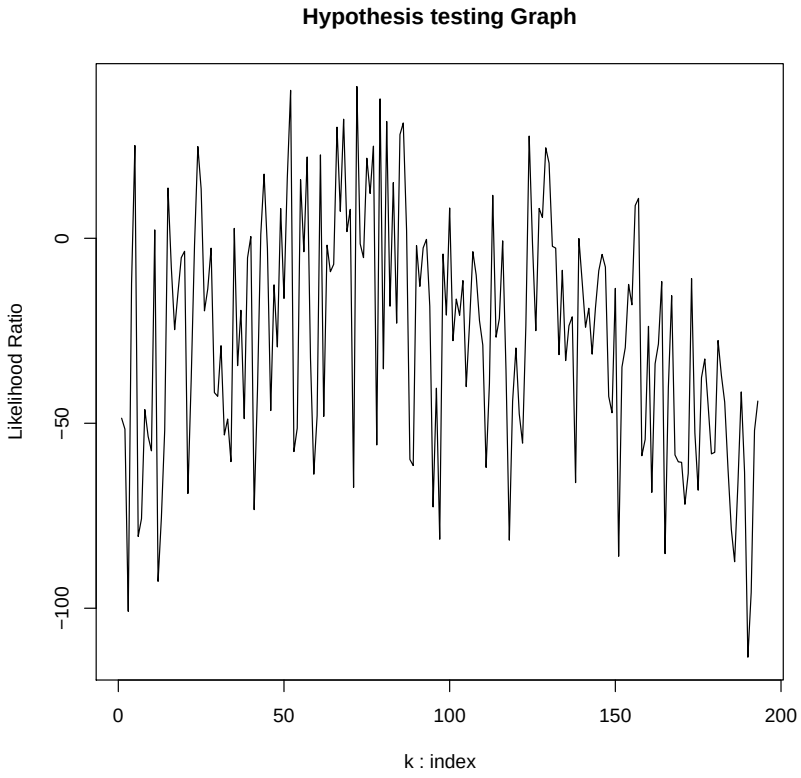

Figure 3.7: *Change Point Detection Graph for the Status covariate*

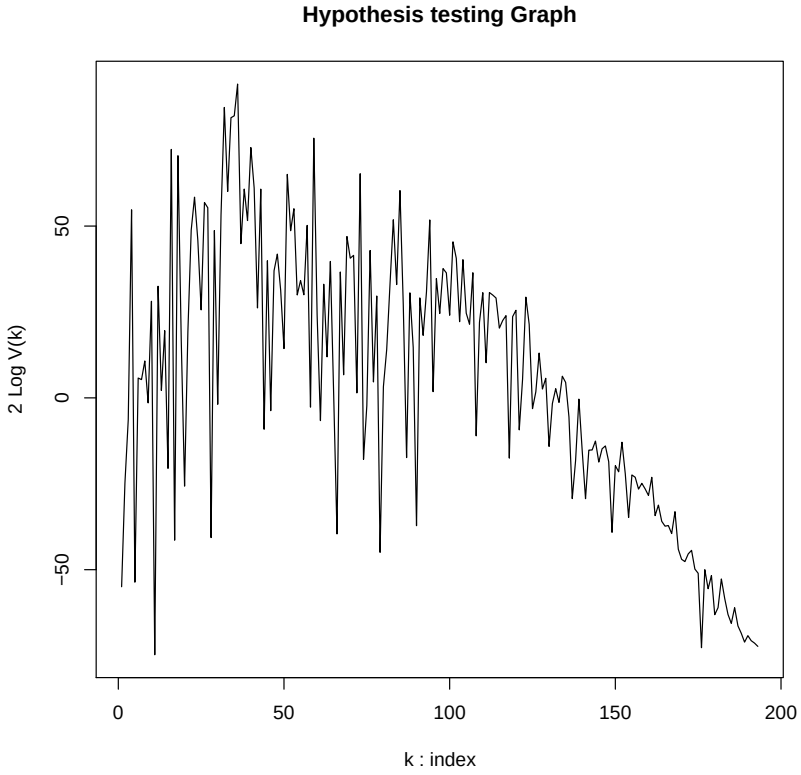

Figure 3.8: *Change Point Detection Graph for the Time covariate*

where

$$\mathbf{X}_1 = \{Treatment, Age, Status\}$$

and

$$\mathbf{X}_2 = \{Time\}$$

.

The likelihood ratio was estimated and the results presented in figure (3.8).

From figure (3.8), there is a strong indication that a change point exists. The maximum value is $Q_{194} = 91.3255$ which gives $Q_{194}^{\frac{1}{2}} = 9.5564$. We are therefore able to reject H_0 in equation (3.1) at $90\%, 95\%$ and 99% confidence. There is a strong indication that patient's follow-up time affected the performance.

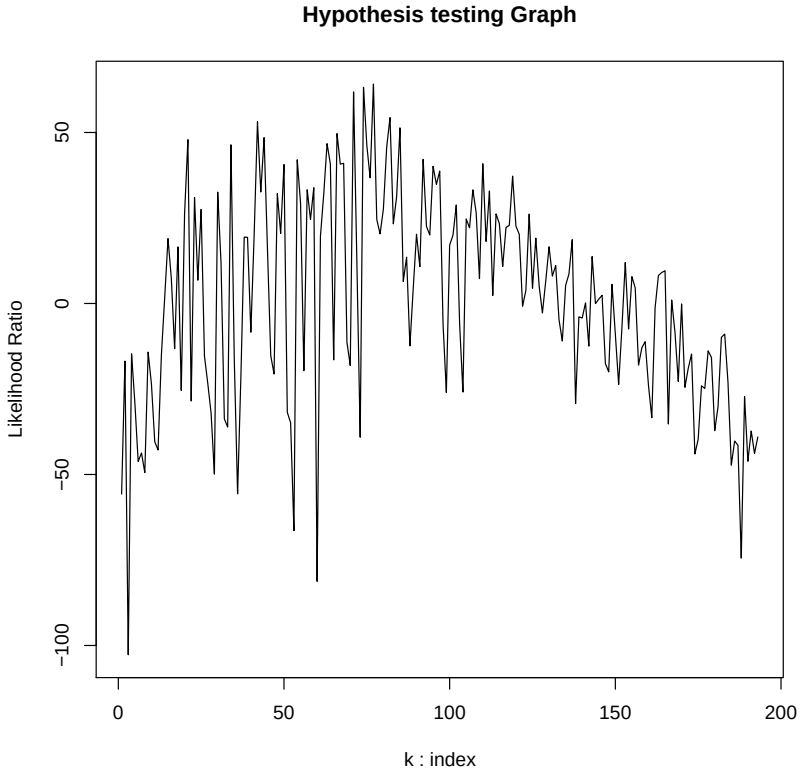

Figure 3.9: *Change Point Detection Graph for the age covariate*

3.14.3 Change Point Detection due to Age

As in sections (3.12.1) and (3.12.2), the observations (**Y**,**X**) were sorted in ascending order of the age variable and the predictor **X** partitioned into two sets so that

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$$

where

$$\mathbf{X}_1 = \{Treatment, Time, Status\}$$

and

$$\mathbf{X}_2 = \{Age\}$$

.

The likelihood ratio was estimated and the results presented in figure (3.9).

From figure (3.9), there is a strong indication that a change point exists. The maximum value is $Q_{194} = 64.1537$ which gives $Q_{194}^{\frac{1}{2}} = 8.0096$. We are

therefore able to reject H_0 in equation (3.1) at $90\%, 95\%$ and 99% confidence. There is a strong indication that patient's age affected the performance.

3.14.4 Change Point Detection due to Treatment

The observations ($\mathbf{Y}$,$\mathbf{X}$) were sorted in ascending order of the Treatment variable and the predictor $\mathbf{X}$ partitioned into two sets so that

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$$

where

$$\mathbf{X}_1 = \{Age, Time, Status\}$$

and

$$\mathbf{X}_2 = \{Treatment\}$$

.

The likelihood ratio was estimated and the results presented in figure (3.10).

From figure (3.10), there is a strong indication that a change point exists. The maximum value is $Q_{194} = 48.2842$ which gives $Q_{194}^{\frac{1}{2}} = 6.9487$. We are therefore able to reject H_0 in equation (3.1) at $90\%, 95\%$ and 99% confidence. There is an indication that the type of treatment had an effect on the performance.

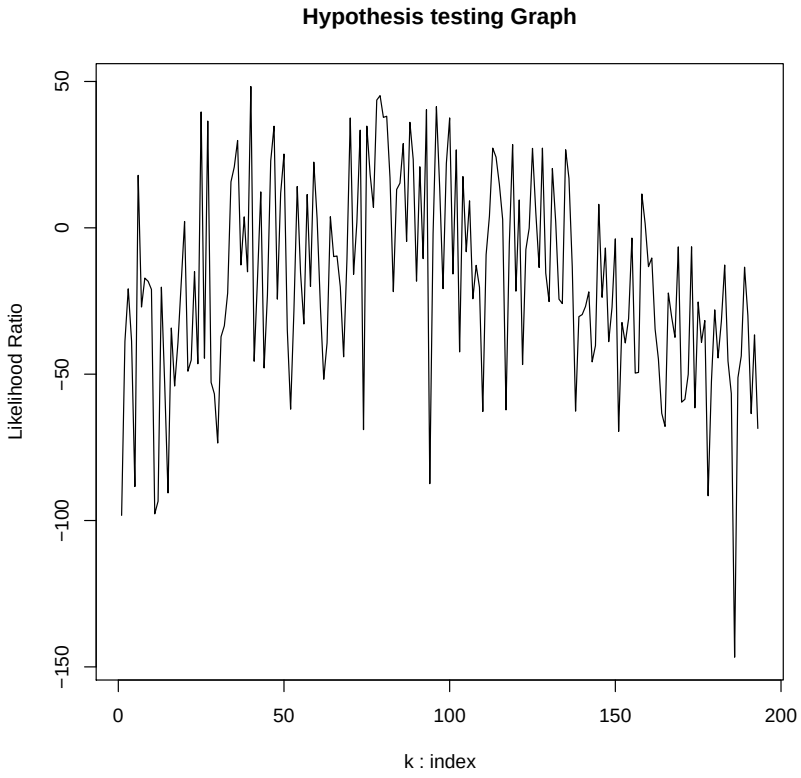

Figure 3.10: *Change Point Detection Graph for the Treatment covariate*

Chapter 4

CHANGE POINT ESTIMATION

We consider the estimation of a change point when it has been detected under equation (3.50). We deal with an independent Bernoulli distributed sequence $(Y_1 \dots Y_n)$.

We discuss and and apply the Maximum Likelihood change point estimation method:

4.1 Maximum Likelihood Method

This method has been studied by many authors. Rukhin [59] studied the asymptotic behaviour of the change point MLE under fixed binomial probabilities as given below:

$$Pr(Y_i = 1) = \begin{cases} \theta_0, i = 1, \dots, K \\ \theta_1, i = K+1, \dots, n \end{cases} \tag{4.1}$$

In this work, Rukhin [59] established the minimum error probability of the change point MLE for fixed binomial probabilities.

Hinkley and Hinkley [37] used the MLE method to estimate the change-point in a sequence of binomial variables when θ_0 and θ_1 are known or unknown.

When θ_0 and θ_1 are known, the likelihood function of $(y_1 \dots y_n)$ is given by:

$$\begin{aligned} L_K(\theta) &= \Pi_{i=1}^{K}[\delta_1(Y_i)\theta_0 + \delta_0(Y_i)(1-\theta_0)]\Pi_{i=K+1}^{n}[\delta_1(Y_i)\theta_1 + \delta_0(Y_i)(1-\theta_1)] \\ & K = 1, \dots, n-1 \end{aligned} \tag{4.2}$$

The maximum likelihood estimate $\hat{K}$ is then given as the value of K which maximizes $L_K(\theta)$ which can be written as:

$$\begin{aligned}\hat{K} &= \arg\max_{1\le K\le n-1}\left\{\sum_{i=1}^{K}[\delta_1(Y_i)\ln\theta_0+\delta_0(Y_i)\ln(1-\theta_0)]+\right. \\ &\left.+\sum_{i=K+1}^{n}[\delta_1(Y_i)\ln\theta_1+\delta_0(Y_i)\ln(1-\theta_1)]\right\},\ K=1,\ldots,n-1\end{aligned} \tag{4.3}$$

When θ_0 and θ_1 (or one of them) are unknown, Hinkley and Hinkley [37] proposed replacing them in equation (4.3)with their estimates so that

$$\begin{aligned}\hat{K} &= \arg\max_{1\le K\le n-1}\left\{\sum_{i=1}^{K}\left[\delta_1(Y_i)\ln\overline{Y}_{1,K}+\delta_0(Y_i)\ln(1-\overline{Y}_{1,K})\right]+\right. \\ &\left.+\sum_{i=K+1}^{n}\left[\delta_1(Y_i)\ln\overline{Y}_{K+1,n}+\delta_0(Y_i)\ln(1-\overline{Y}_{K+1,n})\right]\right\},\ K=1,\ldots,n-1\end{aligned} \tag{4.4}$$

where

$$\overline{Y}_{1,K}=\frac{1}{K}\sum_{i=1}^{K}Y_i \tag{4.5}$$

and

$$\overline{Y}_{K+1,n}=\frac{1}{n-K}\sum_{i=K+1}^{n}Y_i \tag{4.6}$$

However, recent research focuses on conditional mean function where

$$P(Y_i=1|\mathbf{X}_i=\mathbf{x})=\begin{cases}p(\mathbf{x};\boldsymbol{\theta}_0), i=1,\ldots,K\\ p(\mathbf{x};\boldsymbol{\theta}_1), i=K+1,\ldots,n\end{cases} \tag{4.7}$$

where $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$ and K are unknown. Equation (4.7) above can be written as

$$P(Y_i=1|\mathbf{X}_i=\mathbf{x})=p(\mathbf{x};\boldsymbol{\theta}_0)\mathbf{1}_{i\le K}+p(\mathbf{x};\boldsymbol{\theta}_1)\mathbf{1}_{i\ge K+1} \tag{4.8}$$

The regression function $p(\mathbf{x};\boldsymbol{\theta})$ can be linear or non-linear in the parameter vector $\boldsymbol{\theta}$. We deal with the non-linear case.

In line with equation (4.4), $\hat{K}$ is then given by

$$\begin{aligned}\hat{K} &= \arg\max_{1\le K\le n-1}\left\{\sum_{i=1}^{K}\left[\delta_1(Y_i)\ln p(\mathbf{X}_i;\hat{\boldsymbol{\theta}}_0^{1,K})+\delta_0(Y_i)\ln(1-p(\mathbf{X}_i;\hat{\boldsymbol{\theta}}_0^{1,K}))\right]+\right. \\ &\left.+\sum_{i=K+1}^{n}\left[\delta_1(Y_i)\ln p(\mathbf{X}_i;\hat{\boldsymbol{\theta}}_1^{K+1,n})+\delta_0(Y_i)\ln(1-p(\mathbf{X}_i;\hat{\boldsymbol{\theta}}_1^{K+1,n}))\right]\right\} \\ &K=1,\ldots,n-1\end{aligned} \tag{4.9}$$

where $p(\mathbf{x}; \hat{\boldsymbol{\theta}}_0^{1,K})$ is the mean function estimated from $(Y_i, \mathbf{X}_i)_{i=1}^K$. In particular, $\theta_0^{1,K}$ is the parameter vector estimated from $(Y_i, X_i)_{i=1}^K$. Similarly, $p(\mathbf{x}_i; \boldsymbol{\theta}_1^{K+1,n})$ is the mean function estimated from $(Y_i, \mathbf{X}_i)_{i=K+1}^n$.

For Bernoulli outcomes, $E(Y|\mathbf{X} = \mathbf{x}) = p(\mathbf{x}; \boldsymbol{\theta})$ is bounded by 0 and 1. The logistic function as described in chapter 2 is therefore a natural choice for modeling $p(\mathbf{x}; \boldsymbol{\theta})$. As before, we consider the case where

$$P(Y_i = 1|\mathbf{X}_i = \mathbf{x}) = \begin{cases} Z(\mathbf{x}; \boldsymbol{\theta}_0), & i = 1, \ldots, K \\ Z(\mathbf{x}; \boldsymbol{\theta}_1), & i = K+1, \ldots, n \end{cases} \tag{4.10}$$

Similar to Fan *et al* [18] and Frölich [25], we get as the likelihood under the alternative, i.e. under equation (4.10);

$$\begin{aligned} L_{1,K}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \Pi_{i=1}^K \left[\delta_1(Y_i) \left\{Z(\mathbf{X}_i; \boldsymbol{\theta})\right\}\right. \\ &\quad \left. + \delta_0(Y_i) \left\{1 - Z(\mathbf{X}_i; \boldsymbol{\theta})\right\}\right] \\ &\quad * \Pi_{i=K+1}^n \left[\delta_1(Y_i) \left\{Z(\mathbf{X}_i; \boldsymbol{\theta}^*)\right\}\right. \\ &\quad \left. + \delta_0(Y_i) \left\{1 - Z(\mathbf{X}_i; \boldsymbol{\theta}^*)\right\}\right] \end{aligned} \tag{4.11}$$

Analogous to Rukhin [59], we now define the MLE of the change point $\hat{K}_n$ as:

$$\hat{K}_n = \arg_{1 \le k \le (n-1)} \max_{\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \boldsymbol{\Theta}} L_{1,K}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \arg \max_{1 \le k \le (n-1)} L_{1,K}(\hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}_K^*) \tag{4.12}$$

where $\hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}_K^*$ are the ML estimates of $\boldsymbol{\theta}, \boldsymbol{\theta}^*$ under the alternative of a change point at K.

4.2 Simulation Study

The simulation study under H_1 was carried out to investigate the behaviour of the change point estimator under the following considerations:

1. Sample Size
2. Change Point Location
3. Size of Change

Sample Size	$\bar{K}_n$	MSE
$n = 150$	75	3.3815
$n = 200$	100	2.4135

Table 4.1: *Change Point Estimates of K and the Mean Squared Errors (MSE)*

1. Sample Size

The following model was used

$$P(Y_i = 1|\mathbf{X}_i = \mathbf{x}) = \begin{cases} (1 + exp(-(-1.5 + x_1 + x_2)))^{-1}, & i \leq K \\ (1 + exp(-(-1.5 + 2 * x_1 + 1.8 * x_2)))^{-1}, & K < i \leq n \end{cases} \tag{4.13}$$

Two different sample sizes were used, $n = 150$ and $n = 200$. The change point was fixed at half the sample size, i.e. $\lambda = 0.5$ where

$$\lambda = \frac{K}{n}$$

and K is the actual change point. Therefore, for $n = 150$, K was fixed at 75 and for $n = 200$, K was fixed at 100.

We then generated $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ as $uniform[0, 1]$. We then generated the Bernoulli random variables Y_i in line with equation (4.13).

Using equation (4.12), 2000 simulations were carried out to estimate the change point and the results presented in figure (4.1), figure (4.2), figure(4.3) and table (4.1) below.

Figure (4.1) illustrates the maximum log likelihood $\ln L_{1,K}(\hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}^*_K)$ plotted against K for one particular example.

Table (4.1) shows the mean (rounded to integer) and the mean-squared error of the change point estimate. Figures (4.2)-(4.6) show the histograms of the $\hat{K}_n$ for various cases.

Discussion

Using equation (4.12),figure (4.2) and table (4.1), the average estimated change point for $n = 150$ and $K = 75$ was found to be $\bar{K}_{150} = 75$ (rounded to integer) against the true one, $K = 75$. Similarly, from figure (4.3) and table (4.1), the average estimated change point for $n = 200$ and $K = 100$ was found to be $\bar{K}_{200} = 100$ against the true one, $K = 100$. From table (4.1) above, the MSE decreases with increase in sample size. Table (4.1) and Figures (4.2), (4.3) imply that $\bar{K}_n$ is presumably a consistent and asymptotically normal estimate if K is somewhere in the center of the sample. The proof of such a result is postponed to future work.

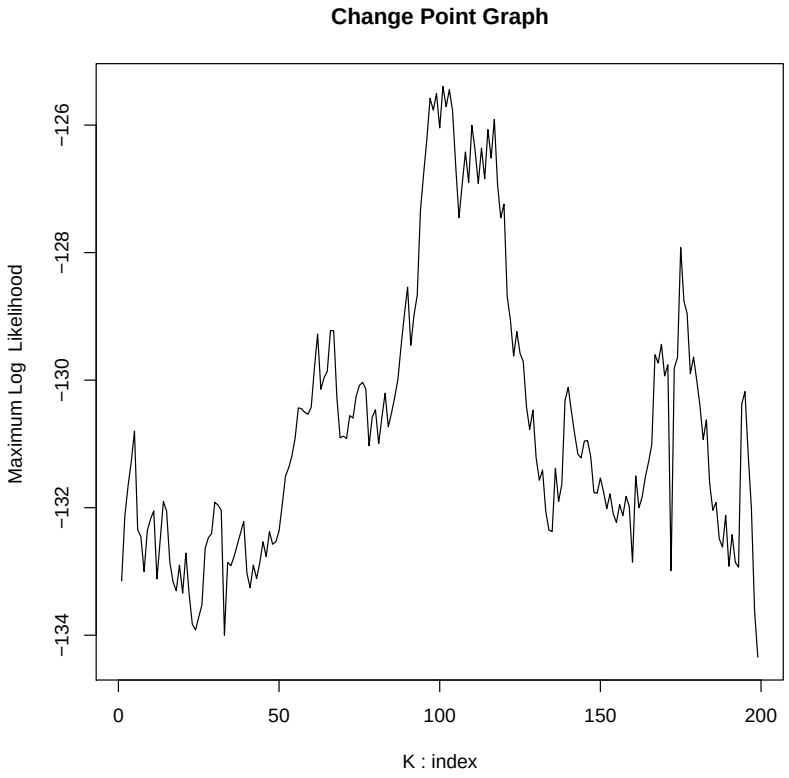

Figure 4.1: *Maximum log likelihood graph for* $n = 200$

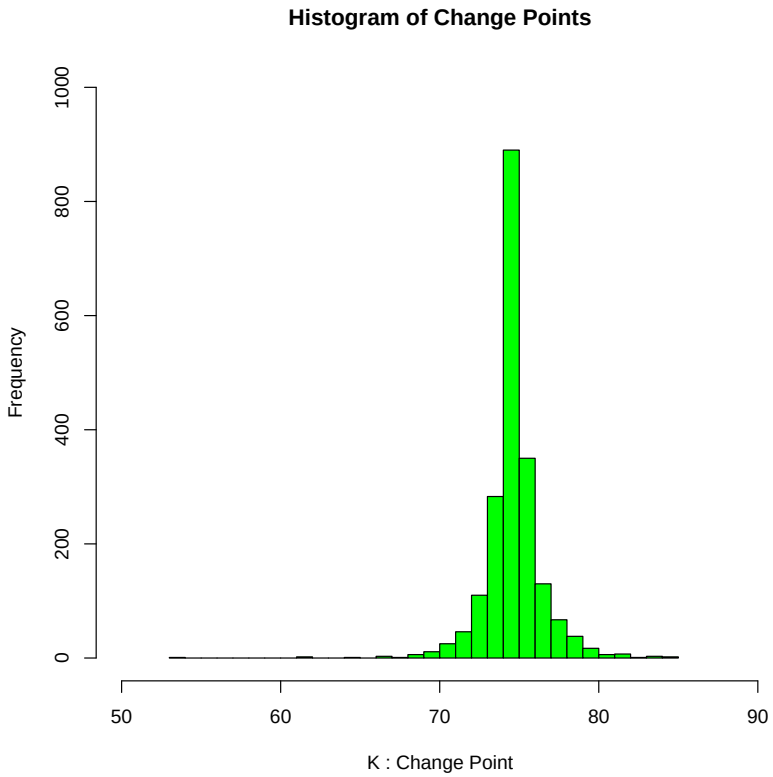

Figure 4.2: *Empirical Distribution of the Change point estimates for* $n = 150$ *and* $K = 75$.

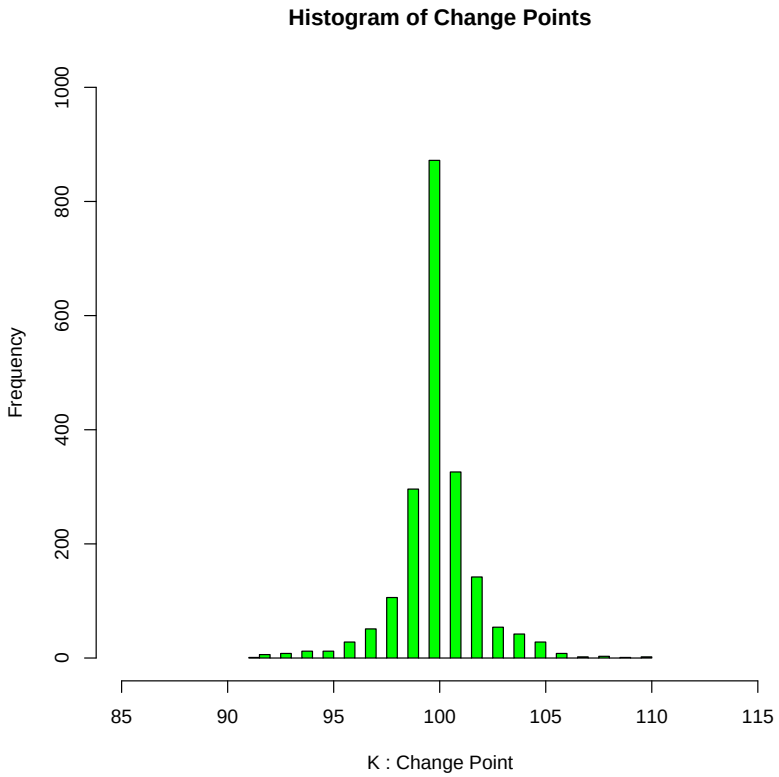

Figure 4.3: *Empirical Distribution of the Change point estimates for* $n = 200$ *and* $K = 100$.

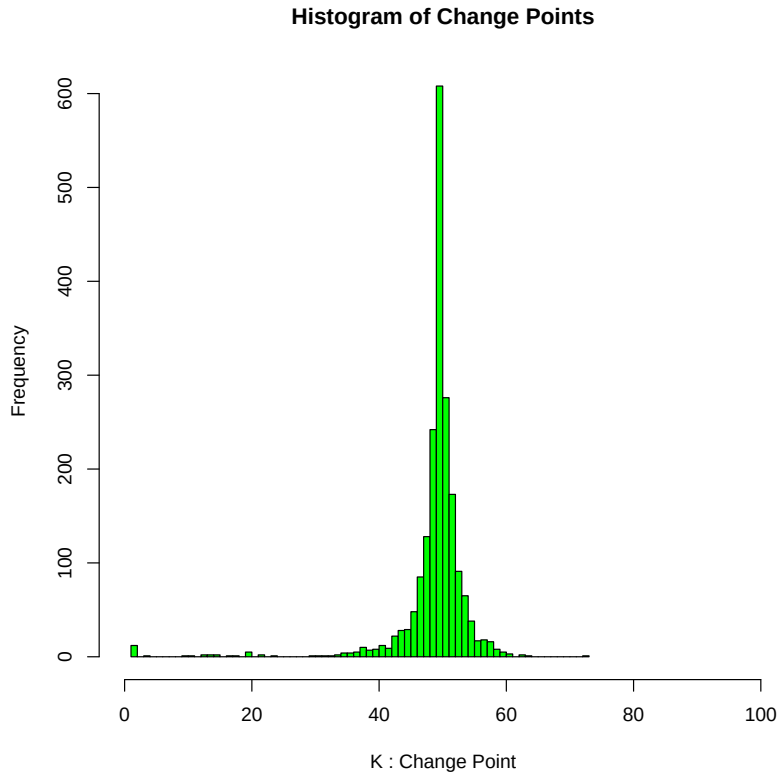

Figure 4.4: *Empirical Distribution of the Change point estimates for* $n = 100$ *when there is a change of* $\Delta = 1.2$.

2. Size of Change

We represent the size of change as Δ where

$$\Delta^2 = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$$

For example, from equation (4.13), we have;

$$\Delta^2 = [-1.5 - (-1.5)\ 1 - 2\ 1 - 1.8]^2 = 3.24$$

. This implies that

$$\Delta = 1.8$$

The Bernoulli random variables were generated in line with equation (4.13). The sample size was fixed at $n = 100$. When there was a change, it was fixed at $K = 50$. The results are found in table (4.2), figure (4.6), figure (4.4) and figure (4.5).

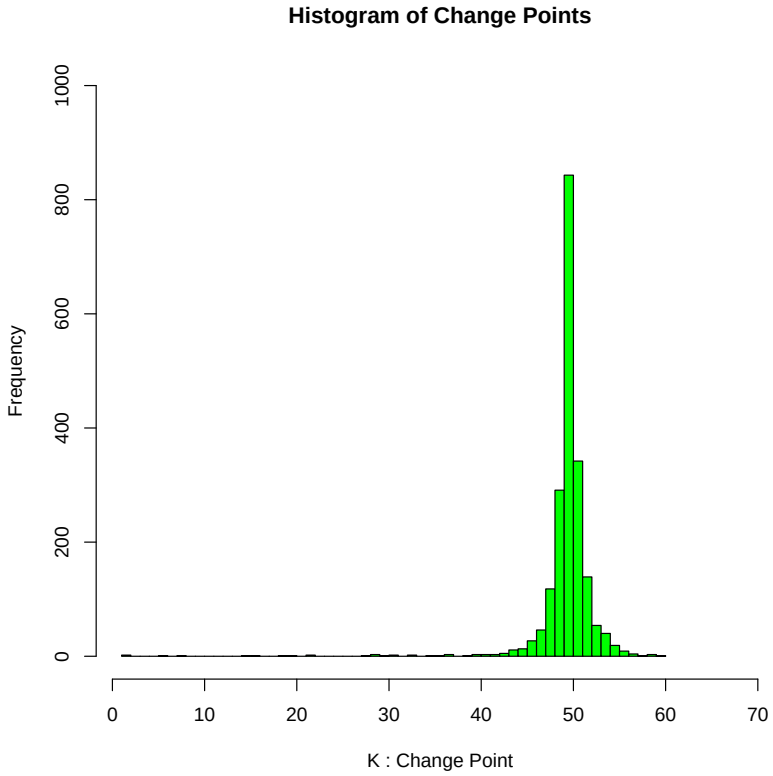

Figure 4.5: *Empirical Distribution of the Change point estimates for* $n = 100$ *when there is a change of* $\Delta = 1.8$.

Size of Change,Δ	MSE
1.2	36.7545
1.8	12.802

Table 4.2: *Change Point Estimates of K and the Mean Squared Errors (MSE) for different sizes of change*

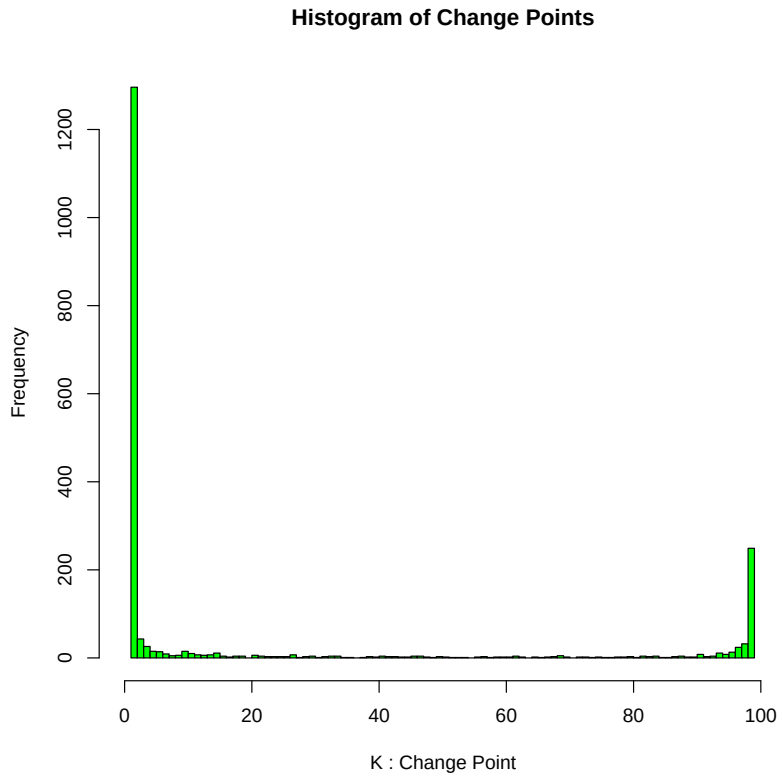

Figure 4.6: *Empirical Distribution of the Change point estimates for* $n = 100$ *when there is no change, i.e* $\Delta = 0$.

Discussion

It is clear from table (4.2) that the MSE decreases with the increase in size of change as expected. Of special interest is the behaviour of the estimator under H_0. Figure (4.6) indicates that the change point estimator concentrates around the two end points of the sample. This implies that under H_0,

$$\frac{\hat{K}_n}{n} \xrightarrow{P} \{0, 1\}$$

. This argument is supported in Gombay and Horvath [29] Theorem 1.2 where they show that under H_0 and when the necessary conditions are satisfied,

$$\frac{\hat{K}_n}{n} \xrightarrow{D} \xi_0$$

where

$$P(\xi_0 = 0) = P(\xi_0 = 1) = \frac{1}{2}$$

This conclusion follows from the fact that

$$\lim_{n\to\infty} P\{\hat{K}_n \leq \frac{n}{\log n} \text{ or } \hat{K}_n \geq n - \frac{n}{\log n}\} = 1$$

and that under H_0

$$\{L_{1,K}(\hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}^*_K), 1 \leq K \leq n-1\} \stackrel{D}{=} \{L_{1,K}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^*), 1 \leq K \leq n-1\}$$

3. Change Point Location

The model used here was

$$P(Y_i = 1|X_i = \mathbf{x}) = \begin{cases} (1 + exp(-(-1.5 + 2 * x_1 + 0.4 * x_2)))^{-1}, & i \leq K \\ (1 + exp(-(-1.5 + 2 * x_1 + 1.6 * x_2)))^{-1}, & K < i \leq n \end{cases} \tag{4.14}$$

The Bernoulli random variables ,Y_i,were generated as in equation (4.14). The sample size was fixed at $n = 100$ while the size of the change was fixed at $\Delta = 1.2$.

In order to study the effect of change point location on the MSE, three different change point locations were used, $K = 20, 50, 80$.

The results are presented in figure(4.7) and table (4.3).

Discussion

From table (4.3), the MSE is lowest when the change point is away from the data edges. It is therefore more probable to estimate a change point accurately from the data edges than when it is at the edges.

$n = 100$, $\Delta = 1.2$	
Actual Change Point, K	MSE
20	12.8405
50	2.816
80	29.1565

Table 4.3: *Change point mean squared errors (MSE) for different change point locations*

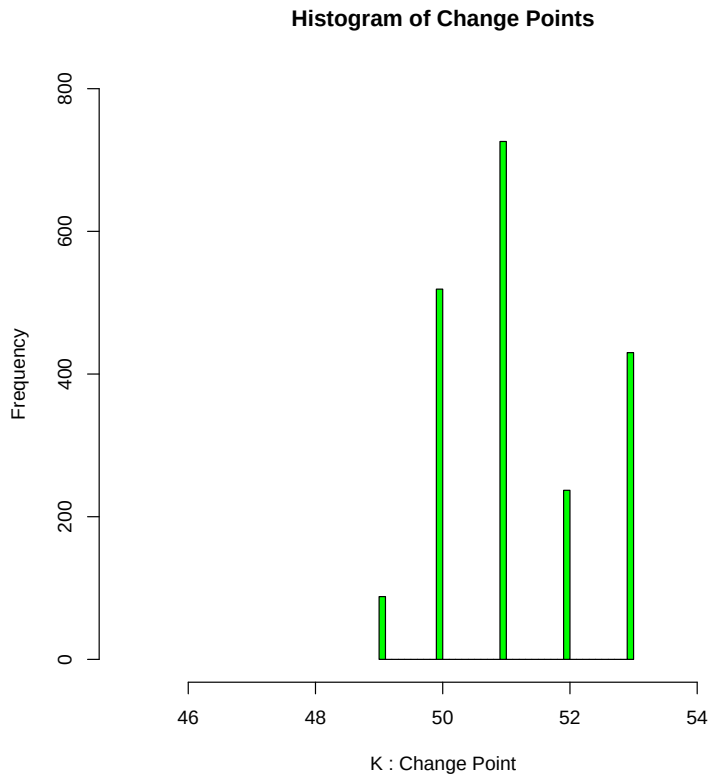

Figure 4.7: *Empirical distribution of the change point estimates for* $n = 100$ *when the actual change point is at* $K = 50$.

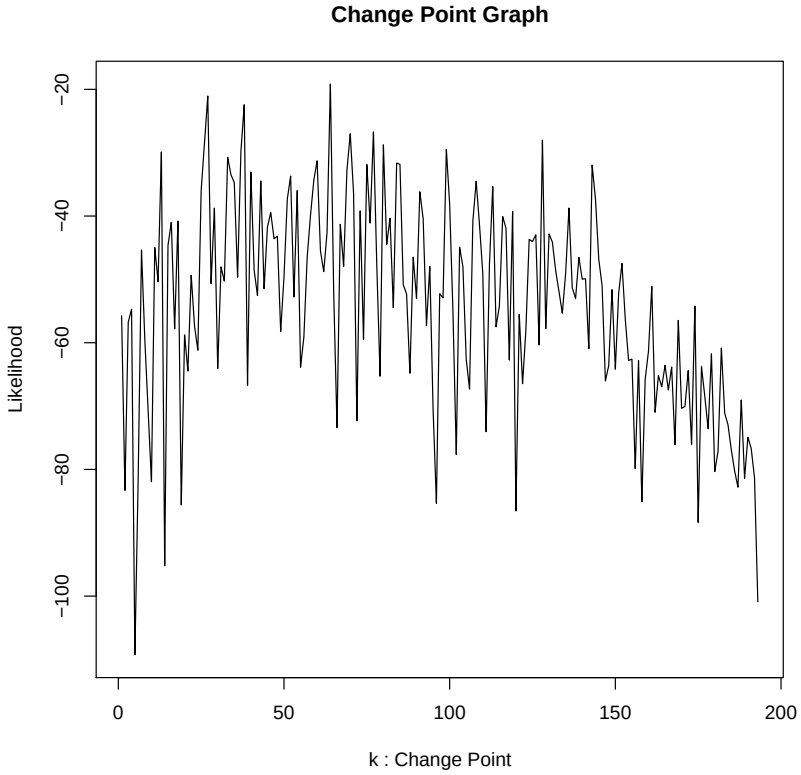

Figure 4.8: *Status Change Point Graph. From this graph,*$\hat{K}_{194} = 64$.

4.3 Real Data Analysis

In this section, we apply the maximum likelihood change point estimation method to the cancer data described in section (3.14).

From section (3.14), change was detected in all the four covariates. We now estimate the location of those changes.

4.3.1 Change Point Estimation due to Status

The data was sorted as in subsection (3.12.1). The change point was estimated using equation (4.12). The results are presented in figure (4.8). From figure (4.8), $\hat{K}_{194} = 64$ which corresponds to the change of status from "0 Status" to "1 Status". This suggests that the patient's performance depended on the status..

There also seems to be another change point around $K = 150$ which corresponds to the change of status from "1 Status" to "2 Status".

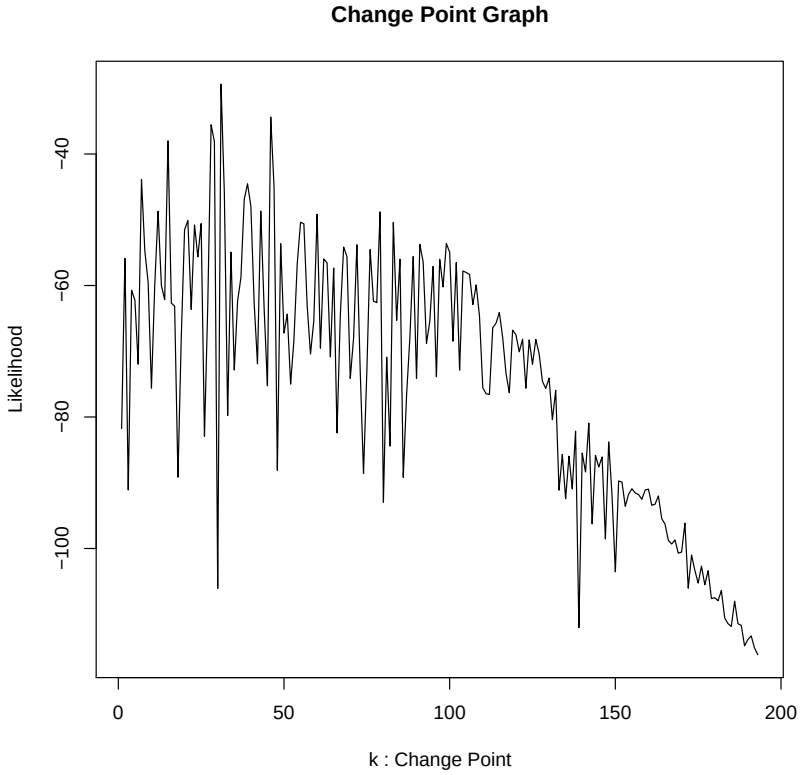

Figure 4.9: *Time Change Point Graph. From this graph,*$\hat{K}_{194} = 31$.

4.3.2 Change Point Estimation due to Time

The data was sorted as in subsection (3.12.2) and the estimation results presented in figure (4.9). From this figure, $\hat{K}_{194} = 31$ which corresponds to a time of 5 weeks. This implies that a follow-up of at least 5 weeks has an effect on the performance.

4.3.3 Change Point Estimation due to Age

The data was sorted as in subsection (3.12.3) and the estimation results presented in figure (4.10). From this figure, $\hat{K}_{194} = 73$ which corresponds to an age of 57 years. This implies that patient's performance changed at around 57 years of age.

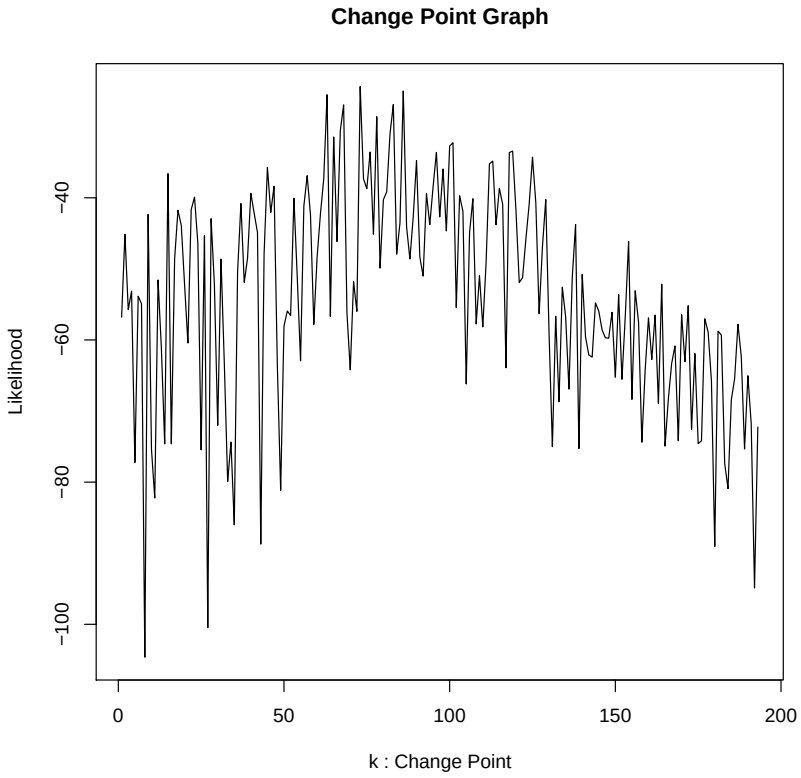

Figure 4.10: *Age Change Point Graph. From this graph,*$\hat{K}_{194} = 73$.

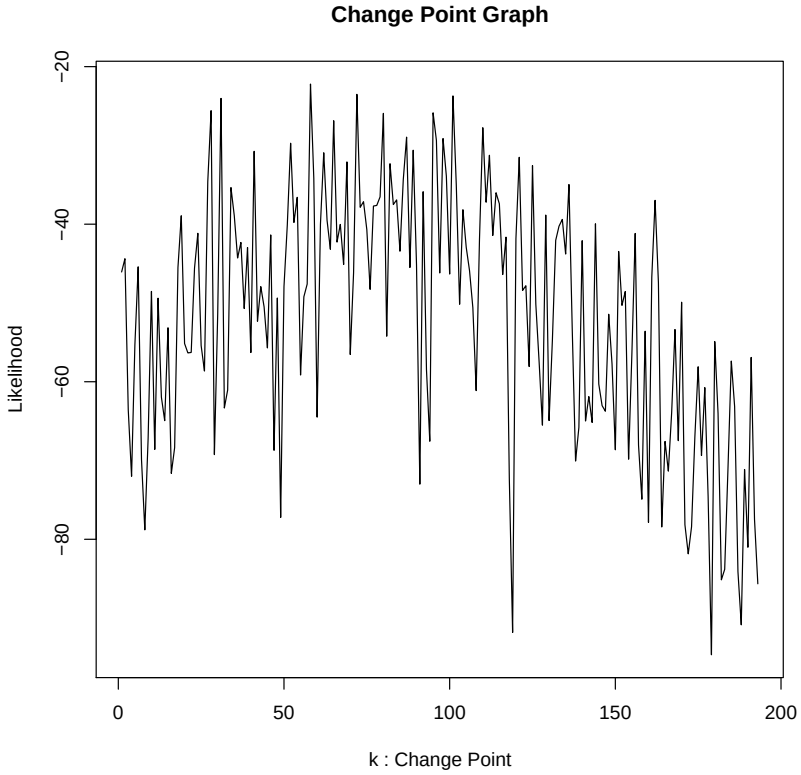

Figure 4.11: *Treatment Change Point Graph. From this graph,*$\hat{K}_{194} = 58$.

4.3.4 Change Point Estimation due to Treatment

The data was sorted as in subsection (3.12.4) and the estimation results presented in figure (4.11). From this figure, $\hat{K}_{194} = 58$ which roughly corresponds to the change from treatment "'A' to treatment "B". The type of treatment therefore affected the performance.

Chapter 5

CONFIDENCE INTERVAL FOR THE CHANGE POINT

Various methods for determining change point confidence intervals exist in the literature. One method involves the asymptotic distribution of $\hat{K} - K$, where K is the true change point and $\hat{K}$ is its estimate. See, for example, Hinkley and Hinkley [37] and Feder [21, 20].

Another approximation method involves the use of bootstrap methods. See for instance Hall [33], Efron and Tibshirani [15], Davison and Hinkley [13] and Pastor-Barriuso et al [57].

Also, log-likelihood ratio method has been used by Cook and Weisberg [10],Zhan *et al* [73] and Pastor and Guallar [56] among others to approximate the Confidence Interval (C.I).

In this study, we discuss and apply the last two methods.

5.1 Construction of profile log-likelihood ratio confidence intervals for the change point

The idea of the profile log-likelihood ratio method is based on observing that, e.g. [10], twice the log likelihood ratio, i.e. $2 \log \Lambda_K^n$, is asymptotically chi-square distributed under the H_0. The degrees of freedom depend on the dimension of $\boldsymbol{\theta}$ under H_0 and H_1. If, for example, H_0 and H_1 differ by one real parameter being 0 or not, we get an asymptotic χ_1^2-distribution

By taking the natural logarithm on both sides of equation (4.11), we have

$$\rho(K, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^*) = \ln(L_{1,K}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^*)) \tag{5.1}$$

From equation (5.1), we have the following likelihood ratio statistic;

$$\eta_K = 2\rho(\hat{K}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^*) - 2\rho(K, \hat{\boldsymbol{\theta}}(K), \hat{\boldsymbol{\theta}}^*(K)) \tag{5.2}$$

where $\hat{\boldsymbol{\theta}}(K)$ and $\hat{\boldsymbol{\theta}}^*(K)$ are the ML estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ given K

Cosequently, the $100(1-\alpha)$ percent confidence interval consists all the values of k in equation (5.2) which satisfy the following condition;

$$\eta_K \leq \chi^2_{1,1-\alpha} \tag{5.3}$$

where $\chi^2_{1,1-\alpha}$ is the $100(1-\alpha)$ percentile of the chi-square distribution with one degree of freedom.

We now discuss the estimation of the upper limit of the confidence interval. The lower part follows a similar procedure.
We begin at the MLE $(\hat{K}, \hat{\boldsymbol{\theta}}_K, \hat{\boldsymbol{\theta}}^*_K)$ and then increase K to $\hat{K}+1$. We then determine the value of $\Lambda_{1,\hat{K}+1}$ in equation (4.11) and consequently equation (5.2) yields the first point in the upper part of the curve. The procedure continues until the condition in equation (5.3) is violated.

5.2 Simulation Study

The model below was used

$$P(Y_i = 1|\mathbf{X}_i = \mathbf{x}) = \begin{cases} (1 + exp(-(-1.5 + 1.8 * x_{1i} + 0.2 * x_{2i})))^{-1}, & i \leq k \\ (1 + exp(-(-1.5 + 2 * x_{1i} + 0.6 * x_{2i})))^{-1}, & k < i \leq n \end{cases} \tag{5.4}$$

where the sample size was fixed at $n = 100$ and the change point at $K = 50$. The Bernoulli random variables were generated in line with equation (5.4) while $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ were generated as *uniform*[0,1]. 2000 simulations of the data in equation (5.4) were done. In each simulation, the profile log-likelihood ratio confidence interval was determined as shown in figure (5.1) below. Table (5.1) represents the percentage mis-coverage under the profile log-likelihood ratio confidence interval method.

Discussion
From figure (5.1), the 90% confidence interval for $\hat{k}$ is 46 - 53. Similarly, the 95% confidence interval for $\hat{k}$ is 45 - 53. Our results indicate that the confidence interval for the change point is not symmetrical. These results are supported by Cook and Weisberg [10].

Table (5.1) shows the percentage of times the Profile log-likelihood ratio Confidence Interval missed the true change point, $K = 50$, on the left and

# Simulations = 2000		
Confidence Level	% Miss Left	% Miss Right
90%	0.0015	0.009
95%	0.0	0.001

Table 5.1: *Results for 2000 confidence interval realizations for the change point K from data of size $n = 100$ generated as in equation (5.4). " %Miss left" represents the percentage of times the left endpoint of the estimated interval was greater than the true change point, $K = 50$. " %Miss Right" represents the percentage of times the right endpoint of the estimated interval was less than the true change point, $K = 50$.*

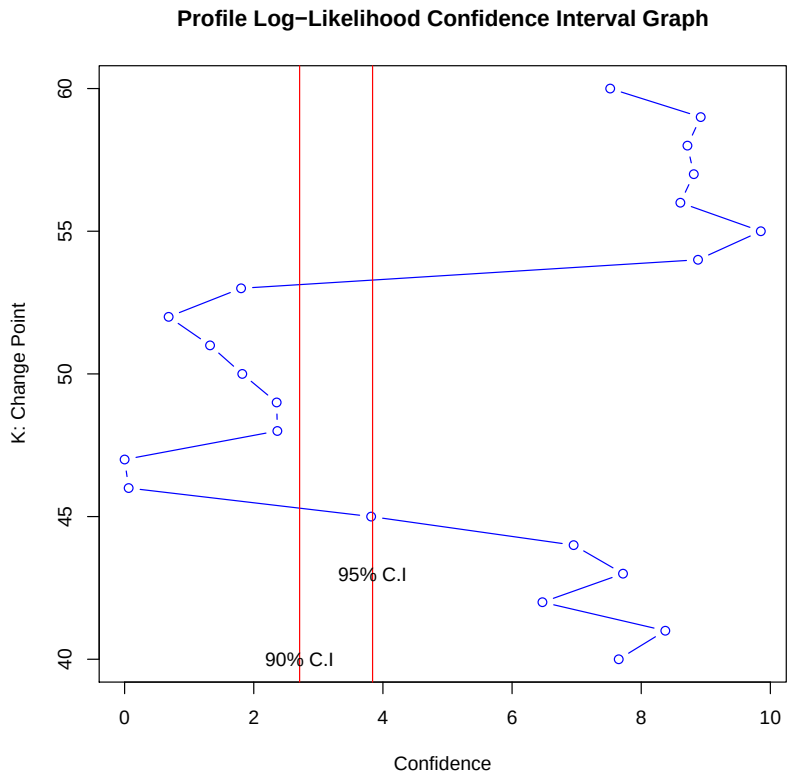

Figure 5.1: *Change Point Confidence Curve. The values of K that satisfy equation (5.3) are found on the left-hand side of a given confidence line*

right hand sides, in 2000 simulated samples. The target mis-coverage for 90% confidence interval is 5% on both sides. Similarly, the target mis-coverage for 95% confidence interval is 2.5% on both sides.

From Table (5.1), it is clear that Profile log-likelihood ratio Confidence Interval method overcovers on the left and on the rightside.

Profile log-likelihood is not without shortfalls. First, many iterations are required to arrive to the global minimum. Also, at each value of K, the LMLEs of θ_0 and θ_1 have to be obtained which is computationally intensive especially for large sample sizes. It was also noted that in case, at a given value of k, the global minimum was not reached, the confidence region had disjoint intervals. This problem can be solved by optimizing the approximation rate.

5.3 Percentile Bootstrap Confidence Interval for the Time of Change

In this section, we approximate the distribution of $\hat{K}_n - K$ using the percentile bootstrap technique.

Percentile Bootstrap Procedure

1. Given the original sample $\{Y_i, \mathbf{X}_i\}_{i=1}^n$, estimate the ML estimates $\hat{K}_n, \hat{\boldsymbol{\theta}}_{\hat{K}_n}$ and $\hat{\boldsymbol{\theta}}^*_{\hat{K}_n}$.

2. From the original covariate vector $\{\mathbf{X}_i\}_{i=1}^n$, get a bootstrap sample $\{\mathbf{X}_i^*\}_{i=1}^n$. This is done by drawing the integers $1, \ldots, n$ with replacement to get a sample from $\{\mathbf{X}_i, \ldots, \mathbf{X}_n\}$.

3. Evaluate $\{Y_i^*\}_{i=1}^n$ corresponding to the bootstrap sample $\{\mathbf{x}_i^*\}_{i=1}^n$ in step 2 above. This is done as follows:

$$Y_i^* \sim B(1, Z(\mathbf{X}_i; \hat{\boldsymbol{\theta}}_{\hat{K}_n}))I_{i \leq \hat{K}_n} + B(1, Z(\mathbf{X}_i; \hat{\boldsymbol{\theta}}^*_{\hat{K}_n}))I_{i > \hat{K}_n} \tag{5.5}$$

4. Using the bootstrap sample $\{Y_i^*, \mathbf{X}_i^*\}_{i=1}^n$, replicate the estimated time of change $\hat{K}_n^*$.

5. Repeat steps 2 to 4 S times. This step gives us S independent bootstrap samples $\{Y_i^{*1}, \mathbf{X}_i^{*1}\}_{i=1}^n, \ldots, \{Y_i^{*S}, \mathbf{X}_i^{*S}\}_{i=1}^n$ from which we get $\hat{K}_n^{*1}, \ldots, \hat{K}_n^{*S}$ respectively.

6. Arrange the bootstrap change point vector $\hat{K}_n^{*1}, \ldots, \hat{K}_n^{*S}$ in ascending order.

From these replicates, we are then able to estimate the distribution function of $\hat{K}_n^* - \hat{K}_n$ where $\hat{K}_n^*$ is the time of change estimate of the re-samples.

Supposing that $K_{\alpha/2}^*$ and $K_{1-\alpha/2}^*$ are the quantiles of $\hat{K}_n^*$ such that

$$P\left(K_n^* \leq K_{\alpha/2}^*\right) = P(K_n^* > K_{1-\alpha/2}^*\big) = \alpha/2 \tag{5.6}$$

we then have

$$P\left(K_{\alpha/2}^* \leq K_n^* \leq K_{1-\alpha/2}^*\right) = 1 - \alpha \tag{5.7}$$

which implies that

$$P\left(K_{\alpha/2}^* - \hat{K}_n \leq K_n^* - \hat{K}_n \leq K_{1-\alpha/2}^* - \hat{K}_n\right) = 1 - \alpha \tag{5.8}$$

Assuming that we can approximate the quantiles of $\hat{k}_n - k$ by the quantiles of $K_n^* - \hat{K}_n$, we have

$$P\left(K_{\alpha/2}^* - \hat{K}_n \leq \hat{K}_n - K \leq K_{1-\alpha/2}^* - \hat{K}_n\right) \approx 1 - \alpha \tag{5.9}$$

By simple calculation, we have

$$P\left(\hat{K}_n - (K_{1-\alpha/2}^* - \hat{K}_n) \leq K \leq \hat{K}_n - (K_{\alpha/2}^* - \hat{K}_n)\right) \approx 1 - \alpha \tag{5.10}$$

As noted in Efron and Tibshirani([15],pp54,pp162), transforming equation (5.14) can give better C.I. We therefore transform the random variable $\hat{K}_n$ using a symmetrical function say, $t()$. We denote this as:

$$\hat{\omega}_n = t(\hat{K}_n) \tag{5.11}$$

Then, from (5.14) we have

$$P\left(\hat{\omega}_n - (\omega_{1-\alpha/2}^* - \hat{\omega}_n) \leq \omega \leq \hat{\omega}_n - (\omega_{\alpha/2}^* - \hat{\omega}_n)\right) \approx 1 - \alpha \tag{5.12}$$

since due to symmetry$(\omega_{1-\alpha/2}^* - \hat{\omega}_n) = -(\omega_{\alpha/2}^* - \hat{\omega}_n)$, we can write equation (5.16) as:

$$P\left(\hat{\omega}_n - (\omega_{\alpha/2}^* - \hat{\omega}_n) \leq \omega \leq \hat{\omega}_n - (\omega_{1-\alpha/2}^* - \hat{\omega}_n)\right) \approx 1 - \alpha \tag{5.13}$$

This reduces to

$$P\left(\omega^*_{\alpha/2} - \leq \omega \leq \omega^*_{1-\alpha/2}\right) \approx 1-\alpha \tag{5.14}$$

Transforming this equation back to the original scale gives

$$P\left(K^*_{\alpha/2} \leq K \leq K^*_{1-\alpha/2}\right) \approx 1-\alpha \tag{5.15}$$

so that

$$P\left(K^*_{n,((S+1)\alpha/2)} \leq K \leq K^*_{n,((S+1)(1-\alpha/2))}\right) \approx 1-\alpha \tag{5.16}$$

Therefore, the $\alpha-$percentile bootstrap C.I is given by:

$$\left(K^*_{n,((S+1)\alpha/2)}, K^*_{n,((S+1)(1-\alpha/2))}\right) \tag{5.17}$$

5.4 Simulation Study

The model below was used

$$P(Y_i = 1|\mathbf{X}_i = \mathbf{x}) = \begin{cases} (1+exp(-(-1.5+2*x_{1i}+0.8*x_{2i})))^{-1}, & i \leq K \\ (1+exp(-(-1.5+2*x_{1i}+1.2*x_{2i})))^{-1}, & K < i \leq n \end{cases} \tag{5.18}$$

where the sample size was fixed at $n = 100$ and the change point at $K = 50$. The Bernoulli random variables were generated in line with equation equation (5.18) while $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ were generated as *uniform*[0,1]. $\hat{K}_n, \hat{\theta}_{0\hat{K}_n}$ and $\hat{\theta}_{1\hat{K}_n}$ were then estimated following the bootstrap procedure above. S=2000 bootstrap replications of $\hat{K}_n$ were then done in line with the bootstrap procedure above. The results are displayed in figure (5.2).

Discussion

The estimated change point from simulated data using equation (5.18) was found to be $\hat{K}_{100} = 50$ against the true one $K = 50$. Table(5.2) represents results from figure (5.3). Just like Profile log-likelihood ratio Confidence Intervals, Percentile Bootstrap Confidence intervals are not symmetrical.

Since the bootstrap percentile method uses the empirical distribution of the estimator of K, it has smaller coverage error than standard interval method which uses asymptotic approximations, see Pastor-Barriuso(2003) for more details.

However, the bootstrap percentile method has a computational obstacle especially for large sample sizes. This is because to accurately estimate the limits of bootstrap percentile intervals, the number of bootstrap samples are

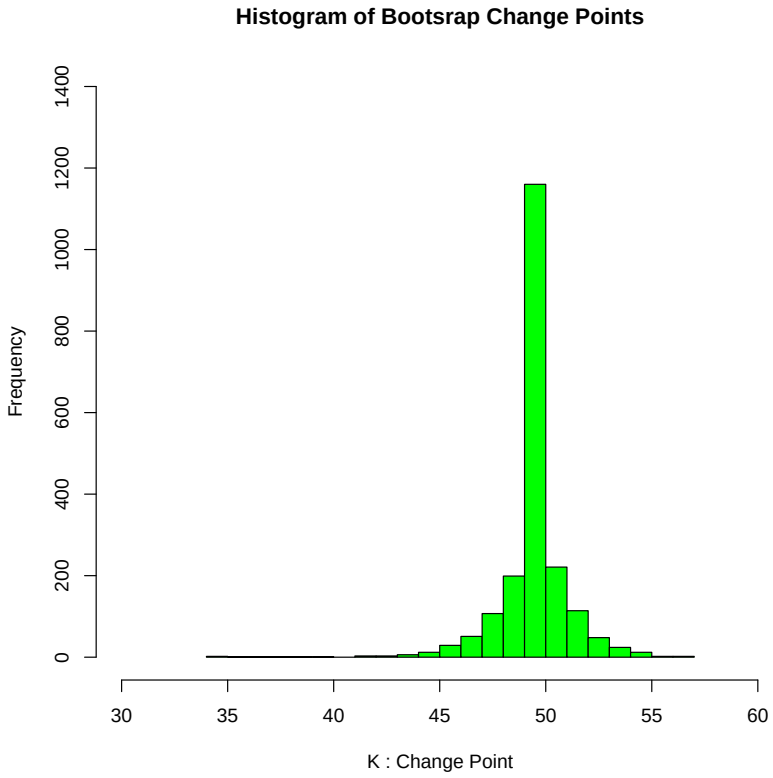

Figure 5.2: *A Histogram of S=2000 bootstrap replications of* $\hat{K}_n$

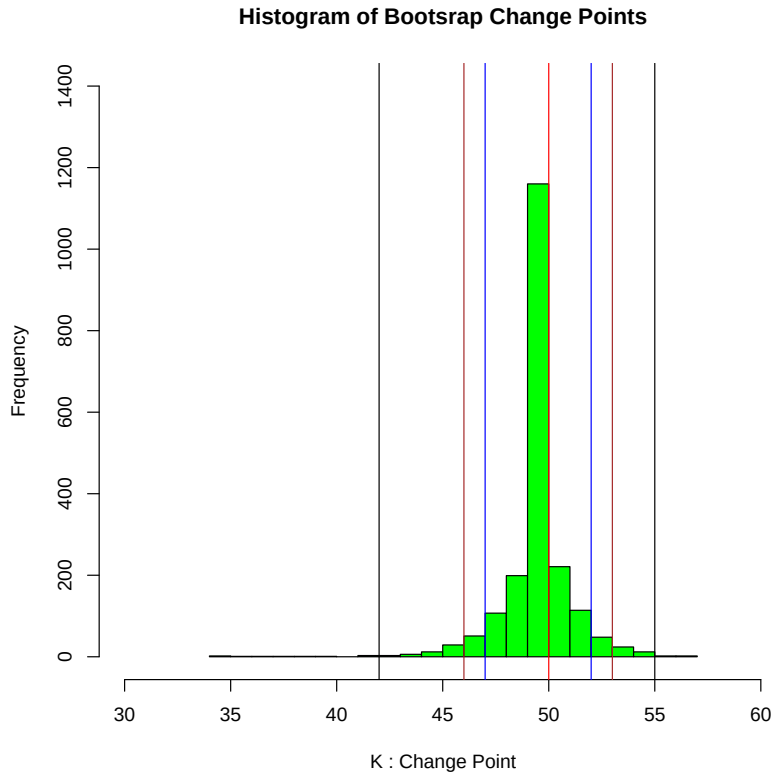

Figure 5.3: *A Histogram of S=2000 bootstrap replications of* $\hat{K}_n$*. The red vertical line represents* $\hat{K}_{100}$*. The two vertical blue lines mark the* 90% *confidence interval, the two vertical brown lines mark the* 95% *confidence interval while the two vertical black lines mark the* 99% *confidence interval.*

usually high (ideally S=1000). Added to this is the fact that the estimation of each of K_n^* is done iteratively which may take long computer time especially for large sample sizes.

In neural network set-up, these obstacles are overcome by optimizing the number of hidden neurons and using a faster Central Processing Unit (CPU).

5.4.1 Coverage Performance

In this section, we investigate whether the percentile method is better than the profile log-likelihood method by comparing their coverage performances.

In both profile log-likelihood and percentile intervals, a $1-2\alpha$ confidence interval $(\hat{k}_{Lower}, \hat{k}_{Upper})$ is expected to have probability α of miss-coverage of the true value K from above or below. That is,

$n = 100, \hat{K}_{100} = 50, S = 2000$	
Confidence Level	Confidence Interval
90%	47 - 52
95%	46 - 53
99%	42 - 55

Table 5.2: *Confidence Interval results for* $S = 2000$ *bootstrap replications of the change point* $\hat{K}_{100} = 50$ *from data of size* $n = 100$ *generated as in equation (5.18).*

Percentile Bootstrap Method		
Confidence Level	% Miss Left	% Miss Right
90%	13.3	4.2
95%	7.5	0.8

Table 5.3: *Results for 120 Percentile Bootstrap Confidence Interval realizations for the change point* K *from data of size* $n = 100$ *generated as in equation (5.4). For each realization,* $S = 1000$ *bootstrap replications of* $\hat{K}_n$ *were done.*

$$\text{Prob}(K < \hat{K}_{Lower}) = \alpha \text{ or } \text{Prob}(K > \hat{K}_{Upper}) = \alpha \tag{5.19}$$

As noted in Efron and Tibshirani [15], approximate confidence intervals can be graded on how accurately they match equation (5.19).

Figure (5.4) , figure (5.5) and table (5.3) represent coverage performance results under percentile bootstrap interval method.

Discussion

Table (5.3) shows the percentage of times the percentile bootstrap intervals missed the true value on the left and right hand sides in 120 realizations. This method slightly over-covers on the right and under-covers on the left.

However, the method clearly performs better compared to the profile log-likelihood which from table((5.1)) greatly over-covers on both the left and right hand sides.

5.5 Real Data Analysis

In section (4.3), the change points for all the four covariates were estimated. In this section, we use the percentile Bootstrap method to determine the 90% , 95% and 99% confidence intervals of the change point estimates in section (4.3).

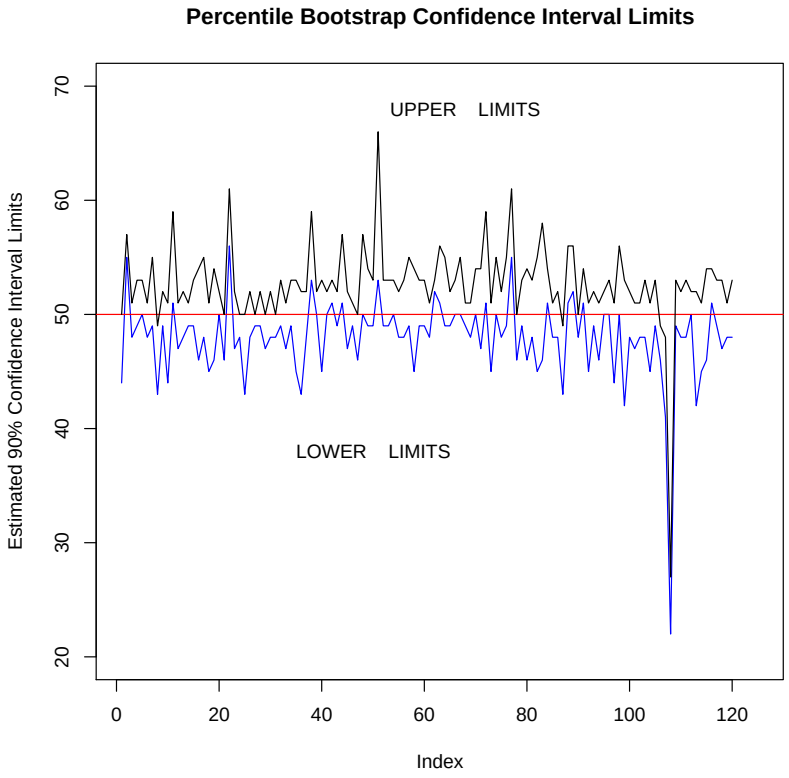

Figure 5.4: *A graph of 120* 90% *Confidence Interval realizations for the Change Point* $K = 50$ *from data of size* $n = 100$ *simulated as in equation (5.18). For each realization,* $S = 1000$ *bootstrap replications of* $\hat{K}_n$ *were done. The red horizontal line represents the true Change point,* $K = 50$*. The blue curve represents the lower* 90% *confidence Interval Limits while the black curve represents the Upper* 90% *confidence interval Limits.*

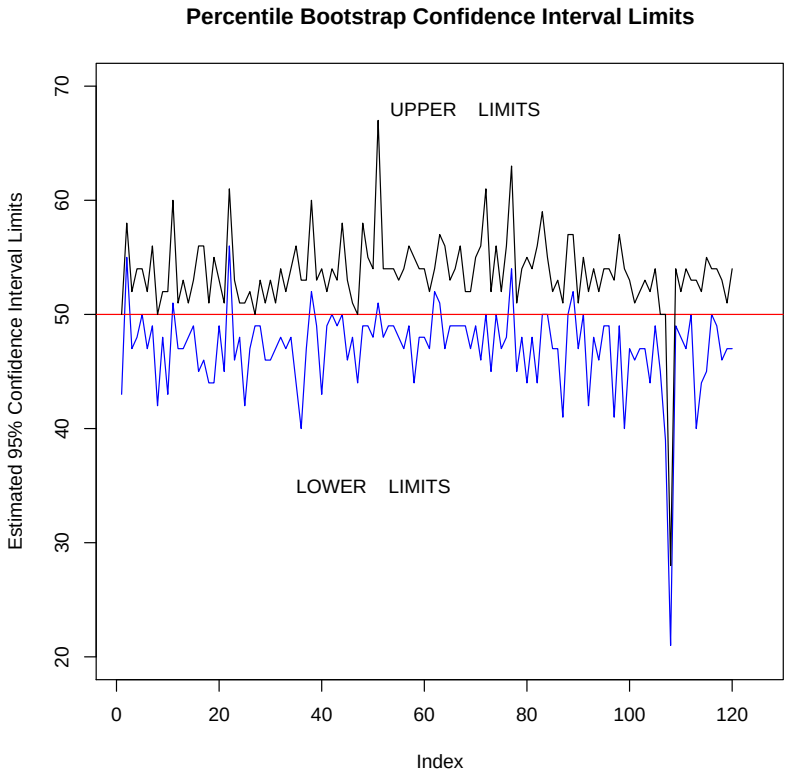

Figure 5.5: *A graph of 120* 95% *Confidence Intervalrealizations for the Change Point* $K = 50$ *from data of size* $n = 100$ *simulated as in equation (5.18). For each realization,* $S = 1000$ *bootstrap replications of* $\hat{K}_n$ *were done. The red horizontal line represents the true Change point,* $K = 50$*. The blue curve represents the lower* 95% *confidence Interval Limits while the black curve represents the Upper* 95% *confidence interval Limits.*

	Confidence		
Covariate	90%	95%	99%
Treatment, $\hat{K}_{194} = 58$	48 - 71	44 - 74	38 - 91
Age, $\hat{K}_{194} = 73$	66 - 92	63 - 95	58 - 102
Time, $\hat{K}_{194} = 31$	7 - 45	3 - 51	1 - 59
Status, $\hat{K}_{194} = 64$	56 - 71	54 - 72	50 - 78

Table 5.4: *Change Point Confidence Interval estimates for the cancer data described in section (3.12). For each covariate,* $S = 1000$ *bootstrap replications of* $\hat{K}_{194}$ *were done.*

The confidence interval estimation was carried out in line with section (4.3) and (5.3). The results are presented in figures (5.6), (5.7), (5.8), (5.9) and table(5.4). For each change point estimate, 1000 replications were done to determine the confidence intervals.

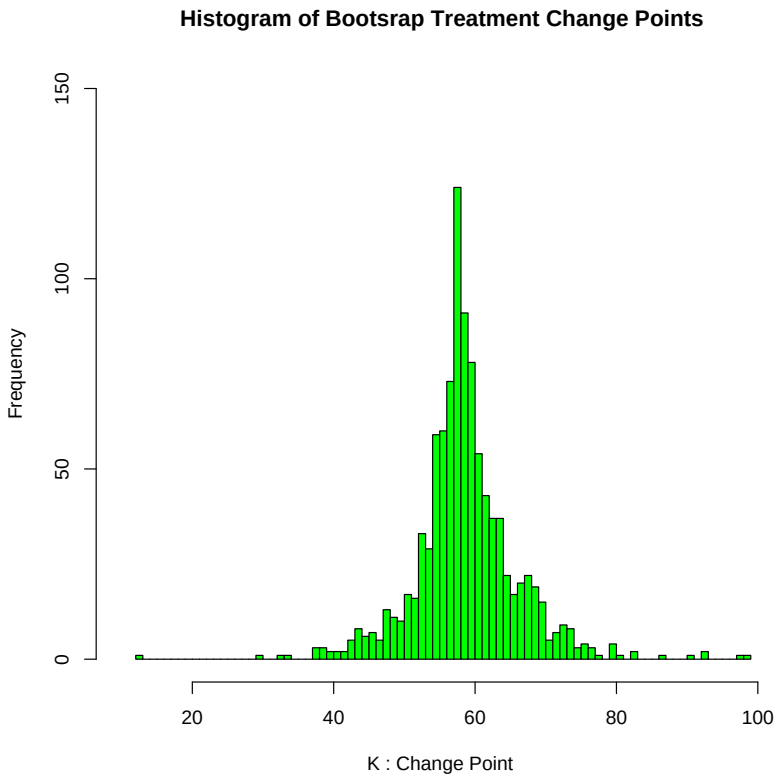

Figure 5.6: *A Histogram of S=1000 bootstrap replications of* $\hat{K}_{194} = 58$

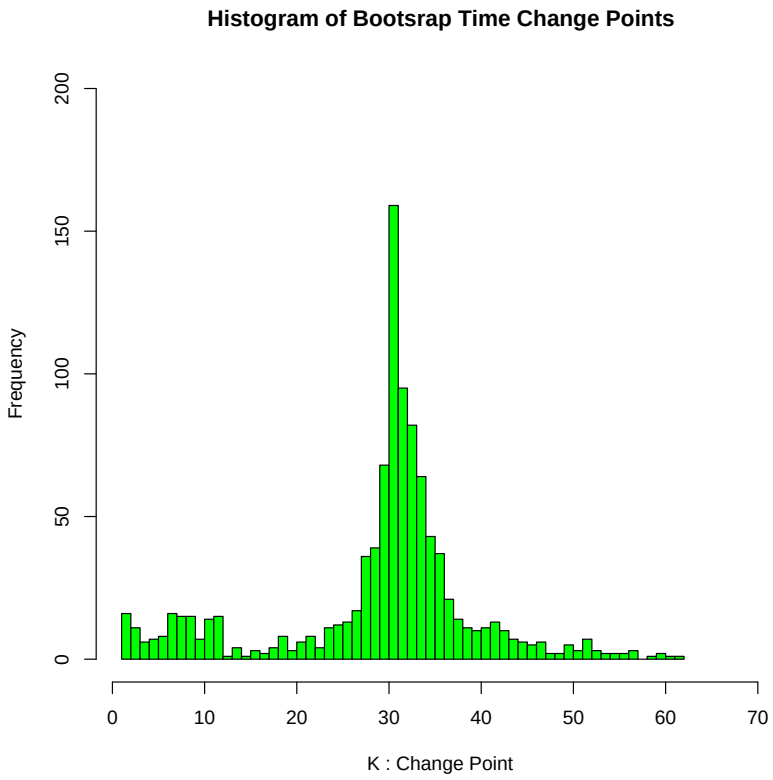

Figure 5.7: *A Histogram of S=1000 bootstrap replications of* $\hat{K}_{194} = 31$

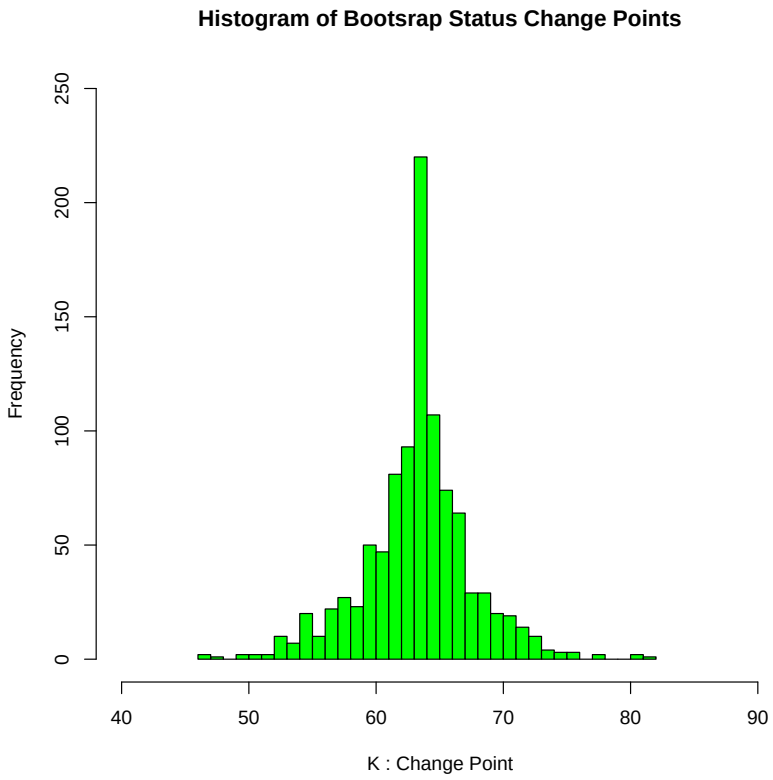

Figure 5.8: *A Histogram of S=1000 bootstrap replications of* $\hat{K}_{194} = 64$

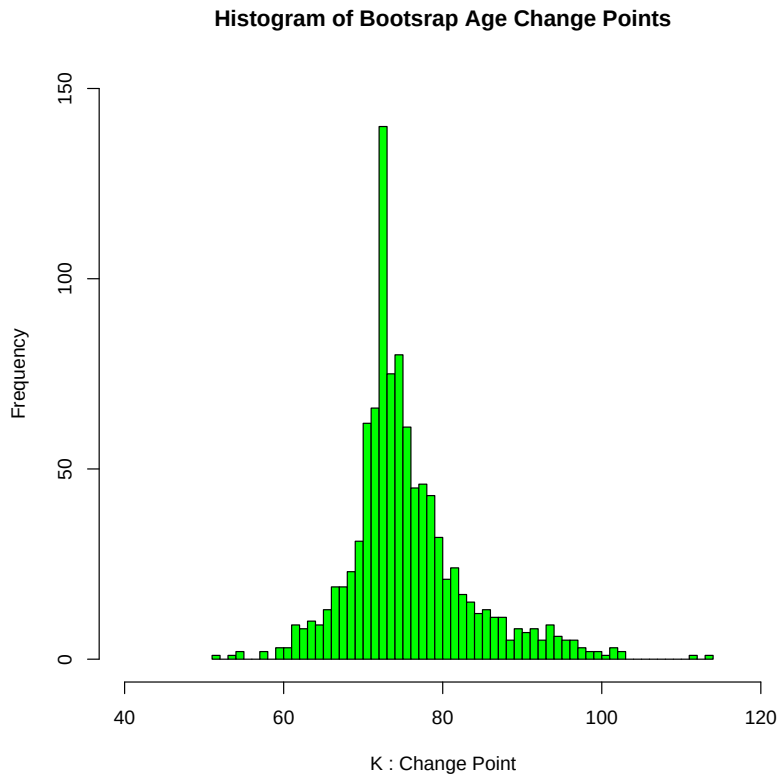

Figure 5.9: *A Histogram of S=1000 bootstrap replications of* $\hat{K}_{194} = 73$

Bibliography

[1] Andrews, D. W. K. Generic uniform convergence. *Econometric Theory 8*, 2 (1992), 241–257.

[2] Antoch, J., and Jarušková, D. On-line statistical process control. In *Multivariate total quality control*, Contrib. Statist. Physica, Heidelberg, 2002, pp. 87–124.

[3] Antoch, J., and Víšek, J. Á. Robust estimation in linear model and its computational aspects. In *Computational aspects of model choice (Prague, 1991)*, Contrib. Statist. Physica, Heidelberg, 1993, pp. 39–104.

[4] Aue, A., Horváth, L., Hušková, M., and Kokoszka, P. Change-point monitoring in linear models. *Econom. J. 9*, 3 (2006), 373–403.

[5] Berkes, I., Gombay, E., Horváth, L., and Kokoszka, P. Sequential change-point detection in GARCH(p, q) models. *Econometric Theory 20*, 6 (2004), 1140–1167.

[6] Braun, J. V., Braun, R. K., and Müller, H.-G. Multiple change-point fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika 87*, 2 (2000), 301–314.

[7] Broyden, C. G. A new method of solving nonlinear simultaneous equations. *Comput. J. 12* (1969/1970), 94–99.

[8] Clark, T. E., and McCracken, M. W. The power of tests of predictive ability in the presence of structural breaks. *J. Econometrics 124*, 1 (2005), 1–31.

[9] Cobb, G. W. The problem of the Nile: conditional solution to a changepoint problem. *Biometrika 65*, 2 (1978), 243–251.

[10] Cook, R. D., and Weisberg, S. Confidence curves in nonlinear regression. *J. Amer. Statist. Assoc. 85*, 410 (1990), 544–551.

[11] Csörgő, M., and Horváth, L. *Limit Theorems in Change-Point Analysis*. pub-WILEY, pub-WILEY:adr, 1997.

[12] Davis, R., and Hsing, T. Testing for a change in the parameter values and order of an autoregressive model. *Ann. Statist. 23* (1995), 282–304.

[13] Davison, A. C., and Hinkley, D. V. *Bootstrap methods and their application*, vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1997. With 1 IBM-PC floppy disk (3.5 inch; HD).

[14] Easterling, D., and Peterson, T. C. . A new method for detecting and adjusting for undocumented discontinuities in climatological time series. *Int. J. Climatol. 15* (1995), 369–377.

[15] Efron, B., and Tibshirani, R. J. *An introduction to the bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

[16] Elsner, J. B., Jagger, T., and F., N. X. Changes in the rates of north atlantic major hurricane activity during the 20th century. *Geophys. Res. Lett. 27* (2000), 1743–1746.

[17] Fan, J., Farmen, M., and Gijbels, I. Local maximum likelihood estimation and inference. *J. R. Stat. Soc. Ser. B Stat. Methodol. 60*, 3 (1998), 591–608.

[18] Fan, J., Farmen, M., and Gijbels, I. Local maximum likelihood estimation and inference. *J. R. Stat. Soc. Ser. B Stat. Methodol. 60*, 3 (1998), 591–608.

[19] Fearnhead, P., and Liu, Z. On-line inference for multiple change-point problems. *J. R. Stat. Soc. Ser. B Stat. Methodol. 69*, 4 (2007), 589–605.

[20] Feder, P. I. The log likelihood ratio in segmented regression. *Ann. Statist. 3* (1975), 84–97.

[21] Feder, P. I. On asymptotic distribution theory in segmented regression problems—identified case. *Ann. Statist. 3* (1975), 49–83.

[22] Fletcher, R. A class of methods for nonlinear programming with termination and convergence properties. In *Integer and nonlinear programming*. North-Holland, Amsterdam, 1970, pp. 157–175.

[23] Franke, J., Härdle, W., and Hafner, C. *Statistics of financial markets*. Universitext. Springer-Verlag, Berlin, 2004. An introduction.

[24] Franke, J., and Neumann, M. H. Bootstrapping neural networks. *Neural Computation 12* (2000), 1929–1949.

[25] Frölich, M. Non-parametric regression for binary dependent variables. *Econom. J. 9*, 3 (2006), 511–540.

[26] Gallant, A. R. Testing a nonlinear regression specification: A non-regular case. *J. Amer. Statist. Assoc. 72* (1977), 523–529.

[27] Goldfarb, D. A family of variable-metric methods derived by variational means. *Math. Comp. 24* (1970), 23–26.

[28] Gombay, E., and Horváth, L. An application of the maximum likelihood test to the change-point problem. *Stochastic Process. Appl. 50*, 1 (1994), 161–171.

[29] Gombay, E., and Horváth, L. Approximations for the time of change and the power function in change-point models. *J. Statist. Plann. Inference 52*, 1 (1996), 43–66.

[30] Gombay, E., and Horváth, L. On the rate of approximations for maximum likelihood tests in change-point models. *J. Multivariate Anal. 56*, 1 (1996), 120–152.

[31] Gordon, L., and Pollak, M. Average run length to false alarm for surveillance schemes designed with partially specified pre-change distribution. *Ann. Statist. 25*, 3 (1997), 1284–1310.

[32] Guan, Z. A semiparametric changepoint model. *Biometrika 91*, 4 (2004), 849–862.

[33] Hall, P. *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.

[34] Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. *Nonparametric and semiparametric models*. Springer Series in Statistics. Springer-Verlag, New York, 2004.

[35] Hinkley, D. V. Inference about the intersection in two-phase regression. *Biometrika 56* (1969), 495–504.

[36] Hinkley, D. V. Inference about the change-point in a sequence of random variables. *Biometrika 57* (1970), 1–17.

[37] Hinkley, D. V., and Hinkley, E. A. Inference about the change-point in a sequence of binomial variables. *Biometrika 57* (1970), 477–488.

[38] Horváth, L. The limit distributions of likelihood ratio and cumulative sum tests for a change in a binomial probability. *J. Multivariate Anal. 31*, 1 (1989), 148–159.

[39] Hsu, C.-C. Change point estimation in regressions with $I(d)$ variables. *Econom. Lett. 70*, 2 (2001), 147–155.

[40] Huber, P. J. *Robust statistics*. John Wiley & Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[41] Hwang, J. T. G., and Ding, A. A. Prediction intervals for artificial neural networks. *J. Amer. Statist. Assoc. 92*, 438 (1997), 748–757.

[42] James, B., James, K. L., and Siegmund, D. Tests for a change-point. *Biometrika 74*, 1 (1987), 71–83.

[43] Jandhyala, V. K., and Fotopoulos, S. B. Capturing the distributional behaviour of the maximum likelihood estimator of a changepoint. *Biometrika 86*, 1 (1999), 129–140.

[44] Jaruskova, D. Some problems with application of change point detection methods to environmental data. *Environmetrics 8* (1997), 469–483.

[45] Joanes, D. N. Reject inference applied to logistic regression for credit scoring. *J. Math. Appl. Bus. Ind. 5* (1993/4), 35–43.

[46] Kim, H.-J., and Siegmund, D. The likelihood ratio test for a change-point in simple linear regression. *Biometrika 76*, 3 (1989), 409–423.

[47] Loader, C. R. Change point estimation using nonparametric regression. *Ann. Statist. 24*, 4 (1996), 1667–1678.

[48] Looney, C. G. *Pattern recognition using neural networks: theory and algorithms for engineers and scientists*. Oxford University Press, Newyork, 1997.

[49] Lund, R., and J., R. Detection of undocumented change points: A revision of the two-phase regression model. *J. Climate 15* (2002), 2547–2554.

[50] MacNeil, and Mao. Change point analysis for mortality rate. *J. Appl. Statist. Sci. 1* (1995), 359–377.

[51] McNelis, P. D. *Neural networks in finance: Gaining predictive edge in the market*. Academic Press, Inc., Orlando, FL, USA, 2004.

[52] Mei, Y. Sequential change-point detection when unknown parameters are present in the pre-change distribution. *Ann. Statist. 34*, 1 (2006), 92–122.

[53] Page, E. S. A test for a change in a parameter occurring at an unknown point. *Biometrika 42* (1955), 523–527.

[54] Page, E. S. On problems in which a change in parameter occurs at an unknown point. *Biometrika 44* (1957), 248–252.

[55] Pan, J., and Chen, J. Application of modified information criterion to multiple change point problems. *J. Multivariate Anal. 97*, 10 (2006), 2221–2241.

[56] Pastor, R., and Guallar, E. Use of two-segmented logistic regression to estimate change-points in epidemiologic studies. *Am. J. Epidemiol. 148*, 7 (1998), 631–642.

[57] Pastor-Barriuso, R., Guallar, E., and Coresh, J. Transition models for change-point estimation in logistic regression. *Statist. Med 22* (2003), 1141–1162.

[58] Pettitt, A. N. A simple cumulative sum type statistic for the change-point problem with zero-one observations. *Biometrika 67*, 1 (1980), 79–84.

[59] Rukhin, A. L. Asymptotic behavior of estimators of the change-point in a binomial probability. *J. Appl. Statist. Sci. 2*, 1 (1995), 1–12.

[60] Seber, G. A. F., and Wild, C. J. *Nonlinear regression*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1989.

[61] Siegmund, D., and Venkatraman, E. S. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist. 23*, 1 (1995), 255–271.

[62] Solow, A. Testing for climate change: an application of the two-phase regression model. *J. Climate Appl. Meteor. 26* (1987), 1401–1405.

[63] Sussmann, H. J. Uniqueness of the weights for minimal feed-forward nets with a given input-output map. *Neural Networks 5* (1992), 589–593.

[64] Swanson, N. R., and White, H. A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks. *J. Bus. Econom. Statist. 13*, 3 (1995), 265–275.

[65] Taylor, W. A. Change-point analysis: A powerful new tool for detecting changes. Available at http://www.variation.com/cpa/tech/changepoint.html, 2000.

[66] Vexler, A., and Gurevich, G. Guaranteed local maximum likelihood detection of a change point in nonparametric logistic regression. *Comm. Statist. Theory Methods 35*, 4-6 (2006), 711–726.

[67] White, H. Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc. 76*, 374 (1981), 419–433.

[68] White, H. Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1 (1982), 1–25.

[69] White, H. "Some asymptotic results for learning in single hidden-layer feedforward network models" [J. Amer. Statist. Assoc. **84** (1989), no. 408, 1003–1013; MR1134490 (92e:62119)]. *J. Amer. Statist. Assoc. 87*, 420 (1992), 1252.

[70] Worsley, K. J. The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika 70*, 2 (1983), 455–464.

[71] Yao, Y. C., and Davis, R. A. The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variables. *Sankhyá Ser. A 48* (1984), 339–353.

[72] Yashchin, E. Change-point models in industrial applications. In *Proceedings of the Second World Congress of Nonlinear Analysts, Part 7 (Athens, 1996)* (1997), vol. 30, pp. 3997–4006.

[73] Zhan, M., Dean, C. B., Routledge, R., Gallaugher, P., Farrell, A. P., and Thorarensen, H. Inference on segmented polynomial models. *Biometrics 52* (1996), 321–327.