ESTIMATION OF CHANGE POINT IN BINOMIAL RANDOM VARIABLES USING NEURAL NETWORKS

MUNDIA SIMON MAINA

DOCTOR OF PHILOSOPHY
(STATISTICS)

JOMO KENYATTA UNIVERSITY OF
AGRICULTURE AND TECHNOLOGY

2014

Estimation of Change Point in Binomial Random Variables using Neural Networks

Mundia Simon Maina

A Thesis Submitted in Fulfilment for the Degree of Doctor of Philosophy in Statistics in the Jomo Kenyatta University of Agriculture and Technology

2014

DECLARATION

This thesis is my original work and has not been presented elsewhere for a degree award.

Signature: ..... Date: .....

This thesis has been submitted with our approval as University Supervisors.

Signature: ..... Date: .....

Dr. Gichuhi A. Waititu

JKUAT, KENYA

Signature: ..... Date: .....

Prof. John M. Kihoro

CUCK, Kenya

DEDICATION

This work is dedicated to my late father, Mundia Mbuthia and my late father-in-law Mwangi Gachigua.

ACKNOWLEDGMENTS

First is to thank the Almighty for giving me a chance to do this work.

I would like to show my deepest gratitude to my supervisors for their guidance and continued support. Their guidance and encouragement enabled to conduct this research.

Special thanks to the teaching staff of the Actuarial and Statistics department of Dedan Kimathi University of Technology for their support throughout the duration of my study.

I owe special thanks to my wife Nyambura, my son Mundia, my daughters Wakarindi and Muthoni who may have wondered why their father was on books most of the time. This thesis would not have been possible without your understanding and encouragement.

Special thanks to everyone who in one way or another contributed to the writing of this thesis. You are all wonderful people and God bless you.

Finally my sincere thanks to Dedan Kimathi University of Technology for their financial support throughout the study.

TABLE OF CONTENTS

DECLARATION . . . . . **ii**

DEDICATION . . . . . **iii**

ACKNOWLEDGMENTS . . . . . **iv**

TABLE OF CONTENTS . . . . . **iv**

LIST OF TABLES . . . . . **viii**

LIST OF FIGURES . . . . . **ix**

LIST OF ABBREVIATIONS . . . . . **xii**

ABSTRACT . . . . . **xiii**

CHAPTER ONE: INTRODUCTION **1**

1.1 Background Information 1

1.2 Literature Review 2

1.2.1 Ways of obtaining data 3

1.2.2 Categorization of data 3

1.2.3 Continuous change point problems 4

1.2.4 Discrete change point problems 6

1.3 Statement of the Problem 7

1.4 Significance of the Study 7

1.5 Objectives of the study 8

1.5.1 General Objectives 8

1.5.2 Specific Objectives 8

CHAPTER TWO: NEURAL NETWORKS AND LOGISTIC REGRESSION **9**

2.1 Introduction 9

2.2 Power polynomials 9

2.3 Orthogonal polynomials 10

2.4 Neural networks 11

2.4.1 The logistic squasher function 14

2.4.2 The hyperbolic tangent function 15

2.4.3 The Gaussian function 17

2.5 Training of the networks 18

2.5.1 The local gradient based search(gradient descent) method 20

2.5.2 Simulated Annealing Search 24

2.5.3 Evolutionary Stochastic Search 25

2.5.4 The stopping Rule 27

2.6 Models for Binomial Data 28

2.6.1 Linear Models 29

2.6.2 Fitting Binomial Data into Linear Model 29

2.7 Models for Binomial Response Data 30

2.7.1 The Logit Transformation 31

2.7.2 The Probit Transformation 31

2.7.3 The Complementary Log-log Transformation 31

2.8 The Logistic Regression Model 32

2.9 Change point detection in binomial variables 34

2.9.1 The likelihood Procedure 34

2.9.2 The Cumulative Sum (CUSUM) Procedure 36

2.9.3 Informational Procedure 37

CHAPTER THREE:CHANGE POINT DETECTION AND ESTIMATION 39

3.1 Introduction 39

3.2 Model Definition 40

3.3 Estimation of the Parameters 40

3.4 Testing for change points 41

3.5 Model Irreducibility 42

3.6 Model Identifiability 45

3.7 Consistency and Asymptotic Normality of Network Parameter Estimates 46

3.8 The Limit Distribution of the Change Point Statistic 57

3.9 Power of the test 74

3.10 Testing for change in misspecified model 75

3.10.1 The general testing for change points in a misspecified model 78

3.11 Change point estimation 85

3.12 Confidence Interval For The Change Point Estimate 87

3.12.1 Profile likelihood method 87

3.12.2 Percentile Bootstrap Confidence Interval 89

CHAPTER FOUR: RESULTS AND DISCUSSIONS 90

4.1 Introduction 90

4.2 Power of the test 90

4.2.1 Change of the power with change point location 91

4.2.2 Change of the power with sample size 94

4.2.3 Change of the power with size of the change 97

4.2.4 Application to real data 99

4.3 Change Point Estimation 101

4.3.1 Real Data Analysis 107
4.3.2 Confidence Interval of Change Point Estimates 109

CHAPTER FIVE: SUMMARY AND RECOMMENDATIONS 111

5.1 Summary of findings 111
5.2 Recommendations for Further Research 112

REFERENCES 112

LIST OF TABLES

Table 3.1: Critical values 73

Table 4.1: Power of the likelihood ratio test from a sample size b= 200 using critical values C1. 92

Table 4.2: Power of the likelihood ratio test from a sample of 200 using critical values C2. 92

Table 4.3: Power of the likelihood ratio test when the change point is at $\frac{b}{4}$ 94

Table 4.4: Power of the likelihood ratio test when the change point is at $\frac{b}{2}$ 95

Table 4.5: Power of the likelihood ratio test when the change point is at $\frac{3b}{4}$ 95

Table 4.6: Power of the likelihood ratio test for different sizes of change and change point locations k. 98

Table 4.7: Beetles Data 99

Table 4.8: Estimated probabilities of death 100

Table 4.9: Confidence Interval results for 1000 bootstrap samples . . 110

LIST OF FIGURES

Figure 2.1: Neural network with one hidden layer of H neurons and D+1 input nodes 13
Figure 2.2: The logistic curve 14
Figure 2.3: The tan-hyperbolic curve 16
Figure 2.4: The standard normal curve 17
Figure 2.5: The logit, probit and the complementary $\log - \log$ transformations of p 32

Figure 4.1: A plot of the power of the test against the location of change point. 93
Figure 4.2: A plot of the power of the test against the size of the sample at $\alpha = 0.01$. 96
Figure 4.3: A plot of the power of the test against the location of the change point at $\alpha = 0.01$ for the changes of size 1.2, 1.5 and 1.8 97
Figure 4.4: A plot of the estimated probabilities of death against dosage 101
Figure 4.5: hypothesis testing graph when the alternative is true . . 102
Figure 4.6: hypothesis testing graph when the alternative is false . . 103
Figure 4.7: loglikelihood graph 103
Figure 4.8: Histogram of likelihood estimates of change point 104
Figure 4.9: Histogram of likelihood estimates of change point when there is no change 105
Figure 4.10: histogram of the biases of the change point estimates . . 105
Figure 4.11: normal curve and histogram together 106
Figure 4.12: qqplot of the change-point estimates 107
Figure 4.13: hypothesis testing graph for the bliss data 108

Figure 4.14:loglikelihood graph for the bliss data 108

Figure 4.15:A histogram of 1000 bootstrap replicates of change point 109

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
ANN	Artificial Neural Network
BUGS	Bayesian inference Using Gibbs Sampling
CUSUM	Cumulative Sum
glm	Generalized Linear Model
m.l.e	maximum likelihood estimator
m.s.e.	mean squared error
SIC	Schartz Information Criterion

ABSTRACT

Statistically, change point is the location or the time point such that observations follow one distribution up to the point and then another afterwards. Change point problems are encountered in our daily life and in disciplines such as economics, finance, medicine, geology, literature among others. Change-point analysis is a powerful tool for determining whether a change has taken place. In this study, change point in binomial random variables whose mean is dependent on explanatory variables is investigated. It is assumed that there was only a single change point in the data. Artificial neural networks are used to estimate the conditional means. Compared with the generalized linear methods the artificial neural network gave better probability estimates. The consistency and the asymptotic distribution of the change point estimator is also investigated, and is found to be asymptotically normally distributed. The limiting distribution of the network based likelihood ratio statistic when change exists is derived and critical regions obtained. Simulated data is used to investigate the power of the test. The test is found to be more powerful when the change is near the center of the data than when it in the edges. The power of the test was found to be affected by the magnitude of the change. The higher the size of the change the higher the chance of detecting it. The power of the test is also found to increase as the size of the sample. In the analysis of real data the change point was found to correspond with the LD50.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Changes occur in everyday life and people need to be aware of them so as to avoid unnecessary losses and to harness transactions. The following examples give a better insight into changes that occur in day to day experiences.

Stock prices fluctuate daily. Though according to economists these changes are normal there are some shifts that are abnormal. Hence a question arises. For instance, "did the post elections violence of 2007 cause a statistically significant change in the stock prices in Kenya?"

In quality control, the quality of a production line is expected to be stable. However due to some reasons the process may fail, and one might be interested in determining the point at which this happened.

In 2003, the then transport minister in Kenya introduced rules in the *Matatu*(passengers service vehicles) transport sector which were flouted a few years later. One would be interested in the investigation of whether statistically these rules had any significant change in the traffic accidents rate.

In mining, analysis is done on the ore samples obtained from different sites. If there is a significant change in the analysed results of the of ore, then the geologist might be interested in the site in which the change took place.

The statistical problem comes in as one requires to determine whether there is a significant change. If there is a significant change then one needs to obtain the point at which the change occurs and estimate the parameters at this point of change. This change may be unique, that is it occurs at only one point or it occurs at several points. Hence the change point analysis problem is twofold.

The first problem being whether significant change exists. This is a hypothesis testing problem. Change occurs when observations follow one distribution up a certain point and then another after that.

Let $X_1, X_2, X_3, \ldots, X_n$ be independent observations from a density $f(x;\theta)$. These observations follow the density $f(x;\theta)$, where θ is the parameter under consideration under normal conditions but when these normal conditions change at some point the observation follow another density $f(x;\theta')$, $\theta \neq \theta'$. To set up a change point problem the acceptance region, where θ resides under normal conditions is first defined. Then the change point analysis problem is equivalent to testing the hypothesis.

$$H_o \quad : \quad \theta_1 = \theta_2 = \ldots = \theta_n = \theta_0$$

Against

$$H_a \quad : \quad \theta_1 = \theta_2 = \ldots = \theta_k \neq \theta_{k+1} = \ldots = \theta_n = \theta' \tag{1.1}$$

where $1 < k < n$ and, k is the unknown change position which has to be estimated if H_0 is not true. If the null hypothesis is not rejected then a change does not exist and therefore the problem stop here. Otherwise one proceeds to the second problem, the estimation of the change point k. Though k may take more than one position in this work situations where only a single change point exists is considered.

1.2 Literature Review

A survey of the studies done on change point analysis problems shows that most of earlier work is concentrated on single change point in the random sequence. Vostrikova (1978) proposed the binary segmentation method and proved its consistency for testing multiple change points. This procedure has three major steps.

1. Test for a single change point. If there is no change then stop. If a change point exists, then obtain the change point location k_1.

2. Test for change in the two subsequences before and after k_1 separately as in step 1.

3. Repeat the process until no further subsequences have change points

The collection of change point locations is say $\{k_1, k_2 \ldots, k_q\}$ and there are q change points. His method has the advantage of detecting the number of change points and their positions simultaneously hence saving a lot of computation time.

1.2.1 Ways of obtaining data

Depending on the way the data is obtained, a change point analysis problem may be classified as either off-line or on-line.

Off-line problems deal with fixed data samples which are first observed and the change detection is done later . Page (1954) introduced this type of change point problem in which a single change point was assumed. This study considers an off-line change point analysis problem with a single change point.

A change point analysis problem is said to be on-line if the independent observations considered initially have the same distribution and the process is said to be in control but at an unknown point the distribution of the observations changes. Then the process is said to be out of control. The setback of on-line change point analysis problems is that the entire data is not taken into account at once. Applications are mainly in quality control in production lines.

1.2.2 Categorization of data

The data obtained may either be discrete or continuous. This makes it possible to categorize change point analysis problems as either discrete or continuous. In

this work discrete data is considered.

The following is a review of work done by various authors on each of these types of change point problem.

1.2.3 Continuous change point problems

The first study on change point was conducted by Page (1954, 1955, 1957) in which he considered continuous inspection schemes and tested the change in a parameter running at an unknown point. Girshick and Rubin (1952) considered a Baye's approach to the quality control method. The two were considering industrial quality control, but it was Kolmogrov et al. (1988) who precisely formulated in a mathematical way of the change point problem.

The following is a review of the work done on specific probabilistic models by various authors. In a one dimensional Gaussian model with a known variance, the problem would be to determine whether the mean of the distribution changes at some point. This problem was first studied by Page (1955, 1957). Chernoff and Zacks (1964), Bhattacharya and Johnson (1973), Gardener (1969), Sen and Srivastava (1975a,b), Gupta and Chen (1996) and Chen and Gupta (1997) have made contributions to the study of this problem. Their interest was to test the hypothesis of the stability of the mean under the assumption that the variance is not changing. The testing procedure depends on whether the nuisance parameter, the variance of the distribution is known or unknown.

In Gaussian models where the variance is not stable inference about changes in variance while the mean remains constant has been studied by Wichern et al. (1976), Hsu (1977), Incl*á*n (1993) and Chen and Gupta (1999). Brown et al. (1975) also studied the situation where both the mean and the variance of a univariate normal distribution change.

In a multi-dimensional Gaussian model, Sen and Srivastava (1973) studied the

problem of a single mean vector change for a sequence of independent normal variables using a Bayesian test statistic. Srivastava and Worsley (1986) used the likelihood ratio test to detect the change in the mean vectors. Zhao et al. (1986a,b) studied the problem of detecting the number of signals in the presence of white noise when the noise covariance matrix is arbitrary. Krishnaiah et al. (1990) used the likelihood method to estimate the change point. James et al. (1992) obtained the asymptotic approximation for the likelihood test and the confidence region for the change in a multivariate normal.

In regression models, many authors have studied the change point problem. Quandt (1958, 1960) derived the likelihood ratio based test for testing and estimating linear models obeying separate regimes. Ferreira (1980) studied a switching regression model from the Bayesian point of view, assuming known numbers of regimes. Brown et al. (1975) brought in the method of recursive residuals to test change points in multiple regression models. Kim (1994) considered a test for change point in a linear regression using the likelihood ratio statistic and studied its asymptotic behavior. Chim Choy and Broemeling (1980) used the Bayesian approach to study a switching linear model. Hobert (1982) also used the Bayesian approach to study simple linear model and multiple regressions. Other continuous models that have been studied include the exponential models where Kander and Zacks (1966) posted a change point problem for the model. Hsu (1979) adopted their results and assumptions and studied the change point problem in a gamma model. Worsley (1986) used the likelihood ratio test to obtain the change in a sequence of independent exponentially distributed random variables. Haccou et al. (1988) used the likelihood ratio test and obtained the asymptotic null distribution of the test statistic while later Haccou and Meelis (1988) gave a procedure for obtaining the number of change points in a sequence of independent exponentially distributed random variables based on partitioning

of the likelihood according to the hierarchy of the sub-hypothesis. Gupta and Ramanayake (1998) studied the epidemic change using the likelihood ratio test. Chen and Gupta (2000) in their monograph have used the informational approach to test for change in exponential random variables.

1.2.4 Discrete change point problems

More work has been done in continuous models than in discrete models. Among the few authors who have contributed in the study of change point problems in discrete models include Hinkley and Hinkley (1970) who studied the change point for a binomial model using the maximum likelihood ratio test. However the mean of the distribution was assumed to be unconditional, that is, the mean was not dependent on some explanatory variables. We wish to study situations where the mean is dependent on the explanatory variables. Smith (1975) considered the same model from a Bayesian approach while Pettitt (1954)used the cumulative sum approach for the same model. Worsley (1983) studied the power of the likelihood ratio and cumulative sum tests for the binomial model. Fu and Curnow (1990) derived the null and non-null distribution of the log likelihood ratio statistic for locating the change point in the binomial model. Chen and Gupta (2000) in their monograph have used both the likelihood ratio and the informational approach to detect the change point in the binomial distribution though again the mean was assumed to be unconditional. They also analyzed data in Hanify et al. (1981) using the informational approach.

The most commonly used tests for testing for change are the maximum likelihood ratio test, the Bayesian test, the cumulative sum test and the informational approach test. The last method introduced by Akaike (1974) is a useful tool in model selection. A model is considered appropriate if it minimizes AIC. However this estimator is not asymptotically consistent. Schwarz (1978) proposed an

asymptotically consistent estimator. Most of the work on change point analysis involves prior assumed distributions.

Waititu (2008) considered bernoulli random variables where the probability of success was depend on a set of explanatory variables and used the ANN to estimate the conditional means of the variables. This work focuses on a non-parametric method of estimating the conditional means where ANN are used to estimate these unknown conditional means of binomial random variables which are dependent on a set of given explanatory variables.

1.3 Statement of the Problem

In this study a sequence of binomial random variables is considered. The probability of success is known to depend on some explanatory variables. To estimate these probabilities a non-parametric method(ANN)is used and the test for change conducted in the sequence. If change exists, the point at which it occurs is determined. The properties of the neural network estimator is investigated and to obtain the critical regions the distribution of the test statistic under the hypothesis of no change has to be determined.

1.4 Significance of the Study

The applications to this include credit scoring in financial institutions, dose-response in biometry and epidemiology. Of interest will be the probabilities of success which will depend on various explanatory variables.

In a random process where change exists, the change may be gradual or abrupt. In abrupt change there is a sudden break in the model parameters. Gradual changes occur when parameters change slowly in the process. These gradual changes have applications in engineering and ecology. This study considers abrupt

change cases with a single change point. To test for change critical regions have to be defined by obtaining the distribution of the test statistic under the null hypothesis. The critical values obtained can be used to construct confidence interval for the change estimates.

1.5 Objectives of the study

1.5.1 General Objectives

The general objective of this study is to detect a change and estimate the position at which the change occurs in a sequence of random binomial observations.

1.5.2 Specific Objectives

Specifically our objectives will be to:-

1. Estimate the conditional binomial probabilities using ANN.
2. Derive the asymptotic distribution of change point likelihood ratio statistic under the hypothesis of no change.
3. Determine the power of the likelihood ratio test for binomial random variables with a change.
4. Derive the asymptotic distribution of the ANN estimator and check its consistency.
5. Validate the model using simulated data.
6. Test and estimate change in real data.

CHAPTER TWO

NEURAL NETWORKS AND LOGISTIC REGRESSION

2.1 Introduction

In this section the theory behind the use of neural networks to approximate parameters of a function. Also modeling of binomial data using logistic regression is looked at. Other parameters approximation methods of available in the literature are:-

2.2 Power polynomials

This is a commonly used method. From the Weierstrass theorem, a polynomial expansion around a set of inputs x with a progressively large power is capable of approximating to a given degree of precision any unknown but continuous function, see Miller et al. (1990). For instance a second-degree polynomial approximation in two variables $[x_1, x_2]$ and it is known that $y = f(x)$ then the approximation formulae becomes

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 \tag{2.1}$$

The major drawback of this method is that it cannot be used for discrete functions. Also as the degree of the polynomial increases the number of parameters to be estimated also increases. This decreases the degrees of freedom for the underlying statistical estimates. This is called the *curse* of dimensionality.

2.3 Orthogonal polynomials

This method is due to Judd (1998) and is based on the sine, the cosine, or the alternative exponential transformations of the variable. The method has been proved to be more efficient than the power polynomials discussed earlier. Before using this method the variables $[y, x]$ have to be transformed into the interval (1,1). The variable x is transformed as

$$x' = \frac{2x - \max(x) - \min(x)}{\max(x) - \min(x)} \tag{2.2}$$

The polynomial approximations are can be represented in recursive manner. Four recursive representation are available in the literature. They are:-

(i) The Chebeychev's polynomial expansion.

The recursive formula is

$$\begin{aligned} R_0(x') &= 1 \\ R_1(x') &= x' \\ R_{i+1}(x') &= 2xR_i(x') - R_{i-1}(x') \end{aligned} \tag{2.3}$$

(ii) The Hermite's polynomial expansion.

The recursive formula is

$$\begin{aligned} R_0(x') &= 1 \\ R_1(x') &= 2x' \\ R_{i+1}(x') &= 2x'R_i(x') - 2R_{i-1}(x') \end{aligned} \tag{2.4}$$

(iii) The Legendre's polynomial expansion.

The recursive formula is

$$\begin{aligned} R_0(x') &= 1 \\ R_1(x') &= 1 - x' \\ R_{i+1}(x') &= \frac{2i+1}{i+1}R_i(x') - \frac{i}{i+1}R_{i-1}(x') \end{aligned} \tag{2.5}$$

(iv) The Laguerre's polynomial expansion.

The recursive formula is

$$\begin{aligned} R_0(x') &= 1 \\ R_1(x') &= 1 - x' \\ R_{i+1}(x') &= \frac{2i+1-x'}{i+1}R_i(x') - \frac{i}{i+1}R_{i-1}(x') \end{aligned} \tag{2.6}$$

Once the polynomial expansion for a given variable x' is obtained, then a linear regression is used to approximate y'. For two variables x_1, x_2 with a second-degree expansion is

$$y' = \sum_{i=1}^{2}\sum_{j=1}^{2} \quad \beta_{ij}R_i(x_1)R_j(x_2) \tag{2.7}$$

where R_i and R_j are the recursive representation of the polynomial expansion. To transform the variable y' back to the interval $(\min(y), \max(y))$ the following expression is used.

$$y = \frac{(y'+1)[\max(y) - \min(y)]}{2} + \min(y) \tag{2.8}$$

For more literature on this method see McNelis (2005a).

2.4 Neural networks

A neural network represents the way in which the human brain processes input sensory data received as input neurons into recognition as output neurons. A neural network is used in forecasting a given target (output) from the information on a set of observed input variable. It uses one or more of the hidden layers in which the input is transformed by a special function called the *activation function*. In this study a feed-forward networks with $D+1$ input nodes and a single layer of H hidden nodes is considered. The input nodes and the hidden layer nodes are connected by the weights $w_{h,d}$, $h \in \{1, 2, \ldots, H\}$ and $d \in \{1, 2, \ldots, D\}$. For

an input vector $\mathbf{x}' = (x_1, \ldots, x_D)$ the following equation describes this input to the h^{th} hidden node.

$$n_h(x;\theta) = w_{h,0} + \sum_{d=1}^{D} w_{h,d} x_d \tag{2.9}$$

The output of the h^{th} hidden node is

$$\psi(n_h(x;\theta)) \tag{2.10}$$

where $\psi(.)$ is the activation function. This forms an input to the output node of the form

$$\zeta(x;\theta) = \alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(n_h(x;\theta)) \tag{2.11}$$

The final output of the network is

$$\varphi(x;\theta) = \alpha_0 + \sum_{h=1}^{H} \alpha_h \psi(n_h(x;\theta)) \tag{2.12}$$

where

$$\theta = (w_{h0}, w_{h1}, \ldots, w_{HD}, \alpha_0, \alpha_1, \ldots, \alpha_H) \tag{2.13}$$

denote the set of the parameters of the network.

A neural network is diagrammatically represented in Figure (2.1). Thus one may think of the network as a mathematical model that consists of

1. The *synapses* or connecting links that provide weights, $w_{h,d}$, to the input values, $x_{i,d}$ for $d = 1, \ldots, D$.

2. An adder that sums the weighted input values to compute an input to the h^{th} node of the hidden layer, $n_h = w_{h,0} + \sum_{d=1}^{D} w_{h,d} x_d$, where $w_{h,0}$ is called the bias (not to be confused with statistical bias in prediction or estimation) is a numerical value associated with the neuron. It is convenient to take the bias as the weight for an input x_0 whose value is always equal to one, so that $n_h = \sum_{d=0}^{D} w_{h,d} x_d$.

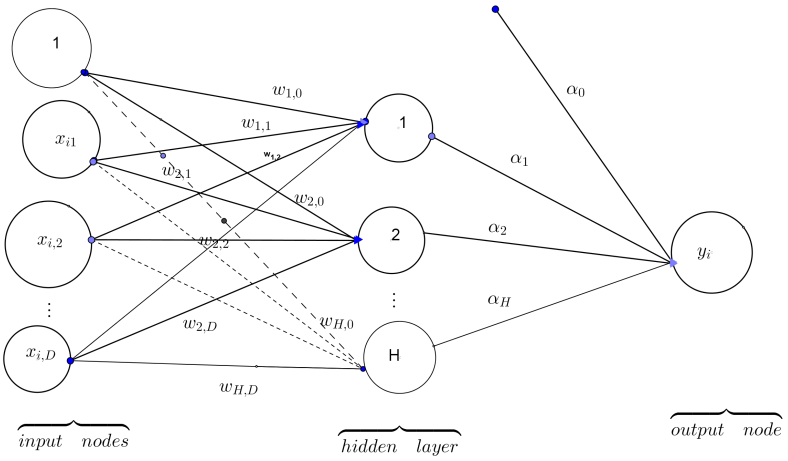

Figure 2.1: Neural network with one hidden layer of H neurons and D+1 input nodes

3. An activation function $\psi(.)$ also called a squashing function that maps n_h to $\psi(n_h)$ the output value of the neuron. This function is monotone. The function transforms its input into the required form. It performs a mathematical operation on the signal output. For instance in this work the interest is to obtain the conditional probabilities and thus one would require a function that has output in the range[0,1].

The number of parameters in a network depends on H, the number of hidden nodes. The total number of parameters in the network in Figure 2.1 is given by $H(D+2)+1$.

Thus the neuron processes the input data by first forming a linear combination of the inputs with their weights. This forms the input to the hidden layer. This in is turn squashed by the activation function to form the output. It represents a very efficient way to model non-linear statistical processes.

The following are some forms of squasher functions commonly used.

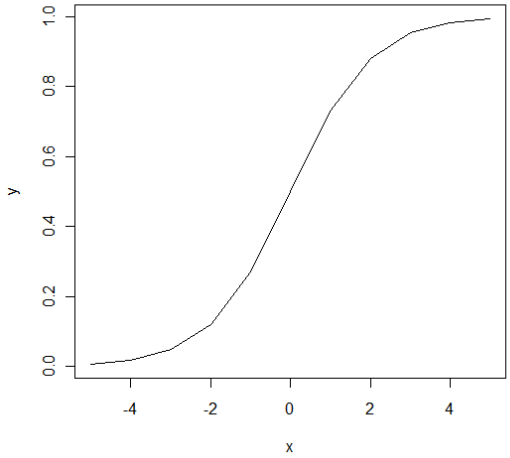

Figure 2.2: The logistic curve

2.4.1 The logistic squasher function

This function is of the form

$$\psi(x) = \frac{1}{1+e^{-x}} \tag{2.14}$$

Note that

$$\begin{aligned} \psi(x) &\rightarrow 0 \quad \text{as} \quad x \rightarrow -\infty \\ \psi(x) &\rightarrow 1 \quad \text{as} \quad x \rightarrow \infty \\ \psi(x) + \psi(-x) &= 1 \end{aligned} \tag{2.15}$$

implying that the function is symmetric sigmoidal an important property of activation functions used for predictive purposes. The function is also differentiable(smooth), an important property as the *learning* of the neural network depends on the gradient of the *error* which is in terms of the weights and the activation function. The activation function is also referred to as the unipolar function. The logistic curve is presented in Figure (2.2). When used as an acti-

vation function in the network defined in equation (2.11) one obtains

$$\psi(\zeta(x;\theta)) = \frac{1}{1+e^{-\zeta(x;\theta)}} = \varphi(x;\theta) \tag{2.16}$$

One advantage of this function is that the output is in the range [0,1] making it appropriate for the estimation of probabilities. Another importance of this function arises from its threshold behavior, which characterizes many types of responses, for example economic responses to changes in fundamental variables. It also reflects the learning behavior. Kauna and Halbert (1994) describes this feature as the tendency of certain types of neurons to be quiescent of modest levels of input activity and only become active after the input activity passes a certain threshold while beyond this, increases in the input activity have little effect. This activation function in our work to estimate the binomial probabilities. Figure 2.2 shows the shape of the logistic function.

2.4.2 The hyperbolic tangent function

Also known as the *tansig* or tanh, it squashes the linear combinations of the inputs to outputs in the range [-1,1]. Its functional form is

$$\psi(x) = \tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.17}$$

When used as an activation function in the network defined in equation(2.11), one obtains

$$\tanh(\zeta(x;\theta)) = \varphi(x;\theta) \tag{2.18}$$

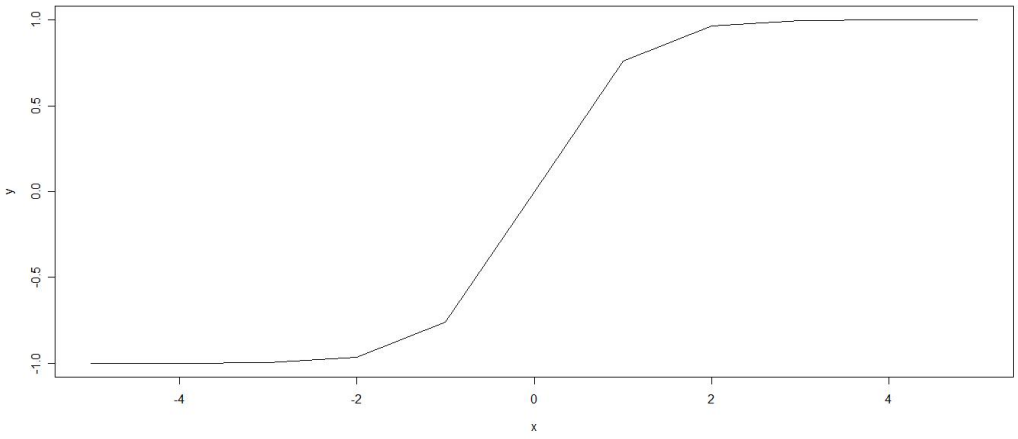

Figure 2.3: The tan-hyperbolic curve

Equation (2.17) may be expressed as

$$\begin{aligned} \tanh x &= \frac{\sinh x}{\cosh x} \\ &= \frac{(e^x - e^{-x})/2}{(e^x + e^{-x})/2} \\ &= \frac{e^x(1 - e^{-2x})}{e^x(1 + e^{-2x})} \\ &= 2\{\frac{1}{1 + e^{-2x}}\} - 1 \end{aligned} \tag{2.19}$$

Note that

$$\begin{aligned} \psi(x) &\rightarrow -1 \quad \text{as} \quad x \rightarrow -\infty \\ \psi(x) &\rightarrow 1 \quad \text{as} \quad x \rightarrow \infty \\ \psi(x) + \psi(-x) &= 0 \end{aligned} \tag{2.20}$$

Figure 2.3 shows the shape of the tan-hyperbolic function.

This activation function is also refereed to as the bipolar and has the same properties as discussed earlier for the unipolar except that its output range is [-1,1]. Hence if used to estimate probabilities it may give inappropriate estimates.

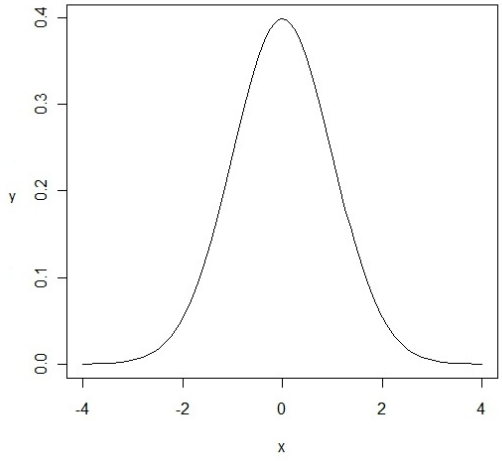

Figure 2.4: The standard normal curve

2.4.3 The Gaussian function

The function is also referred to as the standard normal. Though its slope is much steeper than the logistic function in the intervals [-2,0] and [0,2] but it has little or no response below -2 and above 2. Thus it has a narrow range of x. When used as an activation function in the network defined in equation (2.11), the following is obtain

$$\phi(n_h(x;\theta)) = \int_{-\infty}^{\zeta(x;\theta)} \sqrt{\frac{1}{2\pi}} \; e^{-\frac{\zeta(x;\theta)^2}{2}} dx \tag{2.21}$$

Figure 2.4 shows the shape of the standard normal function

One disadvantage of using this function as an activation is that its output values are in the range [0,0.4]. Thus when using it to estimate probabilities one has to be sure that the required values do not exceed 0.4.

In this study the aim is to estimate the conditional probabilities and thus the logistic function is the most suitable for this purpose.

2.5 Training of the networks

To obtain the required parameter estimates from the network its weights have to be adjusted until the desired goal is achieved. This is called training of the network. The network is said to be learning. Two main methods of training the network are available in the literature; the supervised and non-supervised training.

In unsupervised learning, the weights and biases are modified in response to network input only. The target outputs are not available. At first this seems impractical. How can one train a network if it is not known what it is supposed to do? Most of the algorithms for this type of learning perform some kind of a clustering operation. They learn to categorize the input patterns into a finite number of classes. This is useful in such applications such as vector quantization. Since the desired response is not known the explicit error information cannot be used to improve network's behavior. Thus learning must somehow be accomplished based on observations of responses to inputs that one has have marginal or no knowledge about.

In this mode of learning, a network must discover for itself any possibly existing patterns, regularities, separating properties and any other structural features of the input data. It is while discovering these structural features that the network undergoes change in its parameters. The summary of the rules for this mode of training as follows:-

1. A sample of input vectors. The expected output is not presented to the network.

2. The system learns on its own by discovering and adapting on its own to the structural features of the input data.

3. A stopping rule.

The supervised training is also referred to as the *error* based training since it is based on the comparison between the network's computed output and the expected output. The *error* generated is used to change the network's parameters that result to improved performance.

The training starts with the selection of initial guess or conditions. The aim is to obtain the a set of parameters that minimize the differences between the model predictions $\hat{y}$ and the actual values y. This guess may be anywhere in the parameter space. As the network is trained one may be stuck at a point where the derivative of the curve is zero. This may be a global minimum, a local minimum or an inflexion. Too large an adjustment may bring one near the global minimum or in an inflexion (saddle) point. But a small adjustment may keep one trapped in an inflexion during the training.

Clear-cut solutions for escaping from this kind of a situation are not available but strategies are there in the literature on how to re-estimate the parameters so as to obtain the weights that minimize the loss(error) function given by:-

$$\begin{aligned} \sum_{i=1}^{b}(\hat{y}_i - y_i)^2 &= \sum_{i=1}^{b}(\varphi(x;\theta) - y_i)^2 \\ &= \eta(\theta) \end{aligned} \tag{2.22}$$

where $\varphi(x;\theta)$ is as defined in equation (2.12). The loss function is minimized this with respect to θ, that is one obtains $\min \eta(\theta)$, which is a non-linear function of θ.

Starting with an initial guess θ_0, one trains the network until the best possible solution is obtained within a reasonable amount of training.

The following is a summary of the steps to be followed in the supervised method.

1. A sample of input vectors and an associated output vectors.

2. The selection of initial weights set.

3. A repetitive method that updates the current weights so that the input-output map is optimized.

4. A stopping rule.

In this study the supervised learning method is used since the form the expected outcome is known. Three methods of obtaining $\min \eta(\theta)$ are discussed below.

2.5.1 The local gradient based search(gradient descent) method

This method is based on the minimization of the *errors* which are defined in terms of weights and the activation function of the network. Thus the activation function must be differentiable as the updates of the weights depend on the gradient of the error. The aim will be to minimize the nonlinear error function. After the random selection of θ_0, iteratations on θ are carried until the loss function is minimized by using the first and second order derivatives of the error function with respect to the parameters. This searches for the optimum in the neighborhood of the initial guess. The usual way to iterate is through the quasi- Newton algorithm. To obtain $\eta(\theta_1)$, a second order Taylor expansion is used to give

$$\eta(\theta_1) = \eta(\theta_0) + \nabla_0(\theta_1 - \theta_0) + 0.5(\theta_1 - \theta_0)'\mathbf{M}_0(\theta_1 - \theta_0) \tag{2.23}$$

where ∇_0 is the gradient of the error function with respect to the parameters and $\mathbf{M}_0$ is the Hessian of the error function. If $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,H})$ is the set of

initial guess parameters then

$$\nabla_{\mathbf{0}} = \begin{bmatrix} \frac{\eta(\theta_{0,1}+v_1,\ldots,\theta_{0,H})-(\eta(\theta_{0,1},\ldots,\theta_{0,H}))}{v_1} \\ \vdots \\ \frac{\eta(\theta_{0,1},\ldots,\theta_{0,i}+v_i,\ldots,\theta_{0,H})-(\eta(\theta_{0,1},\ldots,\theta_{0,H}))}{v_i} \\ \vdots \\ \frac{\eta(\theta_{0,1},\ldots,\theta_{0,H}+v_H)-(\eta(\theta_{0,1},\ldots,\theta_{0,H}))}{v_H} \end{bmatrix} \tag{2.24}$$

The denominator v_i is set as $\max(\epsilon, \epsilon\theta_{0,i})$ with $\epsilon = 10^{-6}$

The matrix $\mathbf{M}_0$ is a matrix whose elements are second order derivatives of $\eta(\theta_0)$ with respect to elements of θ_0 and its off-diagonal elements are given by

$$\begin{aligned} \frac{\partial^2\eta}{\partial\theta_{0,i}\partial\theta_{0,j}} &= \frac{1}{v_i v_j}\{\eta(\theta_{0,1},\ldots,\theta_{0,i}+v_i,\theta_{0,j}+v_j,\ldots,\theta_{0,H}) \\ &- \eta(\theta_{0,1},\ldots,\theta_{0,j}+v_j,\ldots,\theta_{0,H})\eta(\theta_{0,1},\ldots,\theta_{0,i}+v_i,\ldots,\theta_{0,H}) \\ &+ \eta(\theta_{0,1},\ldots,\theta_{0,H})\} \end{aligned} \tag{2.25}$$

and the diagonal elements are

$$\begin{aligned} \frac{\partial^2\eta(\theta_0)}{\partial\theta_{i,i}^2} = \frac{1}{v_i^2}\{&\eta(\theta_{0,1},\ldots,\theta_{0,i}+v_i,\ldots,\theta_{0,H}) - 2\eta(\theta_{0,1},\ldots,\theta_{0,H}) \\ &+\eta(\theta_{0,1},\ldots,\theta_{0,i}-v_i,\ldots,\theta_{0,H})\} \end{aligned} \tag{2.26}$$

The direction of iteration 0 to iteration 1 is obtained by minimizing the loss function with respect to $(\theta_1 - \theta_0)$ and the evolution of the parameter set θ is

$$\theta_1 - \theta_0 = \mathbf{M}_0^{-1}\nabla_0 \tag{2.27}$$

The process continues until the training is stopped by either a set tolerance criterion or there is no further change in the error function below the tolerance value or after a specified number of iterations. The drawbacks of this method are:-

(i) It is possible to obtain an inflexion or a local solution rather than the global solution which minimizes the error function. To overcome this, one may

start with a random vector and iterate until a convergence is reached and repeat the process with another random vector and then compare the two results. Also the iterations may be repeated several times until a potential minimum solution over the set of minimum values.

(ii) As the iterations progress, the Hessian matrix may become singular so that it is impossible to obtain its inverse for that particular iteration. This problem is solved by the use of algorithms that approximate the inverse on the basis of the change in gradient relative to the change in parameter.

The following are two examples of algorithms that are gradient based.

2.5.1.1 The Backpropagation Algorithm

Introduced by Werbos (1994). Here the inverse of the Hessian matrix, $-\mathbf{M}_0^{-1}$ is replaced by an identity matrix whose dimension is equal to the number of coefficients in the network multiplied by a learning parameter ρ is used. The following equation gives the relationship between the learning parameter and the Hessian matrix.

$$\begin{aligned} \theta_1 - \theta_0 &= -\mathbf{M}_0^{-1}\nabla_0 \\ &= -\rho\nabla_0 \end{aligned} \tag{2.28}$$

This learning parameter is specified at the beginning of the process and is usually small values in the interval [0.05,0.5]. It may also be endogenous taking different values as the process converges. As in the quasi-Newton algorithm, one may obtain a local rather than a global solution at the point of convergence. The process may be prolonged to convergency by selection of low values of the learning parameter. This is solved by adding a momentum term after n training periods given as

$$\theta_n - \theta_{n-1} = -\rho\nabla_{n-1} + \mu(\theta_{n-1} - \theta_{n-2}) \tag{2.29}$$

The adding of the momentum term, with μ usually set at 0.9 enables the parameter to move fast over the error surface.

The limitation of this algorithm is that it suffers from the trap of local rather than a global minimal. Also low values of the learning parameters may needlessly prolong the convergency process so that convergency may not be guaranteed. For more details see McNelis (2005a) page 69 to 70.

2.5.1.2 The Broyden Fletcher Goldfarb Shanno algorithm

The algorithm was independently developed by the authors Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970). From an initial guess θ_0 and an approximate Hessian matrix $\mathbf{M}_0$ the following steps are repeated until $\eta(\theta)$ converges to the solution.

1. Obtain the direction d_j at the j^{th} stage. This is given by the solution to the equation $\mathbf{M}_j d_j - \nabla\eta(\theta_j) = 0$.

2. Perform a line search to find an acceptable step-size α_j in the direction found in the first step, then update $\theta_{j+1} = \theta_j + \alpha_j \mathbf{d}_j$.

3. Set $\mathbf{s}_j = \alpha_j \mathbf{d}_j$.

4. Let $\mathbf{y}_j = \nabla\eta(\theta_{j+1}) - \nabla\eta(\theta_j)$.

5. $\mathbf{M}_{j+1} = \mathbf{M}_j + \dfrac{\mathbf{y}_j \mathbf{y}_j^T}{\mathbf{y}_j^T} \mathbf{s}_k - \dfrac{\mathbf{M}_j \mathbf{s}_j \mathbf{s}_j^T \mathbf{M}_j}{\mathbf{s}_j^T \mathbf{M}_j \mathbf{s}_j}$.

Practically, $\mathbf{M}_0$ can be initialized with $\mathbf{M}_0 = \mathbf{I}$, the identity matrix of the same dimension so that the first step will be equivalent to a gradient descent, but further steps are more and more refined by $\mathbf{M}_{j+1}$, the approximation to the Hessian.

The first step of the algorithm is carried out using the inverse of the matrix $\mathbf{M}_j$,

which is usually obtained efficiently by applying Sherman Morrison's formula to the fifth step of the algorithm, giving

$$\mathbf{M}_{j+1}^{-1} = \left(\mathbf{I} - \frac{s_j y_j^T}{y_j^T s_j}\right) \mathbf{M}_j^{-1} \left(\mathbf{I} - \frac{y_k s_k^T y_k^T s_k}{y_j^T s_j}\right) + \frac{s_j s_j^T}{y_j^T s_k}.$$

This can be computed efficiently without temporary matrices, since $\mathbf{M}_j^{-1}$ is symmetric, and that $\mathbf{y}_k^{\mathrm{T}} M_k^{-1} \mathbf{y}_k$ and $\mathbf{s}_k^{\mathrm{T}} \mathbf{y}_k$ are scalar, using an expansion such as

$$M_{k+1}^{-1} = M_k^{-1} + \frac{(\mathbf{s}_k^{\mathrm{T}} \mathbf{y}_k + \mathbf{y}_k^{\mathrm{T}} M_k^{-1} \mathbf{y}_k)(\mathbf{s}_k \mathbf{s}_k^{\mathrm{T}})}{(\mathbf{s}_k^{\mathrm{T}} \mathbf{y}_k)^2} - \frac{M_k^{-1} \mathbf{y}_k \mathbf{s}_k^{\mathrm{T}} + \mathbf{s}_k \mathbf{y}_k^{\mathrm{T}} M_k^{-1}}{\mathbf{s}_k^{\mathrm{T}} \mathbf{y}_k}.$$

In statistical estimation problems (such as maximum likelihood or Bayesian inference), confidence intervals for the solution can be estimated from the inverse of the final Hessian matrix. This algorithm to train the network in this work.

2.5.2 Simulated Annealing Search

The method was originally due to Metropolis et al. (1953) and later developed by Kirkpatrick et al. (1983). The idea originates from the theory of statistical mechanics and draws analogy between the annealing of solids and optimization. It differs from the other methods in that it neither uses the derivatives nor the Hessian matrix. The following steps are followed:-

i) With an initial guess θ_0 determine $\eta(\theta_0)$.

ii) Compute the j^{th} iteration temperature as $T(j) = \frac{\bar{T}}{1+ln(j)}$ where $\bar{T}$ is the temperature and cooling schedule parameter.

iii) Randomly perturbate the solution vector to obtain the j^{th} solution vector $\hat{\theta}_j$ and hence compute $\eta(\hat{\theta}_j)$.

iv) Generate from the uniform distribution the probability $P(j)$.

v) Compute the Metropolis ratio, $M_j = \exp(\frac{\eta(\hat{\theta}_j) - \eta(\theta_{j-1})}{T(j)})$.

vi) If $\hat{\theta}_j - \hat{\theta}_{j-1} < 0$,then $\theta_j = \hat{\theta}_j$ otherwise if $P(j) \leq M_j$ then $\theta_j = \hat{\theta}_j$.

vii) Repeat (ii) through (vi) until $j = \bar{T}$ where $j = 1, 2, \ldots, \bar{T}$.

The major drawback of procedure is that it is extremely slow for practical use and the asymptotic convergence is not guaranteed and thus one is not sure of finding the global optimum.

2.5.3 Evolutionary Stochastic Search

This method helps come up with a better initial guess. It reduces the likelihood of one getting trapped in a local minimum.The genetic algorithm which has the following steps is considered below:-

i) *Population creation*, where N(an even number) random vectors of order $H \times 1$ are created.

ii) At random two pairs of the random vectors are selected with replacement. In pairwise combinations the vectors are evaluated according to the sum of squared error function. The pair with the least sum of squares is taken as the one with a better fitness value. The winning vectors (i, j) are retained for *breeding* purposes. This is called the *selection* step.

iii) In this step, called the *crossover*, for each pair the algorithm uses any of the following three different ways with each of them having the same chance of $\frac{1}{3}$ being used. The methods are:-

   a) The *Shuffle* crossover, where $b$random draws are made from a binomial distribution and if the b^{th} draw is a success the parameters $\theta_{i,H}$ and $\theta_{j,H}$ are swapped otherwise no change is made.

   b) The *Arithmetic* crossover, where a random number a is chosen in the interval (0,1). This number is then used to generate a new pair of

the parameter vector that is a linear combination of the original pair vector is formed as $a\theta_{i,H} + (1-a)\theta_{j,H}; (1 - a\theta_{i,H} + a)\theta_{j,H}$.

c) The *Single point change*, where an integer t is randomly selected from the set [1,k-1]. The vectors are then split at integer t and the coefficients to the right of the split point $\theta_{i,t+1}, \theta_{j,t+1}$ are swapped. After the crossover operation, each point of the original vector is associated with two *offsprings* $C1(i)$ and $C2(j)$.

d) *Mutation.* With a small probability, p, which decreases with time each of the elements of the parameter vectors $C1(i)$ and $C2(j)$ are subjected to a mutation. This small probability is given by $p = .15 + \frac{.33}{g}$ for $g = 1, 2, \ldots, g'$ is called the generation number. Michalewicz (1996) proposed the following non-uniform algorithm to be used on an element of a vector if it has to undergo a mutation

$$\tilde{\theta}_{i,H} = \begin{cases} \theta_{i,H} + s[1 - r_2 \quad ^{(1-\frac{g}{g'})^b}] & \text{if} \quad r_1 > \frac{1}{2} \\ \theta_{i,H} - s[1 - r_2 \quad ^{(1-\frac{g}{g'})^b}] & \text{if} \quad r_1 \leq \frac{1}{2} \end{cases} \tag{2.30}$$

where b usually set at 2 is the parameter that governs the non-uniformity of the mutation, r_1 r_2 are two real numbers generated randomly in the interval [0,1] and s is a random number generated from the standard normal.

e) After the mutation the four 'family' members are subjected to a fitness tournament and the pair with the best fitness moving to the next generation while that with the worst fitness are extinguished. The process is repeated with parent i and j returning to the population pool for possible reselection until the next generation is populated by H vectors. This is called the *election tournament* stage.

f) This is an optional process called *elitism* where members of the new

and old generation are compared using the fitness criterion. If the best member of the old generation is dominated by a best member of the new generation then this member displaces the worst member of the new generation making the member eligible for selection in the coming generation.

g) *Convergence*. The process continues for g' generations and since convergence is measured by the fitness value of the best member of each generation g' should be large enough so that there are no changes in the fitness values of the best of several generations. The main disadvantage of this method is that it is slow. The various combinations and permutations of the elements of θ that the method finds optimal at various generations may become very large. Again one has a case of the dimensionality *curse*.

2.5.4 The stopping Rule

When training a network one has to be sure of when to stop. Hence one requires rules that would govern the stoppage of the training. The following are some of the rules commonly used.

(i) $\|\hat{\theta}_{j+1} - \hat{\theta}_j\| < \epsilon$, for $\epsilon > 0$.

Here if the difference between the magnitude of the network's parameters vector at the $(j+1)^{th}$ and the j^{th} iterations is less than a preset small positive number then the training is stopped.

(ii) $|\eta(\hat{\theta}_{j+1}) - \eta(\hat{\theta}_j)| < \epsilon$, for $\epsilon > 0$ but small.

The difference in the training error at the $(j+1)^{th}$ and the j^{th} iterations is obtained. If the absolute value of this difference is less than a preset small positive number then the training is stopped.

(iii) $\eta(\hat{\theta}_j) < \varepsilon_{min}$.

The training is stopped if the training error is less than a pre-assigned lower bound.

(iv) $j > I_{max}$. The training is stopped after a pre-assigned number of iterations.

(v) This rule combines any of the above four rules.

In this study rule(i) where ϵ was set at 10^{-4} and rule(iv) where a maximum of 500 iterations are used.

2.6 Models for Binomial Data

One way of analyzing data is seeking to establish a relationship between the observed response and the explanatory variables. For instance in credit scoring one might be interested in establishing the relationship between the rate of default with the age and the marital status of a client. This is called modeling of the data.

The objective is to come up with a mathematical relationship between the response variable and the explanatory variables together with a measure of uncertainty of the relationship. The model can be expressed as

$$response \quad variable = systematic \quad component + residual \quad component \tag{2.31}$$

where the systematic component summarizes how the variability in the response is explained for by values of the explanatory variables while the residual component takes care of any other unexplained variation. The linear model is considered below.

2.6.1 Linear Models

Let $y_1, \ldots, y_n$ be n observations of a random variable Y that linearly depend on a set of explanatory variables $X_1, \ldots, X_d$ then

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_d x_{di} + \epsilon_i \tag{2.32}$$

where $x_{1i}, \ldots, x_{di} = X_i^t$ are the explanatory variables for the i^{th} observation which are assumed to be known and without any error while $\beta_0, \beta_1, \ldots, \beta_d$ are the unknown parameters in the relationship between y_i and X_i. The last term ϵ_i is the unobservable random variable that represents the residue component assumed to have a mean of zero and a constant variance σ^2. It is also necessary for significance testing to assume that $\epsilon_i \sim N(0, \sigma^2)$.

The implication from the above is that each of the random variables Y_i has a mean that depends on the d explanatory variables. Thus one may denote this as $E(Y_i|X_i)$ where X_i is a vector of the explanatory variables. This makes the systematic component of the model which can be represented as

$$\beta_0 + \beta_1 x_{1i} + \ldots + \beta_d x_{di} = \beta_0 + \sum_{d=1}^{d} \beta_d x_{di} \tag{2.33}$$

2.6.2 Fitting Binomial Data into Linear Model

In binomial data the observed response for the i^{th} group is m_i, $i = 1, \ldots, b$ the number of successes in the group. The proportion $\dfrac{m_i}{n_i} = \tilde{p_i}$ is the approximate probability of success in the i^{th} group. The distribution of the observed response in the of the i^{th} group is $b(n_i, p_i)$ where n_i and p_i are the size and the probability of success of the i^{th} group respectively. Thus the observed response has a mean of $n_i p_i$ and a variance of $n_i p_i (1 - p_i)$.

It is rather natural to find out how the probability of success of the i^{th} group $p_i = E\left[\dfrac{m_i}{n_i}\right]$ can be explained in terms of the explanatory variables. This can be

done by considering the model

$$p_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_d x_{di} \tag{2.34}$$

and using the least square method to obtain the values of $\hat{\beta}_0, \ldots, \hat{\beta}_d$ for which

$$\sum_{i=1}^{b} \left(\frac{m_i}{n_i} - p_i \right)^2 = \sum_{i=1}^{b} \left(\frac{m_i}{n_i} - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_d x_{di} \right)^2 \tag{2.35}$$

is minimized. This approach makes the assumption that the variance of $\tilde{p}_i = \dfrac{p_i(1-p_i)}{n_i}$ is a constant. But this is not necessarily so since the variance depends on the true unknown values of p_i even if the n_i are equal. Normality of the variables is also assumed, though this is not a severe restriction since binomial distribution tends to normal as the sample size becomes large. The main drawback of this approach is that of the fitted values $\hat{p}_i$. The values of the unknown parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d$ are totally unrestricted and can take any value in the interval $(-\infty, \infty)$ and thus the linear combinations

$$\hat{p}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_d x_{di} \tag{2.36}$$

have no guarantee of being in the interval [0,1]. This technicality suggests that the linear models fitted using the least square method may be inappropriate. Hence the need to look into other models which are appropriate for binomial probabilities.

2.7 Models for Binomial Response Data

These models ensure that the fitted probabilities are in the range (0,1) by transforming the response variable so that equation (2.36) predicts the transformed response. Three such transformations are discussed .

2.7.1 The Logit Transformation

The *logit* of the success probability p is defined as

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{2.37}$$

It is noted that the function logit(p) is log-sigmoidal and is symmetric at the point $p = 0.5$. The function is essentially linear between the points $p = 0.2$ and $p = 0.8$ but outside this range it becomes non-linear, see Figure 2.5. This method has the advantage of having a direct interpretation in terms of the logarithm of the odds of a success which has an application in analysis of epidemiological data. It is also appropriate in analysis of data which has been collected retrogressively. When used to transform binomial responses then the responses can be summarized in terms of sufficient statistics.(see Collett (2002), page 58). In this work this transformation in the modeling our data.

2.7.2 The Probit Transformation

The *probit* of a probability p is defined to be the value ϱ for which

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\varrho}\exp(-\frac{1}{2}z^2)dz = p. \tag{2.38}$$

This is the standard normal so that $p = P(z \leq \varrho)$ and $\varrho = \phi^{-1}(p)$. This transformation is similar to the logit as it can be seen from Figure 2.5 but not as convenient in the computational point of view.

2.7.3 The Complementary Log-log Transformation

This transforms a probability of success p into $\log(-\log[1-p])$. The function is not symmetric about $p = 0.5$. The method is most appropriate in situations where the probabilities of success are dealt with in an asymmetric manner. From Figure 2.5, for small values of p the transformation is similar to the logit.

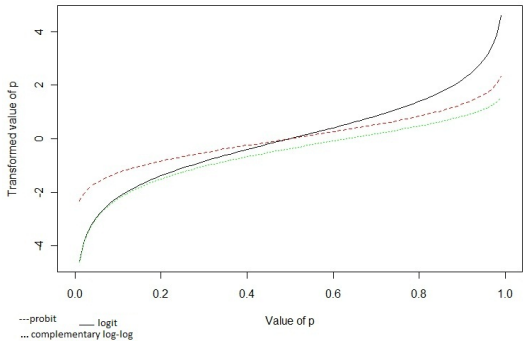

Figure 2.5: The logit, probit and the complementary $\log - \log$ transformations of p

2.8 The Logistic Regression Model

The mathematical concept underlying this model is the *logit* function which is defined as the natural logarithm of the odds ratio. For binomial observations, the probability of an outcome m_i given that the explanatory variable $X = x_i$ has the odds ratio given by

$$\frac{P(m_i|X = x_i)}{1 - P(m_i|X = x_i)} \tag{2.39}$$

where m_i are the number of successes in the i^{th} trial group. Hence the *logit* of the odds ratio is

$$\log \left[\frac{P(m_i|X = x_i)}{1 - P(X_i|X = x_i)} \right] \tag{2.40}$$

As in Chao-Ying and Gary (2002),

$$\begin{aligned} P(m_i|X = x_i) &= \frac{\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_d x_{di})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_d x_{di})} \\ &= \frac{1}{1 + \exp -(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_d x_{di})} \\ &= \frac{1}{1 + \exp -(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ji})} \end{aligned} \tag{2.41}$$

This is called the logistic regression model which can be linearized using the *logit* transformation to obtain

$$\begin{aligned}
\text{logit}(P(m_i|X=x_i)) &= \log\left[\frac{P(m_i|X=x_i)}{1-P(m_i|X=x_i)}\right] \\
&= \log\left[\frac{\frac{1}{1+\exp-(\beta_0+\sum_{j=1}^{d}\beta_j x_{ji})}}{1-\frac{1}{1+\exp(\beta_0+\sum_{j=1}^{d}\beta_j x_{ji})}}\right] \\
&= \log\left[\frac{\frac{1}{1+\exp-(\beta_0+\sum_{j=1}^{d}\beta_j x_{ji})}}{\frac{1+\exp-(\beta_0+\sum_{j=1}^{d}\beta_j x_{ji})-1}{1+\exp-(\beta_0+\sum_{j=1}^{d}\beta_j x_{ji})}}\right] \\
&= \log\left[\frac{1}{\exp-(\beta_0+\sum_{j=1}^{d}\beta_j x_{ji})}\right] \\
&= \log(\exp(\beta_0+\sum_{j=1}^{d}\beta_j x_{ji})) \\
&= \beta_0+\sum_{j=1}^{d}\beta_j x_{ji} \qquad (2.42)
\end{aligned}$$

Thus to estimate $P(m_i|X=x_i)$ is equivalent to estimating the function $g(X,\beta)=\beta_0+\sum_{j=1}^{d}\beta_j x_{ji}$. This may be done using either a parametric or a non-parametric method. In the parametric method one may use the m.l.e. to estimate the parameters. In this work a non-parametric approach where the conditional probability $P(m_i|X=x_i)$ is estimated using the output of a feedforward ANN, with the logistic function as its activation function is considered. Hence

$$\begin{aligned}
P(m_i|X=x_i) &= \frac{\exp(n_h(x;\theta))}{1+\exp(n_h(x;\theta))} \\
&= \frac{1}{1+\exp-(n_h(x;\theta))} \\
&= \varphi(x;\theta) \qquad (2.43)
\end{aligned}$$

as defined in equation(2.16).

2.9 Change point detection in binomial variables

In econometrics, testing for possible structural changes is of importance since a change in the generating process induces an instability in the original model. To detect change, one tests the null hypothesis H_o , which postulates that the model distribution does not change versus the alternative H_a that the model changes at an unknown point k. For the random variables $m_i \quad i = 1, 2, \ldots, b$ such that $m_i \sim b(n_i, p_i)$ and where m_i are number of successes in the i^{th} group of the sample then the model with a change at point $k, \quad 2 \leq k \leq b-1$ will be of the following form.

$$f(m_i, p_i, p_i') = \begin{cases} \begin{pmatrix} n_i \\ m_i \end{pmatrix} & p_i^{m_i}(1-p_i)^{n_i - m_i} \quad i = 1, 2, \ldots, k \\ \begin{pmatrix} n_i \\ m_i \end{pmatrix} & p_i'^{m_i}(1-p_i')^{n_i - m_i} \quad i = k+1, \ldots, b \end{cases} \tag{2.44}$$

Then our hypotheses will be

$$H_o \quad : \quad p_1 = p_2 = \ldots = p_b$$

Against

$$H_a \quad : \quad p_1 = p_2 = \ldots = p_k \neq p_{k+1}' = \ldots = p_b' \tag{2.45}$$

Several testing procedures exist in the literature. Three such procedures are discussed.

2.9.1 The likelihood Procedure

Assuming that the distributional form of the m_i is known upto the point k and that this form remains the same after point k then the likelihood ratio is given

by:-

$$\Lambda_k = \frac{\prod_{i=1}^{k} f(m_i, p_i) \prod_{i=k+1}^{b} f(m_i, p_i')}{\prod_{i=1}^{b} f(m_i, p_i)} \tag{2.46}$$

As in Gombay and Horvath (1994), H_o will be rejected for large values of

$$Q_b = - \max_{2 \le k \le b-1} 2 \log \Lambda_k \tag{2.47}$$

To obtain the maximum likelihood ratio Λ_k, let

$$\begin{aligned} M_k &= \sum_{i=1}^{k} m_i \\ N_k &= \sum_{i=1}^{k} n_i \\ M &= \sum_{i=1}^{b} m_i \\ N &= \sum_{i=1}^{b} n_i \\ M_k' &= M - M_k \\ N_k' &= N - N_k \end{aligned} \tag{2.48}$$

When H_0 is true, the likelihood function is

$$L_o(p_i) = \prod_{i=1}^{b} \binom{n_i}{m_i} p_i^{m_i} (1-p_i)^{n_i - m_i} \tag{2.49}$$

Thus the m.l.e. of p is $\dfrac{M}{N}$.

Similarly when H_a is true, the likelihood function is

$$L_a(p, p', k) = \prod_{i=1}^{k} \binom{n_i}{m_i} p_i^{m_i} (1-p_i)^{n_i - m_i} \prod_{i=k+1}^{b} \binom{n_i}{m_i} p_i'^{m_i} (1-p_i')^{n_i - m_i} \tag{2.50}$$

and the m.l.e. of p' is $\frac{M'_k}{N'_k}$.
The log-likelihood ratio Λ_k is obtained as

$$\begin{aligned}
\Lambda_k &= \log\left[\frac{L_a(\hat{p},\hat{p'})}{L_0(\hat{p})}\right] \\
&= -\sum_{i=1}^{b}\left[m_i\log\frac{M}{N}+(n_i-m_i)\log\left(1-\frac{M}{N}\right)\right] \\
&+ \sum_{i=1}^{k}\left[m_i\log\frac{M_k}{N_k}+(n_i-m_i)\log\left(1-\frac{M_k}{N_k}\right)\right] \\
&+ \sum_{i=k+1}^{b}\left[m_i\log\frac{M'_k}{N'_k}+(n_i-m_i)\log\left(1-\frac{M'_k}{N'_k}\right)\right] \\
&= -M\log\frac{M}{N}-(N-M)\log\frac{N-M}{N} \\
&+ M_k log\frac{M_k}{N_k}+(N_k-M_k)\log\frac{N_k-M_k}{N_k} \\
&+ (M-M_k)\log\frac{M'_k}{N'_k}+[N-M-(N_k-M_k)]\log\frac{N'_k-M'_k}{N'_k} \\
&= -M\log M-(N-M)\log(N-M)+N\log N \\
&+ M_k\log M_k+(N_k-M_k)\log(N_k-M_k)-N_k\log N_k \\
&+ M'_k\log M'_k+(N'_k-M'_k)\log(N'_k-M'_k)-N'_k\log N'_k \qquad (2.51)
\end{aligned}$$

Now H_o will be rejected if $Q_b = \max\limits_{2\le k\le b-1} -2\log\Lambda_k$ is large, that is $Q_b > C$ where C is a constant determined by the null distribution of Q_b, the sample size and α, the level of the test. This method is used in this work. Later in chapter 3 the asymptotic null distribution of Q_b will be given and thus C, the critical values obtained. If the null hypothesis is rejected then the maximum likelihood estimate of the change point position $\hat{k}$ is estimated as $\hat{k} = \arg\max\limits_{2\le k\le b-1} L_a(p,p',k)$.

2.9.2 The Cumulative Sum (CUSUM) Procedure

The CUSUM statistic Q_k at time k is the cumulative sum M_k of the successes upto time k, less the proportion r_kM where $r_k = \frac{N_k}{N}$ divided by the sample

standard deviation. Thus

$$Q_k = \frac{M_k - r_k M}{\sqrt{N p_o (1 - p_o)}} \tag{2.52}$$

It is noted that p_o is the m.l.e. of p under the null hypothesis.
Let $S_k^2 = r_k(1 - r_k)$, then $\frac{Q_k^2}{S_k^2}$ has the usual Pearson χ^2 statistic. Now H_o will be rejected if $Q_k = \max_{2 \le k \le b-1} Q_k$ is large, that is $Q_k > C$ where C is a constant determined by the null distribution of Q_k, the sample size and α, the level of the test.

2.9.3 Informational Procedure

The information criterion is an excellent tool in model selection. If the hypothesis testing problem in equation(2.45)is taken as a model selection problem then one may use this procedure to test for change.
If H_o is true then the likelihood is as in equation(2.49) and $SIC(b)$ is defined as

$$\begin{aligned} SIC(b) &= -2 \log L_o(p_o) + \log b \\ &= -2 \sum_{i=1}^{b} \log \binom{n_i}{m_i} - 2M \log(\frac{M}{N}) \\ &- 2(N - M) \log(\frac{N - M}{N}) + \log b \end{aligned} \tag{2.53}$$

Under H_a the likelihood is as in equation(2.50) and

$$\begin{aligned} SIC(k) &= -2 \log L_a(p, p', k) + \log b \\ &= -2 \sum_{i=1}^{b} \log \binom{n_i}{m_i} - 2M_k \log(\frac{M_k}{N_k}) - 2(N_k - M_k) \log(\frac{N_k - M_k}{N_k}) \\ &- 2M'_k \log(\frac{M'_k}{N'_k}) - 2(N'_k - M'_k) \log(\frac{N'_k - M'_k}{N'_k}) + 2 \log b \end{aligned} \tag{2.54}$$

The null hypothesis will be rejected if

$$SIC(b) > \min_{2 \le k \le b-1} SIC(k) \tag{2.55}$$

or reject H_o if

$$\min_{2\leq k\leq b-1} SIC(k)\Delta(k) < 0 \tag{2.56}$$

where

$$\begin{aligned}
\Delta(k) &= M\log M + (N-M)\log(N-M) - N\log N \\
&- M_k\log M_k - (N_k - M_k)\log(N_k - M_k) \\
&+ N_k\log N_k - M'_k\log M'_k - (N'_k - M'_k)\log(N'_k - M'_k) \\
&+ N'_k\log N'_k - \frac{1}{2}\log\frac{1}{b}
\end{aligned} \tag{2.57}$$

For more details on this method see Gichuhi et al. (2012).

CHAPTER THREE

CHANGE POINT DETECTION AND ESTIMATION

3.1 Introduction

In this section the assumption is that b independent groups of size $n_i \geq 2 \quad i = 1, 2, \ldots, b$ are observed. The observations are of the form $(m_i, X_i), \quad 1 \leq i \leq b$ where $\mathbf{X}_i = [x_{1i}, \ldots, x_{di}]^T \in \Re^d$ are the explanatory variables. The number of successes m_i are independent binomial random variables whose mean depends upon these explanatory variables.

For a standard change point problem the assumption that the observed data (m_i, X_i) are independent and the conditional distribution of $m_i|X_i = x$ is binomial with parameters $n_i, p_i(x)$. The hypotheses are

$H_0 : p_i(x) = p_0(x), \quad 1 \leq i \leq b$

against

$H_a : p_i(x) = p_0(x)$, for some $i \leq k$, and for some $i \geq k+1$, $p_i(x) = p'(x)$

where $2 \leq k \leq b-1$ is the unknown change point and $p_0(x) \neq p'(x)$ are the probabilities before and after the change point respectively.

Thus the general likelihood will be of the form

$$L(m_i, x, p) = \prod_{i=1}^{b} \begin{pmatrix} n_i \\ m_i \end{pmatrix} [p_i(x)]^{m_i} [1 - p_i(x)]^{n_i - m_i} \tag{3.1}$$

Since $p_i(x)$ is not known but is known to depend on the explanatory variables then one may use $\varphi(x; \theta)$ the output of a neural network in equation (2.12) to estimate it.

3.2 Model Definition

The observations (m_i, X_i) are independent binomial random variables whose probability density may be expressed as

$$f(m_i, X_i, p_i) = \begin{pmatrix} n_i \\ m_i \end{pmatrix} [p_i(x)]^{m_i}[1 - p_i(x)]^{n_i - m_i} \tag{3.2}$$

Since the functional form of $p_i(x)$ is not known one may approximate it in the model by replacing it with an output of the neural network defined in equation (2.12)

$$\begin{aligned} \varphi(x;\theta) &= \psi(n_h(x;\theta)) \\ &= \alpha_0 + \sum_{h=1}^{H} \alpha_h \{w_{h0} + \sum_{d=1}^{D} w_{hd} x_d\} \end{aligned} \tag{3.3}$$

where $\theta \in \Omega$ is as in equation (2.13). Assuming that the model is not misspecified, and as in Gombay and Horvath (1996), then (m_i, X_i) has a density function of the form

$$f(m, x;\theta) = \begin{pmatrix} n_i \\ m_i \end{pmatrix} \varphi(x;\theta)^{m_i}[1 - \varphi(x;\theta)]^{n_i - m_i} \tag{3.4}$$

3.3 Estimation of the Parameters

The parameters are estimated using the artificial neural network (ANN). A feed-forward network with a unipolar activation function, ψ as defined in equation (2.14)is used.

Now, using equations (3.2) and (3.3),

$$f(x;\theta) = \begin{pmatrix} n_i \\ m_i \end{pmatrix} \varphi(x;\theta)^{m_i}[1 - \varphi(x;\theta)]^{n_i - m_i} \tag{3.5}$$

so that the log-likelihood is

$$l(x;\theta) = \sum_{i=1}^{b} \left\{ \ln \begin{pmatrix} n_i \\ m_i \end{pmatrix} + \ln \varphi(x;\theta) + (n_i - m_i) \ln(1 - \varphi(x;\theta)) \right\} \tag{3.6}$$

The maximum likelihood estimator for θ will be

$$\hat{\theta} = \arg\max_{\theta \in \Omega} l(x;\theta) \tag{3.7}$$

A solution to equation (3.7) is guaranteed if the following conditions are satisfied.

(a) Ω the parameter space for θ is compact, which is common assumption when dealing with ANN.

(b) The activation function ψ chosen is continuous.

(c) The output of the network is such that $0 < \varphi(x;\theta) < 1$ for all x and θ.

Since $l(x;\theta)$ is continuous in θ for all m_i and attains its maximum on compact sets then the solution is guaranteed.

3.4 Testing for change points

In this section it is assumed that the data (m_i, X_i) are independent and the conditional distribution of $m_i|X_i = x$ is binomial with parameters n_i and p_i. Hence the change point problem will be stated as

$H_0 : p_i(x) = p_0(x), \quad 1 \leq i \leq b$

against

$H_a : p_i(x) = p_0(x)$, for some $i \leq k$, and for some $i \geq k+1, \qquad p_i(x) = p'(x)$

where $2 \leq k \leq b-1$ where k is the unknown change point and $p_0(x) \neq p'(x)$.

The general likelihood function will be of the form

$$L(m,x,p) = \prod_{i=1}^{b} \begin{pmatrix} n_i \\ m_i \end{pmatrix} [p_i(x)]^{m_i} [1 - p_i(x)]^{n_i - m_i} \tag{3.8}$$

The maximum likelihood ratio statistic will be given by

$$\Lambda_k = \frac{L_0(\hat{\theta}_0)}{L_a(\hat{\theta}_k, \hat{\theta}'_k)} \tag{3.9}$$

so that one has

$$\log \Lambda_k = \log L_0(\hat{\theta}_0) - \log L_a(\hat{\theta}_k, \hat{\theta}'_k) \tag{3.10}$$

As in Gombay and Horvath (1996), H_0 will be rejected if and only if

$$Q_b = \max_{2 \le k \le b-1} -2 \log \Lambda_k \ge C \tag{3.11}$$

The limiting distribution of Q_b is derived and the critical values given in a later section.

3.5 Model Irreducibility

A neural network with a fixed number of parameters is reducible if there exists another network with fewer neurons that has exactly the same input-output map. (see Hwang and Ding (1997)).

In this work an ANN model with the logistic function as the activation function is considered. For one to discuss the irreducibility of this model one needs to consider the hyperbolic tangent as it is a symmetric sigmoidal function. From equation(2.19) it is noted that there is a relation between the unipolar and the bipolar functions. If the logistic (unipolar) function is denoted by $\psi_1(x)$ and the hyperbolic tangent (bipolar) by $\psi_2(x)$ then as in equation (2.19),

$$\psi_2(x) = 2\{\psi_1(x)\} - 1 \tag{3.12}$$

From equations (2.9), (2.10), (2.11) and (2.12) the output of the hyperbolic tangent is

$$\begin{aligned}\varphi_1(x;\theta_1) &= \alpha'_0 + \sum_{h=1}^{H} \alpha'_h 2\{\psi_1(x)\} - 1 \\ &= (\alpha'_0 - \sum_{h=1}^{H} \alpha'_h) + \sum_{h=1}^{H} 2\alpha'_h \psi_1(x) \\ &= \varphi(x;\theta) \end{aligned} \tag{3.13}$$

with

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

Therefore it is possible to relate the parameter θ in the logistic function with θ_1 in the hyperbolic tangent if the weights w_{hd} remain the same. Hence irreducibility in θ_1 implies the same in θ.

A network is said to be reducible if at least one of following conditions is satisfied.

.

(a) $\alpha_h = 0$ for some $h = 1, \ldots, H$.

(b) One of the functions $n_h(x;\theta)$ is a constant; or

(c) There exist two indices $i, j \in (h = 1, \ldots, H)$ such that $n_i(x;\theta) = \pm n_j(x;\theta)$.

Note that a reducible θ, with a symmetric sigmoidal activation function leads to a redundant network since it has an input-output map that can be represented by another network by deletion of the h^{th} neuron.

If a network is redundant because of (a) above, then it obvious that the h^{th} neuron makes no contribution in the output and hence it can be deleted without affecting the input-output map.

If a network is redundant because of (b), then $n_h(x;\theta) = c$ one may delete the

h^{th} neuron and replace α_0 with $\alpha_0 + \alpha_h\psi(c)$. This arises if for a fixed value of h, $w_{hd} = 0$ for all $d = 1, \ldots, D$. In this situation then $n_h = w_{h0}$.

If a network is redundant because of (c), then $n_i = \kappa n_j$ where $\kappa = 1$ or -1 and the combined contribution of these two neurons is

$$\begin{aligned}\alpha_i\psi(n_i(x;\theta)) + \alpha_j\psi(n_j(x;\theta)) &= \alpha_i\psi(\kappa n_j(x;\theta)) + \alpha_j\psi(n_j(x;\theta)) \\ &= \kappa\alpha_i\psi(n_j(x;\theta)) + \alpha_j\psi(n_j(x;\theta)) \\ &= (\kappa\alpha_i + \alpha_j)\psi(n_j(x;\theta))\end{aligned}$$

This is due to the fact that the hyperbolic function is an odd function and therefore $\psi(\kappa x) = \kappa\psi(x)$. It is possible to delete the i^{th} neuron and replace α_j by $\kappa\alpha_i + \alpha_j$

To control irrelevant neurons that bring about conditions (a) and (b), one can use the SIC for model selection as in Swanson and White (1995). For a model with h hidden neurons then

$$SIC(h) = \ln\hat{\sigma}^2 + (h(2 + D) + 1)\frac{\ln(n)}{n} \tag{3.14}$$

The first term measures the goodness-of-fit while the second term is the complexity penalty. Using the SIC criterion, one starts 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. One therefore estimate $h + 1$ models in order to choose a model with h neurons. This procedure ensures that $\alpha_h \neq 0$ for all h . Thus one assumes that there exist no two different indexes $i, j \in h = 1, 2, \ldots, H$ such that the functions α_i and α_j are sign equivalent. This assumption solves the irreducibility caused by condition (c) above and thus ensuring that θ is irreducible hence non-redundant. The result translates immediately to the case of a unipolar activation. However even though θ is now irreducible, it is still unidentifiable as discussed in the following section.

3.6 Model Identifiability

Lets represent the weights of our network as α_0 and $\beta_h \quad$ for $\quad h = 1, \ldots, H$ where $\beta_h = (\alpha_h, w_h)$ and $w_h = w_{h,0}, w_{h,1}, \ldots, w_{h,d}$.

A theoretical problem of an ANN is the un-identifiability of the parameters. That is, there are two sets of parameters that the corresponding distributions of (m, X) are identical. The problem of un-identifiability has been looked into by Hwang and Ding (1997) who also noted that every ANN is un-identifiable. Lets consider the following two transformation

1. The permutations of β_h which is equivalent to interchanging any two hidden nodes h_p and h_q where p and q are the node positions and consequently their corresponding weights are also interchanged.

2. The negation of weights of p^{th} hidden node.

These two transformations do not alter the input-output map since

1. The interchange of the labels p and q will obviously not change the output function $\varphi(x;\theta)$. These permutations will give $H!$ different models with the same input-output map.

2. The activation function in equation(2.14) is symmetric i.e. $\varphi(x) = \varphi(-x)$ and hence $(\alpha_0, \beta_1, \ldots, \beta_h, \ldots, \beta_H)$ and $(\alpha_0, \beta_1, \ldots, -\beta_h, \ldots, \beta_H)$ have the same input-output map. These transformations will give 2^H different models with the same input-output map.

The transformations described above can be said to generate a family, τ, which has $2^H H!$ models. In this family $\varphi(x;\theta) = \varphi_\tau(x;\theta)$ and each transformation can be characterised as being a composite function of $\tau_1, \ldots, \tau_H$, where

$$\tau_1(\alpha_0, \beta_1, \ldots, \beta_h, \ldots, \beta_H) \;=\; (\alpha_0 + \alpha_1, -\beta_1, \ldots, \beta_h, \ldots, \beta_H) \qquad (3.15)$$

and

$$\begin{aligned}\tau_i(\alpha_0, \beta_1, \ldots, \beta_h, \ldots, \beta_H) &= (\alpha_0, \beta_h, \beta_2, \ldots, \beta_{h-1}, \beta_1, \beta_{h+1}, \ldots, \beta_H) \\ &\text{for} \quad h = 2, \ldots, H \end{aligned} \tag{3.16}$$

Waititu (2008) has shown in his work that if it is assumed that the model in equation (2.20), that is the bipolar activation function which is a continuous function, then condition A of Hwang and Ding (1997) is satisfied. Thus if θ is irreducible then, it is identifiable up to the family of transformations generated by equation(3.16). This implies that if there exist another $\breve{\theta}$ such that $\varphi(x;\theta) = \varphi(x;\breve{\theta})$ then a transformation exists in equation (3.16) that transforms $\breve{\theta}$ to θ.

3.7 Consistency and Asymptotic Normality of Network Parameter Estimates

In this section the assumption made is that (m_i, X_i) are independent binomial random variables with parameters $n_i, p_i(x) \quad i = 1, \ldots, b$. A neural network output $\varphi(x;\theta)$ is fitted to $p_i(x)$ by minimising the negative of the loglikelihood divided by b. That is the equation

$$l(\theta) = -\frac{1}{b}\sum_{i=1}^{b}\left\{\ln\begin{pmatrix} n_i \\ m_i \end{pmatrix} + m_i \ln\varphi(x;\theta) + (n_i - m_i)(1 - \ln\varphi(x;\theta))\right\} \tag{3.17}$$

is minimized.

The expected value of this target function $E(l(\theta))$ which is denoted by $l_0(\theta)$ is

$$\begin{aligned} l_0(\theta) &= -E\left\{\frac{1}{b}\sum_{i=1}^{b}\left\{\ln\begin{pmatrix} n_i \\ m_i \end{pmatrix} + m_i \ln\varphi(x;\theta) + (n_i - m_i)(1 - \ln\varphi(x;\theta))\right\}\right\} \\ &= -E\left\{\ln\begin{pmatrix} n_1 \\ m_1 \end{pmatrix} + m_1 \ln\varphi(x;\theta) + (n_1 - m_1)(1 - \ln\varphi(x;\theta))\right\} \\ &= -E\left\{\ln\begin{pmatrix} n_1 \\ n_1 p(X_1) \end{pmatrix} + n_1 p(X_1) \ln\varphi(X_1;\theta) + (n_1 - n_1 p(X_1)(1 - \ln\varphi(X_1;\theta))\right\} \end{aligned} \tag{3.18}$$

Assuming that $l_0(\theta)$ has a unique minimum if θ is in the compact set Ω, then this minimum is characterised by

$$\begin{aligned} \nabla l_0(\theta) &= -n_1 E\left\{\frac{p(X_1)}{\varphi(X_1;\theta)} - \frac{1 - p(X_1)}{1 - \varphi(X_1;\theta)}\right\}\nabla\varphi(X_1;\theta) \\ &= 0 \end{aligned} \tag{3.19}$$

Here the fact that the neural network output functions are continuous in x and θ and continuously differentiable in θ so that it is possible to interchange expectation and differentiation is used.

In a correctly specified situation where $p(x) = \varphi(x;\theta')$ for some $\theta' \in \Omega$ then equation (3.19) is solved but in a general situation where there is no true parameter value, θ' is defined as

$$\theta' = \arg\min_{\theta\in\Omega} l_0(\theta) \tag{3.20}$$

For an estimator $\hat{\theta}$ for θ' obtained by minimising equation (3.17), its consistency implies that $\hat{\theta} \to \theta'$ as $b \to \infty$.

The model may expressed as

$$m_i = n_i p(X_i) + \epsilon_i \quad i = 1, \ldots, b \tag{3.21}$$

where the residuals are

$$\epsilon_i = m_i - n_i p(X_i) \quad i = 1, \ldots, b \tag{3.22}$$

Since the observations (m_i, X_i) are independent and $P(m_i|X_i) = \frac{1}{n_i}E(m_i|X_i)$ then $E(\epsilon_i) = 0$ and

$$\begin{aligned} Var(\epsilon_i) &= E[(m_i - n_i p(X_i))^2] \\ &= E\{E[(m_i - n_i p(X_i))^2]|X_i\} \\ &= E\{E[m_i^2 - 2m_i n_i p(X_i) + (n_i p(X_i))^2]|X_i\} \\ &= E\{n_i(n_i - 1)p(X_i) + n_i p(X_i) - (n_i p(X_i))^2\} \\ &= E\{n_i p(X_i)(1 - p(X_i))\} \\ &= \sigma_\epsilon^2 < \infty \end{aligned} \tag{3.23}$$

Also it is noted that $Var(\epsilon_i)$ is independent of θ and $Var(\epsilon_i|X_i) = n_i p(X_i)(1 - p(X_i))$.

The following Uniform Law of Large Numbers will be used in the proof of the consistency of $\hat{\theta}$ Let $U_1, U_2, \ldots$ be independent random vectors in $\Re^D$, $\Omega \subseteq \Re^M$ compact, $\Upsilon : \Re^D \times \Omega \to \Re$ measurable such that

1. $E|\Upsilon(U_1; \theta)| < \infty \quad \forall \theta \in \Omega$

2. $\Upsilon(u; \theta)$ is Lipschitz continuous in θ that is for some $L(u) > 0$

3. $E(L(U_1)) < \infty$

Then $\sup_{\theta \in \Omega} \left| \frac{1}{b} \sum_{i=1}^{b} \Upsilon(U_i; \theta) - E[\Upsilon(U_1; \theta)] \right| \to 0$ in probability. The proof to this theorem is found in Andrews (1992)

Franke and Neumann (2000) in their work discussed nonlinear least square estimates for neural network parameters in which they made some assumptions. The

residuals in equation (3.22) are independent and also bounded in absolute value by m_i. Thus to follow their argument one may reduce these assumptions to

(i) The activation function ψ is bounded and twice continuously differentiable and $E(m_i|X_i = x)$ is also bounded.

This assumption is usually satisfied if the activation function is either unipolar or bipolar .

(ii) $l_0(\theta)$ has a unique global minimum at θ' in the interior of Ω and $\nabla^2 l_0(\theta') = A(\theta')$, which gives the Hessian matrix is positive definite.

This is a standard assumption in regression analysis.

(iii) Ω is chosen such that for some $\delta > 0$, $\delta \leq \varphi(x;\theta) \leq 1 - \delta$ for all $x \in \Re^d$, $\theta \in \Omega$.

This is a standard assumption.

(iv) $X_1, X_2, \ldots$ are independent random vectors with some density $\upsilon(x)$ and $var(X_1) < \infty$.

This is a standard assumption since the observed values of X_1 will have to be finite.

(v) $p(x)$ is continuous and for some $\nu > 0$, $0 < \nu \leq p(x) \leq 1 - \nu < 1$.

This assumption ensures that the experiments do not become degenerate. i.e. Not all the events in the experiment occur with probability of one or zero.

To derive the asymptotic normality of $\hat{\theta} - \theta'$ the two asymptotically independent components of $\hat{\theta} - \tilde{\theta}$ and $\tilde{\theta} - \theta'$ are separately considered where $\tilde{\theta} = \arg\min_{\theta \in \Omega} \tilde{l}(\theta)$

is generated by replacing m_i by $E(m_i|X_i = x)$ in equation (3.17) to obtain

$$\tilde{l}(\theta) = -\frac{1}{b}\sum_{i=1}^{b}\left\{\ln\begin{pmatrix} n_i \\ n_i p_i(x)\end{pmatrix} + n_i p_i(x)\ln\varphi(x;\theta) + (n_i(1-p_i(x)))(1-\ln\varphi(x;\theta))\right\} \tag{3.24}$$

The following theorem which is similar to theorem 1 of Franke and Neumann (2000) is used Suppose assumptions (i)-(v) are satisfied. Let $(m_i|X_i = x) \sim \mathcal{B}(n_i, p_i(x))$. Then as $b \to \infty$, with $\theta, \quad \theta'$ as defined above

$$\sqrt{b}\begin{pmatrix} \hat{\theta}-\tilde{\theta} \\ \tilde{\theta}-\theta' \end{pmatrix} \to d \quad \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\right)$$

that is $\sqrt{b}(\hat{\theta}-\tilde{\theta})$ and $\sqrt{b}(\tilde{\theta}-\theta')$ are asymptotically independent normal random vectors with covariance matrices Σ_1 and Σ_2 respectively, where

$$\Sigma_1 = A^{-1}(\theta')B_1(\theta')A^{-1}(\theta')$$
$$\Sigma_2 = A^{-1}(\theta')B_2(\theta')A^{-1}(\theta')$$

with

$$B_1(\theta') = E\left[\frac{(n_1 p(X_1))(1-p(X_1))}{\varphi^2(X_1;\theta')(1-\varphi(X_1;\theta'))^2}\right]\nabla\varphi(X_1;\theta')\nabla^t\varphi(X_1;\theta')$$
$$B_2(\theta') = E\left[\frac{(n_1 p(X_1)-\varphi(X_1;\theta'))^2}{\varphi^2(X_1;\theta')(1-\varphi(X_1;\theta'))^2}\right]\nabla\varphi(X_1;\theta')\nabla^t\varphi(X_1;\theta')$$

An immediate consequence of this theorem is that $\sqrt{b}(\hat{\theta}-\theta')$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\Sigma_1+\Sigma_2$. In a correctly specified model $\Sigma_2 = \mathbf{0}$ since there is essentially no effect due to the randomness of $X_i's$ implying that the difference $\hat{\theta}-\theta'$ is asymptotically of order smaller than $b^{-0.5}$ while in a misspecified case this difference is of order $b^{-0.5}$. It is also noted that $B_1(\theta')$ contains the variance of m_1 indicating its randomness while $B_2(\theta')$ contains the modeling bias and hence it would be zero if the model were correctly specified.

proof

The proof is done in four parts.

Part I

Using theorem (3.7), and taking $U_i = X_i$ then

$$\Upsilon(X_i;\theta) = -\left[\ln \begin{pmatrix} n_i \\ n_i p(X_i) \end{pmatrix} + n_i p(X_i)\ln\varphi(X_i;\theta) + (n_i - n_i p(X_i))(1-\ln\varphi(X_i;\theta))\right] \tag{3.25}$$

Thus

$$\sup_{\theta\in\Omega}|l(\tilde{\theta}) - l_0(\theta)| = \sup_{\theta\in\Omega}\left|\frac{1}{b}\sum_{i=1}^{n}\Upsilon(X_i;\theta) - E(\Upsilon(X_1;\theta))\right| = o_p(1) \tag{3.26}$$

Similarly

$$l(\theta) - \tilde{l}(\theta) = \frac{1}{b}\sum_{i=1}^{b}\ln\frac{\begin{pmatrix} n_i \\ n_i p(x) \end{pmatrix}}{\begin{pmatrix} n_i \\ m_i \end{pmatrix}} - (m_i - n_i p(x))\ln\frac{\varphi(X_i;\theta)}{1-\varphi(X_i;\theta)}$$

$l(\theta)$ and $\tilde{l}(\theta)$ as defined in equations (3.17) and (3.24).

Taking $U_i = (m_i, X_i)$ one obtains

$$\Upsilon(m_i, x;\theta) = \ln\frac{\begin{pmatrix} n_i \\ n_i p(x) \end{pmatrix}}{\begin{pmatrix} n_i \\ m_i \end{pmatrix}} - (m_i - n_i p(x))\ln\frac{\varphi(X_i;\theta)}{1-\varphi(X_i;\theta)}$$

then

$$\sup_{\theta\in\Omega}|l(\theta) - \tilde{l}(\theta)| = \sup_{\theta\in\Omega}\left|\frac{1}{b}\sum_{i=1}^{n}\Upsilon(m_i, X_i;\theta)\right| = o_p(1) \tag{3.27}$$

as $E(\Upsilon(m_1, X_1;\theta)) = 0$

One has to confirm whether the three conditions of the theorem (3.7) are satisfied in both cases. The activation function ψ is twice continuously differentiable and is bounded and so is $\varphi(x;\theta)$. This is considered in detail in a later section of this

chapter. As the derivative of $\varphi(x;\theta)$ is bounded then for some constant ω so that for all $x \in \Re^D$, $\theta \in \Omega$

$$\begin{aligned} \left|\frac{\partial}{\partial\theta_j}\varphi(x;\theta_j)\right| &\leq \omega \quad \text{if} \quad \alpha_0,\ldots,\alpha_h \quad w_{10},\ldots,w_{h0} \\ \left|\frac{\partial}{\partial\theta_j}\varphi(x;\theta_j)\right| &\leq \omega|x_i| \quad \text{if} \quad w_{1i},\ldots,w_{hi} \quad i=1,\ldots,D. \end{aligned} \tag{3.28}$$

Hence it follows that for a suitable constant $\omega^{'}$

$$\parallel \nabla\varphi(x;\theta) \parallel \leq \omega^{'} \parallel x \parallel$$

In a corresponding manner, for some constant $\omega^{''} > 0$

$$\parallel \nabla \ln\varphi(x;\theta) \parallel = \frac{\parallel \varphi(x;\theta) \parallel}{\varphi(x;\theta)}\omega^{''} \parallel x \parallel$$

and

$$\parallel \nabla \ln(1-\varphi(x;\theta)) \parallel = \frac{\parallel \varphi(x;\theta) \parallel}{1-\varphi(x;\theta)}\omega^{''} \parallel x \parallel$$

Hence for $\Upsilon(x;\theta)$ in equation (3.25)

$$\begin{aligned} |\Upsilon(u;\theta)-\Upsilon(u;\theta')| &\leq \sup_{\theta\in\Omega} \parallel \nabla\Upsilon(u;\theta) \parallel\parallel \theta-\theta' \parallel \\ &\leq \{n_i p(x)+(n_i-n_i p(x))\}\omega^{''} \parallel x \parallel\parallel \theta-\theta' \parallel \\ &= \omega^{''} \parallel x \parallel\parallel \theta-\theta' \parallel \end{aligned} \tag{3.29}$$

The assumption that (m_i, X_i) are independent with finite variance makes conditions (ii) and (iii) of theorem (3.7) to be satisfied with $L(u) = \omega^{''} \parallel u \parallel$.

Also from the third assumption made after the statement of the ULLN theorem and that $0 \leq p(x) \leq 1$ one has that $\Upsilon(u;\theta)$ is uniformly bounded in

$x \in \Re^D, \theta \in \Omega$.

Since m_i are bounded binomial random variables then a similar argument to the above is used for $\Upsilon(x;\theta)$ in equation (3.27) and therefore from equations (3.26) and (3.27)

$$|\hat{\theta} - \tilde{\theta}| = o_p(1) \text{ and } |\tilde{\theta} - \theta'| = o_p(1)$$

Hence it follows by assumption *(ii)* and with increasing probability that $\tilde{\theta}$, $\hat{\theta}$ are interior points in Ω. In particular

$$\nabla l(\hat{\theta}) = \nabla l(\tilde{\theta}) = \nabla l_0(\theta') = 0$$

with probability close to 1 as $b \to \infty$

Part II

With probability close to 1,

$$\begin{aligned} 0 &= \nabla\tilde{l}(\tilde{\theta}) - \nabla\tilde{l}(\theta') + \nabla\tilde{l}(\theta') \\ &= (\tilde{\theta}-\theta')\nabla^2 l_0(\theta') - \frac{1}{b}\sum_{i=1}^{b}\left\{\frac{n_i p(X_i)}{\varphi(X_i;\theta')} - \frac{(n_i-m_i)(1-p(X_i))}{1-\varphi(X_i;\theta')}\right\}\nabla\varphi(X_i;\theta') + F_1 \end{aligned} \tag{3.30}$$

where

$$\begin{aligned} F_1 &= \nabla\tilde{l}(\tilde{\theta}) - \nabla\tilde{l}(\theta') - (\tilde{\theta}-\theta') - \nabla^2\tilde{l}_0(\theta') \\ &+ (\tilde{\theta}-\theta')(\nabla^2\tilde{l}_0(\theta') - \nabla^2 l_0(\theta') \\ &= o_p(\| \ (\tilde{\theta}-\theta') \ \|) \end{aligned} \tag{3.31}$$

But

$$\begin{aligned} 0 &= \nabla l_0(\theta_0) \\ &= n_1 E\left\{\frac{p(X_1)}{\varphi(X_1;\theta')} - \frac{1-p(X_1)}{1-\varphi(X_1;\theta')}\right\}\nabla\varphi(X_1;\theta') \end{aligned} \tag{3.32}$$

and by the central limit theorem the middle term of equation (3.30) is of the order $b^{-0.5}$. Since it is possible to interchange expectations and differentiation and $\varphi(X_1;\theta')$ is bounded and bounded away from zero uniformly in $x \in \Re^D$, $\theta \in \Omega$ then the logarithms in functions $l(\theta)$, $\tilde{l}(\theta)$, $l_0(\theta)$ will all be defined.

Hence equation (3.30) becomes

$$\nabla^2 l_0(\theta')(\tilde{\theta}-\theta') + o_p(\| \ \tilde{\theta}-\theta' \ \|) = O(b^{-0.5}) \tag{3.33}$$

and since $\nabla^2 l_0(\theta')$ the Hessian is positive definite by assumption *(ii)* one has that

$$\| \ \tilde{\theta}-\theta' \ \| = O(b^{-0.5}) \tag{3.34}$$

Replacing $\nabla^2 l_0(\theta')$ with $A(\theta')$ then equation (3.30) becomes

$$\sqrt{b}(\tilde{\theta}-\theta') = A(\theta')^{-1}\frac{1}{\sqrt{b}}\sum_{i=1}^{b}\left\{\frac{n_i p(X_i)}{\varphi(X_i;\theta')} - \frac{(n_i-m_i)(1-p(X_i))}{1-\varphi(X_i;\theta')}\right\}\nabla\varphi(X_i;\theta') + o_p(1) \tag{3.35}$$

and hence for a suitable function s_1 satisfying $E[s_1(X_i)] = 0$ one obtains

$$\sqrt{b}(\tilde{\theta} - \theta') = b^{-.5} \sum_{i=1}^{b} s_1(X_i) + o_p(1) \tag{3.36}$$

Part III

From equations (3.21), (3.27) and that $E(\epsilon_i|X_i) = 0$, then with probability going to 1,

$$\Upsilon(m_i, x; \theta) = \ln \frac{\begin{pmatrix} n_i \\ n_i p(x) \end{pmatrix}}{\begin{pmatrix} n_i \\ m_i \end{pmatrix}} - (\epsilon_i) \ln \frac{\varphi(X_i; \theta)}{1 - \varphi(X_i; \theta)} \tag{3.37}$$

and

$$\begin{aligned} 0 = \nabla l(\hat{\theta}) &= \nabla \tilde{l}(\hat{\theta}) + \nabla\{l(\hat{\theta}) - \tilde{l}(\hat{\theta})\} \\ &= \nabla l(\tilde{\theta}) + \frac{1}{\sqrt{b}} \sum_{i=1}^{b} \Upsilon(m_i, x; \hat{\theta}) \\ &= \nabla l(\tilde{\theta}) - \frac{1}{\sqrt{b}} \sum_{i=1}^{b} (\epsilon_i) \frac{\nabla \varphi(X_i; \hat{\theta})}{\varphi(X_i; \hat{\theta})\{1 - \varphi(X_i; \hat{\theta})\}} \end{aligned} \tag{3.38}$$

As in part II of the proof,

$$\sqrt{b}(\hat{\theta} - \tilde{\theta}) = A^{-1}(\theta') b^{-.5} \sum_{i=1}^{b} \left\{ \frac{\nabla \varphi(X_i; \theta')}{\varphi(X_i; \theta')(1 - \varphi(X_i; \theta'))} \epsilon_i \right\} + o_p(1) \tag{3.39}$$

and hence for a suitable function s_2 satisfying $E s_2(X_i) = 0$ one obtains

$$\sqrt{b}(\hat{\theta} - \tilde{\theta}) = b^{-.5} \sum_{i=1}^{b} s_2(X_i)\epsilon_i + o_p(1) \tag{3.40}$$

Hence for some constants ω_1, ω_2 and for all $x \in \Re^D$

$$\| s_1(x) \| \leq \omega_1 \| x \|, \qquad \| s_2(x) \| \leq \omega_2 \| x \|$$

since $\nabla \parallel \varphi(x;\theta) \parallel$ is bounded. As $E \parallel X_i \parallel$ is finite , ϵ_i bounded and (X_i, ϵ_i) are *i.i.d.* one obtains

$$\sqrt{b}\begin{pmatrix} \hat{\theta}-\tilde{\theta} \\ \tilde{\theta}-\theta' \end{pmatrix} \stackrel{d}{\rightarrow} \quad \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\right) \tag{3.41}$$

as for all f, g

$$\begin{aligned} b\mathrm{Cov}(\tilde{\theta_f}-\theta'_f, \hat{\theta}_g-\tilde{\theta_g}) &= b^{-1}\sum_{i,j=1}^{b} E(s_{1f}(X_i s_{2g}(X_j\epsilon_j) + o_p(1) \\ &= b^{-1}\sum_{i\neq j}^{b} E(s_{1f}(X_i s_{2g}(X_j\epsilon_j) \\ &+ b^{-1}\sum_{i=1} E(s_{1f}(X_i s_{2g}(X_i\epsilon_i) + o_p(1) \\ &= o_p(1) \end{aligned} \tag{3.42}$$

as $E(\epsilon_i|X_i = 0)$

Part IV

The form now of Σ_1 and Σ_2 is now required.

$$\begin{aligned} \Sigma_1 &= E[s_1(X_1)s_1^t(X_1)] \\ &= A^{-1}(\theta')B_1(\theta')A^{-1}(\theta') \end{aligned} \tag{3.43}$$

since $s_1(X_i) \quad i = 1, 2, \ldots, b$ are *i.i.d.* and $E[s_1(X_i)] = 0$ where

$$B_1(\theta') = E\left\{\left[\tfrac{n_1 p(X_1)}{\varphi(X_1;\theta')} - \tfrac{(n_1-m_1)(1-p(X_1))}{1-\varphi(X_1;\theta')}\right]^2 \nabla\varphi(X_1;\theta')\nabla^t\varphi(X_1;\theta')\right\}$$

Similarly as $E[\epsilon_i^2|X_1] = \sigma^2_{\epsilon_i}$ as in equation (3.23) we have

$$\begin{aligned} \Sigma_2 &= E[s_2(X_1)s_2^t(X_1)\epsilon_i^2] \\ &= E[s_2(X_1)s_2^t(X_1)(n_i p(X_1)(1-p(X_1)))] \\ &= A^{-1}(\theta')B_2(\theta')A^{-1}(\theta') \end{aligned} \tag{3.44}$$

where

$$B_2(\theta') = E\left\{\frac{n_1 p(X_1)(1-p(X_1))}{\varphi^2(X_1;\theta')(1-\varphi(X_1;\theta'))^2}\nabla\varphi(X_1;\theta')\nabla^t\varphi(X_1;\theta')\right\}$$

Thus the theorem is proved. The limiting distribution of the change point statistic in equation (3.11) when the null hypothesis is true so as to perform the test of the hypotheses in equation (2.45)is now considered.

3.8 The Limit Distribution of the Change Point Statistic

In their work Gombay and Horvath (1996) gave conditions $C1 - C8$ which have to be satisfied by the probability distribution under consideration. Their enumeration is followed to show that the probability distribution $m_i|X_i = x$, the binomial distribution satisfies these conditions. This probability distribution is of the form

$$f(m_i, x, \theta) = \begin{pmatrix} n_i \\ m_i \end{pmatrix} [\varphi(x;\theta)]^{m_i}[1-\varphi(x;\theta)]^{n_i - m_i} \tag{3.45}$$

For simplicity purpose the subscript i on n and m is dropped from this point onwards in this section.

C1. $f(m, x, \theta)$ generates distinct measures in $\Omega_0 \times \Omega_1$ i.e. the densities $f(m, x, \theta_0)$ and $f(m, x, \theta')$ do not coincide.

proof

Now θ is identifiable from the function $f(m, x, \theta)$ if it is identifiable from $\varphi(x;\theta)$. Conditions for the identifiability are discussed in section 3.6. It is sufficient to assume that

$\theta = (w_{h,0}, w_{1,0}, \ldots, w_{H,D}, \alpha_0, \alpha_1, \ldots, \alpha_H)$ satisfies

c1 $\alpha_h > 0, \quad h = 1, \ldots, H$

c2 $w_h > 0, \quad h = 1, \ldots, H$

c3 $(w_h, w_{h,0}) \neq (w_{\acute{h}}, w_{\acute{h},0})$ for some $h \neq \acute{h}$

The following notations are made to enable us state and prove other conditions

$$\begin{aligned} g(m,x;\theta) &= \log f(m,x;\theta) \\ g_i(m,x;\theta) &= \frac{\partial}{\partial\theta_i} g(m,x;\theta) \\ g_{ij}(m,x;\theta) &= \frac{\partial}{\partial\theta_i\partial\theta_j} g(m,x;\theta) \\ g_{ijk}(m,x;\theta) &= \frac{\partial}{\partial\theta_i\partial\theta_j\partial\theta_k} g(m,x;\theta) \end{aligned} \tag{3.46}$$

C2. For each $k = 1, 2, 3, \ldots, b$, it is possible to find unique values $\hat{\theta_k}$ and $\hat{\theta'_k}$ such that

$$\sum_{1\le j\le k} g_i(m_j, x_j; \theta) = 0 \quad i = 1, 2, \ldots, D \tag{3.47}$$

$$\sum_{k\le j\le b} g_i(m_j, x_j; \theta') = 0 \quad i = 1, 2, \ldots, D \tag{3.48}$$

and

$$\sum_{1\le j\le k} g_i(m_j, x_j; \theta) + \sum_{k\le j\le b} g_i(m_j, x_j; \theta') = 0 \tag{3.49}$$

Note that $\hat{\theta_k}$ and $\hat{\theta'_k}$ are the values that maximize the loglikelihood function.

proof

In the estimation of parameters of neural network the assumption made is that the parameter set is chosen so that there are unique $\hat{\theta_k}$ and $\hat{\theta'_k}$.

It should be noted that if $k = 1 = b$, then there is no change point.

C3. There is an open set $\Omega \subseteq \Re^D$ containing θ_0 such that $g_i(m,x;\theta)$, $g_{ij}(m,x;\theta)$ and $g_{ijk}(m,x;\theta)$, $1 \le i,j,k \le D$ exist and are continuous in θ for all $m, x \in \Re$ and $\theta \in \Omega$.

proof

This condition is satisfied by the fact that $g_i(m,x;\theta)$,$g_{ij}(m,x;\theta)$ and $g_{ijk}(m,x;\theta)$

depend on θ only through $\varphi(x;\theta)$ which is continuously differentiable with respect to θ.

C4. There are functions $M_1(x)$ and $M_2(x)$ such that $|g_i(m,x;\theta)| \leq M_1(x)$, $|g_{ij}(m,x;\theta)| \leq M_2(x)$ and $|g_{ijk}(m,x;\theta)| \leq M_2(x)$ for all $x \in \Re$ and $\theta \in \Omega$, where M_1, M_2 satisfy

$$\sum_{m=0}^{n} \int M_1(x) dv(x) < \infty$$

and

$$E_{\theta_0} M_2(X_1) = \sum_{m=0}^{n} \int (M_2(x) \begin{pmatrix} n \\ m \end{pmatrix} [\varphi(x;\theta_0)]^m dv(x) [1-\varphi(x;\theta_0)]^{n-m}) dv(x) < \infty \tag{3.50}$$

since $0 \leq \varphi(x;\theta) \leq 1$, the latter is satisfied if

$$\sum_{m=0}^{n} \int M_2(x) dv(x) < \infty$$

proof

The proof will be given in three sections

section I

Obtaining the derivative of equation (3.45),

$$\begin{aligned} g_i(m,x;\theta) &= \frac{\partial}{\partial \theta_i} \log f(m,x;\theta) \\ &= m \frac{\varphi_i(x;\theta)}{\varphi(x;\theta)} + [n-m] \frac{\varphi_i(x;\theta)}{1-\varphi(x;\theta)} \end{aligned} \tag{3.51}$$

But $\varphi(x;\theta) = \psi(\zeta(x;\theta))$, implying that

$$\begin{aligned} \varphi_i(x;\theta) &= \frac{\partial}{\partial \theta_i} \psi(\zeta(x;\theta)) \\ &= \psi' \zeta(x;\theta) \frac{\partial}{\partial \theta_i} \zeta(x;\theta) \end{aligned} \tag{3.52}$$

Hence

$$\begin{aligned}
\psi(a) &= \frac{1}{1+e^{-a}} \\
\psi'(a) &= \frac{e^{-a}}{(1+e^{-a})^2} \\
&= \frac{(1+e^{a})e^{-a}}{(1+e^{a})(1+e^{-a})^2} \\
&= \frac{1+e^{-a}}{(1+e^{a})(1+e^{-a})^2} \qquad (3.53)
\end{aligned}$$

and

$$\frac{|\psi'(a)|}{\psi(a)} = \frac{1}{1+e^{a}} \in [0,1] \qquad (3.54)$$

also

$$\frac{|\psi'(a)|}{1-\psi(a)} = \frac{1}{1+e^{-a}} \in [0,1] \qquad (3.55)$$

Now

$$\begin{aligned}
|g_i(m,x;\theta)| &= m\frac{|\varphi_i(x;\theta)|}{\varphi(x;\theta)} + [n-m]\frac{|\varphi_i(x;\theta)|}{1-\varphi(x;\theta)} \\
&= m\frac{|\psi'(\zeta(x;\theta))|}{\psi(\zeta(x;\theta))}\left|\frac{\partial}{\partial\theta_i}\zeta(x;\theta)\right| \\
&+ [n-m]\frac{|\psi'(\zeta(x;\theta))|}{1-\psi(\zeta(x;\theta))}\left|\frac{\partial}{\partial\theta_i}\zeta(x;\theta)\right| \\
&\leq [n]|\frac{\partial}{\partial\theta_i}\zeta(x;\theta)| \quad \text{using equations (3.54)} \quad \text{and} \quad (3.55) \\
& \qquad (3.56)
\end{aligned}$$

Now one requires to determine the bound of $\frac{\partial}{\partial\theta_i}\zeta(x;\theta)$

Now using equation (3.52) and the possible values of the parameter θ_i, one has

For $\theta_i = \alpha_0$

$$\left|\frac{\partial}{\partial\alpha_0}\zeta(x;\theta)\right| = 1 \qquad (3.57)$$

For $\theta_i = \alpha_i, \quad i = 1,2,\ldots,H$

$$\left|\frac{\partial}{\partial\alpha_i}\zeta(x;\theta)\right| = \left|\psi(w_{h0} + \sum_{d=1}^{D} w_{hd}x_d)\right| \leq 1 \qquad (3.58)$$

For $\theta_i = w_{h0}$

$$\begin{aligned}
\left|\frac{\partial}{\partial w_{h0}}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{h0}}\left\{\alpha_0 + \sum_{h=1}^{H}\alpha_h\psi(w_{h0} + \sum_{d=1}^{D} w_{hd}x_d)\right\}\right| \\
&= \left|\alpha_h\psi'(w_{h0} + \sum_{d=1}^{D} w_{hd}x_d)\right| \\
&= \left|\frac{\alpha_h\psi(w_{h0} + \sum_{d=1}^{D} w_{hd}x_d)}{1 + \exp((w_{h0} + \sum_{d=1}^{D} w_{hd}x_d))}\right| \\
&\leq \left|\alpha_h\psi(w_{h0} + \sum_{d=1}^{D} w_{hd}x_d)\right|, \quad \text{since} \quad |1 - \psi(a)| \leq 1 \\
&\leq |\alpha_h|, \quad \text{since} \quad |\psi(a)| \leq 1 \qquad (3.59)
\end{aligned}$$

For $\theta_i = w_{hr}$ for some r

$$\begin{aligned}
\left|\frac{\partial}{\partial w_{hr}}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{hm}}\left\{\alpha_0 + \sum_{h=1}^{H}\alpha_h\psi(w_{h0} + \sum_{r=1}^{D} w_{hr}x_r)\right\}\right| \\
&= \left|\alpha_h x_r\psi'(w_{h0} + \sum_{d=1}^{D} w_{hr}x_r)\right| \\
&= \left|\frac{\alpha_h x_r\psi(w_{h0} + \sum_{m=1}^{D} w_{hr}x_r)}{1 + \exp((w_{h0} + \sum_{r=1}^{D} w_{hr}x_r))}\right| \\
&\leq \left|\alpha_h x_r\psi(w_{h0} + \sum_{r=1}^{D} w_{hr}x_r)\right|, \quad \text{since} \quad |1 - \psi(a)| \leq 1 \\
&\leq |\alpha_h||x_r|, \quad \text{since} \quad |\psi(a)| \leq 1 \qquad (3.60)
\end{aligned}$$

Hence to obtain $C4$ the function

$$M_1(x) = \max(n, \xi\sum_{r=1}^{D}|x_r|) \qquad (3.61)$$

may be used after making the following two assumptions

1. $|\alpha_h| \leq \xi$, $h = 1, \ldots, H$ for all $\theta \in \Omega$ and ξ a constant.

2. $E|X_{r1}| < \infty, r = 1, \ldots, D$ i.e. $E\|X_1\| < \infty$

section II

$g_{ij}(m, x; \theta)$ the second derivative of $g_i(m, x; \theta)$ is now required.

$$g_{ij}(m, x; \theta) = \frac{\partial}{\partial \theta_j} g_i(m, x; \theta) \tag{3.62}$$

Now

$$\begin{aligned} g_{ij}(m, x; \theta) &= \frac{\partial}{\partial \theta_j} \left\{ m \frac{\varphi_i(x; \theta)}{\varphi(x; \theta)} + [n - m] \frac{\varphi_i(x; \theta)}{1 - \varphi(x; \theta)} \right\} \\ &= m \left\{ \frac{\varphi_{ij}(x; \theta)}{\varphi(x; \theta)} - \frac{\varphi_i(x; \theta)\varphi_j(x; \theta)}{[\varphi(x; \theta)]^2} \right\} \\ &+ [n - m] \left\{ \frac{\varphi_{ij}(x; \theta)}{1 - \varphi(x; \theta)} + \frac{\varphi_i(x; \theta)\varphi_j(x; \theta)}{[1 - \varphi(x; \theta)]^2} \right\} \end{aligned} \tag{3.63}$$

But

$$\begin{aligned} \varphi_{ij}(x; \theta) &= \frac{\partial}{\partial \theta_j} \varphi_i(x; \theta) \\ &= \frac{\partial}{\partial \theta_j} \left\{ \psi'(\zeta(x; \theta)) \frac{\partial}{\partial \theta_i} \zeta(x; \theta) \right\} \\ &= \psi''(\zeta(x; \theta)) \frac{\partial}{\partial \theta_i} \zeta(x; \theta) \frac{\partial}{\partial \theta_j} \zeta(x; \theta) \\ &+ \psi'(\zeta(x; \theta)) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \zeta(x; \theta) \end{aligned} \tag{3.64}$$

Also

$$\begin{aligned} \psi''(a) &= \frac{e^{-a}}{(1 + e^{-a})^2(1 + e^a)} - \frac{e^a}{(1 + e^a)^2(1 + e^{-a})} \\ &= \frac{e^{-a} - e^a}{(1 + e^{-a})^2(1 + e^a)^2} \\ &= \frac{(1 + e^{-a}) + (1 + e^a)}{(1 + e^{-a})^2(1 + e^a)^2} \\ &= \frac{1}{(1 + e^{-a})(1 + e^a)^2} - \frac{1}{(1 + e^a)(1 + e^{-a})^2} \end{aligned} \tag{3.65}$$

And hence the deduction is that

$$\begin{aligned} \frac{\psi''(a)}{\psi(a)} &= \frac{1}{(1 + e^a)^2} - \frac{1}{(1 + e^a)(1 + e^{-a})^2} \\ &= (1 - \psi(a))^2 - \frac{\psi'(a)}{\psi(a)} \in [-1, 1] \end{aligned} \tag{3.66}$$

and

$$\begin{aligned}\frac{\psi''(a)}{1-\psi(a)} &= \frac{1}{(1+e^a)(1+e^{-a})^2} - \frac{1}{(1+e^a)^2} \\ &= \frac{\psi'(a)}{\psi(a)} - (1-\psi(a))^2 \in [-1,1] \end{aligned} \tag{3.67}$$

The bound of $g_{ij}(m,x;\theta)$ is obtained as follows,

$$\begin{aligned}|g_{ij}(m,x;\theta)| &= m\left\{\frac{|\psi''(\zeta(x;\theta))|}{\varphi(x;\theta)}\left|\frac{\partial}{\partial\theta_i}\zeta(x;\theta)\right|\left|\frac{\partial}{\partial\theta_j}\zeta(x;\theta)\right|\right\} \\ &+ m\left\{\frac{|\psi'(\zeta(x;\theta))|}{\psi(\zeta(x;\theta))}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\zeta(x;\theta)\right|\right\} \\ &+ m\left\{\left[\frac{|\psi'(\zeta(x;\theta))|}{\psi(\zeta(x;\theta))}\right]^2\left|\frac{\partial}{\partial\theta_i}\zeta(x;\theta)\right|\left|\frac{\partial}{\partial\theta_j}\zeta(x;\theta)\right|\right\} \\ &+ [n-m]\left\{\frac{|\psi''(\zeta(x;\theta))|}{1-\psi(\zeta(x;\theta))}\left|\frac{\partial}{\partial\theta_i}\zeta(x;\theta)\right|\left|\frac{\partial}{\partial\theta_j}\zeta(x;\theta)\right|\right\} \\ &+ [n-m]\left\{\frac{|\psi'(\zeta(x;\theta))|}{1-\psi(\zeta(x;\theta))}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\zeta(x;\theta)\right|\right\} \\ &+ [n-m]\left\{\left[\frac{|\psi'(\zeta(x;\theta))|}{1-\psi(\zeta(x;\theta))}\right]^2\left|\frac{\partial}{\partial\theta_i}\zeta(x;\theta)\right|\left|\frac{\partial}{\partial\theta_j}\zeta(x;\theta)\right|\right\}\end{aligned} \tag{3.68}$$

This implies that

$$|g_{ij}(m,x;\theta)| \leq n\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\zeta(x;\theta)\right| + 2n\left|\frac{\partial}{\partial\theta_i}\zeta(x;\theta)\right|\left|\frac{\partial}{\partial\theta_j}\zeta(x;\theta)\right| \tag{3.69}$$

using equations (3.66) and (3.67)

But from *section I* of the proof,

$$|g_i(m,x;\theta)||g_j(m,x;\theta)| \leq \begin{cases} \text{constant}.|x_i||x_j| & \theta_i = w_{hi}, \theta_j = w_{hj}; i,j \geq 1 \\ \text{constant} & \theta_i = \alpha_i, \theta_j = w_{hj} \\ \text{constant} & \theta_i = \alpha_i, \theta_j = \alpha_j \end{cases} \tag{3.70}$$

and by equation (3.61) one has

$$\left|\frac{\partial}{\partial\theta_i}\zeta(x;\theta)\right|\left|\frac{\partial}{\partial\theta_j}\zeta(x;\theta)\right| \leq M_1^2(x) \tag{3.71}$$

Hence for $E[M_2, X_1] < \infty$ to hold one must have at least have $E[X_{1l}^2] < \infty \quad l = 1, \ldots, d$ or $E||X_1||^2 < \infty$

Next the bound of $\left|\frac{\partial^2}{\partial\theta_i \partial\theta_j}\zeta(x;\theta)\right|$ is required

For $\theta_i = \alpha_i, \theta_j = \alpha_j$ then

$$\left|\frac{\partial^2}{\partial\theta_i \partial\theta_j}\zeta(x;\theta)\right| = 0 \tag{3.72}$$

For$\theta_i = \alpha_i, \theta_j = w_{hr}, h \neq i$ then

$$\left|\frac{\partial^2}{\partial\theta_i \partial\theta_j}\zeta(x;\theta)\right| = 0 \tag{3.73}$$

For $\theta_i = \alpha_h, \theta_j = w_{h0}$ then

$$\begin{aligned}\left|\frac{\partial^2}{\partial\theta_i \partial\theta_j}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{h0}}\psi(w_{h0} + \sum_{d=1}^{D} w_{hd}x_d)\right| \\ &= \left|\psi'(w_{h0} + \sum_{d=1}^{D} w_{hd}x_d)\right| \leq 1 \quad \text{by equation} \quad (3.54)\end{aligned} \tag{3.74}$$

For $\theta_i = \alpha_h, \theta_j = w_{hr}, r > 0$ then

$$\begin{aligned}\left|\frac{\partial^2}{\partial\theta_i \partial\theta_j}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{hr}}\psi(w_{h0} + \sum_{r=1}^{D} w_{hr}x_r)\right| \\ &= \left|\psi'(w_{h0} + \sum_{r=1}^{D} w_{hr}x_r)\right| |x_r| \leq |x_r| \end{aligned} \tag{3.75}$$

since $|\psi'(a)| \leq 1$ by equation

For $\theta_i = \alpha_{hr}, \theta_j = w_{h^*r^*}, h \neq h^*$ then

$$\left|\frac{\partial^2}{\partial\theta_i \partial\theta_j}\zeta(x;\theta)\right| = \left|\frac{\partial}{\partial w_{h^*r^*}}\alpha_h\psi(w_{h0} + \sum_{r=1}^{D} w_{hr}x_r)\right| \tag{3.76}$$

For $\theta_i = \alpha_h, \theta_j = w_{h0}$ then

$$\begin{aligned}\left|\frac{\partial^2}{\partial\theta_i \partial\theta_j}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{h0}}\alpha_h\psi'(w_{h0} + \sum_{d=1}^{D} w_{hd}x_d)\right| \\ &= \left|\alpha_h\psi''(w_{h0} + \sum_{d=1}^{D} w_{hd}x_d)\right| \\ &\leq |\alpha_h| \end{aligned} \tag{3.77}$$

since $|\psi''(a)| \leq 1$

For $\theta_i = w_{h0}, \theta_j = w_{hd}$ then

$$\begin{aligned}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{hd}}\alpha_h\psi'(w_{h0}+\sum_{d=1}^{D}w_{hd}x_d)\right| \\ &= \left|\alpha_h\psi''(w_{h0}+\sum_{m=1}^{D}w_{hd}x_d)x_d\right| \\ &\leq |\alpha_h||x_d| \end{aligned} \tag{3.78}$$

For $\theta_i = w_{hd}, \theta_j = w_{hr}$ then

$$\begin{aligned}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{hd}}\alpha_h\psi'(w_{h0}+\sum_{d=1}^{D}w_{hd}x_d)x_d\right| \\ &= \left|\alpha_h\psi''(w_{h0}+\sum_{r=1}^{D}w_{hr}x_r)x_dx_r\right| \\ &\leq |\alpha_h||x_d||x_r| \end{aligned} \tag{3.79}$$

Hence to have an appropriate bound one requires that $|\alpha_1|,\ldots,|\alpha_H|$ be bounded and $E[|x_d||x_r|] \leq \infty$ i.e. $E\|X_i\|^2 \leq \infty$

section III

In this section $g_{ijk}(m,x;\theta) = \dfrac{\partial}{\partial\theta_k}g_{ij}(m,x;\theta)$ is determined.

$$\begin{aligned} g_{ijk}(m,x;\theta) &= m\frac{\partial}{\partial\theta_k}\left\{\frac{\varphi_{ij}(x;\theta)}{\varphi(x;\theta)} - \frac{\varphi_i(x;\theta)\varphi_j(x;\theta)}{[\varphi(x;\theta)]^2}\right\} \\ &= [m]\frac{\varphi_{ijk}(x;\theta)}{\varphi(x;\theta)} \\ &- [m]\frac{\varphi_i(x;\theta)\varphi_{jk}(x;\theta)+\varphi_j(x;\theta)\varphi_{ik}(x;\theta)+\varphi_k(x;\theta)\varphi_{ij}(x;\theta)}{[\varphi(x;\theta)]^2} \\ &+ [2m]\frac{\varphi_i(x;\theta)\varphi_j(x;\theta)\varphi_k(x;\theta)}{[\varphi(x;\theta)]^3} \\ &+ [n-m]\frac{\varphi_{ijk}(x;\theta)}{[1-\varphi(x;\theta)]^2} \\ &+ [n-m]\frac{\varphi_i(x;\theta)\varphi_{jk}(x;\theta)-\varphi_j(x;\theta)\varphi_{ik}(x;\theta)-\varphi_k(x;\theta)\varphi_{ij}(x;\theta)}{[1-\varphi(x;\theta)]^2} \\ &- 2[n-m]\frac{\varphi_i(x;\theta)\varphi_j(x;\theta)\varphi_k(x;\theta)}{[1-\varphi(x;\theta)]^3} \end{aligned} \tag{3.80}$$

Hence the bound is

$$\begin{aligned}
|g_{ijk}(m,x;\theta)| &\leq [m]\frac{|\varphi_{ijk}(x;\theta)|}{\varphi(x;\theta)} \\
&+ [m]\frac{|\varphi_i(x;\theta)\varphi_{jk}(x;\theta)+\varphi_j(x;\theta)\varphi_{ik}(x;\theta)+\varphi_k(x;\theta)\varphi_{ij}(x;\theta)|}{[\varphi(x;\theta)]^2} \\
&+ [2m]\frac{|\varphi_i(x;\theta)\varphi_j(x;\theta)\varphi_k(x;\theta)|}{[\varphi(x;\theta)]^3} \\
&+ [n-m]\frac{\varphi_{ijk}(x;\theta)}{[1-\varphi(x;\theta)]^2} \\
&+ [n-m]\frac{|\varphi_i(x;\theta)\varphi_{jk}(x;\theta)+\varphi_j(x;\theta)\varphi_{ik}(x;\theta)+\varphi_k(x;\theta)\varphi_{ij}(x;\theta)|}{[1-\varphi(x;\theta)]^2} \\
&+ 2[n-m]\frac{|\varphi_i(x;\theta)\varphi_j(x;\theta)\varphi_k(x;\theta)|}{[1-\varphi(x;\theta)]^3} \qquad (3.81)
\end{aligned}$$

Each of the last four terms in equation above is bounded by constant+constant$|x_q x_r x_s|$ for an appropriate choice of subscripts as in *section II* of this proof.

$\varphi_{ijk}(x;\theta)$ is now worked out

$$\begin{aligned}
\varphi_{ijk}(x;\theta) &= \frac{\partial}{\partial\theta_k}\varphi_{ij}(x;\theta) \\
&= \frac{\partial}{\partial\theta_k}\left\{\psi''(\zeta(x;\theta))\frac{\partial}{\partial\theta_i}(\zeta(x;\theta))\frac{\partial}{\partial\theta_j}(\zeta(x;\theta))+\psi'(\zeta(x;\theta))\frac{\partial^2}{\partial\theta_i\partial\theta_j}\varphi(x;\theta)\right\} \\
&= \psi'''(\zeta(x;\theta))\frac{\partial}{\partial\theta_i}\zeta(x;\theta)\frac{\partial}{\partial\theta_j}\varphi(x;\theta)\frac{\partial}{\partial\theta_k}\varphi(x;\theta) \\
&+ \psi''(\zeta(x;\theta))\frac{\partial^2}{\partial\theta_i\partial\theta_k}\zeta(x;\theta)\frac{\partial}{\partial\theta_j}\varphi(x;\theta) \\
&+ \psi''(\zeta(x;\theta))\frac{\partial^2}{\partial\theta_j\partial\theta_k}\zeta(x;\theta)\frac{\partial}{\partial\theta_i}\varphi(x;\theta) \\
&+ \psi''(\zeta(x;\theta))\frac{\partial^2}{\partial\theta_i\partial\theta_j}\zeta(x;\theta)\frac{\partial}{\partial\theta_k}\varphi(x;\theta) \\
&+ \psi'(\zeta(x;\theta))\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\zeta(x;\theta) \qquad (3.82)
\end{aligned}$$

Now

$$\begin{aligned}\psi'''(a) &= \frac{d}{da}\left\{\frac{1}{(1+e^{-a})(1+e^{a})^2}+\frac{d}{da}-\frac{1}{(1+e^{a})(1+e^{-a})^2}\right\}\\ &= 2\left\{\frac{1}{(1+e^{a})(1+e^{-a})^2}-\frac{1}{(1+e^{a})(1+e^{-a})^3}-\frac{1}{(1+e^{-a})(1+e^{a})^3}\right\}\\ &= 2\{\psi(a)\psi'(a)-\psi^2(a)\psi'(a)-(1-\psi(a))^2\psi'(a)\}2\psi'(a)-2\psi^2(a)+3\psi(a)-1\\ &\in [-2,2] \qquad (3.83)\end{aligned}$$

Using this equation ,

$$\frac{\psi'''(a)}{\psi(a)} = 2\frac{\psi'(a)}{\psi(a)}\{-2\psi^2+3\psi(a)-1\} \in [-2,2] \qquad (3.84)$$

and

$$\frac{\psi'''(a)}{1-\psi(a)} = 2\frac{\psi'(a)}{1-\psi(a)}\{-2\psi^2+3\psi(a)-1\} \in [-2,2] \qquad (3.85)$$

For

$$\frac{|\varphi_{ijk}(x;\theta)|}{\varphi(x;\theta)} \leq constant + constant|x_q x_r x_s| + n\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\varphi(x;\theta)\right| \qquad (3.86)$$

using *section II* of the proof and therefore

$$|g_{ijk}(x;\theta)| \leq constant + constant|x_q x_r x_s| + n\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\zeta(x;\theta)\right| \qquad (3.87)$$

The next task will be to look for the bound of the derivative in (3.86)

For $\theta_i = \alpha_i, \theta_j = \alpha_j, \theta_k = \alpha_k$

$$\left|\frac{\partial^3}{\partial\alpha_i\partial\alpha_j\partial\alpha_k}\zeta(x;\theta)\right| = 0 \qquad (3.88)$$

For $\theta_i = \alpha_i, \theta_j = w_{hd}, \theta_k = w_{hd}, h \neq d$

$$\left|\frac{\partial^3}{\partial\alpha_i\partial w_{hd}\partial w_{hd}}\zeta(x;\theta)\right| = 0 \qquad (3.89)$$

For $\theta_i = \alpha_h, \theta_j = w_{h0}, \theta_k = w_{h0}$

$$\begin{aligned}\left|\frac{\partial^3}{\partial\alpha_h\partial w_{h0}\partial w_{h0}}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{h0}}\psi'(w_{h0}+\sum_{d=1}^{D} w_{hd}x_d)\right|\\ &= |\psi'(w_{h0}+\sum_{d=1}^{D} w_{hd}x_d)|\\ &\leq 1 \qquad (3.90)\end{aligned}$$

For $\theta_i = \alpha_h, \theta_j = w_{h0}, \theta_k = w_{hd} \quad d > 0$

$$\begin{aligned}\left|\frac{\partial^3}{\partial\alpha_h\partial w_{h0}\partial w_{hd}}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{hd}}\psi'(w_{h0}+\sum_{d=1}^{D} w_{hd}x_d)\right| \\ &= |x_d\psi''(w_{h0}+\sum_{d=1}^{D} x_d)| \\ &\leq |x_d| \end{aligned} \tag{3.91}$$

For $\theta_i = \alpha_h, \theta_j = w_{hd}, \theta_k = w_{hd}$

$$\begin{aligned}\left|\frac{\partial^3}{\partial\alpha_h\partial w_{hd}\partial w_{hd}}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{hd}}\psi'(w_{h0}+\sum_{d=1}^{D} w_{hd}x_d)\right| \\ &= |x_d x_d\psi''(w_{h0}+\sum_{d=1}^{D} x_d)| \\ &\leq |x_d x_d| \end{aligned} \tag{3.92}$$

For $\theta_i = w_{h0}, \theta_j = w_{h0}, \theta_k = w_{h0}$

$$\begin{aligned}\left|\frac{\partial^3}{\partial w_{h0}\partial w_{h0}\partial w_{h0}}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{h0}}\psi''\alpha_h(w_{h0}+\sum_{d=1}^{D} w_{hd}x_d)\right| \\ &= |\alpha_h\psi'''(w_{h0}+\sum_{d=1}^{D} x_d)| \\ &\leq 2|\alpha_h| \end{aligned} \tag{3.93}$$

since $|\psi'''(a)| \leq 2$ from equation (3.83)

For $\theta_i = w_{h0}, \theta_j = w_{h0}, \theta_k = w_{hd}$

$$\begin{aligned}\left|\frac{\partial^3}{\partial w_{h0}\partial w_{h0}\partial w_{hd}}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{h0}}\psi''\alpha_h(w_{h0}+\sum_{d=1}^{D} w_{hd}x_d)\right| \\ &= |\alpha_h x_d\psi'''(w_{h0}+\sum_{d=1}^{D} x_d)| \\ &\leq 2|\alpha_h||x_d| \end{aligned} \tag{3.94}$$

Finally for $\theta_i = w_{hr}, \quad \theta_j = w_{hs}, \quad \theta_k = w_{ht}$for $\quad r, s, t \geq 1$

$$\begin{aligned}\left|\frac{\partial^3}{\partial w_{hr}\partial w_{hs}\partial w_{ht}}\zeta(x;\theta)\right| &= \left|\frac{\partial}{\partial w_{hr}}\psi''\alpha_h(w_{h0}+\sum_{d=1}^{D} w_{hd}x_d)\right| \\ &= \left|\frac{\partial}{\partial w_{hr}}\psi'''\alpha_h(w_{h0}+\sum_{d=1}^{D} w_{hd}x_d)x_r x_s x_t\right| \\ &\leq |\alpha_h||x_r x_s x_t| \end{aligned} \tag{3.95}$$

Hence the function

$$M_2(x) = C_1 + \xi_1 \sum_{r,s,t}^{D} |x_r x_s x_t| \tag{3.96}$$

may be used for us to have $C4$ for large enough values of the constants C_1 and ξ_1. Since $|\alpha_1|, \ldots, |\alpha_H|$ are bounded uniformly in $\theta \in \Omega$ by the compactness of Ω, then for the integrability of M_1 and M_2 one needs to assume that M_2 is integrable with respect to the distribution of X_t. This is implied by $E\|X_1\| < \infty$

To enable one to state and proof a later condition one also assumes that

$$E\|X_1\|^4 < \infty \tag{3.97}$$

C5. $E_\theta[g_i(m, X_1; \theta)] = 0$ for all $\theta \in \Omega$ and i.

proof

Before proceeding with the proof it is worthwhile to note that m being a binomial random variable represents the number of successes in the n independent trials, so that $m = \sum_{i=1}^{n} Y_i$ where

$$Y_i = \begin{cases} 1 & \text{if the outcome is a success} \\ 0 & \text{otherwise} \end{cases}$$

Now $Y_i \quad i = 1, \ldots, n$ are independent Bernoulli random variables with $E(Y_i) = \varphi(x;\theta)$ so that $E(m) = \sum_{i=1}^{n} E(Y_i)$. Waititu (2008) in his work

proved that the Bernoulli form of p.d.f. follows conditions $C1$ to $C8$. Therefore to obtain $C5$,

$$\begin{aligned} E[g_i(m, X_1; \theta)] &= E\left\{E[g_i(m, X_1; \theta)|X_1]\right\} \\ &= E\left\{\sum_{j=1}^{n} E[\lambda_i(Y_j, X_1; \theta)|X_1]\right\} \\ &= 0 \end{aligned}$$

where $\lambda_i(Y, X_1; \theta)$ is as in Waititu (2008)

C6. $E_\theta[g_i(m, X_1; \theta)g_j(m, X_1; \theta)] = -E_\theta[g_{ij}(m, X_1; \theta)] = I_{ij}(\theta)$ for all $\theta \in \Omega$, $I_{ij}^{-1}(\theta)$ exist. Both $I_{ij}^{-1}(\theta)$ and $I_{ij}(\theta)$ are continuous in $\theta \in \Omega$, where $I_{ij}(\theta)$ is the information matrix.

proof

$$\begin{aligned} E[g_i(m, X_1; \theta)g_j(m, X_1; \theta)] &= E[E[g_i(m, X_1; \theta)|X_1 g_j(m, X_1; \theta)|X_1]] \\ &= E\left\{\sum_{k=1}^{n} E[\lambda_i(Y_k, X_1; \theta)|X_1 \lambda_j(Y_k, X_1; \theta)|X_1]\right\} \\ &\quad \text{where} \quad \lambda_i(Y, X_1; \theta) \quad \text{is as in Waititu (2008)} \\ &= n\left[\frac{\varphi_i(X_1; \theta)\varphi_j(X_1; \theta)}{\varphi(X_1; \theta)} + \frac{\varphi_i(X_1; \theta)\varphi_j(X_1; \theta)}{1 - \varphi(X_1; \theta)}\right] \end{aligned} \tag{3.98}$$

Similarly

$$\begin{aligned} E[g_{ij}(m, X_1; \theta) &= E[E[g_i(m, X_1; \theta)|X_1] \\ &= E\left\{\sum_{k=1}^{n} E[\lambda_{ij}(Y_k, X_1; \theta)|X_1]\right\} \\ &= -n\left[\frac{\varphi_i(X_1; \theta)\varphi_j(X_1; \theta)}{\varphi(X_1; \theta)} + \frac{\varphi_i(X_1; \theta)\varphi_j(X_1; \theta)}{1 - \varphi(X_1; \theta)}\right] \end{aligned} \tag{3.99}$$

Hence with the assumption of the invertibility of $I_{ij}(\theta)$, one obtains $C6$ With the above assumption, then for a true parameter $\theta \in \Omega_0$, $I_{ij}^{-1}(\theta)$ exists.

C7. $Var[g_{ij}(m, X_1; \theta)] < \infty$ for all i, j.

proof

To prove this condition it is sufficient to show that

$E[(g_{ij}(m, X_1; \theta))^2] < \infty$. From $C4$ $|g_{ij}(m, X_1; \theta)| \leq M_2(X_1)$ and from equation (3.96) depending on i, j and from the assumption (3.97) one has that

$Var[g_{ij}(m, X_1; \theta)] < \infty$

C8. $E|(g_i(m, X_1; \theta))|^q < \infty$ for all i for some $q > 2$.

proof

Now using equation (3.61) and the second assumption in (3.8), $C8$ is obtained with the assumption that $E\|X_1\|^{1+q} < \infty$

If conditions $C1-C8$ hold then under H_o, the test statistic $Q_b = \max_{1<k\leq b-1} -2\log\Lambda_k$ is such that

$P(a(\log b)Q_b^{\frac{1}{2}} \leq x + f(\log b)) = \exp(-2\exp(-x))$ for all $x \in \Re$ where $a(s) = (2\log s)^{\frac{1}{2}}$, $f(s) = 2\log s + \frac{d}{2}\log(\log s) - \log(\Gamma(\frac{d}{2}))$ and d is the dimension of θ. This theorem is similar to Theorem 2.1 of Gombay and Horvath (1996). For some level α and sample sizes the asymptotic critical values from theorem(3.8) are presented as C_1 in Table 3.1. It is noted that $\exp(-2\exp(-x))$ is the square of Gumbel distribution, which is an extreme value distribution and the rate of to these type of distributions convergency is usually slow. Hence theorem (3.8) gives conservative rejection regions when the sample sizes are moderate or small and only works well for large sample sizes. For small sample sizes Gombay and Horvath (1996) derived a further approximation of $Q_b^{\frac{1}{2}}$ If conditions $C1-C8$ hold then under H_o,

$$\left| Q_b^{\frac{1}{2}} - \sup_{\frac{1}{b}\leq t\leq 1-\frac{1}{b}} \left(\frac{B_b^{(d)}(t)}{t(1-t)} \right)^{\frac{1}{2}} \right| = O_p(\exp(-\log b)^{1-\epsilon})$$

for all $0 < \epsilon < 1$ where $B_b^{(d)}$ is a sequence of stochastic process distributed as $B_b^{(d)} = \sum_{1 \le i \le d} B_i^2(t), \quad 0 \le t \le 1$ and $B_i(t)$ are independent Brownian bridges. For $0 \le \alpha \le 1$ define

$$\begin{aligned} q_k \approx q_k(1-\alpha) &= \sup(x : P(Q_b^{\frac{1}{2}} \le x) \le 1-\alpha) \\ \text{and} & \\ v(r,s) &= v(r,s;1-\alpha) \\ &= \sup_{r \le t \le 1-s} \left(x : P\left(\frac{B_b^{(d)}(t)}{t(1-t)} \right)^{\frac{1}{2}} \le x \right) = 1-\alpha \end{aligned} \quad (3.100)$$

It is then shown that $v(r,s)$ is an asymptotically correct critical value of size α

Let conditions $C1 - C8$ and H_o hold.

If $r(b), s(b) \ge \frac{1}{b}$ and $\lim_{b\to\infty} \sup(b(r(b)+s(b)exp(-(\log b)^{1-\varepsilon})) < \infty$ where $0 \le \varepsilon \le 1$ then one has that $\lim_{b\to\infty} P\left(Q_b^{\frac{1}{2}} > v(r(b), s(b))\right) = \alpha$ and $|q_k - v(r(b), s(b))| = o((\log\log b)^{\frac{1}{2}})$ As in Gombay and Horvath (1996) put

$r(b) = s(b) = \frac{(\log b)^{\frac{3}{2}}}{b}$

which makes $v(r,s) \approx q_k$. But there is no known simple formula for the distribution function of $\sup_{r \le t \le 1-s} \left(\frac{B_b^{(d)}(t)}{t(1-t)} \right)^{\frac{1}{2}}$ and its inverted Laplace transform

$$P\left(\sup_{r \le t \le 1-s} \left(\frac{B_b^{(d)}(t)}{t(1-t)} \right)^{\frac{1}{2}} \ge x \right) = \frac{x^d \exp(-x^2/2)}{2^{d/2}\Gamma(d/2)} \left\{ T - \frac{d}{x^2}T + \frac{4}{x^2} + O(\frac{1}{x^4}) \right\} \quad (3.101)$$

where $T = \log \frac{(1-r)(1-s)}{rs}$ is used.

The asymptotic critical values from theorem (3.8)are presented as C_2 in Table 3.1.

Sample size	α	C_1	C_2
50	0.01	5.154013	4.787015
	0.05	4.167178	4.306045
	0.1	3.3731367	4.063449
100	0.01	5.219244	4.854494
	0.05	4.286601	4.385838
	0.1	3.874723	4.151836
150	0.01	5.249661	4.887406
	0.05	4.341763	4.42462
	0.1	3.940813	4.194628
200	0.01	5.268792	4.908558
	0.05	4.467199	4.449472
	0.1	3.982043	4.22199
481	0.01	5.310178	4.73092
	0.05	4.456003	4.274104
	0.1	4.078778	4.049254
500	0.01	5.319912	4.966611
	0.05	4.167178	4.517474
	0.1	4.09062	4.296645

Table 3.1: Critical values

3.9 Power of the test

In this section the assumption made is that the model is correctly specified. The test statistic is $Q_b = \max_{1<k\leq b-1} -2\log\Lambda_k$ where $\Lambda_k = \frac{L(\hat{\Omega}_o)}{L(\hat{\Omega}_a)}$ and $\hat{\Omega}_o$ contains $\hat{\theta}_o$ the m.l.e. of θ under the null hypothesis while $\hat{\Omega}_a$ contains $\hat{\theta}_k$, $\hat{\theta}_{k+1}$ the m.l.e. of θ under the alternative hypothesis before and after the change point respectively. Now the test statistic Q_b is an increasing function of $\max_{1<k\leq b-1} -2\log\Lambda_k$ and therefore the null hypothesis is rejected if Q_b is large, i.e. reject H_o if $Q_b \geq C$ where C is some bound that depends on the size α of the test and b the sample size. If $P_\theta(m_i|X_i)$ is the conditional probability of $m_i = m$ given that $X_i = x$ provided that θ is the true parameter then,

$$\Lambda_k = \prod_{i=1}^{k} \frac{P_{\hat{\theta}_0}(m_i|X_i)}{P_{\hat{\theta}_k}(m_i|X_i)} \prod_{i=k+1}^{b} \frac{P_{\hat{\theta}_0}(m_i|X_i)}{P_{\hat{\theta}_k^*}(m_i|X_i)} \tag{3.102}$$

where $\hat{\theta}_0 \in \hat{\Omega}_o$ and $\hat{\theta}_k, \hat{\theta}_k^* \in \hat{\Omega}_a$

From theorem (3.8), one notes that C grows asymptotically as b and for a given x depending on the size of the test then,

$$\begin{aligned} Q_b &= \frac{(x + f(\log b))^2}{a^2(\log b)} \\ &\approx 2\log b \end{aligned} \tag{3.103}$$

The argument is that the test is consistent in the sense that for a given size α its power converges to 1.

Under the alternative change occurs after a certain fraction of the data. That is there is a change point k, $2 \leq k \leq b-1$ such that as $b \to \infty$, then one has $k, b-k \to \infty$, $\frac{k}{b} = \iota \in (0,1)$.

Let θ_ι, θ_ι^* be the parameter values before and after the change point respectively and θ_0 denote the parameter value under the null hypothesis.

For consistency one has that as $b \to \infty$

$$\hat{\theta}_0 \to \theta_0 \quad ,\hat{\theta}_k \to \theta_\iota \quad \hat{\theta}_k^* \to \theta_\iota^*$$

So that by the law of large numbers

$$\frac{1}{b}\log\Lambda_k \sim \iota E_{\theta\iota}\log\frac{P_{\hat{\theta}_0}(m_i|X_i)}{P_{\hat{\theta}_\iota}(m_i|X_i)} + (1-\iota)E_{\theta_\iota^*}\log\frac{P_{\hat{\theta}_0}(m_i|X_i)}{P_{\hat{\theta}_\iota^*}(m_i|X_i)} \tag{3.104}$$

Under the alternative $\theta_\iota \neq \theta_\iota^*$ and $\theta_0 \neq \theta_\iota^*$, $\theta_0 \neq \theta_\iota$ by the definition of θ_0. If the model is correctly specified and the identifiability assumptions hold then, $P_{\theta_0} \neq P_{\theta_\iota}$, $P_{\theta_0} \neq P_{\theta_\iota{}^*}$. From Jensen's inequality and the fact the logarithm as a function is strictly concave one has that

$$\begin{aligned}
E_{\theta_\iota}\log\frac{P_{\theta_0}(m_i|X_i)}{P_{\theta_\iota}(m_i|X_i)} &= \log E_{\theta_\iota}\frac{P_{\theta_0}(m_i|X_i)}{P_{\theta_\iota}(m_i|X_i)} \\
&= \log\int\int\frac{P_{\theta_0}(m|X)}{P_{\theta_\iota}(m|X)}P_{\theta_\iota}(m|X)d\nu(x)d\mu(x) \\
&= \log\int\int P_{\theta_0}(m|X)d\nu(x)d\mu(x) \\
&= 0
\end{aligned}$$

similar results are obtained for the last term of equation (3.104). Hence for some constant $\gamma > 0$, $\frac{1}{b}\log\Lambda_k \approx -\gamma$. Thus $\log(\Lambda_k)^{-1} \approx b\gamma$. The size of type II error which depends on the power of the test under the alternative vanishes since

$$P(\max_{2\leq k\leq b-1}(\Lambda_k)^{-1} \leq C|H_a) \leq P((\Lambda_k)^{-1} \leq C|H_a) \to 0 \quad \text{as} \quad b\to\infty \tag{3.105}$$

as $(\Lambda_k)^{-1}$ changes as $e^{b\gamma}$ and C changes only as b. Thus the asymptotic power of the test is unity.

3.10 Testing for change in misspecified model

Suppose the form of $p_i(x)$ is not as that of the output of the neural network $\varphi(x;\theta)$. This implies that the model is not correctly specified. Since one still wishes to apply the test, we check how this misspecification affects the Gombay and Horvath (1996) conditions $C1 - C8$ stated earlier.

When the null hypothesis, that is there is no change, the true $p_i(x) = p_o(x)$ and if θ_o is the parameter value for which $\varphi(x;\theta)$ best approximates $p_o(x)$ then

$$\begin{aligned}
\theta_o &= \arg\max_{\theta\in\Omega} E\left\{\frac{1}{b}l(\theta)\right\} \\
&= \arg\max_{\theta\in\Omega} E\left\{\ln\begin{pmatrix} n_1 \\ m_1 \end{pmatrix} + m_1 \ln\varphi(x;\theta) + (n_1 - m_1)(1 - \ln\varphi(x;\theta))\right\} \\
&= \arg\max_{\theta\in\Omega} E\left\{E\left\{\ln\begin{pmatrix} n_1 \\ m_1 \end{pmatrix} + m_1 \ln\varphi(x;\theta) + (n_1 - m_1)(1 - \ln\varphi(x;\theta))\right\}|X_1\right\} \\
&= \arg\max_{\theta\in\Omega} E\left\{\ln\begin{pmatrix} n_1 \\ n_1 p(X_1) \end{pmatrix} + n_1 p_o(X_1)\ln\varphi(X_1;\theta)\right. \\
&+ \left.(n_1 - n_1 p_o(X_1))(1 - \ln\varphi(X_1;\theta))\right\}
\end{aligned} \tag{3.107}$$

Since in a misspecified model the density of (m_i, X_i) is still

$$f(m_i, X_i;\theta) = \ln\begin{pmatrix} n_i \\ m_i \end{pmatrix} + m_i \ln\varphi(x;\theta) + (n_i - m_i)(1 - \ln\varphi(x;\theta)) \tag{3.108}$$

then θ_o may also be expressed as

$$\begin{aligned}
\theta_o &= \arg\max_{\theta\in\Omega} \int\int \ln f(m_i, X_i;\theta) f_o(m_i, X_i) dv(m) du(x) \\
&= \arg\min_{\theta\in\Omega}\left\{-E\ln\frac{f(m_i, X_i;\theta)}{f_o(m_i, X_i)}\right\}
\end{aligned} \tag{3.109}$$

so that θ_o minimizes the Kullback-Leibler distance between the true density $f_o(m_i, X_i)$ and the approximating density $f(m_i, X_i;\theta)$.

The conditions of Gombay and Horvath (1996) under a misspecification situation are now considered.

(C1) Let $\theta_o, \quad \theta_o^* \in \Omega$. If $\theta_o \neq \theta_o^*$ then $\varphi(x;\theta_o \neq \varphi(x;\theta_o^*)$. Further if $\theta_o, \quad \theta_o^*$ are the solution to equation 3.106 for $f_o(m,x)$ and $f_o^*(m,x)$ respectively then $\theta_o \neq \theta_o^*$ implies that $\varphi(x;\theta_o \neq \varphi(x;\theta_o^*)$ and therefore $f_o(m,x) \neq f_o^*(m,x)$. Thus there are no different parameter values corresponding to the same distribution. To obtain the estimates of θ the approximating parametric model is

$$f(m_i, X_i; \theta) = \ln \begin{pmatrix} n_i \\ m_i \end{pmatrix} + m_i \ln \varphi(x;\theta) + (n_i - m_i)(1 - \ln \varphi(x;\theta)) \quad (3.110)$$

Using this condition one has that

$$\theta_o = \arg\max_{\theta \in \Omega} E\{f(m, x; \theta)\} \quad (3.111)$$

(C2) This condition is satisfied if the loglikelihood of the approximating parametric model has an unique maxima. This condition is equivalent to the identifiability condition on Ω.

(C3) This is a regularity condition on $f(m, x; \theta)$ which continues to hold.

(C4) This is another regularity condition on $f(m, x; \theta)$ which continues to hold but one must take into consideration the $E(M_2(m_1, X_1))$ which is with respect to the distribution $f_o(m, x; \theta)$ but

$$\begin{aligned} E_{\theta_0} M_2(X_1) &= \sum_{m=0}^{n} \int (M_2(x) \begin{pmatrix} n \\ m \end{pmatrix} [\varphi(x;\theta_0)]^m dv(x) [1 - \varphi(x;\theta_0)]^{n-m}) dv(x) \\ &< \infty \end{aligned}$$

if $E\|X\|^3 < \infty$

(C5) Using the regularity of $f(m, x; \theta)$ and the definition of θ_0 one has that

$$\nabla E(f(m_1, X_1; \theta_o)) = E\nabla(f(m_1, X_1; \theta_o)) = 0$$

Thus condition (C5) is satisfied.

(C6) It is the equivalent of the Fisher's information matrix in a correctly specified model. In a misspecified model one needs to assume that $I^{-1}(\theta)$ exists for $\theta = \theta_o$.

(C7 and C8) Continue to hold for $f(m, x; \theta)$.

3.10.1 The general testing for change points in a misspecified model

In this section we will digress from the binomial random variable and consider a general situation. Consider the random vectors $\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_b$ with respective densities $f_1(x), f_2(x), \ldots, f_b(x)$. The test is

$$H_0 \quad : \quad f_1(x) = f_2(x) = \ldots = f_b(x)$$

against

$$H_a \quad : \quad f_1(x) = f_2(x) = \ldots = f_k(x) \neq f_{k+1}(x) \tag{3.112}$$

for some $2 \leq k \leq b-1$

The form of $f_j(x)$ is not known but one may approximate it by some parametric density $f(x, \theta_j)$ and use maximum likelihood ratio test in the parametric hypotheses

$$H_0 \quad : \quad \theta_1 = \theta_2 = \ldots = \theta_b$$

against

$$H_a \quad : \quad \theta_1 = \theta_2 = \ldots = \theta_k \neq \theta_{k+1} \tag{3.113}$$

for some $2 \leq k \leq b-1$

Though θ_j is the parameter of the distribution of $\mathbf{X}_j$ it does not completely specify the density $f_j(x)$ as in Gombay and Horvath (1996). If the following identifiability

assumption is imposed then it is possible to test the original change-point problem in a parametric setup.

(a) For $\theta, \theta' \in \Omega$ and $\theta \neq \theta'$ the densities $f(x,\theta)$ and $f(x,\theta')$ do not coincide. Thus if the null hypothesis in equation(3.112) is rejected so is the null hypothesis in equation(3.113).

The aim here is to construct a parametric likelihood ratio test by adopting the misspecified parametric model that $\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_b$ are independent with densities $f(x,\theta_j)$, $j = 1, 2, \ldots, b$. The asymptotic behavior of the likelihood ratio statistic in this misspecified case is considered.

The choice of θ_j is made so as to minimize the Kullback-Leibler distance between $f(x,\theta_j)$ and $f_j(x)$. that is

$$\begin{aligned} \theta_j &= \arg\min_{\theta\in\Omega} -\left\{ \int f_j(x) \log \frac{f(x,\theta)}{f_j(x)} dx \right\} \\ &= \arg\max_{\theta\in\Omega} \int f_j(x) \log f(x,\theta) dx \\ &= E \log f(x,\theta) \end{aligned} \tag{3.114}$$

since the denominator of the logarithmic term is independent of θ. Under the null hypothesis in equation(3.113) one can assume that $\theta_1 = \theta_2 = \ldots = \theta_b = \theta_0$. Further to follow the argument of Gombay and Horvath (1996) similar notations are used to enable one state the other assumptions. Let

$$\begin{aligned} g(m,x;\theta) &= \log f(m,x;\theta) \\ g_i(m,x;\theta) &= \frac{\partial}{\partial\theta_i} g(m,x;\theta) \\ g_{ij}(m,x;\theta) &= \frac{\partial^2}{\partial\theta_i \partial\theta_j} g(m,x;\theta) \\ g_{ijl}(m,x;\theta) &= \frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_l} g(m,x;\theta) \end{aligned} \tag{3.115}$$

(b) For $k = 1, 2, \ldots, b$ there are unique solutions to the quasi likelihood equations, that is there are unique $\hat{\theta}_k$, $\hat{\theta'}_{b-k}$ such that

$\sum_{j=1}^{k} g_i(m,x;\hat{\theta}_k) = \sum_{j=k+1}^{b} g_i(m,x;\hat{\theta'}_{b-k}) = 0$ that is, there are unique quasi m.l.e. for θ and θ'. Thus the quasi-likelihood ratio statistic for testing for change in an approximate parametric model at a given value of k is

$$\begin{aligned}\Lambda_k &= \frac{\sup_{\theta\in\Omega}\prod_{j=1}^{b} f(X_j;\theta)}{\sup_{\theta,\theta'\in\Omega}\prod_{j=1}^{k} f(X_j;\theta)\prod_{j=k+1}^{b} f(X_j;\theta')} \\ &= \frac{\prod_{j=1}^{b} f(X_j;\hat{\theta})}{\prod_{j=1}^{k} f(X_j;\hat{\theta})\prod_{j=k+1}^{b} f(X_j;\hat{\theta'})} \end{aligned} \tag{3.116}$$

Thus $Q_b = \max_{1<k<b} -2\log\Lambda_k$ as the test statistic is considered. Further one requires to put smoothness and moments conditions. Let $\Omega_0 \subseteq \Omega$ be a suitably chosen compact set so that θ_0 is in the interior of Ω_0 and E_0 is the expectation under the null hypothesis that $f_j(x) = f_0(x) \quad j = 1,2,\ldots,b.$

(c) The derivatives of $g(m,x;\theta) = \log f(m,x;\theta)$ with respect to θ in equation (3.115) exist and are continuous in θ for all $\theta \in \Omega_0$, $i,j,k = 1,2,\ldots,D$

(d) There are functions $M_1(x)$ and $M_2(x)$ such that $\int M_1(x)dx < \infty$ and $E_0 M_2(x) < \infty$ so that

$$\begin{aligned} |g_i(m,x;\theta)| &\leq M_1(x) \\ |g_{ij}(m,x;\theta)| &\leq M_2(x) \\ |g_{ijl}(m,x;\theta)| &\leq M_2(x) \end{aligned} \tag{3.117}$$

(e) θ_0 is the unique solution to $E_0\nabla g(X_1;\theta) = 0$ in Ω_0

(f) $\mathbf{A}(\theta) = E_0\nabla^2 g(X_1;\theta)$ and $\mathbf{A^{-1}}(\theta)$ exist and are continuous in $\theta \in \Omega_0$ and $\mathbf{A_0}(\theta)$ is positive definite. $\mathbf{B}(\theta) = E_0\nabla g(X_1;\theta)\nabla^T g(X_1;\theta)$ and $\mathbf{B^{-1}}(\theta)$ exist and are continuous in $\theta \in \Omega_0$

(g) $\mathrm{Var}(g_{ij}(X_1;\theta)) < \infty$

(h) $E_0|g(X_1;\theta)|^\delta < \infty$ with some $\delta > 2$

To derive the asymptotic distribution of Q_b under the null hypothesis in (3.113) the argument of Gombay and Horvath (1994, 1996) is followed and digress only when there is a significant difference between the correctly and incorrectly specified cases. Some preliminary lemmas are given below and use the following abbreviations.

$$\begin{aligned} Z_k &= \sum_{j=1}^{k} \nabla g(X_j;\theta_0) \\ Z'_{b-k} &= \sum_{j=k+1}^{b} \nabla g(X_j;\theta_0)) \end{aligned} \tag{3.118}$$

$$\begin{aligned} \max_{1<k<b} \frac{k}{\log\log k} \left| \hat{\theta}_k - \theta_0 - \frac{1}{k}\mathbf{A^{-1}}(\theta_0) Z_k \right| &= O_p(1) \\ \max_{1<k<b} \frac{k}{\log\log k} \left| \hat{\theta'}_{b-k} - \theta_0 - \frac{1}{b-k}\mathbf{A^{-1}}(\theta_0) Z'_{b-k} \right| &= O_p(1) \end{aligned}$$

The proof to to this lemma is similar to lemma 2.1 of Gombay and Horvath (1994). As in the correctly specified case one has in the misspecified case that $\lim_{k\to\infty} \hat{\theta}_k = \theta_0$ since $\hat{\theta}_k$ is an M-estimate of θ_0. If (3.113) and assumptions (a)-(g) hold then as $b \to \infty$

$$\max_{1<k<b} \frac{k^{.5}}{(\log\log k)^{1.5}} \left| L(\hat{\theta}_k) - L(\theta_0) - \frac{k}{2}(\hat{\theta}_k - \theta_0)^T \mathbf{A}(\theta_0)(\hat{\theta}_k - \theta_0) \right| = O_p(1)$$

$$\max_{1<k<b} \frac{(b-k)^{.5}}{(\log\log(b-k))^{1.5}} \left| L'_{b-k}(\hat{\theta'}_{b-k}) - L'_{b-k}(\theta_0) - \frac{b-k}{2}(\hat{\theta'}_{b-k} - \theta_0)^T \mathbf{A}(\theta_0)(\hat{\theta'}_{b-k} - \theta_0) \right| = O_p(1)$$

where $L(.)$ and $L'(.)$ are the quasi loglikelihood function before and after k Proof of this theorem is similar to lemma 2.2 of Gombay and Horvath (1994). Lemma

3.10.1 and 3.10.1 both imply the following lemma

$$\begin{aligned}
\max_{1<k<b} \frac{k^{.5}}{(\log\log k)^{1.5}} \left| L(\hat{\theta_k}) - L(\theta_0) - \frac{1}{2k}(Z_k)^T \mathbf{A^{-1}}(\theta_0)(Z_k) \right| &= O_p(1) \\
\max_{1<k<b} \frac{(b-k)^{.5}}{(\log\log(b-k))^{1.5}} \left| L'_{b-k}(\hat{\theta'}_{b-k}) - L'_{b-k}(\theta_0) - \frac{b-k}{2}(Z'_{n-k})^T \mathbf{A^{-1}}(\theta_0)(Z'_{n-k}) \right| &= O_p(1)
\end{aligned}$$

Let $\mathbf{\Theta}_l = (\Theta_{l1}, \ldots, \Theta_{ld})^T$, $l = 1, 2, \ldots$ be a sequence of identically and independently distributed random vectors with $E(\mathbf{\Theta}) = 0$ and covariance matrix $E(\mathbf{\Theta}_l \mathbf{\Theta_1}^T) = \mathbf{I}_d$ and $\max_{1<l<d} E|\Theta_{1l}|^{2+\mu} < \infty$ for some $\mu > 0$. Then for all $x \in \Re$

$$\lim_{b\to\infty} P\left[a(\log b) \max_{1<k<b} \left(\frac{1}{b} \sum_{i=1}^{d} \left(\sum_{j=1}^{k} \Theta_{ji} \right)^2 \right)^{\frac{1}{2}} \leq x + f(\log b) \right] = \exp(-2\exp(-x))$$

where $a(s) = (2\log s)^{\frac{1}{2}}$, $f(s) = 2\log s + \frac{d}{2}\log(\log s) - \log(\Gamma(\frac{d}{2}))$ and d is the dimension of θ This theorem is similar to theorem 3.8 and the derivation of the asymptotic distribution of Z_b in a misspecified case is different only in the situation where, with $\Theta = \mathbf{A}^{\frac{1}{2}}(\theta_0)\nabla g(X_j;\theta)$. Now

$$\begin{aligned}
\frac{1}{2k} Z_k^T \mathbf{A^{-1}}(\theta_0) Z_k &= \frac{1}{2k} \left(\sum_{j=1}^{k} \Theta_j \right)^T \left(\sum_{j=1}^{k} \Theta_j \right) \\
&= \frac{1}{2k} \left(\sum_{i,j=1}^{k} \Theta_j^T \Theta_i \right) \\
&= \frac{1}{2k} \left(\sum_{i,j=1}^{k} \sum_{l=1}^{D} \Theta_{jl}^T \Theta_{il} \right) \\
&= \frac{1}{2k} \left(\sum_{l=1}^{D} \left(\sum_{j=1}^{k} \Theta_{jl}^T \Theta_{jl} \right)^2 \right) \qquad (3.119)
\end{aligned}$$

Using theorem 3.10.1 one can derive the asymptotic distribution of

$$\max_{1<k<b} \left(\frac{1}{2k} Z_k^T \mathbf{A^{-1}}(\theta_0) Z_k \right)^{.5}$$

which by Lemma 3.10.1 gives the asymptotic distribution of

$$\max_{1<k<b} \Big(L(\hat{\theta}_k) - L(\theta_0) \Big)^{.5}$$

Similarly one obtains the asymptotic distribution of

$$\max_{1<k<b}\left(L'_{b-k}(\hat{\theta}_{b-k})-L(\hat{\theta})\right)^{.5}$$

which by the argument of Gombay and Horvath (1994) gives the asymptotic distribution of $Z_b^{.5}$.

In the misspecified case $\mathbf{A}(\theta_0)$ and $\mathbf{B}(\theta_0)$ are not usually equal and the test statistic needs to be transformed. Considering a one-dimensional case where $d=1$, $\theta\in\Re$ so that $\mathbf{A}(\theta_0)$ and $\mathbf{B}(\theta_0)$ are scalars then by lemma 3.10.1 and conditions therein one has

$$\begin{aligned}\max_{1<k<b}\frac{k^{.5}}{(\log\log k)^{1.5}}\left|\frac{\mathbf{A}(\theta_0)}{\mathbf{B}(\theta_0)}\left(L(\hat{\theta_k})-L(\theta_0)\right)-\frac{1}{2k}\frac{Z_k^2}{\mathbf{B}(\theta_0)}\right| &= O_p(1)\\ \max_{1<k<b}\frac{(b-k)^{.5}}{(\log\log(b-k))^{1.5}}\left|\frac{\mathbf{A}(\theta_0)}{\mathbf{B}(\theta_0)}\left(L'_{b-k}(\hat{\theta'}_{b-k})-L'_{b-k}(\theta_0)\right)-\frac{1}{2k}\frac{(Z'_{n-k})^2}{\mathbf{B}(\theta_0)}\right| &= O_p(1)\end{aligned}$$

. To apply theorem 3.10.1 replace $\mathbf{A}(\theta_0)$ by $\mathbf{B}(\theta_0)$, so that

$$\frac{(Z_k)^2}{2k\mathbf{B}(\theta_0)}=\frac{1}{2k}\left(\sum_{j=1}^{k}(\Theta_j)^2\right)$$

where $\Theta_j=\mathbf{B}^{-.5}(\theta_0)\nabla f(X_j;\theta_0)$ are iid with mean of zero and variance 1 by the definition of $\mathbf{B}(\theta_0)$. From theorem 3.10.1 and lemma 3.10.1 and replacing $-2\log\Lambda_k$ in the proof of theorem 2.1 in Gombay and Horvath (1994) by $-2\left(\frac{\mathbf{A}(\theta_0)}{\mathbf{B}(\theta_0)}\log\Lambda_k\right)$, then one has the following theorem If the null hypothesis in equation(3.113)is true and assumptions (a)-(g) hold and d=1, the for all $x\in\Re$. Thus

$$\lim_{b\to\infty}P\left(a(\log b)\left[\frac{\mathbf{A}(\theta_0)}{\mathbf{B}(\theta_0)}Z_b\right]^{0.5}\le x+f(\log b)\right)=\exp(-2\exp(-x))$$

where $a(s)=(2\log s)^{\frac{1}{2}}$, $f(s)=2\log s+\frac{d}{2}\log(\log s)-\log(\Gamma(\frac{1}{2}))$ and d is the dimension of θ Since $\mathbf{A}(\theta_0)$ and $\mathbf{B}(\theta_0)$ are unknown, consider replacing θ_0 with its quasi maximum likelihood estimate $\hat{\theta}_b$ and the expectation are replaced by the

sample means. That is

$$\begin{aligned}\hat{\mathbf{A}}(\hat{\theta}_b) &= -\frac{1}{b}\sum_{i=1}^{b}\frac{\partial^2}{\partial\theta^2}g(X_i;\hat{\theta}_b)\\ \hat{\mathbf{B}}(\hat{\theta}_b) &= \frac{1}{b}\sum_{i=1}^{b}\left(\frac{\partial}{\partial\theta}g(X_i;\hat{\theta}_b)\right)^2\end{aligned}$$

If the null hypothesis in equation(3.113)is true and assumptions (a)-(g) hold and d=1, the for all x. Thus

$\lim_{b\to\infty} P\left(a(\log b)\left[\frac{\hat{\mathbf{A}}(\hat{\theta}_b)}{\hat{\mathbf{B}}(\hat{\theta}_b)}Z_b\right]^{0.5} \le x + f(\log b)\right) = \exp(-2\exp(-x))$ *Proof*

Let $\alpha = a(\log b)$, $\beta = f(\log b)$, $F_0 = \frac{\mathbf{A}(\theta_0)}{\mathbf{B}(\theta_0)}$ and $F_b = \frac{\hat{\mathbf{A}}(\hat{\theta}_b)}{\hat{\mathbf{B}}(\hat{\theta}_b)}$

Then

$\alpha(Z_bF_b)^{.5} - \beta = \{\alpha(Z_bF_0)^{.5} - \beta\} + \alpha(Z_b)^{.5}(F_b^{.5} - F_0^{.5})$

$\{\alpha(Z_bF_0)^{.5} - \beta\}$ has its asymptotic distribution given by theorem 3.10.1 so that one needs to show that $\alpha(Z_b)^{.5}(F_b^{.5} - F_0^{.5})$ is $o_p(1)$ as b approaches ∞.

Now $\beta \approx 2\log\log b$ as $b \to \infty$, $\alpha(Z_bF_b)^{.5} = O(\log\log b)$, so that there is need to show that $F_b^{.5} - F_0^{.5}$ is $o_p(\frac{1}{\log\log b})$. Under the null hypothesis in equation(3.113) and theorem 3.10.1 $\hat{\theta}_b \xrightarrow{p} \hat{\theta}_0$ and using the law of large numbers,

$$F_b^{.5} \xrightarrow{p} F_0^{.5} \neq 0 \quad b \to \infty$$

so that

$$F_b^{.5} + F_0^{.5} \xrightarrow{p} 2F_0^{.5}$$

hence

$$F_b^{.5} - F_0^{.5} = \frac{F_b - F_0}{F_b^{.5} + F_0^{.5}}$$

Thus $F_b^{.5} - F_0^{.5}$ is of the same order as $F_b - F_0$.

Using the law of large numbers it is enough to show that $\mathbf{A}(\hat{\theta}_b) - \mathbf{A}(\theta_0) = o_p(\frac{1}{\log\log b})$.

Using first order Taylor's expansion for the first derivative of $g(x)$ and the fact that the second derivative of $g(x)$ is bounded then

$$\begin{aligned}
\mathbf{A}(\hat{\theta}_b) - \mathbf{A}(\theta_0) &= \left| \frac{1}{b}\sum_{i=1}^{b} \frac{\partial^2\theta}{\partial\theta^2} g(X_i;\hat{\theta}_b) - E_0\frac{\partial^2\theta}{\partial\theta^2} g(X_i;\theta_0) \right| \\
&\leq \frac{1}{b}\sum_{i=1}^{b} \left| \frac{\partial^2\theta}{\partial\theta^2} g(X_i;\hat{\theta}_b) - \frac{\partial^2\theta}{\partial\theta^2} g(X_i;\theta_0) \right| \\
&+ \left| \frac{1}{b}\sum_{i=1}^{b} \frac{\partial^2\theta}{\partial\theta^2} g(X_i;\hat{\theta}_0) - E_0\frac{\partial^2\theta}{\partial\theta^2} g(X_1;\theta_0) \right| \\
&\leq \frac{1}{b}\sum_{i=1}^{b} M_2(X_j)|\hat{\theta}_b - \theta_0| \\
&+ \left| \frac{1}{b}\sum_{i=1}^{b} \frac{\partial^2\theta}{\partial\theta^2} g(X_i;\hat{\theta}_0) - E_0\frac{\partial^2\theta}{\partial\theta^2} g(X_1;\theta_0) \right| \qquad (3.120)
\end{aligned}$$

The term $\left| \frac{1}{b}\sum_{i=1}^{b} \frac{\partial^2\theta}{\partial\theta^2} g(X_i;\hat{\theta}_b) - E_0\frac{\partial^2\theta}{\partial\theta^2} g(X_i;\theta_0) \right|$ asymptotically coincide with $E_0 M_2(X_j)|\hat{\theta}_b - \theta_0|$ by law of large numbers and by theorem 3.7 the term $\hat{\theta}_b - \theta_0$ is $O_p(\frac{1}{\sqrt{b}})$. The term $\left| \frac{1}{b}\sum_{i=1}^{b} \frac{\partial^2\theta}{\partial\theta^2} g(X_i;\hat{\theta}_0) - E_0\frac{\partial^2\theta}{\partial\theta^2} g(X_1;\theta_0) \right|$ is $O_p(\frac{1}{\sqrt{b}})$ by the central limit theorem. Hence

$$\begin{aligned}
\mathbf{A}(\hat{\theta}_b) - \mathbf{A}(\theta_0) &= O_p(\frac{1}{\sqrt{b}}) \\
&= O_p(\frac{1}{\log\log b}) \qquad (3.121)
\end{aligned}$$

Thus the theorem is proved and hence for a one dimensional parameter under misspecification the test statistic will be $\hat{Q}_b = \max_{1<k<b} (-2\log\Lambda_k) \frac{\mathbf{A}(\hat{\theta}_b)}{\mathbf{B}(\hat{\theta}_b)}$.

3.11 Change point estimation

In this section the estimation of the change point is considered once it has been established that change exists, that is the alternative hypothesis is true. The maximum likelihood method is used. This method is discussed in the following

section and later simulation studies and analysis of the Bliss (1935) beetles data are conducted.

Several authors have considered this method. Ruhkin and Gary (1995) established the minimum error probability of the change-point maximum likelihood estimates for fixed binomial probabilities. Hinkley and Hinkley (1970) used the same method to estimate the change point when both the probabilities of success before and after the change-point are known. They also consider the situation where these probabilities are unknown and they replace them with their m.l.e. For known probabilities, p_o and p' the likelihood is given by equation(2.46). The m.l.e. $\hat{k}$ is the value of k that maximizes $L_a(p_0, p', k)$ and may be written as

$$\begin{aligned} \hat{k} &= \arg \max_{1<k\leq b-1} \left\{ \sum_{i=1}^{k} \ln \begin{pmatrix} n_i \\ m_i \end{pmatrix} + m_i \ln p_0 + (n_i - m_i) \ln(1-p_0) \right. \\ &+ \left. \sum_{i=k+1}^{b} \ln \begin{pmatrix} n_i \\ m_i \end{pmatrix} + m_i \ln p' + (n_i - m_i) \ln(1-p') \right\} \end{aligned} \tag{3.122}$$

$k = 2, 3, \ldots, b-1$

If the values of p_o and p' are unknown, they are replaced by their maximum likelihood estimates $\frac{M}{N}$ and $\frac{M'_k}{N'_k}$ respectively as defined in equation (2.51).

However our interest is on the conditional probabilities given by

$$P(m_i|X_i = x) = \begin{cases} p_o(x;\theta) & i = 1, 2, \ldots, k \\ p'(x;\theta) & i = k+1, \ldots, b \end{cases} \tag{3.123}$$

These probabilities depend on the explanatory variables X_i. The regression parameters in $p(x;\theta)$ cannot be estimated using the usual linear regression as discussed in section 2.6.2. In this work the neural networks and the logistic regression are used to estimate $p(x;\theta)$.

Now using equation(3.123) the maximum likelihood estimate for k

$$\hat{k} = \arg\max_{1<k\leq b-1} \left\{ \sum_{i=1}^{k} ln \begin{pmatrix} n_i \\ m_i \end{pmatrix} + m_i \ln p(x;\theta) + (n_i - m_i)\ln(1-p(x;\theta)) \right.$$
$$\left. + \sum_{i=k+1}^{b} \ln \begin{pmatrix} n_i \\ m_i \end{pmatrix} + m_i \ln p'(x;\theta) + (n_i - m_i)\ln(1-p'(x;\theta)) \right\}$$

$k = 2, 3, \ldots, b-1$

where $p(x;\theta)$ and $p'(x;\theta)$ are estimated from $(m_i, X_i)_{i=1}^{k}$ and $(m_i, X_i)_{i=k+1}^{b}$ respectively.

Since these conditional probabilities are in the interval(0,1), the logistic function discussed in section 2.4.1 is a suitable choice as an activation function in the network.

3.12 Confidence Interval For The Change Point Estimate

The standard procedure for computing a confidence interval for a parameter in a generalized linear model is by the use of the formula $estimate \pm percentile \times SE(estimate)$, where SE is the standard error. The percentile is selected according to a desired confidence level and a reference distribution. This procedure is commonly referred to as a Wald-type confidence interval. It may work poorly if the distribution of the parameter estimator is markedly skewed or if the standard error is a poor estimate of the standard deviation of the estimator. Various methods for constructing the confidence interval for the change point estimate exist in literature.Two such methods are discussed.

3.12.1 Profile likelihood method

Consider a model with parameters θ and k where k is the parameter of interest and θ is the additional parameter(s) in the model. Denote by $L(\theta; k)$ the likelihood

function. Then the profile likelihood function for k is

$$L_1(\theta; k) = \max_{\theta} L(\theta; k) \tag{3.124}$$

For each value of k , $L_1(\theta; k)$ is the maximum of the likelihood function over θ. Thus, the profile likelihood function is not a likelihood function but each point on the profile likelihood function is the maximum value of a likelihood function. The idea of a profile likelihood confidence interval is to invert a likelihood-ratio test statistic to obtain a CI for the parameter in question. A $100(1-\alpha)\%$ confidence interval for k is the set of all values k_0 such that a two-sided test of the null hypothesis $H_0 : k = k_0$ would not be rejected at the α level of significance. The likelihood ratio test statistic of the hypothesis $H_0 : k = k_0$ (where k_0 is a fixed value) equals the difference between twice the loglikelihood for the full model and twice the loglikelihood for the reduced model which has k fixed at k_0. i.e.

$$2[\log L(\hat{\theta}; \hat{k}) - \log L(\hat{\theta_0}; k_0)] = 2[\log L(\hat{\theta}; \hat{k}) - \log L_1(\theta_0; k_0)] \tag{3.125}$$

where $\hat{k}$ and $\hat{\theta}$ are the maximum likelihood estimates for the full model and $\hat{\theta_0}$ is the maximum likelihood estimates of θ for the reduced model with $k = k_0$. Based on the asymptotic chi-square distribution of the likelihood ratio test statistic, if the null hypothesis is true, then the test will not reject $H_0 : k = k_0$ at the α level of significance if and only if

$$2[\log L(\hat{\theta}; \hat{k}) - \log L_1(\theta_0; k_0)] \leq \chi^2_{1-\alpha}(\nu) \tag{3.126}$$

where ν the degrees of freedom of the test statistic, is the difference between dimension of θ under the full model and that under the reduced model. In this work $\nu = 1$. Thus the null hypothesis is accepted if and only if

$$\log L_1(\theta_0) \geq \log L(\hat{\theta}; \hat{k}) + \frac{\chi^2_{1-\alpha}(1)}{2} \tag{3.127}$$

where $\chi^2_{1-\alpha}(1)$ is the $1-\alpha$ quantile of a χ^2 distribution with 1 d.f. Since $\log L(\hat{\theta}; \hat{k})$ is fixed, one way plot the profile log-likelihood function, $\log L_1(\theta_0)$ and simply look

at the interval for which it exceeds $\log L(\hat{\theta}; \hat{k}) + \frac{\chi^2_{1-\alpha}(1)}{2}$. This is the $100(1-\alpha)\%$ confidence interval for k.

3.12.2 Percentile Bootstrap Confidence Interval

The idea behind bootstrap is to use the data of a sample study at hand as a population, for the purpose of approximating the sampling distribution of a statistic. Re-sample (with replacement) from the sample data at hand and create a large number of samples known as bootstrap samples. The most elementary application of bootstrapping is to produce a large number of copies of a sample statistic, computed from the bootstrap samples. Then, a small percentage, say $100(\alpha/2)\%$, is trimmed off from the lower as well as from the upper end of these numbers. The range of remaining $100(1-\alpha)\%$ values is declared as the confidence limits of the corresponding unknown population summary number of interest, with level of confidence $100(1-\alpha)\%$. This is referred to as bootstrap percentile method. In this work a sample (m_i, X_i), $i = 1, 2, \ldots, b$ is considered and the following is the procedure of obtaining the bootstrap percentile confidence interval.

1. From the original sample estimate the maximum likelihood estimate $\hat{k}$.
2. From the original covariate vector X_i, $i = 1, 2, \ldots, b$ obtain a bootstrap sample $X_{(i)}$ by drawing with replacement the integers $1, 2, \ldots, b$.
3. Calculate $m_{(i)}$ corresponding to the bootstrap sample.
4. Using the bootstrap sample $(m_{(i)}, X_{(i)})$ estimate the change point $\hat{k}_{(i)}$
5. Repeat steps 2 to 4 B times .
6. Arrange the B change point estimates $\hat{k}_{(i)}$ in ascending order.

CHAPTER FOUR

RESULTS AND DISCUSSIONS

4.1 Introduction

In this section both simulated and real data are used in the test for a change point in a sequence of binomial random variables and if the change exists the point at which it occurs is also estimated. In both cases the likelihood method discussed earlier is used. The power of the likelihood test if change exists is first considered.

4.2 Power of the test

The power of a change point test for finite sample size for a specific alternative of one change point was investigated.

The null hypothesis was rejected if the test statistic was large i.e. $Q_b^{0.5} > C$ where C is the asymptotic critical value which depends on the size of the test α and the size b of the sample is obtained using either Theorem 3.8 or 3.8.

For a given level α the power of the test for a specific alternative is the probability of accepting this alternative correctly which is given by:-

$$\kappa(\alpha) = P(Q_b^{0.5} > C|H_a) \tag{4.1}$$

Since the distribution of $Q_b^{0.5}$ under H_a is not known simulations were used to estimate the power of the test as follows:-

For a sample size b, B replicates were made and in each replicate $Q_b^{0.5}$ was estimated. Then the power at α was estimated as:-

$$\hat{\kappa}(\alpha) = \frac{1 + \text{no}(Q_b^{0.5} > C_b(\alpha))}{1 + B} \tag{4.2}$$

where no$(Q_b^{0.5} > C_b(\alpha))$ is the number of times $Q_b^{0.5} > C_b(\alpha)$. $C_b(\alpha)$ is the critical value of the test given in table 3.1

The model was assumed to be of the form $p(m_i|X_i = x) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ and using the logistic regression as in equation (2.40) one has that

$$P(m_i|X_i = x) = \frac{1}{1 + \exp -(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})} \tag{4.3}$$

For simulation purposes, it is assumed that H_a is true and used the following model

$$P(m_i|X_i = x) = \begin{cases} (1 + exp(-(-1.5 + x_{1i} + x_{2i})))^{-1}, & 1 \leq i \leq k \\ (1 + exp(-(-1.5 + 2x_{1i} + 1.8x_{2i})))^{-1}, & k+1 \leq i \leq b \end{cases} \tag{4.4}$$

where the values of β_0, β_1 are arbitrarily picked and β_2 as -1.5, 1 and 1 for $1 \leq i \leq k$. Similarly β_0,β_1 and β_2 as -1.5, 2 and 1.8 for $k + 1 \leq i \leq b$.

For a sample of size b=200, x_{1i} and x_{2i} were generated as *uniform*[0, 1]. n_i, the size of the i^{th} group was generated as the whole part of *uniform*[2, b].

4.2.1 Change of the power with change point location

The relationship between the power of the test and the location of the change point in the data is investigated.

The location of the change point k was placed at 20,40,50,100,150,160 and 180. Then the binomial random variable m_i was generated in line with equation (4.4). 500 simulations were done at each of the change point location. The value of the test statistic $Q_b^{0.5}$ in each of the 500 simulations was computed first using estimates of parameters from a generalized link function and then using a neural network. The critical values $C1$ and $C2$, which were generated using Theorem 3.8 and 3.8 respectively and presented in Table 3.1 were used. The power of the test was estimated using equation (4.2). The results are presented in Tables 4.1

and 4.2 respectively.

	$\hat{\kappa}(\alpha)$						
α	Change points location						
	20	40	50	100	150	160	180
0.01	0.003992	0.4411	0.9142	1	0.9800	0.7625	0.02994
0.01*	0.0099	0.5941	0.9901	1	0.9901	0.8705	0.03237
0.05	0.8323	1	1	1	1	1	0.9661
0.10	1	1	1	1	1	1	1

Table 4.1: Power of the likelihood ratio test from a sample size b= 200 using critical values C1.

	$\hat{\kappa}(\alpha)$						
α	Change points location						
	20	40	50	100	150	160	180
0.01	0.05389	0.9980	1	1	1	1	0.3094
0.01*	0.0791	1	1	1	1	1	0.5049
0.05	0.8762	1	1	1	1	1	0.9741
0.10	1	1	1	1	1	1	1

Table 4.2: Power of the likelihood ratio test from a sample of 200 using critical values C2.

A plot of the power of the test against the location of change point at $\alpha = 0.01$ is presented in Figure 4.1.

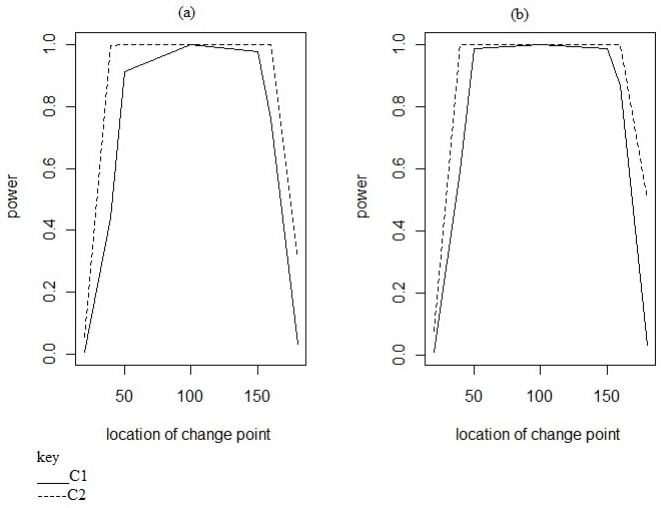

Figure 4.1: A plot of the power of the test against the location of change point.

Results in Table 4.1 and Table 4.2 show that the power of the test is less when the change point is located near the edges of the data. Two method are used to compute the conditional probabilities. Each value of the power of the test at $\alpha = 0.01$ was computed when the parameters were estimated using a parametric method while at $\alpha = 0.01^*$ the parameters were estimated using a neural network. The power of the test is higher when a neural network is used to estimate the parameters. The differences in the power as indicated in Figure 4.1 could be due to the fact that the critical values, C1 are in a squared Gumbel distribution, an extreme value distribution with a slow rate of convergence as noted in Gombay and Horvath (1996). The values in the Figure 4.1(b) were estimated using a parametric method while the values in the Figure 4.1(a) were estimated using a neural network.

When the change point is located in the upper edges, the test has more power compared with the power at the lower edges. This is due to the comparison of an

estimate calculated using a relatively smaller number of observations, in the first k observations and an estimate calculated using a large number of observations, in the last b-k observations.

The test has more power when the change location is near the centre of the data. Thus the test will most probably detect a change when the change point is near the centre. This is due to the comparison of an estimate calculated using an almost equal number of observations before and after the change point. This is as noted by Jaruskova (1997).

4.2.2 Change of the power with sample size

Here the effect of the size of the sample on the power of the test is investigated. The change point k was then put at $\frac{b}{4}$, $\frac{b}{2}$ and $\frac{3b}{4}$ for the samples sizes 50,100,150,200 and 500. For each sample, the power of the test at each change point location was evaluated. 500 simulations were done to determine each estimate and critical values C1 were used. The results are presented in Tables 4.3, 4.4 and 4.5.

	$\hat{\kappa}(\alpha)$				
α	Sample size				
	50	100	150	200	500
0.01	0.005988024	0.001996008	0.005988024	0.9121756	1
0.01*	0.008594	0.009102	0.015620	1	1
0.05	0.001996008	0.02794411	0.998004	1	1
0.10	0.01596806	0.7325349	1	1	1

Table 4.3: Power of the likelihood ratio test when the change point is at $\frac{b}{4}$

.

Table 4.3, Table 4.4 and Table 4.5 indicate that an increase in the sample size increases the power of the test, as expected. As in Table 4.1 and Table 4.2 the

	$\hat{\kappa}(\alpha)$				
α	Sample size				
	50	100	150	200	500
0.01	0.003992016	0.001996008	0.0259481	0.9780439	1
0.01*	0.0023297	0.026902	0.039186	1	1
0.05	0.003992016	0.0998004	1	1	1
0.10	0.02794411	0.8742515	1	1	1

Table 4.4: Power of the likelihood ratio test when the change point is at $\frac{b}{2}$

.

	$\hat{\kappa}(\alpha)$				
α	Sample size				
	50	100	150	200	500
0.01	0.001996008	0.001996008	0.003992016	0.1836327	1
0.01*	0.002583	0.0027153	0.039142	1	1
0.05	0.001996008	0.3812375	1	1	1
0.10	0.06586826	0.998004	1	1	1

Table 4.5: Power of the likelihood ratio test when the change point is at $\frac{3b}{4}$

.

parameter in $\alpha = 0.01$ were estimated using a parametric method while those in $\alpha = 0.01^*$ were estimated using a neural network.

A plot of the power of the test against the size of the sample at $\alpha = 0.01$ is presented in Figure 4.2.

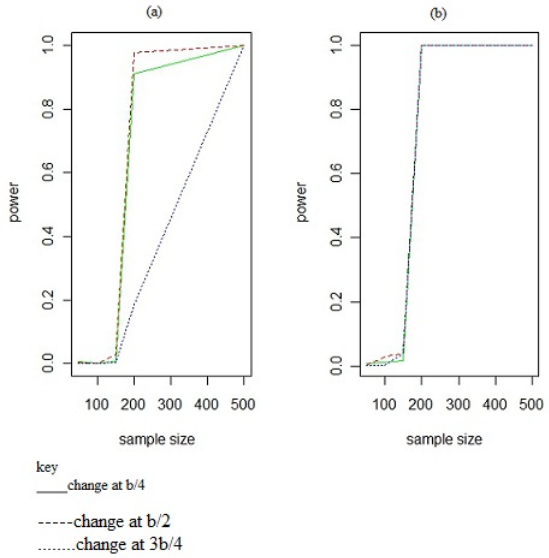

Figure 4.2: A plot of the power of the test against the size of the sample at $\alpha = 0.01$.

As Figure 4.2 shows the loss of power is more due to the size of the sample rather than the location of the change point. This is of importance since it would be desirable to detect a change once it occurs. The values in Figure 4.2(a) were evaluated when the conditional probabilities were estimated using a parametric method while those in Figure 4.2(b) the probabilities were estimated using a neural network.

4.2.3 Change of the power with size of the change

500 further simulations were carried out to investigate the power of the test for a sample size of 200 in relation to the size of the change, denoted as $\triangle$ where,

$$\triangle^2 = \| \theta - \theta^* \|^2 \tag{4.5}$$

θ and θ^* are the parameter values before and after the change point respectively. To compute the power of the test the the critical values, C1 are used. The results are presented in the Table 4.6.

A plot of the power of the test against the location of the change point at $\alpha = 0.01$ for the changes of size 1.2, 1.5 and 1.8 is presented in Figure 4.3.

Figure 4.3 shows that as the size of the change increases the more the chance

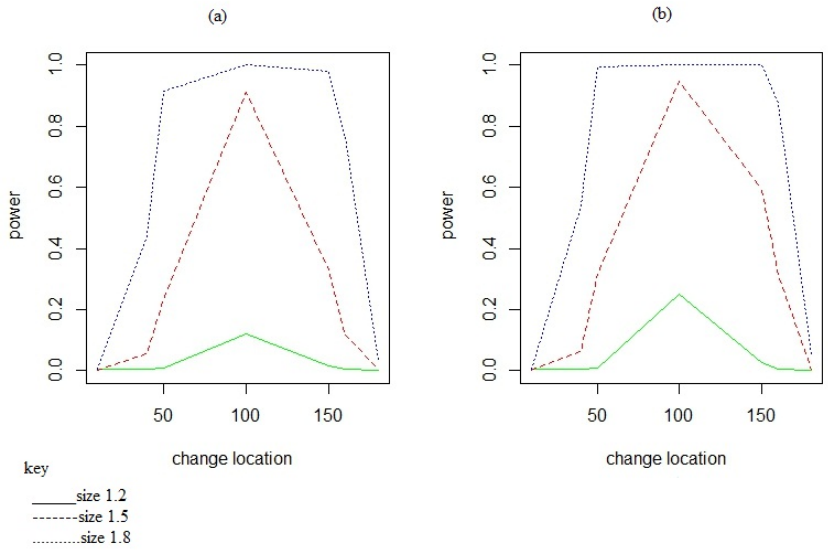

Figure 4.3: A plot of the power of the test against the location of the change point at $\alpha = 0.01$ for the changes of size 1.2, 1.5 and 1.8 .

of detecting it. The values in Figure 4.3(a) were evaluated when the conditional probabilities were estimated using a parametric method while those in Figure 4.3(b) the probabilities were estimated using a neural. It is noted that in all the instances the neural network performs better than the parametric method.

	$\hat{\kappa}(\alpha)$ under C1			
		size of change		
k	α	$\triangle = 1.2$	$\triangle = 1.5$	$\triangle = 1.8$
20	0.01	0.003992016	0.001996008	0.003992
	0.01*	0.00487432	0.00219754	0.00538710
	0.05	0.06387226	0.4530938	0.8323
	0.1	0.8023952	1	1
40	0.01	0.003992016	0.05588822	0.4411
	0.01*	0.00473981	0.06429013	0.53961
	0.05	0.8163673	1	1
	0.1	1	1	1
50	0.01	0.007984032	0.2315369	0.9142
	0.01*	0.00842108	0.3154287	0.99412764
	0.05	0.9520958	1	1
	0.1	1	1	1
100	0.01	0.1197605	0.9121756	1
	0.01*	0.251964	0.9458210	1
	0.05	1	1	1
	0.1	1	1	1
150	0.01	0.01796407	0.3313373	0.9800
	0.01*	0.027210945	0.59430631	1
	0.05	0.9820359	1	1
	0.1	1	1	1
160	0.01	0.003992016	0.1157685	0.7625
	0.01*	0.00492373	0.3154287	0.8764
	0.05	0.8582834	1	1
	0.1	1	1	1
180	0.01	0.001996008	0.003992016	0.02994
	0.01*	0.00284714	0.00572859	0.0529173
	0.05	0.1077844	0.499002	0.9661
	0.1	0.8622754	0.998004	1

Table 4.6: Power of the likelihood ratio test for different sizes of change and change point locations k.

4.2.4 Application to real data

To demonstrate the use of artificial neural networks in the estimation of the conditional means the Bliss (1935) beetles data is used, where batches of adult beetles were exposed to gaseous carbon disulphide for five hours. This data has been extensively used by statisticians in studies of generalized link functions e.g., Prentice and Ross (1976), Stukel (1988) and is used by Spiegelhalter et al. (1996) to demonstrate how *BUGS* handles generalized linear models for binomial data. The data is given in the Table 4.7

Dosage ($CS_2 mg/litre$)	Beetles	Killed
49.057	59	6
52.991	60	13
56.911	62	18
60.842	56	28
64.759	63	52
68.691	59	53
72.611	62	61
76.542	60	60

Table 4.7: Beetles Data

Here the assumption is that $p_i(x) = \beta_0 + \beta_1 X_{1i}$ where $p_i(x)$ is the probability of death due to the i^{th} dose and X_{1i} is the respective dose. Then as in equation (2.43) , the values of $P(m_i|X_1)$ may be estimated.

The dosage at which 50% of the beetles are killed is called the LD50. One may be interested in the determination of this dosage since it indicates a significant change in the structure of the probability of death. From the data the fourth dosage of 60.842 $CS_2 mg/litre$ kills 50% of the beetles. This shows that there

might be a change in the functional structure of probability at the fourth dosage. A comparision of the estimates of conditional means obtained through the parametric method using a generalized linear fit with logit as the link function and those obtained using the neural network is also done. The results are presented in Table 4.8.

Actual probabilities=$\frac{m_i}{n_i}$	Estimates fitted using glm	Estimates fitted using nnet
0.1016949	0.07011985	0.1189710
0.216667	0.16732799	0.1801028
0.2903226	0.34796279	0.3027226
0.500000	0.58696004	0.5230217
0.8253968	0.79040075	0.7834121
0.8983051	0.90945597	0.9394865
0.9838710	0.96386496	0.987661
1.00000	0.98611696	0.9977424

Table 4.8: Estimated probabilities of death

A graph of the estimated probabilities against the dose is given in Figure 4.4. It is evident from this graph that the estimates obtained using neural networks are nearer the actual values than those obtained through the generalized link function.

The probability of death LD50 is 0.5. A horizontal line through this point indicates that the fourth dosage is the LD50 and that the neural network method estimate is nearer the actual dosage than the generalized link function method estimate.

Taking the estimated probabilities from the data as the actual probabilities the mean square error was computed. The generalized link function method had an m.s.e. of 0.002032222 while the neural network estimates had an m.s.e. of

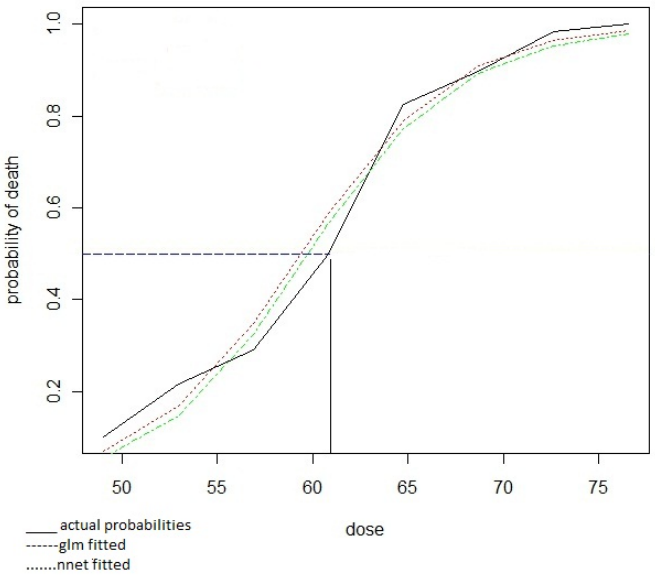

Figure 4.4: A plot of the estimated probabilities of death against dosage

0.0007246531. Thus in terms of m.s.e. the neural network estimates are better than the generalized link function method estimates.

4.3 Change Point Estimation

If change exists then the next problem will be to locate the point at which this change occurs. Hence the assumption in this section is that H_o is rejected and H_a is true . Thus a change exists at a certain point in the data.

For simulation purposes under H_a, the following model was used

$$P(m_i|X_i = x) = \begin{cases} (1 + exp(-(-1.5 + x_{1i} + x_{2i})))^{-1} & 1 \leq i \leq k \\ (1 + exp(-(-1.5 + 2 * x_{1i} + 1.8 * x_{2i})))^{-1} & k + 1 \leq i \leq b \end{cases} \tag{4.6}$$

The change point k was at fixed half way through the data i.e. at $\frac{b}{2}$ for a sample of size $b = 200$. x_{1i} and x_{2i} were generated as *uniform* [0,1]. n_i were generated

as the integer part of *uniform* [2, b]. Then the binomial random variable m_i was generated in line with equation (4.6). A simulation was done when change point was fixed half way through the data with the aim of testing whether change existed. A plot of the test statistic is presented in figure 4.5.

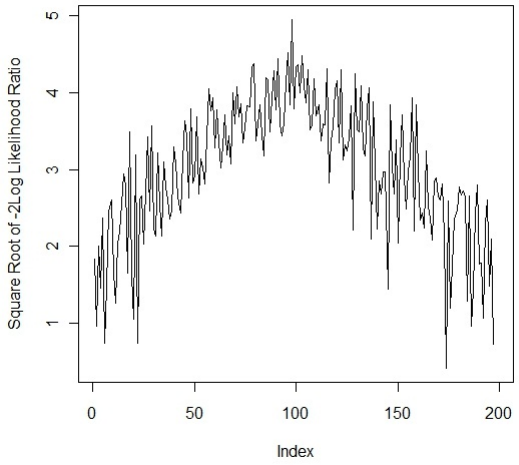

Figure 4.5: hypothesis testing graph when the alternative is true

Using critical values in Table 3.1 with a sample size of 200, it is noted that the hypothesis of change is accepted at 5%.

A further simulation is carried out to test for a change when it was actually not present. Thus one had the same parameters in equation (4.6). A plot of the test statistic is represented in Figure 4.6.

It is evident that the null hypothesis is not rejected at all the three levels of significance. Thus no change is detected.

A further 1000 simulations were carried out with the change point fixed half way through the data. The aim is to estimate the location of the change point as in equation(3.124). A plot of the loglikelihood against the estimated change point for one of the simulations is given in Figure 4.7.

From this graph the maximum of the loglikelihood is near the actual change point

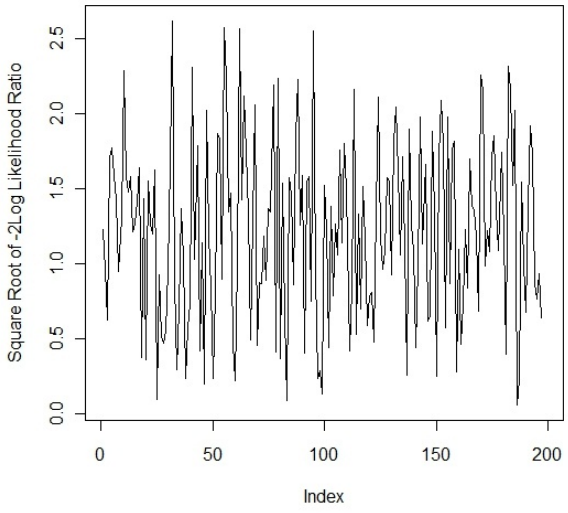

Figure 4.6: hypothesis testing graph when the alternative is false

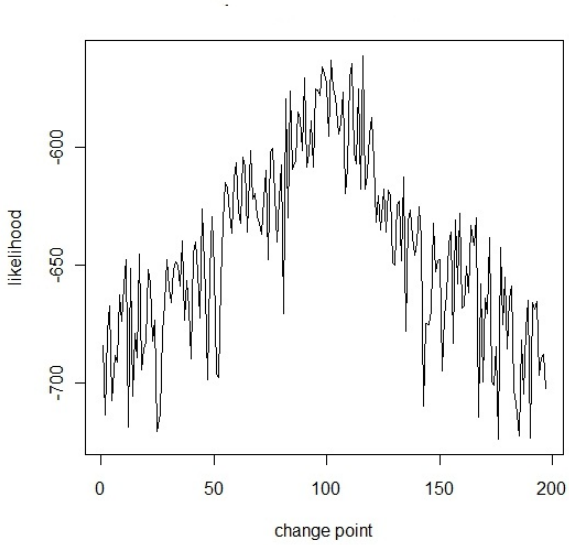

Figure 4.7: loglikelihood graph

of the data, halfway through the data. A histogram of the likelihood estimates of the change points presented in Figure 4.8.

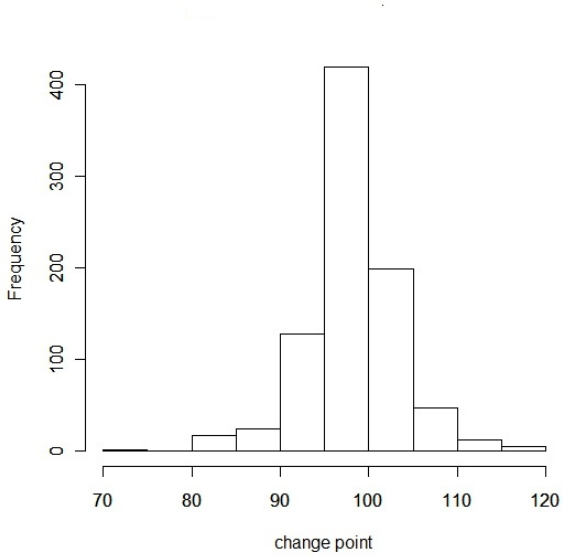

Figure 4.8: Histogram of likelihood estimates of change point

It is evident that most of the change point estimates are near the actual change point. The distribution of these estimates is not symmetric about the change point with more estimates being more to the left of the change point. Thus the unknown distribution of the change point estimates may be positively skewed.
When there is no change in the parameters, the histogram of estimates is presented in Figure 4.9.
The figure shows that the estimates are not concentrated around any point. The estimates seem to be uniformly distributed in the interval [0,200].
The asymptotic properties of the change point estimator are considered. Through simulation the estimate's asymptotic biasedness of the estimates is investigated. Figure 4.10 shows the histogram of the biases of the change point estimate.

The biases have an approximate mean of 0. Thus the change point estimates are

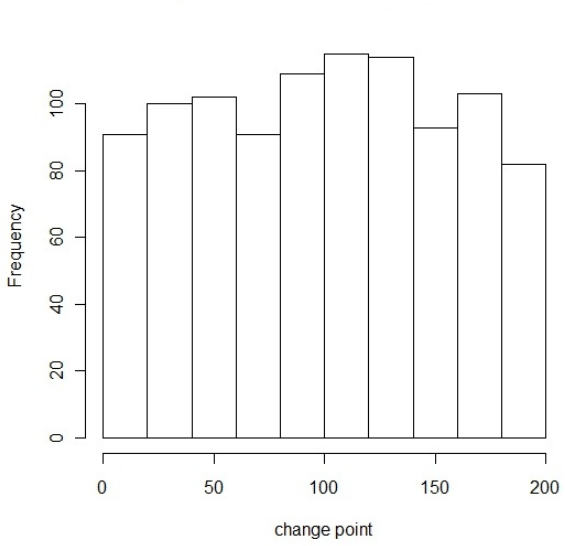

Figure 4.9: Histogram of likelihood estimates of change point when there is no change

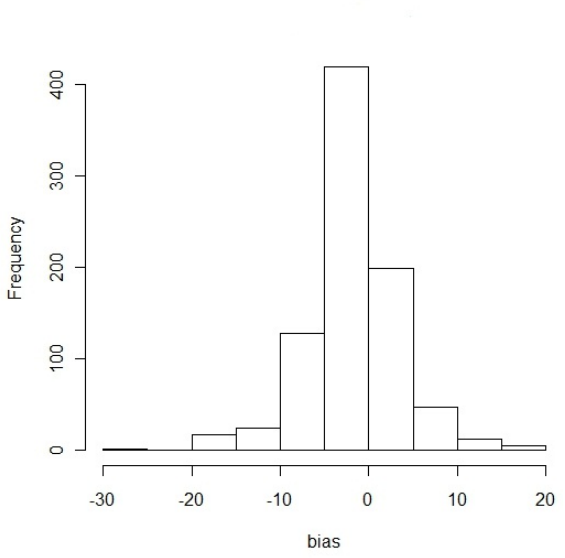

Figure 4.10: histogram of the biases of the change point estimates

asmptotically unbiased as expected since mle are asymptotically unbiased.

To evaluate the goodness of a fit of the normal curve those mean and variance are those of the biases on the same axes with the histogram of the change point estimates is drawn. This is presented in Figure 4.11

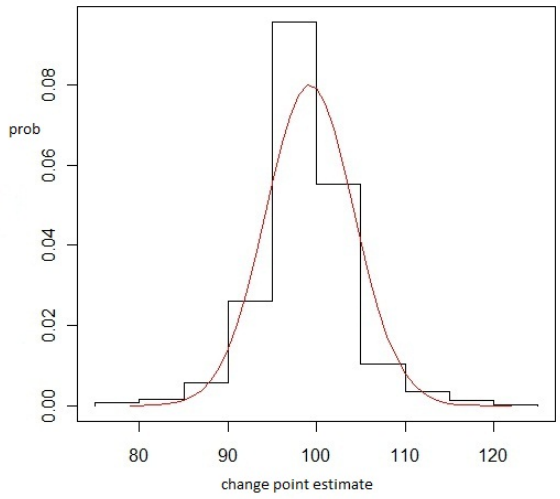

Figure 4.11: normal curve and histogram together

The figure suggests asymptotic normality of the change point estimator. This is in line with the central limit theorem and the law of large numbers.

Further a quantile-quantile plot in Figure 4.12 confirms the normality of the change point estimates.

The Kolmogorov-Smirnov test is performed on the biases of the change point. The test gave a *p-value* of 0.01121. Thus the null hypothesis of normality is accepted at 1%. For samples of sizes 300 and 500, the *p-values* of the same test are 0.01489 and 0.2106 respectively. This shows that the bias of the estimator is asymptotically normally distributed with a mean of zero.

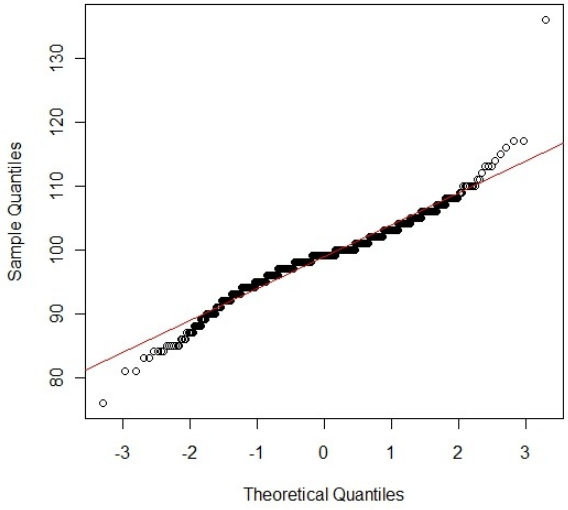

Figure 4.12: qqplot of the change-point estimates

4.3.1 Real Data Analysis

To demonstrate the use of non-parametric methods in this context, the Bliss (1935) beetles data is used, where batches of adult beetles were exposed to gaseous carbon disulphide for five hours. The data is given Table 4.7. Here it is assumed that probability of death $p_i(m_i|X_1) = \beta_0 + \beta_1 X_{1i}$ where m_i is the number of deaths due to the i^{th} dose and X_{1i} is the respective dose. One may wish to determine the point at which the functional form of probability of death of the beetles changes significantly. Using Theorems 3.8 or 3.8, the the critical values for a sample size as 481 were generated and are presented table 3.1. The computed the test statistic as in equation (3.124). The graph of the test statistic is presented in Figure 4.13.

The maximum value of this test statistic of 15.44364 which leads to the rejection the null hypothesis of no change. Figure 4.14 gives the plot of the loglikelihood against the dosage.

The maximum of the curve corresponds to the third change point location the

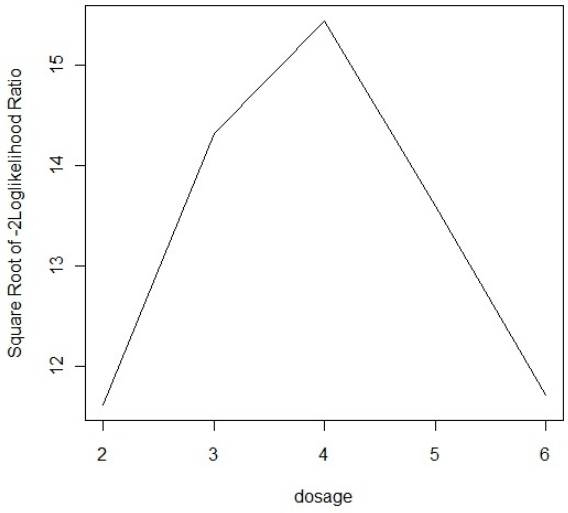

Figure 4.13: hypothesis testing graph for the bliss data

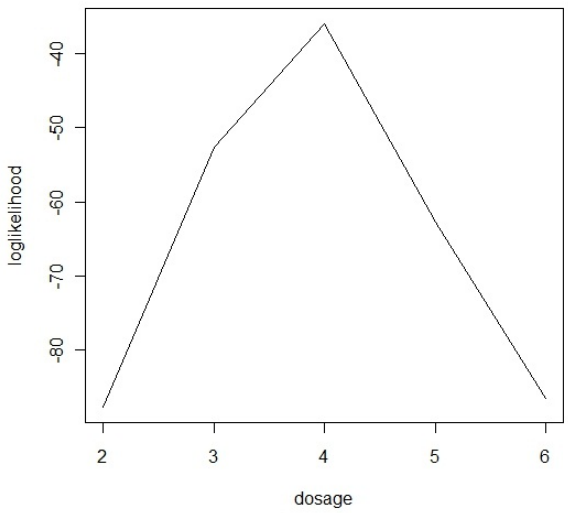

Figure 4.14: loglikelihood graph for the bliss data

fourth dosage. From the data the fourth dosage of 60.842 $CS_2 mg/litre$ kills 50% of the beetles, which is the LD50.

4.3.2 Confidence Interval of Change Point Estimates

The sample size was fixed at 100 and the change point midway through the data. Following the procedure of the bootstrap percentile confidence interval 1000 samples are replicated of which the change point was estimated. For simulation purposes the following model considered.

$$P(m_i|X_i = x) = \begin{cases} (1 + exp(-(-1.5 + x_{1i} + x_{2i})))^{-1} & 1 \leq i \leq k \\ (1 + exp(-(-1.5 + 2 * x_{1i} + 1.8 * x_{2i})))^{-1} & k + 1 \leq i \leq b \end{cases} \tag{4.7}$$

A histogram of the bootstrap change point is represented in Figure 4.15

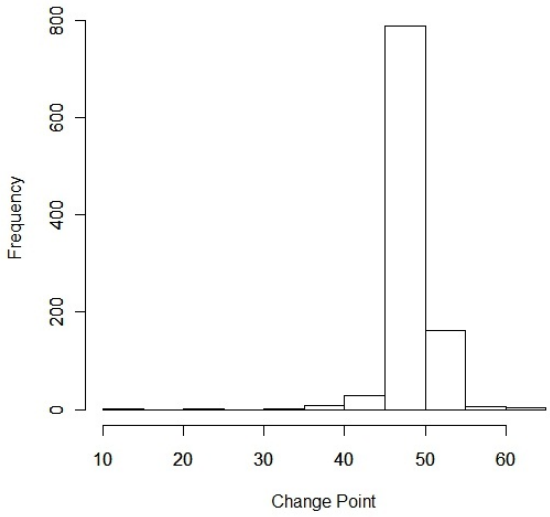

Figure 4.15: A histogram of 1000 bootstrap replicates of change point

4.3.2.1 Coverage Performance

In this section the $1-\alpha$ confidence interval $(\widehat{LCL}, \widehat{UCL})$ which is expected to have a probability α of miss-coverage of the true value of k from below or above is considered. This may be presented as

$$P(k < \widehat{LCL}) = P(k > \widehat{UCL}) = \frac{\alpha}{2} \tag{4.8}$$

A good confidence interval is the one which approximately matches the above equation. The confidence interval and the coverage performance of the above simulation is presented Table 4.9 It is noted again that the distribution of the

Confidence Level(%)	Confidence Interval	%miss-left	%miss-right
90	46-52	4	4.6
95	44-53	1.6	2.2
99	36-59	0.4	0.4

Table 4.9: Confidence Interval results for 1000 bootstrap samples

change point estimates is not symetrical.

CHAPTER FIVE

SUMMARY AND RECOMMENDATIONS

5.1 Summary of findings

This work sought to test for a change in binomial random variable and thereafter estimate the location of the change.

The problem at hand was as follows:-

There b groups and the size i^{th} is n_i. The probabilty of m_i sucesses in the i^{th} group was dependent on some covariates $X_i = (x_i, \ldots, x_D)^t$. Thus the conditional probabilities $p_i(m_i|X_i)$ were related to these coariates. Using logistic regression and a feed-forward single layer network the estimated probabilities in both the simulated and real data are computed. In the real data analysis it is found that the estimates from the neural network were better estimates in terms of m.s.e. than the estimates obtained using generalised linear model.

To test for change the likelihood ratio procedure is used. The null distribution of this statistic is derived the using an approach similar to Gombay and Horvath (1996). The critical values of the test then computed using this distribution.

The power of this test was investigated. The test was found to be powerful when the change was located near the center of the data and losses power when the change point is at the edges of the data. The asypmtotic power of this test was found to be unity implying that the test is consistent. The test was also found to be dependent on the size of the change with the power increasing as the size of the change. As one would expect the power of the test increases as the sample size increases.

The asymptotic properties of the change point estimator was also investigated. Using the approach of Franke and Neumann (2000) the estimator was found to

be asymptotically normally distributed. The change point estimator was also asymptotically consistent. This was further confirmed by the analysis of the simulated data.

Using the Bliss (1935) beetles data, change is tested for and found that change does exist. We estimated the change point and it corresponded with the dosage that killed 50% of the bettles.

5.2 Recommendations for Further Research

This work is a stepping stone for future research in this area. The work considered a sequence of binomial random variables with a single chage point. Of interest would be a situation where multiple change points exists. One would need to test for these changes and estimate the location of these changes. Random variables with other distributional forms may also be considered.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on automatic control*, AC-19(46):716–723.

Andrews, D. (1992). Generic uniform convergence. *Econometric Theory*, 8(2):241–257.

Bhattacharya, G. and Johnson, R. (1973). Non parametric tests for shifts at an unknown time point. *Annals of Mathematical Statistics*, (39):1731–1743.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22:134–167.

Broadslay, B. and Darkhovsky, B. (1993). *Nonparametric Methods in change point Problems*. Kluwer Academic Publishers.

Brown, R., Durbin, J., and Evans, J. (1975). Techniques for testing the constancy of regression relationships with time (with discussion). *Journal of the Royal Statistical Society*, B(37):149–192.

Broyden, C. G. ((1969/1970)). A new method of solving nonlinear simultaneous equations. *Comput. J.*, 12:9499.

Chao-Ying, J. and Gary, M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1):3–5.

Chen, J. and Gupta, A. (1997). Testing and locating variance changes with application to stock prices. *Journal of American Statistical Association*, (92):730–747.

Chen, J. and Gupta, A. (1998a). Information theoretic approach for detecting change in the parametric normal model. *Department of Mathematics and Statistics, Bowling Green State University,Technical report No.98-01.*

Chen, J. and Gupta, A. (1998b). Information theoretic approach for detecting change in the parametric normal model. *Department of Mathematics and Statistics, Bowling Green State University,Technical report No.98-05.*

Chen, J. and Gupta, A. (1999). Change point analysis of a gaussian model. *Statistical Papers*, (40):323–333.

Chen, J. and Gupta, A. (2000). *Parametric Statistical Change point Analysis.* Birkhuser, Boston.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of normal distribution which is subject to changes in time. *Annual of Mathematical Statistics*, (35):999–1018.

Chim Choy, J. and Broemeling, L. (1980). Some bayesian inferences for a changing linear model. *Technometrics*, (22):71–78.

Collett, D. (2002). *Modelling Binary Data.* Chapman and Hall.

Ferreira, P. (1980). A bayesian analysis of a switching regression model. *Journal of American Statistical Association*, (70):370–374.

Fletcher, R. (1970). A class of methods for nonlinear programming with termination and convergence properties in integer and nonlinear programming. *North-Holland, Amsterdam*, page 157175.

Franke, J. and Neumann, M. (2000). Bootstrapping neural networks. *Neural Computation*, 12(12):1929–1949.

Fu, Y. and Curnow, R. (1990). Maximum likelihood estimate of multiple change points. *Biometrika*, (77):563–573.

Gardener, L. (1969). On detecting change in the mean of normal variates. *Annals of Mathematical Statistics*, (40):116–126.

Gichuhi, A. W., Franke, J., and Kihoro, J. (2012). Parametric change point estimation, testing and confidence interval applied in business. *JAGST*, 14(2):136–148.

Girshick, M. and Rubin, H. (1952). A baye's approach to a quality control model. *Annals of Mathematical Statistics*, (23):114–125.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Math. Comp.*, 24:2326.

Gombay, E. and Horvath, L. (1994). An application of the maximum likelihood test to change point problem. *Stochastic Process. Appl.*, (50):161–171.

Gombay, E. and Horvath, L. (1996). On the rate of approximation for maximum likelihood tests in change-points models. *Journal of Multivariate Analysis*, (56):120–152.

Gupta, A. and Chen, J. (1996). Detecting changes of mean in multidimensional normal sequences with application to literature and geology. *Computational Statistics*, (4):211–221.

Gupta, A. and Ramanayake, A. (1998). Change point with linear trend for exponential distribution. *Department of Mathematics and Statistics,Bowling Green State University,Technical Report*, (98-14).

Haccou, P. and Meelis, E. (1988). Testing for the number of change points in a sequence of exponential random variables. *Journal of Statistical Computation and Simulation*, (30):285–298.

Haccou, P., Meelis, E., and Geer, S. (1988). The likelihood ratio test for the change point problem for exponentially distributed random variables. *Stochastic Processes and their Applications*, (27):121–139.

Hanify, J., Metcalf, P., Nobbs, C., and Worsley, K. (1981). Aerial spraying of 2,4,5-t and human birth malformations: An epidemilogical investigation. *Science*, (212):349–351.

Hinkley, D. and Hinkley, E. (1970). Inference about change point in a sequence of binomial variable. *Biometrika*, (57):477–488.

Hobert, D. (1982). A bayesian analysis of a switching linear model. *Journal of econometrics*, (19):77–87.

Hsu, D. (1977). Tests for variance shifts at unknown time point. *Applied Statistics*, (26):179–184.

Hsu, D. (1979). Detecting shifts in parameters in gamma sequences with applications to stock price and air traffic flow analysis. *Journal of the American Statistical Association*, (74):31–40.

Hwang, J. and Ding, A. (1997). Prediction intervals for artificial neural network. *Journal of the American Statistical Association*, 92(438):748–757.

Incl*á*n, C. (1993). Detection of multiple changes of variance using posterior odds. *Journal of Business and Economic Statistics*, (11):189–300.

James, B., James, K., and Siegmund, D. (1992). Asymptotic approximations for likelihood ratio tests and confidence regions for a change point in the mean of a multivariate normal distribution. *Statistical Sinica*, (2):69–90.

Jaruskova, D. (1997). Some problems with application of change point detection methods to enviromental data. *Envirometrics*, 8:469–483.

Judd, K. (1998). Numerical methods in economics. *Cambridge*, MA:MIT Press.

Kander, Z. and Zacks, S. (1966). Test procedures possible changes point in parameters of statistical distributions occurring at unknown points. *Annals of Mathematical Statistics*, (37):1196–1210.

Kauna, C. and Halbert, W. (1994). Artificial neural networks, an econometric perspective. *Econometric Review*, (13):1–91.

Kim, D. (1994). Tests for change point in linear regression. *IMS Lecture Notes, monograph series*, (23):170–176.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimisation by simulated annealing. *Science*, 220:671–680.

Kolmogrov, A. N., Prokhorov, Yu, V., and Shiryaev, A. (1988). Probabilistic-statistical methods for detection of spontaneous effects. *Maths Institute of the Academy of Sciences of the USSR.*, 182:4–23.

Krishnaiah, P., Miao, B., and Zhao, L. (1990). Local likelihoods method in problems related to change points,. *Chinese Annal of Mathematics*, 11B(3):363–375.

McNelis, P. (2005a). *Neural Networks in Finance: Gaining Predictive Edge in the Market*. Elsevier Academic Press.

McNelis, P. (2005b). *Neural Networks in Finance Gaining Predictive Edge in the Market*. Elsevier Academic Press .

Metropolis, N., Rusenbluth, M., Rusenbluth, A., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics.*, 21:1087–1092.

Michalewicz, Z. (1996). *Genetic Algorithms + Data Structures = Evolution Programs*. New York: Springer-Verlag.

Miller, W., Thomas, I., Richard, S., and Paul, J. (1990). Neural networks for control. *Cambridge*, MA:MIT Press.

Page, E. (1954). Continuous inspection schemes. *Biometrika*, (1):100–115.

Page, E. (1955). A test for change in a parameter occurring at unknown point. *Biometrika*, (42):523–527.

Page, E. (1957). On problem in which a change in parameter occurs at an unknown points. *Biometrika*, (442):248–252.

Pettitt, A. N. (1954). A simple cumulative sum type statistic for change point problem with zero-one observations. *Biometrika*, (67):79–84.

Prentice, A. and Ross, L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics*, 32:134–167.

Quandt, R. (1958). The estimation of parameters of a linear regression system that obeys two separate regimes. *Journal of American Statistical Association*, (53):73–88.

Quandt, R. (1960). Tests of hypothesis that a linear regression system obeys two separate regimes. *Journal of American Statistical Association*, (55):324–330.

Ruhkin, A. and Gary, M. (1995). Asymptotic behavior of estimators of change-point in binomial probability. *Applied Statistical science*, 2(1):1–12.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, (6):461–464.

Sen, A. and Srivastava, M. (1973). On multivariate tests for detecting change in mean. *Sankhya*, A(35):173–186.

Sen, A. and Srivastava, M. (1975a). On tests for detecting change in mean. *Annals of Statistics*, (3):98–108.

Sen, A. and Srivastava, M. (1975b). Some one-sided tests in change in level. *Technometrics*, (17):61–64.

Shanno, D. (1970). Conditioning of quasi-newton methods for function minimization. *Math. Comp.*, 24:647656.

Smith, A. (1975). A bayesian approach to inference about change point in a sequence of random variables. *Biometrika*, (62):407–416.

Spiegelhalter, D. J., Thomas, A., Best, N., and Gilks, W. R. (1996). Bugs: Bayesian inference using gibbs sampling, version 0.5 (version ii). *Cambridge, UK:Biostatistics Unit.*

Srivastava, M. and Worsley, K. (1986). Likelihood ratio test for a change in multivariate normal mean. *Journal of American Statistics Association*, (81):199–204.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, 83:426–431.

Swanson, N. R. and White, H. (1995). 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):265275.

Vlachonikalis, I. and Vasdekis, V. (1994). On a class of change point models incovariance structures for growth curves and repeated measurements. *Communication in Statistics- Theory and Methods*, (23):1087–1102.

Vostrikova, L. (1978). Dectecting " disorder" in multidimensional random processes. *Soviet Mathematics Doklady*, (25):55–59.

Waititu, A. (2008). *Nonparametric change point analysis for Bernoulli random variables based on neural networks*. PhD thesis.

Werbos, P. J. (1994). The roots of backpropagation: From ordered derivatives to neural networks and political forecasting. *New York: Wiley Interscience*.

Wichern, D. W., Miller, R., and Hsu, D. (1976). Changes in first order autoregressive time series models with an application. *Applied Statistics*, (25):248–356.

Worsley, K. (1979). On the likelihood ratio test for a shift in location of normal populations. *Applied Statistics*, (4):365–367.

Worsley, K. (1983). The power of the likelihood ratio and cumulative sum tests for change in binomial probability. *Biometrika*, (70):455–464.

Worsley, K. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, 1(73):91–104.

Yao, Q. (1993). Tests for change points with epidemic alternatives. *Biometrika*, (80):179–191.

Zhao, L., Krisknaiah, P., and Bai, Z. (1986a). On detection of the number of signals in the presence of the whitenoise. *Journal of Multivariate Analysis*, (20):1–25.

Zhao, L., Krisknaiah, P., and Bai, Z. (1986b). On detection of the number of signals in the presence of the whitenoise. *Journal of Multivariate Analysis*, (20):26–49.