SOME ASPECTS ON DATA MODELLING

XIAOYING SUN

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

April 2017

Abstract

Statistical methods are motivated by the desire of learning from data. Transaction dataset and time-ordered data sequence are commonly found in many research areas, such as finance, bioinformatics and text mining. In this dissertation, two problems regarding these two types of data: association rule mining from transaction data and structural change estimation in time-ordered sequence, are studied.

Informative association rule mining is fundamental for knowledge discovery from transaction data, for which brute-force search algorithms, e.g., the well-known Apriori algorithm, were developed. However, operating these algorithms becomes computationally intractable in searching large rule space. A stochastic search framework is developed to tackle this challenge by imposing a probability distribution on the association rule space and using the idea of annealing Gibbs sampling. Large rule space of exponential order can still be randomly searched by this algorithm to generate a Markov chain of viable length. This chain contains the most informative rules with probability one. The stochastic search algorithm is flexible to incorporate any measure of interest. Moreover, it reduces computational complexities and large memory requirements.

A time-ordered data sequence may contain some sudden changes at some time points, before and after which the data sequences follow different distributions or statistical models. Change point problems in generalized linear models and distributions of independent random variables are studied respectively. Firstly, to estimate multiple change points in generalized linear models, we convert it into a model selection problem. Then modern model selection techniques are applied to estimate the regression coefficients. A consistent estimator of the number of change points is developed, and an algorithm is provided to estimate the change points. Secondly, to estimate single change point in distributions of independent random variables, a change point estimator is proposed based on empirical characteristic functions. Its consistency is also established.

Keywords: association rule, Gibbs sampling, transaction data, genomic data, multiple change points, GLM, SIS, MCP, segmentation, change point estimator, empirical characteristic function

Acknowledgements

First and foremost, I would like to express my greatest appreciation to my supervisor, Professor Yuehua Wu. Without her valuable guidance, constant encouragement and extensive knowledge, this dissertation would not be possible. I truly respect her brilliant insights and enthusiasm on research.

I would like to extend my sincere appreciation to Professor Huaiping Zhu, Professor Xin Gao and Professor Steven Wang as members of my supervisory committee. My appreciation also goes to all faculty members, staffs and fellow graduate students in the Department of Mathematics and Statistics at York. I also would like to thank Professor Guoqi Qian, Professor Changchun Tan and Professor Xiaoping Shi for their tremendous suggestions on my dissertation.

Last but not least, I deeply thank my parents for their continuous support. A special thank goes to my beloved husband for his patience, kindness and wisdom. This dissertation is dedicated to my lovely daughter who brings me happiness everyday.

Table of Contents

List of Tables

List of Figures

1 Introduction and Notations

Statistical methods are motivated by the desire of learning from data. As the development of computer science and the advent of the information age, data generated in many fields have exploded both in size and complexity of the structure which challenges the field of Statistics and leads to a revolution in the statistical science [Hastie et al., 2009]. Transaction data and time-ordered data sequence are commonly found in many research areas, such as finance, bioinformatics and text mining. Transaction dataset was originally found from market basket analysis. A market basket dataset contains a large collection of items. Each transaction is a basket of items that a customer purchased. Many other types of data can be converted into a transaction dataset. For instance, text data can be converted to a transaction dataset in which each word is an item and each sentence is a basket of items. Time-ordered data sequence is a set of observations on single or multiple random variables over time. For instance, a dataset containing the hourly counts of rental bikes recorded in the bike sharing system from 2011 to 2012 is a time-ordered data sequence. The annual flow of the river Nile at Aswan from 1871 to 1970 is another example. In this dissertation, two problems regarding these two types of data: association rule mining from transaction data and structural change estimation in time-ordered sequence, are studied. These two problems are formally introduced in the next two sections.

1.1 Association Rule Mining from Transaction Dataset

Association rule mining [Agrawal et al., 1993 and Agrawal et al., 1994] in many research areas such as marketing, politics, and bioinformatics is an important task. One of its well-known applications is the market basket analysis. An example of association rule from the basket data might be that "90% of all customers who buy bread and butter also buy milk "[Agrawal et al., 1993], providing important information for the supermarkets management of stocking and shelving. Instead of mining all association rules from a database, an interesting and useful task is to discover the most significant association rules for a given consequent. For a genomic dataset, one might be interested in finding which SNP (single nucleotide polymorphism at certain loci in a gene) variables and their values imply a certain disease. The objective is to identify the most significant association rules for a given item from a transaction dataset according to a given measure for rules.

Constraint-based search is mostly used in current algorithms to mine association rules. For instance, the Apriori algorithm [Agrawal et al., 1993] mines all rules satisfying a user-specified minimum support or minimum confidence, and maximum length. It is difficult to use such an algorithm in a sparse dataset with a large number of items because it either searches through too many rules being computationally infeasible if the constraint is low, or misses the important ones otherwise. Some rule-mining algorithms use well-defined metrics to identify the most significant association rules [Bayardo and Agrawal, 1999]. But, they also use deterministic and exhaustive search, consequently becoming computationally intractable when applied to a large dataset with, say, a few hundred items in the item space.

To formally study the problem, we introduce two types of notations, the set notations and the binary variable notations.

1. Notations using sets:

    - *Item space*: $I = \{I_1, I_2, \ldots, I_m, I_c\}$ and $I_{-c} = \{I_1, I_2, \ldots, I_m\}$. Here, I_c is a given item as the consequent of association rules and I_{-c} is a set of items that could appear in the antecedent of association rules.
    - List of *transactions*: $D = \{t_1, t_2, \ldots, t_n\}$ with $t_j \subset I$, $j = 1, \ldots, n$
    - *Itemset*: $B \subset t_j$ for some j's.
    - *Association rule*: $A \Rightarrow I_c$ with $A \subset I_{-c}$ where A and I_c are the antecedent and consequent of the rule respectively.

- *Support*: $\text{supp}(A) = \frac{|\{t_i \in D | A \subset t_i, i=1,\ldots,n\}|}{n}$, $\text{supp}(A \Rightarrow I_c) = \text{supp}(A \cup I_c)$. Here $|\{t_i \in D \mid A \subset t_i, i = 1, \ldots, n\}|$ denotes the size of the set $\{t_i \in D \mid A \subset t_i, i = 1, \ldots, n\}$.
- *Confidence*: $\text{conf}(A \Rightarrow I_c) = \dfrac{\text{supp}(A \cup I_c)}{\text{supp}(A)}$.

2. Binary variable notations:

- *Binary vector*: $V = (J_1, \ldots, J_m, J_c)$ and $\boldsymbol{J} = (J_1, \ldots, J_m)$ where

$$J_s = \begin{cases} 1, & \text{presence of item } I_s; \\ 0, & \text{absence of item } I_s,\ s = 1, 2, \ldots, m; \end{cases}$$

and

$$J_c = \begin{cases} 1, & \text{presence of item } I_c; \\ 0, & \text{absence of item } I_c. \end{cases}$$

- Each transaction t_i is an observation, $\boldsymbol{v}_i$ of the binary vector V.

In this dissertation, these two notations will be used interchangeably if no confusion will be caused. The collection of all possible association rules for I_c is

$$\mathcal{R}_{I_c} = \{\boldsymbol{J} \Rightarrow I_c \mid \boldsymbol{J} \in \{0,1\}^m \setminus \boldsymbol{0}\}$$

where $\boldsymbol{J}$ denotes a subset of I_{-c} corresponding to 1's in $\boldsymbol{J}$. The objective is to search in $\mathcal{R}_{I_c}$ for the most significant association rule according to a particular measure of association rules. In chapter 2, a random sampling framework is proposed to solve this problem.

1.2 Structural Changes Estimation

Change point (structural change) analysis is the process of detecting distributional changes within time-ordered observations [Matteson and James, 2014]. Applications can be found in many research areas including climate studies, medical and health sciences, financial econometrics and risk management. For instance, change point analysis is used to examine the North Atlantic tropical cyclone record for statistical discontinuities (change points) [Robbins et al., 2012], confirm the effect of the seat belt legislation on the monthly deaths and serious injuries, detect speech signals [Davis et al., 2006], and estimate change points in the 1982 Urakawa-Oki earthquake records [Jin et al., 2011] and temporal discontinuities in the cloud cover data [Lu and Wang, 2012].

Page [1954, 1955] first introduces the undocumented change point problem. Since then, change point problems have been intensively studied in the literature. The change point problems considered in the literature can roughly be categorized into two groups. One group is the change point detection in the distributions of independent random variables (or vectors). Csörgő and Horváth [1997] present methods to detect change points in means or variances of independent random variables. Hušková and Meintanis [2006] propose a nonparametric test statistic to detect a change point in the distributions of an independent univariate sequence. Robbins et al. [2012] develop a test statistic to detect a single change point in a categorical data sequence.

The other group of change point problems is to detect or estimate the change point before and after which the data sequences follow two different models. The single change point detection and estimation in the linear regression models is studied in Csörgő and Horváth [1997]. Antoch et al. [2004] propose a statistic to test structural change in a generalized linear model (GLM). Davis et al. [2006] and Jin et al. [2011] study the multiple structural break estimation and variable selection problem for nonstationary time series models. Lu and Wang [2012] develop a likelihood ratio test for detecting a sudden change in parameters of a cumulative logit model for a multinomial sequence. Jin et al. [2016] propose an algorithm to estimate multiple change points in the linear regression model.

In this dissertation, two change point problems are studied. One is the multiple change points estimation in a generalized linear model in Chapter 3. The other one is the change point estimation in distributions of independent observations in Chapter 4.

The following are some notations used in the Chapter 3 and Chapter 4. A^T denotes the transpose of a matrix A. $\boldsymbol{v}^T$, v_j and $\|\boldsymbol{v}\|$ denote the transpose, j^{th} component and the L_2 norm of a vector $\boldsymbol{v}$, respectively. Let $\boldsymbol{v} = (v_1, v_2, \ldots, v_p)^T$ be a $p \times 1$ vector, $A = (a_{ij}) = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)$ be a $q \times p$ matrix where a_{ij}'s are the elements of A and $\boldsymbol{a}_j$'s are the column vectors of A, and $\mathcal{B} = \{i_1, i_2, \ldots, i_k\}$ be an index set with $1 \leq i_1 \leq \ldots \leq \ i_k \leq p$. Let $|\mathcal{B}|$ denote the size of $\mathcal{B}$ which is equal to k. Denote $\boldsymbol{v}_{[\mathcal{B}]} = (v_{i_1}, \ldots, v_{i_k})^T$, $A_{[\mathcal{B}]} = (\boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_k})$. Let $I_S(t)$ be the indicator function such that $I_S(t) = 1$ if $t \in S$ and $I_S(t) = 0$ otherwise, $a_+ = a$ if $a > 0$ and $a_+ = 0$ otherwise. Denote the inverse function of $f(x)$ as $f^{-1}(x)$. Let $f'(x)$ and $f''(x)$ denote the first and second order derivatives of a univariate function, $f(x)$ with respect to the scalar x, and let $\partial f(\boldsymbol{v})/\partial \boldsymbol{v}$ and $\partial^2 f(\boldsymbol{v})/(\partial \boldsymbol{v} \partial \boldsymbol{v}^T)$ denote the first and second order derivative with respect to the vector $\boldsymbol{v}$. Define $\lfloor x \rfloor$ and $\lceil x \rceil$ as the largest integer smaller than or equal to x and the smallest integer larger than or equal to x respectively. "$\longrightarrow_P$" stands for the convergence in probability. "$\Rightarrow$" means the weak convergence. $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of a standard normal distribution.

2 Boosting Association Rule Mining in Large Transaction Datasets via Gibbs Sampling

In this chapter, a stochastic search framework is presented to mine the most significant association rules from a transaction dataset according to a given measure for rules without information loss. The motivation comes from a genomic dataset of a disease outcome variable and hundreds of SNP variables, and the desire to mine the most significant association rules for the disease outcome. Such dataset can be converted into a transaction dataset for association rule mining since both the response and the predictors are of categorical type. Here the response is a disease outcome having two categories, case (C) and noncase (NC). Each predictor is the so-called SNP variable having 3 categories corresponding to 0, 1, and 2 copy numbers of the minor allele at the loci. In this case, the response variable can be represented by one response item, I_c, and each predictor variable can be represented by three predictor items subject to the constraint that there must be only one of these three items appearing in the transaction. Suppose that the total number of items is m. Then let $I_{-c} = \{I_1, \ldots, I_m\}$ denote the set of predictor items that could appear in the antecedent of rules for I_c.

By the notations introduced in Chapter 1, the collection of all possible association rules for I_c is

$$\mathcal{R}_{I_c} = \{\boldsymbol{J} \Rightarrow I_c \mid \boldsymbol{J} \in \{0,1\}^m \setminus \mathbf{0}\}$$

where $\boldsymbol{J}$ denotes a subset of I_a corresponding to 1's in $\boldsymbol{J}$. The objective is to search in $\mathcal{R}_{I_c}$ for the most significant association rule according to a particular measure of association rules. The following property clearly holds for this transaction dataset:

Property: $0 \leq \text{supp}(\boldsymbol{J} \Rightarrow I_c) \leq \text{conf}(\boldsymbol{J} \Rightarrow I_c) \leq 1$.

Our interest is to find association rules with high confidence and high support. A constraint-based algorithm like the Apriori cuts the rule space into a smaller one by setting up abrupt constraints including minimum support, minimum confidence and maximum length of rules so that the algorithm is feasible. The constraints are very subjective and the algorithm is still computationally challenging when the item space is too large. It is even more difficult when the rules with high confidence have very low support. An example given in [Hämäläinen, 2009] is that the forestry society *FallAll* conducted association rules mining to a dataset of 1,000 observations on marsh sides for providing advice on draining swamps to grow new forests. The

Apriori algorithm was applied to this dataset by specifying the minimum support and confidence as 0.05 and 0.80, respectively. But, a strong association rule of confidence 1.0 and support 0.04 was missed with this set of constraints. In general, mining association rules in a dense dataset can miss important rules and get misinformed by noninformative rules produced due to improper constraints. Because the deterministic search algorithms are not able to cope with the computational intensity and immensity for this dataset, this motivates us to propose a stochastic sampling framework to overcome the difficulty.

2.1 A New Random Sampling Framework

The probability distribution for sampling and searching important association rules entails incorporating both support and confidence of the rules into the procedure. For this, we first define a new measure for association rules in $\mathcal{R}_{I_c}$ and call it the *importance*, which is of the form $g(\boldsymbol{J} \Rightarrow I_c) = f(\text{supp}(\boldsymbol{J} \Rightarrow I_c), \text{conf}(\boldsymbol{J} \Rightarrow I_c))$, for a given association rule $\boldsymbol{J} \Rightarrow I_c$. Here f is a user-specified positive increasing function reflecting certain combined importance of the support and confidence of the rule. Plausible choices of f are the minimum, summation, or product of the support and confidence. Once f is specified, our aim becomes finding the most significant association rules in $\mathcal{R}_{I_c}$ according to the measure $g(\cdot)$ which can be achieved by the following random-sampling-based search procedure.

In light of the non-Bayesian optimization idea of Qian and Field (2002), we propose a probability distribution defined on $\mathcal{R}_{I_c}$ as

$$p_c(\boldsymbol{J}) = P(\boldsymbol{J} \Rightarrow I_c) = \frac{e^{\xi g(\boldsymbol{J} \Rightarrow I_c)}}{\sum_{\tilde{\boldsymbol{J}} \in \{0,1\}^m \setminus \mathbf{0}} e^{\xi g(\tilde{\boldsymbol{J}} \Rightarrow I_c)}}, \tag{2.1}$$

for any $\boldsymbol{J} \in \{0,1\}^m \setminus \mathbf{0}$, where $\xi > 0$ is a tuning parameter. The most important rule in $\mathcal{R}_{I_c}$, denoted as $\boldsymbol{J}_{opt} \Rightarrow I_c$, is the one maximizing $p_c(\boldsymbol{J})$ over $\mathcal{R}_{I_c}$, i.e. $\boldsymbol{J}_{opt} = \arg\max_{\boldsymbol{J} \in \{0,1\}^m \setminus \mathbf{0}} p_c(\boldsymbol{J})$. This implies that $\boldsymbol{J}_{opt}$ can be found (with probability 1) from a random sample of $\boldsymbol{J}$'s generated from $p_c(\boldsymbol{J})$ if the sample size is sufficiently large. It can be proved that $\boldsymbol{J}_{opt}$ appears most frequently and has the largest value of $g(\boldsymbol{J} \Rightarrow I_c)$ in the sample with probability 1. However, generating a random sample from $p_c(\boldsymbol{J})$ is not trivial when m is not small, because the rule space $\mathcal{R}_{I_c}$ becomes huge and the normalizing denominator in $p_c(\boldsymbol{J})$ becomes intractable in evaluation. It turns out that the method of Gibbs sampling can be used to generate random samples from $p_c(\boldsymbol{J})$, where we need all conditional probability distributions of J_s given $\boldsymbol{J}_{-s}$:

$$\begin{aligned}
p_c(J_s = 1 | \boldsymbol{J}_{-s}) &= \frac{p_c(J_s = 1, \boldsymbol{J}_{-s})}{p_c(\boldsymbol{J}_{-s})} \\
&= \frac{p_c(J_s = 1, \boldsymbol{J}_{-s})}{p_c(J_s = 1, \boldsymbol{J}_{-s}) + p_c(J_s = 0, \boldsymbol{J}_{-s})}, \\
p_c(J_s = 0 | \boldsymbol{J}_{-s}) &= 1 - p_c(J_s = 1 | \boldsymbol{J}_{-s})
\end{aligned}$$

for $s = 1, 2, \ldots, m$. Here $\boldsymbol{J}_{-s}$ is the sub-vector of $\boldsymbol{J}$ with J_s removed and $(J_s, \boldsymbol{J}_{-s})$ is the vector with J_s being put back into its original position in $\boldsymbol{J}$.

Then the Gibbs sampling algorithm for generating $\boldsymbol{J}$'s from $p_c(\boldsymbol{J})$ is given as the following:

- Arbitrarily choose an initial vector $\boldsymbol{J}^{(0)} = (J_1^{(0)}, \ldots, J_m^{(0)})$;
- Repeating for $l = 1, 2, \ldots, L$, the antecedent of the rule, $\boldsymbol{J}^{(l)} \Rightarrow I_c$, is obtained by generating $J_s^{(l)}, s = 1, 2, \ldots, m$ sequentially from the Bernoulli distribution $p_c(J_s | J_1^{(l)}, \ldots, J_{s-1}^{(l)}, J_{s+1}^{(l-1)}, \ldots, J_m^{(l-1)})$;
- Return $(\boldsymbol{J}^{(1)}, \ldots, \boldsymbol{J}^{(L)})$ for the association rules sample $\{\boldsymbol{J}^{(l)} \Rightarrow I_c; l = 1, \cdots, L\}$.

The generated sequence $\{\boldsymbol{J}^{(1)}, \cdots, \boldsymbol{J}^{(L)}\}$ is actually a Markov chain with its stationary distribution being $p_c(\boldsymbol{J})$ and it can be shown that the most frequent rule occurring in the generated sample converges to $\boldsymbol{J}_{opt}$ with probability 1 as $L \to \infty$. Moreover, those most significant association rules in $\mathcal{R}_{I_c}$ are more likely to appear the most frequently in the generated sample than other less significant ones, provided that the sample size M is sufficiently large. In the cases that the measures of many significant association rules are large but very close to each other, choosing a larger value for the tuning parameter ξ increases the probability ratio of every two rules, $\frac{p_c(\boldsymbol{J}^{(1)})}{p_c(\boldsymbol{J}^{(2)})} = e^{\xi(g(\boldsymbol{J}^{(1)} \Rightarrow I_c) - g(\boldsymbol{J}^{(2)} \Rightarrow I_c))}$, which helps differentiate the more significant rules from the less significant ones.

We remark that the measure $g(\cdot)$ can be replaced by any other interesting measure of association rules such as lift and leverage [Hämäläinen, 2009]. Thus, a random sample can also be easily generated according to that interesting measure.

Once $\{\boldsymbol{J}^{(1)}, \cdots, \boldsymbol{J}^{(L)}\}$ is generated, the optimal association rules in $\mathcal{R}_{I_c}$, which have the highest probability, can be approximated by the association rules with the near-highest frequencies in the sample. The approximation precision can be achieved as high as one wants provided that the sample size is sufficiently large. Note that if the item space is very large, the generation of a long sample is computationally expensive. However, it is possible that in the random sample of a relatively small size L, the association rules could all be different from each other and each has the same frequency $1/L$. In this case, it is possible that none of the rules is optimal. Instead, we can compute the frequency for each item that ever appeared in the antecedents of the sampled rules. The frequency for item I_s is $\sum_{l=1}^{L} J_s^{(l)}/L$ for $s = 1, 2, \ldots, m$. We would obtain a subset of items that appear most frequently. Then we can apply the Apriori algorithm on the itemset space generated by the selected items to mine the optimal rules. Our simulation study shows that the random sample obtained by the Gibbs sampling method can largely reduce the itemset space for search and retain the most frequent predictor items from the optimal association rules simultaneously.

In the next section, we will elaborate how to use the generated sample of rules.

2.2 Simulation Study

In this section, we present several numerical examples based on simulated data to demonstrate the performance of the random-sampling-based search procedure in different scenarios.

A transaction dataset containing strong association rules can be obtained by using the R package *MultiOrd* [Amatya and Demirtas, 2015] to generate a list of binary vectors from a multivariate Bernoulli distribution of correlated binary random variables with a compatible pair of mean vector $\boldsymbol{p}$ and correlation matrix R [Chaganty and Joe, 2006]. We start with a small dataset to show that our method is able to find the optimal association rules.

Example 1. Suppose a small transaction dataset has $m = 3$ predictor items I_1, I_2, I_3 and one response items I_c. Also suppose that the marginal probability of vector (J_1, J_2, J_3, J_c) is $\boldsymbol{p} = (0.5, 0.5, 0.5, 0.5)$ and the correlation matrix for

(J_1, J_2, J_3, J_c) is

$$R = \begin{pmatrix} 1 & 0 & 0 & 0.8 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.2 \\ 0.8 & 0 & 0.2 & 1 \end{pmatrix}.$$

Then we generate $n = 100$ binary vectors of (J_1, J_2, J_3, J_c) according to (p, R). Then we obtain a transaction dataset containing 100 transactions on 4 items I_1, I_2, I_3, I_c. For each response item, there is in total $2^3 - 1 = 7$ possible association rules. We first use the Apriori algorithm [Hahsler et al., 2005] to mine all association rules of the form $(\boldsymbol{J} \Rightarrow I_c)$ with support and confidence greater than 0 and summarize the results in Table 2.1. We choose $g(\boldsymbol{J}) = \mathrm{supp}(\boldsymbol{J} \Rightarrow I_c) \times \mathrm{conf}(\boldsymbol{J} \Rightarrow I_c)$. Then we use the proposed Gibbs sampling algorithm to generate three random samples of size $L = 1,000$ of association rules from the transaction dataset by choosing $\xi = 3, 6, 10$, respectively. The frequency of each association rule appearing in each sample is shown in Table 2.1. The rank of the frequency conforms to that of $g(\boldsymbol{J})$, showing the good performance of our method. It is easy to see that the frequencies have more power to differentiate the most important rules from the less important ones, as the value of ξ increases. Next we illustrate how to use the random search procedure and how well it performs on three more complex datasets.

Example 2. Consider an item space $I = (I_1, I_2, \ldots, I_{398}, I_c)$ with $m = 398$ predic-

Table 2.1: Association rules and their measurements

Rules	supp	conf	$g(\cdot)$	Frequencies		
				$\xi = 3$	6	10
$I_1 \Rightarrow I_c$	0.47	0.890	0.420	0.242	0.382	0.595
$I_1, I_3 \Rightarrow I_c$	0.28	1.000	0.280	0.190	0.194	0.166
$I_3 \Rightarrow I_c$	0.33	0.650	0.210	0.171	0.155	0.095
$I_1, I_2 \Rightarrow I_c$	0.21	0.910	0.190	0.113	0.093	0.064
$I_1, I_2, I_3 \Rightarrow I_c$	0.11	1.000	0.110	0.101	0.063	0.021
$I_2 \Rightarrow I_c$	0.22	0.470	0.100	0.094	0.057	0.035
$I_2, I_3 \Rightarrow I_c$	0.12	0.570	0.070	0.089	0.056	0.024

"$\cdot$" represents the association rule $\boldsymbol{J} \Rightarrow I_c$.

tor items and one response item. Set each marginal probability as

$$\begin{aligned}\boldsymbol{p} &= \{p_1, p_2, p_3, p_4, \dots, p_{398}, p_c\} \\ &= \{0.8, 0.8, 0.8, 0.2, \dots, 0.2, 0.8\}.\end{aligned}$$

The correlation matrix R between items is set to be an identity matrix except that $R(J_{s_1}, J_{s_2}) = 0.99$ where $s_1, s_2 \in \{1, 2, 3, c\}$. Then we generate $n = 300$ binary vectors from (J_1, J_2, J_3, J_c) according to (p, R). The transaction dataset T_1 is accordingly formed to contain 399 items and 300 transactions.

Example 3. The transaction dataset T_2 has the same item space, the same number of transactions, and the same correlation matrix as T_1 but a different marginal probability vector

$$\begin{aligned} \boldsymbol{p} &= \{p_1, p_2, p_3, p_4, \ldots, p_{20}, p_{21}, \ldots, p_{398}, p_c\} \\ &= \{0.8, 0.8, 0.8, 0.5 \ldots, 0.5, 0.2, \ldots, 0.2, 0.8\}. \end{aligned}$$

Example 4. The transaction database T_3 also has $l = 399$ items and $n = 300$ transactions. The marginal probability vector is

$$\begin{aligned} \boldsymbol{p} &= \{p_1, p_2, p_3, p_4, \ldots, p_{10}, p_{11}, \ldots, p_{398}, p_c\} \\ &= \{0.8, 0.8, 0.8, 0.6 \ldots, 0.6, 0.2, \ldots, 0.2, 0.8\}. \end{aligned}$$

The correlation matrix R is an identity matrix except that

$$\begin{aligned} R(J_{s_1}, J_{s_2}) &= 0.9, \quad \text{for } s_1 \neq s_2;\ s_1,\ s_2 \in \{1, 2, 3, c\}, \\ R(J_{s_1}, J_{s_2}) &= 0.5, \quad \text{for } s_1 \neq s_2;\ s_1,\ s_2 \in \{4, \ldots, 10, c\}, \\ R(J_{s_1}, J_{s_2}) &= 0.5, \quad \text{for } s_1 \in \{1, 2, 3\},\ s_2 \in \{4, 5, \ldots, 10\}. \end{aligned}$$

From the settings of T_1, T_2 and T_3, we see that items I_1, I_2 and I_3 have high support and the antecedents of the important association rules in these datasets most likely contain some of I_1, I_2 and I_3. We now use the Apriori algorithm and the new Gibbs-sampling-based search procedure to see whether we can unveil these attributes in T_1, T_2 and T_3.

To mine the association rules in $\mathcal{R}_{I_c}$ of each transaction dataset, a random sample of 100 association rules is generated from each $\mathcal{R}_{I_c}$ using the new algorithm. We find that the larger ξ is, the more frequently the three items I_1, I_2 and I_3 appear in the generated sample. When $\xi = 100$, all items ever appearing in the sample are I_1, I_2, I_3 and I_{390}. Proportions of the sampled association rules containing each of (I_1, I_2, I_3, I_{390}) from T_1, T_2 and T_3 are shown in Table 2.2. The item I_{390} appears only once in each sample, thus seeming not to have high support in the datasets.

We then apply the Apriori algorithm with the constraint of minimum support 0.05 and minimum confidence 0.6 on the search. This identifies 31,525, 170,600, and 442,191 association rules from T_1, T_2 and T_3 respectively. The 10 most frequent items appearing in these rules for each dataset and their respective proportions of appearance are shown in Table 2.3. For each dataset the top 10 of the identified rules according to $g(\cdot)$ are also calculated and presented in Table 2.4 - 2.6, together with their respective frequencies of appearance in the corresponding random sample generated. Ranks of the top 10 rules in terms of the frequencies in Table 2.4 - 2.6 more or less conform to their ranks in terms of $g(\cdot)$. We find that as the dependence structure of the transaction dataset becomes more complicated, our algorithm can generate a random sample containing the most significant association rules that are confirmed by the Apriori algorithm.

Table 2.2: Items appeared in the random sample for T_1, T_2, T_3

T_1	item	I_{390}	I_3	I_2	I_1
	proportion	0.01	0.43	0.51	0.55
T_2	item	I_{390}	I_3	I_2	I_1
	proportion	0.01	0.43	0.51	0.55
T_3	item	I_{390}	I_2	I_1	I_3
	proportion	0.01	0.55	0.60	0.85

Table 2.3: Top 10 frequent items appearing in the rules identified by the Apriori algorithm for T_1, T_2, or T_3

T_1	item	I_{44}	I_{292}	I_{135}	I_{97}	I_{286}	I_{184}	I_{187}	I_3	I_1	I_2
	proportion	0.019	0.021	0.023	0.024	0.025	0.025	0.027	0.493	0.496	0.500
T_2	item	I_{14}	I_7	I_4	I_{15}	I_8	I_6	I_{13}	I_3	I_1	I_2
	proportion	0.087	0.090	0.091	0.093	0.105	0.130	0.136	0.496	0.499	0.500
T_3	item	I_9	I_4	I_6	I_{10}	I_7	I_5	I_8	I_1	I_2	I_3
	proportion	0.434	0.436	0.438	0.444	0.445	0.445	0.447	0.498	0.499	0.500

From Examples 2-4, we see that our method is capable of finding the most important association rules that also appear most frequently in the random sample generated by properly choosing a large value for ξ. In cases where the item space is large and the support of rules is very low, our proposed algorithm can be combined with the Apriori algorithm to more efficiently tackle the association rule mining task.

2.3 Real Data Application

We apply the proposed Gibbs sampling method to mine a case-control dataset that contains genomic observations for $n = 229$ women, 39 of which are breast cancer cases obtained from the Australian Breast Cancer Family Study (ABCFS) (Dite GS, et al. 2003) and 190 of which are controls from the Australian Mammographic Density Twins and Sisters Study (AMDTSS) (Odefrey F, et al. 2010). The dataset is formed by sampling from a much larger data source from ABCFS and AMDTSS. Each woman in the dataset has 366 genetic observations being the genotype outcomes (from a Human610-Quad beadchip array) of the 366 SNPs on a specific gene pathway suspected to be susceptible to breast cancer. An SNP variable typically takes a value from 0, 1, and 2, representing the number of the minor alleles at the SNP loci. But, in the current dataset there are 31 SNPs, with only 2 of the 3 possible values being observed. Our task is to find out whether there are any SNPs having significant associations with the risk of breast cancer and what these SNPs are. One could use a logistic model to tackle this task. But, it is difficult due to that the number of predictor variables (i.e., SNPs) in the data is much larger than the number of observations, and the SNPs are highly associated with each other due to linkage disequilibrium. Because this dataset can be easily turned into a transaction one, we are able to use an association rule-mining method to undertake the task. The binary transaction dataset converted from our casecontrol dataset contains 1,067 predictor (SNP) items (denoted as $I_1, \ldots, I_{1067}$) and 1 response item I_c (breast cancer or not). It is easy to see that $0 \leq \text{supp}(\boldsymbol{J} \Rightarrow I_c) \leq 0.17$. We choose the measure of association rules as $g(\boldsymbol{J}) = \text{supp}(\boldsymbol{J} \Rightarrow I_c) \times \text{conf}(\boldsymbol{J} \Rightarrow I_c)$. Now our aim is to find the most significant association rules for I_c according to the measure $g(\cdot)$.

For the association rules in $\mathcal{R}_{I_c}$, the support of any of them is not greater than 0.17. Because the support of rules is too low and the item space is very large, the Apriori algorithm cannot cope with the computing intensity and immensity involved, even with the setting of minimum support 0.2 and minimum confidence 1. So, we try to use our proposed method to find the most significant rule with consequent I_c to reduce the size of the item space. The number of items appearing in the generated samples decreases from 1,067 to about 35 by increasing ξ from 10 to 6,000. But, it cannot be further reduced by larger value of ξ. The top 10 frequent items ever appearing in the generated samples are reported in the lower portion of Table 2.7. For illustration purposes we choose $\xi = 6000$, with which the number of distinct items appearing in the random sample is 35. We apply the Apriori algorithm on the subset of transaction dataset including only these 35 items by specifying the minimum support and confidence as 0.2 and 1, respectively. The Apriori algorithm is still not implementable. So, we then single out a subset of 22 items from the 35 items which appeared in at least three fourths of the sampled association rules and cut out a new subset of the original transaction dataset by including only these 22 items in the transactions. By specifying the minimum support and confidence as 0.05 and 0.6, a total number of 286,188 association rules have been found in the new subset transaction data. The top 10 important association rules among them are reported in Table 2.8. From the table, we can see that the measurements of these association rules are very low and close to each other. It is not possible to find out these rules by applying the Apriori algorithm alone. Our proposed Gibbs-sampling-based algorithm can be used to reduce the number of items for mining; the reduced data subset is exactly where the Apriori algorithm can be applied to find the most significant association rules subject to negligible information loss. One could look into these rules or the frequent items in Tables 2.7 and 2.8 to find out the biological meaning behind them.

Table 2.4: Top 10 significant association rules from T_1 and their frequencies in the relevant sample

Association Rules	supp	conf	$g(\cdot)$	frequency
$I_2 \Rightarrow I_c$	0.787	1.000	0.787	0.20
$I_3 \Rightarrow I_c$	0.783	1.000	0.783	0.12
$I_1 \Rightarrow I_c$	0.783	1.000	0.783	0.26
$I_2, I_3 \Rightarrow I_c$	0.783	1.000	0.783	0.12
$I_1, I_2 \Rightarrow I_c$	0.783	1.000	0.783	0.10
$I_1, I_3 \Rightarrow I_c$	0.780	1.000	0.780	0.10
$I_1, I_2, I_3 \Rightarrow I_c$	0.780	1.000	0.780	0.09
$I_3, I_{286} \Rightarrow I_c$	0.213	1.000	0.213	0.00
$I_1, I_{286} \Rightarrow I_c$	0.213	1.000	0.213	0.00
$I_2, I_{286} \Rightarrow I_c$	0.213	1.000	0.213	0.00

"$\cdot$" represents the association rule $\boldsymbol{J} \Rightarrow I_c$.

Table 2.5: Top 10 significant association rules from T_2 and their frequencies in the relevant sample

Association Rules	supp	conf	$g(\cdot)$	frequency
$I_2 \Rightarrow I_c$	0.787	1.000	0.787	0.20
$I_3 \Rightarrow I_c$	0.783	1.000	0.783	0.12
$I_1 \Rightarrow I_c$	0.783	1.000	0.783	0.26
$I_2, I_3 \Rightarrow I_c$	0.783	1.000	0.783	0.12
$I_1, I_2 \Rightarrow I_c$	0.783	1.000	0.783	0.10
$I_1, I_3 \Rightarrow I_c$	0.780	1.000	0.780	0.10
$I_1, I_2, I_3 \Rightarrow I_c$	0.780	1.000	0.780	0.09
$I_1, I_{13} \Rightarrow I_c$	0.450	1.000	0.450	0.00
$I_2, I_{13} \Rightarrow I_c$	0.450	1.000	0.450	0.00
$I_1, I_2, I_{13} \Rightarrow I_c$	0.450	1.000	0.450	0.00

" $\cdot$ " represents the association rule $\boldsymbol{J} \Rightarrow I_c$.

Table 2.6: Top 10 significant association rules from T_3 and their frequencies in the relevant sample

Association Rules	supp	conf	$g(\cdot)$	frequency
$I_1, I_3 \Rightarrow I_c$	0.783	0.996	0.780	0.26
$I_1, I_2, I_3 \Rightarrow I_c$	0.783	0.996	0.780	0.23
$I_3 \Rightarrow I_c$	0.793	0.979	0.777	0.15
$I_2, I_3 \Rightarrow I_c$	0.787	0.987	0.777	0.21
$I_1, I_2 \Rightarrow I_c$	0.783	0.983	0.770	0.08
$I_1 \Rightarrow I_c$	0.783	0.975	0.764	0.03
$I_2 \Rightarrow I_c$	0.787	0.963	0.758	0.03
$I_3, I_8 \Rightarrow I_c$	0.610	0.995	0.607	0.00
$I_3, I_5 \Rightarrow I_c$	0.607	1.000	0.607	0.00
$I_1, I_3, I_8 \Rightarrow I_c$	0.607	1.000	0.607	0.00

"$\cdot$" represents the association rule $\boldsymbol{J} \Rightarrow I_c$.

Table 2.7: Top 10 frequent items appeared in the random sample of association rules for I_c

$\xi = 2700$	item	I_{750}	I_{45}	I_{1004}	I_{42}	I_{389}	I_{804}	I_{191}	I_{193}	I_{214}	I_{711}
	proportion	0.60	0.63	0.70	0.72	0.86	0.92	0.98	0.99	0.99	0.99
$\xi = 3500$	item	I_{914}	I_{750}	I_{42}	I_{389}	I_{1004}	I_{191}	I_{193}	I_{214}	I_{711}	I_{804}
	proportion	0.64	0.71	0.74	0.95	0.97	0.99	0.99	0.99	0.99	0.99
$\xi = 6000$	item	I_{937}	I_{45}	I_{750}	I_{1004}	I_{389}	I_{214}	I_{711}	I_{191}	I_{193}	I_{804}
	proportion	0.65	0.67	0.67	0.84	0.90	0.93	0.96	0.99	0.99	0.99

Table 2.8: Top 10 association rules for I_c after reducing the item space

Association Rules	$\mathrm{supp}(\boldsymbol{J} \Rightarrow I_c)$	$\mathrm{conf}(\boldsymbol{J} \Rightarrow I_c)$	$g(\boldsymbol{J} \Rightarrow I_c)$
$I_7, I_{42}, I_{750}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061
$I_{645}, I_{914}, I_{42}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061
$I_{645}, I_{42}, I_{937}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061
$I_{636}, I_{914}, I_{42}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061
$I_{636}, I_{42}, I_{937}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061
$I_7, I_{45}, I_{750}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061
$I_{645}, I_{914}, I_{45}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061
$I_{645}, I_{937}, I_{45}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061
$I_{636}, I_{914}, I_{45}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061
$I_{636}, I_{937}, I_{45}, I_{1004}, I_{389}, I_{214}, I_{711}, I_{191}, I_{193}, I_{804} \Rightarrow I_c$	0.066	0.938	0.061

3 Simultaneous Multiple Change Points Estimation in Generalized Linear Models

In this chapter, we focus on the problem of multiple change points estimation in GLMs in which the number of change points and their locations are all unknown. In light of Jin et al. [2011], we propose a simultaneous multiple change points estimation method which partitions the data sequence into several segments to construct a new design matrix and estimate the regression coefficients by maximizing a penalized likelihood function. The consistency of the coefficient estimator is established in which the number of parameters in the penalized likelihood function is diverging as the sample size goes to infinity. The nonzero coefficient estimates provide the information about which segments potentially contain a change point. An algorithm is provided to estimate the change point in each possible segment. In this algorithm, we use the test statistic proposed in Antoch et al. [2004] to test if there exists a change point in each possible segment.

The rest of this chapter is organized as follows. In Section 3.1, we present a GLM with multiple change points and describe our methodology. A theorem regarding the consistency of the coefficient estimators is established and its proof is also provided. In Section 3.2, an algorithm is given to obtain the change point estimates. Simulation studies and a real data application are presented in Section 3.3 and Section 3.4 respectively. The test proposed by Antoch et al. [2004] is given in the Appendix A.1.

3.1 Simultaneous Multiple Change Points Estimation

3.1.1 The GLM with Multiple Change Points

Let $(y_{n1}, \boldsymbol{x}_{n1}), (y_{n2}, \boldsymbol{x}_{n2}), \cdots, (y_{nn}, \boldsymbol{x}_{nn})$ be a double-indexed series of random samples where y_{nt} is a scalar response and $\boldsymbol{x}_{nt} = (x_{nt1}, x_{nt2}, \cdots, x_{ntp})^T$ is a vector of covariates for all $t = 1, 2, \cdots, n$. Suppose that for every n and given $\boldsymbol{x}_{nt}$, Y_{nt} has a distribution in the exponential family, taking the form

$$f_{nt}(y_{nt}|\boldsymbol{x}_{nt}) = \exp\left\{\frac{y_{nt}\theta(\boldsymbol{x}_{nt}) - b(\theta(\boldsymbol{x}_{nt}))}{a(\phi)} + c(y_{nt}, \phi)\right\}$$

for some specific function $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. Then the expectation of Y_{nt} given $\boldsymbol{x}_{nt}$ is $\mu_{nt} = E(Y_{nt}|\boldsymbol{x}_{nt}) = b'(\theta(\boldsymbol{x}_{nt}))$ and the variance of Y_{nt} given $\boldsymbol{x}_{nt}$ is $\sigma_{nt}^2 = Var(Y_{nt}|\boldsymbol{x}_{nt}) = a(\phi)b''(\theta(\boldsymbol{x}_{nt}))$.

The GLM is formulated as

$$g(\mu_{nt}) = \sum_{j=1}^{p} \beta_j x_{ntj} = \boldsymbol{x}_{nt}^T \boldsymbol{\beta}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is the vector of parameters, and $g(\cdot)$ is a proper link function. In this dissertation, we consider the canonical link, *i.e.*, $g(\mu_{nt}) = \left(\frac{db}{d\theta}\right)^{-1}(\mu_{nt})$, then $\theta(\boldsymbol{x}_{nt}) = \boldsymbol{x}_{nt}^T\boldsymbol{\beta}$.

Denote the change points as $\{l_{n,1}, l_{n,2}, \cdots, l_{n,s}\}$ satisfying that $0 = l_0 < l_{n,1} < l_{n,2} < \cdots < l_{n,s} < l_{n,s+1} = n$, where s is the total number of change points. Consider the following GLM with multiple change points formulated as

$$g(\mu_{nt}) = \boldsymbol{x}_{nt}^T \boldsymbol{\beta}_i, \qquad l_{n,i-1} < t \leq l_{n,i}, \qquad i = 1, 2, \cdots, s+1, \qquad t = 1, 2, \cdots, n, \tag{3.1}$$

where $\boldsymbol{\beta_i} = (\beta_{i1}, \cdots, \beta_{ip})^T$ is the parameter vector associated with the i^{th} segment $\{l_{n,i-1}, \ldots, l_{n,i}\}$. The objective is to estimate the total number of change points, s, and their locations, $l_{n,1}, l_{n,2}, \cdots, l_{n,s}$.

In model (3.1), the variables depend on the sample size n, and $l_{n,i}$ increases as $n \to \infty$. We assume throughout this chapter that $l_{n,i} = \lfloor \tau_i n \rfloor$, where $\tau_i \in (0, 1)$ for $i = 1, 2 \cdots, s$. Set $\tau_0 = 0$ and $\tau_{s+1} = 1$. For the rest of the chapter, the subscript n is suppressed if there is no confusion.

3.1.2 The Method

In order to estimate the change points in model (3.1), the proposed method is to transform the change points detection problem into a model selection problem by partitioning the data sequence and rewriting model (3.1) into model (3.2), and then utilize modern model selection techniques to estimate the total number of change points, s and the change points l_i's simultaneously. The procedure is described as following.

1. Partition the data sequence into q_n segments, $Q_1 = \{1, 2, \cdots, n-(q_n-1)m\}$ as the first segment with length $n-(q_n-1)m$ satisfying that $m \leq n-(q_n-1)m \leq d_0 m$ for some $d_0 \geq 1$ and $Q_k = \{n-(q_n-k+1)m+1, \cdots, n-(q_n-k)m\}$ as the k^{th} segment with length m for $k = 2, 3, \cdots, q_n$. Then there exist $n_1 < n_2 < \cdots < n_s$ such that $l_i \in Q_{n_i}$ for $i = 1, 2, \cdots, s$.

2. Rewriting model (3.1) in order to incorporate the partition yields the following model

$$g(\mu_t) = \boldsymbol{x}_t^T [\, \boldsymbol{\beta}_1 + \sum_{k=2}^{q_n} \boldsymbol{\delta}_k I_{\{n-(q_n-k+1)m+1,\ldots,\ n\}}(t) \,] - v_t, \tag{3.2}$$

   where

$$\boldsymbol{\delta}_k = \begin{cases} \boldsymbol{\beta}_{i+1} - \boldsymbol{\beta}_i, & \text{for } k = n_i, \quad i = 1, 2, \ldots, s, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$v_t = \begin{cases} \boldsymbol{x}_t^T \boldsymbol{\delta}_k, & \text{for } k = n_i, \quad t \in \{n - (q_n - k + 1)m + 1, \ldots, l_i\}, \\ 0, & \text{otherwise,} \end{cases}$$

$t = 1, 2, \cdots, n$. For the sake of convenience, denote $\varsigma_i = n - (q_n - n_i + 1)m + 1$.

3. Denote $\boldsymbol{g}(\boldsymbol{\mu}) = (\ g(\mu_1),\ g(\mu_2),\ \ldots,\ g(\mu_n)\ )^T$. Let $\mathcal{A} = \cup_{i=0}^{s} B_i$, where $B_i = \{(n_i - 1)p + 1, \ldots, n_i p\},\ i = 1, \ldots, s, B_0 = \{1, \ldots, p\}$ and $\mathcal{A}^c = \{1, \ldots, pq_n\} \backslash \mathcal{A}$. Denote $\boldsymbol{\gamma} = (\ \boldsymbol{\beta}_1^T,\ \boldsymbol{\delta}_2^T,\ \cdots,\ \boldsymbol{\delta}_{q_n}^T\ )^T = (\ \gamma_1,\ \gamma_2,\ \cdots,\ \gamma_{pq_n}\ )^T$, where $\boldsymbol{\gamma}_{[\mathcal{A}^c]} = \boldsymbol{0}$. Now we write model (3.2) in the following matrix form

$$\boldsymbol{g} = Z\boldsymbol{\gamma} - W\boldsymbol{\gamma}. \tag{3.3}$$

Here,

$$\begin{aligned} Z &= [\ \boldsymbol{z}_1,\ \boldsymbol{z}_2,\ \cdots,\ \boldsymbol{z}_n]^T = [\ \tilde{\boldsymbol{z}}_1,\ \tilde{\boldsymbol{z}}_2,\ \cdots,\ \tilde{\boldsymbol{z}}_{pq_n}] \\ &= \begin{pmatrix} Z^{(1)} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ Z^{(2)} & Z^{(2)} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ & & \cdots & & \\ Z^{(q_n)} & Z^{(q_n)} & Z^{(q_n)} & Z^{(q_n)} & Z^{(q_n)} \end{pmatrix}_{n \times (pq_n)}, \\ Z^{(1)} &= (\ \boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots,\ \boldsymbol{x}_{n-(q_n-1)m}\ )^T, \text{ of dimension } (n - (q_n - 1)m) \times p, \\ Z^{(2)} &= (\ \boldsymbol{x}_{n-(q_n-1)m+1},\ \boldsymbol{x}_{n-(q_n-1)m+2},\ \cdots,\ \boldsymbol{x}_{n-(q_n-2)m}\ )^T, \text{ of dimension } m \times p, \\ &\cdots \end{aligned}$$

$$Z^{(q_n)} \;=\; (\; \boldsymbol{x}_{n-m+1},\; \boldsymbol{x}_{n-m+2},\; \cdots,\; \boldsymbol{x}_n\;)^T, \text{ of dimension } m \;\times\; p\;,$$

$$\boldsymbol{z}_t, \qquad t \;=\; 1,\; 2,\; \ldots,\; n \;\text{ are row vectors of } Z,$$

$$\tilde{\boldsymbol{z}}_j, \qquad j \;=\; 1,\; 2,\; \cdots,\; pq_n \;\text{ are column vectors of } Z,$$

and $W_{n\times(pq_n)} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n)^T$ with $\boldsymbol{w}_t = \mathbf{0}$ for $t \notin \{n-(q_n-n_i+1)m+1, \ldots, l_i\}$, otherwise $\boldsymbol{w}_{t[B_i]} = \boldsymbol{x}_t$ and $\boldsymbol{w}_{t[B_i^c]} = \mathbf{0}$, where $t \;=\; 1,\; 2,\; \ldots,\; n$ and i = 1, 2, ..., s.

Then the log-likelihood function for model (3.3) is

$$\mathcal{L}(\boldsymbol{\gamma}) = \sum_{t=1}^{n} \left[\frac{y_t(\boldsymbol{z}_t^T\boldsymbol{\gamma} - \boldsymbol{w}_t^T\boldsymbol{\gamma}) - b(\boldsymbol{z}_t^T\boldsymbol{\gamma} - \boldsymbol{w}_t^T\boldsymbol{\gamma})}{a(\phi)} + c(y_t, \phi) \right].$$

4. Denote $Q(\boldsymbol{\gamma}) = \mathcal{L}_1(\boldsymbol{\gamma}) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_j|)$ where $\mathcal{L}_1(\boldsymbol{\gamma}) = \sum_{t=1}^{n}(\frac{y_t(\boldsymbol{z}_t^T\boldsymbol{\gamma}) - b(\boldsymbol{z}_t^T\boldsymbol{\gamma})}{a(\phi)} + c(y_t, \phi))$. We propose to estimate $\boldsymbol{\gamma}$ in model (3.3) by maximizing the following penalized log-likelihood function

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} Q(\boldsymbol{\gamma}) = \arg\max_{\boldsymbol{\gamma}} \left\{ \mathcal{L}_1(\boldsymbol{\gamma}) - n\sum_{j=1}^{pq_n} p_{\lambda_n,d}(|\gamma_j|) \right\}, \tag{3.4}$$

where $\lambda_n > 0$, $d > 0$, and the penalty function $p_{\lambda_n,d}(\theta)$ is symmetric about $\theta = 0$ and satisfies the following assumptions: $p_{\lambda_n,d}(0) = 0$, $p'_{\lambda_n,d}(\theta) = 0$ if $\theta > \lambda_n d$ and $p'_{\lambda_n,d}(0) = \lambda_n$. There are two penalty functions among others that meet these assumptions. One is the SCAD penalty function defined in Fan and Li [2001] satisfying that $p_{\lambda_n,d}(0) = 0$ and $p'_{\lambda_n,d}(\theta) = \lambda_n\{I_{(0,\lambda_n]}(\theta) +$

$\frac{(d\lambda_n-\theta)_+}{(d-1)\lambda_n} I_{(\lambda_n,\infty)}(\theta)\}$. The other is the MCP penalty defined in Zhang [2010] satisfying that $p_{\lambda_n,d}(\theta) = (\lambda_n\theta - \frac{\theta^2}{2d}) I_{(0,d\lambda_n]}(\theta) + \frac{1}{2} d\lambda_n^2 I_{(d\lambda_n,\infty)}(\theta)$. In this dissertation, we use these two penalty functions for illustration purpose. Other penalty functions may also be used to derive the coefficient estimator.

3.1.3 Consistency of the Proposed Estimator

To study the asymptotic properties of the estimator $\hat{\boldsymbol{\gamma}}$, we assume that there is an underlying true model with true change points $l^*_{n,i} = \lfloor n\tau^*_i \rfloor,\ i = 1, 2, \cdots, s$ and there exist true values of $\boldsymbol{\gamma}_n : \boldsymbol{\gamma}^0_n = (\gamma^0_{n1}, \ldots, \gamma^0_{n,pq_n})^T$ with $\boldsymbol{\gamma}^0_{n[\mathcal{A}^c_n]} = \mathbf{0}$. Note that the dimension of $\boldsymbol{\gamma}_n$ goes to ∞ as $n \to \infty$. To prove the consistency of the estimator $\hat{\boldsymbol{\gamma}}_n$, we employ the techniques developed in Fan and Peng [2004] which proves the asymptotic properties of the maximum nonconcave penalized likelihood estimator with a diverging number of parameters. The following assumptions make the technical proof easy to follow. The first four assumptions are made on both the likelihood term and penalty term. The last one is made on the term involving $\boldsymbol{w}$.

Assumption 1. $\liminf_{n\to\infty} \liminf_{\gamma\to 0^+} p'_{\lambda_n}(\gamma)/\lambda_n > 0$.

Assumption 2. $\lambda_n \to 0$, $\sqrt{n/q_n}\lambda_n \to \infty$ as $n \to \infty$.

Assumption 3. $\min_{j\in\mathcal{A}}\{|\gamma^0_{nj}|/\lambda_n\} \to \infty$ as $n \to \infty$.

Assumption 4. For every n and i, $\{(Y_t, \boldsymbol{x}_t), l_{i-1} < t \le l_i\}$ are independent and identically distributed with probability density $f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta_i})$, which has a common support, and the model is identifiable. Furthermore, they satisfy the following three regularity conditions.

(1) The first and second derivatives of the likelihood function satisfy the joint equations

$$E_{\boldsymbol{\beta}_i}\left\{\frac{\partial \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial \beta_{ij}}\right\} = 0,$$

and

$$E_{\boldsymbol{\beta}_i}\left\{\frac{\partial \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial \beta_{ij}}\frac{\partial \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial \beta_{ik}}\right\} = -E_{\boldsymbol{\beta}_i}\left\{\frac{\partial^2 \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial \beta_{ij}\partial \beta_{ik}}\right\},$$

for $j, k = 1, 2, \ldots, p$.

(2) The Fisher information matrix

$$I(\boldsymbol{\beta}_i) = E_{\boldsymbol{\beta}_i}\left[\left\{\frac{\partial \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial \boldsymbol{\beta}_i}\right\}\left\{\frac{\partial \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial \boldsymbol{\beta}_i}\right\}^T\right]$$

satisfies conditions $0 < C_1 < e_{\min}\{I(\boldsymbol{\beta}_i)\} \leq e_{\max}\{I(\boldsymbol{\beta}_i)\} < C_2 < \infty$ for all n with $e_{\min}\{I(\boldsymbol{\beta}_i)\}$ and $e_{\max}\{I(\boldsymbol{\beta}_i)\}$ denoting the minimum and maximum eigenvalues of $I(\boldsymbol{\beta}_i)$ respectively. For $j, k = 1, 2, \ldots, p$,

$$E_{\boldsymbol{\beta}_i}\left\{\frac{\partial \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial \beta_{ij}}\frac{\partial \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial \beta_{ik}}\right\}^2 < C_3 < \infty$$

and

$$E_{\boldsymbol{\beta}_i}\left\{\frac{\partial^2 \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial \beta_{ij}\partial \beta_{ik}}\right\}^2 < C_4 < \infty.$$

(3) There is a large enough open subset ω_i of $\Omega \in R^p$ which contains the true parameter $\boldsymbol{\beta}_i$, such that for almost all $(Y_t, \boldsymbol{x}_t)$, the density admits all third derivatives $\partial f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)/\partial\beta_{ij}\partial\beta_{ik}\partial\beta_{il}$ for all $\boldsymbol{\beta}_i \in \omega_i$. Furthermore, there are functions M_{njkl} such that

$$|\frac{\partial \log f_{n,i}(y_{l_i}, \boldsymbol{x}_{l_i}, \boldsymbol{\beta}_i)}{\partial\beta_{ij}\partial\beta_{ik}\partial\beta_{il}}| \leq M_{njkl}(y_{l_i}, \boldsymbol{x}_{l_i})$$

for all $\boldsymbol{\beta}_i \in \omega_i$, and

$$E_{\boldsymbol{\beta}_i}\{M^2_{njkl}(y_{l_i}, \boldsymbol{x}_{l_i})\} < C_5 < \infty$$

for all p, n, j, k, l.

These regularity conditions correspond to Assumptions (E) - (G) in Fan and Peng (2004).

Assumption 5. Assume that $\min\{\tau_i^* - \tau_{i-1}^*,\ i = 1, 2, \cdots, s+1\} > \iota > 0$ where ι is a constant. Also assume that $q_n = O(n^{\frac{1}{6}})$ and $l^*_{n,i} - \varsigma_i = O(\sqrt{nq_n})$ where $\varsigma_i = n - (q_n - n_i + 1)m + 1$.

To this end, we state the theorem as follows and its proof is also given.

Theorem 3.1.1 *If Assumptions 1-5 hold, there exists a local maximizer $\hat{\boldsymbol{\gamma}}_n$ to $Q(\boldsymbol{\gamma}_n)$ and $\|\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^0\| = O_p((n/q_n)^{-\frac{1}{2}})$, where $\hat{\boldsymbol{\gamma}}_n$ is the SCAD estimator. Furthermore, we have* $\lim_{n\to\infty} P(\hat{\boldsymbol{\gamma}}_{n[\mathcal{A}_n^c]} = \mathbf{0}) = 1$.

Proof. Consider a ball $\|\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n^0\| \leq M(n/q_n)^{-\frac{1}{2}}$ for some finite M.

$$\begin{aligned}
& Q(\boldsymbol{\gamma}_n) \\
= \; & \mathcal{L}_1(\boldsymbol{\gamma}_n) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_{nj}|) \\
= \; & \sum_{t=1}^{n}(\frac{y_{nt}(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)} + c(y_{nt},\phi)) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_{nj}|) \\
= \; & \sum_{t=1}^{n}(\frac{y_{nt}(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)} + c(y_{nt},\phi)) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_{nj}|) \\
& + \sum_{t=1}^{n}\frac{y_t(\boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)} - \sum_{i=1}^{n}\frac{b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)} \\
= \; & \mathcal{L}(\boldsymbol{\gamma}_n) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_{nj}|) + \sum_{t=1}^{n}\frac{y_t(\boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)} - \sum_{i=1}^{n}\frac{b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)} \\
= \; & \mathcal{L}(\boldsymbol{\gamma}_n) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_{nj}|) + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}}\frac{y_t(\boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)} - \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}}\frac{b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)}
\end{aligned}$$

where $\boldsymbol{w}_{nt} = \mathbf{0}$ for $t \notin \{n - (q_n - n_i + 1)m + 1, \ldots, l_{n,i}\}$.

First, we consider $\|\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n^0\| = M(n/q_n)^{-\frac{1}{2}}$.

$$\begin{aligned}
& Q(\boldsymbol{\gamma}_n) - Q(\boldsymbol{\gamma}_n^0) \\
= \; & (\mathcal{L}(\boldsymbol{\gamma}_n) - \mathcal{L}(\boldsymbol{\gamma}_n^0)) - n\sum_{j=1}^{pq_n}(p_{\lambda_n}(|\gamma_{nj}|) - p_{\lambda_n}(|\gamma_{nj}^0|)) + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}}\frac{y_{nt}(\boldsymbol{w}_{nt}^T(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n^0))}{a(\phi)} \\
& - \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}}\frac{b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n^0)}{a(\phi)} + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}}\frac{b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n^0 - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n^0)}{a(\phi)}
\end{aligned}$$

$$
\begin{aligned}
&= (\mathcal{L}(\boldsymbol{\gamma}_n) - \mathcal{L}(\boldsymbol{\gamma}_n^0)) - n\sum_{j\in\mathcal{A}_n}(p_{\lambda_n}(|\gamma_{nj}|) - p_{\lambda_n}(|\gamma_{nj}^0|)) - n\sum_{j\in\mathcal{A}_n^c}(p_{\lambda_n}(|\gamma_{nj}|) - p_{\lambda_n}(|\gamma_{nj}^0|)) \\
&\quad + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{y_{nt}(\boldsymbol{w}_{nt}^T(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n^0))}{a(\phi)} - \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n^0)}{a(\phi)} \\
&\quad + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n^0 - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n^0)}{a(\phi)}.
\end{aligned}
$$

As $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(|\gamma_{nj}|) \geq 0$, we have

$$
\begin{aligned}
& Q(\boldsymbol{\gamma}_n) - Q(\boldsymbol{\gamma}_n^0) \\
\leq\ & (\mathcal{L}(\boldsymbol{\gamma}_n) - \mathcal{L}(\boldsymbol{\gamma}_n^0)) - n\sum_{j\in\mathcal{A}_n}(p_{\lambda_n}(|\gamma_{nj}|) - p_{\lambda_n}(|\gamma_{nj}^0|)) + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{y_{nt}(\boldsymbol{w}_{nt}^T(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n^0))}{a(\phi)} \\
& - \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n^0)}{a(\phi)} + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n) - b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n^0 - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n^0)}{a(\phi)} \\
\leq\ & [\mathcal{L}(\boldsymbol{\gamma}_n) - \mathcal{L}(\boldsymbol{\gamma}_n^0)] - n\sum_{j\in\mathcal{A}_n}[p'_{\lambda_n}(|\gamma_{nj}^0|)\mathrm{sign}(\gamma_{nj}^0)(\gamma_{nj} - \gamma_{nj}^0) + p''_{\lambda_n}(|\gamma_{nj}^0|)(\gamma_{nj} - \gamma_{nj}^0)^2(1 + o_P(1))] \\
& + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} a(\phi)^{-1}[y_{nt}(\boldsymbol{w}_{nt}^T(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n^0)) - \frac{\partial b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n^*)}{\partial\boldsymbol{\gamma}_n}\boldsymbol{z}_{nt}^T(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n^0) \\
& \quad + \frac{\partial b(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n^* - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n^*)}{\partial\boldsymbol{\gamma}_n}(\boldsymbol{z}_{nt}^T - \boldsymbol{w}_{nt}^T)(\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n^0)] \\
=\ & A_1 + A_2 + A_3
\end{aligned}
$$

where $\|\boldsymbol{\gamma}_n^* - \boldsymbol{\gamma}_n^0\| \leq M(n/q_n)^{-\frac{1}{2}}$.

By the Taylor expansion and Assumption 4, $A_1 = \mathcal{L}(\boldsymbol{\gamma}_n) - \mathcal{L}(\boldsymbol{\gamma}_n^0) = -M^2 O_p(q_n)$. By Assumption 2, $p'_{\lambda_n}(|\gamma_{nj}^0|) = p''_{\lambda_n}(|\gamma_{nj}^0|) = 0$, for $j \in \mathcal{A}_n$ and large n. Then $|A_2| = o_p(q_n^{\frac{1}{2}})$. By Assumption 5, $|A_3| = O_P(\sqrt{nq_n})M(n/q_n)^{-\frac{1}{2}} = O_p(q_n)$. By choosing a sufficiently large M, the first term dominates the other terms. Since A_1 is negative, for $\varepsilon > 0$, there exists a large constant M such that $P\{\sup_{\|\boldsymbol{\gamma}_n-\boldsymbol{\gamma}_n^0\|=M(n/q_n)^{-\frac{1}{2}}} Q(\boldsymbol{\gamma}_n) < Q(\boldsymbol{\gamma}_n^0)\} \geq 1-\varepsilon$. This implies that with probability at least $1-\varepsilon$ there exists a local maximum in the ball $\{\boldsymbol{\gamma}_n : \|\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_n^0\| \leq M(n/q_n)^{-\frac{1}{2}}\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_n^0\| = O_P((n/q_n)^{-\frac{1}{2}})$.

Then we consider for $j \in \mathcal{A}_n^c$,

$$\begin{aligned}
& \frac{\partial Q(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}} \\
= \; & \frac{\partial \mathcal{L}(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}} - np'_{\lambda_n}(|\gamma_{nj}|)\text{sign}(\gamma_{nj}) \\
& + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}}\sum_{r=1}^{s} y_{nt} x_{nt(j-(n_r-1)p)} I_{B_r}(j) - \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{b'(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)} x_{nt(j-(n_i-1)p)} I_{\cup_{k=0}^{i} B_k}(j) \\
& + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{b'(\boldsymbol{z}_{nt}^T\boldsymbol{\gamma}_n - \boldsymbol{w}_{nt}^T\boldsymbol{\gamma}_n)}{a(\phi)} (x_{nt(j-(n_i-1)p)} I_{\cup_{k=0}^{i} B_k}(j) - \sum_{r=1}^{s} x_{nt(j-(n_r-1)p)} I_{B_r}(j)).
\end{aligned}$$

By the standard Taylor expansion of the function $\frac{\partial \mathcal{L}(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}}$ at $\boldsymbol{\gamma}_n^0$, we obtain

$$\begin{aligned}
& \frac{\partial Q(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}} \\
= \; & \frac{\partial \mathcal{L}(\boldsymbol{\gamma}_n^0)}{\partial \gamma_{nj}} + \sum_{j'=1}^{pq_n} (\gamma_{nj'} - \gamma_{nj'}^0) \frac{\partial^2 \mathcal{L}(\boldsymbol{\gamma}_n^0)}{\partial \gamma_{nj}^2}(1 + O_P(1)) - np'_{\lambda_n}(|\gamma_{nj}|)\text{sign}(\gamma_{nj}) + O_P(\sqrt{nq_n}) \\
= \; & O_P(\sqrt{nq_n}) + O_P(\sqrt{nq_n}) - np'_{\lambda_n}(|\gamma_{nj}|)\text{sign}(\gamma_{nj}) + O_P(\sqrt{nq_n}) \\
= \; & n\lambda_n[O_P(\frac{\sqrt{q_n/n}}{\lambda_n}) - \lambda_n^{-1} p'_{\lambda_n}(|\gamma_{nj}|)\text{sign}(\gamma_{nj})]
\end{aligned}$$

by Assumption 1. Since $\frac{\sqrt{q_n/n}}{\lambda_n} \to 0$ by Assumption 2, this entails that the sign of $\frac{\partial Q(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}}$ is determined by the sign of γ_{nj} inside the neighborhood of $\boldsymbol{\gamma}_n^0$ with radius $M(n/q_n)^{-\frac{1}{2}}$ by assumption 3. That is, $\frac{\partial Q(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}} > 0$ for $\gamma_{nj} < 0$ and $\frac{\partial Q(\boldsymbol{\gamma}_n)}{\partial \gamma_{nj}} < 0$ for

$\gamma_{nj} > 0$. Therefore, for any local maximizer $\hat{\boldsymbol{\gamma}}_n$ inside this ball, $\hat{\boldsymbol{\gamma}}_{n\mathcal{A}_n^c} = 0$ with probability tending to one. This completes the proof. □

Let $\hat{\mathcal{A}} = \{j : \hat{\gamma}_j \neq 0\}$. Then the total number of change points is estimated by the size of the set $\{\lceil j/p \rceil, j \in \hat{\mathcal{A}}\}$ which is denoted as $\hat{s}$. Theorem 3.1.1 implies the consistency of $\hat{s}$ to s. It also provides the information that the $\hat{k}_i^{th}$ segment contains a change for each $\hat{k}_i \in \{\lceil j/p \rceil, j \in \hat{\mathcal{A}}\}, j = 1, \ldots, \hat{s}$.

3.2 An Algorithm

Since $\hat{\boldsymbol{\gamma}}_n$ provides the information about which segments potentially contain a change point, we present an algorithm in this section to detect the change point for each possible segment. The algorithm consists of the following steps.

Step 1. First, we test if there exists a change point in the sequence by the test proposed in Antoch, *et al.* [2004]. The details are given in Appendix A.1.

- If there is no change point, set $\tilde{s} = 0$.
- Otherwise, estimate the change point by the estimator in Appendix A.1 and denote it by $\hat{l}$. Then set $\tilde{s} = 1$.

Step 2. Compute the estimate $\hat{\boldsymbol{\gamma}}$ defined in (3.4) by the R Package *SIS* [Fan, *et al.*, 2010] or *cvplogistic* [Jiang and Huang, 2014].

Step 3. Let $\hat{s}$ record the number of change point estimates, $\hat{\mathbf{k}} = \{\hat{k}_1, \hat{k}_2, \ldots, \hat{k}_{\hat{s}}\}$ be a vector containing the change point estimates. Set $\hat{s} = 0$.

- If $\hat{\gamma}_j = 0$ for all $j > p$, go to Step 5.
- Otherwise, set $\tilde{\mathbf{k}} = \{\tilde{k}_1, \tilde{k}_2, \ldots, \tilde{k}_{s^*}\} = \{\lceil \frac{j}{p} \rceil$: for all $j > p$ such that $\hat{\gamma}_j \neq 0\}$ with $\tilde{k}_1 < \tilde{k}_2 < \ldots < \tilde{k}_{s^*}$ which records the segment number that contains possible change point and s^* is the total number of possible change points. Set $l = 1$ where l is from 1 to s^*.

Step 4. Use the test proposed in Antoch, *et al.* [2004] to detect a change point in each segment which possibly contains a change point. The details are given in Appendix A.1. This step is to reduce the overestimation of the number of change points from Step 3 and also can estimate the accuracy of change points.

- If $l > s^*$, go to Step 5.
- Otherwise, test $H_0^{(l)}$ that there is no change point in $g(\mu_t) = \boldsymbol{x}_t^T \boldsymbol{\beta},\ t = n - (q_n - \tilde{k}_l + 2)m + 1, \ldots, \leq n - (q_n - \tilde{k}_l)m$, at the significance level, 5% by Antoch, *et al.* [2004].
  - If the test is not significant, set $l = l + 1$, and repeat Step 4.
  - Otherwise, set $\hat{s} = \hat{s} + 1$, and $\hat{k}_{\hat{s}+1} = \tilde{k}_l$. Then we obtain a change point $\hat{k}_{\hat{s}}$ in this segment.

- $*$ If $\tilde{k}_{l+1} - \tilde{k}_l = 1$, set $l = l + 1$, and repeat Step 4.
- $*$ Otherwise, set $l = l + 2$, and repeat Step 4.

Step 5.

- If $\hat{s} \leq 1$,
  - If $\tilde{s} = 0$, there is no change point.
  - If $\tilde{s} = 1$, there exists one change point and the estimate of this change point, $\hat{k}$ is given by the estimate, $\hat{l}$ in Step 1.
- If $\hat{s} > 1$, the total number of change points is $\hat{s}$ and the estimates of these change points are $\{\hat{k}_1, \hat{k}_2, \ldots, \hat{k}_{\hat{s}}\}$.

In next two sections, data examples are presented to show the performance of the algorithm proposed in this section.

3.3 Simulation Studies

The false alarm rate (Type I error) and the accuracy of the change point estimates derived by the algorithm proposed in section 3.2 are evaluated through Monte Carlo simulations in this section. More specifically, we will calculate the empirical probabilities that the proposed algorithm erroneously detects change points when they actually do not exist. Moreover, we show how frequently the algorithm detects the correct number of change points and how accurately it estimates the change points when they do exist. Two specific generalized linear models, the logistic and the log models, are considered for demonstration purpose.

3.3.1 Two Specific Generalized Linear Models

For the binomial response, $y_t|\boldsymbol{x}_t \sim \text{Binomial}(1, \pi(\boldsymbol{x}_t))$. The density function is

$$f(y_t|\boldsymbol{x}_t) = \pi(\boldsymbol{x}_t)^{y_t}(1-\pi(\boldsymbol{x}_t))^{1-y_t} = \exp\left\{y_t \log\frac{\pi(\boldsymbol{x}_t)}{1-\pi(\boldsymbol{x}_t)} + \log(1-\pi(\boldsymbol{x}_t))\right\}.$$

Then $\theta(\boldsymbol{x}_t) = \log\frac{\pi(\boldsymbol{x}_t)}{1-\pi(\boldsymbol{x}_t)}$, $b(\theta(\boldsymbol{x}_t)) = \log\left(1+e^{\theta(\boldsymbol{x}_t)}\right)$, $\mu_t = b'(\theta(\boldsymbol{x}_t)) = \frac{e^{\theta(\boldsymbol{x}_t)}}{1+e^{\theta(\boldsymbol{x}_t)}}$, and $\sigma_t^2 = b''(\theta(\boldsymbol{x}_t)) = \frac{e^{\theta(\boldsymbol{x}_t)}}{(1+e^{\theta(\boldsymbol{x}_t)})^2}$. So the canonical link function for the Binomial response is $g(\mu_t) = \log(\frac{\mu_t}{1-\mu_t})$.

For the Poisson response, $y_t|\boldsymbol{x}_t \sim \text{Poisson}(\lambda(\boldsymbol{x}_t))$. The density function is

$$f(y_t|\boldsymbol{x}_t) = \frac{\lambda(\boldsymbol{x}_t)^{y_t}e^{-\lambda(\boldsymbol{x}_t)}}{y_t!} = \exp\{y_t \log\lambda(\boldsymbol{x}_t) - \lambda(\boldsymbol{x}_t) - \log(y_t!)\}.$$

Then $\theta(\boldsymbol{x}_t) = \log\lambda(\boldsymbol{x}_t)$, $b(\theta(\boldsymbol{x}_t)) = e^{\theta(\boldsymbol{x}_t)}$, $\mu_t = b'(\theta(\boldsymbol{x}_t)) = e^{\theta(\boldsymbol{x}_t)}$, and $\sigma^2 = b''(\theta(\boldsymbol{x}_t)) = e^{\theta(\boldsymbol{x}_t)}$. So the canonical link function for the Poisson response is $g(\mu_t) = \log(\mu_t)$.

3.3.2 GLMs with No Change Point

To examine the false alarm rate of the proposed algorithm, we consider the following four models, two for the binomial response and the other two for the Poisson response:

$B1 : \log \frac{\mu_t}{1-\mu_t} = -0.7; \qquad B2 : \log \frac{\mu_t}{1-\mu_t} = 12 - 3x_t;$

$P1 : \log(\mu_t) = 2; \qquad P2 : \log(\mu_t) = 2 - x_t$, where $t = 1, \ldots, n$.

All of these four models contain no change point. We first generate x_t from the uniform distribution $\mathrm{U}(0, 9)$ for $B2$ and $\mathrm{U}(0, 1)$ for $P2$. For each model, we generate $1,000$ independent series with length $n = 1,000$. The empirical probabilities that the proposed algorithm erroneously detects change points in the generated sequences are 0.039 for $B1$, 0.084 for $B2$, 0.034 for $P1$, and 0.044 for $P2$. This demonstrates that our algorithm has low false alarm rates for all these four models.

3.3.3 GLMs with One Change Point

In this subsection, the performance of the proposed algorithm is evaluated through Monte Carlo simulations from single change point models. The effect of the difference between two regression functions before and after the change point on the detection power is also studied. Furthermore, we compare the accuracy of the change point estimates derived by the proposed algorithm under the assumption that the number of change points is unknown with that of the test proposed in Antoch, et al. [2004] under the assumption that there is at most one change point.

We consider five models $B3-B7$ for the binomial response and five models $P3-P7$ for the Poisson response:

$B3: \log\frac{\mu_t}{1-\mu_t} = 1.0 - 0.8x_t + (1.9 + 0.1x_t)I_{[501,1000]}(t);$

$B4: \log\frac{\mu_t}{1-\mu_t} = 1.0 - 0.8x_t + (1.9 + 0.2x_t)I_{[501,1000]}(t);$

$B5: \log\frac{\mu_t}{1-\mu_t} = 1.0 - 0.8x_t + (1.9 + 0.3x_t)I_{[501,1000]}(t);$

$B6: \log\frac{\mu_t}{1-\mu_t} = 7 - 2x_t + (4 + 0x_t)I_{[501,1000]}(t);$

$B7: \log\frac{\mu_t}{1-\mu_t} = -0.7 - 0.2x_{1t} - 0.1x_{2t} + (2.0 + 0.3x_{1t} + 0.1x_{2t})I_{[501,1000]}(t);$

$P3: \log(\mu_t) = 2.3 - 1.5x_t + (-0.3 - 0.2x_t)I_{[501,1000]}(t);$

$P4: \log(\mu_t) = 2.3 - 1.5x_t + (-0.4 - 0.2x_t)I_{[501,1000]}(t);$

$P5: \log(\mu_t) = 2.3 - 1.5x_t + (-0.5 - 0.2x_t)I_{[501,1000]}(t);$

$P6: \log(\mu_t) = 8.5 - 2x_t + (0.5 + 0x_t)I_{[501,1000]}(t);$

$P7: \log(\mu_t) = 1.31 - 1.03x_{1t} - 0.56x_{2t} - (0.03 - 0.36x_{1t} - 0.9x_{2t})I_{[501,1000]}(t).$

All of these models contain single change point $l = 500$. First, we generate x_t from the uniform distribution $\mathrm{U}(0,9)$ for $B3-B7$ and $\mathrm{U}(0,1)$ for $P3-P7$, and then generate y_t according to each model for $t = 1, 2, \ldots, n$. The length of the sequence generated from models $B3-B7$ and $P3-P7$ is $n = 1{,}000$. The accuracy of the change point estimates is calculated based on 1000 independent simulations. Let

$\hat{\mathcal{N}}_i^{(M_j)} = \{\hat{t}_1^{(M_j)}, \ldots, \hat{t}_{\hat{s}}^{(M_j)}\}$ contain all change points estimated by the algorithm in the i^{th} simulation based on model M_j with $M = B$ or P for $i = 1, 2, \ldots, 1,000$ and $j = 3, 4, \ldots, 7$. Denote $\tilde{\epsilon}_{M_j} = \{\hat{\mathcal{N}}_i^{(M_j)} : |\hat{\mathcal{N}}_i^{(M_j)}| = 1, i = 1, 2, \ldots, 1,000\}$ for $j = 3, 4, \ldots, 7$ and $M = B$ or P. The results are reported in Table 3.1. Here $|\tilde{\epsilon}_{M_j}|$ denotes the number of simulations from model M_j out of 1000 in which the number of change points has been correctly detected. Let $Acc(l, r) = |\{\hat{k}_i : |\hat{k}_i - l| \leq r, i = 1, \ldots, 1000\}$ with $r = 10$ or 15 denote the number of simulations out of $1,000$ in which the change point estimate $\hat{k}_i$ falls into the interval of length $2r$ centered at the true change point l, for $i = 1, \ldots, 1000$.

The logistic functions for $B3 - B6$, and $P3 - P6$ are plotted in Figures 3.1 and 3.2. From the plots for $B3 - B5$ and $P3 - P5$, we can see that the larger the difference in coefficients (before and after the change points) of each model is, the larger the difference in two regression functions will be. This is also reflected in the accuracy of the change point estimates reported in Table 3.1 for models $B3 - B5$ and Table 3.2 for models $P3 - P5$. Larger difference in two regression functions before and after the change points results in higher power of detecting the correct number of change points and higher level of accuracy in estimating the change point.

However, for different types of response variables, as the values of the coefficients in the model increase, the same difference in the coefficients before and after the change point might have different impacts on the difference of two regression functions before and after the change point. For example, the plot for model $B6$ for the binomial response tells us that even though the difference in the coefficients is larger than that in $B4$, but the absolute value of the coefficient in $B6$ is also larger than that in $B4$, the differences between two logistic functions for model $B6$ is even less than that for $B4$. Therefore, the accuracy of the change point estimates for model $B6$ is lower than that for $B4$. However, for model $P6$ for the Poisson response, the difference in the coefficients is only 0.5, but the values of the coefficients are much bigger than that in model $P3-P5$. This results in that the difference in two log functions before and after the change points for $P6$ is much larger than that for $P3-P5$ since the units for u is 1000 for $P6$, which yields the extremely high detection power and level of accuracy.

For logistic regression models $B3-B7$, we derive both the SCAD estimator and the MCP estimator for illustration purpose. From the results in Table 3.1 and Table 3.2, it is easy to see that both of the SCAD estimator and the MCP estimator perform well in estimating the change points. In Table 3.1 and Table 3.2, Smax refers to the test proposed in Antoch, et al. [2004] under the extra assumption that there is at most one change point in the simulated data sequence. With this extra information, the test performs slightly better than the proposed algorithm in terms of detecting correct number of change points.

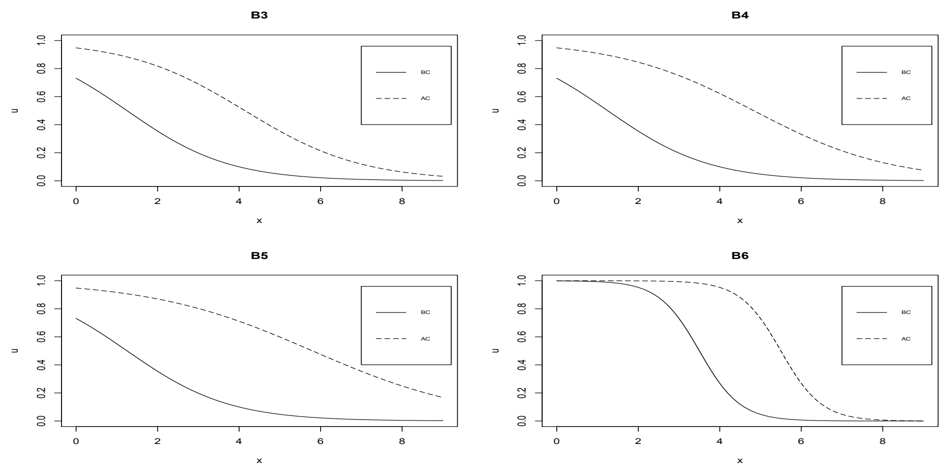

Figure 3.1: The plots of two logistic functions before (BC) and after (AC) the change point for each of models B3-B6

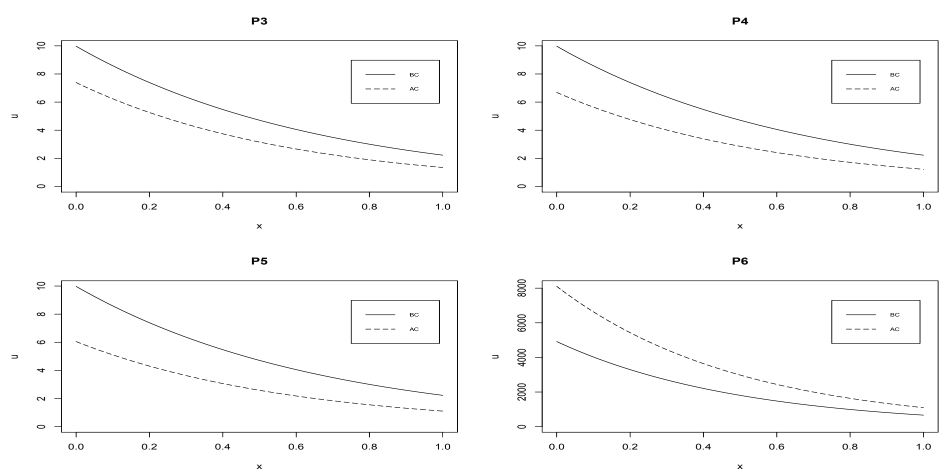

Figure 3.2: The plots of two log functions before (BC) and after (AC) the change point for each of models P3-P6

Table 3.1: Simulation results based on 1000 simulations for $B3 - B7$

	$\lvert\tilde{\epsilon}_{M_j}\rvert$			$Acc(500, 10)$			$Acc(500, 15)$		
M_j	Smax	scad	mcp	Smax	scad	mcp	Smax	scad	mcp
B3	1000	957	943	843	846	843	931	931	926
B4	1000	939	913	973	972	970	994	993	991
B5	1000	962	934	921	919	916	964	962	959
B6	1000	910	942	903	901	901	951	948	951
B7	1000	928	761	1000	994	987	996	1000	997

Table 3.2: Simulation results based on 1000 simulations for $P3 - P7$

	$\lvert\tilde{\epsilon}_{M_j}\rvert$		$Acc(500, 10)$		$Acc(500, 15)$	
M_j	Smax	scad	Smax	scad	Smax	scad
P3	1000	960	859	837	924	916
P4	1000	968	924	905	970	965
P5	1000	975	956	942	986	986
P6	1000	920	1000	1000	1000	1000
P7	1000	925	907	881	959	948

3.3.4 GLMs with Multiple Change Points

The performance of the proposed algorithm is also evaluated in this subsection through Monte Carlo simulations for GLMs with multiple change points. We will estimate how frequently the algorithm detects the correct number of change points and how accurately it estimates the change points when they do exist. We consider the following four models. $B8 - B9$ are for the binomial response and $P8 - P9$ are for the Poisson response.

$B8 : \log \frac{\mu_t}{1-\mu_t} = -0.73+0.14x_t+(2.02+1.34x_t)I_{[513,769]}(t)-(2.15+1.57x_t)I_{[770,1000]}(t).$

$B9 : \log \frac{\mu_t}{1-\mu_t} = 1.58 - 0.79x_t - (2.04 - 0.90x_t)I_{[1428,10000]}(t)$

$+ (2.25 - 0.07x_t)I_{[3085,10000]}(t) - 2.86I_{[4503,10000]}(t) + (1.66 - 0.02x_t)I_{[5913,10000]}(t)$

$- (0.59 + 0.79x_t)I_{[7422,10000]}(t) + (0.67 + 1.27x_t)I_{[8804,10000]}(t).$

$P8 : \log(\mu_t) = 0.31 - 0.11x_t + 0.91I_{[513,769]}(t) - (0.64 - 0.01x_t)I_{[770,1000]}(t).$

$P9 : \log(\mu_t) = 1.58 - 0.79x_t - (2.04 - 0.90x_t)I_{[1428,10000]}(t)$

$+(0.95-0.18x_t)I_{[3085,10000]}(t)-(1.06+0.12x_t)I_{[4503,10000]}(t)+(0.95+0.41x_t)I_{[5913,10000]}(t)$

$- (0.88 + 0.39x_t)I_{[7422,10000]}(t) + (0.87 + 0.30x_t)I_{[8804,10000]}(t).$

Both $B8$ and $P8$ contain two change points located at $t = 512$ and $t = 769$ respectively. Both $B9$ and $P9$ contain 6 change points at $t = 1427, 3084, 4502, 5912, 7421, 8803$ respectively. First, we generate x_t from the uniform distribution $\mathrm{U}(0, 9)$ for $B8 - B9$ and $\mathrm{U}(0, 1)$ for $P8 - P9$, then we generate y_t according to each model for $t =$

$1, 2, \ldots, n$, with $n = 1,000$ for $B8$ and $P8$ and $n = 10,000$ for $B9$ and $P9$. The accuracy of the change point estimates is calculated based on 1000 independent simulations. The results are reported in Table 3.3 for $B8-B9$ and Table 3.4 for $P8-P9$. From the table, it can be seen that our algorithm has a high power in detecting the correct number of multiple change points and a high accuracy in estimating them.

Table 3.3: Simulation results based on 1000 simulations for $B8$ and $B9$

	$\lvert\tilde{\epsilon}_{M_j}\rvert$							
M_j	scad	mcp		scad	mcp		scad	mcp
B8	927	927	$Acc(512, 10)$	916	971	$Acc(512, 15)$	931	988
			$Acc(769, 10)$	994	999	$Acc(769, 15)$	995	1000
B9	824	723	$Acc(1427, 10)$	914	915	$Acc(1427, 15)$	955	956
			$Acc(3084, 10)$	882	884	$Acc(3084, 15)$	933	934
			$Acc(4502, 10)$	986	988	$Acc(4502, 15)$	992	994
			$Acc(5913, 10)$	856	850	$Acc(5913, 15)$	924	920
			$Acc(7422, 10)$	993	993	$Acc(7422, 15)$	998	998
			$Acc(8804, 10)$	957	972	$Acc(8804, 15)$	957	972

Table 3.4: Simulation results based on 1000 simulations for $P8$ and $P9$

M_j	$\lvert\tilde{\epsilon}_{M_j}\rvert$		scad		scad
P8	973	$Acc(512, 10)$	922	$Acc(512, 15)$	958
		$Acc(769, 10)$	885	$Acc(769, 15)$	942
P9	873	$Acc(1427, 10)$	995	$Acc(1427, 15)$	998
		$Acc(3084, 10)$	965	$Acc(3084, 15)$	986
		$Acc(4502, 10)$	990	$Acc(4502, 15)$	998
		$Acc(5913, 10)$	997	$Acc(5913, 15)$	1000
		$Acc(7422, 10)$	982	$Acc(7422, 15)$	998
		$Acc(8804, 10)$	986	$Acc(8804, 15)$	986

3.4 A Real Data Application

In this section, we apply our algorithm on the Bike Sharing data set which contains the hourly counts of rental bikes in years 2011 and 2012 at Washington, D.C., USA. There are three reasons for which we think this data set fits our method. Firstly, the hourly count of rental bikes can be assumed to follow a Poisson distribution which describes such phenomenons. Secondly, the data set has been used in Fanaee-T and Gama [2014] for event labeling which is a process of marking unusual data points as events. Their results show that there are lots of events marked in the hourly counts of rental bikes. So it is suspectable that there exist change points in the mean hourly counts of rental bikes. Our method is applicable to detect those changes. Lastly, there are other variables such as hourly temperatures and hourly measurements of humidity in the data set which might provide some justifications of the changes.

The time series of hourly counts including $17,379$ hours is plotted in Figure 4.2 (upper panel). There are 16 change points in the series detected by our algorithm which are indicated by the vertical lines in Figure 4.2 (upper panel). So the whole time period is divided into 17 intervals by these 16 change points. The means of both the standardized hourly temperatures and the standardized hourly humidity within each time interval separated by the change points are also plotted in Figure

4.2 (lower panel). From Figure 4.2, we can see that for most of the time intervals, the changes in the means of the hourly counts for rental bikes conform with the changes in the means of the hourly temperatures within each time interval. However, for only two time intervals, the 4^{th} and 13^{th} intervals, the count of rental bikes drops while the mean of hourly temperatures increases. We suspect that, in those two time intervals, the increases of the mean of hourly temperatures and the drops of the mean of hourly humidity together caused the drops in the rental counts.

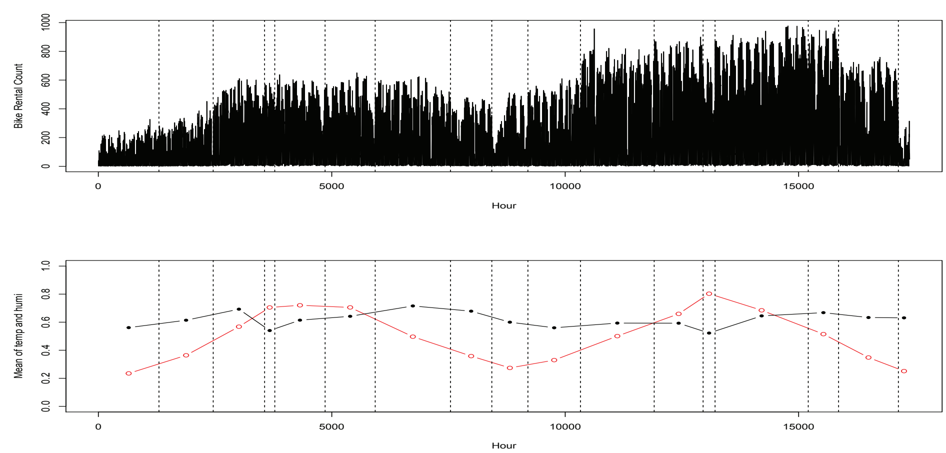

Figure 3.3: The time series plot of the hourly rental bike counts together with the change points (upper panel) and the mean of hourly standardized temperature and hourly standardized humidity within each time interval separated by the change points (lower panel)

4 Nonparametric Change-point Estimators based on Empirical Characteristic Functions

Nonparametric methods play a big role in tackling the problem of a change point in distributions of a data sequence. Most of the nonparametric methods are based either on empirical distributions, U-statistics or quantile functions [Carlstein, 1988, Csörgő and Horváth, 1997, Rafajlowicz, *et al.* 2010, Holmes, *et al.* 2013]. Another nonparametric tool is the empirical characteristic function (ECF). The definition of the ECF was given by Paren [1962]. Kent [1975] studied the weak convergence theorem of the ECF. Since then, the ECF has been applied to solve various statistical problems such as hypodissertation testing for symmetry about the origin, dependence or normality [Feuerverger and Mureika, 1977, Kankainen and Ushakov, 1998, Ushakov, 1999, Epps, 1999, and Koutrouvelis and Meintanis, 1999].

Hušková and Meintanis [2006] proposed a class of test statistics based on the ECF to test if there is a change point in distributions of a sequence of independent random variables. They gave two choices of the weight function for their proposed statistics. They studied the limiting behaviour of the test statistics under both null and alternative hypotheses. Built upon their statistics, a change point estimator is given in this chapter for the same change point problem. The weight function $\omega(t;a)$ under consideration includes the two weight functions from Hušková and Meintanis [2006] plus the weight function used in Matteson and James [2014], where a is a tuning parameter. We will study the consistency of this estimator when the difference between the distributions before and after the change point tends to zero as the sample size goes to infinity.

Simulation results in Hušková and Meintains [2006] showed that the test statistics are robust with respect to the value of the tuning parameter a in the weight function, which, however, is selected from 1 to 4 increased by 1 each time in their simulation study. It is noted that the domain of a in their weight functions ranges from 0 to infinity. The real data example reveals that the change point estimate may be influenced significantly by the value of the tuning parameter a (see Table 4.1 of section 4.3). Thus, accuracy of the change point estimate is in question. To tackle this problem, we propose an algorithm for selecting an appropriate value of a, a_{s}, in order to obtain a change point estimate with a satisfactory accuracy.

The rest of the chapter is organized as follows: In section 4.1, we propose a nonparametric change point estimator in the distributions of a sequence of independent observations in terms of the test statistics given in Hušková and Meintanis (2006) that are based on weighted empirical characteristic functions. In section 4.2, we investigate the asymptotic properties of this estimator assuming that there exists one change point in the data sequence. We present an algorithm for selecting a value a_{s}, for the tuning parameter a which is also justified in section 4.3. We carry out simulation study to evaluate the performance of the change point estimation with use of a_{s} in section 4.4. A real data example is also given there. The proofs of all the theorems are given in the appendix.

4.1 The Change Point Estimator based on the ECF

Let $Y_{n,1}, Y_{n,2}, \ldots, Y_{n,n}$ be a sequence of independent random variables where $Y_{n,j}$ has a distribution function $F_{n,j}$, $j = 1, 2, \ldots, n$. Consider the testing problem

$$H_0 : F_1 = F_{n,1} = F_{n,2} = \cdots = F_{n,n},$$

against

$$H_1 : F_1 = F_{n,1} = \cdots = F_{n,k_0^{(n)}} \neq F_{n,k_0^{(n)}+1} = \cdots = F_{n,n} = F_n, \quad \text{for } k_0^{(n)} < n \quad (4.1)$$

where $k_0^{(n)}$, F_1 and F_n are unknown. $k_0^{(n)}$ is called the change point. For the sake of convenience, the subscript n in $Y_{n,j}$ and $F_{n,j}$ and the superscript n in $k_0^{(n)}$ are all suppressed if there is no confusion.

Hušková and Meintains [2006] developed the following class of test statistics based on the empirical characteristic function and a non-negative weight function $\omega(.)$ with a non-negative tuning parameter a:

$$T_{\omega,\gamma}(k) = \left(\frac{k(n-k)}{n^2}\right)^{\gamma} \frac{k(n-k)}{n} \int_{-\infty}^{\infty} |\phi_k(t) - \phi_k^0(t)|^2 \omega(t) dt, \tag{4.2}$$

where $\gamma \in (0,1]$, $\omega(\cdot)$ satisfies that $0 < \int \omega(t)dt < \infty$, $\phi_k(t)$ and $\phi_k^0(t)$ are ECFs based on $Y_1, \ldots, Y_k$ and $Y_{k+1}, \ldots, Y_n$, respectively, i.e.,

$$\phi_k(t) = \frac{1}{k}\sum_{j=1}^{k} \exp\{itY_j\}, \qquad \phi_k^0(t) = \frac{1}{n-k}\sum_{j=k+1}^{n} \exp\{itY_j\}, \quad j = 1, 2, \ldots, n.$$

Under the alternative hypodissertation, we propose the change point estimator for k_0 as

$$\hat{k}_n = \arg\max_{1 \le k < n} T_{\omega,\gamma}(k). \tag{4.3}$$

Some choices of $\omega(\cdot)$ are

$$\omega_1(t;a) = \frac{1}{2a}\exp\{-a|t|\}, \quad t \in \mathcal{R}^1, \quad a > 0, \tag{4.4}$$

$$\omega_2(t;a) = \frac{\sqrt{a}}{\sqrt{\pi}}\exp\{-at^2\}, \quad t \in \mathcal{R}^1, \quad a > 0, \tag{4.5}$$

or

$$\omega_3(t;a) = \frac{a\ 2^a \Gamma(\frac{1+a}{2})}{2\sqrt{\pi}\Gamma(1-\frac{a}{2})}|t|^{-a-1}, \quad t \in \mathcal{R}^1, \quad a \in (0,2). \tag{4.6}$$

We remark that $\omega_1(t;a)$ and $\omega_2(t;a)$ were given in Hušková and Meintains [2006] while $\omega_3(t;a)$ was used as the weight function in Matteson and James [2014] for obtaining their nonparametric change point estimator in distributions of a sequence of multivariate random variables.

We assume that k_0 satisfies

$$k_0 = \lfloor n\tau_0 \rfloor, \quad \tau_0 \in [\kappa_1, \kappa_2] \quad \text{for some} \quad 0 < \kappa_1 \le \kappa_2 < 1. \tag{4.7}$$

This is a conventional assumption made in change point detection problems [Csörgő & Horváth, 1997]. The estimator for τ_0 is given by

$$\hat{\tau}_n = \frac{\hat{k}_n}{n} = \frac{1}{n} \arg \max_{1 \le k < n} T_{\omega,\gamma}(k). \tag{4.8}$$

4.2 Consistency of the Change Point Estimator

Define

$$\begin{aligned} \Delta_n &= \int \left\{ \left(\int \cos(tx) d(F_1(x) - F_n(x)) \right)^2 + \left(\int \sin(tx) d(F_1(x) - F_n(x)) \right)^2 \right\} \omega(t) dt \\ &= E[h(Y_1, Y_2)] - 2E[h(Y_1, Y_{k_0+1})] + E[h(Y_{k_0+1}, Y_{k_0+2})], \end{aligned} \tag{4.9}$$

and $h(x,y) = \int \cos(t(x-y))\omega(t)dt$. In this section, we will study consistency of the change point estimator $\hat{\tau}_n$ under the assumption that $\Delta_n \to 0$. Denote

$$\begin{aligned}
\tilde{h}(Y_r, Y_s) &= h(Y_r, Y_s) - E[h(Y_r, Y_s)|Y_r] - E[h(Y_r, Y_s)|Y_s] + E[h(Y_r, Y_s)],\\
\overline{h}(Y_r, Z_1) &= E[h(Y_r, Z_1)|Y_r] - E[h(Y_r, Z_1)],\\
\overline{h}(Y_r, Z_2) &= E[h(Y_r, Z_2)|Y_r] - E[h(Y_r, Z_2)], \qquad (4.10)
\end{aligned}$$

where Z_1 and Z_2 are independent of $Y_1, Y_2, \ldots, Y_n$ and follow the distributions F_1 and F_n respectively. To simplify the notation, $T_{\omega,\gamma}(k)$ is abbreviated by $T(k)$. The theorem is given as follows, and its proof is also provided.

Theorem 4.2.1 *Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of independent random variables, where $Y_1, \ldots, Y_{k_0}$ have a common distribution function F_1, and $Y_{k_0+1}, \ldots, Y_n$ have a common distribution function F_n. Assume that k_0 satisfies* (4.7) *and $\gamma \in (0,1]$. If Δ_n defined in* (4.9) *satisfies that $\Delta_n \longrightarrow 0$ and*

$$n\Delta_n^2 \to \infty, \qquad as \;\; n \longrightarrow \infty, \qquad (4.11)$$

then, as $n \to \infty$,

$$\hat{\tau}_n \longrightarrow_P \tau_0. \qquad (4.12)$$

Proof: Since $T(k) \le |T(k) - ET(k)| + ET(k)$, and $ET(k_0) \le |ET(k_0) - T(k_0)| + T(k_0)$, by the triangle inequality, it is easy to show that

$$ET(k_0) - ET(k) \le 2 \max_{1 \le k < n} |T(k) - ET(k)| + T(k_0) - T(k). \qquad (4.13)$$

Let $c_{k,n}(\gamma) = \left(\frac{k(n-k)}{n^2}\right)^{\gamma} \frac{k(n-k)}{n}$, $\quad k = 1, 2, \ldots, n-1$, then $T(k) = c_{k,n}(\gamma)Q_k$, where

$$Q_k = \frac{1}{k^2}\sum_{r,s=1}^{k} h(Y_r, Y_s) + \frac{1}{(n-k)^2}\sum_{r,s=k+1}^{n} h(Y_r, Y_s) - \frac{2}{k(n-k)}\sum_{r=1}^{k}\sum_{s=k+1}^{n} h(Y_r, Y_s) \tag{4.14}$$

For $k \le k_0$, Q_k can be decomposed as follows:

$$\begin{aligned} Q_k \;=\; & \frac{1}{k^2}\sum_{r=1}^{k} h(Y_r, Y_r) + \frac{1}{(n-k)^2}\sum_{r=k+1}^{n} h(Y_r, Y_r) + \frac{1}{k^2}\sum_{r=1}^{k}\sum_{s=1, s\neq r}^{k} h(Y_r, Y_s) \\ & + \frac{1}{(n-k)^2}\left[\sum_{r=k+1}^{k_0}\sum_{s=k+1, s\neq r}^{k_0} + \sum_{r=k_0+1}^{n}\sum_{s=k_0+1, s\neq r}^{n} + 2\sum_{r=k+1}^{k_0}\sum_{s=k_0+1}^{n}\right] h(Y_r, Y_s) \\ & - \frac{2}{k(n-k)}\sum_{r=1}^{k}\left[\sum_{s=k+1}^{k_0} + \sum_{s=k_0+1}^{n}\right] h(Y_r, Y_s). \end{aligned}$$

So

$$\begin{aligned} EQ_k \;=\; & \frac{n}{k(n-k)}\int \omega(t)dt + \frac{(n-k_0)^2}{(n-k)^2}\left[E[h(Y_1, Y_2)] - 2E[h(Y_1, Y_{k_0+1})] + E[h(Y_{k_0+1}, Y_{k_0+2})]\right] \\ & + \left[\frac{k-k_0}{(n-k)^2} - \frac{1}{k}\right] E[h(Y_1, Y_2)] - \frac{n-k_0}{(n-k)^2}E[h(Y_{k_0+1}, Y_{k_0+2})], \end{aligned} \tag{4.15}$$

where

$$E[h(Y_1, Y_2)] = \int \left\{\left(\int \cos(tx)dF_1(x)\right)^2 + \left(\int \sin(tx)dF_1(x)\right)^2\right\}\omega(t)dt,$$

and

$$E[h(Y_{k_0+1}, Y_{k_0+2})] = \int \left\{\left(\int \cos(tx)dF_n(x)\right)^2 + \left(\int \sin(tx)dF_n(x)\right)^2\right\}\omega(t)dt.$$

Then we have, as $k \le k_0$,

$$
\begin{aligned}
ET(k) - ET(k_0) &= \left[\left(\frac{k(n-k)}{n^2}\right)^\gamma - \left(\frac{k_0(n-k_0)}{n^2}\right)^\gamma\right] \int \omega(t)dt \\
&+ \left[\left(\frac{k(n-k)}{n^2}\right)^\gamma \frac{k(n-k_0)}{n-k} - \left(\frac{k_0(n-k_0)}{n^2}\right)^\gamma k_0\right] \frac{(n-k_0)}{n}\Delta_n \\
&+ \left[\left(\frac{k(n-k)}{n^2}\right)^\gamma \left(\frac{k(k-k_0)}{n(n-k)} - \frac{n-k}{n}\right) + \left(\frac{k_0(n-k_0)}{n^2}\right)^\gamma \frac{n-k_0}{n}\right] E[h(Y_1, Y_2)] \\
&- \left[\left(\frac{k(n-k)}{n^2}\right)^\gamma \frac{k(n-k_0)}{n(n-k)} - \left(\frac{k_0(n-k_0)}{n^2}\right)^\gamma \frac{k_0}{n}\right] E[h(Y_{k_0+1}, Y_{k_0+2})]. \qquad (4.16)
\end{aligned}
$$

It is easy to conclude that from (4.11) the second term is the dominating one in (4.16). Using the mean value theorem, we obtain that

$$ET(k) - ET(k_0) = g_1'(\xi_1)(\tau - \tau_0)n\Delta_n + o_p(n\Delta_n), \qquad (4.17)$$

where $g_1'(\cdot)$ is the first order derivative of $g_1(\cdot)$ with $g_1(x) = (1-\tau_0)^2 x^{\gamma+1}(1-x)^{\gamma-1}$, and $\tau \le \xi_1 \le \tau_0$. Similar arguments yield that, as $k > k_0$

$$ET(k) - ET(k_0) = g_2'(\xi_2)(\tau - \tau_0)n\Delta_n + o_p(n\Delta), \qquad (4.18)$$

where $g_2'(\cdot)$ is the first order derivative of $g_2(\cdot)$ with $g_2(x) = \tau_0^2 x^{\gamma-1}(1-x)^{\gamma+1}$, and $\tau_0 \le \xi_2 \le \tau$. Combining (4.13), (4.16)-(4.18), we obtain that

$$
\begin{aligned}
n\Delta_n|\tau - \tau_0|\delta + o_p(n\Delta_n) &\le ET(k_0) - ET(k) \\
&\le 2 \max_{1 \le k < n} |T(k) - ET(k)| + T(k_0) - T(k), \quad (4.19)
\end{aligned}
$$

where $\delta = \min\{g_1'(\xi_1), g_2'(\xi_2)\}$. Since $\hat{\tau}_n = \hat{k}_n/n$, $T(\hat{k}_n) \ge T(k_0)$, and T is nonnegative, by replacing τ by $\hat{\tau}_n$ in (4.19), we have

$$n\Delta_n|\hat{\tau}_n - \tau_0|\delta + o_p(n\Delta_n) \le 2 \max_{1\le k<n} |T(k) - ET(k)|. \tag{4.20}$$

In order to show the consistency of change point estimator $\hat{\tau}_n$, we consider the probability $P\left(|\hat{\tau}_n - \tau_0| > \varepsilon\right),\ \forall\varepsilon > 0$. It is easily to see from (4.20) that

$$\begin{aligned} P\left(|\hat{\tau}_n - \tau_0| > \varepsilon\right) \ &\le\ P\left(\max_{1\le k<k_0} |T(k) - ET(k)| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \\ &\quad +P\left(\max_{k_0<k<n} |T(k) - ET(k)| > \frac{n\varepsilon\delta\Delta_n}{2}\right). \end{aligned} \tag{4.21}$$

Because of the symmetry, we only show $P\left(\max_{1\le k\le k_0} |T(k) - ET(k)| > \dfrac{n\varepsilon\delta\Delta_n}{2}\right) \to 0$ as $n \to \infty$. The remaining part is analogous and thus is omitted.

We start with that $P\left(\max_{1\le k\le k_0} |T(k) - ET(k)| > \dfrac{n\varepsilon\delta\Delta_n}{2}\right)$. If $k \le k_0$, by (4.14),

$$T(k) - ET(k) = A_1 + A_2 + \cdots + A_{12}, \tag{4.22}$$

with

$$A_1 = \left(\tfrac{k(n-k)}{n^2}\right)^\gamma \tfrac{1}{k} \sum_{r=1}^{k} \sum_{s=1,s\ne r}^{k} \tilde{h}(Y_r, Y_s), \quad A_2 = \left(\frac{k(n-k)}{n^2}\right)^\gamma \frac{1}{n-k} \sum_{r=k+1}^{n} \sum_{s=k+1,s\ne r}^{n} \tilde{h}(Y_r, Y_s),$$

$$A_3 = \left(\tfrac{k(n-k)}{n^2}\right)^\gamma \tfrac{1}{n} \sum_{r=1}^{n} \sum_{s=1,s\ne r}^{n} \tilde{h}(Y_r, Y_s), \quad A_4 = \left(\frac{k(n-k)}{n^2}\right)^\gamma \frac{2(n-k_0)}{n} \sum_{r=1}^{k} \overline{h}(Y_r, Z_1),$$

$$A_5 = -\left(\tfrac{k(n-k)}{n^2}\right)^\gamma \tfrac{2(n-k)}{nk} \sum_{r=1}^{k} \overline{h}(Y_r, Z_1), \quad A_6 = -\left(\frac{k(n-k)}{n^2}\right)^\gamma \frac{2k(n-k_0)}{n(n-k)} \sum_{r=k+1}^{k_0} \overline{h}(Y_r, Z_1),$$

$$A_7 = -\left(\tfrac{k(n-k)}{n^2}\right)^\gamma \tfrac{2k}{n(n-k)} \sum_{r=k+1}^{k_0} \overline{h}(Y_r, Z_1), \quad A_8 = \left(\frac{k(n-k)}{n^2}\right)^\gamma \frac{2k(n-k_0)}{n(n-k)} \sum_{r=k_0+1}^{n} \overline{h}(Y_r, Z_2),$$

$$A_9 = -\left(\tfrac{k(n-k)}{n^2}\right)^\gamma \tfrac{2k}{n(n-k)} \sum_{r=k_0+1}^{n} \overline{h}(Y_r, Z_2), \quad A_{10} = \left(\frac{k(n-k)}{n^2}\right)^\gamma \frac{2k(n-k_0)}{n(n-k)} \sum_{r=k+1}^{k_0} \overline{h}(Y_r, Z_2),$$

$$A_{11} = -\left(\tfrac{k(n-k)}{n^2}\right)^\gamma \tfrac{2k(n-k_0)}{n(n-k)} \sum_{r=k_0+1}^{n} \overline{h}(Y_r, Z_1), \quad A_{12} = -\left(\frac{k(n-k)}{n^2}\right)^\gamma \frac{2(n-k_0)}{n} \sum_{r=1}^{k} \overline{h}(Y_r, Z_2),$$

where Z_1 and Z_2 have the distribution functions F_1 and F_n, respectively, and are independent of $Y_1, Y_2, \ldots, Y_n$.

Next we investigate each term in (4.22). Towards this end, we consider the following statistics

$$S_k(\tilde{h}) = \sum_{1 \le i < j \le k} \tilde{h}(Y_i, Y_j), \quad k = 1, 2, \ldots, n,$$

where $\tilde{h}$ is defined in (4.10). Since $E\left[S_{k+1}(\tilde{h}) | Y_1, Y_2, \ldots, Y_k\right] = S_k(\tilde{h})$ for $k = 1, 2, \ldots, n-1$, $\{S_k, \sigma(Y_1, \ldots, Y_k); k = 1, 2, \ldots, n\}$ is a martingale, where $\sigma(Y_1, \ldots, Y_k)$ denotes the $\sigma-$field generated by $Y_1, \ldots, Y_k$. Then by the Hájek-Rényi-Chow inequality

$$\begin{aligned} &P\left(\max_{1 \le k \le k_0} |A_1| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le P\left(\max_{1 \le k \le k_0} \frac{|S_k(\tilde{h})|}{k^{1-\gamma}} > \frac{n^{1+\gamma}\varepsilon\delta\Delta_n}{4}\right) \\ &\le \frac{c}{n^2\varepsilon^2\delta^2\Delta_n^2} \left\{\frac{1 + I_{\{\gamma=1/2\}} \log n}{n^{\min(2\gamma,1)}}\right\} \le \frac{c}{n^2\varepsilon^2\delta^2\Delta_n^2}. \end{aligned}$$

Similar arguments yield that

$$P\left(\max_{1 \le k \le k_0} |A_2| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n^2\varepsilon^2\delta^2\Delta_n^2},$$

and

$$P\left(\max_{1\le k\le k_0}|A_3| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n^2\varepsilon^2\delta^2\Delta_n^2}.$$

Since each of $\{E(h(Y_r, Z_1)|Y_r) - Eh(Y_r, Z_1),\ r = 1, 2, \ldots, k_0\}$, $\{E(h(Y_r, Z_1)|Y_r) - Eh(Y_r, Z_1),\ r = k_0 + 1, \ldots, n\}$, $\{E(h(Y_r, Z_2)|Y_r) - Eh(Y_r, Z_2),\ r = 1, 2, \ldots, k_0\}$, and $\{E(h(Y_r, Z_2)|Y_r) - Eh(Y_r, Z_2),$

$r = k_0 + 1, \ldots, n\}$ is an identically distributed and independent sequence of random variables with zero mean and finite variance, the application of the Hájiek-Rényi-Chow inequality leads to

$$\begin{aligned}
&P\left(\max_{1\le k\le k_0}|A_4| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n\varepsilon^2\delta^2\Delta_n^2},\\
&P\left(\max_{1\le k\le k_0}|A_5| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n^{2+2\gamma}\varepsilon^2\delta^2\Delta_n^2}\sum_{k=1}^{m}\frac{1}{k^{2-2\gamma}} \le \frac{c}{n^2\varepsilon^2\delta^2\Delta_n^2}.
\end{aligned}$$

Similarly, we can obtain that

$$\begin{aligned}
&P\left(\max_{1\le k\le k_0}|A_6| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n\varepsilon^2\delta^2\Delta_n^2}, \quad P\left(\max_{1\le k\le k_0}|A_7| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n^2\varepsilon^2\delta^2\Delta_n^2}\\
&P\left(\max_{1\le k\le k_0}|A_8| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n\varepsilon^2\delta^2\Delta_n^2}, \quad P\left(\max_{1\le k\le k_0}|A_9| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n^2\varepsilon^2\delta^2\Delta_n^2},\\
&P\left(\max_{1\le k\le k_0}|A_{10}| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n\varepsilon^2\delta^2\Delta_n^2}, \quad P\left(\max_{1\le k\le k_0}|A_{11}| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n\varepsilon^2\delta^2\Delta_n^2},\\
&P\left(\max_{1\le k\le k_0}|A_{12}| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c}{n\varepsilon^2\delta^2\Delta_n^2}.
\end{aligned}$$

Thus, we have

$$P\left(\max_{1\le k\le k_0}|T(k) - ET(k)| > \frac{n\varepsilon\delta\Delta_n}{2}\right) \le \frac{c_0}{\varepsilon^2\delta^2 n\Delta_n^2}. \tag{4.23}$$

By (4.11), (4.21) and (4.23), it follows that $\lim_{n\to\infty} P\left(|\hat{\tau}_n - \tau_0| > \varepsilon\right) = 0$, i.e. $\hat{\tau}_n \to_P \tau_0$.

□

4.3 An Algorithm for Selecting an Appropriate Value for a

We now present a real data example to demonstrate how the change point estimate can be affected by the choice of a. Consider the Nile data, a time series of the annual flow of the river Nile at Aswan from 1871 to 1970 [Cobb , 1978, Dumbgen, 1991, Balke, 1993], which has a change in year 1898 corresponding to the 28th observation in the data sequence detected in Zeileis et al. [2003]. The data is depicted in Figure 1. For the purpose of illustration, we assume that the observations are independent as in Cobb [1978]. We use (4.3) with respective weight functions $\omega_1(t;a)$, $\omega_2(t;a)$, and $\omega_3(t;a)$ for different values of a to estimate the change point. The resulted change point estimates are reported in Table 4.1.

It can be seen from Table 4.1 that the value of a has a large impact on the accuracy of the change point estimate. An inappropriate a may result in a misleading estimate. In practice, we have no information about the change point in a given data sequence. However a needs to be prechosen in order to find the change point estimate by (4.3). As shown above, different values of a might result in different change point estimates. Thus it is important to select a value from a set of possible values of a such that the

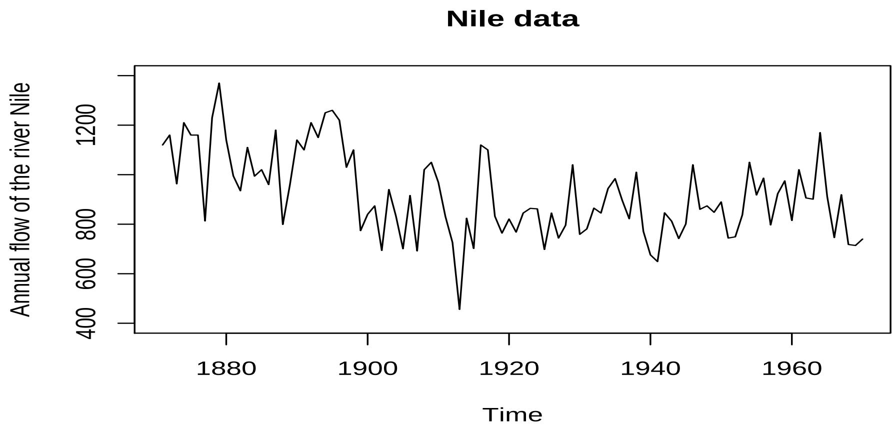

Figure 4.1: The Nile data

Table 4.1: Estimated change point $\hat{k}_n$ using different weight function $\omega(t;a)$ with different values of a and a fixed $\gamma = 0.5$

$\omega_1(t;a)$	a	1	2	3	4	5	6	7	$\cdots$	100
	$\hat{k}_n$	47	48	48	48	48	28	28	$\cdots$	28
$\omega_2(t;a)$	a	1	2	3	$\cdots$	22	23	24	$\cdots$	100
	$\hat{k}_n$	48	48	48	48	48	28	28	$\cdots$	28
$\omega_3(t;a)$	a	0.001	0.002	$\cdots$	0.009	0.01	0.02	0.03	$\cdots$	2
	$\hat{k}_n$	47	47	$\cdots$	48	28	28	28	$\cdots$	28

resulted change point estimate has a satisfactory performance. Such an appropriate choice of a is denoted as a_s in this paper, where the subscript "s" is taken from the first letter of "*s*election". We propose the following algorithm for finding a_s.

Step 1 Let $Y_1, Y_2, \ldots, Y_{k_0}, Y_{k_0+1}, \ldots, Y_n$ be a given data sequence with the change point located at k_0 and $\mathcal{A} = \{a_1, a_2, ..., a_\ell\}$ be a set of possible values for a. For each a_i from the set $\mathcal{A}$, we obtain $\hat{k}_{a_i} = \arg\max_k T_{\gamma,w}(k)$.

Step 2 Compute the mean of $\hat{k}_{a_i}$, $i = 1, 2, \ldots, \ell$ as $\bar{\hat{k}} = \frac{1}{\ell}\sum_{i=1}^{\ell} \hat{k}_{a_i}$. Then $a_\mathrm{s} = \arg\min_{a_i} |\hat{k}_{a_i} - \bar{\hat{k}}|$.

From the proposed algorithm, it can be seen that a_s is dependent on the data sequence and hence random. a_s might not give us the best change point estimate but it will provide an improved performance over a fixed one, which is not only justified in Proposition 4.3.1, but also confirmed by the simulation study in the next section.

Proposition 4.3.1 *Given a data sequence* $Y_1, Y_2, \ldots, Y_{k_0}, Y_{k_0+1}, \ldots, Y_n$ *with the change point located at* k_0 *and* $\mathcal{A} = \{a_1, a_2, ..., a_\ell\}$ *be a set of possible values for* a*. Then there exists at least one point* $a^* \neq a_s$ *in* $\mathcal{A}$ *such that* $|\hat{k}_{a_s} - k_0| \leq |\hat{k}_{a^*} - k_0|$.

Proof: Suppose that $k_0 \geq \bar{\hat{k}}$.

$$|\hat{k}_{a_\mathrm{s}} - k_0| = |\hat{k}_{a_\mathrm{s}} - \bar{\hat{k}} + \bar{\hat{k}} - k_0| \leq |\hat{k}_{a_\mathrm{s}} - \bar{\hat{k}}| + |\bar{\hat{k}} - k_0| \leq |\hat{k}_{a_i} - \bar{\hat{k}}| + |\bar{\hat{k}} - k_0|.$$

The last inequality holds true for any $a_i \in \mathcal{A}$ by the definition of a_s. There always exists at least one point $a^* \neq a_\mathrm{s}$ in $\mathcal{A}$ such that $\hat{k}_{a^*} \leq \min(\hat{k}_{a_\mathrm{s}}, \bar{\hat{k}})$. Therefore,

$$|\hat{k}_{a_\mathrm{s}} - k_0| \leq |\hat{k}_{a^*} - \bar{\hat{k}}| + |\bar{\hat{k}} - k_0| \leq \bar{\hat{k}} - \hat{k}_{a^*} + k_0 - \bar{\hat{k}} = |\hat{k}_{a^*} - k_0|. \tag{4.24}$$

Similarly, we can show (4.24) for the case that $k_0 < \bar{\hat{k}}$. The proof is completed. □

4.4 Numerical Examples

In this section, we carry out a simulation study to investigate the performance of $\hat{k}_n$ obtained via (4.3) when using different values of a including a_s in terms of accuracy of the change point estimate. In addition, we apply (4.3) with $a = a_\mathrm{s}$ to the Nile data.

4.4.1 Simulation Studies

We perform a simulation study to compare the change point estimate obtained via (4.3) using a set of fixed values of a and a_s. The following is the details of the simulation study.

(1) Generate data $Y_1, Y_2, \ldots, Y_{k_0}$ from the distribution F_1 and $Y_{k_0+1}, \ldots, Y_n$ from the distribution F_n with one change point located at $k_0 = 30, 50$, or 70, where $n = 100$. Three cases of F_1 are considered: Case 1: the normal distribution

$N(0,1)$; Case 2: the laplace distribution $L(0,1)$; Case 3: the gamma distribution $G(1,1)$. Correspondingly, we consider $F_n(x) = F_1((x-b)/d)$ for $b=1$, and $d=1$ or $\sqrt{2}$.

(2) For a chosen weight function $\omega(t;a)$ and a given set of possible values of a, say $\mathcal{A}$, first execute the step 1 of the algorithm given in section 4.3 and obtain $\{\hat{k}_a,\ a \in \mathcal{A}\}$, and then execute the step 2 of this algorithm to obtain a_{s}. Compute the change point estimate $\hat{k}_{a_{\mathrm{s}}}$.

(3) Repeat (1)-(2) 1000 times and then compute the number of times that the change point estimate falls into the interval $[k_0-\delta, k_0+\delta]$ for $\delta = 5, 10, 15$.

In this simulation study, γ is set as 0.5, $\mathcal{A}$ is chosen as $\{1,2,3,\ldots,15\}$ for both ω_1 and ω_2 but $\{0.2, 0.4, \ldots, 2\}$ for ω_3. Similarly as in Chapter 3, let $Acc(k_0,\delta)$ denote the number of $\hat{k}_a$ out of 1000 that fell into the interval centered at k_0 with length 2δ. The simulation results are reported in Table 4.2 to 4.19, which show that the value of a has a large impact on the accuracy of the change point estimate for all three weight functions. From these tables, it can been seen that the change point estimate obtained by using a_{s} always outperforms the change point estimates obtained by using some values of a, and has the best performance in some cases. It can also be observed that the weight function ω_3 performed better than both ω_1 and ω_2 in terms of the accuracy of change point estimation overall.

We know from Hušková and Meintanis [2006] that the role of the tuning parameter a is to control the rate of decay of the weight function. We remark that for simple presentation, we have only presented the simulation results for using $a \leq 11$. As a matter of fact, the accuracy of the change point estimate using $a > 11$ is almost the same as the one using $a = 11$ for the weight function being ω_1 or ω_2, and the change point estimates using either ω_1 or ω_2 perform similarly when a goes to infinity.

Table 4.2: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_1 when F_1 is $N(0,1)$ and F_n is $N(1,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	706	733	743	752	755	761	762	762	760	762	761	752
		873	899	904	907	905	908	909	909	907	908	909	906
		927	941	944	949	951	951	951	951	950	951	951	951
ω_1	$k_0 = 50$	725	749	763	771	772	770	770	772	772	772	773	773
		895	915	928	931	934	931	930	930	930	930	930	935
		964	970	970	971	972	973	973	973	973	973	973	972
	$k_0 = 70$	691	730	742	744	742	744	745	745	745	745	745	740
		856	872	879	887	888	891	891	891	891	891	891	888
		926	935	942	944	942	942	940	937	937	936	937	942

Table 4.3: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_2 when F_1 is $N(0,1)$ and F_n is $N(1,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	734	745	753	754	762	761	762	762	762	760	760	760
		899	906	906	905	909	908	909	909	909	908	907	909
		940	948	949	950	952	951	951	951	951	950	950	952
ω_2	$k_0 = 50$	754	765	769	772	770	771	768	770	772	772	772	770
		915	927	930	934	932	931	930	930	930	930	930	932
		970	969	971	972	972	972	973	973	973	973	973	972
	$k_0 = 70$	725	741	741	742	743	744	745	746	745	745	745	744
		868	880	885	887	890	891	891	892	891	890	890	891
		933	944	942	942	941	940	939	937	937	937	937	942

Table 4.4: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_3 when F_1 is $N(0,1)$ and F_n is $N(1,1)$.

	a	0.2	0.4	0.5	0.7	0.9	1.1	1.3	1.5	1.6	1.8	2	a_s
	$k_0 = 30$	676	703	710	721	731	738	747	740	748	750	745	733
		825	847	854	864	867	874	881	880	880	884	889	872
		894	911	915	919	919	926	934	936	933	937	936	928
ω_3	$k_0 = 50$	773	775	786	792	796	797	802	803	805	807	809	799
		931	934	938	946	950	953	952	950	950	951	952	954
		976	974	974	976	977	980	981	980	981	981	980	981
	$k_0 = 70$	680	706	715	718	721	733	745	748	742	743	744	734
		836	850	861	866	870	878	886	890	891	891	892	879
		912	925	934	937	938	942	945	950	949	949	950	944

Table 4.5: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_1 when F_1 is $N(0,1)$ and F_n is $N(1,2)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	636	643	630	613	599	580	566	549	542	533	528	600
		819	824	809	794	784	771	753	735	729	719	711	777
		890	895	878	863	850	841	825	812	805	797	790	848
ω_1	$k_0 = 50$	727	735	717	706	688	667	653	648	638	625	615	686
		895	904	890	881	867	853	846	835	828	814	806	879
		953	956	955	948	938	932	925	919	915	909	901	942
	$k_0 = 70$	730	762	767	755	742	724	714	703	681	675	673	743
		901	921	915	902	887	873	867	856	848	842	840	896
		950	962	963	952	945	942	936	929	921	917	917	947

Table 4.6: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_2 when F_1 is $N(0,1)$ and F_n is $N(1,2)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	647	638	624	610	596	580	578	568	560	550	546	596
		827	817	803	791	778	771	766	754	745	737	735	781
		897	887	868	858	846	841	836	825	821	813	811	848
ω_2	$k_0 = 50$	740	718	712	701	683	680	661	651	651	644	644	682
		907	894	885	875	865	862	851	845	843	835	832	865
		958	959	950	943	938	936	930	926	925	920	919	939
	$k_0 = 70$	767	769	763	752	737	727	721	712	706	698	691	738
		925	920	908	899	886	876	874	866	859	856	853	890
		964	965	958	950	946	944	942	936	934	929	926	947

Table 4.7: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_3 when F_1 is $N(0,1)$ and F_n is $N(1,2)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	617	636	629	617	609	588	564	551	525	502	470	580
		789	799	789	779	775	755	733	716	689	666	633	753
		867	868	857	849	844	831	811	796	772	745	716	831
ω_3	$k_0 = 50$	749	743	739	725	708	701	687	670	642	605	567	698
		910	905	903	885	877	871	856	849	822	786	752	869
		966	960	958	946	939	934	921	915	896	870	843	936
	$k_0 = 70$	743	757	762	762	760	748	728	717	686	665	636	753
		884	900	903	906	905	900	891	881	859	841	810	902
		938	948	949	954	957	954	954	949	928	916	898	956

Table 4.8: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_1 when F_1 is $L(0,1)$ and F_n is $L(1,1)$, the distribution of $Y+1$ with $Y \sim L(0,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	676	680	667	652	646	639	630	622	616	614	613	640
		830	841	834	830	823	819	812	805	801	798	798	823
		896	906	900	900	897	893	889	882	880	877	876	900
ω_1	$k_0 = 50$	702	718	702	689	682	674	673	667	664	663	659	687
		885	890	880	868	862	855	853	851	847	846	844	870
		952	947	938	935	934	930	933	930	925	925	923	942
	$k_0 = 70$	658	670	666	670	662	653	653	649	643	639	634	661
		829	835	827	822	818	814	813	811	804	801	801	820
		904	903	896	895	895	889	887	885	881	878	879	899

Table 4.9: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_2 when F_1 is $L(0,1)$ and F_n is $L(1,1)$, the distribution of $Y+1$ with $Y \sim L(0,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	674	665	646	645	642	638	635	629	623	621	618	639
		835	836	824	823	819	815	813	808	804	803	801	819
		902	901	898	900	895	891	890	887	883	881	879	896
ω_2	$k_0 = 50$	708	697	696	687	678	677	674	668	670	666	666	681
		882	880	868	865	859	856	855	853	852	848	848	862
		941	938	930	932	932	931	931	931	932	927	924	935
	$k_0 = 70$	658	663	665	664	660	653	650	651	648	646	647	655
		824	825	821	815	816	812	810	810	807	806	806	814
		898	895	891	893	893	888	887	884	883	883	883	893

Table 4.10: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_3 when F_1 is $L(0,1)$ and F_n is $L(1,1)$, the distribution of $Y+1$ with $Y \sim L(0,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	657	673	670	664	670	667	659	647	634	615	586	661
		813	829	826	824	829	825	820	811	801	786	763	824
		885	903	899	894	900	892	888	881	875	857	841	892
ω_3	$k_0 = 50$	719	717	711	700	685	679	663	650	639	630	610	682
		907	895	886	879	866	859	847	834	822	814	795	864
		960	950	949	946	937	938	934	924	918	911	896	940
	$k_0 = 70$	656	676	683	679	671	666	651	639	626	605	585	664
		829	841	842	841	842	839	840	831	819	802	789	839
		897	902	911	913	912	910	910	900	890	878	867	909

Table 4.11: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_1 when F_1 is $L(0,1)$ and F_n is $L(1, \sqrt{2})$, the distribution of $\sqrt{2}Y + 1$ with $Y \sim L(0,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	587	594	575	568	560	548	533	528	517	506	501	552
		773	775	763	754	739	733	715	710	704	699	689	740
		859	860	846	837	827	820	808	801	796	786	776	831
ω_1	$k_0 = 50$	676	677	667	653	647	642	624	620	605	602	591	655
		872	860	854	847	839	832	819	809	800	797	788	844
		945	937	928	921	915	912	907	897	889	882	876	921
	$k_0 = 70$	633	642	637	636	633	618	610	609	601	596	594	632
		814	818	815	811	808	793	780	775	772	770	768	810
		896	896	896	895	889	876	865	863	861	863	858	891

Table 4.12: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_2 when F_1 is $L(0,1)$ and F_n is $L(1,\sqrt{2})$, the distribution of $\sqrt{2}Y+1$ with $Y \sim L(0,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_{s}
	$k_0 = 30$	590	582	569	563	556	556	544	534	533	529	524	557
		766	765	755	744	736	734	728	716	717	713	710	737
		854	849	834	830	824	822	818	811	809	806	801	828
ω_2	$k_0 = 50$	666	671	662	648	646	639	628	626	621	617	614	653
		860	857	848	840	837	831	821	818	814	809	803	840
		937	931	921	913	912	911	907	907	903	898	893	919
	$k_0 = 70$	623	626	632	633	625	622	617	613	609	608	607	627
		812	809	806	808	798	793	789	782	781	777	776	805
		893	891	891	891	880	878	875	869	865	864	865	887

Table 4.13: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_3 when F_1 is $L(0,1)$ and F_n is $L(1, \sqrt{2})$, the distribution of $\sqrt{2}Y + 1$ with $Y \sim L(0,1)$.

	a	0.2	0.4	0.5	0.7	0.9	1.1	1.3	1.5	1.6	1.8	2	a_s
	$k_0 = 30$	597	611	606	597	587	571	547	529	508	489	466	562
		759	768	765	751	748	735	713	693	678	660	638	728
		840	844	847	840	839	822	801	782	764	746	731	816
ω_3	$k_0 = 50$	692	690	667	656	639	618	603	583	559	526	504	620
		879	876	853	842	823	804	783	768	742	709	686	808
		952	945	933	923	915	906	895	885	864	836	816	910
	$k_0 = 70$	627	649	649	657	650	639	623	611	581	552	526	637
		805	818	817	825	822	817	810	800	772	755	725	816
		888	900	903	905	902	900	894	885	868	850	823	902

Table 4.14: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_1 when F_1 is $G(1,1)$ and F_n is $G(4, \frac{1}{2})$, the distribution of $Y + 1$ with $Y \sim G(1,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	957	925	898	873	857	841	824	815	809	808	802	864
		993	985	973	958	949	940	934	930	926	924	920	949
		999	994	990	986	981	974	969	967	964	963	960	979
ω_1	$k_0 = 50$	929	903	889	870	853	838	826	821	816	817	815	862
		981	975	974	969	965	960	953	950	944	944	943	970
		997	994	991	991	990	987	983	980	977	977	976	991
	$k_0 = 70$	845	840	835	825	809	804	794	786	784	778	775	816
		939	938	936	931	919	914	905	900	898	896	892	919
		973	973	973	970	962	957	952	946	946	942	939	964

Table 4.15: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_2 when F_1 is $G(1,1)$ and F_n is $G(4, \frac{1}{2})$, the distribution of $Y + 1$ with $Y \sim G(1,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	920	887	873	857	845	833	822	817	814	809	808	844
		981	967	959	949	942	937	932	931	928	926	925	941
		993	988	985	981	976	972	967	967	965	964	963	976
ω_2	$k_0 = 50$	895	880	860	847	839	832	825	821	822	815	815	839
		973	970	967	966	960	956	953	950	951	944	944	964
		990	991	990	990	987	985	984	982	981	977	977	990
	$k_0 = 70$	835	830	821	809	805	800	794	788	785	783	784	808
		937	939	930	920	914	911	905	900	898	898	900	915
		970	974	969	963	957	956	953	947	945	944	946	958

Table 4.16: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_3 when F_1 is $G(1, 1)$ and F_n is $G(4, \frac{1}{2})$, the distribution of $Y + 1$ with $Y \sim G(1, 1)$.

	a	0.2	0.4	0.5	0.7	0.9	1.1	1.3	1.5	1.6	1.8	2	a_s
	$k_0 = 30$	956	953	946	930	922	907	874	849	803	767	724	906
		995	995	992	988	985	979	965	955	937	919	894	979
		998	998	996	996	995	994	991	989	977	965	945	994
ω_3	$k_0 = 50$	917	918	915	913	908	899	881	864	844	826	795	898
		982	981	979	978	975	973	971	965	956	948	939	974
		998	996	996	994	992	992	989	987	984	983	979	993
	$k_0 = 70$	831	841	841	835	829	829	818	815	797	786	777	832
		929	937	937	933	925	920	916	907	895	891	879	923
		972	977	977	976	971	969	965	958	950	948	934	969

Table 4.17: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_1 when F_1 is $G(1,1)$ and F_n is $G(\frac{3+2\sqrt{2}}{2}, 2\sqrt{2}-2)$, the distribution of $\sqrt{2}Y+1$ with $Y \sim G(1,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	952	941	909	890	872	857	842	834	827	823	819	875
		991	987	974	965	955	948	941	935	931	929	929	953
		997	994	990	985	979	975	971	968	965	963	962	979
ω_1	$k_0 = 50$	937	931	912	902	894	885	877	873	868	866	866	904
		983	985	982	981	979	976	974	972	971	969	969	985
		996	996	996	996	995	993	993	992	990	989	989	996
	$k_0 = 70$	863	870	875	876	872	871	867	864	864	866	866	882
		950	957	958	962	957	958	956	956	957	957	956	961
		981	983	984	987	986	989	986	987	987	988	988	990

Table 4.18: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_2 when F_1 is $G(1,1)$ and F_n is $G(\frac{3+2\sqrt{2}}{2}, 2\sqrt{2}-2)$, the distribution of $\sqrt{2}Y+1$ with $Y \sim G(1,1)$.

	a	1	2	3	4	5	6	7	8	9	10	11	a_s
	$k_0 = 30$	931	906	888	873	865	856	846	840	836	833	831	863
		986	972	965	956	952	947	943	938	935	935	933	951
		993	990	986	980	978	974	972	970	968	968	967	978
ω_2	$k_0 = 50$	918	906	896	892	887	885	878	874	874	870	868	890
		982	981	978	979	978	975	972	972	972	971	971	980
		994	995	995	995	993	992	992	992	992	991	991	993
	$k_0 = 70$	861	876	876	871	868	868	866	866	864	865	865	869
		954	960	962	958	956	956	954	957	956	956	956	958
		981	986	987	987	986	987	984	987	987	987	986	986

Table 4.19: $Acc(k_0, \delta)$ for $\delta = 5$ (top entry), 10 (middle entry) and 15 (bottom entry) by using the weight function ω_3 (lower part) when F_1 is $G(1, 1)$ and F_n is $G(\frac{3+2\sqrt{2}}{2}, 2\sqrt{2} - 2)$, the distribution of $\sqrt{2}Y + 1$ with $Y \sim G(1, 1)$.

	a	0.2	0.4	0.5	0.7	0.9	1.1	1.3	1.5	1.6	1.8	2	a_s
	$k_0 = 30$	947	944	934	920	916	893	868	833	806	783	748	898
		993	990	987	987	983	973	963	945	932	923	898	975
		996	996	995	995	994	992	990	981	973	968	946	993
ω_3	$k_0 = 50$	936	940	939	936	932	920	917	912	894	882	860	928
		988	989	992	990	988	987	988	984	978	975	970	989
		999	999	999	998	998	997	996	997	997	996	995	998
	$k_0 = 70$	861	870	879	879	885	886	884	881	874	872	871	887
		951	954	956	952	957	955	953	955	953	950	950	957
		987	987	987	985	985	982	981	981	978	978	981	983

4.4.2 A Real Data Application

In this subsection, we revisit the Nile data discussed in section 4. We employ all three weight functions with a_s chosen from $\{1, 2, \ldots, 100\}$ for both ω_1 and ω_2 but $\{0.2, 0.4, \ldots, 2\}$ for ω_3. We set γ to be either $0, 0.5$, or 1. They have all detected that the change point is located at the 28th observation, corresponding to the year 1898, which is the same as that detected in Zeileis *et al.* (2003).

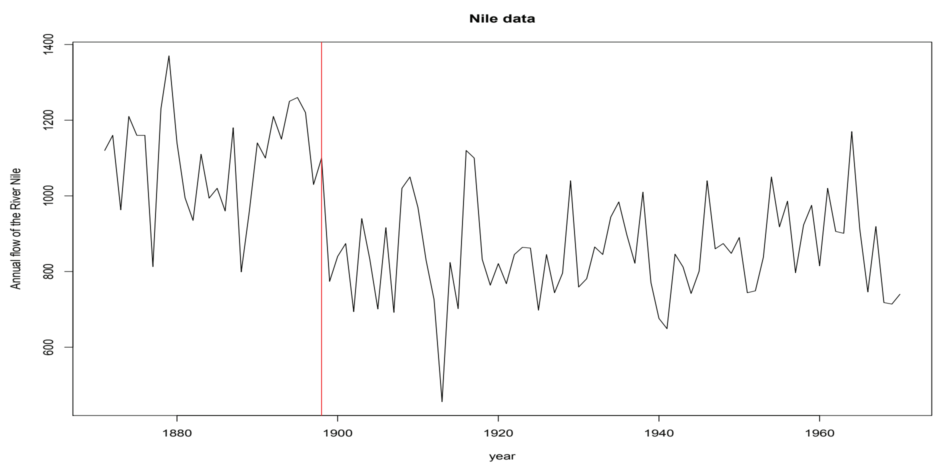

Figure 4.2: The time series plot of the annual flow of river Nile at Aswan from 1871 to 1970

5 Conclusions and Future Work

In this chapter, we summarize the results in this dissertation and introduce some possible future working problems.

5.1 Conclusions

In this dissertation, we investigate association rule mining from a transaction dataset and structural changes estimation in a time-ordered data sequence.

Firstly, we develop a new random sampling framework which imposes a probability distribution on the rule space and proposes to mine a random sample of rules from this probability distribution instead of mining the entire rule space. The annealing Gibbs sampling algorithm is adopted to randomly sample rules. It guarantees that the random sample contains the most significant rule with probability one. The sampling framework is flexible to incorporate any measure of interest for rules. Carefully designed simulation studies and a novel application of the method to a genomic data has shown the power of the new framework.

Secondly, structural changes estimation in GLMs is considered in the dissertation. A novel idea of matrix segmentation is introduced to transform the structural change problem into a model selection problem. A consistent estimator of coefficients is developed and an algorithm to estimate change points is also provided. Simulation studies show that this algorithm has low false alarm rate and high level of accuracy in estimating change points. This methodology can be used to estimate structural changes in distribution parameters of exponential family and coefficients of GLMs.

Lastly, structural change estimation in distributions of independent random variables is considered. A consistent change point estimator is proposed based on empirical characteristic functions. An algorithm to select an appropriate value for the tunning parameter a is also provided. The accuracy of this estimator is shown through carefully designed simulations for three different distributions and three different weight functions. The methodology can used to estimate changes in distribution parameters and distribution functions of independent random variables.

5.2 Future Work

In the area of transaction data mining, there are three possible working directions. The first one is to incorporate more measures into our algorithm since there are more and more measures of association rules proposed in literatures to measure different aspects of the rules and meet their own needs. For example, in Hahsler, et al. [2005], the apriori function can do association rules mining according to various of measures. The second one is to apply our method on real data analysis in areas of business, medical studies and economics. Many datasets from those research areas can be converted to a transaction dataset and the research question becomes mining association rules given some consequent. Then our method is applicable to such problems. Lastly, mining the most significant rule for a given consequent can be viewed as selecting a subset of features according to certain criterion. It is worth investigating how to turn this random sampling framework into a feature selection and grouping technique for transaction dataset.

In the area of change point analysis, two problems can be considered. In the dissertation, we consider the change point problem in GLMs for independent observations. However the data sequence may be correlated in time [Fokianos, et al., 2014]. So the procedure for estimating multiple change points in GLMs may be extended to estimate multiple change points in GLMs with AR(p)-type autocorrelations. There are some methods developed to detect change points in the climate data [Wang, et al., 2007]. However, there are a few methods invented to detect change points in the spatio-temporal data which draws a dramatically increasing attention due to their wide availabilities in many research areas including environmental study, climate change and biology. A model based change point detection method to detect change points in the spatial-temporal data is another possible working topic.

Bibliography

Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Rec*, **22**, 207–216.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases (Morgan Kaufmann, San Francisco)*, 487–499.

Amatya, A. and Demirtas, H. (2005). MultiOrd: An R package for generating correlated ordinal data. *Commun Stat Simul Comput*, **44**, 1683–1691.

Antoch, J., Gregoire, G. and Jarušková, D. (2004). Detection of structural changes in generalized linear models. *Statistics* & *probability letters*, **69**, 315–332.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Rev. Econ. Stat.*, **79**, 551–563.

Balke, N.S. (1993). Detecting level shifts in time series. *J. Bus. Econom. Statist*, **11**, 81–92.

Bayardo, R.J. and Agrawal, R. (1999). Mining the most interesting rules. *5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, New York)*, 145–154.

Bhattacharya, P.K. (1987). Maximum likelihood estimation of a change point in the distribution of independent random variables: general multiparameter case. *J. Multivariate Anal.*, **23**, 183–208.

Bhattacharya, P.K. and Brockwell, P.J. (1976). The minimum of an additive process with applications to signal estimation and storage theory. *Probab. Theory Related Fields*, **37**, 51–75.

Brodskij, B.E. and Darchovskij, B.S. (2000). *Non-parametric Statistical Diagnosis: Problems and Methods.* Kluwer Academic Publishers.

Carlstein, E. (1988). Nonparametric change point estimation. *Ann. Stat.*, **16**, 188–197.

Chaganty, N.R. and Joe, H. (2006). Range of correlation matrices for dependent Bernoulli random variables. *Biometrika*, **93**, 197–206.

Chen, J. and Gupta, A.K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media.

Chen, X.R. (1988). Testing and interval estimation in a change-point model allowing at most one change. *Sci China Ser A-Math*, **30**, 817–827.

Cobb, G.W. (1978). The problem of the nile: conditional solution to a change point problem. *Biometrika*, **65**, 243–251.

Csörgő, M. and Horváth, L. (1997). *Limit theorems in change-point analysis.* John Wiley & Sons Inc.

Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, **101**, 223–239.

Dite, G.S., et al. (2003). Familial risks, early-onset breast cancer, and BRCA1 and BRCA2 germline mutations. *J Natl Cancer Inst*, **95**, 448–457.

Dumbgen, L. (1991). The asymptotic behavior of some nonparametric change point estimators. *Ann. Stat.*, **19**, 1471–1495.

Epps, T.W. (1999). Limiting behaviour of the ICF test for normality under Gram-Charlier alternatives. *Statist. Probab. Lett.*, **42**: 175–184.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J., et al. (2010). *SIS*: Sure Independence Screening.

Fan, J. and Heng, P. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, **32**, 928–961.

Fanaee-T, H. and Joao, G. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, **2**, 113–127.

Feuerverger, A. and Mureika, R.A. (1997). The empirical characteristic function and its applications. *Ann. Stat.*, **5**, 88–97.

Fokianos, K., Gombay, E. and Hussein, A. (2014). Retrospective change detection for binary time series models. *Journal of Statistical Planning and Inference*, **145**, 102–112.

Gombay, E. (2001). U-statistics for change under alternatives. *J. Multivariate Anal.*, **78**, 139–158.

Hahsler, M., Grn, B. and Hornik, K. (2005). arules A computational environment for mining association rules and frequent item sets. *J Stat Softw*, **14**, 1–25.

Hämäläinen, W. (2009). Statapriori: An efficient algorithm for searching statistically significant association rules. *Knowl Inf Syst*, **23**, 373–399.

Hinkley, D. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**: 1–17

Hastie, T., Tibshirani and R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York, Springer.

Holmes, M., Kojadinovic, I. and Quessy, J. (2013). Nonparametric tests for change-point detection á la Gombay and Horváth, *J. Multivar. Anal.*, **115**, 16–32.

Hušková, M. and Meintanis, S.G. Change point analysis based on empirical characteristic functions. *Metrika*, **63**, 145–168.

Hušková, M. and Meintanis, S.G. (2013). Tests for the multivariate k-sample problem based on the empirical characteristic function. *J. Nonparametr. Stat.*, **20**, 263–277.

Inclan, C. and Tiao, G.C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Amer. Statist. Assoc*, **89**, 913–923.

Jiang, D. and Huang, J. (2014). Majorization minimization by coordinate descent for concave penalized generalized linear models. *Statistics and computing*, **24**, 871–883.

Jin, B., Shi, X. and Wu, Y. (2011). A novel and fast methodology for simultaneous multiple structural break estimation and variable selection for nonstationary time series models. *Stat Comput*, **2**, 221–231.

Jin, B., Wu, Y. and Shi, X. Consistent two-stage multiple change-point detection in linear models. *Canadian Journal of Statistics*, **44**, 161–179.

Jin, B.S., et al. (2014). Estimator of a change point in single index models. *Science China: Mathematics*, **57**, 1701–1712.

Kankainen, A. and Ushakov, N.G. (1998). A consistent modification of a test for independence based on the empirical characteristic function. *J. Math. Sci.*, **89**, 1486–1494.

Kent, J.T. (1975). A weak convergence theorem for the empirical characteristic function. *J. Appl. Probab.*, **12**, 515–523.

Koutrouvelis, I.A. and Meintanis, S.G. (1999). Testing for stability based on the empirical characteristic function with applications to financial data. *J. Stat. Comput. Simul*, **64**, 275–300.

Lu, Q. and Wang, XL. (2012). An extended cumulative logit model for detecting a shift in frequencies of sky-cloudiness conditions. *Journal of Geophysical Research*, **117**, 1–11.

Matteson, D.S. and James, N.A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.*, **109**, 334–345.

Odefrey, F., et al. (2010). Common genetic variants associated with breast cancer and mammographic density measures that predict disease. *Cancer research*, **70**, 1449–1458.

Page, ES. (1954). Continuous inspection schemes. *Biometrika*, **41**, 100–115.

Page, ES. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.*, **33**, 1065–1076.

Qian, G. and Field, C. (2000). Using MCMC for logistic regression model selection involving large number of candidate models. *Monte Carlo and Quasi-Monte Carlo Methods 2000, eds Fang K-T, et al. (Springer, Berlin)*, 460-474.

Qian, G., Shi, X. and Wu, Y. (2014). A Statistical Test of Change Point in Mean that Almost Surely Has Zero Error Probabilities. *Aust. N. Z. J. Stat.*, **55**, 435–454.

Rafajlowicz, E., Pawlak, M. and Steland, A. (2010). Nonparametric sequential change-point detection by a vertically trimmed box method. *IEEE Trans. Inform. Theory*, **56**, 3621–3634.

Shi, X., Wu, Y. and Miao, B. (2009). Strong convergence rate of estimators of change point and its application. *Comput. Stat. Data Anal.*, **53**, 990–998.

Ushakov, N.G. (1999). *Selected topics in characteristic functions*. Walter de Gruyter.

Wang X., Wen, H. and Wu, Y. (2007). Penalized maximal t test for detecting undocumented mean change in climate data series *Journal of Applied Meteorology and Climatology*, **46**, 916–931.

Yao, Y. (1987). Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *Ann. Statist.*, **13**, 1321–1328.

Zeileis, A., et al. (2003). Testing and dating of structural changes in practice. *Comput. Statist. Data Anal.*, **44**, 109–123.

A Appendix

A.1 A single change point detection and estimation in GLM

Consider the following model

$$g(\mu_t) = \begin{cases} \boldsymbol{x}_t^T \boldsymbol{\beta}, & t = 1, 2, \ldots, l, \\ \boldsymbol{x}_t^T \boldsymbol{\beta}^*, & t = l+1, l+2, \ldots, n. \end{cases}$$

Test $H_0: \; l = n$ and $H_1: \; l < n$.

The test statistic proposed in Antoch, *et al.* [2004] is summarized as follows. The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined as the solution of the following system of equations: $\sum_{t=1}^{n}(y_t - g^{-1}(\boldsymbol{x}_t^T\boldsymbol{\beta}))x_{tj} = 0, \; j = 1, 2, \ldots, p$. Then $\hat{\mu}_t = b'(\boldsymbol{x}_t^T\hat{\boldsymbol{\beta}})$ and $\hat{\sigma}^2 = a(\phi)b''(\boldsymbol{x}_t^T\hat{\boldsymbol{\beta}})$, where ϕ is assumed to be known. Let $\hat{S}(\tilde{l}) = \sum_{t=1}^{\tilde{l}}(y_t - \hat{\mu}_t)^T\boldsymbol{x}_t$, $\hat{F}(\tilde{l}) = \sum_{t=1}^{\tilde{l}} \hat{\sigma}_t^2 \boldsymbol{x}_t\boldsymbol{x}_t^T$, $\hat{F}(n) = \sum_{t=1}^{n} \hat{\sigma}_t^2 \boldsymbol{x}_t\boldsymbol{x}_t^T$, and $\hat{D}(\tilde{l}) = \hat{F}(\tilde{l}) - \hat{F}(\tilde{l})\hat{F}(n)^{-1}\hat{F}(\tilde{l})^T$. Assume that there exists k_0 such that $\hat{D}(\tilde{l})$ is positive definite for all $k_0 < \tilde{l} < n - k_0$. The test statistic is $T = \max_{k_0 < \tilde{l} < n-k_0} \hat{S}(\tilde{l})^T\hat{D}(\tilde{l})^{-1}\hat{S}(\tilde{l})$. They also showed that under

H_0, the limiting distribution of the test statistic is

$$P(T \leq 2\log\log n + (p+1)\log\log\log n + 2t - 2log\Gamma(\frac{p+1}{2})) \to \exp\{-2e^{-t}\}.$$

The asymptotic critical value for the test statistic at a given significance level can be obtained from this limiting distribution.

In the case that H_0 is rejected, the estimate of l is given by

$$\hat{l} = \arg\max_{k_0<\tilde{l}<n-k_0} \hat{S}(\tilde{l})^T \hat{D}(\tilde{l})^{-1} \hat{S}(\tilde{l}).$$

A.2 Hájiek-Rényi-Chow inequality

Lemma A1. (Hájek-Rényi-Chow inequality.) Suppose that $\{X_n, n \geq m\}, 1 \leq m \leq n$, is a martingale difference sequence. Let $\sigma_n^2 = EX_n^2$ and $c_1 \geq c_2 \geq \cdots \geq c_n > 0$. Define $S_n = \sum_{j=1}^{n} X_j$. Then for any $x > 0$, we have

$$P(\max_{m \leq j \leq n} c_j |S_j| \geq x) \leq \frac{1}{x^2}\left[m c_m^2 \sigma_m^2 + \sum_{j=m+1}^{n} c_j^2 \sigma_j^2 \right].$$