Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

STATISTICAL LEARNING METHODS FOR MINING MARKETING AND BIOLOGICAL DATA

by
Jie Zhang

Nowadays, the value of data has been broadly recognized and emphasized. More and more decisions are made based on data and analysis rather than solely on experience and intuition. With the fast development of networking, data storage, and data collection capacity, data have increased dramatically in industry, science and engineering domains, which brings both great opportunities and challenges. To take advantage of the data flood, new computational methods are in demand to process, analyze and understand these datasets.

This dissertation focuses on the development of statistical learning methods for online advertising and bioinformatics to model real world data with temporal or spatial changes. First, a collaborated online change-point detection method is proposed to identify the change-points in sparse time series. It leverages the signals from the auxiliary time series such as engagement metrics to compensate the sparse revenue data and improve detection efficiency and accuracy through "smart" collaboration. Second, a task-specific multi-task learning algorithm is developed to model the ever-changing video viewing behaviors. With the ℓ_1-regularized task-specific features and jointly estimated shared features, it allows different models to seek common ground while reserving differences. Third, an empirical Bayes method is proposed to identify $3'$ and $5'$ alternative splicing in RNA-seq data. It formulates alternative $3'$ and $5'$ splicing site selection as a change-point problem and provides for the first time a systematic framework to pool information across genes and integrate various information when available, in particular the useful junction read information, in order to obtain better performance.

STATISTICAL LEARNING METHODS FOR MINING MARKETING AND BIOLOGICAL DATA

by
Jie Zhang

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2017

APPROVAL PAGE

STATISTICAL LEARNING METHODS FOR MINING MARKETING AND BIOLOGICAL DATA

Jie Zhang

Dr. Zhi Wei, Dissertation Advisor — Date
Associate Professor of Computer Science, NJIT

Dr. James M. Calvin, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Usman W. Roshan, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Antai Wang, Committee Member — Date
Associate Professor of Mathematical Sciences, NJIT

Dr. Zhigen Zhao, Committee Member — Date
Associate Professor of Statistical Science, Temple University

BIOGRAPHICAL SKETCH

Author: Jie Zhang

Degree: Doctor of Philosophy

Date: May 2017

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2017
- Bachelor of Engineering in Software Engineering,
  Nanjing University, Nanjing, P. R. China, 2012

Major: Computer Science

Presentations and Publications:

Jie Zhang, Zhi Wei, Zhigen Zhao, and Zhenyu Yan, "A Beta-Binomial mixture model for conversion rate estimation in online advertising," *IEEE Transactions on Systems, Man, and Cybernetics*, in submission.

Jie Zhang, Zhi Wei, Pixu Shi, and Jun Chen, "A distance-based omnibus test for mediation effects of a high-dimensional mediator with application to microbiome data," *The American Journal of Human Genetics*, in submission.

Jianglan Liu, Mizuho Fukunaga-Kalabis, Gao Zhang, Clemens Krepler, et al., "LPAR signaling activated in neural crest stem cells promotes melanoma cell growth, invasion and therapy resistance," *Cancer Cell*, under review.

Markus V. Heppt, Joshua X. Wang, Denitsa M. Hristova, Zhi Wei, et al., "MSX1-induced neural crest-like reprograming promotes melanoma progression," *Journal of Investigative Dermatology*, under review.

Gao Zhang, Lawrence Wu, Ilgen Mender, Michal Barzily, et al., "Therapeutic targeting of telomerase prolongs control of therapy resistant melanomas," *Cancer Cell*, under review.

Hezhe Lu, Shujing Liu, Gao Zhang, Bin Wu, et al., "PAK signaling drives acquired drug resistance to MAPK inhibitors in BRAF-mutant melanomas," *Nature*, under review.

Bo Zhu, Shuyang Chen, Hongshen Wang, Juxiang Cao, et al., "DOT1L is a melanocyte lineage specific caretaker tumor suppressor," *Nature Cell Biology*, under review.

Stephan Wagner, Gao Zhang, Michela Perego, Mizuho Fukunaga-Kalabis, et al., "Tumor-associated B-cells induce therapy resistance of melanoma," *Nature Communications*, under review.

Jie Zhang, Zhi Wei, Zhenyu Yan, Mengchu Zhou, and Abhishek Pani, "Collaborated online change-point detection in sparse time series with application to online advertising," *IEEE Transactions on Systems, Man, and Cybernetics*, under review.

Jie Zhang, Zhigen Zhao, Kai Zhang, and Zhi Wei, "A feature sampling strategy for analysis of high dimensional genomic data," *The Fifteenth Asia Pacific Bioinformatics Conference*, accepted.

Jie Zhang, Kuang Du, Ruihua Cheng, Zhi Wei, et al., "Reliable gender prediction based on user's video viewing behavior," *Proceedings of IEEE International Conference on Data Mining*, pp 649-658, 2016.

Jie Zhang, and Zhi Wei, "An empirical Bayes change-point model for identifying $3'$ and $5'$ alternative splicing by next-generation RNA sequencing," *Bioinformatics*, vol. 32, pp 1823-1831, 2016.

Batool Shannan, Andrea Watters, Quan Chen, Stefan Mollin, et al., "PIM kinases as therapeutic targets against advanced melanoma," *Oncotarget*, vol. 7, No. 34, 2016.

Gao Zhang, Dennie T Frederick, Lawrence Wu, Zhi Wei, et al., "Targeting mitochondrial biogenesis to overcome drug resistance to MAPK inhibitors," *The Journal of Clinical Investigation*, vol. 126, pp 1834-1856, 2016.

Jie Zhang, Zhi Wei, Zhenyu Yan, and Abhishek Pani, "Collaborated online change-point detection in sparse time series for online advertising," *Proceedings of IEEE International Conference on Data Mining*, pp 1099-1104, 2015.

Turki Turki, Muhammad Ihsan, Nouf Turki, Jie Zhang, et al., "Top-k parametrized boost," *Mining Intelligence and Knowledge Exploration*, vol. 8891, pp 91-98, 2014.

To My Beloved Parents and Wife.

ACKNOWLEDGMENT

First and foremost, I wish to take this opportunity to express my heartfelt appreciation to my advisor Dr. Zhi Wei. It is his generous help and sustained encouragements that bring me the courage to overcome the difficulties in the road of pursuing my dream and hunting for the scientific truth. Dr. Wei is my friend, my mentor and my family. It is his tremendous efforts, invaluable guidance, and infinite patience that enable me to bring this dissertation to its culmination. I will always be indebted to Dr. Wei for generously encouraging and supporting me in this critical stage of my life.

Second, I am extremely grateful to Dr. James M. Calvin, Dr. Usman W. Roshan, Dr. Antai Wang and Dr. Zhigen Zhao for serving on my committee. They have provided me with academic advice inside and outside my research field. This dissertation would not have been possible without their invaluable guidance and generous help. In addition, I would like to extent special thanks to my collaborators, Dr. Mengchu Zhou from the Department of Electrical and Computer Engineering, Dr. Zhenyu Yan and Dr. Abhishek Pani from Adobe Systems Incorporated, Dr. Jun Chen from Mayo Clinic. I wish to thank Dr. Cristian M. Borcea, Dr. Ali Mili, Dr. David Nassimi, Dr. George Olsen and Dr. James Geller for supporting and helping me all the time. I would also like to thank my fellow graduate students and lab mates for their assistance and support.

Third, I truly appreciate my family for being my emotional anchor. I thank my parents for their endless love, support, and encouragement. I also wish to thank my lovely wife who has been quietly supporting me all the time.

Last, but not the least, I am thankful to all who have helped me directly or indirectly for the past five years. It is their support and encouragement that give me the courage and strength to pursue my Ph.D. degree abroad.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

The value of data has been broadly recognized and emphasized nowadays [41]. More and more decisions are made based on the data and analysis rather than solely on experience and intuition [44]. As evidenced by popular news media, e.g., the Economist, the New York Times and the National Public Radio, companies have benefited from the value of Big Data to guide decisions, trim costs and lift scales [41, 44]. For example, Walmart and Kohl's analyze sales, pricing, demographic and weather data to tailor product selections at different stores and determine the timing of price markdowns [44]; U.P.S. analyzes truck delivery times and traffic patterns to optimize routing [44].

With the fast development of networking, data storage, and the data collection capacity, data have increased in a dramatic scale in industry, science and engineering domains, which brings both great opportunities and challenges [82]. As reported, 2.5 quintillion bytes of data are created daily and 90% of the data in the world today are produced within the past few years [72, 82]. A report from McKinsey Global Institute points out that it would need 140,000 to 190,000 more employees with deep analytical expertise and 1.5 million more data-literate managers [44]. At the same time, new computational methods are in demand to process, analyze and understand these datasets.

Online advertising, an industry responsible for hundreds of billions of dollars yearly [3], benefits a lot from the Big Data technology. It plays a critical role in the Web ecosystem [25] and is emerging as a primary business for major technology companies such as Google, Facebook, Microsoft and Adobe [85]. Online advertising delivers promotional marketing messages to consumers through online media. It often

involves both a publisher or media provider, who integrates advertisements into online contents, and an advertiser, who provides the advertisements to be displayed on the publisher's contents [1]. Advertisers are usually motivated to fine-tune their ad spending strategies to drive the highest return on investment and maximize their key performance indicator (KPI). Media providers want to match the best ads to suitable audiences and improve the advertising effectiveness by displaying more pertinent ads to targeted audiences [85]. Thus, statistical models are in demand to help different practitioners to analyze relevant data and optimize their strategies and objectives.

However, the ever-changing data patterns are not in favor of the modeling [90]. Because of the dynamic nature of the online ad environment, data may change gradually or rapidly along the time. It should be noted that the temporal or spatial change of the data has been widely observed in industry and academia. For example, the Click Through Rate (CTR) and Revenue Per Click (RPC) may change remarkably due to the marketing events such as promotion and new product announcement [90]; The structure and activity of the microorganisms residing in the human gut are affected and changed due to the long-term dietary [16]. In addition, the change itself is an important pattern that researchers may want to capture in solving their specific problems. Taking bioinformatics as an example, Wei et al. propose to identify the $3'$ UTR length changes (or $3'$ UTR switching), as it plays a critical role in regulating the stability, localization and translation of mRNA. Thus, for both industry applications and basic science research, new methods, which could address the change explicitly by performing the change-point detection or implicitly by considering it in building the model, are required and desirable.

This dissertation focuses on the development of computational methods for analyzing and modeling the ever-changing data, with application to industry problems (online advertising) and basic science research (bioinformatics). First, a collaborated online change-point detection method is proposed to perform online change-point

detection in sparse time series. Through effectively leveraging and coordinating with auxiliary time series, such as the engagement metrics in online advertising, it can quickly and accurately identify the change-points in sparse and noisy time series data. It could greatly improve the precision of the predictive model by providing accurate change-point information and, therefore, help users build more accurate systems to measure ad performance. Second, the user's online video viewing behaviors are explored for gender prediction and a novel task-specific multi-task learning algorithm is developed to model the ever-changing viewing behaviors. It extends the conventional multi-task learning methods by introducing the task-specific features and, therefore, allows different models to seek common ground while reserving differences. Third, an empirical Bayes change-point model is proposed to identify alternative $3'$ and $5'$ splice sites (SS) in next-generation RNA sequencing data. Specifically, the alternative $3'$ SS and $5'$ SS problem is formulated as a change-point problem. The proposed empirical Bayes method could efficiently pool information across genes to improve detection efficiency. In addition, a flexible testing framework is provided for users to address different levels of questions, namely, whether alternative $3'$ SS or $5'$ SS happens, and/or where it happens.

This dissertation is organized in the following manner. Chapter 2 discusses the background and related work of the online change-point detection for online advertising, gender prediction based on user's video viewing behaviors, together with the identification of alternative $3'$ SS and $5'$ SS from the RNA-seq data. Chapter 3 introduces the collaborated online change-point detection method for sparse time series. Chapter 4 proposes the task-specific multi-task learning algorithm for predicting the users' genders based on their video viewing behaviors. Chapter 5 develops an empirical Bayes change-point model for identifying $3'$ and $5'$ alternative splicing sites in RNA-seq studies. Finally, Chapter 6 summarizes the contribution of this dissertation and discusses future directions for the research.

CHAPTER 2

BACKGROUND

2.1 Online Change-point Detection for Online Advertising

In online advertising, advertisers are motivated to optimize their allocation of dollars and advertising strategies through internal or external platforms to drive the highest return on investment (ROI) and maximize their key performance indicator (KPI). For example, Adobe Media Optimizer (AMO) is such a platform that integrates various statistical models to help advertisers manage, forecast and mathematically optimize their paid media, e.g., ad campaigns in search, display, as well as social media. It provides a consolidated point of view about how statistical models and algorithmic solutions are performing together with online media to accurately forecast ad performance. Based on the modern portfolio theory, AMO also employs a portfolio optimization approach borrowed from finance risk management [46] to optimize ad spending and bidding price numerically and deliver global optimal ad bidding strategies for advertisers across multiple ad channels.

A crucial step in ad optimization is to build accurate predictive models and predict various critical quantities that measure ad performance. The quantities may be cost-related, such as the number of impressions during a time interval (e.g., a day), Click Through Rate (CTR), Cost Per Thousand Impressions (CPM) and Cost Per Click (CPC), as well as revenue-related, such as Conversion Rate (CR), Revenue Per Thousand Impressions (RPM) and Revenue Per Click (RPC). Practitioners often face a variance-bias dilemma when integrating and leveraging historical data in building those predictive models. On the one hand, they can select a long window of historical data for utilizing as much data as possible to reduce estimation uncertainty. However, due to the highly dynamic nature of online ad environment, the data may shift quickly

and dramatically, making the model suffer from severe bias once the data pattern changes. On the other hand, the modelers can simply ignore long historical data and only focus on a short data window, in order to avoid the potential bias problem. Nevertheless, the variance of the model may largely increase due to the limited size of training data. How to address these issues remains a formidable challenge to both academic researchers and industrial practitioners.

One interesting way to address it is to detect the change-points before building predictive models. Given the information of change-points, which indicate the drastic changes in the data pattern, practitioners can easily apply appropriate strategies to optimize the variance-bias tradeoff. Specifically, if a change-point is identified, practitioners may simply ignore or apply a larger decay rate to the data before the change-point to reduce the bias incurred by data pattern changes. On the other hand, if no change-point is detected, practitioners can safely leverage a long historical window of data and enjoy variance reduction brought by it.

However, a common problem for online advertising is that data are generally very sparse. With sparse observations, online change-point detection often becomes challenging. Sparse and noisy data often lead to a high level of false discoveries when using a loose cutoff in determining change-points. Imposing a strict cutoff for avoiding false discoveries too strenuously causes a long delay between the time of the occurrence of a change-point and the time it is detected, and may miss some real change-points as well [56].

The data sparsity problem is even more severe for revenue or conversion related data (e.g., RPC time series) in the online advertising application here. Taking Search Engine Marketing (SEM) as an example, Figure 2.1 shows a typical online advertising funnel, in which the volumes of events become sparser and sparser from top to bottom.

The revenue or conversion related events, located at the most bottom of funnel, are much sparser than the previous steps in the funnel (e.g., about 4 order of

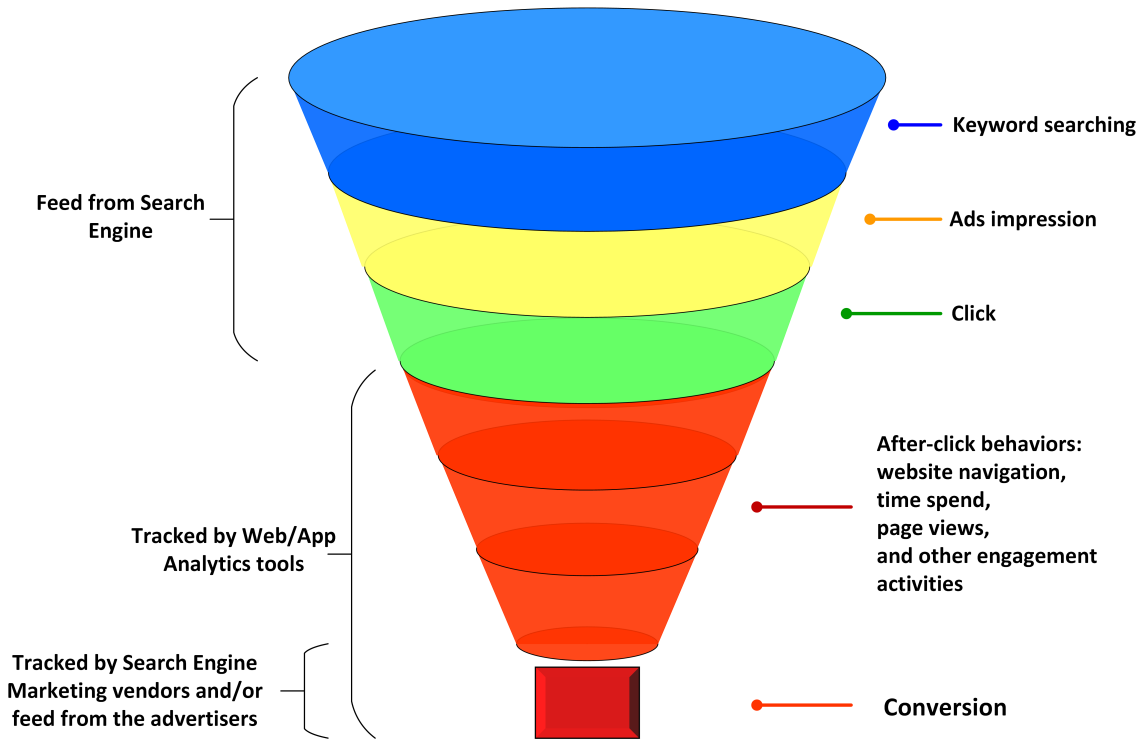

Figure 2.1 Data funnel in online advertising.

magnitude less than impression data and 2 order of magnitude less than click data). For example, it is common that, for an advertiser, over 90% of keywords have fewer than 5 conversions per day in average. This sparsity poses a great challenge in detecting the revenue or conversion change-points of a keyword.

While the revenue data are sparse, the on-site user behavior events, located in the middle part of the funnel, are much richer. After a user enters an advertiser's website by clicking an ad, the user may do a variety of activities before considering a purchase or subscription. Those activities are called engagement metrics, which are usually correlated with but much richer than final conversion events.

2.2 Gender Information, Video Viewing Behavior and Online Advertising

Among various demographic traits, gender has been a critical factor in market segmentation strategy [78] and plays a crucial role in precisely targeting the potential

consumers in online advertising and ecommerce [34, 49]. For example, companies which provide fragrances, skin care, and makeup, mainly target female audiences, while razor producers like Philips focus almost exclusively on male audiences. Chinese advertisers may require a certain percentage of their audiences to be female in their commercial contracts with publishers, as women are often thought of as the major buyers in China [88]. In addition, third party measurements, for instance, from Nielsen and comScore in USA or from Miaozhen and AdMaster in China[1], are widely applied by advertisers to monitor how many ads, sold by digital media sellers, get delivered to audiences with the targeted gender.

However, digital media sellers often do not have access to registered data for their users. Even when people need to register to consume digital content, they may not provide the correct information. For these reasons, targeting based partially on self-reported registered data may not score well against more accurate third party measurements. With this in mind, many companies have started using media consumption data to model users' demographics. Since people of different genders have different media preferences, this model provides significant enhancements over random guesses.

2.3 Demographic Prediction

Earlier demographic prediction was mainly built upon the analysis of the association between people's demographic attributes and their linguistics writing and speaking styles [40, 52, 59]. For instance, Otterbacher et al. applied logistic regression on movie reviews from IMDB [52]. Feng et al. utilized logistic regression to infer user's gender based on video tags and keywords [22]. With the rise of web services and social media, exemplified by Google, Bing, Facebook and Twitter, researchers begin to predict demographic information with users' online activities. Hu et al. [31]

[1]Nielsen: `http://www.nielsen.com/`; comScore: `http://www.comscore.com/`; Miaozhen: `http://miaozhen.com/`; AdMaster: `http://www.admaster.com.cn/`.

made a first approach to predict users' genders and ages from their web browsing behaviors. Supervised regression model was trained to estimate every Webpage's demographic tendency, i.e., the probability distribution of the ages and genders of a given Webpage's readers, and then Bayesian framework was employed to predict user's age and gender based on the age and gender tendency of the Webpages that he/she had browsed. Followed by [8], Bi et al. proposed to infer users' demographic information from their query history based on labelled Facebook Likes data. Because only Facebook Likes data had labelled users, they matched the Facebook Likes with search queries by using Open Directory Project categories, and transfered the model trained on Facebook Likes to predict users' genders and ages based on their search query histories. Burger et al. [12] sampled millions of tweets from Twitter and applied Balanced Winnow2 algorithm for gender prediction of the unlabeled Twitter users. Culotta et al. [15] proposed to predict the demographics of Twitter users based solely on whom they followed through regression models. Other behaviors that have been investigated include mobile communication patterns [18, 89] and purchase behavior [74]. However, little research has been conducted, in the scenario of gender prediction, to model the ever-changing data patterns and discriminant features, and to control the Type I error rate.

2.4 Identification of 3′ and 5′ Alternative Splicing from RNA-Seq

Alternative splicing plays an important role in building complex organisms from a limited number of genes. They provide a major mechanism for enhancing transcriptome and proteome diversity, and critically regulate various biological functions [36, 38]. Researchers observe that more than 90% of human genes undergo alternative splicing (AS), a much higher percentage than anticipated [10, 73]. Of various alternative splice forms, alternative 3′ SS and 5′ SS are particularly important and constitute more than 30% of all AS events as revealed by RNA-seq [73]. Several

studies have found that alternative 3′ SS and 5′ SS events are relevant to many diseases. By analyzing alternative 3′ SS and 5′ SS events, researchers can obtain precious diagnostic and prognostic information for therapies [27, 65].

Thanks to high-throughput RNA-seq, genome-wide quantitative studies on AS events become feasible [53, 73]. Quite a few computational methods have been developed to detect and identify AS events. These methods can be roughly classified into three categories based on their strategies. The first category, represented by Cufflinks [70], perform differential splicing detection based on transcript quantification, which is the most challenging. Short reads, sampled from RNA-seq, can be aligned to multiple transcripts due to the similarity and overlaps between alternative transcripts [33, 42]. It makes the expression estimation of individual transcript an undetermined problem. In addition, various sampling biases, including position-specific biases [11, 43, 58, 83] and sequence-specific biases [58, 71] in the RNA-seq data, incur daunting difficulties for accurate transcript quantification. Consequently, the efficiency of these methods is diminished by the uncertainty in transcript quantification.

The second category of methods aims to detect differential splicing by testing differential expression of the annotated events obtained from existing splicing databases. Representative examples in this category include ALEXA-seq [26], MISO [37], MATS [61] and SpliceTrap [80]. Among them, MISO employs a statistical model to estimate expression of alternatively spliced exons and isoforms [37]; MATS leverages a Bayesian statistical framework to flexibly test the hypothesis of differential alternative splicing patterns [61]. These methods may work well when splicing events are well and accurately annotated. They are not applicable to detect novel AS events not cataloged yet in existing annotation databases.

The third category of methods, including DiffSplice [32], DEXSeq [2] and FDM [64], utilize splice junction read information to overcome the annotation dependency

limitation. The performances of these methods are highly dependent on the number and quality of splice junction reads. Sequencing costs often set limits to sequencing depth and coverage of RNA-seq data sets [63] and, consequently, performance of these junction read-based methods. Furthermore, the number and quality of aligned splice junction reads will also rely on sequencing technology as well as read-mapping tools [21].

Very recently, Wang et al. [75] propose a change-point model which requires no annotation information. It relies on characterizing the coverage change for detecting alternative polyadenylation (APA). In principle, it can be applied for detecting $3'/5'$ AS events. However, compared with APA, a key difference for $3'/5'$ alternative splicing is that junction read information can be useful and utilized for locating splice sites. For example, a simple strategy for calling $3'/5'$ AS events is to identify locations supported by at least N independent splice junction reads with different alignment start positions [76]. It is noted that when sequencing depth is not enough, a significant proportion of $3'/5'$ AS events may not be covered by junction reads, in particular when using a stringent N threshold for ensuring quality. Due to the junction read coverage limitation, there is room for improvement even when sequencing depth is high. Exon read coverage may be used as clues for $3'/5'$ AS events. Thus, using both coverage and junction read information may improve both sensitivity and specificity of AS $3'/5'$ calls compared with relaying solely on junction reads or read coverage. It is therefore desirable to develop a method that can systematically integrate both junction read information and coverage information. From a methodology point of view, Wei's method is a frequestist approach and fails to pool and exploit information across many genes under investigation. In addition, it tests only whether there is a change-point, but not where the change-point is. Thus, its change-point location estimation does not guarantee any multiplicity control.

CHAPTER 3

COLLABORATED ONLINE CHANGE-POINT DETECTION IN SPARSE TIME SERIES FOR ONLINE ADVERTISING

3.1 Introduction

Online advertising delivers promotional marketing messages to consumers through online media. Advertisers often have the desire to optimize their advertising spending strategies in order to drive the highest return on investment and maximize their key performance indicator. To build accurate ad performance predictive models, it is crucial to detect the change-points in the historical data and apply appropriate strategies to address the data pattern shift problem. However, with sparse data, which is common in online advertising and some other applications, online change-point detection is very challenging.

This chapter proposes a novel collaborated online change-point detection method. Through effectively leveraging and coordinating with auxiliary time series, such as the engagement metrics introduced above, it can quickly and accurately identify the change-points in sparse and noisy time series data. In addition, the proposed method could help to improve the accuracy of predictive models by providing accurate change-point information. Simulations and real data experiments have been conducted to justify and demonstrate the effectiveness of the new method.

3.2 Motivation and Data Overview

Without loss of generality, it focuses on detecting the change-points in Revenue Per Click (RPC) Time Series (TS) in Search Engine Marketing throughout this chapter. The same idea can also be applied to other revenue related TS or other advertising channels. Because of the data sparsity issue, the signal-to-noise ratio of RPC time series is not high enough for effective online change-point detection. On the other

Top Funnel | Middle Funnel | Bottom Funnel

Keyword	Clicks	Engagement metrics	Conversions
Keyword No. 1	5	• Visited 5 pages per click in average • No bounces • Watched video twice • Revisited the website later after close the browser • ……	0
Keyword No. 2	5	• 2 bounces out of 3 clicks; • Less one page view per click in average; • No revisits • ……	0

Figure 3.1 Aggregated data of two keywords in a day.

hand, as shown in Figure 2.1, the engagement metrics, which are usually precursors of final conversions, are much richer. Thus, it is natural to consider that RPC and those engagement metrics may have significant positive correlations.

Figure 3.1 shows an example of data of two keywords collected in a given day. The point estimates of RPC of the two keywords in the day are both zero due to zero conversions. However, the engagement metrics indicate that users who enter the website through keyword 1 are more engaged with the website or products than those who enter through keyword 2. If practitioners believe in the correlation hypothesis between RPC and engagement metrics, they may want to assign different RPC estimates to the two keywords (e.g., keyword 1 looks more promising than keyword 2). Therefore, the basic idea is to leverage the signals in those richer engagement metrics TS to compensate the sparse RPC TS and improve detection efficiency and accuracy through "smart" collaboration between the "target" RPC TS and the "auxiliary" engagement metrics TS.

There are many possible types of engagement metrics and some of them can be industry or website-specific. In this chapter, three most commonly used engagement metrics, Time Spend Per Click (TSPC), Page View Per Click (PVPC) and Bounce

Rate (BR), are selected as examples. Time Spend Per Click measures the total time a user spends on the website after entering the website through an ad click and before shutting down the browser. While longer time may indicate higher engagement, it does not fully capture a user's activeness on the website. Thus, Page View Per Click is introduced, which measures the total number of pages viewed by the user after entering the website. In addition, the Bounce Rate is used as the third metric. A bounce means that a user exits the website immediately after the click. For convenience, the engagement metrics TS are denoted as auxiliary TS and the RPC TS is the target TS. The latter is tracked by Adobe Media Optimizer (AMO) while the former are tracked by Adobe Analytics. All the Metrics as well as RPC are averaged across the users by day to generate the TSs for each keyword.

3.3 Collaborated Online Change-point Detection

The extremely sparse and noisy RPC TS will incur great difficulty for existing algorithms to accurately detect change-points based on such revenue data alone. Engagement metrics collected by analytics tools, such as TSPC, PVPC, and BR, can be informative for detecting RPC changes. This section proposes a collaborated online change-point detection method to leverage this information for detecting sequential change-points in the sparse RPC TS data. Specifically, each individual engagement metric may have its own relevance to RPC. Instead of trying to use these raw auxiliary TSs directly and separately, the new method first combines them into one single TS. Then it coordinates the target TS with the combined auxiliary TS for more accurate change-point detection.

3.3.1 Combining Auxiliary TSs

Let $Y = (y_1, y_2, \cdots, y_t)$ and $A^{(i)} = (a_1^{(i)}, a_2^{(i)}, \cdots, a_t^{(i)})$, $i \in \{1, 2, 3\}$, denote the RPC TS and the auxiliary TSs for a keyword, respectively. To leverage the signals of the

auxiliary TS to inform the target TS, the linear model is utilized to characterize the relation between the RPC TS and the auxiliary TS, and extract most relevant signals from the raw data.

$$y_i = \beta_0 + \sum_{j=1}^{K} \beta_j \times a_i^{(j)} + \epsilon_i,$$

where K is the number of auxiliary TSs and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the random error. Ridge regression with L_2 penalty is utilized to estimate the coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_K)^T$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{K} a_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^{K} \beta_j^2 \right\}, \tag{3.1}$$

where λ is the tuning parameter and y_i and $a_i^{(j)}$ represent the data in RPC TS and jth auxiliary TS, respectively. Parameter λ is selected through 10-fold cross-validation for the application. Ridge regression shrinks the regression coefficients by imposing the L_2 penalty on their size [28]. It can effectively reduce the variance of parameter estimation and obtain more stable models compared with ordinary least squares (OLS) estimation [84]. This is particularly desired when the sample size is small, which is the case for the current application. Other regressions such as LASSO [68] and Elastic Net [93] may be considered too when scenarios change.

Though it aims to perform change-point detection on keyword-level RPC TS, in order to obtain more reliable estimates of the regression coefficients, the data are aggregated and the model is fitted at a higher level, e.g., ad group, campaign or campaign group. After fitting the model, all the keywords under the same group use the same coefficients to combine the auxiliary TSs to obtain the combined auxiliary TS, which is a stable estimation of the original target TS.

Given keyword-level auxiliary TSs, the combined auxiliary TS for each keyword is $\hat{Y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_t)$ and

$$\hat{y}_t = \hat{\beta}_0 + \sum_{j=1}^{K} a_t^{(j)} \hat{\beta}_j, \tag{3.2}$$

where $a_t^{(j)}$ is the t-th data in the j-th auxiliary TS.

3.3.2 Online Change-point Detection

A likelihood-based method [84] is employed by the proposed framework as a component for online change-point detection. It was first proposed to detect changes in the mean within a sequence of normally distributed observations by Hinkley [30], who derives the asymptotic distributions of the maximum likelihood estimate and the likelihood ratio statistic for testing hypotheses. It was further extended to detect changes in variance [14, 35]. In addition, a popular R [57] package, **changepoint** [39], has implemented this likelihood-based framework for performing change-point detection. It is convenient for implementation and yields good performance as demonstrated in the experiments later.

It is noted that this online change-point detection module is relatively independent in the whole framework. Other existing online change-point detection algorithms can be employed to replace the likelihood-based method and perform the online change-point detection function. This is another advantage of the proposed method. Namely, it gives users the flexibility to choose suitable online change-point detection algorithms to fit their particular applications.

LRT for Single Change-point Detection This section considers the following likelihood ratio test (LRT) for single change-point detection. Let $X_{1:N} = (x_1, x_2, \cdots, x_N)$ denote a sequence of N observations ordered in time, where x_i represents the observed value at time i. The change-point τ divides the whole sequence into two homogeneous segments $X_{1:\tau}$ and $X_{(\tau+1):N}$, in which the observations are independently and identically distributed (i.i.d.). Let H_0 be the null hypothesis that there is no change-point in sequence $X_{1:N}$ and H_1 be the alternative hypothesis that there is a

single change-point. The testing statistic Λ for LRT is defined as

$$\begin{aligned}\Lambda &= -2\log\left(\frac{\mathcal{L}(X_{1:N}|H_0)}{\mathcal{L}(X_{1:N}|H_1)}\right) \\ &= 2\left[\log\left(\mathcal{L}(X_{1:N}|H_1)\right) - \log\left(\mathcal{L}(X_{1:N}|H_0)\right)\right],\end{aligned} \tag{3.3}$$

where $\mathcal{L}(X_{1:N}|H_0)$ and $\mathcal{L}(X_{1:N}|H_1)$ are the maximum likelihoods under null hypothesis H_0 and alternative hypothesis H_1, respectively. And

$$\begin{aligned}\mathcal{L}(X_{1:N}|H_0) &= \prod_{i=1}^{N} f(x_i|\hat{\theta}_0), \\ \mathcal{L}(X_{1:N}|H_1) &= \max_{\tau}\left(\prod_{i=1}^{\tau} f(x_i|\hat{\theta}_1) \times \prod_{i=\tau+1}^{N} f(x_i|\hat{\theta}_2)\right),\end{aligned} \tag{3.4}$$

where $\hat{\theta}_0$, $\hat{\theta}_1$ and $\hat{\theta}_2$ are the maximum likelihood estimates of the parameters and $\tau \in \{1, 2, \cdots, N-1\}$ indicates a change-point. The testing procedure also involves choosing an appropriate threshold, C, such that users reject the null hypothesis H_0 if and only if $\Lambda > C$. If H_0 is rejected, the change-point is estimated as

$$\hat{\tau} = \arg\max_{\tau}\left(\prod_{i=1}^{\tau} f(x_i|\hat{\theta}_1) \times \prod_{i=\tau+1}^{N} f(x_i|\hat{\theta}_2)\right). \tag{3.5}$$

This is a general LRT in that the model $f(X_{1:N};\theta)$ is to be specified according to a data model. In this application, Gaussian models are utilized to characterize data fluctuation along time.

It is noted that it is not trivial to select an appropriate threshold C. For most distributions, including the Gaussian models, there is no closed form for C that can guarantee multiplicity control because it seeks through a series of dependent models (i.e., change-point candidates) for a maximum statistic. An exception is the binomial model for count data [79]. In addition, falsely reported change-points (false positives) and falsely missed change-points (false negatives) incur different costs that are generally hard to quantify. For example, in this application, it is hard to compare the impacts of false positives and false negatives on the RPC predictive model

users aim to optimize. It is noted that the ultimate goal is to use the change-point model for improving the RPC predictive model. Focusing on change-point detection accuracy only may fail to consider the different costs of different types of errors in the change-point detection procedure and lead to a sub-optimal solution.

Therefore, the parameter C is tuned through 10-fold cross-validation by minimizing the prediction error of the RPC predictive model with the reported change-points. More specifically, users choose an initial value based on the Bayesian information criterion (BIC) [9, 60], and then conduct greedy search to find the best threshold C through 10-fold cross-validation.

Online Change-point Detection There are now two sets of time series, the original RPC TS and the combined auxiliary TS derived from multiple auxiliary TSs. Instead of further combining them, the proposed method performs online change-point detection on the target TS and combined TS separately, and then coordinates the detection decision in a later stage, which could utilize the original signal of the target TS and simultaneously take the advantage of the more stable combined auxiliary TS. Algorithm 1 shows the detailed steps of the proposed online change-point detection algorithm. It tries to detect one change-point at a time. When a new observation has been received, a decision, whether there is a change-point, is made based on the observations received so far. When a change-point is identified, the proposed algorithm starts to detect the next change-point from the observations received after the newly identified change-point.

Technically, it may detect change-points for a sequence of any length (>1). In practice, however, testing on a sequence with too few observations makes little sense. As noted in [29], industrial practitioners need to gather a modest number of observations to acquire an initial verification of the assumptions of their models before starting formal change-point testing. When the data are extremely sparse and noisy,

Algorithm 1: Online Change-Point Detection Algorithm

initialize change-point set $T = \emptyset$, sequence $X = \emptyset$, last change-point $\tau' = 0$, least-segment-length $L = 10$ and minimum-distance $D = \frac{L}{2}$

repeat

  get new data x_t and add it to X

  if $length(X[(\tau' + 1) : t]) \geq L$ **then**

    perform single change-point detection on segment $X[(\tau' + 1) : t]$

    if τ is identified and $(\tau - \tau') \geq D$ **then**

      update $\tau' = \tau$

      add τ to T

    end if

  end if

  output T for coordination

until no data available

e.g., RPC TS, a large number of observations could help reduce false discoveries and are generally desired. However, collecting more data requires more time and causes a long delay between the time of the occurrence of the change-point and the time it is detected. The experimental results show that initializing least-segment-length $L = 10$ is a good compromise for balancing the tradeoff to detect a change-point as early as possible and to make an accurate detection with few false discoveries. A minimum distance $D = \frac{L}{2}$ between two adjacent change-points is also set to require a reasonable distance between them.

3.3.3 Coordination Strategies

This section proposes two strategies to coordinate the results:

- **Strategy 1:** Report a change-point if and only if it is identified as a change-point in both RPC TS and combined auxiliary TS;
- **Strategy 2:** Report a change-point if Strategy 1's condition is met or if it is identified as a change-point in either TS and its combined statistic is significant.

Strategy 1 requires the change-point supported by both RPC TS and combined auxiliary TS. Strategy 2 relaxes this constraint by considering the points that may be extremely significant in one TS but marginally significant in the other. Thus, Strategy 2 is relatively lenient and may report more change-points than Strategy 1 given that they have the same significance threshold.

Let T_0 and T_1 denote the two sequences of change-points detected from the original RPC TS and the combined auxiliary TS, respectively. Because the target TS data are extremely sparse and noisy, it is rare that the change-points detected from both TSs exactly locate at the same position. Thus, a distance threshold δ is defined, and if the change-points detected from the RPC TS and the combined auxiliary TS are *close* to each other (distance $< \delta$), the proposed method treats them as the same

change-point and reports their middle position as the final change-point position. Note that $\delta = 5$ in all the experiments. With this relaxed match rule, Strategy 1 is formally defined as

$$T^* = \{\frac{\tau + \tau'}{2} |\ \tau \in T_0,\ \exists \tau' \in T_1,\ |\tau - \tau'| < \delta\}, \tag{3.6}$$

where T_0, T_1 and T^* represent the change-point sets for RPC TS, combined auxiliary TS and final reports, respectively.

Let

$$\begin{aligned} \Lambda^*_\tau &= \max_{(\tau-\delta)<\upsilon<(\tau+\delta)} \{\Lambda_\upsilon\}, \\ \upsilon^*_\tau &= \arg\max_{(\tau-\delta)<\upsilon<(\tau+\delta)} \Lambda_\upsilon, \end{aligned} \tag{3.7}$$

where Λ^*_τ represents the maximum testing statistic for all positions close to τ (distance $< \delta$), and υ^*_τ represents the position with the maximum testing statistic. The Strategy 2 is defined as

$$\begin{aligned} T^{**} = \ & T^* \cup \\ & \{\frac{\tau + \upsilon^*_{1\tau}}{2} |\ \tau \in T_0, \Lambda_{0\tau} + \Lambda^*_{1\tau} > 2C^*\} \cup \\ & \{\frac{\upsilon^*_{0\tau} + \tau}{2} |\ \tau \in T_1,\ \Lambda_{1\tau} + \Lambda^*_{0\tau} > 2C^*\}, \end{aligned} \tag{3.8}$$

where $\Lambda_{0\tau}$ and $\Lambda_{1\tau}$ represent the testing statistics from the RPC TS and combined auxiliary TS, respectively; $\Lambda^*_{0\tau}$ and $\Lambda^*_{1\tau}$ are the maximum testing statistics for all positions close to τ in the RPC TS and combined auxiliary TS, respectively; C^* is the threshold for the combined statistics. A larger threshold for the combined statistic is advisable to reduce false discoveries, and T^{**} reduces to T^* when $C^* \to +\infty$. Note that $C^* = 1.5C$ for all the experiments.

3.4 Experiment

This section first performs simulation studies to investigate the numerical performance of the proposed method. Then the real life experiments are conducted with the data

from Adobe Media Optimizer and Adobe Analytics. It compares the performance of four different methods:

M0: No change-point detection;

M1: Online change-point detection with RPC TS only;

M2: Collaborated online change-point detection with coordination Strategy 1;

M3: Collaborated online change-point detection with coordination Strategy 2.

Two metrics are used to evaluate the performance of the competing methods:

Metric 1: Detection accuracy of change-points;

Metric 2: Prediction accuracy of the predictive model after leveraging the detected change-points.

Metric 1 measures the detection accuracy of the competing methods mainly for simulation studies, in which the true locations of change-points are known. Metric 2 measures the prediction accuracy of the predictive model which utilizes the information of the estimated change-points. As shown in Figure 3.2, when equipped with the change-point model, the predictive model utilizes the change-point information to make the differential use of historical data. M0, which does not perform any change-point detection, is used as a baseline method to compare the performance of the same predictive model when integrated with different change-point methods.

There are many different implementations of RPC predictive models ranging from the most advanced machine learning methods to the simplest point estimation methods. Since the predictive models themselves are not the focus of this chapter, for illustration purpose, two popular and easy-to-implement models are used:

- **Predictive Model 1**: Predict with average;

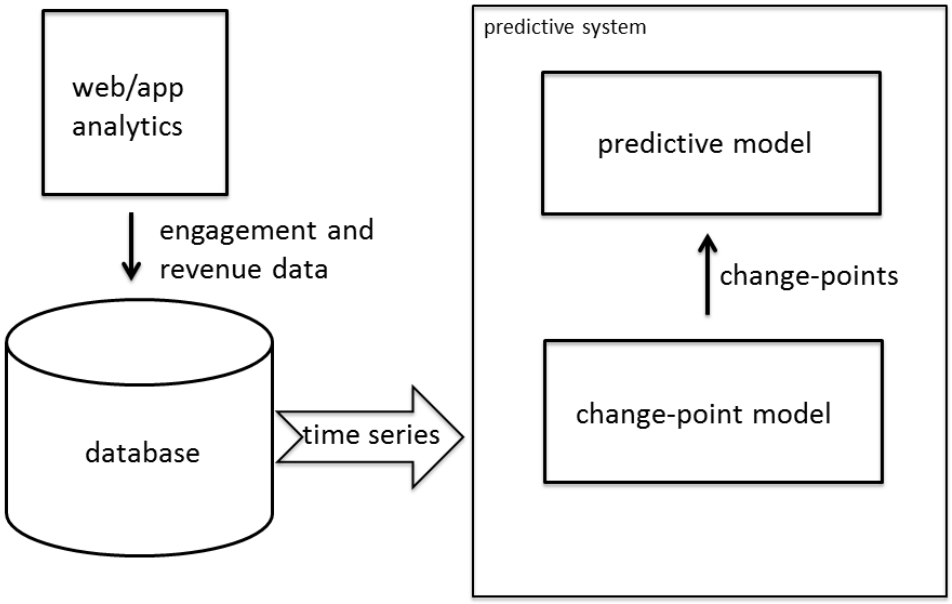

Figure 3.2 Structure of predictive system when integrated with the change-point model.

- **Predictive Model 2**: Predict with weighted average based on time decay (half-life =30 days).

Given a collected RPC segment $Y_{i:j}$, predicting with average (Model 1) can be simply viewed as predicting the next time period RPC using the mean of the previous data

$$y_{next} = \frac{\sum_{k=i}^{j} y_k}{j - i + 1},$$

while predicting with decay (Model 2) predicts next RPC using the weighted average of the previous data

$$y_{next} = \frac{\sum_{k=i}^{j} y_k w_k}{\sum_{k=i}^{j} w_k},$$

where $w_k = \gamma^{(j-k)}$ reflects time decay and $i \leq k \leq j$. Note that the decay rate $\gamma = 0.02284$, such that the half-life is equal to 30 days.

It should be noted that Model 1 does not address any change in the data while Model 2 can account for gradual change in the data. However, Model 2 can not cope well with sharp changes in the data pattern. Each time a change-point is detected, a simple strategy, i.e., ignoring the data prior to the detected change-point, is applied in the experiments to help predictive model focus on the most relevant data. In real implementation, more sophisticated strategy, e.g., increasing the decay rate, can be employed.

3.4.1 Simulation Studies

This section simulates the data by mimicking the data generating procedure in online advertising. First, it samples the number of clicks n from a Poisson distribution, the number of conversions c from a Binomial distribution and the Revenue Per Conversion

r from a Normal distribution

$$\begin{cases} n \sim Pois(\lambda), \\ c \sim B(n, P), \\ r \sim \mathcal{N}(\mu_0, \sigma_0), \end{cases} \tag{3.9}$$

where λ is the expected number of clicks per day, P is the conversion rate, and μ_0 and σ_0 are mean and standard deviation for Revenue Per Conversion. Then, it computes the Revenue Per Click

$$y = \frac{\sum_{i=1}^{c} r_i}{n}. \tag{3.10}$$

The auxiliary TS is simulated by first sampling n observations and then computing their average value as one data point

$$\begin{aligned} x^{(i)} &\sim \mathcal{N}(\mu_i, \sigma_i), \\ a^{(i)} &= \frac{\sum_{j=1}^{n} x_j^{(i)}}{n}, \end{aligned} \tag{3.11}$$

where μ_i and σ_i ($i \in \{1, 2, 3\}$) are mean and standard deviation, respectively. Note that

$$\mu_i = \beta_0^{(i)} + \beta_1^{(i)} \times P + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 1)$. Positive $\beta_1^{(i)}$ indicates the positive correlation between auxiliary TS and RPC TS, while negative $\beta_1^{(i)}$ represents the negative correlation between auxiliary TS and RPC TS. In the simulation studies, all parameters, λ, μ_i and σ_i ($i \in \{0, 1, 2, 3\}$), are chosen such that the mean and variance are close to real data. Given all other parameters, conversion rate P is used to control the signal level of the simulated RPC TS. Larger conversion rate means more conversions generated from the same number of clicks making the RPC TS less sparse and noisy, while smaller conversion rate reduces the number of conversions and makes the RPC TS sparser

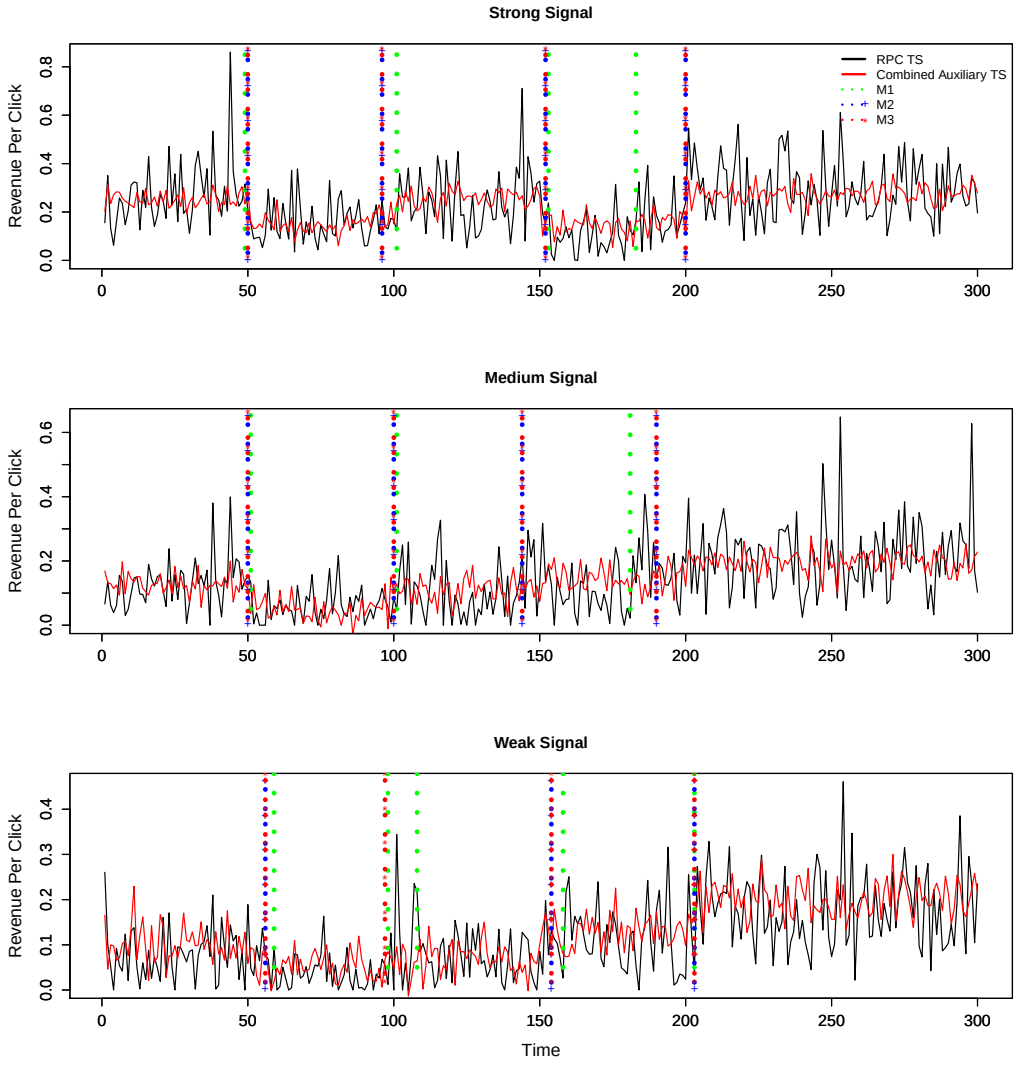

Figure 3.3 Change-point detection results for simulated data.

and noisier. In addition, P is also used to set change-points by employing different P values to simulate the data before and after the change-points.

Three scenarios, with different parameters, are simulated, generating datasets with strong, medium and weak signals by controlling conversion rate, representing low, medium and high sparsity, respectively. Each scenario has 50 randomly generated datasets, with time series length = 300. To make a complete comparison between different methods, four change-points are set in generated TS at time 50, 100, 150 and 200, respectively.

Table 3.1 Averaged Number of False Positives and False Negatives (#FP, #FN)

Method	**Strong Signal (#FP, #FN)**	**Medium Signal (#FP, #FN)**	**Weak Signal (#FP, #FN)**
M1	(1.54, 1.78)	(1.58, 2.2)	(1.64, 2.54)
M2	(0.08, 1.3)	(0.14, 1.8)	(0.22, 2.16)
M3	(0.72, 0.62)	(0.42, 1.68)	(0.62, 1.9)

Detection Accuracy Figure 3.3 shows examples of the change-point detection results for different scenarios with strong, medium and weak signal datasets, respectively. The black curve and red curve are the original RPC TS and combined auxiliary TS, respectively. Dotted green, blue and red vertical lines indicate the change-points identified by M1, M2 and M3, respectively. Here are a few remarks. Firstly, the combined auxiliary TS (red solid line) is less noisy and retains the characteristics of the original RPC TS, indicating that a better inference of change-points may be obtained by employing the combined auxiliary TS together with the original RPC TS. Secondly, M2 and M3 are better than M1, as M1 reports false discoveries as shown in Figure 3.3. Thirdly, M3 detects more change-points than M2. For example, M2 misses the change-point at time 100 in weak signal dataset (Figure 3.3: Weak Signal). This is expected because, compared with Strategy 1, the Strategy 2 is less strict and allows more change-points to be reported. Table 4.1 summarizes the average results for different scenarios using different change-point detection methods. In Table 4.1, the two numbers in each pair of parentheses are the average number of false discoveries (false positives) and average number of missed real change-points (false negatives), respectively. It is clear that M2 and M3 are better than M1 in all scenarios. First, M1, which is based on RPC TS only, has more false discoveries than the collaborated methods, M2 and M3, in all scenarios. Secondly, M1 also misses more real change-points in all scenarios compared with M2 and M3. Thirdly, comparing

Table 3.2 Average Prediction Error Ratios for Simulation Experiments

Prediction Method	**M0**	**M1**	**M2**	**M3**
Predict with average	1.0	0.877	0.871	0.861
Predict with decay	1.0	0.904	0.898	0.890

M2 and M3, we can find that M2 has fewer false discoveries but more missed real change-points than M3, which is expected, as M2 uses a much stricter coordination strategy than M3.

Prediction Accuracy This section further compares the prediction errors of the same predictive system when integrated with M0, M1, M2 and M3, respectively. Among them, M0 is used as a baseline method, representing the predictive model without using any change-point information.

Fifty datasets with weak signal are simulated. Table 3.2 summarizes the average prediction error ratios $\bar{\varepsilon}_i/\bar{\varepsilon}_0$ for the same predictive system when integrated with different change-point algorithms Mi for $i \in \{1, 2, 3\}$. If $\bar{\varepsilon}_i/\bar{\varepsilon}_0 < 1$, method Mi could help to improve the accuracy of the predictive system. The smaller $\bar{\varepsilon}_i/\bar{\varepsilon}_0$ is, the larger the improvement is. As shown in Table 3.2, the average prediction errors for systems integrated with change-point methods are smaller than the baseline errors of predictive model integrated with M0, as corresponding average prediction error ratios are smaller than one ($\bar{\varepsilon}_i/\bar{\varepsilon}_0 < 1$). In addition, the predictive model integrated with the collaborated online change-point detection methods, M2 and M3, are better than the same predictive model integrated with M1, as $\bar{\varepsilon}_j/\bar{\varepsilon}_0 < \bar{\varepsilon}_1/\bar{\varepsilon}_0,\ j \in \{2, 3\}$. It should be noted that the proposed methods, M2 and M3, could help the predictive model to obtain over 10% improvement in prediction accuracy than M1 (as shown in Table 3.2).

Figure 3.4 plots the prediction error ratios for the simulated 50 datasets. The

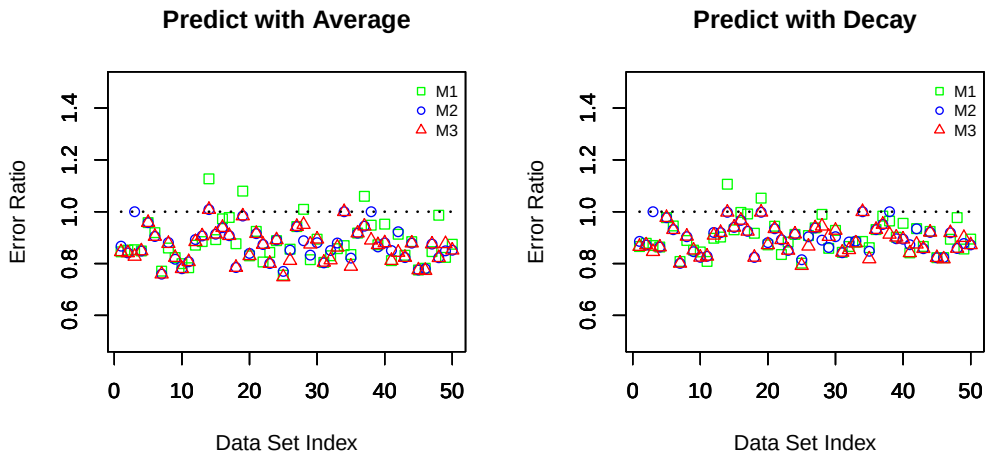

Figure 3.4 Prediction error ratios for predictive model integrated with different change-point detection methods on 50 datasets.

left and right subgraphs show the results for model predicting the next RPC data by averaging historical data and by weighted average with decay rate $\gamma = 0.02284$, respectively. Each point represents an prediction error ratio ($\bar{\varepsilon}_i/\bar{\varepsilon}_0$) on one dataset. Specifically, green, blue and red points represent the prediction error ratios for the predictive system integrated with M1, M2 and M3, respectively. The dotted horizontal line $Y = 1$ is the baseline such that points below it represent smaller prediction errors and points above it represent larger prediction errors compared with the baseline. In Figure 3.4, most of the points, which represent prediction error ratios for M1, M2 and M3 on 50 different datasets, are under the baseline $Y = 1$ indicating that the predictive model could improve its precision by efficiently employing the change-point information. In addition, there are several green points above the dotted baseline ($\bar{\varepsilon}_1/\bar{\varepsilon}_0 > 1$) indicating that, in some cases, system, integrated with M1, has a worse performance than the baseline, while all blue and red points are under or at least at the baseline ($\bar{\varepsilon}_j/\bar{\varepsilon}_0 \leq 1,\ j \in \{2, 3\}$). Thus, the predictive system integrated with M2 and M3 tend to have a much more stable performance than the system integrated with M1.

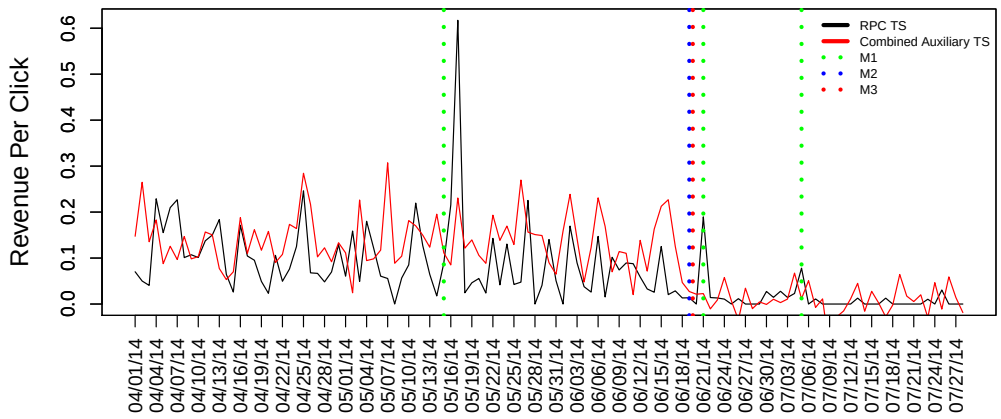

Figure 3.5 Change-point detection results for real data experiments.

Paired two-sample t-tests are also performed on prediction results to compare the performance of the predictive models when integrated with different change-point algorithms. It shows that M1, M2 and M3 are significantly better than baseline method (p-values $< 10^{-5}$). In addition, both M2 and M3 are better than M1. Among them, M3 is significantly better than M1 (p-value $= 0.02$).

3.4.2 Real Data Experiments

For real data experiments, 228 keywords from three advertisers across different industries are randomly selected to evaluate the competing methods.

Detection Accuracy Figure 3.5 shows an example of the change-point detection results for the RPC TS. As shown in Figure 3.5, M1 detects 3 change-points, while M2 and M3 only report one change-point at the date close to June 18, 2014. Since the advertiser had a new product released on June 18, 2014, it is very likely that the date, June 18, 2014, is a real change-point in this RPC TS. It is with high probability that M1 reports 2 false discoveries considering the spikes in the RPC TS near the change-points. After investigation, there are no particular events in those days. Thus, they

Table 3.3 Average Prediction Error Ratios for Real Data Experiments

Method	Advertiser 1	Advertiser 2	Advertiser 3
M0	(1.0, 1.0)	(1.0, 1.0)	(1.0, 1.0)
M1	(0.933, 0.951)	(1.025, 1.146)	(0.970, 0.969)
M2	(0.911, 0.938)	(0.868, 1.024)	(0.954, 0.958)
M3	(0.911, 0.929)	(0.904, 1.005)	(0.938, 0.936)

are concluded as false positives. These falsely reported change-points may provide misleading information to the predictive model and consequently impose negative impact on the predictive model's performance.

Prediction Accuracy This section further integrates M1, M2 and M3 into a predictive system to demonstrate that more persistent and stable improvements can be gained for the predictive model by leveraging accurate change-points detected from M2 and M3 compared with M1, which reports more false discoveries in the sparse and noisy data. Table 3.3 summarizes the average prediction error ratios across the keywords for each Advertiser. In Table 3.3, the two numbers in each pair of parentheses are the average prediction error ratios for predictive model using different prediction methods, predicting with average and predicting with decay, respectively. Here are a few remarks. Firstly, though M1 improves the performance of the predictive system in Advertiser 1 and 3, it causes much worse results for Advertiser 2 compared with baseline M0. Secondly, the system, integrated with M2 or M3, has a better or comparable performance compared with the baseline for all Advertisers, indicating that the proposed method is more appropriate for sparse and noisy data and can improve the performance of the predictive system persistently and stably. The detail results for 228 keywords are shown in Figure 3.6. It is obvious that M2 and M3 greatly improve the system's prediction accuracy, since the majority of the points are below

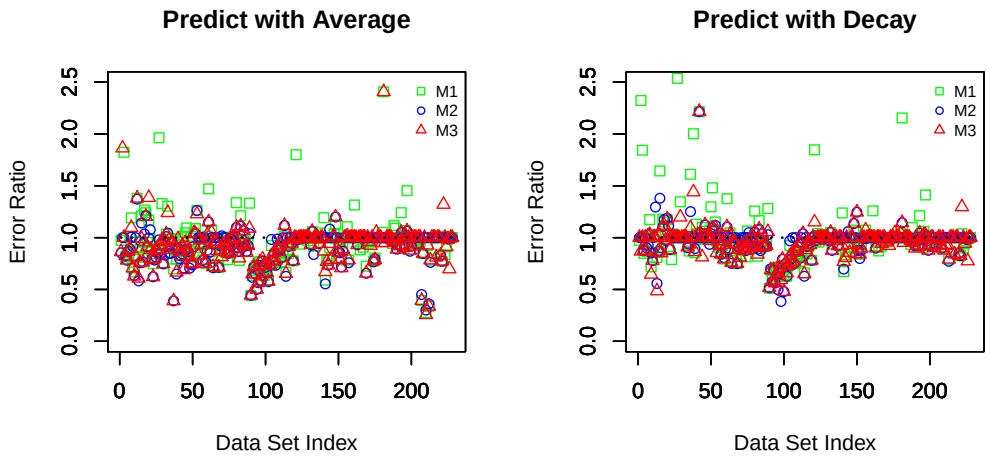

Figure 3.6 Prediction error ratios for predictive model integrated with different change-point detection methods on 228 keywords.

the dotted baseline $Y = 1$ and others are at the baseline. Though M1 also improves the prediction accuracy in most of the keywords, it reports many false discoveries when the data is sparse and noisy, which eventually hurts the system's performance and causes many green points above the baseline as shown in the Figure 3.6.

Paired two-sample t-tests show that M1, M2 and M3 can significantly improve the system's prediction accuracy (p-values < 0.01). Furthermore, the system integrated with M2 or M3 gives significantly better results than system integrated with M1 (p-values $< 10^{-5}$), which demonstrates the power and advantages of the collaborated online change-point detection method.

3.5 Conclusion

This chapter proposes a collaborated online change-point detection method for sparse time series. By leveraging the auxiliary time series, it can quickly and accurately identify the changes in the revenue data and enable the predictive model to use historical data intelligently. Experimental results have demonstrated the benefits of using the proposed algorithm in improving the precision of the predictive model in

online advertising. Based on its efficiency in real data applications, it is in the process of being deployed in AMO at Adobe.

CHAPTER 4

RELIABLE GENDER PREDICTION BASED ON USER'S VIDEO VIEWING BEHAVIOR FOR ONLINE ADVERTISING

4.1 Introduction

With the growth of the digital advertising market, it has become more important than ever to target the desired audiences. Among various demographic traits, gender information plays a key role in precisely targeting the potential consumers in online advertising and ecommerce. However, such personal information is generally unavailable to digital media sellers.

This chapter investigates the problem of gender prediction based on users' online video viewing behavior. Considering the ever-changing data patterns and related features, it proposes a novel task-specific multi-task learning algorithm to efficiently leverage historical data and obtain decent performance. To achieve high-precision predictions, it further proposes Bayes testing and decision procedures to identify desired users with controlled false discovery rate (FDR). Comprehensive experiments show that the proposed method can deliver the best performance over alternative methods.

4.2 Analyzing Video Viewing Behavior

To analyze the video viewing patterns, 35187, 70031 and 78996 users, registered with identification cards, were randomly sampled in August, September and October 2015, respectively, and their viewing logs were extracted from PPTV. It contains 543,240 distinct videos and more than 20 million video viewing logs. Table 4.1 summarizes the number of users and videos sampled in each month. It collected 8871, 18563 and 21057 female users in August, September and October, respectively, which accounted for 25.2%, 26.5% and 26.7% of the total sampled users in that month.

Table 4.1 Summary of the Sampled Users and Videos

Month	Users			Videos
	Male	Female	Total	
Aug	26,316	8,871	35,187	301,079
Sept	51,468	18,563	70,031	342,034
Oct	57,939	21,057	78,996	377,499

4.2.1 User's Viewing Behavior and Preference

Based on the video tags that PPTV utilizes to manage its video resources, it could be viewed as 11 large categories, which contain thousands of videos, plus many small groups. In addition, an additional category of "Others" is created to collect all small group videos. Figure 4.1 summarizes the distribution of popularity, measured by the number of views, of different video categories. TV Series, Animation, Movies, and News, which are favored by different audiences, account for 78.4% of total views, while Sports (Sp for short), Games (Ga for short), Variety Shows (Va for short), and Stars and Entertainment (Star for short) are only applicable to a specific group of audiences and are less popular.

Category-level Gender Preference This section summarizes the proportion of female audiences for each category in Figure 4.2 and the corresponding male users' proportion could be computed as $P_{male} = 1 - P_{female}$. The red dashed horizontal line (Y=0.252) is served as a baseline representing the proportion of female users in the collected data. Note that September and October actually have a slightly higher percentage of female users than August. For simplicity, the same baseline is applied for all these three months. Obviously, the proportion of female audiences in some categories significantly deviates from the baseline, which exhibits strong gender preference. Specifically, Sports and Games exhibit very low proportions of

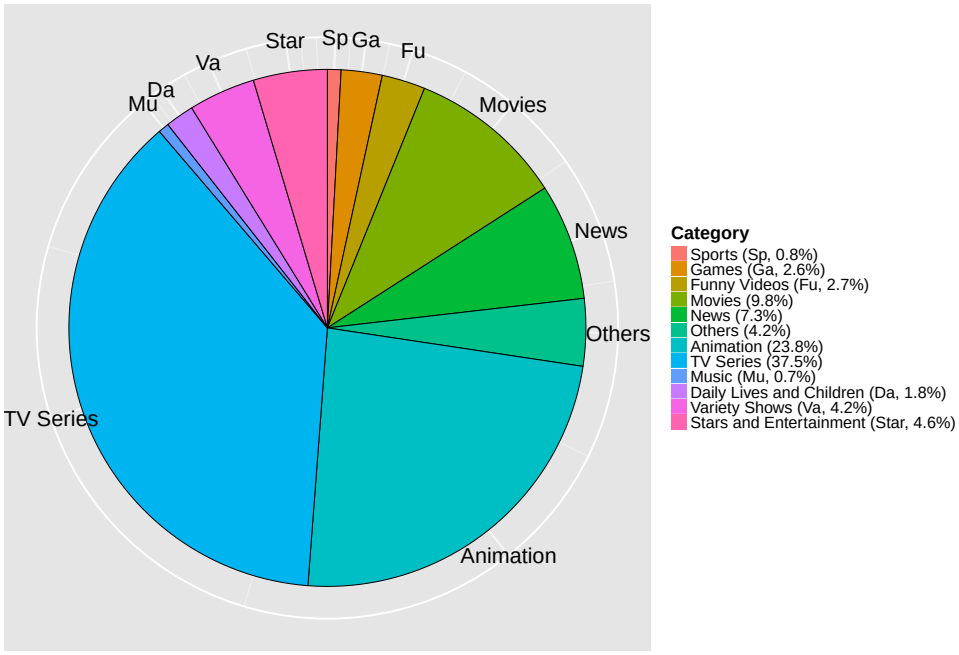

Figure 4.1 Popularities of different video categories.

female audiences. In contrast, Variety Shows, and Stars and Entertainment are more attractive to female audiences. An extreme case is that the "Hallyu Channels" (a sub-category of Stars and Entertainment), which is about South Korean actors and stars, focuses exclusively on female audiences. Additionally, female audiences also pay close attention to the Daily Lives and Children category. It is interesting that Sports and Games, which only account for 3.4% of total views (See Figure 4.1), exhibit stronger demographic preference than popular TV dramas and movies, and therefore, are more informative for gender prediction.

Video-level Gender Preference Many videos are equally popular among female and male audiences. But also quite a few videos are watched more by female than male, or vice versa. To find these gender-discriminative videos, this section performs Fisher's exact test [23] on each video. Let H_0 and H_1 be the null hypothesis

that the gender distributions are the same among people who watch the video and the people who don't, and the alternative hypothesis that the proportion of female audiences is either higher or lower among people who watch the video than the people who don't (denoted as background), respectively. FDR [7] adjustment is used to provide multiplicity control. It detects 8188, 13952 and 15764 videos in August, September and October, respectively, which demonstrate significant difference (adjusted p-value ≤ 0.05) in the proportions of female audiences compared with the background. The odds ratio ($OR = \frac{odds_{viewed}}{odds_{not\ viewed}}$, where $odds = \frac{\#female\ users}{\#male\ users}$) is further computed for each significant video. Table 4.2 summarizes the number of significant videos in each odds ratio interval. Note that the case $OR = 1$ represents the null hypothesis of no gender distribution difference, and is not shown in Table 4.2.

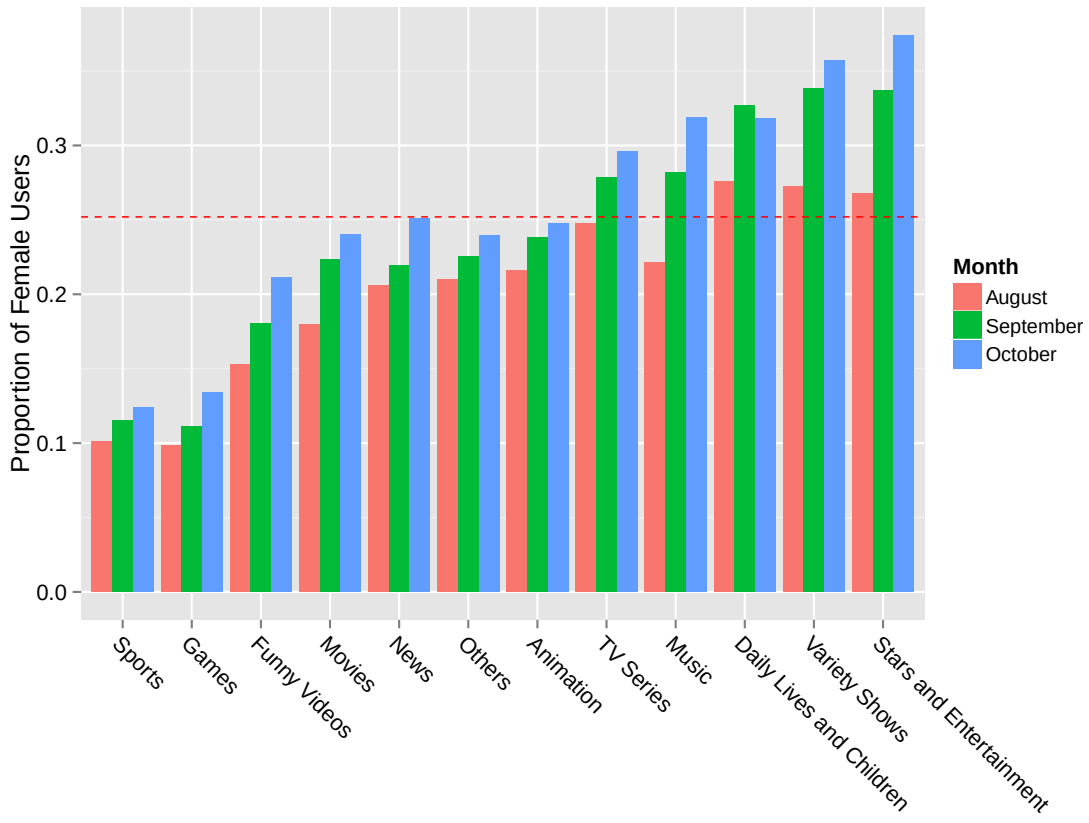

Figure 4.2 Proportion of female audiences in each category.

Table 4.2 Summary of the Significant Videos

Month	$OR < 1$	$1 < OR \leq 3$	$OR > 3$	Total
Aug	4,110	1,356	2,722	8,188
Sept	6,670	2,338	4,944	13,952
Oct	6,502	3,196	6,066	15,764

$OR < 1$ suggests a lower proportion of female audiences compared to the background, while the $OR > 1$ implies a higher percentage of female audiences. Considering the *odds* for background is close to $\frac{1}{3}$, $OR > 3$ indicates that more female users than male users watch the video. It is interesting that there are many significant videos with $OR > 3$ especially in September and October when more data are sampled (See Table 4.1). Thus, researchers could treat videos with strong gender preference as discriminant features and infer users' gender information based on whether or not they watch the video and the number of times they watch it.

4.3 Challenge and Motivation

The data pattern and discriminant videos (features) keep changing over time, making it hard to build efficient models. For example, in Table 4.2, let S_{Aug}, S_{Sept} and S_{Oct} denote sets of significant videos for August, September and October. Then, $|S_{Aug} \cap S_{Sept}| = 4254$, $|S_{Sept} \cap S_{Oct}| = 6621$ and $|S_{Aug} \cap S_{Sept} \cap S_{Oct}| = 3022$. Approximately, $1 - \frac{|S_{Aug} \cap S_{Sept}|}{|S_{Aug}|} = 48.0\%$ and $1 - \frac{|S_{Sept} \cap S_{Oct}|}{|S_{Sept}|} = 52.5\%$ of significant videos from August and September become nonsignificant in the next month, respectively, indicating that a large proportion of previously significant and valuable videos (features) may become nonsignificant or unrelated in the next time period. In addition, for those videos that remain significant, their magnitude of effects may change over time. Both the reduced number of views and the change in audiences' gender distribution for those videos

may lead to such changes. It should be also noted that 3022 videos could maintain their significance for three months in Table 4.2.

The continuous emergence of new videos may contribute to such ever-changing data patterns and features. Additionally, video recommendation system provides a profound effect on the views of videos [92]. Zhou et al. observed that the related video recommendation was the main source of views for the majority of the videos [92]. An ordinary video may quickly become popular when promoted by the recommendation system. However, its popularity may not last long, as the recommender system may quickly switch to promote other newly uploaded videos. The intrinsic property of the video and user's watching habit may also affect the data patterns. For example, news and live program, which are marked with immediacy, may only be watched in a short time period after they are uploaded. In contrast, high quality movies and TV dramas could maintain stable views for a long time, as users usually like to "binge-watch" these videos [77].

Videos with short-term popularity and long-term popularity are treated as "ordinary videos" and "classical videos", respectively. Because previous ordinary videos are rarely viewed by any audience later, employing this part of historical data for current gender prediction will inevitably introduce many unrelated features and degrade the performance. In contrast, previous viewing records, generated from classical videos, may contribute to the current task. In the following sections, a novel multi-task learning algorithm is proposed to model ordinary videos and classical videos separately and efficiently.

4.4 Reliable Gender Prediction

4.4.1 Problem Formulation

As shown in Figure 4.3, historical data can be divided into successive time intervals $t_1, \cdots, t_L$, where t_L represents the current time period. Each interval t_i contains

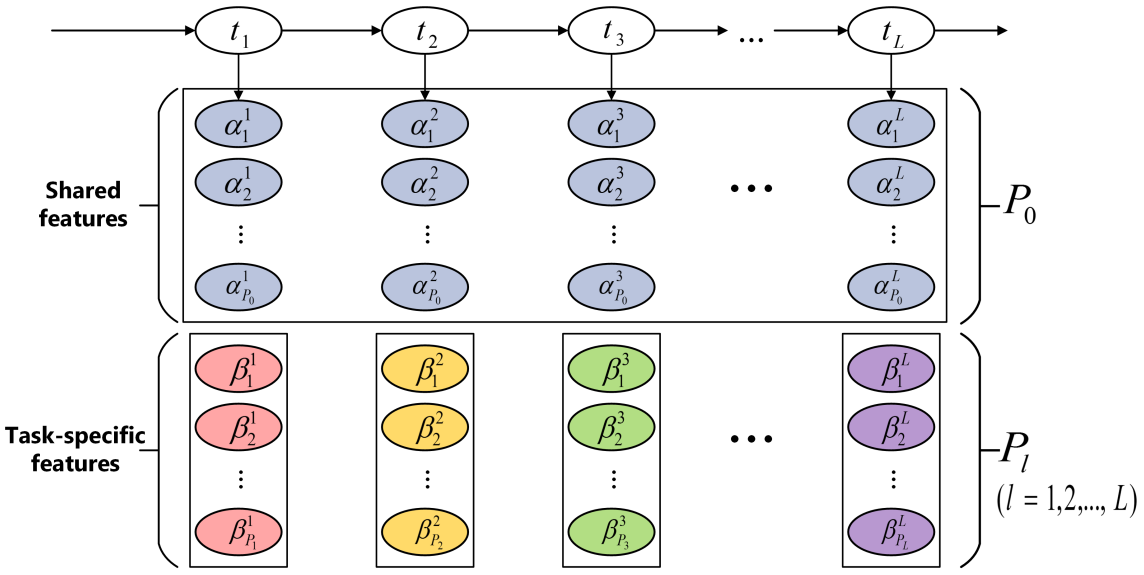

Figure 4.3 Task-specific multi-task learning model for modeling users' video viewing behavior.

both transient ordinary videos, which locally belong to one specific time interval, and classical videos, which are shared across different time intervals. In order to maximize the benefits brought by shared classical videos while eliminating interference from previous transient videos, it is formulated as a multi-task learning problem. Models trained in successive time intervals are treated as multiple related tasks so that the shared classical videos (shared features) are jointly estimated across multiple tasks (See Figure 4.3). In addition, task-specific features are introduced into the conventional multi-task learning framework to capture the effects of transient ordinary videos. It is worth noting that, though multiple models are jointly trained, only the one trained in current interval t_L is utilized to make predictions.

4.4.2 Task-specific Multi-task Learning

Suppose that there are L tasks indexed from 1 to L and, for each task l, the training set consists of N_l i.i.d. samples $\{(\boldsymbol{x}_i^l, y_i^l)\}_{i=1}^{N_l}$, where $\boldsymbol{x}_i^l \in \mathbb{R}^{P_0+P_l}$ represents the i-th training data and $y_i^l \in \mathbb{R}$ denotes the corresponding output. P_0 and P_l are the

number of shared features (shared classical videos) and the number of task-specific features (local transient videos) for task l, respectively. Let $\boldsymbol{X}^l = [\boldsymbol{x}_1^l, \boldsymbol{x}_2^l, \cdots, \boldsymbol{x}_{N_l}^l]^{\mathrm{T}} \in \mathbb{R}^{N_l \times (P_0 + P_l)}$ be the data matrix for task l and $\boldsymbol{y}^l = [y_1^l, y_2^l, \cdots, y_{N_l}^l]^{\mathrm{T}} \in \mathbb{R}^{N_l}$ be the corresponding output vector.

Let $\boldsymbol{w}^l = \begin{bmatrix} \boldsymbol{\alpha}^l \\ \boldsymbol{\beta}^l \end{bmatrix}$ be the parameter vector for task l, where $\boldsymbol{\alpha}^l = (\alpha_1^l, \cdots, \alpha_{P_0}^l)^{\mathrm{T}} \in \mathbb{R}^{P_0}$ and $\boldsymbol{\beta}^l = (\beta_1^l, \cdots, \beta_{P_l}^l)^{\mathrm{T}} \in \mathbb{R}^{P_l}$ are the model parameter vectors for shared and task-specific features, respectively. Let $\boldsymbol{A} = [\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \cdots, \boldsymbol{\alpha}^L]$ be the $P_0 \times L$ parameter matrix for the shared features across L tasks and $\boldsymbol{A}_k \in \mathbb{R}^L$ represent the k-th row of matrix $\boldsymbol{A}$. Let $\boldsymbol{B} = [\boldsymbol{\beta}^{1\mathrm{T}}, \cdots, \boldsymbol{\beta}^{L\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{(\sum_{l=1}^L P_l)}$ be the parameter vector for the task-specific features of L tasks. Figure 4.3 provides a pictorial representation of the features and parameters of the proposed model.

The ℓ_1/ℓ_2 norm regularization is introduced to globally select and estimate the shared features. Furthermore, the ℓ_1 norm regularization is employed to obtain element-wise sparsity in task-specific features. Putting them together with logistic regression would yield the optimization problem

$$\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} J(\boldsymbol{x}_i^l, y_i^l, \boldsymbol{w}^l) + \lambda_1 \|\boldsymbol{A}\|_{2,1} + \lambda_2 \|\boldsymbol{B}\|_1 \tag{4.1}$$

where $J(\boldsymbol{x}_i^l, y_i^l, \boldsymbol{w}^l) = -\left(y_i^l \cdot \boldsymbol{x}_i^{l\mathrm{T}} \boldsymbol{w}^l - \log(1 + e^{\boldsymbol{x}_i^{l\mathrm{T}} \boldsymbol{w}^l})\right)$ is the negative log-likelihood, $\|\boldsymbol{A}\|_{2,1} = \sum_{k=1}^{P_0} \|\boldsymbol{A}_k\|_2$ is the ℓ_1/ℓ_2 norm of the matrix $\boldsymbol{A}$ and $\|\boldsymbol{B}\|_1 = \sum_{l=1}^{L} \|\boldsymbol{\beta}^l\|_1$ is the ℓ_1 penalty introduced for task-specific features. Note that Expression (4.1) is the sum of convex functions and is therefore convex.

Here, the ℓ_1/ℓ_2 norm is utilized to regularize the shared features in the proposed method. It performs joint covariate selection that, depending on the tuning parameter λ_1, an entire group of shared features may enter into or drop out of the multiple models simultaneously [4, 47, 51, 62]. One obvious extension is to use the ℓ_p norms for $1 \leq p \leq \infty$ and generalize to ℓ_1/ℓ_p norm regularizations [51]. Modelers could choose

p based on how much a priori feature sharing among the tasks, from none ($p = 1$) to complete ($p = \infty$) [51]. The ℓ_1 norm regularization is also applied on task-specific features to obtain feature-wise sparsity [68].

The new method relaxes the constraints of conventional multi-task learning methods by introducing the task-specific features. It allows different models to seek common ground while reserving differences. With the proposed method, modelers are encouraged to discover and leverage important task-specific information and domain knowledge, which are usually very helpful for improving the performance.

4.4.3 Bayes Testing and Decision Procedure

It is important for industry people to get reliable and high-precision results. Classical classification algorithms may label all the audiences and provide us with unsatisfactory precision. Inspired by multiple hypothesis testing, this section formulates the classification problem as two separate detection problems:

Q1: Female Detection: which users are female users?

Q2: Male Detection: which users are male users?

The intuition is to label the user only when people are confident. Users without sufficient viewing data or evidence support will be marked as indecision currently and labeled later [67]. For both questions, it aims to find as many desired users as possible, subject to the constraint that the false discovery rate (FDR) or Type I error rate [7, 66] is controlled at a user-specified level α, namely, FDR $\leq \alpha$.

Given the estimated parameter $\hat{\boldsymbol{w}}^L = \begin{bmatrix} \hat{\boldsymbol{\alpha}}^L \\ \hat{\boldsymbol{\beta}}^L \end{bmatrix}$, we have

$$\begin{cases} P(y_i = Female|\boldsymbol{x}_i; \hat{\boldsymbol{w}}^L) = \dfrac{e^{\boldsymbol{x}_i^{\mathrm{T}} \cdot \hat{\boldsymbol{w}}^L}}{1 + e^{\boldsymbol{x}_i^{\mathrm{T}} \cdot \hat{\boldsymbol{w}}^L}} \\ P(y_i = Male|\boldsymbol{x}_i; \hat{\boldsymbol{w}}^L) = \dfrac{1}{1 + e^{\boldsymbol{x}_i^{\mathrm{T}} \cdot \hat{\boldsymbol{w}}^L}} \end{cases}. \tag{4.2}$$

Bayes testing and decision procedures are proposed for **Q1** and **Q2**, respectively.

A Bayes testing and decision procedure for ***Q1***:

1. Order users by $P(y_i = Male|\boldsymbol{x}_i; \hat{\boldsymbol{w}}^L) = 1 - P(y_i = Female|\boldsymbol{x}_i; \hat{\boldsymbol{w}}^L)$ in an ascending order and denote them by $\mu^{(1)}, \mu^{(2)}, \cdots, \mu^{(N)}$.

2. Let $k = \max\{j : \frac{1}{j}\sum_1^j \mu^{(j)} \leq \alpha\}$.

3. Report users U_i ($U_i \in \mathcal{G}^{Female}$) as female users, where $\mathcal{G}^{Female} = \{i : P(y_i = Male|\boldsymbol{x}_i; \hat{\boldsymbol{w}}^L) \leq \mu^{(k)}\}$.

A Bayes testing and decision procedure for ***Q2***:

1. Order users by $P(y_i = Female|\boldsymbol{x}_i; \hat{\boldsymbol{w}}^L) = 1 - P(y_i = Male|\boldsymbol{x}_i; \hat{\boldsymbol{w}}^L)$ in an ascending order and denote them by $\nu^{(1)}, \nu^{(2)}, \cdots, \nu^{(N)}$.

2. Let $k = \max\{j : \frac{1}{j}\sum_1^j \nu^{(j)} \leq \alpha\}$.

3. Report users U_i ($U_i \in \mathcal{G}^{Male}$) as male users, where $\mathcal{G}^{Male} = \{i : P(y_i = Female|\boldsymbol{x}_i; \hat{\boldsymbol{w}}^L) \leq \nu^{(k)}\}$.

As shown in the experiments, the proposed Bayes decision procedure could precisely control the FDR at the nominal level with valid posterior probabilities.

4.5 Experiment

4.5.1 Experiment Settings

This section runs real data experiments to investigate the numerical performance of the proposed method. It compares the performance of four different methods:

M1: naive Bayes classifier (nBayes);

M2: feature selection + naive Bayes classifier (F+nBayes);

M3: logistic regression with ℓ_1 penalty (LR);

M4: task-specific multi-task learning method (tMulti).

Naive Bayes classifier and logistic regression with ℓ_1 penalty are two popular approaches in the world of big data modeling [45] and have been widely applied by various practitioners to solve practical problems in industry. In addition, they provide probability estimates[1], which is desired for controlling the FDR or Type I error rate. In the experiment, Laplace smoothing is applied to naive Bayes classifiers to obtain smoother and better performance. Since logistic regression with ℓ_1 penalty (**M3**) and the proposed multi-task learning method (**M4**) do variable selection automatically, to compare them fairly, naive Bayes classifier with feature selection is also included as a competing method in the experiment. Fisher's exact test [23] is conducted for each video to select those with gender preference (p-value ≤ 0.05) as input features for the naive Bayes classifier (**M2**). Through feature selection, it could greatly eliminate the irrelevant features, and therefore, improve performance. Note that the proposed method (**M4**) will reduce to the ℓ_1-regularized logistic regression (**M3**) if all features are treated as task-specific features.

Three metrics are applied to evaluate the performance of competing methods:

Metric 1: area under the ROC curve (AUC);

Metric 2: Precision@K and Recall@K for the identified top K female users;

Metric 3: Sensitivity at the nominal FDR level α.

Metric 1 measures the discrimination, that is, the ability of a classifier to correctly classify those female or male users. Since industry people are more interested in

[1]SVM doesn't directly provide probability estimates. Although Platt scaling could be utilized to calibrate the binary SVM's scores by fitting an additional logistic regression on the scores, it is known to have theoretical issues and the probability estimates may be inconsistent with the scores [55, 81].

detecting female users, **Metric 2** is introduced to clearly evaluate how well a model can identify female users. Let z_M and z_F denote the collection of male users and female users, respectively. Sort the predicted users in descending order according to their posterior probabilities $P(y_i = Female|\boldsymbol{x}_i)$ and report the top K users (denoted as z_K) as identified female users. Let Precision@K $= \frac{|z_K \cap z_F|}{K}$ and Recall@K $= \frac{|z_K \cap z_F|}{|z_F|}$. **Metric 3** measures the sensitivity (or recall) for each method at the nominal FDR level α. As a part of the commercial system, one of the most important metrics is the reliability that it could provide controllable results. **Metric 3** is designed for this purpose and evaluates the competing methods in two aspects, namely, whether the method could control the FDR at the nominal level, and what the sensitivity is.

It sampled 35187, 70031 and 78996 registered users in August, September and October 2015, respectively, and extracted more than 20 million video viewing records from PPTV. Invalid records, e.g., viewing time < 5 seconds, were removed, as such records were generated by accidentally clicking on a video. Efficiently leveraging the historical data plays an important role in obtaining decent performance. Here, it considers up to 3 weeks' previous data and empirically takes each week as a time interval. It is natural for the proposed method to jointly train multiple models based on previous weeks' and the current week's data. Classical movies and TV dramas are treated as shared features, while transient ordinary videos like news, sports, and stars and entertainment are treated as task-specific features. However, for naive Bayes method and logistic regression, there is no single preferred way to utilize the historical data. To compare them completely, different ways of leveraging the historical data are investigated for these competing methods. Specifically, it considers three strategies:

S1: use the current data only;

S2: use both the current data and the previous data;

S3: use the current data and only part of the previous data.

Table 4.3 Summary of the Testing Users

Testing Users	Aug E1	Aug E2	Aug E3	Sept E1	Sept E2	Sept E3	Oct E1	Oct E2	Oct E3
Male	3883	3946	3919	8087	8069	8065	6405	6348	6373
Female	1397	1353	1339	3033	2951	3051	2404	2367	2388
Total	5280	5299	5258	11120	11020	11116	8809	8715	8761

For **S1**, it only uses the current week's data to train the model and drop all previous data. **S2** and **S3** both leverage the previous data but in different manners. **S2** utilizes all previous data and video viewing records, while **S3** only extracts and utilizes the viewing records whose corresponding videos are viewed by some audiences in the current week.

4.5.2 Experiment Results

Experiments are repeatedly conducted 9 times from August to October. Each time it randomly chooses a time point. The 7 days after the selected time are considered as the "current week", while the data collected before it are treated as "previous weeks' data". As mentioned earlier, only up to 3 weeks' previous data are considered in the experiments. The current week's newly sampled users are treated as testing users and their viewing records, which are only located in the current week, are used as testing data. Other users are treated as training users and their viewing records collected in both current and previous weeks are treated as training data. Note that there is no overlap between the testing users and the training users.

The parameters of different methods, e.g., regularization parameters, are tuned through 5-fold cross-validation with the training data. Table 4.3 summarizes the testing users sampled at each testing point. It randomly performs three experiments in each month denoted as **Aug E1**, **Aug E2**, **Aug E3** and so on. Note that more than $75,000$ testing users are used to evaluate the competing methods in the experiments.

Table 4.4 Performance of the Competing Methods in Terms of Area under the ROC Curve (AUC)

Method	Aug E1	Aug E2	Aug E3	Sept E1	Sept E2	Sept E3	Oct E1	Oct E2	Oct E3
nBayes+S1	*0.7246*	*0.7251*	*0.7383*	*0.7386*	*0.7279*	*0.7312*	*0.7164*	*0.7321*	*0.7080*
nBayes+S2	0.6080	0.6182	0.6132	0.6795	0.6672	0.6751	0.5662	0.5779	0.5803
nBayes+S3	0.6124	0.6262	0.6194	0.6800	0.6680	0.6758	0.5714	0.5859	0.5918
F+nBayes+S1	*0.7268*	*0.7292*	*0.7415*	*0.7425*	*0.7296*	*0.7341*	*0.7217*	*0.7379*	*0.7186*
F+nBayes+S2	0.6143	0.6256	0.6208	0.6804	0.6681	0.6745	0.5687	0.5806	0.5831
F+nBayes+S3	0.6187	0.6334	0.6275	0.6808	0.6688	0.6752	0.5728	0.5889	0.5931
LR+S1	*0.7365*	*0.7345*	*0.7391*	*0.7536*	*0.7511*	*0.7473*	*0.7360*	*0.7570*	*0.7297*
LR+S2	0.6572	0.6554	0.6681	0.6966	0.7007	0.7033	0.6165	0.6330	0.6313
LR+S3	0.6581	0.6571	0.6669	0.6979	0.7018	0.7042	0.6193	0.6368	0.6345
tMulti	**0.7545**	**0.7467**	**0.7601**	**0.7674**	**0.7646**	**0.7650**	**0.7610**	**0.7764**	**0.7551**

It costs approximately 5 hours for a 64-bit server with two 6-Core 2.93GHz CPUs and 60 GB RAM to train the proposed model.

Metric 1 Table 4.4 summarizes the experiment results of the proposed method (tMulti) and other competing methods (nBayes, F+nBayes and LR) combined with different strategies (**S1**, **S2** and **S3**) in terms of the area under the ROC curve (AUC). Here are a few remarks. Firstly, method combined with strategy 1 (**S1**) yields much better performance than the corresponding method combined with other strategies (**S2** and **S3**). This observation is consistent among logistic regression with ℓ_1 penalty (LR) and naive Bayes methods (nBayes and F+nBayes). As stated earlier, data collected at different time intervals usually exhibit significantly different patterns (See Table 4.2). Not only the discriminant features could change dramatically, but also their coefficients may shift over time. Thus, naively leveraging the previous data will have an adverse effect. Secondly, logistic regression with ℓ_1 norm regularization outperforms the naive Bayes methods in most cases. Thirdly, the proposed method (tMulti) demonstrates the best performance among all competing methods. It provides an advanced way to leverage the historical data. For different time intervals,

Table 4.5 Precision and Recall over Top K (K=250, 500, 750 and 1000) Identified Female Users

Precision@250	**Aug E1**	**Aug E2**	**Aug E3**	**Sept E1**	**Sept E2**	**Sept E3**	**Oct E1**	**Oct E2**	**Oct E3**
nBayes+S1	0.644	0.628	0.632	0.676	0.616	0.708	0.636	0.648	0.620
F+nBayes+S1	0.652	0.636	0.656	0.668	0.624	0.700	0.632	0.640	0.624
LR+S1	0.720	0.736	0.736	**0.752**	0.704	0.768	0.768	0.744	0.716
tMulti	**0.780**	**0.748**	**0.740**	**0.752**	**0.708**	**0.788**	**0.792**	**0.768**	**0.740**
Recall@250	**Aug E1**	**Aug E2**	**Aug E3**	**Sept E1**	**Sept E2**	**Sept E3**	**Oct E1**	**Oct E2**	**Oct E3**
nBayes+S1	0.115	0.116	0.118	0.056	0.052	0.058	0.066	0.068	0.065
F+nBayes+S1	0.117	0.118	0.122	0.055	0.053	0.057	0.066	0.068	0.065
LR+S1	0.129	0.136	0.137	**0.062**	0.059	0.063	0.080	0.079	0.075
tMulti	**0.140**	**0.138**	**0.138**	**0.062**	**0.060**	**0.065**	**0.082**	**0.081**	**0.077**
Precision@500	**Aug E1**	**Aug E2**	**Aug E3**	**Sept E1**	**Sept E2**	**Sept E3**	**Oct E1**	**Oct E2**	**Oct E3**
nBayes+S1	0.586	0.570	0.606	0.642	0.608	0.664	0.600	0.600	0.584
F+nBayes+S1	0.602	0.604	0.626	0.638	0.596	0.646	0.606	0.614	0.588
LR+S1	0.656	0.630	0.616	0.712	0.674	0.734	0.710	0.710	0.680
tMulti	**0.674**	**0.636**	**0.648**	**0.724**	**0.698**	**0.744**	**0.724**	**0.714**	**0.694**
Recall@500	**Aug E1**	**Aug E2**	**Aug E3**	**Sept E1**	**Sept E2**	**Sept E3**	**Oct E1**	**Oct E2**	**Oct E3**
nBayes+S1	0.210	0.211	0.226	0.106	0.103	0.109	0.125	0.127	0.122
F+nBayes+S1	0.215	0.223	0.234	0.105	0.101	0.106	0.126	0.130	0.123
LR+S1	0.235	0.233	0.230	0.117	0.114	0.120	0.148	0.150	0.142
tMulti	**0.241**	**0.235**	**0.242**	**0.119**	**0.118**	**0.122**	**0.151**	**0.151**	**0.145**
Precision@750	**Aug E1**	**Aug E2**	**Aug E3**	**Sept E1**	**Sept E2**	**Sept E3**	**Oct E1**	**Oct E2**	**Oct E3**
nBayes+S1	0.561	0.540	0.557	0.612	0.592	0.639	0.561	0.581	0.553
F+nBayes+S1	0.552	0.545	0.577	0.605	0.591	0.631	0.572	0.584	0.556
LR+S1	0.592	0.559	0.575	0.685	0.660	0.687	0.655	0.641	0.647
tMulti	**0.624**	**0.579**	**0.597**	**0.704**	**0.685**	**0.692**	**0.676**	**0.677**	**0.653**
Recall@750	**Aug E1**	**Aug E2**	**Aug E3**	**Sept E1**	**Sept E2**	**Sept E3**	**Oct E1**	**Oct E2**	**Oct E3**
nBayes+S1	0.301	0.299	0.312	0.151	0.150	0.157	0.175	0.184	0.174
F+nBayes+S1	0.296	0.302	0.323	0.150	0.150	0.155	0.178	0.185	0.175
LR+S1	0.318	0.310	0.322	0.169	0.168	0.169	0.204	0.203	0.203
tMulti	**0.335**	**0.321**	**0.335**	**0.174**	**0.174**	**0.170**	**0.211**	**0.215**	**0.205**
Precision@1000	**Aug E1**	**Aug E2**	**Aug E3**	**Sept E1**	**Sept E2**	**Sept E3**	**Oct E1**	**Oct E2**	**Oct E3**
nBayes+S1	0.527	0.502	0.530	0.610	0.585	0.606	0.544	0.566	0.532
F+nBayes+S1	0.525	0.517	0.529	0.592	0.583	0.606	0.562	0.569	0.538
LR+S1	0.547	0.519	0.541	0.658	0.634	0.654	0.613	0.614	0.609
tMulti	**0.567**	**0.540**	**0.553**	**0.678**	**0.663**	**0.669**	**0.626**	**0.632**	**0.616**
Recall@1000	**Aug E1**	**Aug E2**	**Aug E3**	**Sept E1**	**Sept E2**	**Sept E3**	**Oct E1**	**Oct E2**	**Oct E3**
nBayes+S1	0.377	0.371	0.396	0.201	0.198	0.199	0.226	0.239	0.223
F+nBayes+S1	0.376	0.382	0.395	0.195	0.198	0.199	0.234	0.240	0.225
LR+S1	0.392	0.384	0.404	0.217	0.215	0.214	0.255	0.259	0.255
tMulti	**0.406**	**0.399**	**0.413**	**0.224**	**0.225**	**0.219**	**0.260**	**0.267**	**0.258**

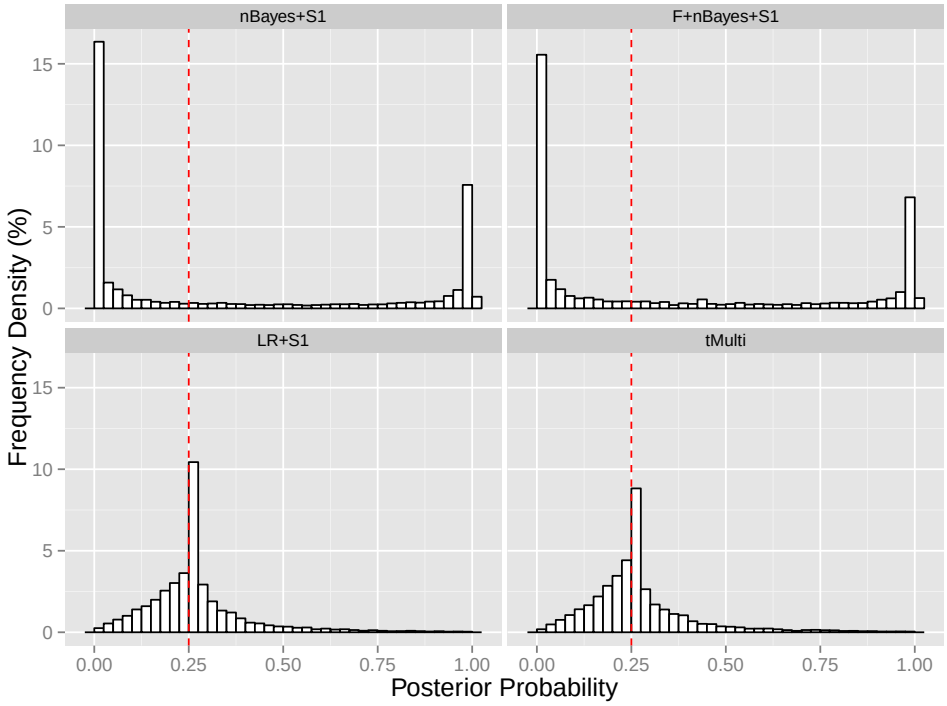

Figure 4.4 Histograms of the posterior probabilities $P(y_i = Female|\boldsymbol{x}_i)$.

it introduces the task-specific features to capture those important but transient videos and jointly estimate the shared features to obtain better performance.

Metric 2 The Precision@K and Recall@K for top K (K=250, 500, 750 and 1000) identified female users are summarized in Table 4.5. It should be noted that accurately detecting female users is more challenging, as the ratio of women to men is approximately 1 : 3 and it only yields a prediction accuracy of 0.25 for female users with a random guess. In Table 4.5, it only reports the results for methods combined with strategy 1 (**S1**), as strategy 1 (**S1**) yields the best performance for other competing methods (See Table 4.4). The experiments conducted in September and October may have larger precisions but with lower recalls than the results in August. This is because there are more female testing users in September and October (See Table 4.3). It is obvious that logistic regression with ℓ_1 norm regularization outperforms the naive Bayes methods in most cases, as ℓ_1 norm regularization could

greatly reduce the risk of overfitting and yields better coefficient estimation. In addition, such regularizations could help to obtain valid posterior probabilities for testing users and enable reliable decisions (See **Metric 3**). As shown in Table 4.5, the proposed method (tMulti) beats all of other methods for both precision and recall in all settings (K=250, 500, 750 and 1000). It gains approximately 2% and more than 5% improvements in precision compared with logistic regression with ℓ_1 penalty (LR+S1) and naive Bayes methods (nBayes+S1 and F+nBayes+S1), respectively. Considering there are more than 190 million monthly active users watching videos on PPTV, such improvements could bring a dramatic increase in commercial value.

Metric 3 A month is randomly selected to evaluate **Metric 3**. The last week of the selected month is treated as the "current week", while the previous 3 weeks are treated as "previous weeks". It repeats the experiment in this setting 10 times

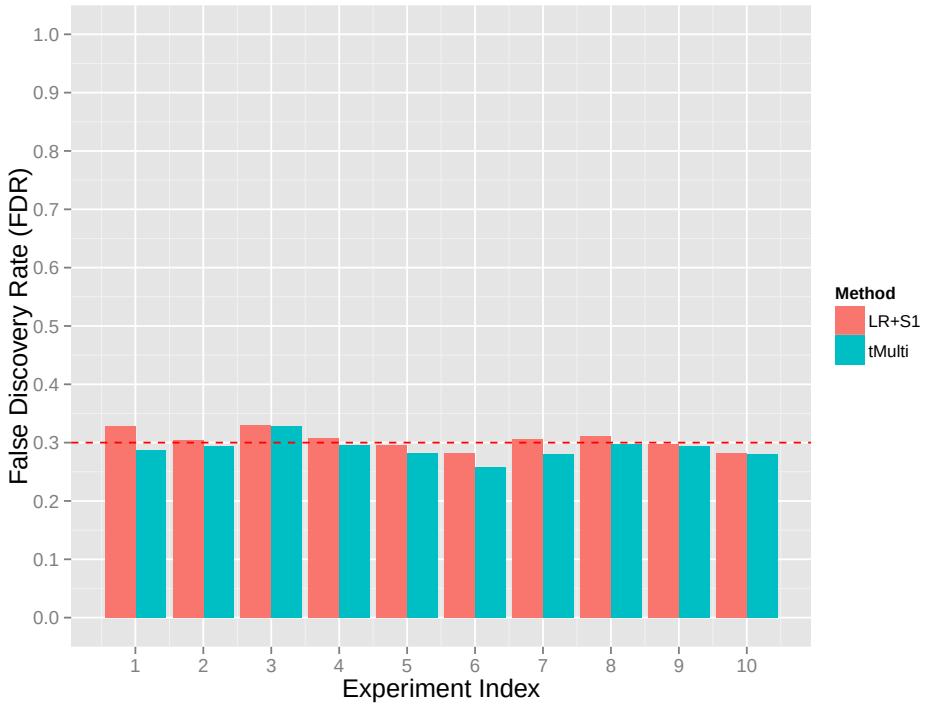

Figure 4.5 False discovery rates for 10 random experiments (at nominal level $\alpha = 0.3$).

to evaluate the **Metric 3**. Each time, it randomly divides the sampled users into training users (80%) and testing users (20%) in advance and remove all testing users' viewing records from previous data. Figure 4.4 plots the histograms of posterior probabilities $P(y_i = Female|\boldsymbol{x}_i)$ of testing users for different methods. The red dashed vertical line indicates the proportion of female users (approximately 25%) in the testing data. Obviously, regularization-based methods (LR+S1 and tMulti) exhibit more reasonable distribution of posterior probabilities among all testing users. In contrast, naive Bayes methods (nBayes+S1 and F+nBayes+S1), which make unrealistic independence assumptions, push probabilities toward 0 and 1 and over-estimate the posterior probabilities $P(y_i = Female|\boldsymbol{x}_i)$ [50]. As shown in Figure 4.4 (nBayes+S1 and F+nBayes+S1), the majority of testing users are located in intervals $[0, 0.1]$ and $[0.9, 1]$. Such over-estimations incur great difficulty for researchers to choose a reasonable cutoff to classify users accurately. For example, if they choose 0.9 as a cutoff and report the users with $P(y_i = Female|\boldsymbol{x}_i) \geq 0.9$ as female users, naive Bayes methods will still report too many users and the precision is usually less than 0.5. Thus, the naive Bayes method will not be described in detail in following paragraphs, as it can't guarantee any multiplicity control.

Figure 4.5 summarizes the false discovery rates of competing methods (LR+S1 and tMulti). To save space, it presents the results at the nominal FDR level $\alpha = 0.3$ for female users as an example. Note that it delivers similar results for male users or other FDR levels. Clearly, both methods when combined with the Bayes decision procedure could precisely control FDR at the nominal level $\alpha = 0.3$, which demonstrates both the validity of posterior probabilities and the effectiveness of the proposed Bayes decision procedure. The sensitivities of competing methods at the nominal FDR level $\alpha = 0.3$ are summarized in Table 4.6. To clearly present the results, the improvements of the proposed method (tMulti) are also calculated. Obviously, the proposed method could achieve 7.26% improvements on average in

Table 4.6 Sensitivities of Competing Methods (LR+S1 and tMulti) at the Nominal FDR Level $\alpha = 0.3$ for 10 Random Experiments E1-10

Method	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**	**E10**
LR+S1	0.128	0.155	0.137	0.143	0.143	0.140	0.146	0.142	0.152	0.145
tMulti	0.149	0.162	0.146	0.158	0.152	0.146	0.155	0.153	0.160	0.152
Improvement	**16.4%**	**4.5%**	**6.6%**	**10.5%**	**6.3%**	**4.3%**	**6.2%**	**7.7%**	**5.3%**	**4.8%**

terms of the sensitivity (or recall) while controlling the FDR (or Type I error rate) at the nominal level.

4.6 Conclusion

This chapter investigates the feasibility and challenges of gender prediction based on users' video viewing behavior. It proposes a novel task-specific multi-task learning algorithm to efficiently leverage training data and obtain decent performance. Inspired by multiple hypothesis testing, it further proposes Bayes decision procedures to identify female and male users, respectively, which could precisely control the Type I error rate at a user-specified level. Experiment results have justified the effectiveness and reliability of the proposed method.

CHAPTER 5

AN EMPIRICAL BAYES CHANGE-POINT MODEL FOR IDENTIFYING 3′ AND 5′ ALTERNATIVE SPLICING BY NEXT-GENERATION RNA SEQUENCING

5.1 Introduction

Next-generation RNA sequencing (RNA-seq) has been widely used to investigate alternative isoform regulations. Among them, alternative 3′ splice site (SS) and 5′ SS account for more than 30% of all alternative splicing (AS) events in higher eukaryotes. Recent studies have revealed that they play important roles in building complex organisms and have a critical impact on biological functions which could cause disease. Quite a few analytical methods have been developed to facilitate alternative 3′ SS and 5′ SS studies using RNA-seq data. However, these methods have various limitations and their performances may be further improved.

This chapter proposes an empirical Bayes change-point model for identifying 3′/5′ AS events. The new approach requires no annotation information and is applicable to detect novel aberrant splicing events. Compared with previous methods, it has several unique merits. First of all, it does not rely on annotation information. Instead, it provides for the first time a systematic framework to integrate read coverage information and junction read or annotation information, when available, in order to obtain better performance. Secondly, an empirical Bayes model is utilized to efficiently pool information across genes for improving detection efficiency. Thirdly, it provides a flexible testing framework in which the user can choose to address different levels of questions, namely, whether alternative splicing happens, and/or where it happens. Simulation studies and applications to real data have demonstrated that the proposed method is powerful and accurate.

5.2 Methods

5.2.1 Alternative 3′ SS and 5′ SS Selection and Change-point Problem

The alternative 3′ SS and 5′ SS selection problem and the change-point model are illustrated via toy examples in Figure 5.1. As shown in Figure 5.1, different mRNA isoforms, with different expression levels, are generated from a single gene through the alternative selection of 3′ SS or 5′ SS. The common regions (constitutive exons), shared by the two isoforms, are expected to have a higher expression level than the extended regions (spliced regions). As a result, for RNA-seq data, the common regions will have higher short read densities than the extended regions, and the exons with alternative 3′ SS or 5′ SS will have change-points at their 3′ or 5′ splice sites as illustrated in Figure 5.1. Thus, researchers can detect exons with alternative 3′ SS and 5′ SS and their splice sites by detecting the change-points where the short read densities change.

Let $S_i = (x_{i1}, x_{i2}, \cdots, x_{iL_i})$ be a sequence of observations ordered in position, where x_{ij} is the number of reads (read-count) whose first base mapped to exon i at position j. Following previous literature [5, 17, 86], the change-points divide the sequence of observations into K unknown homogeneous segments, $\Pi = (\Pi_1, \cdots, \Pi_K)$, such that the data is independent across different segments

$$p(S_{i,1:L_i}|\Pi) = \prod_{k=1}^{K} p(S_{i,\Pi_k}).$$

For the alternative 3′ SS and 5′ SS problem, it further assumes $K = 1$ or 2, namely, expecting there is at most one change-point in a read-count sequence. Let $\rho_i \in \{0, 1, 2, \cdots, L_i - 1\}$ denote the change-point position for sequence S_i. Specifically, $\rho_i = 0$ indicates there is no change-point and $\rho_i = \tau (\tau > 0)$ means that there is a change-point at position τ, before which read-counts in $S_{i,1:\tau} = (x_{i1}, \cdots, x_{i\tau})$ follow one homogeneous distribution and after which read-counts in $S_{i,(\tau+1):L_i} =$

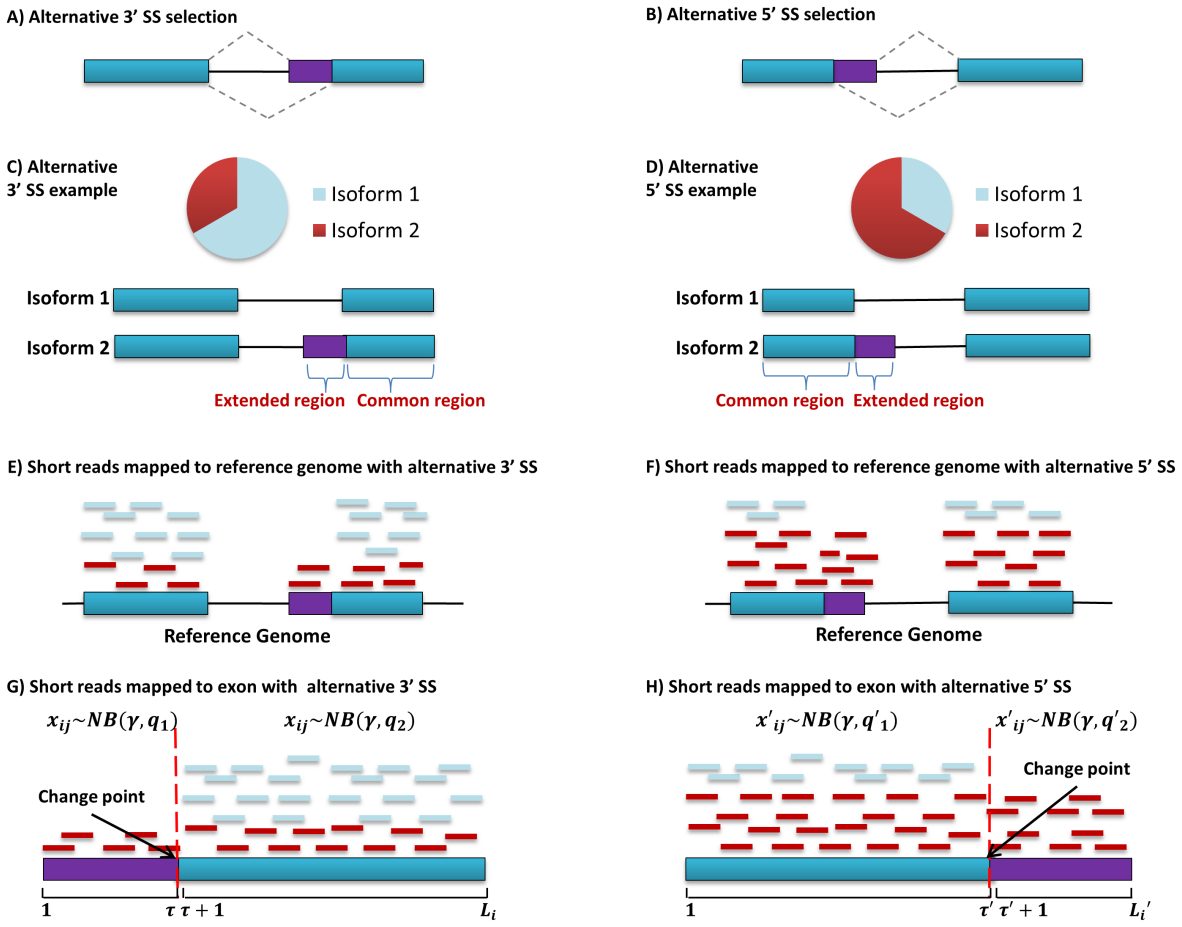

Figure 5.1 Illustration and notation of change-point model for alternative 3′ SS and 5′ SS problem. **A)** and **B)** show two AS events: alternative 3′ SS and 5′ SS selection, respectively. Blue rectangles represent constitutive exons (common regions) and purple rectangles represent alternatively spliced regions (extended regions). Solid lines and dashed lines indicate the introns and splicing options, respectively. **C)** and **D)** are examples of isoforms generated from alternative 3′ SS and 5′ SS selection, respectively. In **C)**, isoform 1 has a higher expression level, while, in **D)**, isoform 2 has a higher expression level. **E)** and **F)** show the results of mapping short reads to the reference genome, respectively. The reads from isoform 2 are marked as dark red, while reads from isoform 1 are marked as blue. **G)** and **H)** show the detailed results of the exons that contain alternative 3′ SS and 5′ SS. Because of the alternative 3′ SS or 5′ SS, the common region shared by the two isoforms has a higher gene expression level than the extended region. Thus, the average number of short reads (read-count) mapped to the common region will be larger than the one for extended region. This generates a change-point at the splice site, which partitions the whole region into two different homogeneous segments with different average read-counts.

$(x_{i(\tau+1)}, \cdots, x_{iL_i})$ follow another homogeneous distribution, as described in details below.

5.2.2 Negative Binomial-Beta Model

Considering the over-dispersion of RNA-seq data, the Negative Binomial (NB) distribution is utilized to characterize the observed read-counts for each segment. As shown in Figure 5.1, if there is no alternative $3'$ SS or $5'$ SS in exon i, the read-counts across the whole exon i are generated from a single model $NB(r, q_{i0})$. Otherwise, the splice site τ divides the read-count sequence into two homogeneous parts such that two Negative Binomial distributions $NB(r, q_{i1})$ and $NB(r, q_{i2})$ are involved in modeling the data. Formally,

$$x_{ij}|\rho_i \sim \begin{cases} NB(r, q_{i0}), & \text{if } \rho_i = 0, \\ NB(r, q_{i1}), & \text{if } \rho_i = \tau \text{ and } j \leq \tau, \\ NB(r, q_{i2}), & \text{if } \rho_i = \tau \text{ and } j > \tau, \end{cases}$$

where $q_{ik} \sim Beta(\alpha, \beta)$. The hierarchical structure of this Negative Binomial-Beta model is illustrated in Figure 5.2. Differently from a conventional Bayesian approach, the hyperparameters α and β are estimated from the data using an empirical Bayes approach. From the Negative Binomial-Beta model, the probability density function is calculated as

$$\begin{aligned} f(x_{ij}|r, q_i) &= \binom{x_{ij} + r - 1}{x_{ij}} q_i^r (1 - q_i)^{x_{ij}} \\ f(q_i|\alpha, \beta) &= \frac{q_i^{\alpha-1}(1 - q_i)^{\beta-1}}{B(\alpha, \beta)}. \end{aligned} \tag{5.1}$$

Integrating out the unknown segment specific parameter q_i, the likelihood of x_{ij} is

$$\begin{aligned} f(x_{ij}|r, \alpha, \beta) &= \int_q f(x_{ij}|r, q_i) \times f(q_i|\alpha, \beta) \mathrm{d}q \\ &= \binom{x_{ij} + r - 1}{x_{ij}} \frac{B(r + \alpha, x_{ij} + \beta)}{B(\alpha, \beta)}. \end{aligned} \tag{5.2}$$

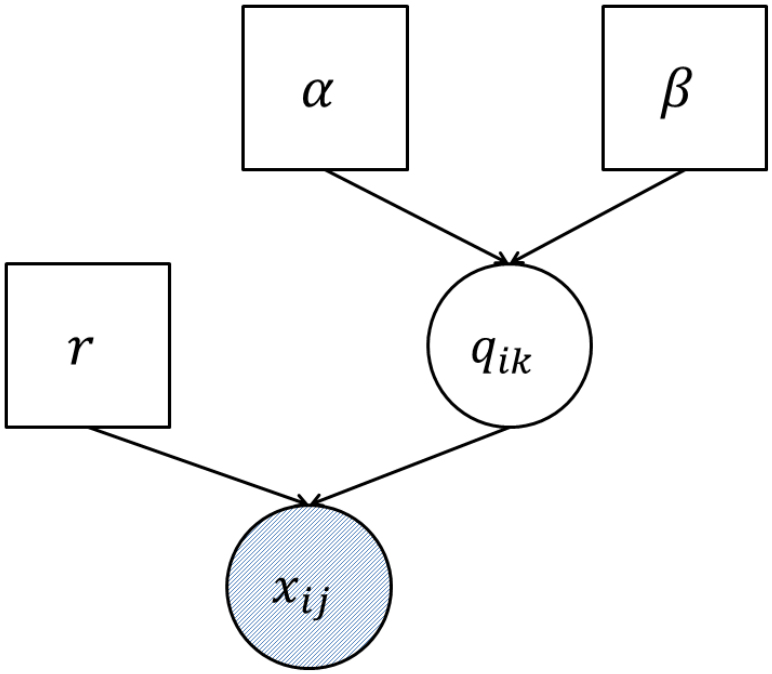

Figure 5.2 Hierarchical structure of Negative Binomial-Beta model.

Since observations in the same segment are independently and identically distributed (i.i.d.), the likelihood for a homogeneous segment $S_{i,j:k} = (x_{ij}, x_{i(j+1)} \cdots, x_{ik})$ can be computed as

$$\begin{aligned} f_0(S_{i,j:k}|r,\alpha,\beta) &= \int_{q_i} \prod_{l=j}^{k} f(x_{il}|r,q_i) \times f(q_i|\alpha,\beta)\mathrm{d}q \\ &= \left[\prod_{l=j}^{k} \binom{x_{il}+r-1}{x_{il}}\right] \\ &\times \frac{B\left((k-j+1)r+\alpha, \sum_{l=j}^{k} x_{il}+\beta\right)}{B(\alpha,\beta)}. \end{aligned} \tag{5.3}$$

If there is no change-point in the sequence S_i, the likelihood is

$$f(S_i|\rho_i = 0, r, \alpha, \beta) = f_0(S_{i,1:L_i}|r,\alpha,\beta). \tag{5.4}$$

When there is a change-point at τ, the likelihood is

$$f(S_i|\rho_i = \tau, r, \alpha, \beta) = f_0(S_{i,1:\tau}|r, \alpha, \beta) \times f_0(S_{i,(\tau+1):L_i}|r, \alpha, \beta). \tag{5.5}$$

5.2.3 Prior Information and Hot Points

When no additional information is available, every position has the same prior probability of being a change-point. Suppose that each sequence S_i has a change-point with a prior probability P, then the prior probability for each candidate position is

$$Pr(\rho_i; P) = \begin{cases} 1 - P, & \text{if } \rho_i = 0, \\ \frac{P}{L_i - 1}, & \text{if } \rho_i = 1, 2, \cdots, L_i - 1. \end{cases}$$

If additional information is available, e.g., splice junction reads or isoform annotation, the proposed method assigns different weights to different candidate positions allowing them to have different prior probabilities as derived from extra information. It assigns weight $W \geq 1$, which can be estimated from data or pre-specified by the user, to hot points and weight 1 to ordinary positions. Then the prior probability for each position will be

$$Pr(\rho_i; P, W) = \begin{cases} 1 - P, & \text{if } \rho_i = 0, \\ P \times \frac{w_{i\rho_i}}{\sum_{j=1}^{L_i - 1} w_{ij}}, & \text{if } \rho_i = 1, 2, \cdots, L_i - 1, \end{cases}$$

where w_{ij} is the weight assigned to position j in sequence i, (i, j) for short, and

$$w_{ij} = \begin{cases} 1, & \text{if } (i, j) \text{ is ordinary position}, \\ W, & \text{if } (i, j) \text{ is hot point}. \end{cases}$$

This weighting scheme allows flexible weight assigning strategies. It is very useful when the user has different kinds of prior information. Assuming that various information has additive effects on the weight of a candidate position, the proposed method can make full use of all kinds of information. Suppose there are m different

kinds of information, it assigns weights to candidate positions as follows

$$w_{ij} = 1 + \beta^{(1)}\delta_{ij}^{(1)} + \beta^{(2)}\delta_{ij}^{(2)} + \cdots + \beta^{(m)}\delta_{ij}^{(m)},$$

where $\beta^{(k)}$ $(k = 1, 2, \cdots, m)$ measures the additive effect of information k and $\delta_{ij}^{(k)} = I$ {position (i, j) supported by information k}. Moreover, interactions of different information can be considered and added into the weight assigning procedure by introducing interaction terms $\beta^{(kl)}\delta_{ij}^{(kl)}$ $(k, l = 1, 2, \cdots, m)$. For example, given the annotations and splicing reads, a weight assigning strategy can be

$$w_{ij} = 1 + \beta^{(1)}\delta_{ij}^{(1)} + \beta^{(2)}\delta_{ij}^{(2)} + \beta^{(12)}\delta_{ij}^{(12)},$$

where $\beta^{(1)}$, $\beta^{(2)}$ and $\beta^{(12)}$ measure the additive effects of splice junction reads, isoform annotation and their interactions, respectively.

By assigning different weights to different candidate positions and distinguishing hot points from ordinary ones, it can efficiently leverage domain knowledge and improve the performance. Since all parameters are estimated from data through the empirical Bayes approach, it doesn't matter if the domain knowledge is dubious or totally wrong. It is noted that there are trade-offs between simple *versus* sophisticated strategies. Adopting more sophisticated strategy will make better usage of prior information on one hand, but on the other hand, it will introduce more parameters and cause difficulty in parameter estimation. Thus, appropriate strategies need to be chosen to balance these trade-offs for different applications.

5.2.4 Empirical Bayes Estimator

Empirical Bayes estimates combine the Bayesian and frequentist reasoning that the prior probability is estimated frequentistically in order to perform Bayesian inferences [19]. This kind of combination not only provides the Bayesian accurate, objective and data-related prior information, but also enables frequentists to obtain more

test efficiency in solving scientific problems [20, 91]. Let Φ denote the set of the parameters for the Negative Binomial-Beta model (r, α and β) and the parameters for characterizing prior information (P and W). The maximum likelihood estimation of Φ, applied to the total N sequences, is

$$\hat{\Phi} = \arg\max_{\Phi} \ \log\left(\prod_{i=1}^{N}\sum_{\rho_i=0}^{L_i-1} Pr(\rho_i|\Phi) f(S_i|\rho_i,\Phi)\right).$$

The optimization algorithm L-BFGS-B [13], a limited-memory modification of the BFGS quasi-Newton method with box constraints, is applied to estimate the parameters Φ.

5.2.5 Empirical Bayes Testing and Decision Procedure

There are two questions of interest that can be addressed by the proposed method:

Q1: Detection, which genes have change-points?

Q2: Identification, where is the change-point, if any?

For Q1, it only cares about whether there is a change-point or not, and is not concerned about where the change-point locates. For Q2, it aims to find the accurate location of the change-point. In other words, if it correctly detects a sequence with change-point but wrongly locates the change-point position, it is still considered as an error. Given the estimated parameters $\hat{\Phi}$, the posterior probability is

$$Pr(\rho_i = \tau|S_i;\hat{\Phi}) = \frac{Pr(\rho_i=\tau, S_i;\hat{\Phi})}{\sum_{j=0}^{L_i-1} Pr(\rho_i=j, S_i;\hat{\Phi})}.$$

Let

$$\begin{aligned}\pi_{i0} = &\quad Pr(\rho_i = 0|S_i;\hat{\Phi}),\\ \pi_i^* = &\quad \max\{Pr(\rho_i=\tau|S_i;\hat{\Phi})\}, \quad \tau = 1,2,\cdots,L_i-1.\end{aligned}$$

It is desirable to control the false discovery rate (FDR) [7] at a nominal level α and find as many sequences with change-points as possible. To obtain this goal, this

section proposes the following two empirical Bayes testing and decision procedures for Q1 and Q2, respectively.

An empirical Bayes testing and decision procedure for Q1:

1. Order sequences by π_{i0} in an ascending order and denote them by $\pi_0^{(1)}, \pi_0^{(2)}, \cdots, \pi_0^{(N)}$.
2. Let $k = \max\{j : \frac{1}{j}\sum_1^j \pi_0^{(j)} \leq \alpha\}$.
3. Report sequences S_i ($S_i \in \mathcal{G}^{Detection}$) to have a change-point, where $\mathcal{G}^{Detection} = \{i : \pi_{i0} \leq \pi_0^{(k)}\}$.

An empirical Bayes testing and decision procedure for Q2:

1. Order sequences by $(1 - \pi_i^*)$ in an ascending order and denote them by $\pi_*^{(1)}, \pi_*^{(2)}, \cdots, \pi_*^{(N)}$.
2. Let $k = \max\{j : \frac{1}{j}\sum_1^j \pi_*^{(j)} \leq \alpha\}$.
3. Report sequences S_i ($S_i \in \mathcal{G}^{Identification}$) to have a change-point at position τ_i^*, where $Pr(\rho_i = \tau_i^* | S_i; \hat{\Phi}) = \pi_i^*$, and $\mathcal{G}^{Identification} = \{i : (1 - \pi_i^*) \leq \pi_*^{(k)}\}$.

5.3 Experiment

5.3.1 Simulation Settings

This section first performs simulation studies to investigate the numerical performance of the proposed method. It randomly generates $N = 500$ sequences, each with length $L_i = 100$. $P * N$ sequences are selected to have change-points. It simulates two scenarios, the first one without hot points and the second one with hot points. For the first scenario, it randomly picks one position with equal probability to be the change-point for all the $P * N$ selected sequences. For the second scenario, it sets positions 25, 50 and 75 as hot points with weight $W = 32$ while the other points with weight $W = 1$. one half of the selected $P * N$ sequences have change-points

at hot points and the other half don't. As a result, there are $(1-P)*N+2*P*N=(1+P)N$ homogeneous sequence segments in total. The read-count data from each segment are generated from a Negative Binomial distribution with parameters estimated from real data [48]. Prior probability P is varied from 0.05 to 0.7 ($P=0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7$) to make a complete comparison. The simulation is repeated 100 times for each parameter setting, and the averaged FDR and sensitivity are reported.

It considers two versions of the proposed method. The first one does not consider hot points by setting $W=1$ rather than estimating it. The second one considers hot points by allowing W to be estimated from data. The method in [75] essentially scans the whole sequence and selects a position exhibiting the most dramatic difference as a potential change-point to be determined by a statistical test. Following their strategy, a frequentist testing procedure is implemented as a competing method to be compared with the proposed empirical Bayes method. Specifically, this frequentist method scans the whole sequence and finds a position with the most significant difference as quantified by rank-sum testing statistic. It is an extreme testing statistic and its original p-value may not be valid any more. As a result, this frequentist method couldn't guarantee multiplicity control. To make comparison, the exons are ranked based on their maximum rank-sum testing statistic and then the same number of significant exons are reported.

5.3.2 Simulation Results

Figure 5.3 shows the results for the scenario without hot points. First, the model without considering hot points (NB model without HP), which is ideal for this scenario, can control FDR precisely at the nominal level 0.1 for all settings. Second, the model considering hot points (NB model with HP) can also control the FDR at nominal level 0.1 for all settings. In addition, it demonstrates similar sensitivity as

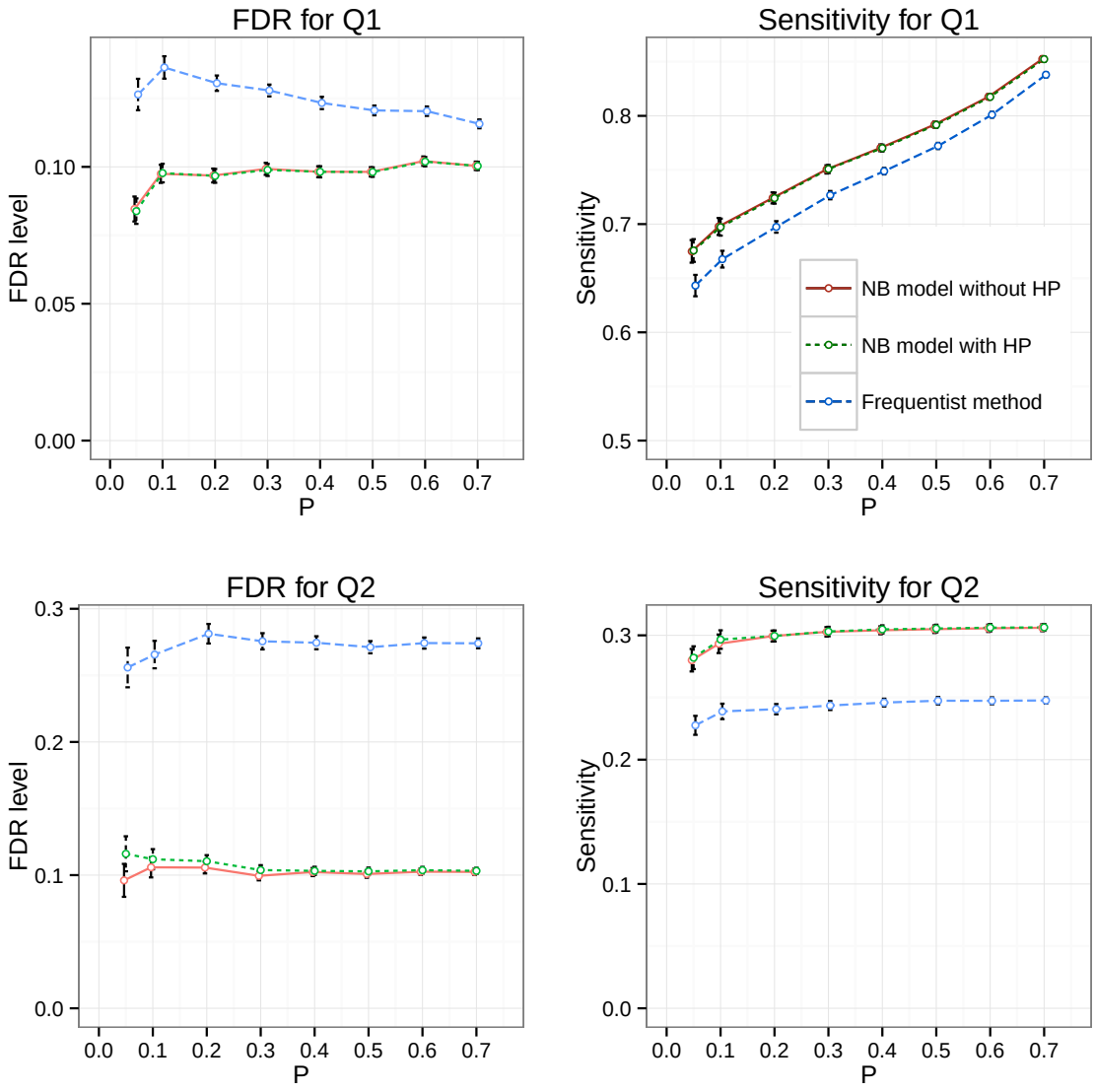

Figure 5.3 Results for different methods applied on data set without hot points. "NB Model with HP" represents change-point model considering hot points; "NB Model without HP" represents change-point model without considering hot points.

the ideal model which has the information of W=1. This comparable performance suggests that the model considering hot points, which estimates W from data, is more general and robust. Third, both of the empirical Bayes models outperform the frequentist method which demonstrates a higher FDR while with a lower sensitivity. Fourth, the sensitivity for the detection problem (Q1) is much higher than the identification problem (Q2), which is expected, as the latter indeed is more challenging. Figure 5.4 shows the results for the scenario with hot points simulated. Again, the empirical Bayes models, considering hot points or not, both demonstrate significantly better performance than the frequentist method. The model considering hot points is the optimal model. It can precisely control FDR at the nominal level and shows the best performance. It is noted that the model without considering hot points erroneously set W=1, and, as a result, it either couldn't guarantee FDR control for the detection problem (Q1), or has a lower sensitivity than the correct model for the identification problem (Q2).

From the results previously mentioned, in comparison with the model considering hot points *versus* the one without considering hot points, the former is comparable when applied to data without hot points, and better than the latter when applied to data with hot points. Therefore, the model considering hot points is robust and superior.

5.3.3 Real Data Experiments

The proposed method is applied to analyze a real dataset in this section. Flockhart et al. [24] conducted whole transcriptome RNA-seq to study melanoma cell migration. Their RNA-seq datasets have been deposited to the NCBI Gene Expression Omnibus (GEO) database (`http://www.ncbi.nlm.nih.gov/geo/`) with ID GSE33092. There are 69,925,376 paired-end reads in the control sample SRR354040 from primary human melanocytes infected with RFP lentivirus, and 62,884,955 paired-end reads in

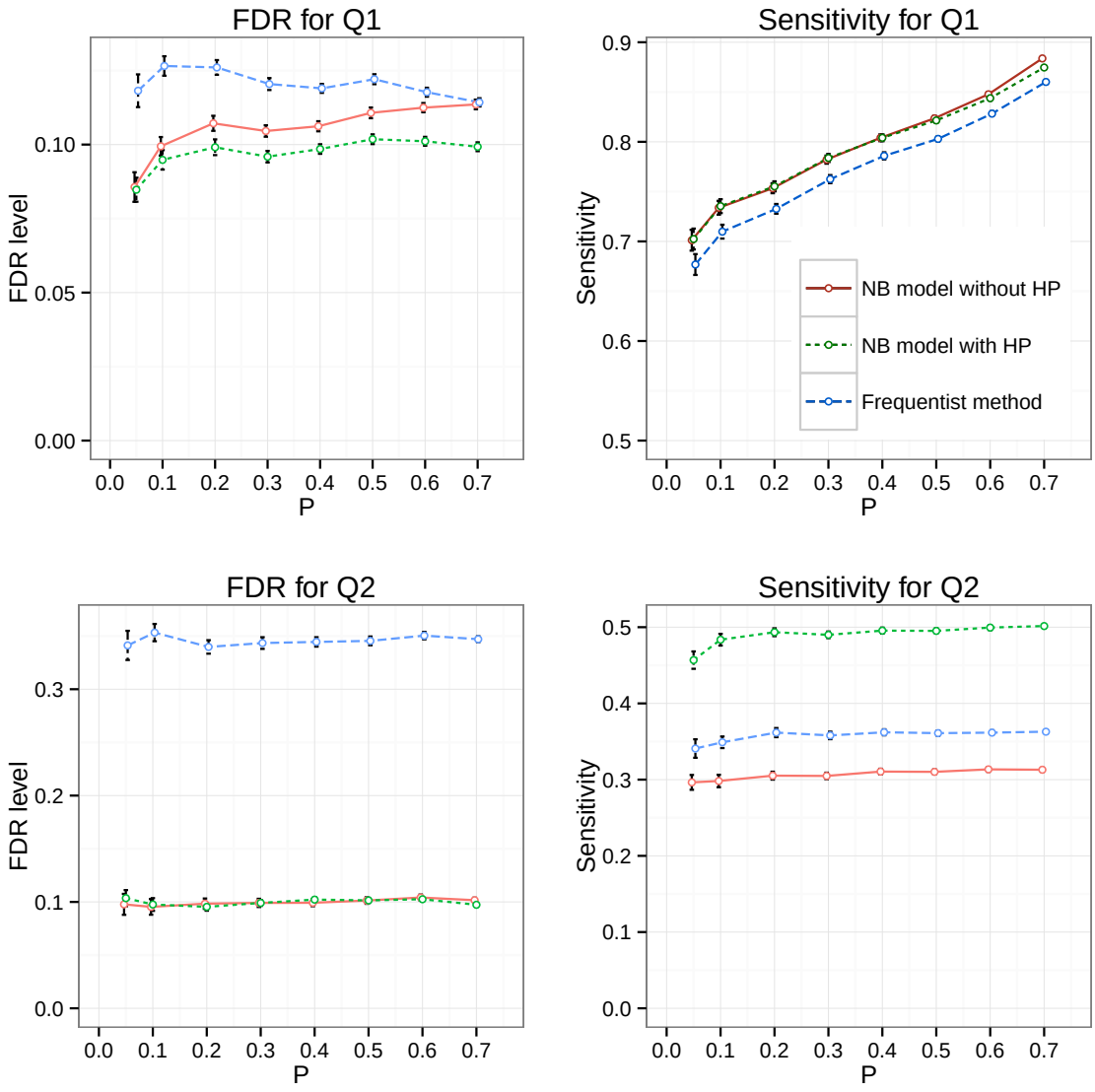

Figure 5.4 Results for different methods applied on data set with hot points. "NB Model with HP" represents change-point model considering hot points;"NB Model without HP" represents change-point model without considering hot points.

Table 5.1 Estimated Parameters for Different Samples

Data Set	P	r	α	β	W
SRR354040	0.354	2.21	1.02	1.22	9.90
SRR354042	0.348	0.86	1.20	1.57	10.00

the case sample SRR354042 from primary human melanoma sample [24]. The raw reads were downloaded from GEO and then aligned to the hg19 reference genome using the popular RNA-seq mapping tool Tophat [69] v1.3.1 with default settings. Exons with short reads found in both samples are used for further analysis. As a result, 62209 exons from 13290 distinct genes remain.

The read-count data were calculated for each position, and then binned every 5 BPs as one point to reduce the effect of sparsity and noise. In addition, junction read-supported positions are treated as hot points, whose weights will be estimated from the data. Table 5.1 shows the estimated parameters for the two samples. Note that the estimated weights of hot points are bigger than 1 ($\hat{W} > 1$), which indicates that the positions supported by junction reads do have higher prior probabilities than other locations. The proposed model can capture and make use of this information effectively.

To find biologically meaningful $3'/5'$ AS events, it tries to detect the exons that have change-points in one sample but not in the other sample. Under the FDR level $\alpha = 0.05$, the proposed method detects 7222 such exons. As a comparison, the tool developed by Wang et al. [75] and the simple strategy, which counts only junction reads, are also applied to analyze this dataset. Wei's tool only detects 3366 exons with significant changes between the two samples at the same FDR level, which suggests a lower power compared to the proposed method. For the simple strategy, it reports a $3'/5'$ AS event if it is supported by one or more junction reads. This strategy detects 796 exons that contain $3'/5'$ AS events in one sample but not in the other sample.

Table 5.2 Results for Real Data Experiments

Method	# Total Detected AS	# Novel AS	# AS Supported by AceView	Supporting Rate
EB Change-point	7222	1988	5234	72.5%
Frequentist Method (Wei's)	3366	1058	2308	68.6%
Simple Strategy	796	24	772	97.0%

This improved sensitivity of the proposed method over the simple strategy shows that utilizing junction reads together with read coverage information could obtain a better performance than using junction reads only.

The detected AS events are further compared with the isoforms cataloged in the AceView database. When the algorithm detects an AS event and there is one in the matching AceView exon as well, it means this detected AS event is supported by AceView. It is noted that this is not an experimental validation but serves as a proxy to the "truth". As summarized in Table 5.2, 5234 out of the 7222 AS events the proposed method finds are supported by AceView and 1988 AS events are novel, while for Wei's method, 2308 out of the 3366 AS events are supported and 1058 AS events are novel, and for the simple strategy, 772 AS events are supported and 24 AS events are novel. In contrast, for the whole genome, 64.7% exons contain AS events annotated in the AceView database. In this section, the proposed model only uses the junction reads as side information and leaves the AceView annotations for evaluation purpose. The exons reported by the proposed method have a statistically higher supporting rate of 72.5%.

Finally, following [75], the gene set enrichment analysis (GSEA) is conducted based on the genes with $3'/5'$ AS events reported by the proposed model in order to evaluate the results from a systems biology point of view. The canonical pathways definitions (Version 4) are downloaded from the Molecular Signatures Database (http://www.broadinstitute.org/gsea/msigdb/index.jsp). The proposed method identifies 12 significantly enriched pathways as shown in Table 5.3 at an FDR level of 0.05. Interestingly, many of them are relevant to melanoma or cancer.

For example, recent studies demonstrate that activation of the Smad2/3 pathway has inhibitory effects on tumor cell plasticity of melanoma [54]. Regulation of the actin cytoskeleton contributes to cancer cell migration and invasion [87]. The Notch signaling pathway plays a key role in melanoma growth and progression [6]. The meaningful GSEA results from a systems biology point of view provide supplementary support to the proposed method. In addition, they may provide insight into the role of $3'/5'$ AS in these pathways.

5.4 Conclusion

This chapter proposes an empirical Bayes change-point model to identify $3'/5'$ AS events. Simulation studies and real data application have demonstrated that the proposed method is powerful, accurate and efficient for analyzing the next-generation RNA sequencing data. Compared with previous methods, the new approach does not rely on annotation information. Instead, it provides for the first time a systematic framework to characterize coverage change while being capable of integrating other information, in particular the junction read information which is very helpful for detecting $3'/5'$ AS events.

It utilizes an empirical Bayes model to efficiently pool information across genes. The Negative Binomial-Beta model, which allows the over-dispersion in the real data, could estimate the hyperparameters from data efficiently. This makes the model more powerful compared with frequentist methods, as it applies Bayesian inference [91]. Since the hyperparameters are estimated frequentistically from data, it also overcomes the defects of subjective priors of Bayesian methods. In addition, it provides a flexible testing framework in which the user can choose to address different levels of questions, namely, whether alternative splicing happens, and/or where it happens. This gives users more flexibility in solving real problems. When exact splice sites are hard to determine, user could choose to only report the exons that contain alternative splicing.

In addition, a Bayesian confidence interval for the splicing point can be constructed based on the posterior probabilities if it is of the user's particular interest.

Table 5.3 Gene Set Enrichment Analysis Results

Canonical Pathway	P-Value
PID_SMAD2_3NUCLEAR_PATHWAY	3.82E-05
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	7.02E-05
PID_MET_PATHWAY	1.54E-04
KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM	1.68E-04
SIG_BCR_SIGNALING_PATHWAY	2.19E-04
REACTOME_DEVELOPMENTAL_BIOLOGY	2.24E-04
BIOCARTA_VDR_PATHWAY	2.30E-04
PID_NECTIN_PATHWAY	2.47E-04
KEGG_PATHWAYS_IN_CANCER	2.50E-04
KEGG_FOCAL_ADHESION	2.94E-04
KEGG_NOTCH_SIGNALING_PATHWAY	4.06E-04
PID_HES_HEY_PATHWAY	4.06E-04

CHAPTER 6

CONCLUSIONS AND FUTURE WORKS

This dissertation focuses on the development of statistical learning methods for mining marketing and biological data. The main contributions of this dissertation are as listed below:

First, a collaborated online change-point detection method is developed for identifying the change-points in sparse time series. By leveraging the auxiliary time series, it can quickly and accurately identify the changes in the revenue data and enable the predictive model to use historical data intelligently. With the improved accuracy, advertisers could further optimize their biding strategies and increase the revenue.

Second, a novel task-specific multi-task learning algorithm is proposed to help media providers predict users' gender information from their video viewing behaviors. Compared with the traditional multi-task learning algorithms, it combines the ℓ_1 regularized task-specific features and ℓ_1/ℓ_2 regularized shared features to model the ever-changing user's watching behaviors. It brings considerable flexibility for practitioners to incorporate domain knowledge into their models. In addition, Bayes testing and decision procedures are proposed to report as many desired users as possible, while controlling the false discovery rate (FDR) or Type I error rate at a user-specified level.

Finally, an empirical Bayes change-point model is proposed to identify $3'$ and $5'$ alternative splicing from RNA-seq data. It provides for the first time a systematic framework to integrate various information when available, in particular the useful junction read information, in order to obtain better performance in change-point detection. An empirical Bayes method is utilized to efficiently pool information across

genes to improve detection efficiency. It also provides a flexible testing framework in which the user can choose to address different levels of questions, namely, whether alternative $3'$ SS or $5'$ SS happens, and/or where it happens.

Future work lies in the following directions:

First, except the traditional changes in mean, variance, sequence length and so on, the periodic change is a special and usually pretty useful pattern in industry and academic data. Modeling it appropriately could further improve the performance of existing systems.

Second, the available data may demonstrate various biases. For example, recent studies have revealed that RNA-seq data sampled from the transcriptome exhibit various biases, including position-specific and sequence-specific biases [32]. These biases may incur great difficulties in detecting change-points and cause false positive reports. Additional efforts are required to circumvent this problem by improving the data collection procedure and building more robust models.

Third, the proposed methods mainly focus on one change-point at a time. It is noted that there can be more than one change-points in a sequence. In principle, the proposed procedures could be extended to search for more change-points. However, seeking the optimal model for multiple change-points would impose great computational cost. Computational time is linear to scan for one possible change-point, and becomes factorial when considering multiple change-points. There is also a caveat of overfitting to consider. Because of these implications, the potential gain may not necessarily warrant seeking a perfect model. The extension for multiple change-points is left for future work.

BIBLIOGRAPHY

[1] Online advertising. https://en.wikipedia.org/wiki/Online_advertising. Accessed: 2015-09-01.

[2] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from rna-seq data. *Genome Res.*, 22(10):2008–17, Oct 2012.

[3] Sebastian Angel and Michael Walfish. Verifiable auctions for online ad exchanges. In *ACM SIGCOMM Computer Communication Review*, volume 43, pages 195–206. ACM, 2013.

[4] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[5] Daniel Barry and John A Hartigan. A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

[6] Barbara Bedogni. Notch signaling in melanoma: Interacting pathways and stromal influences that enhance notch targeting. *Pigment Cell & Melanoma Research*, 27(2):162–168, 2014.

[7] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 289–300, 1995.

[8] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 131–140, 2013.

[9] Christopher M Bishop. *Neural networks for pattern recognition*. Oxford University Press, Oxford, UK, 1995.

[10] Benjamin J Blencowe, Sidrah Ahmad, and Leo J Lee. Current-generation high-throughput sequencing: Deepening insights into mammalian transcriptomes. *Genes & Development*, 23(12):1379–1386, 2009.

[11] Regina Bohnert and Gunnar Rätsch. A tool for rna-seq-based transcript quantitation. *Nucleic Acids Research*, 38(suppl 2):W348–W351, 2010.

[12] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[13] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[14] Jie Chen and Arjun K Gupta. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92(438):739–747, 1997.

[15] Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. Predicting the demographics of twitter users from website traffic data. In *Proceedings of the International Conference on Web and Social Media*, 2015.

[16] Lawrence A David, Corinne F Maurice, Rachel N Carmody, David B Gootenberg, Julie E Button, Benjamin E Wolfe, Alisha V Ling, A Sloan Devlin, Yug Varma, Michael A Fischbach, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, 2014.

[17] David GT Denison. *Bayesian methods for nonlinear classification and regression*, volume 386. John Wiley & Sons, 2002.

[18] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 15–24. ACM, 2014.

[19] Bradley Efron. Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, 100(469):1–5, 2005.

[20] Bradley Efron. *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, Cambridge, UK, 2010.

[21] Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Gunnar Rätsch, Nick Goldman, Tim J Hubbard, Jennifer Harrow, Roderic Guigó, et al. Systematic evaluation of spliced alignment programs for rna-seq data. *Nature Methods*, 10(12):1185–1191, 2013.

[22] Tingting Feng, Yuchun Guo, Yishuai Chen, Xiaoying Tan, Ting Xu, Baijun Shen, and Wei Zhu. Tags and titles of videos you watched tell your gender. In *IEEE International Conference on Communications*, pages 1837–1842. IEEE, 2014.

[23] Ronald A Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, pages 87–94, 1922.

[24] Ross J Flockhart, Dan E Webster, Kun Qu, Nicholas Mascarenhas, Joanna Kovalski, Markus Kretz, and Paul A Khavari. Brafv600e remodels the melanocyte transcriptome and induces bancr to regulate melanoma cell migration. *Genome Res.*, 22(6):1006–14, Jun 2012.

[25] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *Proceedings of the Conference on Internet Measurement Conference*, pages 141–148. ACM, 2013.

[26] Malachi Griffith, Obi L Griffith, Jill Mwenifumbo, Rodrigo Goya, A Sorana Morrissy, Ryan D Morin, Richard Corbett, Michelle J Tang, Ying-Chen Hou, Trevor J Pugh, et al. Alternative expression analysis by rna sequencing. *Nature Methods*, 7(10):843–847, 2010.

[27] Suzan M Hammond and Matthew JA Wood. Genetic therapies for rna mis-splicing diseases. *Trends in Genetics*, 27(5):196–205, 2011.

[28] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 2. Springer, 2009.

[29] Douglas M Hawkins, Qiu Peihua, and Wook Kang Chang. The changepoint model for statistical process control. *Journal of Quality Technology*, 35(4):355–366, 2003.

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

[31] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International Conference on World Wide Web*, pages 151–160. ACM, 2007.

[32] Yin Hu, Yan Huang, Ying Du, Christian F Orellana, Darshan Singh, Amy R Johnson, Anaïs Monroy, Pei-Fen Kuan, Scott M Hammond, Liza Makowski, et al. Diffsplice: The genome-wide detection of differential splicing events with rna-seq. *Nucleic Acids Research*, 41(2):e39–e39, 2013.

[33] Yan Huang, Yin Hu, Corbin D Jones, James N MacLeod, Derek Y Chiang, Yufeng Liu, Jan F Prins, and Jinze Liu. A robust method for transcript quantification with rna-seq data. *Journal of Computational Biology*, 20(3):167–187, 2013.

[34] Bernard J Jansen, Kathleen Moore, and Stephen Carman. Evaluating the performance of demographic targeting using gender in sponsored search. *Information Processing & Management*, 49(1):286–302, 2013.

[35] Tang Jen and Arjun K Gupta. On testing homogeneity of variances for gaussian models. *Journal of Statistical Computation and Simulation*, 27(2):155–173, 1987.

[36] Auinash Kalsotra and Thomas A Cooper. Functional consequences of developmentally regulated alternative splicing. *Nature Reviews Genetics*, 12(10):715–729, 2011.

[37] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat. Methods*, 7(12):1009–15, Dec 2010.

[38] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: Diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355, 2010.

[39] Rebecca Killick and Idris Eckley. Changepoint: An r package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.

[40] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

[41] Alexandros Labrinidis and Hosagrahar V Jagadish. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12):2032–2033, 2012.

[42] Bo Li and Colin N Dewey. Rsem: Accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[43] Bo Li, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.

[44] Steve Lohr. The age of big data. *New York Times*, 11, 2012.

[45] Steve Lohr. Data-ism: The revolution transforming decision-making, consumer behavior, and almost everything else. 2015.

[46] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[47] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[48] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat. Methods*, 5(7):621–8, Jul 2008.

[49] Partha Mukherjee and Bernard J Jansen. The gender-brand effect of key phrases on user clicks in sponsored search. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 1845–1850. ACM, 2013.

[50] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632. ACM, 2005.

[51] Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2006.

[52] Jahna Otterbacher. Inferring gender of movie reviewers: Exploiting writing style, content and metadata. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 369–378. ACM, 2010.

[53] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–5, Dec 2008.

[54] E Pardali, DWJ van der Schaft, E Wiercinska, A Gorter, PCW Hogendoorn, AW Griffioen, and P Ten Dijke. Critical role of endoglin in tumor cell plasticity of ewing sarcoma and melanoma. *Oncogene*, 30(3):334–345, 2011.

[55] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

[56] Aleksey S Polunchenko and Alexander G Tartakovsky. State-of-the-art in sequential change-point detection. *Methodology and Computing in Applied Probability*, 14(3):649–684, 2012.

[57] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[58] Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3):R22, 2011.

[59] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.

[60] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[61] Shihao Shen, Juw Won Park, Jian Huang, Kimberly A Dittmar, Zhi-xiang Lu, Qing Zhou, Russ P Carstens, and Yi Xing. Mats: A bayesian framework for flexible detection of differential alternative splicing from rna-seq data. *Nucleic Acids Res.*, 40(8):e61, Apr 2012.

[62] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[63] David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

[64] Darshan Singh, Christian F Orellana, Yin Hu, Corbin D Jones, Yufeng Liu, Derek Y Chiang, Jinze Liu, and Jan F Prins. Fdm: A graph-based statistical method to detect differential transcription using rna-seq data. *Bioinformatics*, 27(19):2633–2640, 2011.

[65] Ravi K Singh and Thomas A Cooper. Pre-mrna splicing in disease and therapeutics. *Trends in Molecular Medicine*, 18(8):472–482, 2012.

[66] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[67] Wenguang Sun and Zhi Wei. Hierarchical recognition of sparse patterns in large-scale simultaneous inference. *Biometrika*, pages 267–280, 2015.

[68] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[69] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: Discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[70] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

[71] Ernest Turro, Shu-Yi Su, Ângela Gonçalves, LJ Coin, Sylvia Richardson, Alex Lewin, et al. Haplotype and isoform specific expression estimation using multi-mapping rna-seq reads. *Genome Biol*, 12(2):R13, 2011.

[72] Ben Walker. Every day big data statistics: 2.5 quintillion bytes of data created daily. *VCloudNews. April*, 5, 2015.

[73] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6, Nov 2008.

[74] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. Your cart tells you: Inferring demographic attributes from purchase data. pages 173–182, 2016.

[75] Wei Wang, Zhi Wei, and Hongzhe Li. A change-point model for identifying 3'utr switching by next-generation rna sequencing. *Bioinformatics*, 30(15):2162–2170, Aug 2014.

[76] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, Jan 2009.

[77] Kelly West. Unsurprising: Netflix survey indicates people like to binge-watch tv. *Cinema Blend*, 2013.

[78] Lori D Wolin and Pradeep Korgaonkar. Web advertising: Gender differences in beliefs, attitudes and behavior. *Internet Research*, 13(5):375–385, 2003.

[79] KJ Worsley. The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika*, 70(2):455–464, 1983.

[80] Jie Wu, Martin Akerman, Shuying Sun, W Richard McCombie, Adrian R Krainer, and Michael Q Zhang. Splicetrap: A method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, 27(21):3010–3016, 2011.

[81] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004.

[82] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107, 2014.

[83] Zhengpeng Wu, Xi Wang, and Xuegong Zhang. Using non-uniform read distribution models to improve isoform expression inference in rna-seq. *Bioinformatics*, 27(4):502–508, 2011.

[84] J. L. Xu, W. Su, and M. Zhou. Likelihood-ratio approaches to automatic modulation classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(4):455–469, July 2011.

[85] Tianbing Xu. *Online advertising: A large scale computing perspective*. University of California, Irvine, USA, 2013.

[86] Xiang Xuan and Kevin Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th International Conference on Machine Learning*, pages 1055–1062. ACM, 2007.

[87] Hideki Yamaguchi and John Condeelis. Regulation of the actin cytoskeleton in cancer cell migration and invasion. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1773(5):642–652, 2007.

[88] Zhibo Yin, Lingxiao Zhang, Xinheng Fan, and Wei Li. What chinese female online shoppers need. In *Cross-Cultural Design*, pages 498–508. Springer, 2014.

[89] Josh Jia-Ching Ying, Yao-Jen Chang, Chi-Min Huang, and Vincent S Tseng. Demographic prediction based on users mobile behaviors. *Mobile Data Challenge*, 2012.

[90] Jie Zhang, Zhi Wei, Zhenyu Yan, and Abhishek Pani. Collaborated online change-point detection in sparse time series for online advertising. In *IEEE International Conference on Data Mining*, pages 1099–1104. IEEE, 2015.

[91] Zhigen Zhao, Wei Wang, and Zhi Wei. An empirical bayes testing procedure for detecting variants in analysis of next generation sequencing data. *The Annals of Applied Statistics*, 7(4):2229–2248, 2013.

[92] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of youtube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pages 404–410. ACM, 2010.

[93] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.