Summer 2023

Advances in Differentially Methylated Region Detection and Cure Survival Models

Daniel Ahmed Alhassan
Missouri University of Science and Technology

Department: Mathematics and Statistics

ADVANCES IN DIFFERENTIALLY METHYLATED REGION DETECTION AND CURE SURVIVAL MODELS

by

DANIEL AHMED ALHASSAN

A DISSERTATION

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

MATHEMATICS WITH STATISTICS EMPHASIS

2023

Approved by:

Akim Adekpedjou, Advisor
Gayla R. Olbricht, Co-Advisor
V. A. Samaranayake
Xuerong Wen
Matthew Thimgan

PUBLICATION DISSERTATION OPTION

This dissertation consists of the following two articles, formatted in the style used by the Missouri University of Science and Technology which are both intended for submission as follows:

Paper I: Pages 54-94 are intended for submission to refereed archival journal.

Paper II: Pages 95-139 are intended for submission to refereed archival journal.

ABSTRACT

This dissertation focuses on two areas of statistics: DNA methylation and survival analysis. The first part of the dissertation pertains to the detection of differentially methylated regions in the human genome. The varying distribution of gaps between succeeding genomic locations, which are represented on the microarray used to quantify methylation, makes it challenging to identify regions that have differential methylation. This emphasizes the need to properly account for the correlation in methylation shared by nearby locations within a specific genomic distance. In this work, a normalized kernel-weighted statistic is proposed to obtain an optimal amount of "information" from neighboring locations to detect those differences. The large sample properties of the proposed statistic are also studied. Simulation studies show that the proposed method captures the true length of differentially methylated regions more accurately than a widely used existing method.

The second focus area of the dissertation pertains to mixture cure models in survival analysis. Mixture cure models are those based on a cured and uncured population. The choice of models for the cured and uncured is crucial in developing statistical models for this type of population. In this work, a flexible mixture cure model is proposed that incorporates a generalized partially linear single-index model for modeling the cure part and an additive hazard model for the uncured. This model proves to be particularly effective when dealing with large data sets where an underlying baseline mechanism is expected. In such cases, limited models like the logistic regression are not suitable. By employing the additive hazard model, the proposed approach offers an alternative for modeling cure survival data, especially when hazard differences are of interest and the proportional hazard assumption is violated.

ACKNOWLEDGMENTS

I extend my deepest gratitude to the Infinite, Magnificent, and Almighty God for His unwavering support throughout the course of my PhD studies.

My sincere appreciation is directed towards my advisors, Dr. Adekpedjou and Dr. Olbricht. Their consistent encouragement, support, and patience during my research have been invaluable.

I am also grateful to my committee members for their time, presence and inputs in my dissertation. Special recognition is due to Dr. Samaranayake, who showed immediate interest in me and patiently answered all my questions, academic or otherwise. My gratitude also extends to Dr. Thimgan for his stimulating conversations and warm hospitality during my office visits.

My heartfelt thanks go to my family - my mother, Christiana, my father, Mr. Alhassan, my brother David, and Uncle Richard. Their unwavering support has been my pillar of strength. I am also indebted to Professor Gabriel Asare Okyere, whose teachings, guidance, and mentorship made my education in the US possible.

I wish to acknowledge my dear friend, colleague, and sister, Dr. Priscilla (Codjoe) Bonnah, for her wisdom, support, and assistance over the years. I am equally thankful to my good friend, Dr. Charles Amponsah, for his insightful questions that challenged my ideas and thought processes. I also extend my sincere appreciation to Dr. Nicholas and Samantha Dadzie for their mentorship throughout my studies.

Lastly, I want to express my gratitude to my church family, First Love Church. Their fellowship and fervent prayers have been a source of strength and inspiration throughout this journey. This dissertation would not have been possible without their support.

TABLE OF CONTENTS

Page

PUBLICATION DISSERTATION OPTION iii

ABSTRACT iv

ACKNOWLEDGMENTS v

LIST OF ILLUSTRATIONS x

LIST OF TABLES xii

SECTION

1. INTRODUCTION 1

2. BACKGROUND TO PAPER I 2

2.1. EPIGENETICS AND DNA METHYLATION 2

2.1.1. Background 2

2.1.2. DNA Methylation and Diseases 3

2.2. DNA METHYLATION TECHNOLOGIES 3

2.2.1. Bisulfite Sequencing 3

2.2.2. Bisulfite Microarray Technologies 4

2.3. DNA METHYLATION DATA 9

2.4. PREPROCESSING OF METHYLATION DATA 11

2.4.1. Quality Control 11

2.4.2. Normalization 12

2.5. DIFFERENTIAL METHYLATION TESTING METHODS 14

2.5.1. Site-level Testing 14

2.5.2. Region-level Testing 17

2.5.2.1. Bump Hunter 18

2.5.2.2. Probe Lasso 19

2.5.2.3. DMRcate 19
2.6. MOTIVATION 21
2.7. MATHEMATICAL BACKGROUND 22

3. BACKGROUND TO PAPER II 31
3.1. SURVIVAL ANALYSIS AND THE FEATURES OF SURVIVAL DATA 31
3.1.1. Definition of Survival Quantities 32
3.1.2. Survival Models 33
3.1.2.1. Cox Proportional Hazards model 34
3.1.2.2. Additive Hazard model 34
3.1.2.3. Accelerated Failure Time model 35
3.2. CURE MODELS 36
3.2.1. Background and History 36
3.2.2. Survival Analysis in the Presence of a Cure Fraction: Key Quantities and Concepts 38
3.2.3. Mixture Cure Models: A Brief Review of Modeling Approaches 40
3.3. IDENTIFIABILITY OF THE MIXTURE CURE MODEL 42
3.4. THE EM ALGORITHM 43
3.5. MOTIVATION 44
3.6. GENERALIZED PARTIALLY LINEAR SINGLE INDEX MODEL 45
3.7. ESTIMATION OF THE PROPOSED MIXTURE CURE MODEL 46

PAPER

I. DIFFERENTIAL METHYLATED REGION DETECTION VIA AN ARRAY-ADAPTIVE NORMALIZED KERNEL-WEIGHTED MODEL 54
ABSTRACT 54
1. INTRODUCTION 55

2. METHODS 61

2.1. SITE-LEVEL DIFFERENTIAL METHYLATION TESTING WITH LIMMA 61

2.2. A GENERAL LOCALLY-WEIGHTED STATISTIC 61

2.3. ASYMPTOTIC RESULTS 62

2.4. NORMALIZED KERNEL-WEIGHTED STATISTIC 63

2.5. MODELING $S(x_i)$ VIA SATTERTHWAITE'S APPROXIMATION 66

2.6. STEP-BY-STEP SUMMARY OF THE FADMR & AADMR DETECTION APPROACH 68

3. RESULTS 69

3.1. SIMULATION STUDY 69

3.2. EVALUATION CRITERIA 70

3.3. SIMULATION RESULTS 71

3.4. REAL DATA EXAMPLE 76

3.4.1. Data Extraction 76

3.4.2. Oral Squamous Cell Carcinoma (OSCC) 77

3.5. SUMMARY OF RESULTS 79

4. CONCLUSION 80

ACKNOWLEDGEMENTS 81

REFERENCES 82

II. A GENERALIZED PARTIALLY LINEAR SINGLE-INDEX ADDITIVE HAZARD MIXTURE CURE MODEL 95

ABSTRACT 95

1. INTRODUCTION 96

2. THE MODEL AND ESTIMATION 101

2.1. IDENTIFIABILITY OF MODEL 102

2.2. ESTIMATION AND THE EM ALGORITHM 102

2.3. A COMPUTATIONAL ALGORITHM 107
3. SIMULATION STUDIES 109
4. REAL DATA EXAMPLE 122
5. CONCLUSIONS 125
REFERENCES 127

SECTION

4. SUMMARY AND CONCLUSIONS 140

APPENDIX 142

REFERENCES 149

VITA 162

LIST OF ILLUSTRATIONS

Figure Page

2.1. The Infinium Assay Workflow - The Infinium assay workflow proceeds from input DNA to automated genotype report (or methylation status for methylation arrays) with a total assay turnaround time of three days. Source: (Illumina, 2017). 5

2.2. Two Infinium probe types: Infinium I (top) and Infinium II (bottom) Source: Illumina (2012). 7

2.3. Illumina Infinium HumanMethylation 450 BeadChip. Source: Illumina Inc. (2012). 8

3.1. A heterogeneous population with cured and uncured sub-populations. 38

3.2. A schematic that describes the group (censored or uncensored) in which the cured and uncured may arise. 39

3.3. Kaplan and Meier (1958) estimator of the survival function for the breast cancer dataset of Wang *et al.* (2005) (+ : censored observations). 43

PAPER I

1. Probe Spacing distribution on the 450K array truncated at 1000bp to ease visualization. 65

2. Probe Spacing distribution on the EPIC array truncated at 1000bp to ease visualization. 66

3. Distribution of $\beta-$values from one of 1000 datasets showing the parameter space is well explored. 73

4. Large treatment effect ($\Delta\beta = 0.2$): Precision, recall and F1 score metrics based on EO criteria. Boxplots of results across the 1000 simulated datasets. 73

5. Large treatment effect ($\Delta\beta = 0.2$): Precision, recall and F1 score metrics based on AO criteria. Boxplots of results across the 1000 simulated datasets. 74

6. Small treatment effect ($\Delta\beta = 0.09$): Precision, recall and F1 score metrics based on EO criteria. Boxplots of results across the 1000 simulated datasets. 74

7. Small treatment effect ($\Delta\beta = 0.09$): Precision, recall and F1 score metrics based on AO criteria. Boxplots of results across the 1000 simulated datasets. 75

8. Density of CpGs in Significant DMRs from one of 1000 datasets for the three methods (DMRcate, faDMR and aaDMR). 76

9. Venn diagram of significant DMRs identified by the three methods from OSCC data. 79

PAPER II

1. Boxplots of the Average Squared Error (ASE) for three models under the logistic scenario for $n = 250$. 113

2. Boxplots of the Average Squared Error (ASE) for three models under the logistic scenario for $n = 500$. 114

3. Boxplots of the Average Squared Error (ASE) for three models under the logistic scenario for $n = 1000$. 114

4. Boxplots of the Average Squared Error (ASE) for three models under the logistic scenario for $n = 2500$. 115

5. Boxplots of the Average Squared Error (ASE) for the three models under the sine bump scenario for $n = 250$. 115

6. Boxplots of the Average Squared Error (ASE) for the three models under the sine bump scenario for $n = 500$. 116

7. Boxplots of the Average Squared Error (ASE) for the three models under the sine bump scenario for $n = 1000$. 116

8. Boxplots of the Average Squared Error (ASE) for the three models under the sine bump scenario for $n = 2500$. 117

9. Kaplan Meier survival plot of the DRS study showing a plateau. 124

LIST OF TABLES

Table Page

PAPER I

1. Large treatment effect ($\Delta\beta = 0.2$): A confusion matrix comparing the results from three methods (DMRcate, faDMR and aaDMR) with true DMRs based on EO criteria averaged across 1000 datasets. Sig. means statistically significant DMRs; Not Sig. indicates the number of regions that are not statistically significant..... 71

2. Large treatment effect ($\Delta\beta = 0.2$): A confusion matrix comparing the results from three methods (DMRcate, faDMR and aaDMR) with true DMRs based on AO criteria averaged across 1000 datasets. Sig. means statistically significant DMRs; Not Sig. indicates the number of regions that are not statistically significant..... 72

3. Small treatment effect ($\Delta\beta = 0.09$): Confusion matrix comparing three methods (DMRcate, faDMR, and aaDMR) to true DMRs based on AO criteria, averaged across 1000 datasets. Sig. represents statistically significant DMRs; Not Sig. indicates non-statistically significant regions. 75

PAPER II

1. Parameters for the incidence and latency sub-model, the cure fraction and the censoring rates..... 111

2. Coefficient estimates and standard deviations (sd) of partial linear term from the GPLSI-AH compared to LC model under the three censoring schemes and four sample sizes where the true relationship is logistic. 119

3. Coefficient estimates and standard deviations (sd) of partial linear term from the GPLSI-AH compared to LC model under the three censoring schemes and four sample sizes for the Sine Bump model. 120

4. Bias and variance of $\hat{\beta}$ for the GPLSI-AH, SIC and LC cure models under the three censoring schemes and four sample sizes. 121

5. Description of variables in the Diabetic Retinopathy Study data used for our analysis..... 124

6. Parameter estimates and standard errors for GPLSI-AH, SIC and LC cure models..... 125

1. INTRODUCTION

This dissertation focuses on the development of novel statistical methods in two different areas: DNA methylation analysis and cure survival models. The work is organized into two papers, one for each topic. In this dissertation, a separate section is devoted to providing background information for each paper. The two papers are then provided, followed by a conclusion section.

2. BACKGROUND TO PAPER I

The first part of this dissertation involves the development of statistical methods for detecting regions of differential methylation in the genome. This section focuses on the biological and mathematical information needed to fully appreciate the ideas in Paper I.

2.1. EPIGENETICS AND DNA METHYLATION

This section explores the epigenetic modification of DNA and its impact on gene expression and cellular processes.

2.1.1. Background. Epigenetics is the study of how one's behaviors and the changes in environmental characteristics affect the way genes work. Epigenetic changes are reversible and do not affect the deoxyribonucleic acid (DNA) sequence (Fernandez *et al.*, 2021). However, they can affect the activity of genes (gene expression) by turning genes off and on (Cedar, 1988). There are different types of epigenetic modifications, such as DNA methylation and histone modifications, that can occur in the genome. The focus of this work is on DNA methylation.

DNA methylation occurs when a methyl (CH3) group is added to the fifth carbon of a cytosine on the DNA sequence. In mammals, DNA methylation is known to typically occur symmetrically in a sequence where a cytosine (C) nucleotide base is followed by a guanine (G) nucleotide base. These locations are commonly referred to as CpG loci or sites. The "p" stands for the phosphate bond in between the C and G base (Laurent *et al.*, 2010; Ramsahoye *et al.*, 2000). Most methylation in mammals occurs at CpG sites but it can occur in places other than the CpG sequence. This non-CpG methylation mostly occurs in embryonic stem cells (Ichiyanagi *et al.*, 2013).

2.1.2. DNA Methylation and Diseases. The role of DNA methylation in human diseases was first seen in the area of genomic imprinting. Genomic imprinting is the situation where only one copy of the genes (whether maternal or paternal) is expressed. Diseases such as the Beckwith-Wiedemann, Prader-Willi and the Angelman syndromes, have been associated with the loss of imprinting (Procter *et al.*, 2006; Rossignol *et al.*, 2006). Moreover, DNA methylation patterns can change over time and are influenced by genetic factors and environmental changes. DNA methylation plays a complex role in disease etiology. Numerous studies have highlighted its significance, particularly in relation to the progression of cancer, diabetes, and aging (Jin and Liu, 2018). DNA methylation has the ability to modify how genes are expressed differently between individuals with and without a disease. Detecting locations in the genome that differ in methylation patterns between disease and healthy groups is a growing area of research that can aid in understanding the mechanism of complex diseases such as cancer. Statistical methods are essential in identifying genomic locations with significant association between DNA methylation and disease status.

2.2. DNA METHYLATION TECHNOLOGIES

Several technologies have been developed for quantifying DNA methylation across the genome. Bisulfite sequencing and bisulfite microarrays are the two most widely used technologies. These technologies are described in the following two sections, with a more detailed discussion about the bisulfite microarrays since they are the focus of this work.

2.2.1. Bisulfite Sequencing. Bisulfite sequencing begins with the DNA being treated with sodium bisulfite, which converts unmethylated cytosines to uracil while leaving methylated cytosines unaffected. The bisulfite-treated DNA is subjected to PCR amplification, which is used to make several copies of the bisulfite-treated DNA fragments. The PCR results are then subjected to DNA sequencing

methods such as Sanger sequencing or high-throughput next-generation sequencing (NGS) platforms, which determines the nucleotide (A, T, C, G) sequence of the DNA fragments. Unmethylated cytosines are read as thymines during sequencing, whereas methylated cytosines are read as cytosines. The methylation status of individual cytosine sites is established and quantified by comparing the sequenced fragments to a reference genome (Frommer *et al.*, 1992; Susan *et al.*, 1994). The gold standard for quantifying DNA methylation is Whole Genome Bisulfite Sequencing (WGBS) because it achieves the most comprehensive coverage of a genome (Jeong *et al.*, 2017). Though it is cheaper now than initially, the WGBS is still considered expensive compared to microarray technologies (Crary-Dooley *et al.*, 2017).

2.2.2. Bisulfite Microarray Technologies. The most popular microarray technology for quantifying DNA methylation was developed by Illumina. The Illumina microarray or BeadArray technology provides an alternative user-friendly and cost-effective approach to large scale epidemiological studies involving the human genome. It uses silica beads, which are coated with multiple copies of oligonucleotide capture probes that consist of nucleotide sequences corresponding to a specific location in the genome. In this approach, sodium bisulfite is first applied to DNA fragments. As described previously, this sodium bisulfite treatment converts unmethylated cytosines to uracils and leaves the methylated cytosines unchanged (Du *et al.*, 2010). The uracils are later read as thymines through PCR amplification. This helps determine the methylation status at each CpG site. Next, is the hybridization step. Bisulfite-converted DNA fragments from the biological sample are passed over the BeadChip and each fragment binds to a complementary sequence in the probes represented on the array, stopping one base before the CpG locus of interest. This positioning ensures that subsequent processes, such as single-nucleotide extension or detection, query the CpG site accurately to capture the methylation status at that specific locus. The design of the probes and their complementary binding to

the bisulfite-converted DNA fragments enable the BeadChip platform to detect and quantify DNA methylation levels at single-nucleotide resolution. A typical workflow of the Infinium assay technology is shown in Figure 2.1.

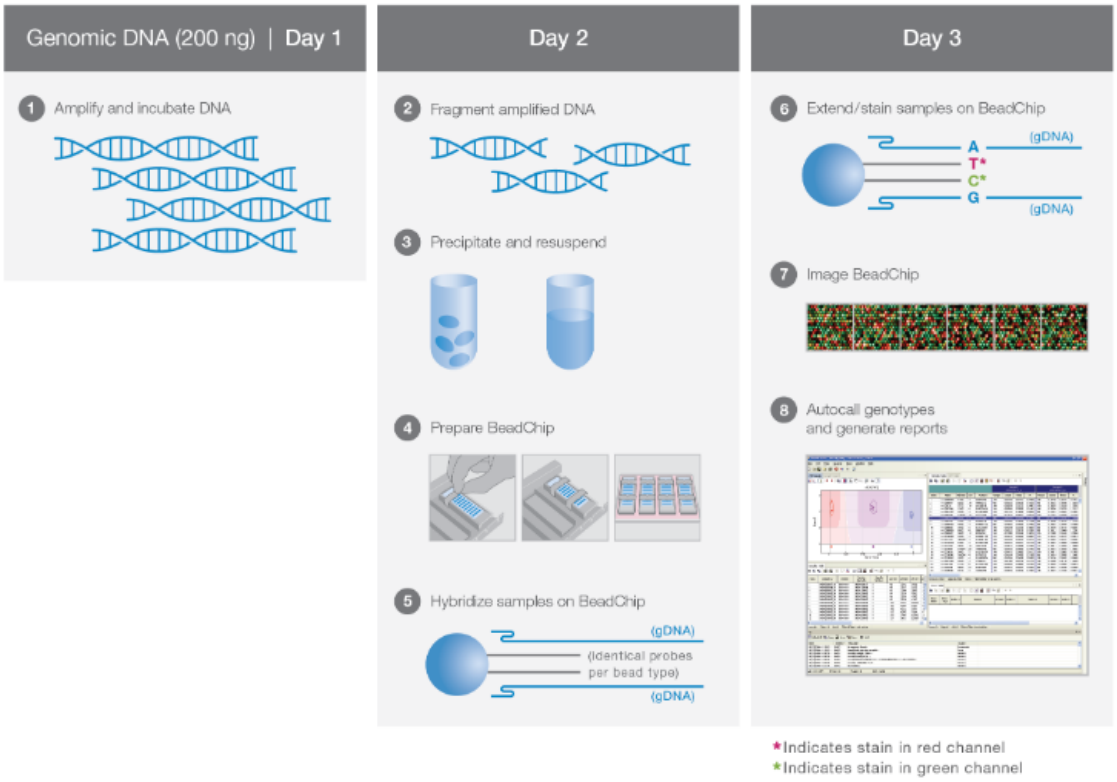

Figure 2.1. The Infinium Assay Workflow - The Infinium assay workflow proceeds from input DNA to automated genotype report (or methylation status for methylation arrays) with a total assay turnaround time of three days. Source: (Illumina, 2017).

The first BeadChip technology used to study methylation was the Illumina Infinium HumanMethylation27 (27K) array introduced in 2008. Measurements were taken at approximately 27,000 CpG sites, which cover approximately 14,000 genes. The 27K (which housed 12 samples per array) was one of the first Infinium BeadChips used to perform epigenome-wide association studies (EWAS). EWAS research aids in understanding the mechanisms behind various diseases and identifying biomarkers for cancers (Pidsley *et al.*, 2016). In 2011, Illumina introduced the Infinium HumanMethylation450 BeadChip (450K) that interrogated 485,577 CpG sites, and

kept 94% of the CpGs on the 27K array. The coverage on the 450K included a set of diverse genomic regions, such as the CpG Islands (CGI) (regions with a high frequency of CpG sites), shores (regions that are 2kb upstream and downstream of CpG islands), shelves (regions that are further away from the CpG islands, specifically 2kb away from the shores), 5'UTR (upstream Untranslated Region), 3'UTR (downstream Untranslated Region), FANTOM4 promoters (4th phase of Functional Annotation of the Mammalian Genome project that aims to identify all functional elements in mammalian genomes) and some enhancer regions (Bibikova *et al.*, 2011; Pidsley *et al.*, 2016). Organizations like The Cancer Genome Atlas (`https://www.cancer.gov/ccg/research/genome-sequencing/tcga`) and the Gene Expression Omnibus (`https://www.ncbi.nlm.nih.gov/geo/`) of the National Center for Biotechnology Information (NCBI) used the 450K platform to profile samples for different types of cancer.

The 450K uses two different probe types that could impact the analysis. The main types of probes used in the Infinium assay are the Infinium I and II probe designs. An illustration of these probe types can be found in Figure 2.2. For each CpG position, two readings are obtained: a methylated intensity and an unmethylated intensity. The way these intensities are measured depends on the specific type of probe used at the position. The Infinium I probe design, which measures approximately 30% of the CpG sites on the 450K, uses two separate probes for each CpG site, one for the methylated signal and one for the unmethylated signal. Both signals are measured using the same color, resulting in two one-color (red-red or green-green) assays. On the other hand, the Infinium II probe design, which measures the remaining $\sim 70\%$ of the CpG sites, uses a single probe to measure both the methylated and unmethylated signals, but each signal is reported in a different color, resulting in a two-color (red

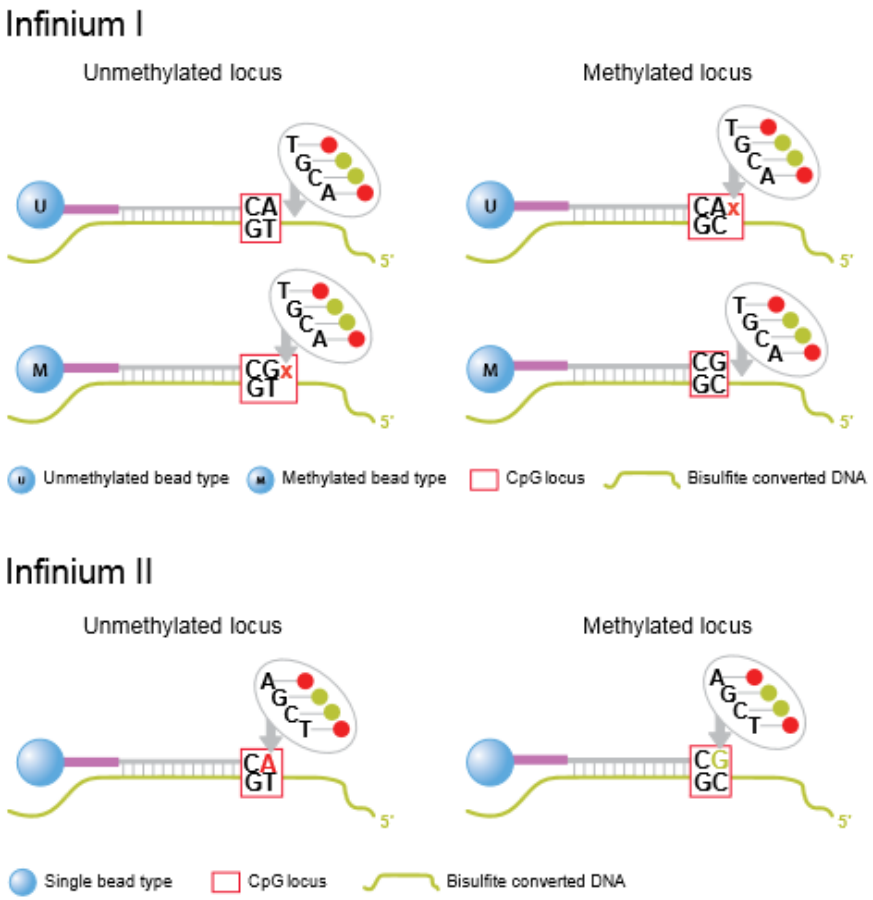

Figure 2.2. Two Infinium probe types: Infinium I (top) and Infinium II (bottom) Source: Illumina (2012).

and green) assay (Bibikova *et al.*, 2011). This unique combination of probe designs allows for a comprehensive and accurate assessment of methylation status across the genome.

Physically, the 450K is a small, rectangular piece of glass about the size of a standard microscope slide. Its surface is thoroughly covered with thousands of tiny spots or wells, each containing many tiny beads attached to specific DNA probes (Sandoval *et al.*, 2011). While these individual beads or probes are not visible to the human eye, under a microscope, a grid-like pattern of spots can be observed. The BeadChip is divided into sections, each capable of analyzing a different DNA sample.

Depending on the specific version, it might be divided into 8 or 12 sections, each physically separated from the others by a small gap. A barcode or other identifier is usually present on the edge of the BeadChip for tracking purposes (Sandoval *et al.*, 2011). See Figure 2.3 for the structure of the 450K BeadChip.

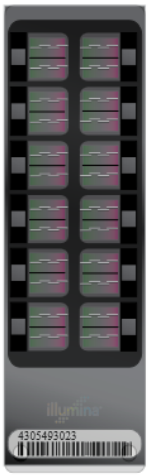

Figure 2.3. Illumina Infinium HumanMethylation 450 BeadChip. Source: Illumina Inc. (2012).

The 450K array missed essential regulatory regions on the genome and in 2016, Illumina introduced the EPIC array that targeted over 850,000 CpG sites; covering more than 90% of the sites on the 450K. It also contained more than 350,000 CpG loci identified as potential regulatory enhancers by two significant projects, FANTOM5 and ENCODE (Moran *et al.*, 2016; Pidsley *et al.*, 2016). FANTOM5 is an international research consortium that aims to identify all functional elements in mammalian genomes (The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014), while ENCODE is a public research consortium focused on identifying all functional elements in the human genome (ENCODE Project Consortium, 2012). The inclusion of these additional CpG loci in the EPIC array has expanded the ability to study the role of DNA methylation in gene regulation, contributing to a deeper understanding of the association between DNA methylation and diseases (McCartney

et al., 2016; Pidsley *et al.*, 2016). It is also notable that data collected at each CpG locus on the EPIC array maintains a high correlation with the 450K array, ensuring consistency across studies using the two different arrays (Pidsley *et al.*, 2016).

2.3. DNA METHYLATION DATA

There are two main file formats for Illumina methylation BeadChip 450K and EPIC array data: (i) raw (*.idat) data stores the raw intensities for each probe and (ii) *.txt data is obtained after preprocessing of the intensities has been completed. In line with The Cancer Genomic Atlas protocol, idat (intensity data) files are referred to as Level 1 data and *.txt files, level 3 data. There are several R packages built to handle these two main types of data. The raw idat files must undergo preprocessing to correct for technical variation and potential biases, such as background noise, ensuring that subsequent analyses are based on accurate and reliable methylation measurements (see section 2.4 for details). Two popular R packages *illuminaio* and *wateRmelon* are designed to read, process, and analyze the raw and preprocessed data from the Illumina methylation BeadChip (Pidsley *et al.*, 2013; Smith *et al.*, 2013).

Irrespective of the data format, there are two general resulting methylation data formats for analysis purposes. The first of the two is the β value. This represents the ratio of methylated intensities to the total methylated and unmethylated intensities. They range from 0 (fully unmethylated) to 1 (fully methylated) and have a highly skewed distribution (Du *et al.*, 2010). Mathematically, the β value for the ith CpG site is expressed as:

$$\beta_i = \frac{\max(\gamma_{i,\text{methy}}, 0)}{\max(\gamma_{i,\text{unmethy}}, 0) + \max(\gamma_{i,\text{methy}}, 0) + \alpha}, \tag{2.1}$$

where $\gamma_{i,\text{methy}}$ and $\gamma_{i,\text{unmethy}}$ represent the intensities measured by the i^{th} methylated and unmethylated probes respectively. α is an offset constant added to adjust the β values in the case where both methylated and unmethylated intensities are near 0. Illumina recommends an α value of 100. The $\max(\gamma_{i,\text{methy}}, 0)$ and $\max(\gamma_{i,\text{unmethy}}, 0)$ are used to avoid any possible negative values that may arise. The β values typically follow a beta distribution if one assumes the intensities are gamma distributed.

The second data type is the log ratio of the methylated intensities to the unmethylated intensities, referred to as the M-value. Mathematically it is expressed as follows:

$$M_i = \log_2 \frac{\max(\gamma_{i,\text{methy}}, 0) + \alpha}{\max(\gamma_{i,\text{unmethy}}, 0) + \alpha}. \tag{2.2}$$

Here, the offset parameter, α is set to a default value of 1. From (2.2), it is clear that an M-value close to 0 indicates similar intensity levels between the two probes. Positive M-values indicate more methylated than unmethylated intensities and vice versa for negative M-values. The relationship between the β and M values is shown below, where it is clear that the M-values have a support in $\mathcal{R}$:

$$M_i = \log_2 \left(\frac{\beta_i}{1 - \beta_i} \right). \tag{2.3}$$

Both data types are relevant. The β values are preferred when it comes to biological interpretation because they provide an intuitive measure of the methylation level at each CpG site and as such they are useful for visualizing and communicating methylation data (Du *et al.*, 2010). However, the M-values have better statistical properties, such as homoscedasticity of variance, an assumption required of most commonly used statistical methods and the same support as Gaussian distribution. For these reasons, M-values are preferred for statistical data analysis (Du *et al.*, 2010).

2.4. PREPROCESSING OF METHYLATION DATA

The typical preprocessing of DNA methylation data involves quality control via probe filtering and normalization (Wang *et al.*, 2018).

2.4.1. Quality Control. Prior to analysis, probe quality control must be performed to examine the success of the bisulfite conversion and array hybridization. Chen *et al.* (2012) found probes on the Illumina BeadChips that target CpG sites that overlap single nucleotide polymorphisms (SNPs), known as polymorphic CpGs. This can potentially introduce bias in the measurement of methylation levels, as the presence of SNPs can affect the binding efficiency of the probe and thus influence the success of bisulfite conversion. They also concluded that approximately 6% of probes generate fake signals because they target repetitive sequences or hybridize to multiple genomic regions. In other situations, a CpG site may be mutated and hence does not match its intended complimentary probe, so it should be excluded before downstream analysis. Thus, probes that meet any of these criteria are filtered out since they are of questionable quality. In many DNA methylation studies, probes on the X and Y chromosomes are also often filtered out to avoid potential sex biases in the downstream analysis (McGregor *et al.*, 2016).

Detection p-values are also used for quality control as they are a common technique used in the analysis of methylation data to distinguish true signal from background noise. They are calculated for each probe on the array and provide an estimate of the probability that the observed signal intensity for a given probe is distinguishable from the background. For Illumina's methylation arrays, the detection p-value is calculated based on the intensity signals of negative control probes, which are designed to not hybridize to any genomic DNA sequence (Aryee *et al.*, 2014). The distribution of these negative control intensities is used to estimate the background noise, and a z-test is performed to calculate a p-value for each probe on the array. A low detection p-value indicates that the signal intensity for a given probe

is significantly higher than the background noise, suggesting that the probe is reliably "detected" (Aryee *et al.*, 2014). The conventional cutoff for detection p-values is typically set at 0.01 or 0.05, although this can vary depending on the specific study or analysis. Probes with detection p-values above this cutoff are considered to be "undetected" and are often excluded from further analysis due to the low confidence in their signal intensities. An alternative approach to calculating detection p-values can be found in the work of Heiss and Just (2019).

2.4.2. Normalization. Normalization is the process of adjusting the raw intensity values of the methylation probes to correct for technical variation and potential biases. When done, it ensures that the differences observed in methylation levels across samples are reflective of true biological differences rather than technical artifacts or biases. To allow more probes to fit on the 450K and EPIC arrays, the Infinium II probe/assay was introduced in addition to the Infinium I that was used on the 27K. This created a two probe/assay design for both 450K and EPIC arrays. Though these assays are complimentary to each other, they possess very different chemistries and different dynamic ranges. Moreover, the 450K array uses 70% Infinium II probes/assays, and this leads to a potential type II probe bias during the analysis stage. After poor quality probes are filtered out, within-array normalization is the next step in the preprocessing phase of the DNA methylation analysis. This typically includes correction of type II probe bias, background correction and color-bias adjustment. For a detailed list of background correction methods developed, see Triche *et al.* (2013). The *lumi* package in R provides methods for background correction and the *methylumi* R package provides methods for color-bias adjustment. According to Dedeurwaerder *et al.* (2011), applying these methods improves the quality of data. The probe type-bias is considered the most critical of the preprocessing

techniques because it has the greatest potential of reducing the quality of data passed on for further analysis. Due to this, many efforts were made to develop methods to correct this type of bias. Some of the popular methods are described below.

The first method is the peak-based correction (PBC) method (Dedeurwaerder *et al.*, 2011). The Infinium II probes are less accurate and reproducible compared to the Infinium I probes. The PBC method re-scales methylation intensities of the type II probes to match the same modes as the type I probes. This method of bias correction is known to be less robust when the density distribution of β values do not show well defined peaks (Maksimovic *et al.*, 2012). In 2012, Maksimovic *et al.* (2012) introduced the Subset-quantile Within Array Normalization (SWAN), a two-step procedure that assumes the same intensity distribution when probes have the same number of CpGs. Another implicit assumption is that the differences in the methylation intensity distributions between probe types I and II, represent technical differences between them. The first step in the SWAN method is to identify an average quantile distribution using a subset of probes marked to be similar in terms of CpG content. Following this step, the intensities of the remaining probes must be adjusted via interpolation. Other within-sample normalization methods that were not used in this dissertation include the subset quantile normalization (SQN) and Beta Mixture Quantile normalization (BMIQ), proposed by Touleimat and Tost (2012) and Teschendorff *et al.* (2013), respectively.

One characteristic of experiments with large sample sizes is unwanted batch effects. Batch effects are known to be common sources of non-biological variation. Any equipment (such as a microarray chip) that is made in batches and used in an experiment is one potential type of batch effect. These effects are typically unmeasured but can impact analysis. In 2014, Fortin *et al.* (2014) proposed the functional normalization procedure, an unsupervised technique, that attempts to remove non-biological variation by adjusting for covariates estimated from control probes. The

focus of this method is to normalize methylation data that have global methylation changes, such as in cancer and normal samples. For a systematic study and comprehensive list of between and within-array normalization methods and the associated R packages where such methods could be implemented, see Wang *et al.* (2015) and Shiah *et al.* (2017).

2.5. DIFFERENTIAL METHYLATION TESTING METHODS

Epigenome-wide association studies (EWAS) were intended to identify disease risk factors by using statistical techniques to test associations between disease states and DNA methylation at individual CpG sites or genomic regions.

2.5.1. Site-level Testing. Site-level testing refers to testing for differences in methylation patterns between groups representing differences in conditions (e.g. disease vs. healthy) on a base-pair level. CpG sites that possess significant differences in methylation levels are called Differentially Methylated Positions (DMP). Several methods have been developed in the literature to test for differential methylation at each CpG locus. Different methods are needed for microarray and bisulfite sequencing data. This section focuses on popular methods for DNA methylation microarray data, as that is the focus of this work. Robinson *et al.* (2014) and references therein provide a review of site-level testing methods developed for bisulfite sequencing data.

The Linear Models for Microarray data (*limma*) by Smyth (2004) is arguably the most commonly used approach by many researchers to test for differential methylation at individual CpG sites. It was initially developed as a general and practical approach for identifying genes that are differentially expressed across conditions in designed microarray experiments. However, the method readily applies to DNA methylation data. *limma* uses a linear modeling approach combined with the empirical Bayes technique developed by Efron and Morris (1972) to borrow information across genomic locations to estimate site-level variance. In some microarray experiments,

the number of samples is small and in such cases inferences made are not stable. The *limma* framework solves this problem through the use of a moderated statistic which employs the posterior residual standard deviation in place of the ordinary standard deviation. This shrinks CpG site-wise residual sample variances towards a pooled estimate leading to much more stable inferences, especially in small sample-sized experiments (Ritchie *et al.*, 2015; Smyth, 2004). In terms of site-level testing, *limma* can be explained in the following steps:

1. Fit a linear model at the ith CpG site. Consider a response vector of M-values $\mathbf{y}_i^T = (y_{i1}, \ldots, y_{in})$ for the ith CpG site. The model $\mathbf{y}_i = X\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i$, assumes that $E(\mathbf{y}_i) = X\boldsymbol{\alpha}_i$ where X is a design matrix, $\boldsymbol{\alpha}_i$ is a coefficient vector, and $\boldsymbol{\epsilon}_i$ is an error vector. Further, assume $\text{var}(\mathbf{y}_i) = W_i\sigma_i^2$ where W_i is a known non-negative definite weight matrix.

2. Define the hypothesis of interest. This is achieved by defining contrasts of the coefficients that are assumed to be of biological interest in the form $\boldsymbol{B}_i = C^\top \boldsymbol{\alpha}_i$ where C is a contrast matrix. Let V_i be a positive definite matrix not depending on s_i^2. The contrast estimators are $\hat{\boldsymbol{B}}_i = C^T\hat{\boldsymbol{\alpha}}_i$ with estimated covariance matrices $\text{var}\left(\hat{\boldsymbol{B}}_i\right) = C^T V_i C s_i^2$. It is of interest to test whether individual contrast values B_{ig} are equal to zero or not. The hypothesis being tested by *limma* is:

$$H_0 : B_{ig} = 0 \tag{2.4}$$

$$H_1 : B_{ig} \neq 0 \tag{2.5}$$

where g is the specific contrast to be tested at the ith CpG site.

3. Test the hypothesis of interest. This is achieved using the moderated t-statistic given by

$$\tilde{t}_i = \frac{\hat{B}_{ig}}{\tilde{s}_i \sqrt{v_{ig}}} \tag{2.6}$$

where v_{ig} is the gth diagonal element of $C^\top V_i C$. This statistic estimates s_i, the posterior residual standard deviation, using empirical Bayes as follows. The unknown variances σ_i^2 are assumed to have a prior distribution given by

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0} \tag{2.7}$$

where s_0^2 is a prior estimator of the variance and d_0 is the degrees of freedom Next, this prior distribution is updated using observed data to obtain the posterior distribution whose mean given by

$$\tilde{s}_i^2 = E\left(\sigma_i^2 \mid s_i^2\right) = \frac{d_0 s_0^2 + d_i s_i^2}{d_0 + d_i}. \tag{2.8}$$

d_i denotes the residual degrees of freedom for the linear model for CpG site i.

4. Calculate p-values. The moderated t-statistics are then used to calculate p-values. Due the multiple hypotheses that are tested (one for each CpG site), the p-values are then adjusted for multiple testing to control the false discovery rate (Benjamini and Hochberg, 1995a).

In 2012, Wang *et al.* (2012) introduced the Illumina Methylation Analyzer (IMA) for both site and region-level testing. It uses a Wilcoxon rank-sum test in their R package, *IMA*, which is based on β values as their default method for testing for differences in methylation between treatment and control groups. *IMA* also provides a Student's t-test option as well. In the same year, Barfield *et al.* (2012) developed CpGassoc, an R package, that uses a fixed or mixed effects model to test for methylation differences at CpG sites. Some characteristics of their method in-

clude: the ability to handle covariates, the ability to model batch effects as fixed or as random effects among others. Other methods existing in the literature for site-level testing include COHCAP (Warden *et al.*, 2013a), MethLAB (Kilaru *et al.*, 2012) and penalized logistic regression for DNA methylation data with case-control by Sun and Wang (2012) among others.

2.5.2. Region-level Testing. While site-level testing can be informative (Weaver *et al.*, 2004), a more biologically relevant and statistically appropriate approach involves testing across clusters of CpG sites, referred to as *regions*. These regions of methylation frequently align with biologically functional units, such as genes. The term "statistically appropriate" in this context refers to the increased statistical power achieved through region-level testing, which increases the ability to detect continuous methylation differences across a given region (Robinson *et al.*, 2014). Certain diseases have even been associated with differentially methylated regions (DMRs) in the epigenome. For instance, a type of colorectal cancer, called CpG island methylator phenotype (CIMP) cancer, has a high frequency of methylated genes (Lao and Grady, 2011).

Regions can be defined in two different ways. Some methods use predefined genomic regions (CGIs, Open sea, CGI shores, among others). Other methods define regions based on the data or array. These user-defined regions typically utilize statistics or p-values (adjusted for multiple testing) from site-level tests to group contiguous sites together and form regions based on some criteria (Mallik *et al.*, 2019). Though the predefined region-level testing approach can reduce the number of tests that need to be considered in controlling the false discovery rate (FDR), the problems it possesses are two-fold: (1) regions are comprised of only a subset of the probes presented on the array, which may bias results and (2) it requires defining genomic regions prior to evaluation, which forces DMRs to have artificial start and end points. Two examples of predefined region-level testing methods include *IMA* by Wang *et al.*

(2012) and COHCAP by Warden *et al.* (2013a). The next three subsections describe common user-defined DMR testing methods for array-based data, which is the focus of this work. For a good review of differential methylation detection methods for bisulfite sequencing data, see (Piao *et al.*, 2021; Shafi *et al.*, 2018).

2.5.2.1. Bump Hunter. The bump hunting method proposed by Jaffe *et al.* (2012) is a DMR identification tool that takes into consideration the correlations of methylation levels between nearby CpG sites (co-methylation). This method starts off by first fitting a linear regression model at each CpG site as in (2.9)

$$Y_{ji} = \mu_i + \beta_i X_j + \sum_{k=1}^{p} \gamma_{ki} Z_{jk} + \sum_{l=1}^{q} a_{li} W_{jl} + \varepsilon_{ji} \tag{2.9}$$

where Y_{ji} = M-values at CpG site i for individual j, μ_i represents the population-level DNA methylation profile of the healthy group, X_j = disease status for individual j, β_i = measures association between X_j and Y_{ji} at CpG site i, Z_{jk} = measured confounders (e.g. sex, age, race), γ_{ki} = effect of confounder k at CpG site i, W_{jl} = unmeasured confounders (e.g. batch effects), a_{li} effect of unmeasured confounder l at site i and ε_{ji} = error.

Next, the coefficient, β_i, representing the difference in average methylation levels between two groups (e.g. disease vs. healthy) at each CpG site, is then used to implement the bump hunting methods in the following steps (Jaffe *et al.*, 2012):

1. Estimate $\beta^*(t)$, a smooth function via LOWESS (Locally Weighted Scatterplot Smoothing) smoothing using the β_i's.

2. Use the smooth function $\beta^*(t)$ to estimate the regions R_n, $n = 1, 2, \cdots, N$ for which $\beta^*(t) \neq 0 \ \forall t \in R_n$. R_n are then contiguous intervals (genomic regions) for which methylation levels at consecutive measured CpG sites are significantly different between the groups.

3. Use permutation to assign statistical uncertainty to each estimated region.

A key element in their model is accounting for batch effects (via surrogate variable analysis, Leek and Storey (2007)). The Bump Hunter procedure is implemented in the R/Bioconductor packages *bumphunter* and *minfi*.

2.5.2.2. Probe Lasso. The Probe Lasso DMR detection method developed by Butcher and Beck (2015), fits a linear model by regressing M values on group (e.g. disease vs healthy) at each CpG site. This DMR detection method was developed under the motivation that probe spacing on the beadchip arrays is not uniform with respect to gene feature or genomic annotation. That is, probes near the transcription Start Site (TSS) are the most densely spread compared to intergenic regions (IGRs). Thus, probe lasso works by generating dynamic, flexible boundaries or windows around each probe based on the type of genetic/epigenetic feature the probe is located in (e.g. TSS500, Gene body, etc.). Based on this feature, it "throws" a lasso (constructs a genomic window or region) around each probe centered at the target locus with the size of the lasso depending on the type of genetic/epigentic feature where the probe is located. A region is selected if the number of significant probes within the probe-lasso bounds is at least equal to a user-specified threshold. Next, a p-value is estimated for each region using Stouffer's method (Stouffer *et al.*, 1949) to assign weights to the individual p-values based on a correlation matrix of β values (Butcher and Beck, 2015). Probe Lasso is implemented in the R/Bioconductor package *ChAMP*.

2.5.2.3. DMRcate. The DMRcate method developed by Peters *et al.* (2015), first fits a linear model by regressing the M-values on group status (e.g. disease vs. healthy) at each CpG site using the empirical Bayes technique from the *limma* R/Bioconductor package. Then the DMR identification is done following the steps below:

1. A statistic, $Y_i = t_i^2$ is calculated at each CpG site, i, where t is the moderated t-statistic (see section 2.5.1) from the linear model fit, which denotes the group effect or the site-level difference in methylation.

2. Apply a kernel-based Gaussian smoother (2.10) with bandwidth λ (scaled by some factor C, for $C \in \mathcal{R}^+$) at each location on the Y_i's:

$$S_{KY}(i) = \sum_{j=1}^{n} K_{ij} Y_j \tag{2.10}$$

   where $K_{ij} = \exp\left(\frac{-[x_i - x_j]^2}{2\sigma^2}\right)$, $x_1 < x_2 < \cdots < x_n$ are CpG site positions for some chromosome, and $\sigma = \lambda/C$. This Gaussian kernel smoothing is employed as a way of borrowing the co-methylation (similar methylation profiles) information known to exist among "nearby" CpG sites.

3. Compute p-values for the local kernel-weighted statistic $S_{KY}(i)$, at each CpG site using a moment-matching technique via the method of Satterthwaite (Satterthwaite, 1946).

4. Apply the multiple comparison correction via Benjamini and Hochberg (Benjamini and Hochberg, 1995b) to obtain False Discovery Rate (FDR) corrected p-values.

5. Combine nearby contiguous significant CpG sites that are within λ nucleotides from each other to form regions.

6. Use the minimum p-value within a DMR as the representative p-value for that region.

The DMRcate method is implemented in the R/Bioconductor package *DMRcate* (Peters *et al.*, 2015).

2.6. MOTIVATION

In the early part of 2015, Li *et al.* (2015) reported in their article titled, "An evaluation of statistical methods for DNA methylation microarray data analysis", that amongst the methods evaluated they would recommend the region-level Bump Hunter method over the site-level empirical Bayes method (such as *limma*) when DNA methylation levels are correlated across CpG loci, as their method improved statistical power. It has since been clear the advantage of DMR detection over DMP detection. Later that same year, Probe Lasso and DMRcate methods were published in the literature. Simulation studies comparing Bump Hunter, Probe Lasso and DMRcate in the paper that proposed the DMRcate method, showed that DMRcate outperformed its counterparts based on the area under the precision-recall curve (AUPRC) metric (Peters *et al.*, 2015). The AUPRC is a metric that provides a single summary measure of the DMR detection method's performance across all possible thresholds for defining a DMR. A perfect DMR detection method has an AUPRC of 1, while a method that is no better than random has an AUPRC equal to the proportion of CpG sites that are in true DMRs. Four years later, Mallik *et al.* (2019) performed a comprehensive evaluation of Bump Hunter, Probe Lasso, DMRcate, and comb-p (a Python-based DMR identifier) methods under 60 different parameter settings. They evaluated the precision, recall, F1 score, the AUPRC, the type I error rate (see Ma and He (2013) for definitions of these terms) and other metrics. They concluded that DMRcate and comb-p were the two best methods of the four methods they compared in terms of precision, power and execution time. The results of these two studies perhaps provide evidence that DMRcate's novel way of detecting DMRs is a useful lens through which to view DMR detection.

All aforementioned DMR detection methods rely on findings from Eckhardt *et al.* (2006), which says that methylation levels between CpGs that are within 1000 bp are highly correlated. In recent studies done by Sun and Sun (2019) on within-sample

co-methylation of normal tissues, they concluded that co-methylation regions are as short as a few hundred bp. Another recent preliminary analysis of methylation patterns between consecutive CpG sites by Sun *et al.* (2019), revealed that co-methylation region lengths differed significantly for unmethylated and methylated states. These recent findings reveal the somewhat complex nature of the co-methylation patterns in DNA methylation data.

In this work, a general locally-weighted statistic for DMR detection is proposed. DMRcate is a special case of this general statistic. Next, a normalized kernel-weight is proposed as a superior way of borrowing information from nearby CpG sites. Asymptotic properties of this statistic are then studied. The method is then evaluated via simulation studies and applied to a dataset from NCBI's Gene Expression Omnibus database. In summary, a new method is developed that better captures or borrows the right amount of information from nearby CpG sites in detecting DMRs.

2.7. MATHEMATICAL BACKGROUND

In this section, mathematical details are provided to facilitate the understanding of the proposed DMR detection method. These details are utilized in the method development. These results can be found in numerous texts, including Härdle *et al.* (1991), Silverman (1986), Casella and Berger (2021), Mood *et al.* (1974) among others. The following definitions, theorems, and discussions are from Mood *et al.* (1974) and Casella and Berger (2021). First the distributions used in this work are defined along with some results about relationships between these distributions.

Definition 1 (Normal Distribution)**.** *A random variable X is defined to be normally distributed if its density is given by*

$$f_X(x) = f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, \tag{2.11}$$

where $-\infty < x < \infty$. *The parameters* μ *(mean) and* σ *(standard deviation) satisfy* $-\infty < \mu < \infty$ *and* $\sigma > 0$. *Any distribution defined by a density function given in (2.11) is called a normal or Gaussian distribution. Notation:* $X \sim N(\mu, \sigma^2)$.

If a normal random variable has mean 0 and variance 1, it is called a *standard* normal random variable, denoted Z.

Definition 2 (Gamma Distribution)**.** *If* X *has density given by*

$$f_X(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \tag{2.12}$$

where $0 \leq x < \infty$, $\alpha > 0$, $\beta > 0$ *then* X *is defined to have a gamma distribution with mean,* $\alpha\beta$ *and variance,* $\alpha\beta^2$. $\Gamma(\cdot)$ *is the gamma function defined by:*

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx.$$

Definition 3 (Chi-square Distribution)**.** *If* X *is a random variable with density*

$$f_X(x) = \frac{x^{\nu/2-1} e^{-x/2}}{\Gamma(\nu/2) 2^{\nu/2}}, \tag{2.13}$$

where $0 \leq x < \infty$ *and* ν *is a positive integer then* X *is defined to have a chi-square distribution with* ν *degrees of freedom.*

Note that a chi-square density is a particular case of a gamma density with gamma parameters α and β equal, respectively, to $\nu/2$ and 2.

Another distribution of practical importance is the student's t-distribution which is defined as the ratio of a standard normally distributed random variable to the square root of an independently distributed chi-square random variable divided by its degrees of freedom.

Theorem 1 (Student's t-distribution)**.** *If Z has a standard normal distribution, if U has a chi-square distribution with ν degrees of freedom, and if Z and U are independent, then*

$$T = \frac{Z}{\sqrt{U/\nu}} \tag{2.14}$$

has a Student's t-distribution with ν degrees of freedom.

Next, a theorem is given for the F-distribution, which is defined as the ratio of two independent chi-square random variables divided by their respective degrees of freedom.

Theorem 2 (F-distribution)**.** *Let U be a chi-square random variable with μ degrees of freedom; let V be a chi-square random variable with ν degrees of freedom, and let U and V be independent. Then the random variable*

$$F = \frac{U/\mu}{V/\nu} \tag{2.15}$$

is distributed as an F distribution with μ and ν degrees of freedom with density defined in (2.16)

$$f_F(x) = \frac{\Gamma[(\mu+\nu)/2]}{\Gamma(\mu/2)\Gamma(\nu/2)} \left(\frac{\mu}{\nu}\right)^{\mu/2} \frac{x^{(\mu-2)/2}}{[1+(\mu/\nu)x]^{(\mu+\nu)/2}}. \tag{2.16}$$

We note that if F is an $F-$distributed random variable with μ and ν degrees of freedom then $E(F) = \dfrac{\nu}{\nu-2}$ for $\nu > 2$ and $Var(F) = \dfrac{2\nu^2(\mu+\nu-2)}{\mu(\nu-2)^2(\nu-4)}$ for $\nu > 4$.

Another important result used in this work is the relationship between between the gamma distribution and the beta distribution in the next corollary.

Corollary 1 (Gamma to Beta Relationship)**.** *Let X and Y be two independently and identically distributed random variables each having a Gamma distribution with parameters α and β as in definition 2, then the random variable*

$$W = \frac{X}{X+Y} \tag{2.17}$$

has a beta distribution (Cramér, 2016) denoted $W \sim Beta(\alpha, \alpha)$ with density function

$$f(w) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha)} w^{\alpha-1}(1-w)^{\alpha-1} \ , \ 0 < w < 1. \tag{2.18}$$

Proposition 1 (F to Chi-square Relationship)**.** *Let Y be a random variable with probability density function (PDF) given in* (2.16) *so that Y is defined as in* (2.15)*. Then as $\nu \to \infty$,*

$$Y \xrightarrow{d} \chi^2_\mu/\mu \tag{2.19}$$

where χ^2_μ is a chi-squared distribution with μ degrees of freedom and $\xrightarrow{d}$ means converges in distribution (refer to Definition 6 for details).

The next background information provided is important to the large-sample results that underpin this work. The following definitions, concepts and theorems can be found in Billingsley (1995); DasGupta (2008); Jiang (2010); Lehmann (2004); Resnick (1999).

Definition 4 (Convergence in Probability)**.** *Let $\{X_n\}$, X be a sequence of random variables. Then $\{X_n\}$ converges in probability to X as $n \to \infty$ ($\{X_n\} \xrightarrow{p} X$) if for each $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\{X_n\} - X| > \epsilon) = 0. \tag{2.20}$$

The limiting random variable X may be a constant; in which case we write $\{X_n\} \xrightarrow{p} c$ (a constant). This means for large n, there is almost no variation in the random variable X.

A statistical version of convergence in probability is the concept of consistency of a statistic or an estimator, which is defined in corollary 2.

Corollary 2 (Consistency)**.** *Let $T_1, T_2, \cdots, T_n, \cdots$ be a sequence of estimators of $\tau(\theta)$. The sequence $\{T_n\}$ is defined to be weakly consistent if for every $\epsilon > 0$ the following is satisfied:*

$$\lim_{n\to\infty} P(|T_n - \tau(\theta)| > \epsilon) = 0 \tag{2.21}$$

for every θ in Θ.

Convergence in probability is weak and merely requires that the probability difference becomes small. A much stronger form of convergence is almost sure convergence, which is defined below.

Definition 5 (Almost Sure Convergence)**.** *A sequence of random variables $\{X_n\}$ converges with probability 1 (almost surely) to a random variable X ($\{X_n\} \xrightarrow{a.s.} X$) if*

$$P([\omega : \lim_{n\to\infty} \{X_n\}(\omega) = X(\omega)]) = 1. \tag{2.22}$$

Another aspect of large-sample theory deals with convergence in distribution, which involves approximations to probability distributions and the limit theorems that underlie these approximations.

Definition 6 (Convergence in Distribution)**.** *Suppose $\{X_n\}$, X are random variables. Then $\{X_n\}$ converges in distribution to X as $n \to \infty$ ($\{X_n\} \xrightarrow{d} X$) if*

$$\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x) = F(x) \tag{2.23}$$

for each continuity point of the distribution function $F(x)$. Note this refers to distributions with cumulative distribution functions (CDFs), F_n, converging to a distribution function F. i.e. $F_n(x) \to F(x)$ at all continuity points x of F.

The concept of uniform tightness is crucial to the study of convergence of distributions of random variables.

Definition 7 (Uniform Tightness)**.** *A collection of random variables $\{X_\alpha\}_{\alpha \in A}$ is uniformly tight if $\forall \epsilon > 0$, there exists $M < \infty$ such that:*

$$\sup_{\alpha} P(|X_\alpha| \geq M) \leq \epsilon.$$

The next theorem pertains to the generalization of the famous Central Limit Theorem to the case where the summands are independent but not identically distributed.

Theorem 3 (Lindeberg-Feller)**.** *Let $\{X_n, n \geq 1\}$ be independent (but not necessarily identically distributed) random variables and suppose X_k has distribution F_k, and that $E(X_k) = 0$, $Var(X_k) = \sigma_k^2$. Define*

$$Y_n = X_1 + X_2 + \ldots + X_n \tag{2.24}$$

$$s_n^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2 = Var\left(\sum_{i=1}^{n} X_i\right). \tag{2.25}$$

$\{X_k\}$ satisfies the Lindeberg-Feller condition (uniform integrability) if for all $t > 0$ as $n \to \infty$ we have

$$\frac{1}{s_n^2} \sum_{i=1}^{n} E\left(X_k^2 \mathbf{1}_{[|X_k/s_n| > t]}\right) \to 0. \tag{2.26}$$

The Lindeberg-Feller condition implies the Uniform Asymptotic Negligibility (UAN) condition below:

$$\max_{1\leq k\leq n} \frac{\sigma_k^2}{s_n^2} \to 0 \quad as \quad n \to \infty. \tag{2.27}$$

The Lindeberg-Feller condition in Theorem 3 further implies:

$$\frac{Y_n}{s_n} \xrightarrow{d} \mathcal{N}(0,1) \tag{2.28}$$

where $\mathcal{N}(0,1)$ is a normal random variable with mean 0 and variance 1. Note that (2.28) is the so-called Lindeberg-Feller Central Limit Theorem (CLT).

Usually an easier condition to check than Lindeberg-Feller condition is the Lyapounov condition which is stated in the next corollary.

Corollary 3 (Lyapounov Condition)**.** *Let $\{X_n, n \geq 1\}$ be an independent sequence of random variables satisfying $E(X_k) = 0$, $Var(X_k) = \sigma_k^2 < \infty$, $s_n^2 = \sum_{k=1}^n \sigma_k^2$. If for some $\delta > 0$:*

$$\frac{\sum_{i=1}^n E|X_k|^{2+\delta}}{s_n^{2+\delta}} \to 0, \tag{2.29}$$

then the Lindeberg-Feller condition (Theorem 3) holds and hence the CLT.

Now a Lemma and Theorem (due to Adler and Rosalsky (1991)) are introduced that pertain to situations when weighted sums of independent and identically distributed (iid) random variables obey the weak law of large numbers.

Lemma 1. *Let $\{Y, Y_n\}$ be iid random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and let $\{a_n, n \geq 1\}$ and $\{b_n, n \geq 1\}$ be constants with $a_n \neq 0$, $b_n > 0$, $n \geq 1$ and $\left\{c_n = \frac{b_n}{|a_n|}, n \geq 1\right\}$.*

If

$$nP(|Y_n| > c_n) = o(1) \tag{2.30}$$

and either

$$c_n \uparrow, \frac{c_n}{n} \downarrow, \sum_{j=1}^{n} a_j^2 = o(b_n^2), \quad \text{and} \quad \sum_{j=1}^{n} \left(\frac{c_j}{j}\right)^2 = \mathcal{O}\left(\frac{b_n^2}{\sum_{j=1}^{n} a_j^2}\right) \tag{2.31}$$

or

$$\frac{c_n}{n} \uparrow \quad \text{and} \quad \sum_{j=1}^{n} a_j^2 = \mathcal{O}\big(na_n^2\big) \tag{2.32}$$

then

$$\sum_{j=1}^{n} a_j^2 E\left(Y^2 \mathbf{1}_{\{|Y| \le c_n\}}\right) = o(b_n^2) \tag{2.33}$$

where the symbols $u_n \uparrow$ or $u_n \downarrow$ are used to indicate that the given numerical sequence $\{u_n, n \ge 1\}$ is monotone increasing or decreasing, respectively.

Theorem 4. *Let $\{Y_n, n \ge 1\}$ be iid random variables and let $\{a_n, n \ge 1\}$ and $\{b_n, n \ge 1\}$ be constants satisfying $a_n \ne 0$, $b_n > 0$, $n \ge 1$ and either (2.31) or (2.32) hold. If (2.30) holds, then*

$$\frac{\sum_{j=1}^{n} a_j \left(Y_j - EY\mathbf{1}_{\{|Y| \le c_n\}}\right)}{b_n} \xrightarrow{p} 0. \tag{2.34}$$

The next definition pertains to the concept of a Kernel function, which holds significance as the method proposed in Paper I is based upon it.

Definition 8 (Kernel)**.** *A kernel is any measurable weighting function which satisfies the following conditions. For every* $x \in \mathcal{R}$*:*

$$K(x) = K(-x) \tag{K.1}$$

$$\int_{\mathcal{R}} K(t)dt = 1 \tag{K.2}$$

$$\int_{\mathcal{R}} tK(t)dt = 0 \tag{K.3}$$

$$\int_{\mathcal{R}} t^2K(t)dt < \infty \tag{K.4}$$

$$\int_{\mathcal{R}} K^2(t)dt < \infty. \tag{K.5}$$

Note that the symmetry condition in (K.1) implies that $\int tK(t)dt = 0$.

3. BACKGROUND TO PAPER II

The second part of this dissertation deals with statistical methods for time-to-event data that includes a cure fraction. The primary characteristic of time-to-event data is, as the name suggests, the *length of time* until an event occurs. This type of data is often referred to as survival data in the field of statistics. In classical survival analysis, it is typically assumed that all subjects or units in the population will eventually experience the event of interest. However, this is not always the case. When some subjects will *never experience the event of interest*, they are referred to as the cure fraction. The following sections are dedicated to presenting survival analysis both without and with a cure fraction, along with estimation methods for cure survival models.

3.1. SURVIVAL ANALYSIS AND THE FEATURES OF SURVIVAL DATA

Survival analysis aims to describe and model time-to-event data, that is, data concerning the duration between a starting point and a predetermined event of interest. This duration is often referred to as the *survival time*, *event time*, or *failure time*. The event in question is not always death, as is typically the case in a medical context; it can also be the recurrence of a disease. In fact, survival analysis has been applied in other fields such as economics (e.g., the time until an unemployed person finds a new job) and engineering (e.g., the time until a machine breaks down) (Kleinbaum and Klein, 2012).

Survival data are characterized by two key features: the survival time, denoted by T (a positive continuous random variable), and a censoring feature, which renders classical statistical methods unsuitable for this type of data. Censoring arises when the exact survival time of some subjects is unknown. There are various types of

censoring, including right censoring, left censoring, and interval censoring. However, this work focuses on *random right censoring*, which refers to situations where the censoring time is a random variable. This corresponds to scenarios where all participants can enter and exit the study at different times. If the random censoring time is denoted by C, then the observed data consists of the observed time, $Y = \min(T, C)$, and the censoring indicator, $\Delta = I(T \leq C)$ (Lee and Wang, 2003).

When participants are required to meet a condition or to have experienced the event to be included in the analysis, this is referred to as *truncation*. Truncation is more severe than censoring because it implies that some individuals do not appear in the dataset at all. This work focuses solely on right-side censorship without truncation, which is representative of many real-world scenarios (Chiou *et al.*, 2019).

The censoring is assumed to be uninformative, meaning that the distribution of the censoring times does not depend on the parameters appearing in the survival distribution, and that T and C are independent given the covariates. This assumption is crucial as it allows the censoring mechanism to be disregarded when modeling the survival time, simplifying the analysis. However, if this assumption is violated, it could lead to biased estimates (Therneau *et al.*, 2000).

3.1.1. Definition of Survival Quantities. In classical statistical analysis, the probability density function, $f(t)$, and the cumulative distribution function, $F(t) = P(T \leq t)$, are usually of interest. However, in survival analysis, the mathematical functions of interest are the survival and hazard functions (Lee and Wang, 2003; Price, 2000).

The *survival function*, $S(t)$, is given by

$$S(t) = P(T > t) = 1 - F(t), \quad t \geq 0. \tag{3.1}$$

The survival function is the probability that an individual survives beyond time t. It represents the probability that the event of interest has not yet occurred by time t. The *hazard rate* or *hazard function*, $\lambda(t)$, is given by

$$\lambda(t) = \lim_{\epsilon \to 0} \frac{P(t \leq T < t + \epsilon \mid T \geq t)}{\epsilon}, \quad t \geq 0, \tag{3.2}$$

$$= f(t)/S(t) \tag{3.3}$$

where $\lambda(t) \geq 0$. The hazard function is the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. It can be interpreted as the risk of the event occurring at the next instant.

Additionally, the *cumulative hazard function*, $\Lambda(t)$ is defined as

$$\Lambda(t) = \int_0^t \lambda(u)du, \quad t \geq 0. \tag{3.4}$$

The cumulative hazard function can be interpreted as the accumulated risk of the event occurring up to time t. It is a measure of the total risk experienced by the individual up to time t.

These functions are fundamental to survival analysis and are often visualized to provide insights into the data. For instance, a plot of the survival function can show how the probability of survival changes over time, while a plot of the hazard function can show how the risk of the event changes over time (Kleinbaum and Klein, 2012; Lee and Wang, 2003).

3.1.2. Survival Models. When modeling the effect of covariates on the survival or the hazard, commonly used models include the Cox Proportional Hazard (Cox, 1972), the Additive Hazard (Aalen, 1980; Huffer and McKeague, 1991; Lin and Ying, 1994) and the Accelerated Failure Time models (Klein and Moeschberger,

2003). An overview of these three methods is given in the following sections. Other less popular models include the Accelerated Hazard (Chen and Wang, 2000) and the Proportional Odds (Bennett, 1983) models.

3.1.2.1. Cox Proportional Hazards model. The Cox Proportional Hazards (PH) model, developed by Cox (1972), is a widely used semi-parametric model for incorporating covariate information that may influence failure time. Let $\boldsymbol{Z}$ represent a vector of covariates and $\boldsymbol{\beta}$ a vector of regression coefficients. The model is then defined by the hazard relationship:

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t) \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}\right) \tag{3.5}$$

where $\lambda_0(t)$ is an unspecified baseline hazard common to all individuals. The model assumes a multiplicative effect of the covariate on the baseline hazard and proportional hazard ratios for all subjects. Using the relationships between survival quantities, the cumulative hazard function and survival function can be derived as:

$$\Lambda(t \mid \boldsymbol{Z}) = \Lambda_0(t) \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}\right) \tag{3.6}$$

and

$$S(t \mid \boldsymbol{Z}) = S_0(t)^{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}\right)} \tag{3.7}$$

respectively, where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ and $S_0(t) = P(T > t \mid \boldsymbol{Z} = 0) = \exp\{-\Lambda_0(t)\}$ represent the baseline cumulative hazard and baseline survival functions, respectively. Inferences about $\boldsymbol{\beta}$ are made via the partial likelihood of Cox (1975). See (Cox and Oakes, 1984; Kleinbaum and Klein, 2012) for details.

3.1.2.2. Additive Hazard model. In the additive hazard (AH) model, the hazard function for a given set of covariates is expressed as the sum of the baseline hazard function and the regression function of covariates. This is in contrast to the

PH model, which uses the product of these two components. Various forms of the additive hazard model have been proposed. Aalen (1980, 1989) proposed a fully non-parametric additive hazard, while Lin and Ying (1994) proposed the semi-parametric additive hazard model (3.8), a special case of Huffer and McKeague (1991)'s partly non-parametric model, given by:

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}. \tag{3.8}$$

Subsequently, the cumulative hazard function is given by:

$$\Lambda(t \mid \boldsymbol{Z}) = \int_0^t \left(\lambda_0(u) + \boldsymbol{\beta}^\top \boldsymbol{Z}\right) du. \tag{3.9}$$

Lin and Ying (1994) developed estimation and inference methods for the regression coefficient vector $\boldsymbol{\beta}$ using a counting process framework.

3.1.2.3. Accelerated Failure Time model. The Accelerated Failure Time (AFT) model, first introduced by Cox (1972), offers an alternative to the Cox PH and Additive Hazard models. Unlike its counterparts, the AFT model defines the survival function as:

$$S(t \mid \boldsymbol{Z}) = S_0 \left\{ \frac{t}{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}\right)} \right\}, \tag{3.10}$$

where $S_0(\cdot)$ is a parametric baseline survival function. The covariates act multiplicatively on t, thereby accelerating or decelerating the event rate. See Kalbfleisch and Prentice (2011) and Klein and Moeschberger (2003) for a thorough discussion on AFT models).

While the AFT model offers valuable insights, the additive and proportional hazard models are the principal frameworks used to investigate the relationship between risk variables and the time to an event (Lin and Ying, 1994). The additive hazard and PH models, underpinned by robust empirical evidence, offer solid em-

pirical foundations and provide complementary insights into this association. The Cox PH model has gained popularity in the literature, partly due to its theoretical properties and easy implementation in statistical software like R and SAS. However, there are situations where the additive hazards model is more preferable. One such scenario is when the cumulative hazard is minimal, indicating infrequent events. In such cases, the change in cumulative hazard reflects the difference in disease risk attributable to exposure, also known as the excess or attributable risk. In such cases, an additive hazard regression model may be preferable (Madadizadeh *et al.*, 2017).

3.2. CURE MODELS

The following sections will provide an overview of cure models, including their historical development, elucidation of survival quantities in the context of cure, and an introduction to the mixture cure model.

3.2.1. Background and History. In epidemiological studies focusing on the time until relapse of a specific disease, some subjects may never experience the disease, especially when monitored over a long period. In such cases, these individuals are referred to as *cured* or *immune* to that condition. Classical survival analysis techniques applied to time-to-event data with a cure fraction can result in an overestimation of censored observations (Amico and Van Keilegom, 2018). To account for the presence of a cure fraction in the data, cure models were developed. These models are not limited to medical research, from which they derive their name, but are also employed in various other fields, including engineering, economics, and the humanities (Maller and Zhou, 1996). For example, in engineering, there are machines that never fail, and researchers are interested in understanding the time until a machine breaks down. Similarly, in economics, there are jobless individuals who may never find employment, and the event of interest is the duration until they secure a job. Other studies investigate topics such as the time it takes to resolve an issue, the

lifespan of a bank before failure, the time of purchase for a new product by a client, or the time until a freed convict is arrested again. In each of these scenarios, there is a possibility that individuals will never experience the event of interest. The lifetimes of these *immune* subjects, in relation to the event under study, are represented in the censored portion of the data. This circumstance has given rise to the importance of cure models as an important area of study.

The analysis of survival data with a cure fraction has a long history, dating back to the early 1950s. Boag (1949) and Berkson and Gage (1952) were pioneers in formulating explicit models for such data. Boag (1949) estimated the proportion of cured breast cancer patients by defining a patient as cured if they had a five-year survival rate. Berkson and Gage (1952) introduced the concept of the susceptible and cured groups, dividing the population into two categories. A group of treated subjects were considered cured if they had the same survival distribution as the general population who had never had the disease of interest.

There are two main classes of cure models: mixture cure models and promotion time (or bounded cumulative hazard) cure models. Mixture cure models assume that the population consists of both susceptible individuals who may experience the event and immune individuals who will never experience the event. The survival function in these models is a mixture of the survival functions of the susceptible and immune individuals. On the other hand, promotion time cure models assume that all individuals are initially susceptible to the event, but some individuals may require an extremely long time to experience the event, effectively making them cured in practical terms. The survival function in these models is derived from a bounded cumulative hazard function (Amico and Van Keilegom, 2018; Peng and Yu, 2021). This dissertation will focus discussions on the mixture cure models, however the reader should see Peng and Yu (2021) for discussions on the promotion time cure models.

In their excellent text *Survival Analysis with Long Term Survivors*, Maller and Zhou (1996) provide a comprehensive discussion of the early development of mixture cure models and their theoretical underpinnings. However, the field of cure models has developed immensely and a more recent discussion of the current methods and their implementations can be found in Peng and Yu (2021).

3.2.2. Survival Analysis in the Presence of a Cure Fraction: Key Quantities and Concepts. Consider a population comprising both cured and susceptible groups in relation to an event of interest. This division results in two subpopulations, where individuals are either cured with a probability of $1-p$ (exhibiting a degenerate survival function of 1) or uncured with a probability of p (possessing a proper survival function, $S_u(t)$) (Amico and Van Keilegom, 2018; Price, 2000). Refer to Figure 3.1 for illustration.

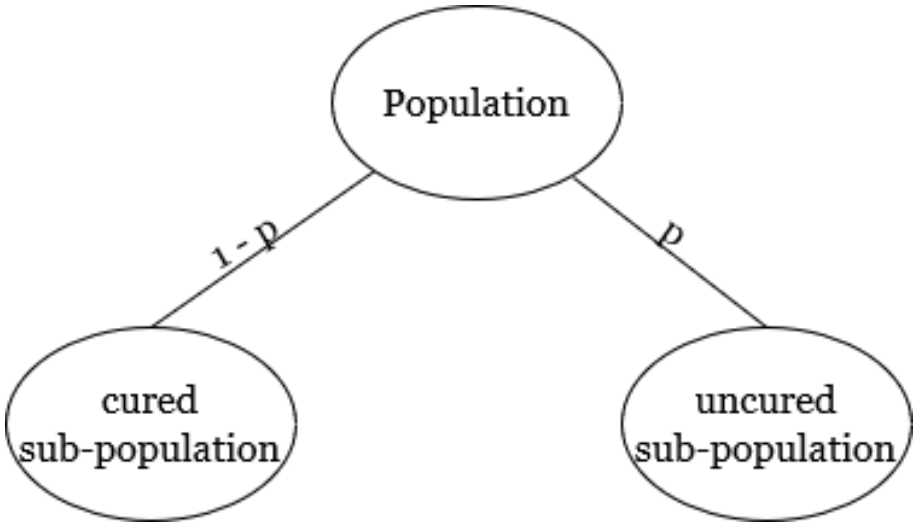

Figure 3.1. A heterogeneous population with cured and uncured sub-populations.

Let T denote the random variable representing the time until the event of interest occurs and $T = \infty$ is allowed in order to represent the situation when the event never happens. T is subject to random right censoring. So that $Y = \min(T, C)$ is observed and $\Delta = I(T \leq C)$, where C is a random censoring time and $I(\cdot)$ is the indicator variable. Consider a population that contains a cure fraction, it is assumed

that $\lim_{t\to\infty} S(t) > 0 := 1 - p$, where $1 - p$ is sometimes referred to as the *cured proportion* or *cure rate*. Assume that there is an indicator variable, $B = I(T < \infty)$, called the *cure status* defined as in (3.11):

$$B = \begin{cases} 1, & T < \infty \text{ with probability } p \\ 0, & T = \infty \text{ with probability } 1 - p. \end{cases} \tag{3.11}$$

At the end of a study, information is only available on the censoring status of the subjects. Consequently, B is not fully observed and is considered a latent variable. For an uncensored observation ($\Delta = 1$), it is obvious $B = 1$. However, censoring affects both the cured (because the event "never" happens) and uncured subjects, because follow-up cannot be infinite (Figure 3.2).

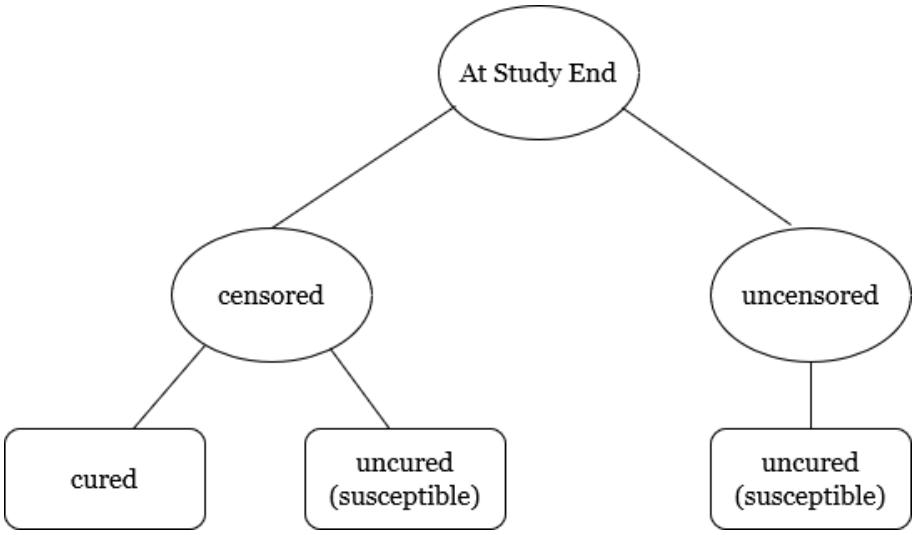

Figure 3.2. A schematic that describes the group (censored or uncensored) in which the cured and uncured may arise.

Let $\boldsymbol{X}$ be a d dimensional set of covariates, $\boldsymbol{Z}$ be a q dimensional set of covariates identical to $\boldsymbol{X}$ or partially or completely different from $\boldsymbol{X}$. Let $S(t|\boldsymbol{X}, \boldsymbol{Z})$ in (3.12) be the (improper) survival function of the (heterogeneous) population described

in Figure 3.1:

$$S(t|\boldsymbol{X},\boldsymbol{Z}) = P(T > t|\boldsymbol{X},\boldsymbol{Z}) = 1 - p(\boldsymbol{X}) + p(\boldsymbol{X})S_u(t|\boldsymbol{Z}) \tag{3.12}$$

where $p(\boldsymbol{X}) = P(B = 1|\boldsymbol{X} = \boldsymbol{x})$ is the probability of being susceptible (incidence) and $S_u(t|\boldsymbol{Z}) = P(T > t|\boldsymbol{Z} = \boldsymbol{z}, B = 1)$ is a proper conditional survival function of susceptibles (latency). Similarly one can write the model as $F(t \mid \boldsymbol{X}, \boldsymbol{Z}) = p(\mathbf{X})F_u(t \mid \boldsymbol{Z})$, where $F(t) = 1 - S(t)$ depends on $\boldsymbol{Z}$ (and not on $\boldsymbol{X}$). Parametric, semi-parametric, and non-parametric families of mixture cure models result from additional model assumptions on $p(\boldsymbol{X})$ and $S_u(t \mid \boldsymbol{Z})$. While the logistic model is typically used to describe the $1 - p(\boldsymbol{X})$ cure rate, other different models have been presented to explain the survivability of the susceptibles.

3.2.3. Mixture Cure Models: A Brief Review of Modeling

Approaches. The work of Boag (1949) initiated the mixture cure model where the incidence was modeled as a constant. Farewell (1982) was the first to introduce covariates in the incidence by assuming a logistic model for the probability of uncured, $p(\boldsymbol{x}) = \exp\left(\gamma^\top \boldsymbol{x}\right) / \left\{1 + \exp\left(\gamma^\top \boldsymbol{x}\right)\right\}$. In a parametric mixture cure model, both the incidence sub-model, $p(\boldsymbol{X})$, and the latency sub-model, $S_u(t \mid \boldsymbol{Z})$, are fully parametrically specified. Since the incidence part involves the probability parameter p, common models that allow different link functions between p and the functional form of some effect $\boldsymbol{\gamma}^\top \boldsymbol{X}$ such as the logistic model via logit link, complementary log-log link and probit link models, among others have been employed (Peng and Yu, 2021). López-Cheda *et al.* (2017) assumed a fully nonparametric mixture cure model where they estimated the cure rate using the kernel estimator of Xu and Peng (2014) (see Amico and Van Keilegom (2018) for details). The latency sub-model incorporates the assumption of a conditional distribution $(T|B = 1)$ for the group that is susceptible to the event of interest. This establishes a functional relationship that depends on $\boldsymbol{Z}$,

a vector of covariates. One common approach to achieve this is through the use of the proportional hazards (PH) assumption. This results in the survival function for the susceptible individuals being expressed as: $S_u(t|\boldsymbol{Z}) = S_0(t)^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})}$.

A relatively popular latency sub-model in the literature is the accelerated failure time (AFT) latency sub-model, which adopts the AFT assumption introduced by Cox and Oakes (1984). The functional relationship of the AFT model is given by: $\log(T|B = 1) = \boldsymbol{\beta}^\top \boldsymbol{X} + \sigma\epsilon$, where σ represents the scale parameter and ϵ is the error term. The baseline survival function $S_0(t)$, which is parametrically specified, is defined by the probability $P(e^{\sigma\epsilon} > t)$. In this case, the conditional distribution of $T|B = 1$ follows the AFT model: $S_u(t|\boldsymbol{Z}) = S_0(te^{-\boldsymbol{\beta}^\top \boldsymbol{Z}})$.

When the latency sub-model is modeled nonparametrically or semiparametrically while keeping the incidence parametric, it leads to a semi-parametric mixture cure model. This is the most popular type of cure model in in the literature. The Cox proportional hazards (PH) model (Cox, 1972) has been extensively used as the model for the conditional survival function, resulting in $S_u(t|\boldsymbol{Z})$ taking the form of (3.7). The first such work for this type of model is attributable to Kuk and Chen (1992), who adapted the marginal likelihood approach proposed by Kalbfleisch and Prentice (1973) to estimate the model parameters. Later, Peng and Dear (2000) and Sy and Taylor (2000) introduced the standard approach to mixture cure modeling, which utilizes the Expected-Maximization (EM) algorithm (Dempster *et al.*, 1977) and assumes the model depends on the latent variable B defined in (3.11). A comprehensive review article by Amico and Van Keilegom (2018) provides a concise overview of cure models, including an discussion of the estimation techniques proposed by Sy and Taylor (2000).

3.3. IDENTIFIABILITY OF THE MIXTURE CURE MODEL

Identifiability of statistical models refers to the ability to uniquely determine model parameters from observed data. This is a critical feature of mixture cure model estimation and inference, particularly for the semiparametric models (Amico and Van Keilegom, 2018), as the latency component $S_u(\cdot)$ is left unspecified. A general and informal rule that applies to all cure models stipulates that the follow-up period of the study needs to be extensive enough. The estimated survival function should have a lengthy plateau comprised of numerous observations that have been right-censored. The fraction of cured subjects may be assessed by noting whether or not the "plateau" in the survival function plot for the whole population comprises solely cured subjects. When the plateau remains constant for a long period of time without declining even gradually, as shown in Figure 3.3, one can be considerably certain that all uncured individuals had their event before the plateau began, and so the cure fraction corresponds to the height of the plateau. In other words, the maximum possible event time should be less than the maximum possible censoring time. Formally, this means (with covariates omitted):

$$\tau_{F_u} < \tau_G \tag{3.13}$$

where $F_u = 1 - S_u$, G is the censoring distribution, $\tau_G = \inf t : G(t) = 1$, and $\tau_F = \inf t : F(t) = 1$ for some distribution F (Maller and Zhou, 1996). Sy and Taylor (2000) proposed that when estimating (part of) the latency sub-model nonparametrically, the survival function should be forced to reach 0 at the longest survival time, called the zero-tail constraint. For a thorough discussion of the identifiability of cure models, see Hanin and Huang (2014).

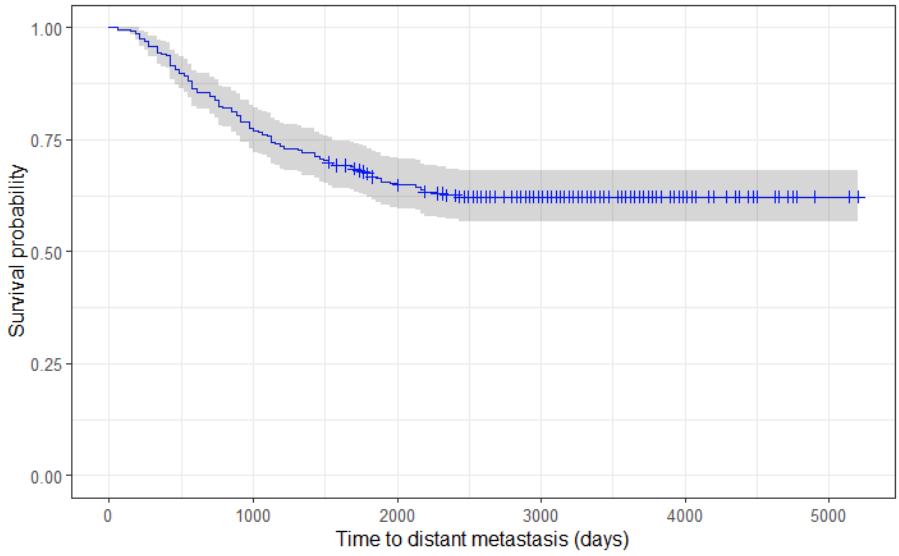

Figure 3.3. Kaplan and Meier (1958) estimator of the survival function for the breast cancer dataset of Wang *et al.* (2005) (+ : censored observations).

3.4. THE EM ALGORITHM

With a brief review of mixture cure models mentioned and identifiability stated, it is now important to describe the tool used for obtaining maximum likelihood estimate, namely the Expectation-Maximization (EM) algorithm. The EM algorithm introduced by Dempster *et al.* (1977) is a powerful widely-used technique for obtaining the maximum likelihood estimates of model parameters when there exists "incomplete data" or where some variables are not observed. The general idea of the EM algorithm is simple. Consider the vector of variables $\boldsymbol{X}$ (parameterized by an unknown vector, $\boldsymbol{\theta}$) and $\boldsymbol{B}$ ($\boldsymbol{B}$ partially observed through $\boldsymbol{X}$). The complete-data likelihood is given by $\mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{B})$. If it is of interest to estimate $\boldsymbol{\theta}$ via maximum likelihood, then the maximum likelihood estimator $\hat{\boldsymbol{\theta}} = \arg\max_\theta \log \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{X})$ is complex. The EM algorithm offers a reasonable solution to (1) estimate the complete-data likelihood from $\boldsymbol{X}$ (E-step) and (2) maximize the estimated log complete-data likelihood (M-step). This is done iteratively and at the $(m+1)^{\text{th}}$ iteration the two steps are:

1. E-step (Expectation): Compute the expected value of the complete data log-likelihood, given the current parameter estimates and the observed data. Mathematically, this can be expressed as:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) = E_{\boldsymbol{B}} \left\{ \log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{B}) \mid \mathbf{X}, \boldsymbol{\theta}^{(m)} \right\}$$

   where $\boldsymbol{\theta}^{(m)}$ are the current estimates of the parameters at the m^{th} iteration.

2. M-step (Maximization): Update the parameter estimates by maximizing the expected complete data log-likelihood $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)})$ with respect to the parameter vector $\boldsymbol{\theta}^{(m)}$:

$$\boldsymbol{\theta}^{(m+1)} = \arg\max_{\theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta^{(m)}}).$$

This process is repeated until convergence to a maximum of the likelihood (McLachlan and Krishnan, 2007).

3.5. MOTIVATION

In the literature on (mixture) cure models, limited work has been done using the semi-parametric additive hazard model to model the susceptibles, resulting in $S_u(t|\boldsymbol{Z}) = e^{-\Lambda_u(t|\boldsymbol{Z})}$, where $\Lambda_u(t \mid \boldsymbol{Z})$ is defined as in (3.9). Currently, there is no latency sub-model that employs the semi-parametric additive hazard model via the EM algorithm. In an extensive simulation study conducted by Legrand (2021), it was concluded that when the proportional hazards assumption is not met due to the presence of a cure fraction, the mixture cure model should be used. However, there is currently no readily implementable alternative method when the proportional hazards assumption is not appropriate or when there is reason to believe that the covariate effects are additive (linear) on the baseline hazard function of the susceptibles. The work in this dissertation fills this gap. For a thorough review of situations where ad-

ditive hazards seem favorable to the PH model, please see *Survival and Event History Analysis: A Process Point of View* by (Aalen *et al.*, 2008, p.155). For the incidence sub-model, the logistic model has been extensively used and preferred due to its ease of interpretation. However, it is too restrictive as it forces the relationship between the log odds ratio and the covariates to be linear. Amico *et al.* (2019) proposed the single-index model as a flexible to capture non-linear or non-logistic relationships. However, it is not always easy to interpret covariates of interest. This work proposes a *generalized partially linear single-index model* (GPLSIM) as an alternative. This model combines interpretability and flexibility well. See section 3.6 for details on this model.

3.6. GENERALIZED PARTIALLY LINEAR SINGLE INDEX MODEL

In this section the Generalized Partially Linear Single-Index Model (GPLSIM) is introduced as the model proposed for the cured fraction. The GPLSIM is a semiparametric version of the generalized linear model (GLM) which was first proposed by Carroll *et al.* (1997). In the GPLSIM the unknown regression function $\mu(\boldsymbol{X_1}, \boldsymbol{X_2}) = E(Y \mid \boldsymbol{X_1}, \boldsymbol{X_2})$ is modeled via a link function H by:

$$H^{-1}(\mu(\boldsymbol{X_1}, \boldsymbol{X_2})) = g\left(\boldsymbol{\alpha}^\top \boldsymbol{X_1}\right) + \boldsymbol{\gamma}^\top \boldsymbol{X_2}, \quad \text{with } \|\boldsymbol{\alpha}\| = 1, \tag{3.14}$$

where H is a known monotone function, Y is the response variable, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are covariates, g is a unknown link function, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$ are coefficient vectors. Note that $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are covariates split from $X = (\boldsymbol{X_1}, \boldsymbol{X_2})$. To ensure identifiability and estimability, it is assumed that the first element of $\boldsymbol{\alpha}$ is positive. To complete the specification of (3.14), it is assumed that $Var(Y \mid \boldsymbol{X_1}, \boldsymbol{X_2}) = V\left\{H\left(g\left(\boldsymbol{\alpha}^\top \boldsymbol{X_1}\right) + \boldsymbol{\gamma}^\top \boldsymbol{X_2}\right)\right\}$ where V is a known positive function. In a GLM, $\mu(\boldsymbol{X_1}, \boldsymbol{X_2})$ is modeled linearly via

a link function h by:

$$H(\mu(\boldsymbol{X_1}, \boldsymbol{X_2})) = \boldsymbol{\alpha}^\top \boldsymbol{X_1} + \boldsymbol{\gamma}^\top \boldsymbol{X_2}, \tag{3.15}$$

where H is usually taken to be the canonical link function (see McCullagh and Nelder (2019)). However, as Carroll *et al.* (1997) recounts, the model in (3.15) is not sophisticated enough to capture the true relationship between the response variable and covariates. As can be noted, (3.14) subsumes (3.15), which allows some predictors to be modeled linearly while others modeled non-linearly. Special cases of the model in (3.14) also result in well-known models. When $\boldsymbol{\alpha} = \mathbf{1}$, a vector with all elements being 1, then (3.14) reduces to the (generalized) partially linear model (GPLM) (Härdle *et al.*, 2004). When $\boldsymbol{\gamma} = \mathbf{0}$ (no predictors in $\boldsymbol{X_2}$), (3.14) is simply the semiparametric single-index model (SIM) studied by Hardle *et al.* (1993), which solves the "curse of dimensionality" problem in purely non-parametric regression.

3.7. ESTIMATION OF THE PROPOSED MIXTURE CURE MODEL

This section pertains to the estimation of the proposed model: generalized partially linear single-index additive hazard (GPLSI-AH) model. The (log) likelihood (3.16) in classical survival analysis involves contributions from censored and uncensored groups and is defined below:

$$\ell(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} \left[f\left(Y_i \mid \boldsymbol{Z}_i\right) \right]^{\Delta_i} \times \prod_{i=1}^{n} \left[S\left(Y_i \mid \boldsymbol{Z}_i\right) \right]^{1-\Delta_i}. \tag{3.16}$$

The censored contribute through the survival function, while the uncensored contribute through the density function. Following the definition of the survival function $S(\cdot)$ and the density function $f(\cdot)$ of the mixed population in (3.12), the log likelihood

(3.16) can be rewritten as (3.17):

$$\ell(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} \left[p\left(\boldsymbol{X}_i\right) f_u\left(Y_i \mid \boldsymbol{Z}_i\right) \right]^{\Delta_i} \times \prod_{i=1}^{n} \left[1 - p\left(\boldsymbol{X}_i\right) + p\left(\boldsymbol{X}_i\right) S_u\left(Y_i \mid \boldsymbol{Z}_i\right) \right]^{1-\Delta_i} \quad (3.17)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \lambda_0, g)^\top$ and $f_u\left(Y_i \mid \boldsymbol{Z}_i\right) = \lambda_u\left(Y_i \mid \boldsymbol{Z}_i\right) S_u\left(Y_i \mid \boldsymbol{Z}_i\right)$, which depends on the unobserved uncured status B, and hence the parameters cannot yet be estimated. Note that $\boldsymbol{\beta}, \lambda_0$ are the coefficient vector and baseline hazard defined in the additive hazard model (see section 3.1.2.2). The likelihood in (3.17), assumes an equal contribution from the censored but cured and the censored but uncured, a rare case in the presence of a cure fraction (Amico *et al.*, 2019). Sy and Taylor (2000) used the EM algorithm of Dempster *et al.* (1977) to handle the partially observed B. In the same vein, this work defines the complete-data likelihood (3.18) as follows:

$$\begin{aligned} L_c(\boldsymbol{\theta}) = & \underbrace{\prod_{i=1}^{n} \left\{ p\left(\boldsymbol{X}_i\right) S_u\left(Y_i \mid \boldsymbol{Z}_i\right) \right\}^{B_i(1-\Delta_i)}}_{\text{censored \& uncured}} \underbrace{\prod_{i=1}^{n} \left\{ 1 - p\left(\boldsymbol{X}_i\right) \right\}^{(1-B_i)(1-\Delta_i)}}_{\text{censored \& cured}} \\ & \times \underbrace{\prod_{i=1}^{n} \left\{ p(\boldsymbol{X}_i) f_u\left(Y_i \mid \boldsymbol{Z}_i\right) \right\}^{B_i \Delta_i}}_{\text{uncensored \& uncured}} . \end{aligned} \quad (3.18)$$

In this likelihood, there are contributions from the *censored & uncured*, *censored & cured* and *uncensored & uncured* (See Figure 3.2). Expanding, combining like terms and replacing $f_u\left(Y_i \mid \boldsymbol{Z}_i\right)$ with $\lambda_u\left(Y_i \mid \boldsymbol{Z}_i\right) S_u\left(Y_i \mid \boldsymbol{Z}_i\right)$ in (3.18) gives:

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ p\left(\boldsymbol{X}_i\right)^{B_i} S_u\left(Y_i \mid \boldsymbol{Z}_i\right)^{B_i} \lambda_u\left(Y_i \mid \boldsymbol{Z}_i\right)^{B_i \Delta_i} \right\} \prod_{i=1}^{n} \left\{ 1 - p\left(\boldsymbol{X}_i\right) \right\}^{(1-B_i)} . \quad (3.19)$$

Taking the log gives

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \{B_i \log p\left(\boldsymbol{X}_i\right) + B_i \log S_u\left(Y_i \mid \boldsymbol{Z}_i\right) + B_i \Delta_i \log \lambda_i\left(Y_i \mid \boldsymbol{Z}_i\right) \\
&\quad + (1 - B_i) \log\left(1 - p\left(\boldsymbol{X}_i\right)\right)\} \\
&= \underbrace{\sum_{i=1}^{n} \{B_i \log p\left(\boldsymbol{X}_i\right) + (1 - B_i) \log\left(1 - p\left(\boldsymbol{X}_i\right)\right)\}}_{\text{incidence}} \\
&\quad + \underbrace{\sum_{i=1}^{n} B_i \left\{\Delta_i \log \lambda_u\left(Y_i \mid \boldsymbol{Z}_i\right) + \log S_u\left(Y_i \mid \boldsymbol{Z}_i\right)\right\}}_{\text{latency}} \\
&= \ell_1(\boldsymbol{\theta}) + \ell_2(\boldsymbol{\theta})
\end{aligned} \tag{3.20}$$

The "log-likelihood" contains the partially known cure status B that will be replaced with it's estimate via the conditional expectation of the complete log-likelihood with respect to B given the observed data, $\mathcal{D}$, and current estimates of the parameters $\boldsymbol{\theta}^{m-1}$. The expectation of B gives the same expression since the complete log-likelihood is a linear function of B. Recall that the cure status B, is a Bernoulli random variable as defined in (3.11). Therefore in the E-step of the EM algorithm, $E(B|\mathcal{D}, \boldsymbol{\theta}^{m-1})$ is given by:

$$\begin{aligned}
E(B|\mathcal{D}, \boldsymbol{\theta}^{m-1}) &= 1 \times P(T < \infty|\mathcal{D}, \boldsymbol{\theta}^{m-1}) + 0 \times P(T = \infty|\mathcal{D}, \boldsymbol{\theta}^{m-1}) \\
&= P(T < \infty|\mathcal{D}, \boldsymbol{\theta}^{m-1}) \\
&= (1 - \Delta_i)\, P\left(B_i = 1 \mid Y_i, \Delta_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{m-1}\right) \\
&\quad + \Delta_i P\left(B_i = 1 \mid Y_i, \Delta_i = 1, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{m-1}\right)
\end{aligned} \tag{3.21}$$

where the second equation can the expressed as a sum, in the final equation, using censoring indicator notation. That expression represents the probability of susceptible (uncured) given the parameter estimates at the $(m-1)$th iteration, which may constitute censored or uncensored observations. Further, the second expres-

sion in the last equation is 1. This is due to the fact that if an individual experiences the event, then it is known that they belong to the uncured group so, $P\left(B_i = 1 \mid Y_i, \Delta_i = 1, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{m-1}\right) = 1$.

Therefore (3.21) reduces to:

$$E(B|\mathcal{D}, \boldsymbol{\theta}^{m-1}) = (1 - \Delta_i)\, P\left(B_i = 1 \mid Y_i, \Delta_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{m-1}\right) + \Delta_i. \tag{3.22}$$

The probability expression in equation (3.22) can be expressed using the conditional probability rule as:

$$\begin{aligned} P\left(B_i = 1 \mid Y_i, \Delta_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{m-1}\right) &= \frac{P\left(B_i = 1, Y_i, \Delta_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{m-1}\right)}{P\left(Y_i, \Delta_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{m-1}\right)} \\ &= \frac{p^{(m-1)}(\boldsymbol{X}_i) S_u^{(m-1)}\left(Y_i \mid \boldsymbol{Z}_i\right)}{1 - p^{(m-1)}(\boldsymbol{X}_i) + p^{(m-1)}(\boldsymbol{X}_i) S_u^{(m-1)}\left(Y_i \mid \boldsymbol{Z}_i\right)}. \end{aligned} \tag{3.23}$$

Substituting the result obtained from (3.23) into (3.22) it can be seen that:

$$\begin{aligned} E(B|\mathcal{D}, \boldsymbol{\theta}^{m-1}) &= \Delta_i + (1 - \Delta_i) \left[\frac{p^{(m-1)}(\boldsymbol{X}_i) S_u^{(m-1)}\left(Y_i \mid \boldsymbol{Z}_i\right)}{1 - p^{(m-1)}(\boldsymbol{X}_i) + p^{(m-1)}(\boldsymbol{X}_i) S_u^{(m-1)}\left(Y_i \mid \boldsymbol{Z}_i\right)} \right] \\ &:= W_i^m. \end{aligned} \tag{3.24}$$

The log likelihood (3.20) suggests that the optimization problem can be carried out separately for the incidence ($\ell_1(\boldsymbol{\theta})$) and the latency ($\ell_2(\boldsymbol{\theta})$) parts of the mixture cure model. For the incidence part, notice that the likelihood, (3.25) has the form of the likelihood of a GLM:

$$\ell_1(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ B_i \log p\left(\boldsymbol{X}_i\right) + (1 - B_i) \log\left(1 - p\left(\boldsymbol{X}_i\right)\right) \right\}. \tag{3.25}$$

For the proposed GPLSIM, $p(\boldsymbol{X}) = H\left(g\left(\boldsymbol{\alpha}^\top \boldsymbol{X_1}\right) + \boldsymbol{\gamma}^\top \boldsymbol{X_2}\right)$. Substituting into (3.25) gives

$$\begin{aligned}\ell_1(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big\{ B_i \log H\left(g\left(\boldsymbol{\alpha}^\top \boldsymbol{X_{1i}}\right) + \boldsymbol{\gamma}^\top \boldsymbol{X_{2i}}\right) \\ + (1 - B_i) \log\left(1 - H(g\left(\boldsymbol{\alpha}^\top \boldsymbol{X_{1i}}\right) + \boldsymbol{\gamma}^\top \boldsymbol{X_{2i}})\right)\Big\}.\end{aligned} \tag{3.26}$$

A penalized splines (P-splines) (Eilers and Marx, 1996) approach that follows the work of Yu *et al.* (2017) is utilized to model the unknown univariate function $g(\cdot)$ and it is estimated by a linear combination of truncated power spline bases:

$$\begin{aligned} g(u) &= \varphi_0 + \varphi_1 u + \cdots + \varphi_p u^p + \sum_{k=1}^{K} \varphi_{p+k} \left(u - v_k\right)_+^p \\ &= \boldsymbol{\varphi}^\top \mathbf{S}(u) \end{aligned} \tag{3.27}$$

where $\mathbf{S}(u) = \left\{1, u, \ldots, u^p, (u - v_1)_+^p, \ldots, (u - v_K)_+^p\right\}$ are spline bases with K knots placed at $(v_1, \ldots, v_K)$, and $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \ldots, \varphi_{p+K})^\top$ are spline coefficients to be estimated. There are several options for a basis in the `mgcv` R package (Wood, 2012). Other spline basis such as the thin plate regression spline (Wood, 2003) or cubic regression spline can be employed especially for smaller sample sizes, as they tend to give the best mean squared error performance at the expense of longer computational time. As pointed out by Ruppert *et al.* (2003), for P-splines, the spline basis used and the choice of knots are less important than the smoothing parameter. Hence the default knots placements suggested in the `mgcv` R package, at the sample quantiles of the predictors, is used. Increasing the number of knots (usually fewer than the number of observations) can accurately approximate flexible functions very fast. In such situations, a smoothing parameter can be added to control roughness and pre-

vent overfitting. In this work, the maximum likelihood criterion of Anderssen and Bloomfield (1974) is used in estimating the smoothing parameter, as it produced good results.

To elaborate on the estimation of the GPLSIM, notice that B, the cure indicator acts as a response variable in the "GLM-like" likelihood. $g, \boldsymbol{\alpha}, \boldsymbol{\gamma}$ need to be estimated through an iterative algorithm. Using the truncated power bases splines in (3.27), (3.26) becomes:

$$\begin{aligned}\ell_1(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\varphi}) = \sum_{i=1}^{n} \Big\{ & B_i \log H \left\{ \boldsymbol{\varphi}^\top \mathbf{S} \left(\boldsymbol{\alpha}^\top \boldsymbol{X}_{1i} \right) + \boldsymbol{\gamma}^\top \boldsymbol{X}_{2i} \right\} \\ & + (1 - B_i) \log \left(1 - H \left\{ \boldsymbol{\varphi}^\top \mathbf{S} \left(\boldsymbol{\alpha}^\top \boldsymbol{X}_{1i} \right) + \boldsymbol{\gamma}^\top \boldsymbol{X}_{2i} \right\} \right) \Big\}.\end{aligned} \tag{3.28}$$

In an EM-like algorithm, the values $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\varphi}})$ are obtained that maximize the likelihood

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\varphi}}) = \underset{\substack{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\varphi} \\ \|\boldsymbol{\alpha}\| = 1}}{\arg\max}\, \ell_1(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\varphi}). \tag{3.29}$$

The second part of the likelihood, $\ell_2(\boldsymbol{\theta})$ is the likelihood of the additive hazard model (3.30) with an extra weighting term B_i.

$$\ell_2(\boldsymbol{\theta}) = \sum_{i=1}^{n} B_i \left\{ \Delta_i \log \lambda_u \left(Y_i \mid \boldsymbol{Z}_i \right) + \log S_u \left(Y_i \mid \boldsymbol{Z}_i \right) \right\}. \tag{3.30}$$

For estimation purposes we proceed in the manner of Lin and Ying (1994). B_i acts as a weight in this estimation method. It is estimated via the EM-like algorithm and denotes the probability that the individual is cured or not. The additive hazard model of Lin and Ying (1994) can be written in counting process form. For n independent subjects where the counting process $N_i(t)$ counts the number of events up to time t, the intensity function for the model $\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}$ is given by:

$$R_i(t) d\Lambda(t \mid \boldsymbol{Z}) = R_i(t) \{ d\Lambda_0(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}(t) dt \} \tag{3.31}$$

where $R_i(t)$ is the at-risk process (1 if the individual is at risk at time $t, 0$ otherwise). Multiplying both sides by the weighting term, B_i, and using the definition of a martingale: $dM(t) = dN(t) - R(t)d\Lambda(t)$ (where $R(t)d\Lambda(t)$ is the cumulative intensity process), it can be seen that:

$$\begin{aligned} dM_i(t) &= B_i dN_i(t) - B_i R_i(t) d\Lambda(t \mid \boldsymbol{Z}) \\ &= B_i dN_i(t) - B_i R_i(t)\{d\Lambda_0(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}(t)dt\}. \end{aligned} \tag{3.32}$$

Next, the martingale property is applied. This property states that $E[dM(t) \mid \mathcal{F}(t)] = 0$, where $\mathcal{F}(t)$ is the filtration up to time t. For all individuals, it can be seen that the estimating equation $\frac{1}{n}\sum_{i=1}^{n} dM_i(t) = 0$. That is,

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} dM_i(t) \\ &= \sum_{i=1}^{n} B_i \left[dN_i(t) - R_i(t)\left\{d\Lambda_0(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)dt\right\}\right]. \end{aligned} \tag{3.33}$$

For fixed $\boldsymbol{\beta}$ and assuming B is known, the solution to (3.33) is

$$\hat{\Lambda}_0(t \mid \boldsymbol{\beta}) = \int_0^t \frac{\sum_{i=1}^{n} B_i \left\{dN_i(u) - R_i(u)\boldsymbol{\beta}^\top \boldsymbol{Z}_i(u)du\right\}}{\sum_{i=1}^{n} B_i R_i(u)}. \tag{3.34}$$

Using the definition of a martingale and multiplying (3.32) by $R_i(t)Z_i(t)$ so that the left-hand side of this equation represents the expected value of the covariates at the event times, it can be seen that:

$$\begin{aligned} U(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \int_0^\infty R_i(t)\boldsymbol{Z}_i(t)dM_i(t) \\ &= \sum_{i=1}^{n} \int_0^\infty B_i \left[R_i(t)\boldsymbol{Z}_i(t)dN_i(t) - R_i(t)\boldsymbol{Z}_i(t)\left\{d\Lambda_0(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)dt\right\}\right]. \end{aligned} \tag{3.35}$$

Plugging in the estimate of baseline hazard and simplifying, (3.35) becomes:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} B_i \boldsymbol{Z}_i(t) \left\{ dN_i(t) - R_i(t) d\hat{\Lambda}_0(t \mid \boldsymbol{\beta}) - R_i(t) \boldsymbol{\beta}^{\top} \boldsymbol{Z}_i(t) dt \right\}. \tag{3.36}$$

Again, the resulting estimating function is that obtained in Lin and Ying (1994) with the weighting term, B_i. Setting (3.36) equal to the $q \times 1$ vector $\mathbf{0}$ produces the estimating equation for $\boldsymbol{\beta}$ whose solution is $\hat{\boldsymbol{\beta}} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{D}}$ with :

$$\hat{\mathbf{A}} = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} B_i R_i(t) \left\{ \boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t) \right\}^{\otimes 2} dt \tag{3.37}$$

and

$$\hat{\mathbf{D}} = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\infty} B_i \left\{ \boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t) \right\} dN_i(t) \tag{3.38}$$

where $\bar{\boldsymbol{Z}}(t) = \sum_{i=1}^{n} B_i R_i(t) \boldsymbol{Z}_i(t) / \sum_{i=1}^{n} B_i R_i(t)$. Consequently a natural estimator for the survival function of the uncured sub-population is

$$\hat{S}_u(t \mid \boldsymbol{z}) = \exp\left\{ -\hat{\Lambda}_0(t \mid \hat{\boldsymbol{\beta}}) - \int_0^t \hat{\boldsymbol{\beta}}^{\top} \boldsymbol{z}(u) du \right\}. \tag{3.39}$$

PAPER

I. DIFFERENTIAL METHYLATED REGION DETECTION VIA AN ARRAY-ADAPTIVE NORMALIZED KERNEL-WEIGHTED MODEL

Daniel Ahmed Alhassan
Department of Mathematics & Statistics
Missouri University of Science and Technology
Rolla, Missouri 65409–0050

ABSTRACT

A differentially methylated region (DMR) is a genomic region that has significantly different methylation patterns between biological conditions. Identifying DMRs between different biological conditions is critical for developing disease biomarkers. Although methods for detecting DMRs in microarray data have been introduced, developing methods with high precision, recall, and accuracy in determining the true length of DMRs remains a challenge. In this study, we propose a normalized kernel-weighted model to account for similar methylation profiles using the relative probe distance from "nearby" CpG sites. We also extend this model by proposing an array-adaptive version in attempt to account for the differences in probe spacing between Illumina's Infinium 450K and EPIC bead array respectively. We also study the asymptotic results of our proposed statistic. We compare our approach with a popular DMR detection method via simulation studies under large and small treatment effect settings. We also discuss the susceptibility of our method in detecting the true length of the DMRs under these two settings. Lastly, we demonstrate the biological usefulness of our method when combined with pathway analysis methods

on oral cancer data. We have created an R package called *idDMR*, downloadable from GitHub repository with link: https://github.com/DanielAlhassan/idDMR, that allows for the convenient implementation of our array-adaptive DMR method.

Keywords: Differentially methylated regions, DNA Methylation, Illumina 450K and EPIC, Kernel smoothing

1. INTRODUCTION

DNA methylation is an important epigenetic mechanism used by cells to control gene expression (Eden and Cedar, 1994). It plays an important role in many biological processes such as somatic cell (Greenberg and Bourc'his, 2019) and embryonic development (Breton-Larrivée *et al.*, 2019). Aberrant DNA methylation patterns have been linked to complex diseases like cancer and diabetes (Maghbooli *et al.*, 2014). DNA methylation refers to the addition of a methyl group to a DNA base. In mammals, it is known to occur at cytosine sites when followed by a guanine nucleotide (called a CpG site) (Ehrlich and Wang, 1981). Whole-genome bisulfite sequencing (WGBS) is the gold standard for measuring methylation status in any organism. It can capture more than 28 million CpGs (Shu *et al.*, 2020), providing genome-wide coverage, whereas microarrays focus on a subset of the genome, targeting specific genomic regions. Despite the massive reduction in cost of WGBS in recent years, it is still expensive when employed in large-scale epidemiological studies. Microarrays have become more popular for most epigenome-wide association studies (EWAS). They provide an economically feasible (Laird, 2010) means to explore the associations between DNA methylation and complex diseases (Szyf, 2012). These studies aim to better understand the connection between DNA methylation and human health through identifying markers associated with diseases.

Illumina Infinium HumanMethylation BeadChip technology is the most widely used array-based technology for EWAS studies. The Illumina platform estimates the methylation status using two probes (methylated and unmethylated) at each CpG site to measure the methylation intensities (Weisenberger *et al.*, 2008). There are two ways of quantifying methylation output from the Illumina BeadChip assay: (1) the β-value and (2) the M-value. The β-value measures the percentage of methylation and hence ranges from 0 (unmethylated) to 1 (fully methylated). The M-value is calculated from the β-value using the following relationship:

$$M = \log_2 \left(\frac{\beta}{1 - \beta} \right). \tag{1}$$

Though the β-value is biologically preferred when it comes to interpretation, the M-value is statistically more appropriate. Most classical statistical methods, like the general linear model, used in analyzing high-throughput experiments assume equal variances of populations and normality of errors. The M-values approximately have equal variance and have the same support as the Gaussian distribution. Thus, a statistic based on M-values is more appropriate when using such methods (Du *et al.*, 2010).

Over the years, Illumina has been improving their DNA methylation assay by increasing the number of CpG sites that can be interrogated, starting with the Infinium HumanMethylation 27K (∼27,000 CpG sites), Infinium HumanMethylation 450K (∼480,000 CpG sites) to the most recent Infinium HumanMethylationEPIC (∼850,000 CpG sites). Despite the Infinium HumanMethylationEPIC (herein termed "EPIC") being the most recent, the Infinium HumanMethylation 450K (herein termed "450K") is still used, perhaps due to the similarity in analyzing data collected from both assays. Additionally, many large-scale projects such as the cancer genome atlas (TCGA) utilized 450K arrays, resulting in a wealth of this type of data.

As the technology evolves, future arrays may have different probe gap distributions and a method the readily adapts to the type of array is needed. In this article, we will focus on 450K array data but also provide a method that adapts to the EPIC and possibly future Infinium assays. An important characteristic of the two arrays (450K and EPIC) is that the methylation intensities are measured using either the Infinium I assay or Infinium II assay which have different chemistries. For a detailed description of these two assays, see (Ill, 2015) and references therein. Owing to the different chemistries but complimentary strengths of the two designs, data preprocessing and normalization is critical (Wang *et al.*, 2018). Several normalization methods (Fortin *et al.*, 2014; Maksimovic *et al.*, 2012; Triche *et al.*, 2013) exist in the literature and though no single one always outperforms the other, some methods are built ideally for some specific cases. For instance, the functional normalization method (Fortin *et al.*, 2014) is best suited for cases where global differences are expected, such as in treatment-control studies. We employed this normalization technique in our simulation and data application example.

Differences in DNA methylation between samples (e.g. cancer and normal) can be measured at single CpG sites, referred to as differentially methylated loci (position) (DML/DMP), and over contiguous sites, referred to as differentially methylation regions (DMRs). Despite the many situations where researchers are interested in site-level testing (Weaver *et al.*, 2004), a more useful form is one that involves testing over a region due to the increase in statistical power and biological interpretation (Chen *et al.*, 2016; Robinson *et al.*, 2014). Methylation status between nearby CpG sites are highly correlated (co-methylated) (Eckhardt *et al.*, 2006; Zhang *et al.*, 2015) and this information is employed when collapsing contiguous sites to form DMRs. Regions could be either predefined or user-defined. Differential DNA methylation in predefined regions based on genomic annotations such as CpG Islands (CGI), TSS200 (regions from transcription start site to 200 bases upstream), TSS1500 (200–1500 bases up-

stream of the TSS), Open sea and CGI shores have special biological interpretations (Wang *et al.*, 2012; Warden *et al.*, 2013b; Wu *et al.*, 2013; Zhang *et al.*, 2011). However, they possess two problems: (1) they comprise only a subset of the 450K/EPIC probes for DMR detection which provide less room for knowledge discovery and (2) they require defining genomic regions prior to evaluation, forcing DMRs to have artificial start and end points. A user-defined region, however, allows for flexibility and knowledge discovery as regions can be defined based on some criteria such as median distance between probes. The method we propose falls within this group.

Many user-defined DMR detection methods such as *Probe Lasso*, *Bump Hunter*, *DMRcate* among others, have been proposed (Butcher and Beck, 2015; Jaffe *et al.*, 2012; Peters *et al.*, 2015; Sofer *et al.*, 2013; Zhang *et al.*, 2018). However, no one approach always outperforms the other. The *Probe Lasso* method (Butcher and Beck, 2015) capitalizes on the uneven spacing of probes based on genomic annotation on the array. It calls DMRs based on the probe density so that subsequent analysis do not entirely focus on the dense regions alone. Though the DMR calling framework is purported to be dynamic or flexible, it still forces DMRs to have artificial start and end points within a gene feature. It is a user-defined region method between gene features but a predefined region method within gene features, hence it may fail to detect other novel DMRs when they do exist (lack statistical power). *Bump Hunter* (Jaffe *et al.*, 2012), employs surrogate variable analysis to handle batch effects, a unique feature in their method, as samples are usually not collected at the same time point. However, in an extensive simulation study under 60 different parameter settings (Mallik *et al.*, 2019) comparing the popular methods (*Probe Lasso* and *DMRcate* included), it was revealed that *Bump Hunter* was slow, lacked power (under large and small effect size) and ranked last in terms of precision. *DMRcate* (Peters *et al.*, 2015) is a novel method that only uses the spatial distribution of probes to call DMRs and given a window, borrowing information from nearby CpG sites using a Gaussian

kernel. *DMRcate* has gained much popularity in the literature due to its particularly superior predictive performance compared to *Probe Lasso* and *Bump Hunter* (Mallik *et al.*, 2019). Despite this success, there is still a higher tendency to incur bias due to irregularly spaced CpG sites and further lack the ability to detect all true DMRs that may exist. Given a specific window, highly dense regions are more likely to be detected as all nearby CpG sites will each receive weights that are very close to one. However, less dense regions, may receive no weights at all. This bias towards denser regions leads to the high detection of DMRs in those regions but also a low sensitivity in detecting true DMRs that may exist in less dense regions.

Furthermore, the high weights and subsequently smoother estimates obtained in the high dense regions is not entirely realistic as it does not account for the contribution of each nearby CpG site in obtaining the smoothed estimate. That is, if three adjacent sites A, B, and C are considered neighbors because they are within some genomic distance of each other, then we must consider the relative contribution from these neighbors rather than the raw contribution when attempting to smooth site-level statistics at A. Accounting for the relative contribution could reduce the bias in detecting DMRs to a level sufficient to detect true DMRs in the high dense regions while also improving statistical power to detect DMRs in less dense regions.

The aforementioned DMR detection methods, all reference Eckhardt *et al.* (2006), which first mentioned a strong correlation between methylation levels and CpG distance within a 1000 base pairs (bp). More recently, Sun and Sun (2019) found that co-methylation (similar methylation profiles) within normal tissues were as short as a few hundred base pairs. Another recent preliminary analysis of methylation patterns between consecutive CpG sites in a breast cancer study by Sun *et al.* (2019) revealed that co-methylation region lengths differed significantly for unmethylated

and methylated states. A thorough analysis of co-methylation patterns using breast cancer data was done by (Sun *et al.*, 2022). One notable finding was that the co-methylation patterns on chromosome X were different from the other chromosomes.

The low statistical power due to the bias and lack of flexibility in the aforementioned methods coupled with these recent findings on co-methylation suggest the need for a much more flexible and less-biased DMR detection method. To this end, the goals of this manuscript are three-fold: (1) we propose a general locally-weighted statistic via which co-methylation information can be incorporated to site-level statistics for DMR detection, (2) we show some large sample properties of statistics of this form under some regularity conditions and (3) we develop a new method for DMR detection (a specific case of the general locally-weighted statistic), which reduces the bias due to irregular spaced CpG sites and increases the sensitivity in detecting true DMRs. Furthermore, we formulate an array adaptive version of the method (*aaDMR*) to better capture the somewhat complex co-methylation among CpG sites, and that adapts to the spacing on each chromosome for 450K, EPIC and possibly future arrays.

The rest of the paper is organized as follows. In the methods section we introduce the general locally-weighted statistic and briefly state some asymptotic results for it. We then introduce the normalized kernel statistic as a specific case of this general statistic and use it in DMR detection. In the results section we perform a simulation study to compare the performance of our proposed DMR detection methods with *DMRcate*, apply our method to an oral cancer dataset and briefly study the biological relevance of our methods through a pathway analysis. The conclusion section contains the summary of our findings and highlights potential areas of further research.

2. METHODS

2.1. SITE-LEVEL DIFFERENTIAL METHYLATION TESTING WITH LIMMA

Limma (Smyth, 2004) stands for "linear models for microarray data" and it is the most widely used method for microarray analysis. It involves fitting linear models and is commonly used in the analysis of DNA methylation data. We use limma for our work to obtain differential methylation signals at individual CpG sites. To this end, consider testing $H_0 : \mu_T - \mu_N = 0$ against $H_1 : \mu_T - \mu_N \neq 0$ where μ_T, μ_N are average methylation M-values from tumor and normal samples respectively at a CpG site obtained from the respective true percent of methylation β_T and β_N based on (1). We fit a linear model with the M-values as the response, the condition (tumor or normal) as the predictor, along with any covariates of interest. The specific contrast to test the above hypotheses is conducted and the empirical Bayes techniques is then implemented to obtain robust t estimates called moderated t-statistics. The DMR detection methods we discuss below rely on the robust site-level tests from limma.

2.2. A GENERAL LOCALLY-WEIGHTED STATISTIC

DNA methylation levels are correlated, at least for nearby CpG sites (Eckhardt *et al.*, 2006; Zhang *et al.*, 2015). Hence in determining the DMRs, we propose smoothing limma's moderated t-statistic as a suitable way to capture the correlation information among nearby CpG sites. To this end, we propose the general locally-weighted statistic (2), motivated by *DMRcate*'s kernel-weighted statistic. For a chromosome, define the locally weighted statistic $S(x_i)$ as:

$$S(x_i) = Y_i + \sum_{j \neq i}^{n} w_j(x_i) Y_j \tag{2}$$

where x_i denotes the position of CpG site i (site of interest where smoothing is happening) and j (neighboring sites) respectively, n is the number of sites within some specified genomic distance, $Y_i($ or $Y_j)$ denotes some function of a statistic from a site-level testing (such as the moderated t-statistic from limma (Smyth, 2004)), and $w_j(x_i)$, some appropriate weighting function or mechanism, used to account for the interdependencies between nearby CpG sites. For the purposes of our work, we define $Y = T^2$ where T is a random variable representing limma's moderated t-statistic (Smyth, 2004). We state in passing, that when $w_j(x_i) = K_{ij}$, defined as Gaussian kernel weights as in Peters *et al.* (2015), we obtain *DMRcate's* kernel-weighted statistic $S_{KY(i)}$. The EPIC array has ∼400,000 CpG sites more than 450K array and so within some specified genomic distance, EPIC is likely to contain more sites than the 450K. With this in mind, we investigate the asymptotic behavior of $S(x_i)$ when the number of nearby CpG sites increases within some specified genomic distance of x_i (or as the array technology improves).

2.3. ASYMPTOTIC RESULTS

This subsection pertains to the asymptotic results of the proposed statistic $S(x_i)$. More specifically, we state a situation under which $S(x_i)$ is consistent and obeys a modified central limit theorem. A generalized version of the consistency result is due to Adler and Rosalsky (1991).

Theorem 1 (Consistency)**.** *Let* $\{Y_j,\ j > 1\}$ *be independent F-distributed random variables obtained from site-level testing via limma's moderated F-statistic (*t^2*-statistic) and* $\{w_j,\ j > 1\}$ *be constants satisfying* $\sum_{j=1}^{n} w_j^2 = \mathcal{O}(nw_n^2)$*. Further, at CpG site*

x_i, let $S(x_i) = \sum_{j=1}^{n} w_j(x_i) Y_j$ where

$$w_j(x_i) = \begin{cases} 1, & j = i \\ t \in (0,1), & j \neq i \end{cases}.$$

Then,

$$\frac{\sum_{j=1}^{n} w_j \left(Y_j - EY \mathbf{1}_{\{|Y| \leq n^2\}} \right)}{w_n n^2} \xrightarrow{p} 0.$$

Theorem 2. *Let $\{Y_j,\ j > 1\}$ be independent F-distributed random variables obtained from site-level testing via limma's moderated F-statistic (t^2-statistic) and let $\{w_j,\ j > 1\}$ be constants satisfying $0 < w_j < 1$ with $\sum_{j=1}^{n} w_j = 1$. At CpG site x_i for observed y_i, define $S(x_i) = y_i + \sum_{j \neq i}^{n} w_j Y_j$. Then*

$$y_i + \frac{\sum_{j \neq i}^{n} w_j \left(Y_j - E(Y) \right)}{\sqrt{\sum_{j \neq i}^{n} w_j^2 \sigma_j^2}} \xrightarrow{d} y_i + Z \quad as \quad n \to \infty \tag{3}$$

where $Var\left(\sum_{j \neq i}^{n} w_j Y_j \right) = \sum_{j \neq i}^{n} w_j^2 \sigma_j^2$ and $Z \sim N(0,1)$.

Theorems 1 and 2 apply to a specific case of (2). More specifically the theorems hold for DMR testing case where Y_j is the moderated F-statistic (Smyth, 2004). The proofs are provided in the supplementary text (Appendix: Supplementary File 1).

2.4. NORMALIZED KERNEL-WEIGHTED STATISTIC

We propose a specific weighting function $w_j(x_i)$ (simply w_j) called the normalized kernel-weight function (4) as a realistic way of incorporating the interdependencies among nearby CpG sites. Kernel smoothers allow for a flexible degree of smoothing via its smoothing parameter, known as the *bandwidth*, which is an effective way of incorporating the shared methylation profiles at nearby CpG sites. Our

rationale for the normalized kernel (NK) is two fold. First, if $A < B < C$, where A, B, C are CpG sites, then within some "reasonable genomic distance" (determined by the bandwidth), if A is a neighbor to B, and B a neighbor to C, then A and C are neighbors as well. So in smoothing or weighting the statistic at A for instance, one needs to consider the relative contribution from the nearby CpG sites since there is shared co-methylation among the neighbors. Secondly, the NK reduces the bias towards dense regions in the DMR detection process through a "fair" distribution of the weights thereby increasing the sensitivity in detecting DMRs in less dense regions when they do exist. With the caveat that the total of all contributed weights must equal one, our method maintains the property that sites closer to the reference site contribute more weight than those further away. One major difference between our proposed method and DMRcate lies in the rationale behind the statistics proposed. We hypothesize that within some genomic distance all neighboring sites contain the totality of information to smooth a site-level statistic. However, *DMRcate* advocates for using raw contributions from neighboring sites in smoothing a site-level statistic. As previously mentioned, DMRcate procedure is biased to CpG dense regions due to their statistic (Mallik *et al.*, 2019). The major advantage in our approach is the increase sensitivity or power in picking up DMRs that do exist in less dense regions while maintaining our precision in picking up DMRs in CpG-dense regions. For convenience, we employ the Gaussian kernel, $K(z) = \exp\left(-\frac{z^2}{2}\right)$, and define the weighting function as:

$$w_j(x_i) = \frac{K\left(\frac{x_j - x_i}{h}\right)}{\sum_{j \neq i}^{n} K\left(\frac{x_j - x_i}{h}\right)}, \tag{4}$$

where $K(\cdot)$ is the kernel, h is the bandwidth and x_i, x_j denote the position of CpG sites i and j respectively. More specifically, we address the complex co-methylation patterns in the manner stated below.

(i) First, we employ the normalized kernel-weight (4) to address the interdependencies in the methylation levels at nearby sites. In order to readily capture the benefit gained by using our weighting measure, we keep a fixed bandwidth/ kernel size, h of 500bp as do Peters *et al.* (2015). We call this approach the fixed-spacing array DMR (*faDMR*) detection method.

(ii) Our first approach essentially assumes equal spacing between the probes on each chromosome which is far from truth. See Figures 1 and 2 for the distribution of probe gaps on the 450K and EPIC respectively. In addition to the uneven spacing at the chromosomal level, co-methylation patterns are different for different chromosomes (Sun *et al.*, 2022), suggesting that using a different kernel size, h, for each chromosome may prove useful, since h acts as a measure of spread. To that end, we propose an array-adaptive DMR (*aaDMR*) detection method, one that chooses h to equal the median probe spacing on each chromosome.

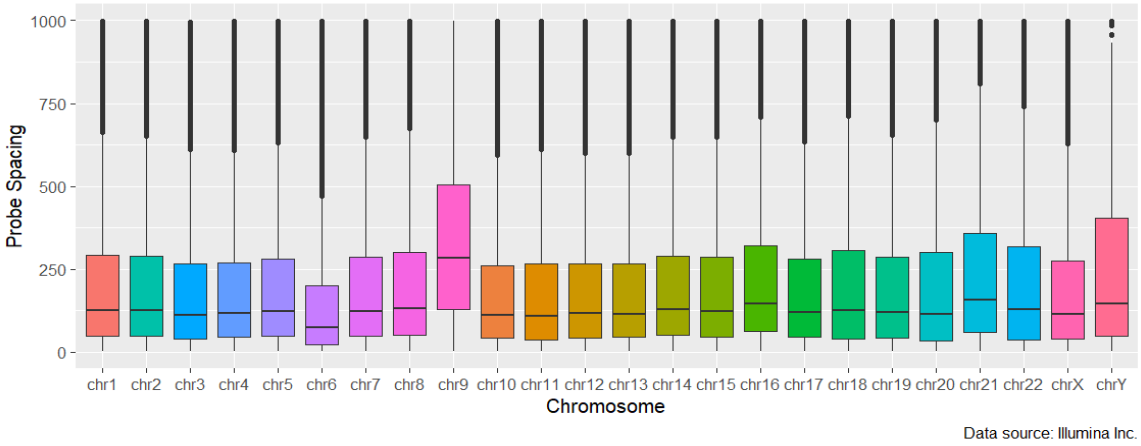

Figure 1. Probe Spacing distribution on the 450K array truncated at 1000bp to ease visualization.

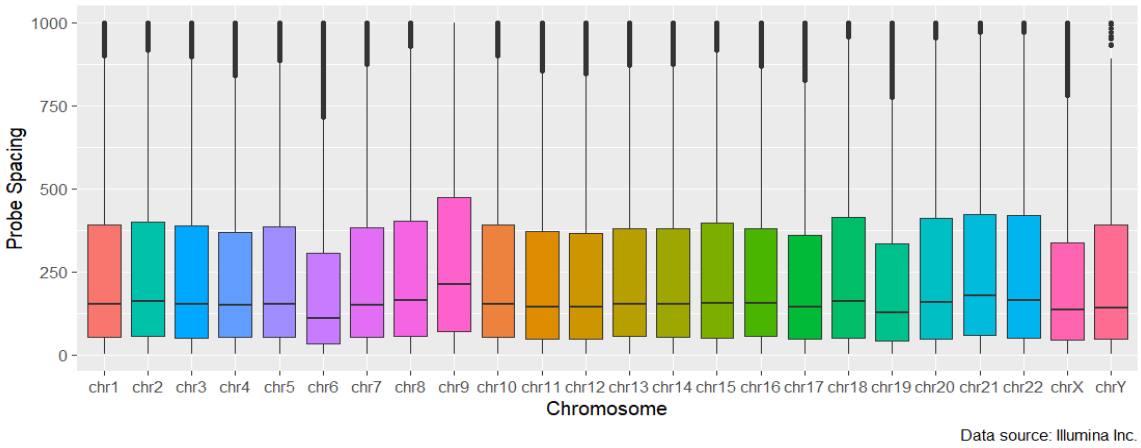

Figure 2. Probe Spacing distribution on the EPIC array truncated at 1000bp to ease visualization.

2.5. MODELING $S(x_i)$ VIA SATTERTHWAITE'S APPROXIMATION

We describe the process of modeling our normalized kernel-weighted statistic and mention that the process is similar to that used by *DMRcate* (Peters *et al.*, 2015). Let $x_1 < x_2, \cdots, < x_n$, be CpG sites for an individual chromosome. Under the appropriate null hypothesis, $T \sim t_\nu$ where ν is the degrees of freedom (after empirical Bayes' adjustment) so that $Y \sim F_{(1,\nu)}$. It is obvious that (2) is a weighted linear combination of F-distributed random variables which is mathematically complex to model (Peters *et al.*, 2015). Owing to the empirical Bayes method used by *limma* (Smyth, 2004), ν is relatively large even for small sample size situations. We capitalize on this so that as $\nu \to \infty$, $Y \xrightarrow{d} \chi^2_1$.

We can now view (2) as a linear combination of scaled χ^2_1 random variables. Assuming that $S(x_i)$ is approximately distributed as a scaled chi-squared random variable of the form $p_{x_i}\chi^2_{q_{x_i}}$, we utilize the rule suggested by Satterthwaite (Satterthwaite, 1946) by matching the first two central moments to obtain p_{x_i} and q_{x_i} in (6). The first two central moments of $S(x_i)$ are $E(S(x_i)) = E\left(Y_i + \sum_{j=1}^n w_j Y_j\right) = 1 + \sum_{j=1}^n w_j$ and $Var(S(x_i)) = Var\left(Y_i + \sum_{j=1}^n w_j Y_j\right) = 2\left(1 + \sum_{j=1}^n w_j^2\right)$ respectively. We adopt the

notation p_{x_i} and q_{x_i} to make clear that the constants are obtained on a CpG-site-level and hence may differ from one site to another. The mean and variance of $p_{x_i}\chi^2_{q_{x_i}}$ are given by $p_{x_i}q_{x_i}$ and $2p^2_{x_i}q_{x_i}$ respectively.

By matching the first two central moments to the mean and variance stated, we have:

$$\begin{cases} p_{x_i}q_{x_i} & = 1 + \sum_{j=1}^{n} w_j \\ \\ 2p^2_{x_i}q_{x_i} & = 2\left(1 + \sum_{j=1}^{n} w_j^2\right). \end{cases} \tag{5}$$

Solving (5) leads to

$$\begin{cases} p_{x_i} & = \dfrac{1 + \sum_{j=1}^{n} w_j^2}{1 + \sum_{j=1}^{n} w_j} \\ \\ q_{x_i} & = \dfrac{(1 + \sum_{j=1}^{n} w_j)^2}{1 + \sum_{j=1}^{n} w_j^2}. \end{cases} \tag{6}$$

Now, since $S(x_i) \sim p_{x_i}\chi^2_{q_{x_i}}$ then, $\dfrac{S(x_i)}{p_{x_i}} \sim \chi^2_{q_{x_i}}$. We compare the observed values of $\dfrac{S(x_i)}{p_{x_i}}$ to a χ^2 distribution with q_{x_i} degrees of freedom to obtain p-values for our local estimator. Next, we apply the Benjamini-Hochberg (BH) correction (Benjamini and Hochberg, 1995a) to control the false discovery rate (FDR) across all site-level tests. CpG sites with a BH-corrected p-value less than the significance level of $\alpha = 0.05$ are retained. As a last step, we group significant CpG sites (retained from the previous procedure) that are within g genomic distance from each other to form DMRs. In collapsing these contiguous sites into regions, we defaulted to $g = 1000$

bp as do DMRcate (Peters *et al.*, 2015) and Bumphunter (Jaffe *et al.*, 2012). To quantify statistical uncertainty to the identified DMRs, we take the CpG site with the minimum p-value as a representative p-value for that region.

2.6. STEP-BY-STEP SUMMARY OF THE FADMR & AADMR DETECTION APPROACH

(a) Obtain site-level moderated observed t-statistics, t_j, using limma (Smyth, 2004).

(b) Calculate observed $y_j = t_j^2$ for CpG site j.

(c) Use the normalized kernel weight (4) to weight observed y_j to obtain smoothed locally weighted statistic $S(x_i)$ (2). For the faDMR method, h is set to 500bp. For the aaDMR method, h is taken to equal the median probe spacing on each chromosome.

(d) Use Satterthwaite's approximation to model $S(x_i)$ to obtain unadjusted p-values.

(e) Apply BH correction to obtain adjusted p-values.

(f) Filter out any CpG site with an adjusted p-value less than $\alpha = 0.05$.

(g) Specify g, the agglomerate parameter, and collapse significant sites from (f) that are within g base pairs of each other to form DMRs.

(h) Take the minimum p-value as a representative p-value across CpG sites in the region. This p-value is used to order the regions in terms of the strength of significance.

3. RESULTS

3.1. SIMULATION STUDY

To validate our method and investigate its performance, we perform a simulation study and compare our method with *DMRcate*. We simulated 1000 repetitions of a 450K dataset, each with 10 control and 10 treatment samples yielding a 450K $\times$ 20 matrix. In each set of simulated data, we randomly assigned 2,136 promoter regions comprising TSS200 and TSS1500 as true DMRs out of 21,363 regions. Half of these were hypermethylated (i.e., methylation higher in treatment than control) and the other half hypomethylated (i.e., methylation lower in treatment than control). For DMRs, $\beta-$values were simulated from a beta distribution with mode equal to some specified beta level. For non-DMRs, probes were classified as completely methylated or unmethylated based on array data from The Cancer Genomic Atlas (TCGA) on cholangiocarcinoma (CHOL) (Center for Cancer Genomics - National Cancer Institute, n.d.). For non-DMRs, $\beta-$values were simulated from a beta distribution with parameters manually chosen to match the average modes of two methylation statuses of CHOL data. A detailed description of the simulation study can be found in the supplementary text (Appendix: Supplementary File 1). We investigated two different treatment effects (large and small). For the large effect, we set the true methylation difference ($\Delta\beta = 0.2$) to be exactly 0.2 (as in *DMRcate* (Peters *et al.*, 2015)). However, (Mallik *et al.*, 2019) reported that the common DMR testing methods such as *DMRcate*, lacked power to detect small effect sizes. Consequently, we compared our methods to *DMRcate* with a true methylation difference of 0.09 (small effect) while maintaining other aspects of the simulation unchanged. All analyses were based on the M-values (see (1)). Our simulation study was set up in a similar way to *DMRcate*. In addition, we obtained a histogram for one of the 1000 datasets to investigate how well the parameter space is explored (see Figure 3).

3.2. EVALUATION CRITERIA

We compare our methods (*faDMR* and *aaDMR*) with *DMRcate* under its best (default) performance setting (see Peters *et al.* (2015)) to assess their performance using three criteria (precision, recall and F1-score). As previously stated, we compare *faDMR* to *DMRcate* to readily capture the benefit gained from using the normalized kernel. Next we compare *faDMR* and *aaDMR* to capture the advantage of using a bandwidth that adapts to the array (and chromosome). We consider two forms of overlap in each criteria that follows: (1) equal overlap (EO) - where start and end positions of significant DMRs match that of true DMRs; (2) any overlap (AO) - where significant DMRs intersect true DMRs. When a significant DMR from any method does not overlap a true DMR, we call that a false positive (FP). Similarly, we define a true positive as an overlap between a significant DMR and a true DMR. When a true DMR is not significant, we call it a false negative (FN). The three criteria we use to evaluate performance are:

(i) Recall (power): We estimate power by dividing the number of true positives (TP) by (TP + FN), ie. 2136 true DMRs. That is, recall $= \dfrac{TP}{TP + FN}$.

(ii) Precision: We estimate precision by dividing the number of true positive DMRs by (TP + FP), i.e. the total number of significant DMRs found by our methods. That is, precision $= \dfrac{TP}{TP + FP}$.

(iii) F1 score: This metric combines precision and recall into a single metric ranging from 0 to 1, where 1 represents perfect precision and recall. Some methods have a tendency to prioritize reducing false positives (increasing precision) at the expense of recall (more false negatives). We use this metric, which weights precision and recall equally, because neither precision nor recall alone tells the complete story. We compute F1 score as, $F_1 = 2 \times \dfrac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ (Haibo and Yunqian, 2013; Mallik *et al.*, 2019).

3.3. SIMULATION RESULTS

Based on the EO criteria (see Table 1), the type I error rates (i.e., percent of FPs) for our methods are around the expected 5% (5.16% for *faDMR* and 5.35% for *aaDMR*) with DMRcate a little above (6.07%). The story is reversed using the AO criteria (see Table 2) with all methods well controlling the type I error rate at less than 5% (0 FP for *DMRcate*, 2 FPs for *faDMR* and 1 FP for *aaDMR*). Both *faDMR* and *aaDMR* detected substantially more DMRs compared to DMRcate from of a total of 2136 (see Tables 1 and 2).

Table 1. Large treatment effect ($\Delta\beta = 0.2$): A confusion matrix comparing the results from three methods (DMRcate, faDMR and aaDMR) with true DMRs based on EO criteria averaged across 1000 datasets. Sig. means statistically significant DMRs; Not Sig. indicates the number of regions that are not statistically significant.

DMRcate	True DMR	No DMR	Total
Sig	**284**	1168	**1452**
Not Sig.	1852	18059	19911
Total	2136	19227	21363
faDMR	**True DMR**	**No DMR**	**Total**
Sig.	**753**	992	**1745**
Not Sig.	1383	18235	19618
Total	2136	19227	21363
aaDMR	**True DMR**	**No DMR**	**Total**
Sig.	**772**	1030	**1802**
Not Sig.	1364	18197	19561
Total	2136	19227	21363

The precision, power and F1-score metrics all indicate *faDMR* and *aaDMR* methods are outperforming *DMRcate*. Based on the EO criteria, DMRcate performed poorly on precision, power and F1 score with median of less than 0.2 on all three criteria across the 1000 simulated datasets. Though power and precision are less than 50% for all methods, it's readily improved with median values in *faDMR* and *aaDMR* (averaged across 1000 datasets) more than twice that of *DMRcate* (see Figure 4). At

Table 2. Large treatment effect ($\Delta\beta = 0.2$): A confusion matrix comparing the results from three methods (DMRcate, faDMR and aaDMR) with true DMRs based on AO criteria averaged across 1000 datasets. Sig. means statistically significant DMRs; Not Sig. indicates the number of regions that are not statistically significant.

DMRcate	True DMR	No DMR	Total
Sig	**1452**	0	**1452**
Not Sig.	684	19227	19911
Total	2136	19227	21363
faDMR	True DMR	No DMR	Total
Sig.	**1743**	2	**1745**
Not Sig.	393	19225	19618
Total	2136	19227	21363
aaDMR	True DMR	No DMR	Total
Sig.	**1801**	1	**1802**
Not Sig.	335	19226	19561
Total	2136	19227	21363

the time of writing, we have not come across an article in this domain that utilizes the EO criteria. Most authors tend to favor the less strict AO criteria (Mallik *et al.*, 2019). Based on the AO criteria, all three methods perform well in terms of precision with DMRcate sacrificing power (less than 0.7) for nearly perfect precision. *faDMR* and *aaDMR* both improve power (between 0.8 and 0.9) without hurting precision (see Figure 5). In allowing for the array-adaptive case, we were able to slightly improve power by 3% (comparing *faDMR* to *aaDMR* under AO criteria). Comparing *faDMR* and *aaDMR* using the EO criteria, we notice a slight increase in power ($\sim 1\%$).

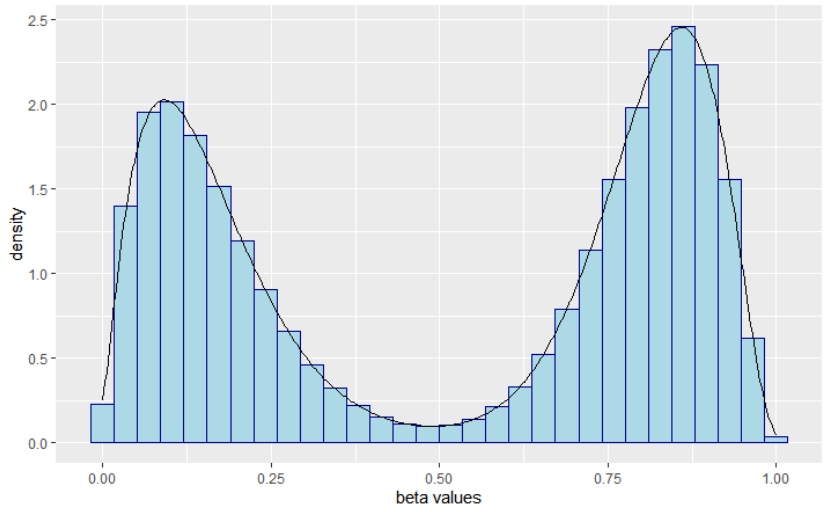

Figure 3. Distribution of $\beta-$values from one of 1000 datasets showing the parameter space is well explored.

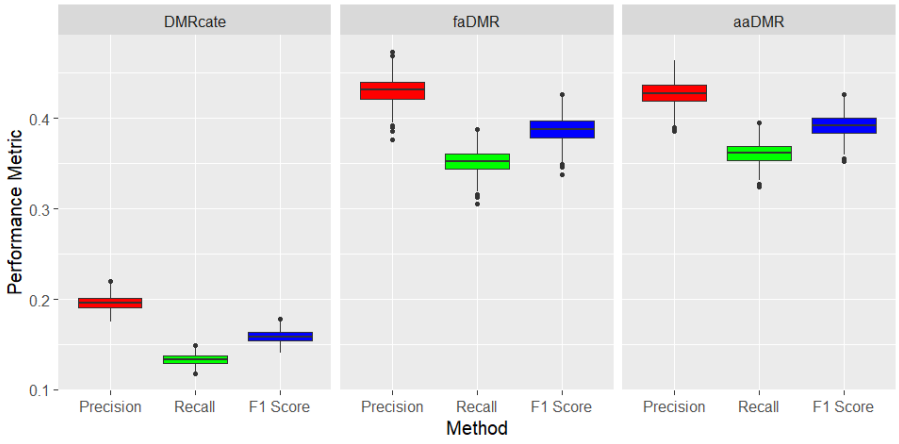

Figure 4. Large treatment effect ($\Delta\beta = 0.2$): Precision, recall and F1 score metrics based on EO criteria. Boxplots of results across the 1000 simulated datasets.

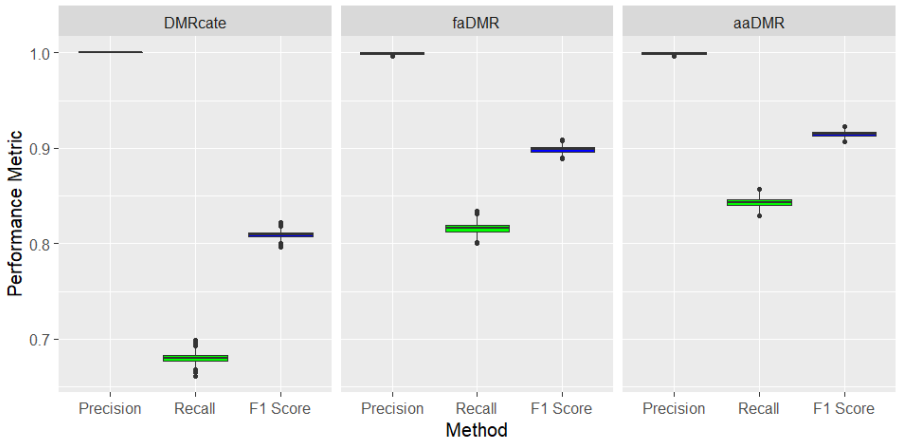

Figure 5. Large treatment effect ($\Delta\beta = 0.2$): Precision, recall and F1 score metrics based on AO criteria. Boxplots of results across the 1000 simulated datasets.

For the small methylation difference (treatment effect) of $\Delta\beta = 0.09$, all methods performed well on precision (based on the AO criteria), but very poor in terms of power. This is usually the case with power for small treatment effects. However, we highlight that our DMR detection methods performed relatively better in terms of power than *DMRcate* (see Table 3 and Figure 7). All methods under the EO criteria performed poorly, yet our methods still do better, especially on precision with median greater than 0.35 compared to the median precision of less than 0.2 for *DMRcate* across the 1000 simulated datasets (see Figure 6).

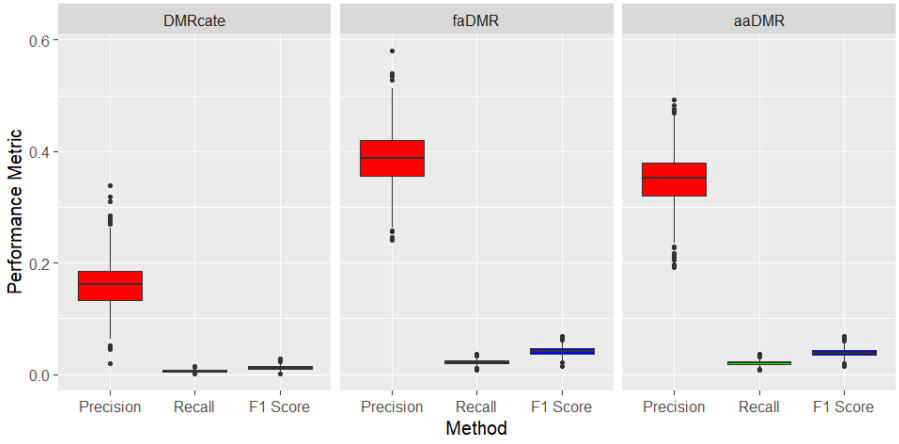

Figure 6. Small treatment effect ($\Delta\beta = 0.09$): Precision, recall and F1 score metrics based on EO criteria. Boxplots of results across the 1000 simulated datasets.

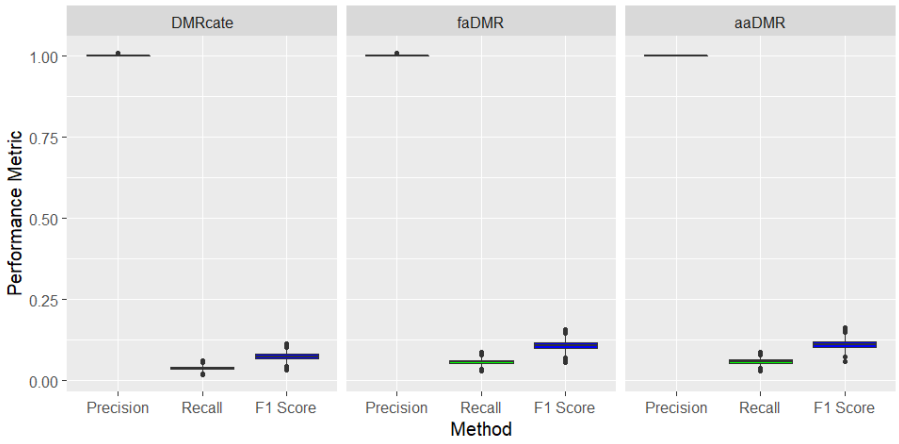

Figure 7. Small treatment effect ($\Delta\beta = 0.09$): Precision, recall and F1 score metrics based on AO criteria. Boxplots of results across the 1000 simulated datasets.

Table 3. Small treatment effect ($\Delta\beta = 0.09$): Confusion matrix comparing three methods (DMRcate, faDMR, and aaDMR) to true DMRs based on AO criteria, averaged across 1000 datasets. Sig. represents statistically significant DMRs; Not Sig. indicates non-statistically significant regions.

DMRcate	True DMR	No DMR	Total
Sig	**82**	0	**82**
Not Sig.	2054	19227	21281
Total	2136	19227	21363
faDMR	**True DMR**	**No DMR**	**Total**
Sig.	**123**	0	**123**
Not Sig.	2013	19227	21240
Total	2136	19227	21363
aaDMR	**True DMR**	**No DMR**	**Total**
Sig.	**127**	0	**127**
Not Sig.	2009	19227	21236
Total	2136	19227	21363

In Figure 8, we compare the density of the CpGs in the DMRs obtained from one of our simulated datasets. As previously stated, these results suggest that our methods are able to detect DMRs in lower density regions compared to DMRcate.

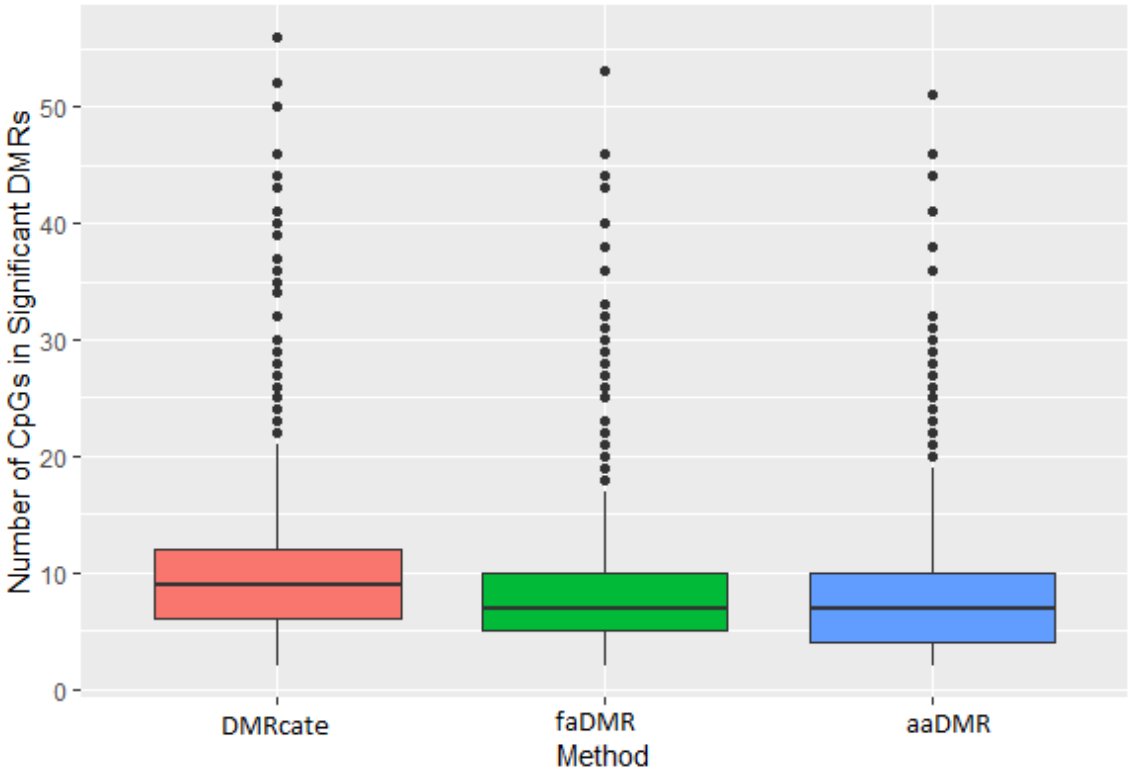

Figure 8. Density of CpGs in Significant DMRs from one of 1000 datasets for the three methods (DMRcate, faDMR and aaDMR).

3.4. REAL DATA EXAMPLE

3.4.1. Data Extraction. We downloaded the Oral Squamous Cell Carcinoma (OSCC) dataset (Basu *et al.*, 2017) from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository. The OSCC dataset (with GEO accession number GSE87053) was generated using the Illumina Infinium 450K Human DNA methylation Beadchip v1.2. A total of 21 samples were collected, with 10 paired tumor and adjacent normal tissues and 1 unpaired tissue. We used the 10 paired tissues for our analysis. Information on human papillomavirus (HPV) status (5 HPV positive, 5 HPV negative) and sex (3 females and 7 males) was also provided.

3.4.2. Oral Squamous Cell Carcinoma (OSCC). Oral cancer is the most common form of head and neck squamous cell cancer. According to the American Cancer Society, about 1 in 60 men and 1 in 140 women have a lifetime risk of developing oral cavity cancer. Aberrant DNA methylation patterns have been found to be associated with OSCC (Basu *et al.*, 2017).

We applied our methods and *DMRcate* to the OSCC data of paired tumor and adjacent normal tissues after we had accounted for the paired data in limma's test-statistic. We adhered to the standard quality control steps such as normalization and probe filtering using the *preprocessFunnorm* (Fortin *et al.*, 2014) function in the *minfi* R/Bioconductor package (Aryee *et al.*, 2014). In addition to this, we controlled for the HPV status and sex in our modeling. All three methods found 13,697 DMRs. *faDMR* and *aaDMR* found more common DMRs (1825) than either method with DMRcate (341 with *aaDMR* and 47 with *faDMR*, see Figure 9). In terms of the number of uniquely identified DMRs, the results were consistent with the simulation study, as *aaDMR* detected the most DMRs (with 828 unique ones) followed by *faDMR* (with 198 unique ones) and then DMRcate (with 247 unique ones) (see Figure 9).

With the large number of DMRs detected, we sought to determine the biological relevance (i.e. the pathways enriched and the associated unique differentially methylated genes) via a KEGG pathway analysis, using the *goregion* function in the *missMethyl* package (Maksimovic *et al.*, 2021). We employed this new function in the *missMethyl* package because in identifying enriched genes, it annotates probes to genes in a way that accounts for the bias towards genes with more measured CpG sites on the array ("probe-bias") and corrects for "multi-gene bias" (Maksimovic *et al.*, 2021) thereby improving the type I error rate (see (Maksimovic *et al.*, 2021) for details). For this downstream analysis, we narrowed our discussion to comparing *aaDMR* and *DMRcate*. Whereas 30 significantly affected pathways were identified from the *aaDMR* procedure, 22 significantly affected pathways were found with *DM-*

Rcate. Nineteen of the 22 significantly affected pathways by *DMRcate* were also reported by *aaDMR*. For the 19 pathways identified by both methods, the evidence of differentially methylated genes was stronger in *aaDMR*. We attribute this to the higher power *aaDMR* has over *DMRcate*, as shown in the simulation study. Moreover, we found that the 11 unique pathways determined by *aaDMR* were related to the immune and nervous system, suggesting that the genes affected by the OSCC disease map to these systems. In the case of *DMRcate*, the 3 unique pathways were related to immune and sensory systems, and cardiovascular disease. The associated genes from the unique pathways identified by *aaDMR* and *DMRcate* were further checked with a wide list of databases using the interactive Enrichr enrichment analysis tool (Chen *et al.*, 2013; Kuleshov *et al.*, 2016; Xie *et al.*, 2021). Our analysis revealed that pathways detected by *aaDMR* contained the AKT serine-threonine protein kinase family (AKT1, AKT2 and AKT3) which has been linked to oral cancer. In their study of the specific role of AKT in OSCC, (Roy *et al.*, 2019) revealed that the silencing of AKT1 and AKT2 genes decreased the expression of proteins regulating cancer cell survival. The frequency of appearances for the three genes suggest their importance in OSCC.

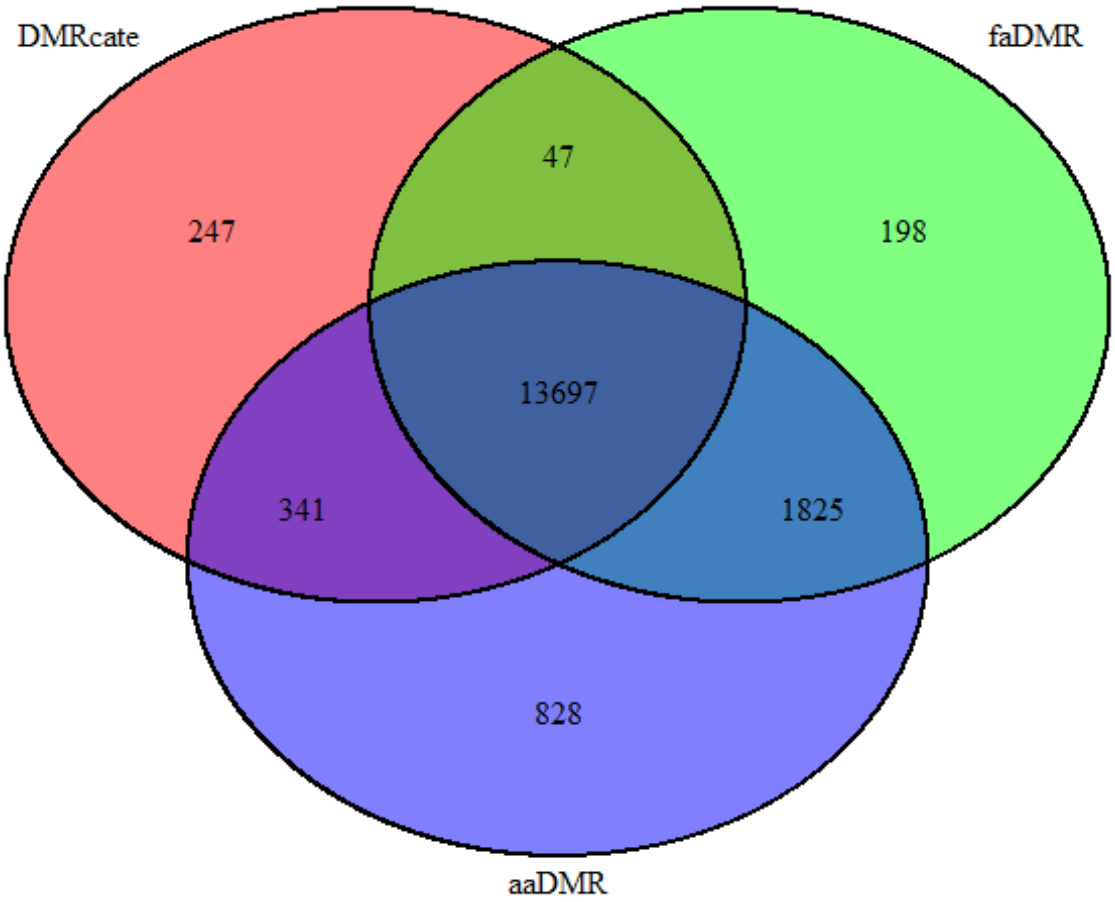

Figure 9. Venn diagram of significant DMRs identified by the three methods from OSCC data.

3.5. SUMMARY OF RESULTS

There were many common DMRs among the three methods compared in both the simulation study and OSCC data example. However *aaDMR* determined more unique DMRs due to its higher power. Using the EO criteria revealed that there is more room for improvement as all methods performed poorly with less than 50% precision and power for the large treatment effect and below 40% precision and 10% power for the small treatment effect. Despite this issue, we point out that our methods are more likely to detect the true length of a DMR than *DMRcate*, as our simulation

results using the EO criteria reveal that *aaDMR* has better precision and recall performance. This finding supports the list of genes that are significantly affected in the pathway analysis when our methods are applied.

4. CONCLUSION

The development of diseases is influenced by a number of factors, some of which are genetic in origin and some of which are environmental. Additionally, by comparing the changes in DNA methylation patterns between disease and normal samples, epigenetic markers like DNA methylation give us a way to understand how diseases are formed. However, this area is not fully understood and methods that reveal unique genomic regions and genes that are enriched by changes in methylation patterns are desirable as they provide researchers with newer areas to explore. We discovered that our method, *aaDMR*, has the capacity to provide researchers with additional pathways to investigate while also revealing particular genes that are significantly impacted by the presence of a condition, thereby assisting researchers in selecting the precise genes that need further analysis. Due to cost and resource constraints, high-throughput experiments frequently have small to medium sample sizes. This means that there is less statistical power to detect DMRs that do exist, which affects the reliability of studies and statistical data analysis results. In such cases, our method may be preferred because it has a higher statistical power to detect a DMR than *DMRcate*, especially in low treatment effect settings.

In summary, we have developed a general class of locally-weighted estimators for use in DMR detection and shown its consistency and asymptotic normality. We have proposed the normalized kernel-weight methods within this general class, which have a higher power to detect a true DMR than *DMRcate* without sacrificing precision for large treatment effect and a higher precision than DMRcate for a low treatment effect. Essentially, our methods demonstrate two things: (1) that using the

normalized kernel-weight is a better way to borrow information from neighboring sites and account for co-methylation than *DMRcate* (revealed by comparing *faDMR* with *DMRcate*) and (2) that accounting for co-methylation using the adaptive-array technique increases the susceptibility of detecting true methylation differences (revealed by comparing *aaDMR* with *faDMR*).

ACKNOWLEDGMENTS

We would like to thank Dr. Matthew Thimgan (Associate Professor, Department of Biological Sciences at Missouri University of Science and Technology) for helpful discussions on our methods and pathway analysis.

REFERENCES

'Illumina methylation beadchips achieve breadth of coverage using 2 Infinium chemistries,' Technical Report Pub. No. 270-2012-00, Illumina, Inc..,, 2015, techsupport@illumina.com.

Aalen, O., 'A model for nonparametric regression analysis of counting processes,' in 'Mathematical statistics and probability theory,' pp. 1–25, Springer, 1980.

Aalen, O., Borgan, O., and Gjessing, H., *Survival and event history analysis: a process point of view*, Springer Science & Business Media, 2008.

Aalen, O. O., 'A linear regression model for the analysis of life times,' Statistics in medicine, 1989, **8**(8), pp. 907–925.

Adler, A. and Rosalsky, A., 'On the weak law of large numbers for normed weighted sums of iid random variables,' International Journal of Mathematics and Mathematical Sciences, 1991, **14**(1), pp. 191–202.

Amico, M. and Van Keilegom, I., 'Cure models in survival analysis,' Annual Review of Statistics and Its Application, 2018, **5**, pp. 311–342.

Amico, M., Van Keilegom, I., and Legrand, C., 'The single-index/cox mixture cure model,' Biometrics, 2019, **75**(2), pp. 452–462.

Anderssen, R. and Bloomfield, P., 'A time series approach to numerical differentiation,' Technometrics, 1974, **16**(1), pp. 69–75.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A., 'Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays,' Bioinformatics, 2014, **30**(10), pp. 1363–1369, doi:10.1093/bioinformatics/btu049.

Barfield, R. T., Kilaru, V., Smith, A. K., and Conneely, K. N., 'Cpgassoc: an r function for analysis of dna methylation microarray data,' Bioinformatics, 2012, **28**(9), pp. 1280–1281.

Basu, B., Chakraborty, J., Chandra, A., Katarkar, A., Baldevbhai, J. R. K., Dhar Chowdhury, D., Ray, J. G., Chaudhuri, K., and Chatterjee, R., 'Genome-wide DNA methylation profile identified a unique set of differentially methylated immune genes in oral squamous cell carcinoma patients in India,' Clin Epigenetics, 2017, **9**, p. 13.

Benjamini, Y. and Hochberg, Y., 'Controlling the false discovery rate: a practical and powerful approach to multiple testing,' Journal of the Royal statistical society: series B (Methodological), 1995a, **57**(1), pp. 289–300.

Benjamini, Y. and Hochberg, Y., 'Controlling the false discovery rate: a practical and powerful approach to multiple testing,' Journal of the Royal statistical society: series B (Methodological), 1995b, **57**(1), pp. 289–300.

Bennett, S., 'Analysis of survival data by the proportional odds model,' Statistics in medicine, 1983, **2**(2), pp. 273–277.

Berkson, J. and Gage, R. P., 'Survival curve for cancer patients following treatment,' Journal of the American Statistical Association, 1952, **47**(259), pp. 501–515.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., *et al.*, 'High density dna methylation array with single cpg site resolution,' Genomics, 2011, **98**(4), pp. 288–295.

Billingsley, P., *Probability and Measure*, John Wiley and Sons, second edition, 1986.

Billingsley, P., *Probability and measure*, John Wiley & Sons, 1995.

Boag, J. W., 'Maximum likelihood estimates of the proportion of patients cured by cancer therapy,' Journal of the royal statistical society series b-methodological, 1949, doi:10.1111/j.2517-6161.1949.tb00020.x.

Breton-Larrivée, M., Elder, E., and McGraw, S., 'DNA methylation, environmental exposures and early embryo development,' Anim Reprod, Oct 2019, **16**(3), pp. 465–474.

Butcher, L. M. and Beck, S., 'Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data,' Methods, Jan 2015, **72**, pp. 21–28.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P., 'Generalized partially linear single-index models,' Journal of the American Statistical Association, 1997, **92**(438), pp. 477–489.

Casella, G. and Berger, R. L., *Statistical inference*, Cengage Learning, 2021.

Cedar, H., 'Dna methylation and gene activity.' Cell, 1988, **53**(1), pp. 3–4.

Center for Cancer Genomics - National Cancer Institute, 'The Cancer Genome Atlas Program (TCGA),' Retrieved from `https://www.cancer.gov/ccg/research/genome-sequencing/tcga`, n.d., accessed June 8, 2023.

Chen, D. P., Lin, Y. C., and Fann, C. S., 'Methods for identifying differentially methylated regions for sequence- and array-based data,' Brief. Funct. Genomics, 2016, **15**(6), pp. 485–490, ISSN 20412657, doi:10.1093/bfgp/elw018.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A., 'Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool,' BMC Bioinformatics, Apr 2013, **14**, p. 128.

Chen, Y. A., Choufani, S., Grafodatskaya, D., Butcher, D. T., Ferreira, J. C., and Weksberg, R., 'Cross-reactive DNA microarray probes lead to false discovery of autosomal sex-associated DNA methylation,' Am J Hum Genet, Oct 2012, **91**(4), pp. 762–764.

Chen, Y. Q. and Wang, M.-C., 'Analysis of accelerated hazards models,' Journal of the American Statistical Association, 2000, **95**(450), pp. 608–618.

Chiou, S. H., Austin, M. D., Qian, J., and Betensky, R. A., 'Transformation model estimation of survival under dependent truncation and independent censoring,' Statistical methods in medical research, 2019, **28**(12), pp. 3785–3798.

Cox, D. R., 'Regression models and life-tables,' J. Roy. Statist. Soc. Ser. B, 1972, **34**, pp. 187–220, ISSN 0035-9246.

Cox, D. R., 'Partial likelihood,' Biometrika, 1975, **62**(2), pp. 269–276.

Cox, D. R. and Oakes, D., *Analysis of survival data*, Chapman and Hall/CRC, 1984.

Cramér, H., *Mathematical Methods of Statistics (PMS-9), Volume 9*, Princeton university press, 2016.

Crary-Dooley, F. K., Tam, M. E., Dunaway, K. W., Hertz-Picciotto, I., Schmidt, R. J., and LaSalle, J. M., 'A comparison of existing global dna methylation assays to low-coverage whole-genome bisulfite sequencing for epidemiological studies,' Epigenetics, 2017, **12**(3), pp. 206–214.

Craven, P. and Wahba, G., 'Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation,' Numerische mathematik, 1978, **31**(4), pp. 377–403.

DasGupta, A., *Asymptotic theory of statistics and probability*, Springer Science & Business Media, 2008.

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F., 'Evaluation of the Infinium Methylation 450K technology,' Epigenomics, Dec 2011, **3**(6), pp. 771–784.

Dempster, A. P., Laird, N. M., and Rubin, D. B., 'Maximum likelihood from incomplete data via the em algorithm,' Journal of the Royal Statistical Society: Series B (Methodological), 1977, **39**(1), pp. 1–22.

Dirick, L., Claeskens, G., Vasnev, A., and Baesens, B., 'A hierarchical mixture cure model with unobserved heterogeneity for credit risk,' Econometrics and Statistics, 2022, **22**, pp. 39–55.

Du, P., Zhang, X., Huang, C. C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M., 'Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis,' BMC Bioinformatics, Nov 2010, **11**, p. 587.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S., 'DNA methylation profiling of human chromosomes 6, 20 and 22,' Nat Genet, Dec 2006, **38**(12), pp. 1378–1385.

Eden, S. and Cedar, H., 'Role of DNA methylation in the regulation of transcription,' Current opinion in genetics & development, 1994, **4**(2), pp. 255–259.

Efron, B. and Morris, C., 'Empirical bayes on vector observations: An extension of stein's method,' Biometrika, 1972, **59**(2), pp. 335–347.

Ehrlich, M. and Wang, R. Y.-H., '5-methylcytosine in eukaryotic DNA,' Science, 1981, **212**(4501), pp. 1350–1357.

Eilers, P. H. and Marx, B. D., 'Flexible smoothing with b-splines and penalties,' Statistical science, 1996, **11**(2), pp. 89–121.

ENCODE Project Consortium, 'An integrated encyclopedia of dna elements in the human genome,' Nature, 2012, **489**(7414), p. 57, doi:10.1038/nature11247.

Farewell, V. T., 'A model for a binary variable with time-censored observations,' Biometrika, 1977, **64**(1), pp. 43–46.

Farewell, V. T., 'The use of mixture models for the analysis of survival data with long-term survivors,' Biometrics, 1982, pp. 1041–1046.

Felizzi, F., Paracha, N., Pöhlmann, J., and Ray, J., 'Mixture cure models in oncology: a tutorial and practical guidance,' PharmacoEconomics-Open, 2021, **5**, pp. 143–155.

Fernandez, A., O'Leary, C., O'Byrne, K. J., Burgess, J., Richard, D. J., and Suraweera, A., 'Epigenetic mechanisms in dna double strand break repair: A clinical review,' Frontiers in Molecular Biosciences, 2021, **8**, p. 685440.

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., Greenwood, C. M., and Hansen, K. D., 'Functional normalization of 450k methylation array data improves replication in large cancer studies,' Genome Biology, 2014, **15**(12), p. 503, doi:10.1186/s13059-014-0503-2.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L., 'A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands,' Proc Natl Acad Sci U S A, Mar 1992, **89**(5), pp. 1827–1831.

Greenberg, M. V. and Bourc'his, D., 'The diverse roles of DNA methylation in mammalian development and disease,' Nature reviews Molecular cell biology, 2019, **20**(10), pp. 590–607.

Haibo, H. and Yunqian, M., 'Imbalanced learning: foundations, algorithms, and applications,' Wiley-IEEE Press, 2013, **1**, p. 27.

Hanin, L. and Huang, L.-S., 'Identifiability of cure models revisited,' Journal of Multivariate Analysis, 2014, **130**, pp. 261–274.

Hardle, W., Hall, P., and Ichimura, H., 'Optimal smoothing in single-index models,' The annals of Statistics, 1993, **21**(1), pp. 157–178.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., *et al.*, *Nonparametric and semiparametric models*, volume 1, Springer, 2004.

Härdle, W. K. *et al.*, *Smoothing techniques: with implementation in S*, Springer Science & Business Media, 1991.

Heiss, J. A. and Just, A. C., 'Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses,' Clin Epigenetics, 01 2019, **11**(1), p. 15.

Hsu, W.-W., Todem, D., and Kim, K., 'A sup-score test for the cure fraction in mixture models for long-term survivors,' Biometrics, 2016, **72**(4), pp. 1348–1357.

Huffer, F. W. and McKeague, I. W., 'Weighted least squares estimation for aalen's additive risk model,' Journal of the American Statistical Association, 1991, **86**(413), pp. 114–129.

Huster, W. J., Brookmeyer, R., and Self, S. G., 'Modelling paired survival data with covariates,' Biometrics, 1989, pp. 145–156.

Ichiyanagi, T., Ichiyanagi, K., Miyake, M., and Sasaki, H., 'Accumulation and loss of asymmetric non-CpG methylation during male germ-cell development,' Nucleic Acids Res, Jan 2013, **41**(2), pp. 738–745.

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A., 'Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies,' Int J Epidemiol, Feb 2012, **41**(1), pp. 200–209.

James, G., Witten, D., Hastie, T., and Tibshirani, R., *An introduction to statistical learning*, volume 112, Springer, 2013.

Jeong, M., Guzman, A. G., and Goodell, M. A., 'Genome-wide analysis of dna methylation in hematopoietic cells: Dna methylation analysis by wgbs,' Acute Myeloid Leukemia: Methods and Protocols, 2017, pp. 137–149.

Jiang, J., *Large sample techniques for statistics*, Springer Science & Business Media, 2010.

Jin, Z. and Liu, Y., 'DNA methylation in human diseases,' Genes Dis., 2018, **5**(1), pp. 1–8, ISSN 23523042, doi:10.1016/j.gendis.2018.01.002.

Kalbfleisch, J. D. and Prentice, R. L., 'Marginal likelihoods based on cox's regression and life model,' Biometrika, 1973, **60**(2), pp. 267–278.

Kalbfleisch, J. D. and Prentice, R. L., *The statistical analysis of failure time data*, John Wiley & Sons, 2011.

Kaplan, E. L. and Meier, P., 'Nonparametric estimation from incomplete observations,' Journal of the American statistical association, 1958, **53**(282), pp. 457–481.

Khaliq, A., Waqas, A., Nisar, Q. A., Haider, S., and Asghar, Z., 'Application of ai and robotics in hospitality sector: A resource gain and resource loss perspective,' Technology in Society, 2022, **68**, p. 101807.

Kilaru, V., Barfield, R. T., Schroeder, J. W., Smith, A. K., and Conneely, K. N., 'MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data,' Epigenetics, Mar 2012, **7**(3), pp. 225–229.

Klein, J. P. and Moeschberger, M. L., *Survival analysis: techniques for censored and truncated data*, volume 1230, Springer, 2003.

Kleinbaum, D. G. and Klein, M., *Survival analysis a self-learning text*, Springer, third edition, 2012.

Kuk, A. Y. and Chen, C.-H., 'A mixture model combining logistic regression with proportional hazards regression,' Biometrika, 1992, **79**(3), pp. 531–541.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., and Ma'ayan, A., 'Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,' Nucleic Acids Res, 07 2016, **44**(W1), pp. W90–97.

Laird, P. W., 'Principles and challenges of genome-wide dna methylation analysis,' Nature Reviews Genetics, 2010, **11**(3), pp. 191–203.

Lao, V. V. and Grady, W. M., 'Epigenetics and colorectal cancer,' Nat Rev Gastroenterol Hepatol, Oct 2011, **8**(12), pp. 686–700.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Sung, K. W. K., Rigoutsos, I., Loring, J., and Wei, C. L., 'Dynamic changes in the human methylome during differentiation,' Genome Res., 2010, **20**(3), pp. 320–331, ISSN 10889051, doi:10.1101/gr.101907.109.

Lee, E. T. and Wang, J., *Statistical methods for survival data analysis*, volume 476, John Wiley & Sons, 2003.

Leek, J. T. and Storey, J. D., 'Capturing heterogeneity in gene expression studies by surrogate variable analysis,' PLoS Genet, Sep 2007, **3**(9), pp. 1724–1735.

Legrand, C., *Advanced survival models*, Chapman and Hall/CRC, 2021.

Lehmann, E. L., *Elements of large-sample theory*, Springer Science & Business Media, 2004.

Li, C.-S. and Lu, M., 'A lack-of-fit test for generalized linear models via single-index techniques,' Computational Statistics, 2018, **33**, pp. 731–756.

Li, C.-S. and Taylor, J. M., 'Smoothing covariate effects in cure models,' Communications in Statistics-Theory and Methods, 2002, **31**(3), pp. 477–493.

Li, D., Xie, Z., Pape, M. L., and Dye, T., 'An evaluation of statistical methods for DNA methylation microarray data analysis,' BMC Bioinformatics, jul 2015, **16**(1), ISSN 14712105, doi:10.1186/s12859-015-0641-x.

Lin, D. Y. and Ying, Z., 'Semiparametric analysis of the additive risk model,' Biometrika, 1994, **81**(1), pp. 61–71.

Liu, A., Jiang, C., Liu, Q., Yin, H., Zhou, H., Ma, H., and Geng, Q., 'The inverted u-shaped association of caffeine intake with serum uric acid in us adults,' The journal of nutrition, health & aging, 2022, **26**(4), pp. 391–399.

López-Cheda, A., Cao, R., Jácome, M. A., and Van Keilegom, I., 'Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models,' Computational Statistics & Data Analysis, 2017, **105**, pp. 144–165.

Lu, W., 'Maximum likelihood estimation in the proportional hazards cure model,' Annals of the Institute of Statistical Mathematics, 2008, **60**, pp. 545–574.

Ma, Y. and He, H., 'Imbalanced learning: foundations, algorithms, and applications,' 2013.

Madadizadeh, F., Ghanbarnejad, A., Ghavami, V., Bandamiri, M. Z., and Mohammadianpanah, M., 'Applying additive hazards models for analyzing survival in patients with colorectal cancer in fars province, southern iran,' Asian Pacific journal of cancer prevention: APJCP, 2017, **18**(4), p. 1077.

Maghbooli, Z., Larijani, B., Emamgholipour, S., Amini, M., Keshtkar, A., and Pasalar, P., 'Aberrant DNA methylation patterns in diabetic nephropathy,' J Diabetes Metab Disord, 2014, **13**, p. 69.

Maksimovic, J., Gordon, L., and Oshlack, A., 'SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips,' Genome Biology, 2012, **13**(6), p. R44, doi:10.1186/gb-2012-13-6-r44.

Maksimovic, J., Oshlack, A., and Phipson, B., 'Gene set enrichment analysis for genome-wide DNA methylation data,' Genome Biol, 06 2021, **22**(1), p. 173.

Maller, R. A. and Zhou, X., *Survival analysis with long-term survivors*, volume 525, Wiley New York, 1996.

Mallik, S., Odom, G. J., Gao, Z., Gomez, L., Chen, X., and Wang, L., 'An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays,' Brief Bioinform, 11 2019, **20**(6), pp. 2224–2235.

McCartney, D. L., Walker, R. M., Morris, S. W., McIntosh, A. M., Porteous, D. J., and Evans, K. L., 'Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationepic beadchip,' Genomics data, 2016, **9**, pp. 22–24.

McCullagh, P. and Nelder, J. A., *Generalized linear models*, Routledge, 2019.

McGregor, K., Bernatsky, S., Colmegna, I., Hudson, M., Pastinen, T., Labbe, A., and Greenwood, C. M., 'An evaluation of methods correcting for cell-type heterogeneity in dna methylation studies,' Genome biology, 2016, **17**(1), pp. 1–17.

McLachlan, G. J. and Krishnan, T., *The EM algorithm and extensions*, John Wiley & Sons, 2007.

Mood, A., Graybill, F., and Boes, D., '(1974), introduction to the theory of statistics,' 1974.

Moran, S., Arribas, C., and Esteller, M., 'Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences,' Epigenomics, 2016, **8**(3), pp. 389–399.

Patilea, V. and Van Keilegom, I., 'A general approach for cure models in survival analysis,' 2020.

Peng, Y. and Dear, K. B., 'A nonparametric mixture model for cure rate estimation,' Biometrics, 2000, **56**(1), pp. 237–243.

Peng, Y. and Yu, B., *Cure Models: Methods, Applications, and Implementation*, Chapman and Hall/CRC, 2021.

Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., V Lord, R., Clark, S. J., and Molloy, P. L., 'De novo identification of differentially methylated regions in the human genome,' Epigenetics Chromatin, 2015, **8**, p. 6.

Piao, Y., Xu, W., Park, K. H., Ryu, K. H., and Xiang, R., 'Comprehensive Evaluation of Differential Methylation Analysis Methods for Bisulfite Sequencing Data,' Int J Environ Res Public Health, 07 2021, **18**(15).

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C., 'A data-driven approach to preprocessing illumina 450k methylation array data,' BMC genomics, 2013, **14**(1), pp. 1–10.

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., and Clark, S. J., 'Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling,' Genome Biol, 10 2016, **17**(1), p. 208.

Price, D., *Survival Models for Heterogeneous Populations with Cure*, Ph.D. thesis, Emory University, 2000.

Procter, M., Chou, L.-S., Tang, W., Jama, M., and Mao, R., 'Molecular diagnosis of prader–willi and angelman syndromes by methylation-specific melting analysis and methylation-specific multiplex ligation-dependent probe amplification,' Clinical chemistry, 2006, **52**(7), pp. 1276–1283.

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., and Jaenisch, R., 'Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a,' Proc. Natl. Acad. Sci. U. S. A., 2000, **97**(10), pp. 5237–5242, ISSN 00278424, doi:10.1073/pnas.97.10.5237.

Resnick, S. I., *A probability path*, Birkhäuser Boston, 1999.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K., 'limma powers differential expression analyses for RNA-sequencing and microarray studies,' Nucleic Acids Res, Apr 2015, **43**(7), p. e47.

Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X., 'Statistical methods for detecting differentially methylated loci and regions,' Frontiers in genetics, 2014, **5**, p. 324.

Rossignol, S., Steunou, V., Chalas, C., Kerjean, A., Rigolet, M., Viegas-Pequignot, E., Jouannet, P., Le Bouc, Y., and Gicquel, C., 'The epigenetic imprinting defect of patients with beckwith—wiedemann syndrome born after assisted reproductive technology is not restricted to the 11p15 region,' Journal of medical genetics, 2006, **43**(12), pp. 902–907.

Roy, N. K., Monisha, J., Padmavathi, G., Lalhruaitluanga, H., Kumar, N. S., Singh, A. K., Bordoloi, D., Baruah, M. N., Ahmed, G. N., Longkumar, I., Arfuso, F., Kumar, A. P., and Kunnumakkara, A. B., 'Isoform-Specific Role of Akt in Oral Squamous Cell Carcinoma,' Biomolecules, 06 2019, **9**(7).

Ruppert, D., Wand, M. P., and Carroll, R. J., *Semiparametric regression*, 12, Cambridge university press, 2003.

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., and Esteller, M., 'Validation of a dna methylation microarray for 450,000 cpg sites in the human genome,' Epigenetics, 2011, **6**(6), pp. 692–702.

Satterthwaite, F. E., 'An approximate distribution of estimates of variance components,' Biometrics, Dec 1946, **2**(6), pp. 110–114.

Shafi, A., Mitrea, C., Nguyen, T., and Draghici, S., 'A survey of the approaches for identifying differential methylation using bisulfite sequencing data,' Brief Bioinform, 09 2018, **19**(5), pp. 737–753.

Shang, S., Liu, M., Zeleniuch-Jacquotte, A., Clendenen, T. V., Krogh, V., Hallmans, G., and Lu, W., 'Partially linear single index cox regression model in nested case-control studies,' Computational statistics & data analysis, 2013, **67**, pp. 199–212.

Shiah, Y. J., Fraser, M., Bristow, R. G., and Boutros, P. C., 'Comparison of pre-processing methods for Infinium HumanMethylation450 BeadChip array,' Bioinformatics, Oct 2017, **33**(20), pp. 3151–3157.

Shu, C., Zhang, X., Aouizerat, B. E., and Xu, K., 'Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells,' Epigenetics Chromatin, 11 2020, **13**(1), p. 51.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, volume 26, CRC Press, 1986.

Smith, M. L., Baggerly, K. A., Bengtsson, H., Ritchie, M. E., and Hansen, K. D., 'illuminaio: An open source idat parsing tool for illumina microarrays,' F1000Research, 2013, **2**.

Smyth, G. K., 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments,' Stat Appl Genet Mol Biol, 2004, **3**, p. Article3.

Sofer, T., Schifano, E. D., Hoppin, J. A., Hou, L., and Baccarelli, A. A., 'A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure,' Bioinformatics, Nov 2013, **29**(22), pp. 2884–2891.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M., 'The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1,' 1949.

Sun, H. and Wang, S., 'Penalized logistic regression for high-dimensional DNA methylation data with case-control studies,' Bioinformatics, May 2012, **28**(10), pp. 1368–1375.

Sun, L., Namboodiri, S., Chen, E., and Sun, S., 'Preliminary Analysis of Within-Sample Co-methylation Patterns in Normal and Cancerous Breast Samples,' Cancer Inform, 2019, **18**, p. 1176935119880516.

Sun, L. and Sun, S., 'Within-sample co-methylation patterns in normal tissues,' BioData Min, 2019, **12**, p. 9.

Sun, S., Dammann, J., Lai, P., and Tian, C., 'Thorough statistical analyses of breast cancer co-methylation patterns,' BMC Genom Data, 04 2022, **23**(1), p. 29.

Susan, J. C., Harrison, J., Paul, C. L., and Frommer, M., 'High sensitivity mapping of methylated cytosines,' Nucleic acids research, 1994, **22**(15), pp. 2990–2997.

Sy, J. P. and Taylor, J. M., 'Estimation in a cox proportional hazards cure model,' Biometrics, 2000, **56**(1), pp. 227–236.

Szyf, M., 'Dna methylation signatures for breast cancer classification and prognosis,' Genome medicine, 2012, **4**(3), pp. 1–12.

Taylor, H. L., 'Physical activity: is it still a risk factor?' Preventive medicine, 1983, **12**(1), pp. 20–24.

Taylor, J. M., 'Semi-parametric estimation in failure time mixture models,' Biometrics, 1995, pp. 899–907.

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S., 'A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data,' Bioinformatics, Jan 2013, **29**(2), pp. 189–196.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 'A promoter-level mammalian expression atlas,' Nature, 2014, **507**(7493), pp. 462–470.

Therneau, T. M., Grambsch, P. M., Therneau, T. M., and Grambsch, P. M., *The cox model*, Springer, 2000.

Touleimat, N. and Tost, J., 'Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation,' Epigenomics, Jun 2012, **4**(3), pp. 325–341.

Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., and Siegmund, K. D., 'Low-level processing of Illumina Infinium DNA Methylation BeadArrays,' Nucleic Acids Res, Apr 2013, **41**(7), p. e90.

Wang, D., Yan, L., Hu, Q., Sucheston, L. E., Higgins, M. J., Ambrosone, C. B., Johnson, C. S., Smiraglia, D. J., and Liu, S., 'Ima: an r package for high-throughput analysis of illumina's 450k infinium methylation data,' Bioinformatics, 2012, **28**(5), pp. 729–730.

Wang, L. and Cao, G., 'Efficient estimation for generalized partially linear single-index models,' 2018.

Wang, T., Guan, W., Lin, J., Boutaoui, N., Canino, G., Luo, J., Celedón, J. C., and Chen, W., 'A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data,' Epigenetics, 2015, **10**(7), pp. 662–669.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Foekens, J. A., 'Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,' Lancet, 2005, **365**(9460), pp. 671–679.

Wang, Z., Wu, X., and Wang, Y., 'A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip,' BMC Bioinformatics, 04 2018, **19**(Suppl 5), p. 115.

Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y. C., 'COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis,' Nucleic Acids Res, Jun 2013a, **41**(11), p. e117.

Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y.-C., 'Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis,' Nucleic acids research, 2013b, **41**(11), pp. e117–e117.

Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R., Dymov, S., Szyf, M., and Meaney, M. J., 'Epigenetic programming by maternal behavior,' Nat Neurosci, Aug 2004, **7**(8), pp. 847–854.

Weisenberger, D., Van Den Berg, D., Pan, F., Berman, B., and Laird, P., 'Comprehensive dna methylation analysis on the illumina infinium assay platform,' Illumina, San Diego, 2008.

Wood, S., 'mgcv: Mixed gam computation vehicle with gcv/aic/reml smoothness estimation,' 2012.

Wood, S. N., 'Thin plate regression splines,' Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2003, **65**(1), pp. 95–114.

Wu, D., Gu, J., and Zhang, M. Q., 'Fastdma: an infinium humanmethylation450 beadchip analyzer,' PloS one, 2013, **8**(9), p. e74275.

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., and Ma'ayan, A., 'Gene Set Knowledge Discovery with Enrichr,' Curr Protoc, Mar 2021, **1**(3), p. e90.

Xu, J. and Peng, Y., 'Nonparametric cure rate estimation with covariates,' Canadian Journal of Statistics, 2014, **42**(1), pp. 1–17.

Yu, Y. and Ruppert, D., 'Penalized spline estimation for partially linear single-index models,' Journal of the American Statistical Association, 2002, **97**(460), pp. 1042–1054.

Yu, Y., Wu, C., and Zhang, Y., 'Penalised spline estimation for generalised partially linear single-index models,' Statistics and Computing, 2017, **27**, pp. 571–582.

Zeng, Z., Gao, Y., Li, J., Zhang, G., Sun, S., Wu, Q., Gong, Y., and Xie, C., 'Violations of proportional hazard assumption in cox regression model of transcriptomic data in tcga pan-cancer cohorts,' Computational and Structural Biotechnology Journal, 2022, **20**, pp. 496–507.

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E., 'Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements,' Genome Biol, Jan 2015, **16**, p. 14.

Zhang, Y., Liu, H., Lv, J., Xiao, X., Zhu, J., Liu, X., Su, J., Li, X., Wu, Q., Wang, F., *et al.*, 'Qdmr: a quantitative method for identification of differentially methylated regions by entropy,' Nucleic acids research, 2011, **39**(9), pp. e58–e58.

Zhang, Y., Wang, S., and Wang, X., 'Data-Driven-Based Approach to Identifying Differentially Methylated Regions Using Modified 1D Ising Model,' Biomed Res. Int., 2018, **2018**, ISSN 23146141, doi:10.1155/2018/1070645.

Zhao, Y., Lee, A. H., Yau, K. K., Burke, V., and McLachlan, G. J., 'A score test for assessing the cured proportion in the long-term survivor mixture model,' Statistics in medicine, 2009, **28**(27), pp. 3454–3466.

II. A GENERALIZED PARTIALLY LINEAR SINGLE-INDEX ADDITIVE HAZARD MIXTURE CURE MODEL

Daniel Ahmed Alhassan
Department of Mathematics & Statistics
Missouri University of Science and Technology
Rolla, Missouri 65409–0050

ABSTRACT

When modeling survival data with a large number of cured patients, the semiparametric mixture cure model that assumes the additive hazards model for the susceptibles has received less attention compared to its proportional hazard (PH) counterpart. However, the PH assumption is often violated in practice, and its violation can lead to biased results. The logistic structure used for the incidence model is also considered too narrow, especially in very large studies. In this work, we propose a generalized partially linear single-index model for the incidence and an additive hazard model for the latency. We describe a computational approach for our mixture cure model, which combines the computational methods for the generalized partially linear single-index model with the semiparametric additive hazards model. This approach can be easily implemented in statistical software. To demonstrate the effectiveness of our model, we present a survival example for diabetic retinopathy patients. The proposed model provides a flexible framework for analyzing the incidence and latency of diseases, making it particularly useful in medical, economic and finance research fields. The computational approach presented in this work enables efficient estimation and inference, making it a valuable tool for analyzing large datasets.

Keywords: mixture cure models, single-index, generalized partial linear, EM algorithm, additive hazard models, P-splines.

1. INTRODUCTION

The study of time-to-event data is an important subject in statistics. Classical survival analysis presupposes that if someone is monitored for an extended period of time, they will ultimately experience the event of interest. This assumption is not always true. Even more recently, due to the improvement of life through better health care, jobs, etc., this assumption has proven unrealistic in many situations, and researchers have become interested in studying failure time data in the presence a of cure fraction. Consider situations where a patient will "never" suffer the relapse of a disease. Such situations are more prevalent now due to better healthcare, for example in oncology (Felizzi *et al.*, 2021). Another example from economics is monitoring the time until an unemployed person finds a new job. The recent advancement of technology due to artificial intelligence has seen many lose their jobs (Khaliq *et al.*, 2022) and as such, until the unemployed obtain the necessary skills, they are likely never to actually find a new job.

When a certain fraction of the population will never experience the event of interest, they are considered "long-term survivors", "cured" or "immune" (Maller and Zhou, 1996) and as such, their survival times are infinite. As it is in classical survival analysis, it is impossible to observe subjects for an infinite time; hence, at the end of the study, some subjects may be right-censored (that is, only a lower bound for their survival time is known). In addition, for a mixed population of cured and uncured subjects, censored observations will consist of both cured and uncured subjects, so the overall survival may no longer indicate a steady fall to zero. More to the point, at the study's end, the censored group will contain both the cured and uncured, and the task of fitting cure models takes into account this characteristic. Examining the Kaplan-Meier estimator (Kaplan and Meier, 1958) of a survival curve is a simple approach to tell if a certain set of data has a subset of long-term survivors. A cure

model could be a suitable and practical method of data analysis if the survival curve plateaus towards the end of the study at a value higher than zero (Sy and Taylor, 2000).

The study of cure models dates back to Boag (1949), Berkson and Gage (1952) and Farewell (1982). They studied what has become known as *mixture cure models*. In the presence of covariates, the survival function of the population, $S(t|\boldsymbol{X}, \boldsymbol{Z}) = P(T > t|\boldsymbol{X}, \boldsymbol{Z})$ of survival time T given covariates $(\boldsymbol{X}, \boldsymbol{Z})$ is given by:

$$S(t|\boldsymbol{X}, \boldsymbol{Z}) = P(T > t|\boldsymbol{X}, \boldsymbol{Z}) = 1 - p(\boldsymbol{X}) + p(\boldsymbol{X})S_u(t|\boldsymbol{Z}) \tag{1}$$

where $p(\boldsymbol{X}) = P(B = 1|\boldsymbol{X} = \boldsymbol{x})$ is the probability of being susceptible (uncured) (also referred to as "incidence"), $S_u(t|\boldsymbol{Z}) = P(T > t|\boldsymbol{Z} = \boldsymbol{z}, B = 1)$ is a proper conditional survival function of the susceptible group (also called the "latency") and $B = I(T < \infty)$ is the partially observed uncured status with indicator function, $I(\cdot)$.

Different models considered for the latency and incidence groups include parametric, semi-parametric, and fully non-parametric mixture cure models. Fully parametric models were first considered by Boag (1949) and Berkson and Gage (1952) who considered a constant incidence and employed the log-normal and exponential models for the latency, respectively. Farewell (1977) later introduced the covariates in the incidence by employing a logistic regression model for $p(\boldsymbol{x}) = \dfrac{\exp(\boldsymbol{\gamma}^\top \boldsymbol{x})}{(1 + \exp(\boldsymbol{\gamma}^\top \boldsymbol{x}))}$. Semi-parametric mixture cure models have been studied extensively by Kuk and Chen (1992), Sy and Taylor (2000), Peng and Dear (2000), and Lu (2008). In all four of these articles, the authors considered the Cox proportional hazard model (Cox, 1972) for the latency sub-model and employed the logistic model for the incidence sub-model while proposing different estimation methods. For a completely non-parametric mixture cure model, the main contribution is due to López-Cheda *et al.* (2017) and Patilea and Van Keilegom (2020). In both the parametric and semi-parametric cases,

the incidence has been extensively modeled using a logistic form and has received this much attention due to the ease of estimation and interpretation as well as its availability in many statistical software packages such as R (Amico *et al.*, 2019). However, it has been noted that the relationship between the covariates and the response is not always linear, and logistic regression can perform poorly in such cases (James *et al.*, 2013). In addition, if the true shape of the cure rate is not an S-shape, then the logistic model may not perform well (Amico *et al.*, 2019). In fact, in a study in Finland, the probability of coronary heart disease was found to be similar between sedentary and very active men but for moderately active men, this probability was doubled (Taylor, 1983). More recently, Liu *et al.* (2022) found that the association of caffeine intake with serum uric acid in US adults was inverted U-shaped. Also, Li and Taylor (2002) employed a generalized partial linear additive model for the incidence but the "curse of dimensionality" is a problem.

To handle such practical situations, Amico *et al.* (2019) proposed the single-index model for $p(\boldsymbol{x}) = g(\gamma^\top \boldsymbol{x})$ which employs an unknown smooth link function $g(\cdot)$ estimated nonparametrically using kernel smoothing methods. Despite their ability to handle complex relationships between covariates and the response and to solve the curse of dimensionality problem faced by fully non-parametric models, single-index methods can be difficult to interpret, especially in cases where there are many covariates involved. It can be challenging to understand the individual contributions of important covariates. There are many situations where the researcher may be interested in knowing the effect of certain crucial covariates on the cure probability while also allowing flexible modeling of $p(\boldsymbol{x})$. In large clinical studies, some covariates may be "nuisance" factors while others are very relevant to the researcher. Thus, it would be advantageous if one could enjoy the interpretability of a logistic model as well as the flexibility of a single-index model (SIM). To this end, we introduce the

generalized partial-linear single-index model (GPLSIM) for $p(\boldsymbol{x})$ given by

$$p(\boldsymbol{x}) = p\left(\boldsymbol{x_1}, \boldsymbol{x_2}\right) = H\{g\left(\boldsymbol{\alpha}^\top \boldsymbol{x_1}\right) + \boldsymbol{\gamma}^\top \boldsymbol{x_2}\} \text{ with } \|\boldsymbol{\alpha}\| = 1 \tag{2}$$

where $H(\cdot)$ is a known monotonic function, $g(\cdot)$ is an unspecified link function, $\boldsymbol{\alpha}^\top \boldsymbol{x_1}$ is called the *index*, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are coefficient vectors and the covariate vector,$\boldsymbol{x}$, is possibly split into $\boldsymbol{(x_1, x_2)}$. The restriction $\|\boldsymbol{\alpha}\| = 1$ is required for identifiability (see section 2.1).

GPLSIMs have numerous advantages and have been used in several contexts in the literature. As noted before, they were proposed to solve the problem of linearity and handle model misspecification. In addition, the GPLSIM is a natural extension to the SIM by allowing discrete covariates to be modeled in the linear term (Wang and Cao, 2018). The decision to include a covariate in the single-index term or the linear term can solely be a result of domain theory in the field of use (Härdle *et al.*, 2004). In their partially linear single-index proportional hazard model, Shang *et al.* (2013) included in the linear component, covariates of interest to yield easily interpreted results and a single-index component to effectively adjust for multiple confounders. Carroll *et al.* (1997) first introduced the GPLSIM and employed kernel smoothing methods for estimating parameters in the maximum quasi likelihood methodology. Yu and Ruppert (2002), Yu *et al.* (2017) and Wang and Cao (2018) advocate for the use of splines, which are known to be stable and computationally expedient. We also advocate for the use of P-splines likelihood estimation for GPLSIM, as investigated by Yu *et al.* (2017). Our reason is purely for ease of implementation of our methods using already existing R software packages such as `mgcv` (Wood, 2012).

In survival analysis literature, the Cox PH (Cox, 1972) model has been extensively used for the latency sub-model. This proportional hazards assumption is usually violated in practice (Zeng *et al.*, 2022) but is frequently used since it forces

the hazard rate to stay within its natural boundaries (Aalen *et al.*, 2008) and has desirable theoretical properties (Lin and Ying, 1994). The additive hazard model of Lin and Ying (1994) has been suggested as a useful alternative in cases where the PH assumption is violated. As a result, we propose the additive hazard model for the latency. Our goal is to provide the user with the tools to fit the additive hazards mixture cure model while relying on the existing software packages in R. Our work is partly motivated by the `smcure` package in R and will serve as an alternative to fitting mixture cure models when one suspects the PH assumption for the uncured is violated.

In addition to the medical context where survival data analysis is prominent, another field that can greatly benefit from the proposed method is finance. Specifically, in credit risk modeling, researchers often aim to simultaneously model the probability of a customer defaulting and the time to default. In such cases, the method we propose can be highly valuable. The generalized partially linear single-index model (GPLSIM) can effectively handle the complex relationships involved in modeling the probability of default. It provides flexibility and accommodates various factors that influence default probabilities. On the other hand, the additive hazard model serves as an alternative to the Cox proportional hazards (PH) model when modeling the time to default. This allows for more accurate and comprehensive analysis of credit risk. By combining the GPLSIM and additive hazard models, researchers in finance can gain a deeper understanding of credit risk dynamics and make more informed decisions.

The proposed method offers a powerful tool for analyzing default probabilities and time to default simultaneously, enhancing the accuracy and reliability of credit risk models. The rest of the article is organized as follows: In Section 2, we describe our proposed model, and state the results for the identifiability of the model, including the penalized maximum likelihood based estimation procedure and our proposed

algorithm. In Section 3, we present numerical studies to illustrate the finite sample performance of the proposed estimator. A brief discussion is given on the issue of spline smoother selection. We provide a real-world data example in Section 4 and finally conclude the study in Section 5.

2. THE MODEL AND ESTIMATION

Let T be the random variable representing the time until the event of interest occurs. T is subject to random right censoring. We observe $Y = \min(T, C)$, the follow-up time, and $\Delta = I(T \leq C)$ where C is the random censoring time and $I(\cdot)$ is the indicator variable. We assume also that T and C are independent given covariates $(\boldsymbol{X}^\top, \boldsymbol{Z}^\top)^\top$ of dimension d and q respectively. The observed data consists of $\mathcal{D} = \{(Y_i, \Delta_i, \boldsymbol{X_i}, \boldsymbol{Z_i})\ (i = 1, \ldots, n)\}$ which we assume to be n independent and identically distributed (i.i.d) realizations of $\{(Y, \Delta, \boldsymbol{X}, \boldsymbol{Z})\}$. Suppose in a population the survival function is characterized by the mixture cure model in (1). We assume that $p(\boldsymbol{x})$ has the GPLSIM form described in (2). We allow the d dimensional set of covariates, $\boldsymbol{X}$, to be possibly split into: $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2})$ with the sum of dimensions equal to d. We consider the semiparametric additive hazard of Lin and Ying (1994) for the latency, with hazard function given by:

$$\lambda_u(t \mid \boldsymbol{z}) = \lambda_0(t) + \boldsymbol{\beta}^\top \boldsymbol{z} \tag{3}$$

where $\lambda_o(t)$ is the baseline hazard and $\boldsymbol{\beta}$ the coefficient vector of parameters associated with covariate $\boldsymbol{Z}$. The proper conditional survival function is of the form

$$S_u(t \mid \boldsymbol{z}) = \exp\left\{-\Lambda_0(t) - \int_0^t \boldsymbol{\beta}^\top \boldsymbol{z}\ du\right\} \tag{4}$$

with $\Lambda_0(t)$, a totally unspecified cumulative baseline hazard, and $S_u(t \mid \boldsymbol{z})$ is a proper survival function because as $t \to \infty$, $S_u(t \mid \boldsymbol{z}) \to 0$.

2.1. IDENTIFIABILITY OF MODEL

For model parameters to possibly be estimated, they must be able to be identified in a unique way. The union of the identifiability conditions for the mixture cure model, the GPLSIM, and the additive hazard model yields the set of identifiability conditions for our model. The identifiability conditions for the GPLSIM are essentially those captured in Carroll *et al.* (1997) and Li and Lu (2018):

1. The function g is differentiable and not constant on the support $\boldsymbol{\alpha}^\top \boldsymbol{x_1}$.
2. There is no intercept term in $\boldsymbol{\alpha}^\top \boldsymbol{x_1}$.
3. $g(0) = 0$ and $\|\boldsymbol{\alpha}\| = 1$ with the first nonzero element being positive.

For the mixture cure model we impose the following (Amico and Van Keilegom, 2018):

1. The cure threshold, $\tau < \infty$, exists such that $T > \tau \iff T = \infty$, and $P(C > \tau \mid \boldsymbol{X}, \boldsymbol{Z}) > 0$ for almost all $\boldsymbol{X}$ and $\boldsymbol{Z}$.
2. For all $\boldsymbol{x}, 0 < p(\boldsymbol{x}) < 1$.

For the additive hazard model, there is no intercept term in $\boldsymbol{\beta}$.

2.2. ESTIMATION AND THE EM ALGORITHM

The (log) likelihood (5) in classical survival analysis involves contributions from censored and uncensored groups. The censored contribute through the survival function, while the uncensored contribute through the density function, as follows:

$$\ell(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} \left[p\left(\boldsymbol{X}_i\right) f_u\left(Y_i \mid \boldsymbol{Z}_i\right) \right]^{\Delta_i} \times \prod_{i=1}^{n} \left[1 - p\left(\boldsymbol{X}_i\right) + p\left(\boldsymbol{X}_i\right) S_u\left(Y_i \mid \boldsymbol{Z}_i\right) \right]^{1-\Delta_i} \quad (5)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \lambda_0, g)^\top$ and $f_u\left(Y_i \mid \boldsymbol{Z}_i\right) = \lambda_u\left(Y_i \mid \boldsymbol{Z}_i\right) S_u\left(Y_i \mid \boldsymbol{Z}_i\right)$ which depends on the unobserved uncured status B. Thus the parameters cannot be estimated yet. The likelihood in (5), assumes an equal contribution from the censored but cured and

the censored but uncured, a rare case in the presence of a cure fraction (Amico *et al.*, 2019). Sy and Taylor (2000) used the EM algorithm of Dempster *et al.* (1977) to handle the partially observed B. Similarly we used the EM to handle B. To that end the complete-data likelihood (6) is defined as:

$$\begin{aligned} L_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ p\left(\boldsymbol{X}_i\right) S_u\left(Y_i \mid \boldsymbol{Z}_i\right) \right\}^{B_i(1-\Delta_i)} \prod_{i=1}^{n} \left\{ 1 - p\left(\boldsymbol{X}_i\right) \right\}^{(1-B_i)(1-\Delta_i)} \\ \times \prod_{i=1}^{n} \left\{ p(\boldsymbol{X}_i) f_u\left(Y_i \mid \boldsymbol{Z}_i\right) \right\}^{B_i\Delta_i} . \end{aligned} \tag{6}$$

In the likelihood function presented above, there are contributions from three distinct groups: *the censored and uncured*, *the censored and cured*, and the *uncensored and uncured*. Simplifying (6) gives:

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ p\left(\boldsymbol{X}_i\right)^{B_i} S_u\left(Y_i \mid \boldsymbol{Z}_i\right)^{B_i} \lambda_u\left(Y_i \mid \boldsymbol{Z}_i\right)^{B_i\Delta_i} \right\} \prod_{i=1}^{n} \left\{ 1 - p\left(\boldsymbol{X}_i\right) \right\}^{(1-B_i)} . \tag{7}$$

Taking log and simplifying further results in:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \left\{ B_i \log p\left(\boldsymbol{X}_i\right) + \left(1 - B_i\right) \log\left(1 - p\left(\boldsymbol{X}_i\right)\right) \right\} \\ &\qquad + \sum_{i=1}^{n} B_i \left\{ \Delta_i \log \lambda_u\left(Y_i \mid \boldsymbol{Z}_i\right) + \log S_u\left(Y_i \mid \boldsymbol{Z}_i\right) \right\} \\ &= \ell_1(\boldsymbol{\theta}) + \ell_2(\boldsymbol{\theta}). \end{aligned} \tag{8}$$

The "likelihood" contains a partially observed cure status B, which needs to be estimated together with the parameters of interest in an EM-like algorithm. To that end, in the E-step, of the EM algorithm $E(B|\mathcal{D}, \boldsymbol{\theta}^{m-1})$ is given by:

$$\begin{aligned} E(B|\mathcal{D}, \boldsymbol{\theta}^{m-1}) &= 1 \times P(T < \infty|\mathcal{D}, \boldsymbol{\theta}^{m-1}) + 0 \times P(T = \infty|\mathcal{D}, \boldsymbol{\theta}^{m-1}) \\ &= P(T < \infty|\mathcal{D}, \boldsymbol{\theta}^{m-1}) \\ &= (1 - \Delta_i)\, P\left(B_i = 1 \mid Y_i, \Delta_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{m-1}\right) + \Delta_i \end{aligned} \tag{9}$$

where $\mathcal{D}$ is the observed data and $\boldsymbol{\theta}^{m-1}$ is the "current" estimates of the parameters at the mth iteration. The probability expression in last equation can be expressed as:

$$P\left(B_i = 1 \mid Y_i, \Delta_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}^{m-1}\right) = \frac{p^{(m-1)}(\boldsymbol{X}_i) S_u^{(m-1)}\left(Y_i \mid \boldsymbol{Z}_i\right)}{1 - p^{(m-1)}(\boldsymbol{X}_i) + p^{(m-1)}(\boldsymbol{X}_i) S_u^{(m-1)}\left(Y_i \mid \boldsymbol{Z}_i\right)}. \tag{10}$$

Substituting (10) into (9) results in:

$$\begin{aligned} E(B|\mathcal{D}, \boldsymbol{\theta}^{m-1}) &= \Delta_i + (1 - \Delta_i) \left[\frac{p^{(m-1)}(\boldsymbol{X}_i) S_u^{(m-1)}\left(Y_i \mid \boldsymbol{Z}_i\right)}{1 - p^{(m-1)}(\boldsymbol{X}_i) + p^{(m-1)}(\boldsymbol{X}_i) S_u^{(m-1)}\left(Y_i \mid \boldsymbol{Z}_i\right)}\right] \\ &:= W_i^m. \end{aligned} \tag{11}$$

As with other mixture cure models, we can maximize $\ell_1(\boldsymbol{\theta})$ and $\ell_2(\boldsymbol{\theta})$ separately. For $\ell_1(\boldsymbol{\theta})$, the triple $(g, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ need to be estimated. g is estimated using a penalized splines (P-splines) approach (Eilers and Marx, 1996) with a truncated power bases splines (Yu *et al.*, 2017). The `mgcv` R package (Wood, 2012) offers a convenient way with different basis options to estimate the index function g. Other spline bases, such as the thin plate regression spline (Wood, 2003) or cubic regression spline, can be employed especially for smaller sample sizes, as they tend to give the best mean squared error performance at the expense of longer computational time (Wood, 2012).

g can represented as a linear combination of truncated power spline bases:

$$\begin{aligned} g(u) &= \varphi_0 + \varphi_1 u + \cdots + \varphi_p u^p + \sum_{k=1}^{K} \varphi_{p+k} \left(u - v_k\right)_+^p \\ &= \boldsymbol{\varphi}^\top \mathbf{S}(u) \end{aligned} \tag{12}$$

where $\mathbf{S}(u) = \left\{1, u, \ldots, u^p, (u - v_1)_+^p, \ldots, (u - v_K)_+^p\right\}$ are spline bases with K knots placed at $(v_1, \ldots, v_K)$, and $\boldsymbol{\varphi} = (\varphi_0, \varphi_1, \ldots, \varphi_{p+K})^\top$ are spline coefficients to be estimated. For our GPLSIM, $p\left(\boldsymbol{X}_i\right) = H\left\{\boldsymbol{\varphi}^\top \mathbf{S}\left(\boldsymbol{\alpha}^\top \boldsymbol{X}_{1i}\right) + \boldsymbol{\gamma}^\top \boldsymbol{X}_{2i}\right\}$. Replacing the single-index term, $g(\cdot)$, with its spline representation in (12) we obtain:

$$\begin{aligned} \ell_1(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\varphi}) = \sum_{i=1}^{n} \Big\{ & B_i \log H\left\{\boldsymbol{\varphi}^\top \mathbf{S}\left(\boldsymbol{\alpha}^\top \boldsymbol{X}_{1i}\right) + \boldsymbol{\gamma}^\top \boldsymbol{X}_{2i}\right\} \\ & + (1 - B_i) \log\left(1 - H\left\{\boldsymbol{\varphi}^\top \mathbf{S}\left(\boldsymbol{\alpha}^\top \boldsymbol{X}_{1i}\right) + \boldsymbol{\gamma}^\top \boldsymbol{X}_{2i}\right\}\right)\Big\}. \end{aligned} \tag{13}$$

This estimation is done iteratively in an EM algorithm and the estimates $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\varphi}})$ are the solution to:

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\varphi}}) = \underset{\substack{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\varphi} \\ \|\boldsymbol{\alpha}\| = 1}}{\arg\max}\, \ell_1(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\varphi}). \tag{14}$$

For the estimation of the additive hazard model, we proceed in the manner of Lin and Ying (1994). Notice that the estimate $\ell_2(\cdot)$ is the likelihood of the additive hazard model with B_i serving as a weight. Using counting process notation, the counting process $N_i(t)$ which records the number of events up to time t for individual can be decomposed so that

$$N_i(t) = M_i(t) + \int_0^t R_i(u) d\Lambda(t \mid \boldsymbol{Z}) \tag{15}$$

where $M_i(\cdot)$ is a martingale, $R_i(t)$ is the at-risk process (1 if the individual is at risk at time $t, 0$ otherwise). The entire second term is the compensator of $N_i(t)$. The intensity function for the $N_i(t)$ is given by:

$$\begin{aligned} dM_i(t) &= B_i dN_i(t) - B_i R_i(t) d\Lambda(t \mid \boldsymbol{Z}) \\ &= B_i dN_i(t) - B_i R_i(t)\{d\Lambda_0(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}(t) dt\}. \end{aligned} \tag{16}$$

Note that the extra B_i is a weight that captures whether or not the individual belongs to the susceptible group. Applying the martingale property, $E[dM(t) \mid \mathcal{F}(t)] = 0$, where $\mathcal{F}(t)$ is the filtration up to time t, we have for all individuals:

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{i=1}^{n} dM_i(t) \\ &= \sum_{i=1}^{n} B_i \left[dN_i(t) - R_i(t) \left\{ d\Lambda_0(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}_i(t) dt \right\} \right]. \end{aligned} \tag{17}$$

Solving the above estimating equation for fixed $\boldsymbol{\beta}$ and known B, we obtain an estimator for $\Lambda_0(\cdot)$ given by:

$$\hat{\Lambda}_0(t \mid \boldsymbol{\beta}) = \int_0^t \frac{\sum_{i=1}^n B_i \left\{ dN_i(u) - R_i(u) \boldsymbol{\beta}^\top \boldsymbol{Z}_i(u) du \right\}}{\sum_{i=1}^n B_i R_i(u)}. \tag{18}$$

If the baseline hazard is known, $\boldsymbol{\beta}$ is estimated from the following estimating function (assuming cure status, B is known)

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\infty B_i \boldsymbol{Z}_i(t) \left\{ dN_i(t) - R_i(t) d\hat{\Lambda}_0(t \mid \boldsymbol{\beta}) - R_i(t) \boldsymbol{\beta}^\top \boldsymbol{Z}_i(t) dt \right\}. \tag{19}$$

Setting (19) equal to the $q \times 1$ vector $\mathbf{0}$ produces the estimating equation for $\boldsymbol{\beta}$ whose solution is $\hat{\boldsymbol{\beta}} = \hat{\mathbf{A}}^{-1} \hat{\mathbf{D}}$ with

$$\hat{\mathbf{A}} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\infty B_i R_i(t) \left\{ \boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t) \right\}^{\otimes 2} dt \tag{20}$$

and

$$\hat{\mathbf{D}} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty} B_i\left\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\right\} dN_i(t) \tag{21}$$

where $\bar{\boldsymbol{Z}}(t) = \sum_{i=1}^{n} B_i R_i(t)\boldsymbol{Z}_i(t) / \sum_{i=1}^{n} B_i R_i(t)$. Consequently, a natural estimator for the survival function of the uncured sub-population is

$$\hat{S}_u(t \mid \boldsymbol{z}) = \exp\left\{-\hat{\Lambda}_0(t \mid \hat{\boldsymbol{\beta}}) - \int_0^t \hat{\boldsymbol{\beta}}^{\top}\boldsymbol{z}(u)du\right\}. \tag{22}$$

2.3. A COMPUTATIONAL ALGORITHM

In this section, we present our computational algorithm (see Algorithm 1), which we refer to as the Expectation Maximization & Estimation (EME) algorithm. The algorithm combines the traditional Expectation (E) step with maximum likelihood estimation for the GPLSI model used in the incidence component, and estimating equations for the additive hazard model used in the latency component. The EME algorithm consists of the following steps:

1. Initialize the model parameters.

2. Perform the E-step: Estimate the latent variables and update the incomplete data likelihood.

3. Perform the ME-step: Update the model parameters by maximizing (M) the complete data likelihood and solving estimating equations (E) respectively.

4. Repeat steps 2 and 3 until convergence is achieved or a maximum number of iterations is reached.

By iteratively updating the model parameters based on the observed data and latent variables, the EME algorithm effectively estimates the parameters of the mixture cure model.

Algorithm 1 Expectation, Maximization & Estimation (EME)

Input: $\mathcal{D} = (y, \delta, \boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{z})$, diff= 1000, eps = 10^{-7}, emmax = 100.

initialize parameters for $\boldsymbol{\alpha_0}, \boldsymbol{\gamma_0}$ using standard logistic regression.

$p_0 \leftarrow \frac{exp(\alpha_0^\top \boldsymbol{x_1} + \gamma_0^\top \boldsymbol{x_2})}{1+exp(\alpha_0^\top \boldsymbol{x_1} + \gamma_0^\top \boldsymbol{x_2})}$ ▷ initialize p

$w_0 \leftarrow \delta$ ▷ initialize weights

fit the GPLSI model with w_0 as response

obtain $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, p

fit the additive hazard (AH) model with data $(y, \delta, \boldsymbol{z})$ and weight, w_0

obtain $\boldsymbol{\beta}$, baseline survival s and s_{u}

$y_{\max} \leftarrow \max(y)$ where $\delta = 1$

i $\leftarrow 1$

while (diff $<$ eps & $i <$ emmax) **do**

  $w \leftarrow \delta + (1-\delta) \times \frac{(p \times s_{\text{u}})}{(1-p+p \times s_{\text{u}})}$ ▷ update weights

  Ensure $w \leftarrow 0$ when $\delta = 0$ & $y > y_{\max}$ ▷ zero tail constraint

  fit GPLSI, AH models and obtain $\boldsymbol{\alpha}_{\text{update}}$, $\boldsymbol{\gamma}_{\text{update}}$, p_{update}, $\boldsymbol{\beta}_{\text{update}}$, s_{update}, $s_{\text{u,update}}$

  diff $= \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\text{update}}\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_{\text{update}}\|^2 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_{\text{update}}\|^2 + \|s_{\text{u}} - s_{\text{u,update}}\|^2$

  $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}_{\text{update}}$

  $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}_{\text{update}}$

  $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma}_{\text{update}}$

  $s_{\text{u}} \leftarrow s_{\text{u,update}}$

end while

Despite the conclusion from Yu *et al.* (2017)'s study using P-splines on the GPLSIM, we found the maximum likelihood criteria (Anderssen and Bloomfield, 1974) for selecting the smoothing parameter to be favorable compared to the Generalized Cross-Validation criteria of Craven and Wahba (1978).

3. SIMULATION STUDIES

In this section, two simulation studies are conducted to evaluate the performance of our model. More precisely we perform simulation studies for the following objectives:

1. To evaluate the finite sample performance of the model under different right censoring loads. We consider how well the uncure or cure probabilities are estimated and how well parameters are estimated.

2. To study briefly the sensitivity of our estimates with varying sample sizes.

3. To compare our generalized partially linear single-index additive hazard (GPLSI-AH) mixture cure model with the Single-Index/Cox (SIC) cure model of Amico *et al.* (2019) and the logistic/Cox (LC) cure model in the `smcure` R package. Particularly, we compare incidence sub-models under the two scenarios: when the ground truth is the logistic structure and the Sine Bump model of Carroll *et al.* (1997).

The data for the incidence sub-model is generated according to two scenarios each with a different link function. The first scenario is the logistic link function of the form

$$p(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{\gamma}^\top \boldsymbol{x}\right)}{1 + \exp\left(\boldsymbol{\gamma}^\top \boldsymbol{x}\right)}. \tag{23}$$

We considered four independent covariates: X_1 and X_2 drawn from a standard normal distribution and $X_3 \sim \text{Ber}(0.3)$ and $X_4 \sim \text{Ber}(0.6)$. The parameters of the model were set to $\boldsymbol{\gamma} = (-1.5, 0.5, 2.3, -1.3)$ with 1.4 as the intercept. This is essentially the same parameter setting used in Amico *et al.* (2019). The second scenario is the Sine

Bump model (Carroll *et al.*, 1997) of the form

$$h(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sin\left\{\pi\left(\boldsymbol{\alpha}^\top \boldsymbol{x_1} - A\right)/(B-A)\right\} + \gamma^\top \boldsymbol{x_2} \tag{24}$$

so that

$$p(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{\exp(\sin\left\{\pi\left(\boldsymbol{\alpha}^\top \boldsymbol{x_1} - A\right)/(B-A)\right\} + \gamma^\top \boldsymbol{x_2})}{1 + \exp(\sin\left\{\pi\left(\boldsymbol{\alpha}^\top \boldsymbol{x_1} - A\right)/(B-A)\right\} + \gamma^\top \boldsymbol{x_2})} \tag{25}$$

where $\boldsymbol{\alpha} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$ with each covariate X_1, X_2, X_3 from independent uniform in $(0,1)$; $\gamma = 0.3$; X_4 is a binary predictor with 1 for even observations and 0 otherwise; $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$.

For the latency sub-model, we considered one covariate $Z \sim \text{Ber}(0.6)$ independent of X_1, X_2, X_3, X_4. The survival times T were generated by solving the following equation:

$$-\log(\text{U}) = qT + \beta^\top Z \tag{26}$$

where $\text{U} \sim U(0,1)$ and parameters $\beta = 1.5$ and $q = 3$. The censoring time, was generated from an exponential distribution with probability density function, $f(t) = \lambda_c \exp(-\lambda_c t)$, independent of $(\boldsymbol{X}, Z, T)$. To access the impact of different levels of censoring on the performance of the model, we considered three levels of censoring: low, mid, and high, with $\sim 5\%$ increase from the previous. For the two scenarios, we have summarized the parameter settings in the Table 1

Table 1. Parameters for the incidence and latency sub-model, the cure fraction and the censoring rates

Scenario	Incidence	Latency	Cure rate	Censoring level	Censoring rate	λ_c
Logistic	$X_1, X_2 \sim N(0,1)$	$\beta = 1.5$	32%	low	34%	0.15
	$X_3 \sim \text{Ber}(0.3)$, $X_4 \sim \text{Ber}(0.6)$	$Z \sim \text{Ber}(0.6)$	32%	mid	40%	0.5
	$\gamma = (-1.5, 0.5, 2.3, -1.3)$		32%	high	45%	0.9
Sine Bump	$X_1, X_2, X_3 \sim U(0,1)$	$\beta = 1.5$	32%	low	34%	0.15
	$X_4 = 0$ for odd observations and 1 for even	$Z \sim \text{Ber}(0.6)$	32%	mid	40%	0.6
	$\boldsymbol{\alpha} = (\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$, $\gamma = 0.3$		32%	high	45%	0.92

It is worth mentioning that too heavy censoring may lead to having less observations beyond the cure threshold $Y_{(r)}$ and hence we do not consider substantially heavy censoring. For each scenario and censoring rate we consider 4 samples sizes, $n = \{250, 500, 1000, 2500\}$, to yield 24 total settings (2 model scenarios, 3 censoring rates, 4 sample sizes). We fit the GPLSI-AH, SIC and LC cure models for each dataset. For each setting we obtained 500 replicates (or datasets).

For identification of the mixture cure model, we followed the suggestion of Taylor (1995) throughout this paper. In our novel EME algorithm, two sets of initial values were sought, one for the single index terms and one for the partial linear terms. Our approach does not require starting values for the latency sub-model. To evaluate the performance of our model, we consider two metrics: the average square error (ASE) for the incidence sub-model given by (27) and the bias and variance of $\hat{\beta}$ for the latency sub-model. The ASE is defined as:

$$ASE(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n}\left\{H\left(g\left(\boldsymbol{\alpha}^\top \boldsymbol{x_{1i}}\right) + \boldsymbol{\gamma}^\top \boldsymbol{x_{2i}}\right) - H\left(\hat{g}\left(\hat{\boldsymbol{\alpha}}^\top \boldsymbol{x_{1i}}\right) + \hat{\boldsymbol{\gamma}}^\top \boldsymbol{x_{2i}}\right)\right\}^2. \qquad (27)$$

The box plots for the ASE of $p(x)$ for all three models are shown in Figures 1–8. As expected, the LC cure model outperforms the SIC cure model when the ground truth is logistic regression, irrespective of the sample size. On the other hand, the

ASE of $p(x)$ for the GPLSI-AH model is roughly the same as for the LC cure model, particularly for low and mid censoring levels, but a little higher for high censoring levels. It is worth noting that the continuous covariates entered the GPLSI-AH model via the single index term, while the categorical covariates entered via the partial linear term. When the true model is the Sine Bump model, the GPLSI-AH performs best, followed by the SIC and LC models, as expected. It is worth mentioning that, though the errors reduce across all three models with increased sample size, we observed more variability in the errors for the SIC model. The SIC and LC cure models seem to suffer heavily from model misspecification. When there is a doubt that the relationship is logistic, specifying a logistic model for the incidence is costly. In addition, when there is doubt about assuming a SIC model, we advocate for using the GPLSI-AH model due to its minimal cost in error due to misspecification. Generally, for all three models, the precision of the estimates decreases as the censoring rate increases. As pointed out in much of the literature on mixture cure models, as the censoring rate gets farther from the cure rate, more observations will have W_i^m between 0 and 1 with a substantial amount of them close to zero. This increases the uncertainty in the estimation of $p(x)$. Furthermore, with assuming the additive hazard model for the latency, this uncertainty is especially high because, W_i^m plays the role a as weight in the semiparametric additive hazard model fitting process at each iteration and can heavily affect parameter estimates obtained.

As was previously mentioned, one of the benefits of GPLSI-AH is that the variables that enter the model through the partial linear component can be interpreted just as the LC model when the real relationship is logistic. The coefficient estimates of our partial linear terms are close to those of the LC model when the ground truth is logistic, with the exception of $\hat{\gamma}_1$ for $n = 250$ and high censoring. Similarly, the precision of the estimates decreases with higher censoring and increases with larger sample sizes, with GPLSI-AH estimates benefiting somewhat more from the

increase in sample size. Overall, GPLSI-AH provides a reliable and accurate method for modeling partial linear relationships, particularly when the true relationship is logistic. However, researchers should be aware that high levels of censoring and smaller sample sizes may lead to decreased precision in the estimates (see Table 2).

Next, we also compare the partial linear term in the GPLSI-AH with that in the LC model for the Sine Bump model (see Table 3). The results are comparable with the GPLSI-AH estimates, benefiting more from an increase in sample size. GPLSI-AH estimates seem to be affected more by the increase in censoring compared to the LC model. This is perhaps due to the fact that the estimate W_i^m will be equal to $\left\{p^{(m-1)}\left(\boldsymbol{X}_i\right) S_u^{(m-1)}\left(Y_i \mid Z_i\right)\right\} / \left\{1-p^{(m-1)}\left(\boldsymbol{X}_i\right)+\ p^{(m-1)}\left(\boldsymbol{X}_i\right) S_u^{(m-1)}\left(Y_i \mid Z_i\right)\right\}$ and, will thus depend on $S_u^{(m-1)}$ and $p^{(m-1)}$. The dependence of the estimate, W_i^m on $S_u^{(m-1)}$ is more pronounced, which in turn heavily relies on the initialized values for W_i as they serve as weights in the additive hazard model.

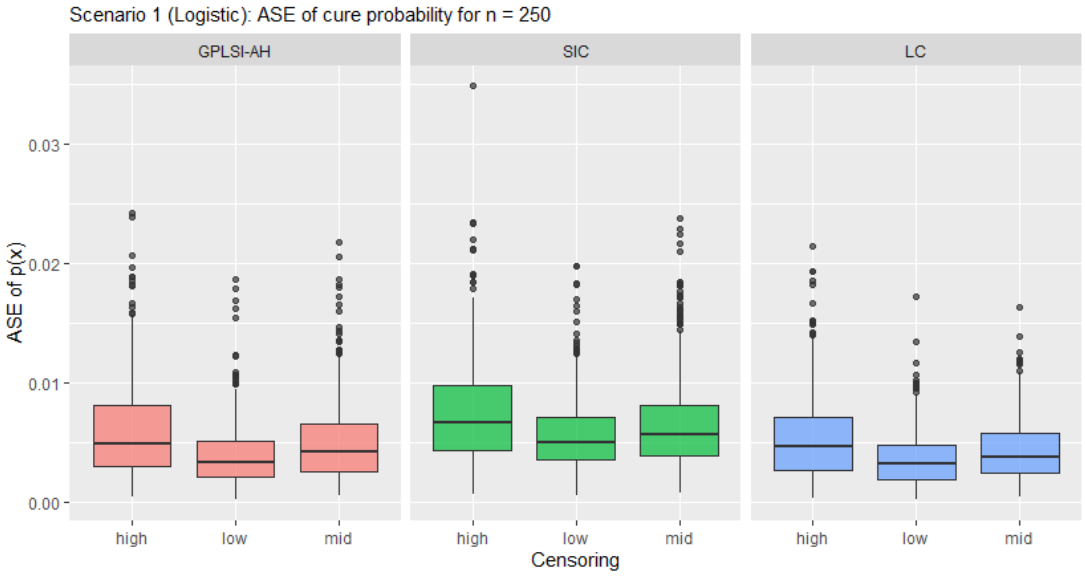

Figure 1. Boxplots of the Average Squared Error (ASE) for three models under the logistic scenario for $n = 250$.

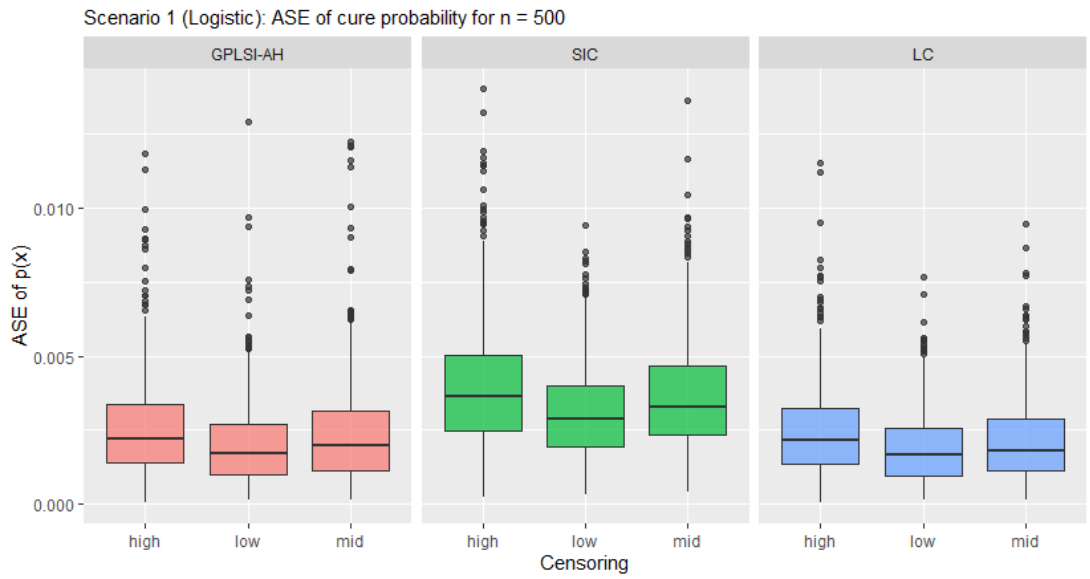

Figure 2. Boxplots of the Average Squared Error (ASE) for three models under the logistic scenario for $n = 500$.

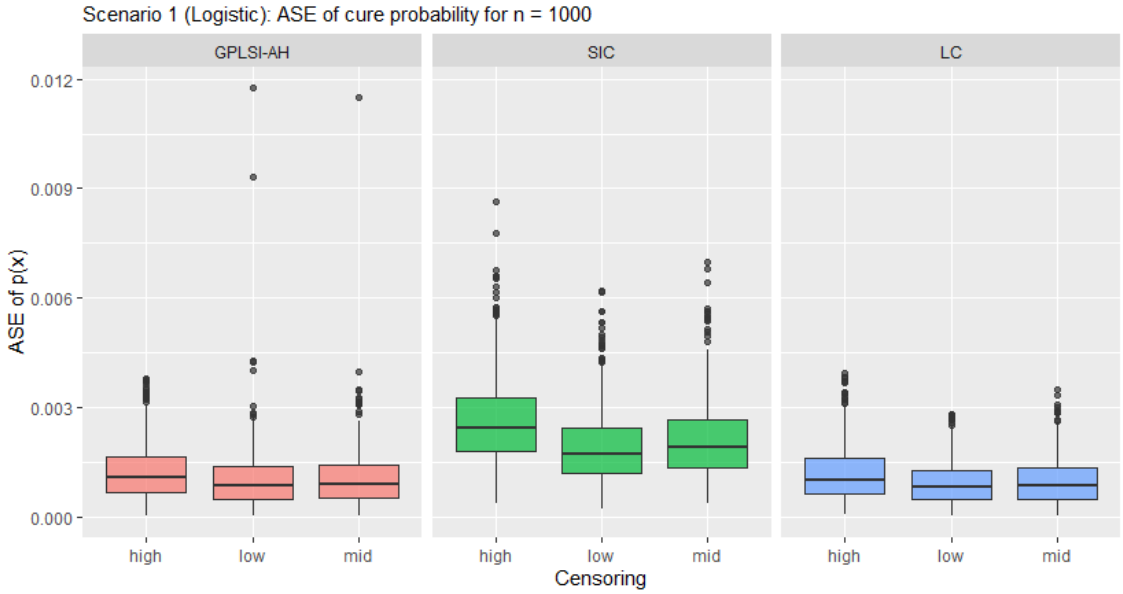

Figure 3. Boxplots of the Average Squared Error (ASE) for three models under the logistic scenario for $n = 1000$.

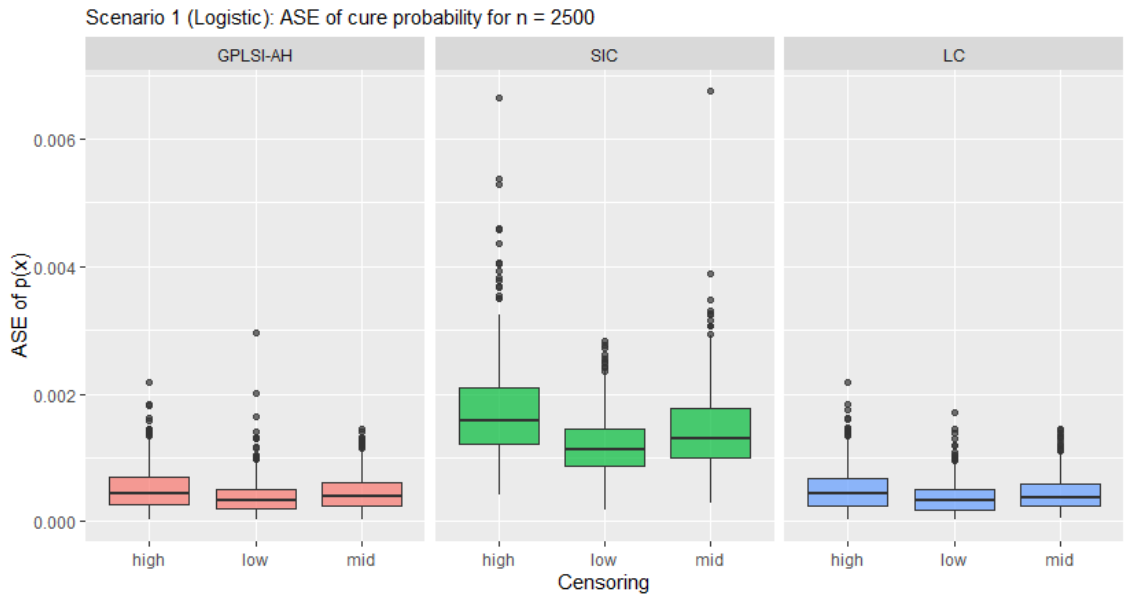

Figure 4. Boxplots of the Average Squared Error (ASE) for three models under the logistic scenario for $n = 2500$.

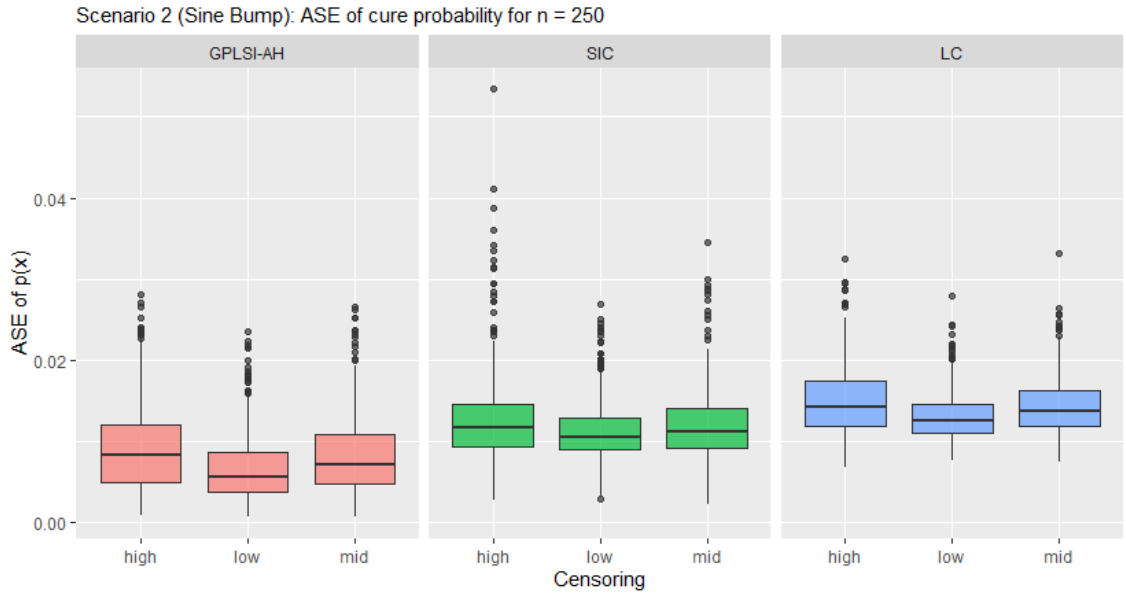

Figure 5. Boxplots of the Average Squared Error (ASE) for the three models under the sine bump scenario for $n = 250$.

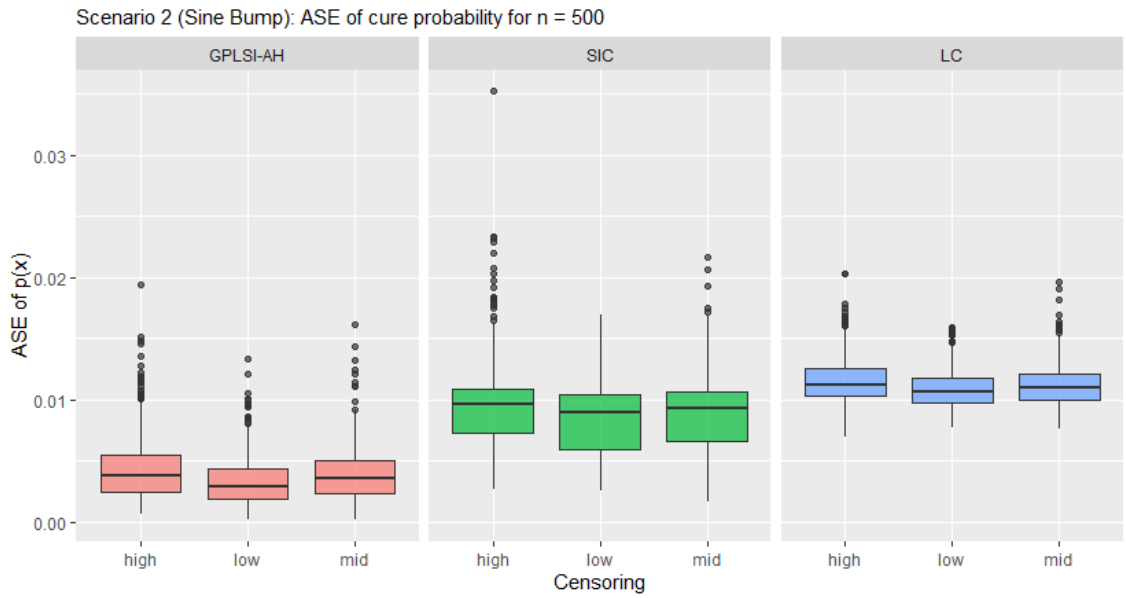

Figure 6. Boxplots of the Average Squared Error (ASE) for the three models under the sine bump scenario for $n = 500$.

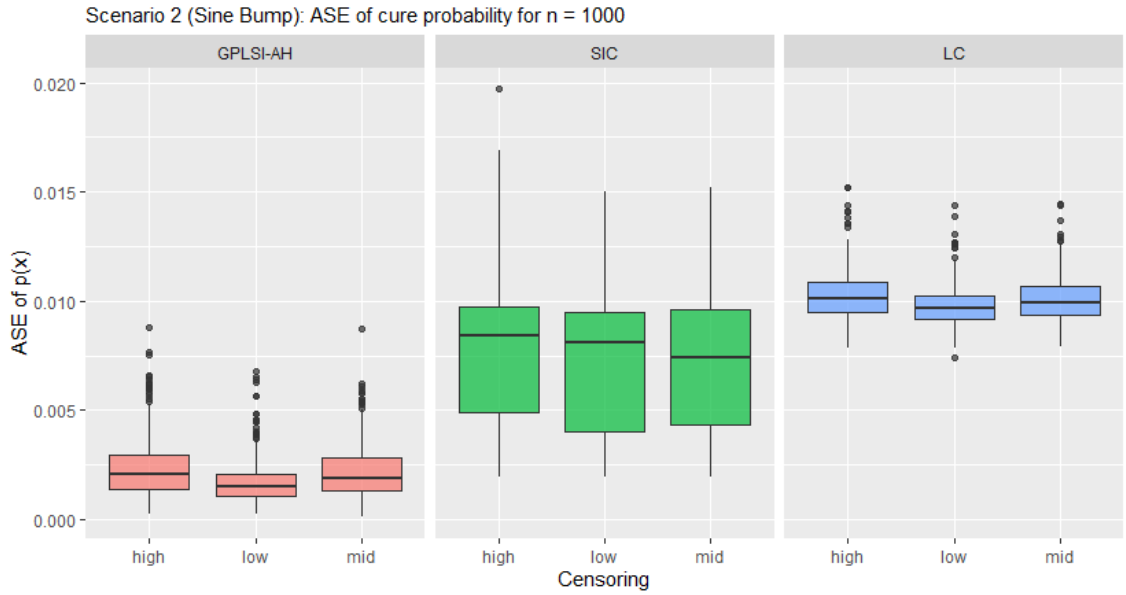

Figure 7. Boxplots of the Average Squared Error (ASE) for the three models under the sine bump scenario for $n = 1000$.

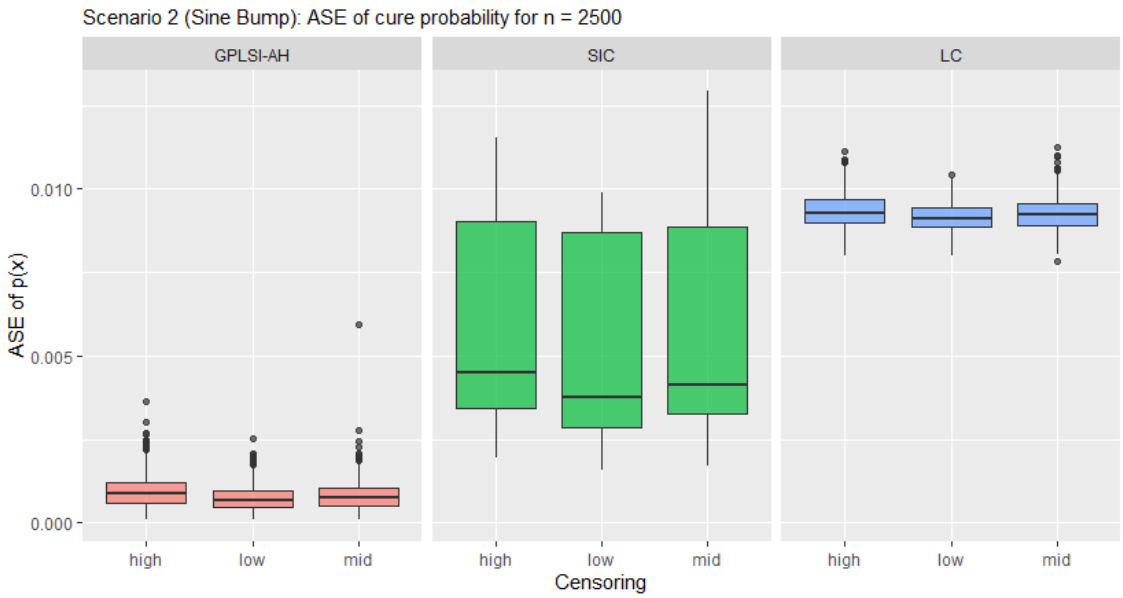

Figure 8. Boxplots of the Average Squared Error (ASE) for the three models under the sine bump scenario for $n = 2500$.

For the latency sub-model, where the true model is the semi-parametric additive hazard model, we assess the performance of parameters using the bias and variance of $\hat{\beta}$ (see Table 4). The LC and SIC model parameter estimates for the latency sub-model are also presented here to demonstrate how biased inferences can be made when the model is misspecified, as they both assume the Cox PH model. Currently, to the best of our knowledge, the software packages for mixture cure model fitting assumes the Cox PH model, which has been shown to be hardly validated in practice. The bias of $\hat{\beta}$ is small for all censoring schemes and sample sizes. It does not appear that the estimate of β is affected as much by the true nature of the cure probability as one would expect. In addition, the variance gets smaller with larger sample sizes. It is worth mentioning that the variance is notably higher for high censoring rates. This is due to the fact that a higher λ_c results in a higher rate of censorship, making the plateau less populated than the true cure rate. In such cases, W_i^m may vary between 0 and 1, and that can affect our parameter estimate for the additive hazard latency sub-model. To control for this variance, we propose using a

different set of initial values for W_i in situations when the censoring rate is high. When fitting the GPLSI-AH cure models to data with heavy censoring we propose modifying the strategy in this manner: set $W_o = 1$ when $\Delta = 1$. For the rest of W_0 randomly assign 1 or 0 based on a $W \sim \text{Ber}(p)$ where p is the censoring rate.

It is quite obvious that estimating a GPLSI-AH cure model is computationally expensive. It requires more computational time than the LC model. However, we observed the time required to fit a GPLSI-AH model to be reasonably close to the SIC model.

Table 2. Coefficient estimates and standard deviations (sd) of partial linear term from the GPLSI-AH compared to LC model under the three censoring schemes and four sample sizes where the true relationship is logistic.

		Low censoring		Mid censoring		High censoring	
n	PL term	estimate	sd	estimate	sd	estimate	sd
250	$\gamma_{1_{\text{GPLSI}}}$	2.475	0.525	2.432	0.629	3.201	5.083
	$\gamma_{1_{\text{LC}}}$	2.348	0.512	2.374	0.659	2.430	0.699
	$\gamma_{2_{\text{GPLSI}}}$	-1.338	0.439	-1.386	0.458	-1.396	0.464
	$\gamma_{2_{\text{LC}}}$	-1.346	0.471	-1.271	0.392	-1.400	0.529
n	PL term	estimate	sd	estimate	sd	estimate	sd
500	$\gamma_{1_{\text{GPLSI}}}$	2.359	0.450	2.403	0.407	2.391	0.435
	$\gamma_{1_{\text{LC}}}$	2.337	0.391	2.382	0.416	2.425	0.467
	$\gamma_{2_{\text{GPLSI}}}$	-1.332	0.293	-1.332	0.312	-1.319	0.318
	$\gamma_{2_{\text{LC}}}$	-1.328	0.272	-1.353	0.336	-1.349	0.341
n	PL term	estimate	sd	estimate	sd	estimate	sd
1000	$\gamma_{1_{\text{GPLSI}}}$	2.334	0.262	2.311	0.277	2.297	0.335
	$\gamma_{1_{\text{LC}}}$	2.329	0.241	2.332	0.299	2.304	0.311
	$\gamma_{2_{\text{GPLSI}}}$	-1.286	0.189	-1.297	0.204	-1.296	0.243
	$\gamma_{2_{\text{LC}}}$	-1.301	0.210	-1.299	0.201	-1.328	0.260
n	PL term	estimate	sd	estimate	sd	estimate	sd
2500	$\gamma_{1_{\text{GPLSI}}}$	2.329	0.167	2.320	0.180	2.341	0.204
	$\gamma_{1_{\text{LC}}}$	2.292	0.177	2.308	0.191	2.323	0.217
	$\gamma_{2_{\text{GPLSI}}}$	-1.306	0.134	-1.324	0.127	-1.294	0.140
	$\gamma_{2_{\text{LC}}}$	-1.313	0.119	-1.318	0.129	-1.306	0.147

Table 3. Coefficient estimates and standard deviations (sd) of partial linear term from the GPLSI-AH compared to LC model under the three censoring schemes and four sample sizes for the Sine Bump model.

		Low censoring		Mid censoring		High censoring	
n	PL term	estimate	sd	estimate	sd	estimate	sd
250	γ_{GPLSI}	0.315	0.317	0.304	0.341	0.319	0.371
	γ_{LC}	0.304	0.312	0.288	0.334	0.308	0.351
n	PL term	estimate	sd	estimate	sd	estimate	sd
500	γ_{GPLSI}	0.304	0.219	0.330	0.216	0.343	0.217
	γ_{LC}	0.294	0.216	0.314	0.208	0.328	0.214
n	PL term	estimate	sd	estimate	sd	estimate	sd
1000	γ_{GPLSI}	0.296	0.142	0.320	0.154	0.324	0.158
	γ_{LC}	0.287	0.135	0.309	0.151	0.311	0.153
n	PL term	estimate	sd	estimate	sd	estimate	sd
2500	γ_{GPLSI}	0.311	0.085	0.297	0.107	0.296	0.111
	γ_{LC}	0.298	0.084	0.286	0.106	0.283	0.110

Table 4. Bias and variance of $\hat{\beta}$ for the GPLSI-AH, SIC and LC cure models under the three censoring schemes and four sample sizes.

			Low censoring		Mid censoring		High censoring	
n	Scenario	Model	Bias	variance	Bias	variance	Bias	variance
250	logistic	GPLSI-AH	0.015	0.429	0.009	0.447	0.024	0.557
	logistic	SIC	-1.090	0.031	-1.095	0.033	-1.087	0.042
	logistic	LC	-1.092	0.031	-1.097	0.032	-1.092	0.041
	SB	GPLSI-AH	-0.012	0.425	0.052	0.572	0.022	0.572
	SB	SIC	-1.098	0.030	-1.079	0.041	-1.094	0.042
	SB	LC	-1.098	0.030	-1.079	0.041	-1.094	0.042
	Scenario	Model	Bias	variance	Bias	variance	Bias	variance
500	logistic	GPLSI-AH	0.019	0.193	-0.032	0.212	0.032	0.236
	logistic	SIC	-1.089	0.014	-1.102	0.016	-1.085	0.018
	logistic	LC	-1.089	0.014	-1.102	0.016	-1.087	0.018
	SB	GPLSI-AH	-0.000	0.212	0.012	0.257	0.035	0.286
	SB	SIC	-1.093	0.016	-1.094	0.019	-1.088	0.021
	SB	LC	-1.093	0.016	-1.095	0.019	-1.088	0.021
	Scenario	Model	Bias	variance	Bias	variance	Bias	variance
1000	logistic	GPLSI-AH	-0.017	0.097	0.006	0.102	0.012	0.116
	logistic	SIC	-1.101	0.008	-1.091	0.008	-1.090	0.009
	logistic	LC	-1.101	0.008	-1.092	0.008	-1.091	0.009
	SB	GPLSI-AH	-0.004	0.094	0.023	0.132	0.006	0.124
	SB	SIC	-1.096	0.007	-1.089	0.010	-1.094	0.010
	SB	LC	-1.096	0.007	-1.089	0.010	-1.094	0.010
	Scenario	Model	Bias	variance	Bias	variance	Bias	variance
2500	logistic	GPLSI-AH	0.006	0.036	-0.005	0.043	0.002	0.056
	logistic	SIC	-1.093	0.003	-1.096	0.003	-1.094	0.004
	logistic	LC	-1.093	0.003	-1.096	0.003	-1.095	0.004
	SB	GPLSI-AH	0.013	0.040	0.007	0.052	0.006	0.059
	SB	SIC	-1.091	0.003	-1.092	0.004	-1.094	0.005
	SB	LC	-1.091	0.003	-1.092	0.004	-1.094	0.005

4. REAL DATA EXAMPLE

We provide an example using a small-scale study on diabetic retinopathy to demonstrate the application of the proposed method. However, it is important to note that the true strength and advantages of the method are best observed in larger-scale studies, such as the Framingham Heart study (Carroll *et al.*, 1997) and the UK bank personal loans data (Dirick *et al.*, 2022). In the case of the Framingham Heart study, there is no specific time-to-event variable available for analysis, which limits the applicability of the proposed method in that particular context. Regarding the UK bank personal loans data, we encountered challenges in obtaining the data from the authors, preventing us from conducting an analysis using that dataset. However, we believe that the proposed method has potential utility in credit risk modeling and similar financial contexts, as it can simultaneously model the probability of default and the time to default.

The Diabetic Retinopathy Study Data: People with diabetes can get diabetic retinopathy, which is when the tiny blood vessels in their eyes get damaged. In the United States and many other developed countries, this can cause blindness, especially in people under 60 years old. In 1971, the Diabetic Retinopathy Study (DRS) was started to find out if laser treatment could keep people with diabetic retinopathy from going blind. Patients in the study had diabetic retinopathy in both eyes, but they could still see fairly well with both eyes. One eye of each patient was treated with a laser at random, while the other eye was not (Huster *et al.*, 1989). The original study followed 1,742 patients for several years. The study looked at how well laser treatment worked by seeing if the patients' vision got worse over time, to the point where they couldn't see better than 5/200 on two separate follow-up visits. The version of the subset of the dataset used in the analysis was obtained from

https://www.mayo.edu/research/documents/diabeteshtml/DOC-10027460/. This consists of $n = 197$ diabetic subjects (see Table 5 for variables used in this study).

The focus of this analysis is to determine the long- and short term effects of the covariates using mixture cure models. For instance, how does a subject's age affect the probability of losing their eye sight, and how does this affect how long it takes to lose your eye sight?

Figure 9 shows the Kaplan and Meier (1958) survival curve with a plateau. We performed a Maller & Zhou test for sufficient follow-up and found evidence of enough follow up to detect the presence of cured subjects. About 73% of the observations were censored, with 12% of observations on the plateau, giving enough reason to consider a cure model.

Next, we fit all models to the data (see Table 6). The standard errors of the parameter estimates for all models have been computed using bootstrapping based on 100 bootstrap samples. For the GPLSI-AH model, age and risk are entered through the single-index and others through the partial linear term. Following the approach of Amico *et al.* (2019), we standardized all covariates in the incidence (probability of blindness) to make them comparable. Age and treated eye variables significantly affected the incidence sub-model at a 5% significance level for the GPLSI-AH and SIC models. The type of eye treated also has a significant impact on the probability of (un)cure for the LC model. For the latency sub-model, none of the predictors significantly affects the time to blindness. For any two subjects who are uncured, our estimate suggests that there is no significant difference in their hazards to blindness if one was treated on the right eye and the other on the left.

Table 5. Description of variables in the Diabetic Retinopathy Study data used for our analysis.

Variables	Description
Subject id	Unique id for subjects
laser type	1 = xenon, 2 = argon
treated eye	1 = right, 2 = left
age at diagnosis	Subject's age at diagnosis of diabetes
risk group	subject's risk of diabetic retinopathy on a scale of 6 -12
status	0 = censored, 1 = blindness
time	follow-up time

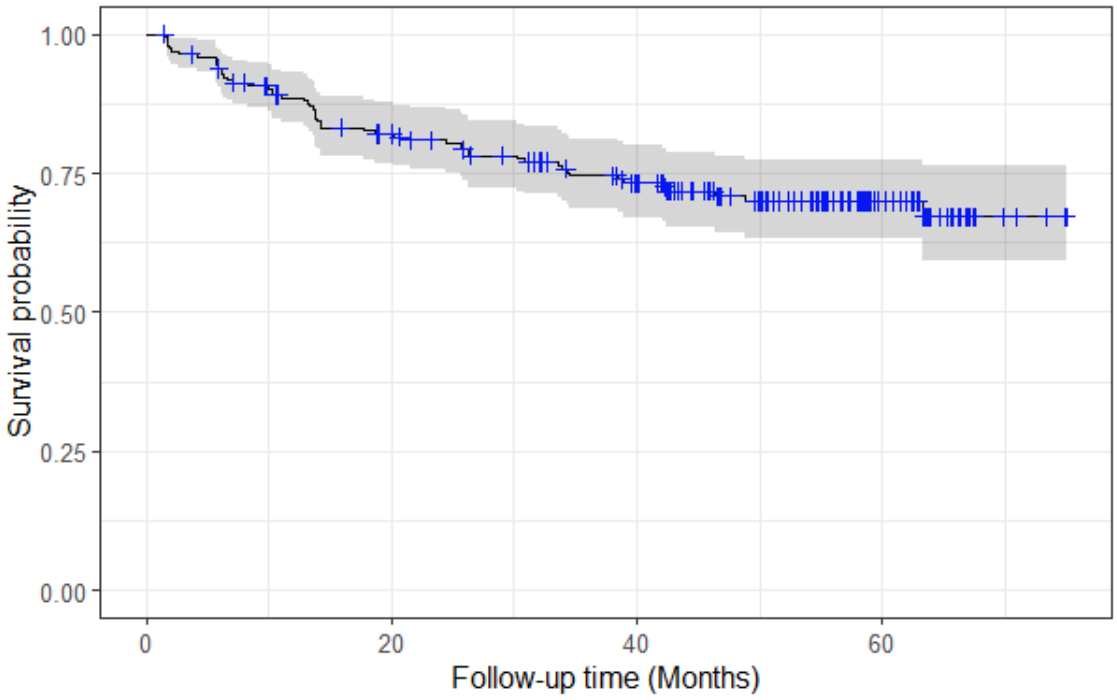

Figure 9. Kaplan Meier survival plot of the DRS study showing a plateau.

Table 6. Parameter estimates and standard errors for GPLSI-AH, SIC and LC cure models.

	GPLSI-AH cure model			SIC cure model			LC cure model		
Incidence	Estimate	Std.Error	p-value	Estimate	Std.Error	p-value	Estimate	Std.Error	p-value
(intercept)	-	-	-	-	-	-	-0.7854	0.1385	0.0000
age	0.9047	0.2414	0.0002	-0.4537	0.2255	0.0442	-0.1787	0.2166	0.4092
risk	-0.4261	0.3755	0.2565	-0.0810	0.3189	0.7994	0.0126	0.2168	0.9538
laser type	0.0718	0.1874	0.7017	0.3967	0.3425	0.2468	0.1924	0.1993	0.3344
treated eye	0.5691	0.1975	0.0040	0.7939	0.2677	0.0030	0.6552	0.2358	0.0054
Latency	Estimate	Std.Error	p-value	Estimate	Std.Error	p-value	Estimate	Std.Error	p-value
age	-0.0007	0.00057	0.1996	-0.0164	0.3150	0.9584	-0.0171	0.0211	0.4180
risk	0.0044	0.0061	0.4689	0.1443	0.2463	0.5579	0.1298	0.1483	0.3811
laser type	-0.0096	0.0134	0.4718	-0.4806	0.2400	0.0452	-0.5607	0.4756	0.2384
treated eye	-0.0177	0.0257	0.4908	-0.6389	0.3594	0.0755	-0.7574	0.6054	0.2109

5. CONCLUSIONS

We proposed a GPLSI-AH mixture cure model, which is a versatile semi-parametric model. We stated the proposed model's identifiability. The GPLSI for the uncured probability is estimated using a P-splines-based maximum likelihood estimation approach with an EM algorithm. The simulation demonstrated that our method performs well in terms of estimating the uncure probability and outperforms other methods when the true link function is not logistic. In terms of the uncured probability, our technique appears to be less sensitive to model misspecification. Despite the many advantages of using a cure model, it is generally recommended that one ensures there is some contextual evidence for the presence of a cure fraction (Legrand, 2021). Several strategies have been proposed. According to Taylor (1995), a good indication of the presence of a cure fraction is when the Kaplan-Meier (KM) survival curve levels out with a lengthy plateau containing a "high" number of data

points. Several attempts have been made to formally test whether one can confidently apply a cure model. Originally, Maller and Zhou (1996), proposed assessing the identifiability condition as a formal test for the presence of a sufficiently long follow-up. Other works for testing the presence of cure can be found in Zhao *et al.* (2009) and Hsu *et al.* (2016). However, as pointed out by Legrand (2021) there is not a clear-cut, widely available hypothesis test for the evidence of a cure fraction, and the current recommendation is to rely on a visual inspection of the tail of the KM survival curve. Assuming that there are many observations in the plateau of the KM curve, then in the case of heavy right censoring, we advocate for our parameter initialization strategy. However, further work through extensive simulations needs to be done to investigate and provide a better approach or rule for applying our initialization strategy. Through, this we can develop a test for the presence of cure. The GPLSI-AH can be considered a diagnostic tool to investigate misspecification of the incidence of the mixture cure model. Future work will focus on developing a test for a parametric form of $p(\cdot)$ possibly using some likelihood based arguments or information criteria.

REFERENCES

'Illumina methylation beadchips achieve breadth of coverage using 2 Infinium chemistries,' Technical Report Pub. No. 270-2012-00, Illumina, Inc..,, 2015, techsupport@illumina.com.

Aalen, O., 'A model for nonparametric regression analysis of counting processes,' in 'Mathematical statistics and probability theory,' pp. 1–25, Springer, 1980.

Aalen, O., Borgan, O., and Gjessing, H., *Survival and event history analysis: a process point of view*, Springer Science & Business Media, 2008.

Aalen, O. O., 'A linear regression model for the analysis of life times,' Statistics in medicine, 1989, **8**(8), pp. 907–925.

Adler, A. and Rosalsky, A., 'On the weak law of large numbers for normed weighted sums of iid random variables,' International Journal of Mathematics and Mathematical Sciences, 1991, **14**(1), pp. 191–202.

Amico, M. and Van Keilegom, I., 'Cure models in survival analysis,' Annual Review of Statistics and Its Application, 2018, **5**, pp. 311–342.

Amico, M., Van Keilegom, I., and Legrand, C., 'The single-index/cox mixture cure model,' Biometrics, 2019, **75**(2), pp. 452–462.

Anderssen, R. and Bloomfield, P., 'A time series approach to numerical differentiation,' Technometrics, 1974, **16**(1), pp. 69–75.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A., 'Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays,' Bioinformatics, 2014, **30**(10), pp. 1363–1369, doi:10.1093/bioinformatics/btu049.

Barfield, R. T., Kilaru, V., Smith, A. K., and Conneely, K. N., 'Cpgassoc: an r function for analysis of dna methylation microarray data,' Bioinformatics, 2012, **28**(9), pp. 1280–1281.

Basu, B., Chakraborty, J., Chandra, A., Katarkar, A., Baldevbhai, J. R. K., Dhar Chowdhury, D., Ray, J. G., Chaudhuri, K., and Chatterjee, R., 'Genome-wide DNA methylation profile identified a unique set of differentially methylated immune genes in oral squamous cell carcinoma patients in India,' Clin Epigenetics, 2017, **9**, p. 13.

Benjamini, Y. and Hochberg, Y., 'Controlling the false discovery rate: a practical and powerful approach to multiple testing,' Journal of the Royal statistical society: series B (Methodological), 1995a, **57**(1), pp. 289–300.

Benjamini, Y. and Hochberg, Y., 'Controlling the false discovery rate: a practical and powerful approach to multiple testing,' Journal of the Royal statistical society: series B (Methodological), 1995b, **57**(1), pp. 289–300.

Bennett, S., 'Analysis of survival data by the proportional odds model,' Statistics in medicine, 1983, **2**(2), pp. 273–277.

Berkson, J. and Gage, R. P., 'Survival curve for cancer patients following treatment,' Journal of the American Statistical Association, 1952, **47**(259), pp. 501–515.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., *et al.*, 'High density dna methylation array with single cpg site resolution,' Genomics, 2011, **98**(4), pp. 288–295.

Billingsley, P., *Probability and Measure*, John Wiley and Sons, second edition, 1986.

Billingsley, P., *Probability and measure*, John Wiley & Sons, 1995.

Boag, J. W., 'Maximum likelihood estimates of the proportion of patients cured by cancer therapy,' Journal of the royal statistical society series b-methodological, 1949, doi:10.1111/j.2517-6161.1949.tb00020.x.

Breton-Larrivée, M., Elder, E., and McGraw, S., 'DNA methylation, environmental exposures and early embryo development,' Anim Reprod, Oct 2019, **16**(3), pp. 465–474.

Butcher, L. M. and Beck, S., 'Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data,' Methods, Jan 2015, **72**, pp. 21–28.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P., 'Generalized partially linear single-index models,' Journal of the American Statistical Association, 1997, **92**(438), pp. 477–489.

Casella, G. and Berger, R. L., *Statistical inference*, Cengage Learning, 2021.

Cedar, H., 'Dna methylation and gene activity.' Cell, 1988, **53**(1), pp. 3–4.

Center for Cancer Genomics - National Cancer Institute, 'The Cancer Genome Atlas Program (TCGA),' Retrieved from `https://www.cancer.gov/ccg/research/genome-sequencing/tcga`, n.d., accessed June 8, 2023.

Chen, D. P., Lin, Y. C., and Fann, C. S., 'Methods for identifying differentially methylated regions for sequence- and array-based data,' Brief. Funct. Genomics, 2016, **15**(6), pp. 485–490, ISSN 20412657, doi:10.1093/bfgp/elw018.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A., 'Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool,' BMC Bioinformatics, Apr 2013, **14**, p. 128.

Chen, Y. A., Choufani, S., Grafodatskaya, D., Butcher, D. T., Ferreira, J. C., and Weksberg, R., 'Cross-reactive DNA microarray probes lead to false discovery of autosomal sex-associated DNA methylation,' Am J Hum Genet, Oct 2012, **91**(4), pp. 762–764.

Chen, Y. Q. and Wang, M.-C., 'Analysis of accelerated hazards models,' Journal of the American Statistical Association, 2000, **95**(450), pp. 608–618.

Chiou, S. H., Austin, M. D., Qian, J., and Betensky, R. A., 'Transformation model estimation of survival under dependent truncation and independent censoring,' Statistical methods in medical research, 2019, **28**(12), pp. 3785–3798.

Cox, D. R., 'Regression models and life-tables,' J. Roy. Statist. Soc. Ser. B, 1972, **34**, pp. 187–220, ISSN 0035-9246.

Cox, D. R., 'Partial likelihood,' Biometrika, 1975, **62**(2), pp. 269–276.

Cox, D. R. and Oakes, D., *Analysis of survival data*, Chapman and Hall/CRC, 1984.

Cramér, H., *Mathematical Methods of Statistics (PMS-9), Volume 9*, Princeton university press, 2016.

Crary-Dooley, F. K., Tam, M. E., Dunaway, K. W., Hertz-Picciotto, I., Schmidt, R. J., and LaSalle, J. M., 'A comparison of existing global dna methylation assays to low-coverage whole-genome bisulfite sequencing for epidemiological studies,' Epigenetics, 2017, **12**(3), pp. 206–214.

Craven, P. and Wahba, G., 'Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation,' Numerische mathematik, 1978, **31**(4), pp. 377–403.

DasGupta, A., *Asymptotic theory of statistics and probability*, Springer Science & Business Media, 2008.

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F., 'Evaluation of the Infinium Methylation 450K technology,' Epigenomics, Dec 2011, **3**(6), pp. 771–784.

Dempster, A. P., Laird, N. M., and Rubin, D. B., 'Maximum likelihood from incomplete data via the em algorithm,' Journal of the Royal Statistical Society: Series B (Methodological), 1977, **39**(1), pp. 1–22.

Dirick, L., Claeskens, G., Vasnev, A., and Baesens, B., 'A hierarchical mixture cure model with unobserved heterogeneity for credit risk,' Econometrics and Statistics, 2022, **22**, pp. 39–55.

Du, P., Zhang, X., Huang, C. C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M., 'Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis,' BMC Bioinformatics, Nov 2010, **11**, p. 587.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S., 'DNA methylation profiling of human chromosomes 6, 20 and 22,' Nat Genet, Dec 2006, **38**(12), pp. 1378–1385.

Eden, S. and Cedar, H., 'Role of DNA methylation in the regulation of transcription,' Current opinion in genetics & development, 1994, **4**(2), pp. 255–259.

Efron, B. and Morris, C., 'Empirical bayes on vector observations: An extension of stein's method,' Biometrika, 1972, **59**(2), pp. 335–347.

Ehrlich, M. and Wang, R. Y.-H., '5-methylcytosine in eukaryotic DNA,' Science, 1981, **212**(4501), pp. 1350–1357.

Eilers, P. H. and Marx, B. D., 'Flexible smoothing with b-splines and penalties,' Statistical science, 1996, **11**(2), pp. 89–121.

ENCODE Project Consortium, 'An integrated encyclopedia of dna elements in the human genome,' Nature, 2012, **489**(7414), p. 57, doi:10.1038/nature11247.

Farewell, V. T., 'A model for a binary variable with time-censored observations,' Biometrika, 1977, **64**(1), pp. 43–46.

Farewell, V. T., 'The use of mixture models for the analysis of survival data with long-term survivors,' Biometrics, 1982, pp. 1041–1046.

Felizzi, F., Paracha, N., Pöhlmann, J., and Ray, J., 'Mixture cure models in oncology: a tutorial and practical guidance,' PharmacoEconomics-Open, 2021, **5**, pp. 143–155.

Fernandez, A., O'Leary, C., O'Byrne, K. J., Burgess, J., Richard, D. J., and Suraweera, A., 'Epigenetic mechanisms in dna double strand break repair: A clinical review,' Frontiers in Molecular Biosciences, 2021, **8**, p. 685440.

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., Greenwood, C. M., and Hansen, K. D., 'Functional normalization of 450k methylation array data improves replication in large cancer studies,' Genome Biology, 2014, **15**(12), p. 503, doi:10.1186/s13059-014-0503-2.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L., 'A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands,' Proc Natl Acad Sci U S A, Mar 1992, **89**(5), pp. 1827–1831.

Greenberg, M. V. and Bourc'his, D., 'The diverse roles of DNA methylation in mammalian development and disease,' Nature reviews Molecular cell biology, 2019, **20**(10), pp. 590–607.

Haibo, H. and Yunqian, M., 'Imbalanced learning: foundations, algorithms, and applications,' Wiley-IEEE Press, 2013, **1**, p. 27.

Hanin, L. and Huang, L.-S., 'Identifiability of cure models revisited,' Journal of Multivariate Analysis, 2014, **130**, pp. 261–274.

Hardle, W., Hall, P., and Ichimura, H., 'Optimal smoothing in single-index models,' The annals of Statistics, 1993, **21**(1), pp. 157–178.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., *et al.*, *Nonparametric and semiparametric models*, volume 1, Springer, 2004.

Härdle, W. K. *et al.*, *Smoothing techniques: with implementation in S*, Springer Science & Business Media, 1991.

Heiss, J. A. and Just, A. C., 'Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses,' Clin Epigenetics, 01 2019, **11**(1), p. 15.

Hsu, W.-W., Todem, D., and Kim, K., 'A sup-score test for the cure fraction in mixture models for long-term survivors,' Biometrics, 2016, **72**(4), pp. 1348–1357.

Huffer, F. W. and McKeague, I. W., 'Weighted least squares estimation for aalen's additive risk model,' Journal of the American Statistical Association, 1991, **86**(413), pp. 114–129.

Huster, W. J., Brookmeyer, R., and Self, S. G., 'Modelling paired survival data with covariates,' Biometrics, 1989, pp. 145–156.

Ichiyanagi, T., Ichiyanagi, K., Miyake, M., and Sasaki, H., 'Accumulation and loss of asymmetric non-CpG methylation during male germ-cell development,' Nucleic Acids Res, Jan 2013, **41**(2), pp. 738–745.

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A., 'Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies,' Int J Epidemiol, Feb 2012, **41**(1), pp. 200–209.

James, G., Witten, D., Hastie, T., and Tibshirani, R., *An introduction to statistical learning*, volume 112, Springer, 2013.

Jeong, M., Guzman, A. G., and Goodell, M. A., 'Genome-wide analysis of dna methylation in hematopoietic cells: Dna methylation analysis by wgbs,' Acute Myeloid Leukemia: Methods and Protocols, 2017, pp. 137–149.

Jiang, J., *Large sample techniques for statistics*, Springer Science & Business Media, 2010.

Jin, Z. and Liu, Y., 'DNA methylation in human diseases,' Genes Dis., 2018, **5**(1), pp. 1–8, ISSN 23523042, doi:10.1016/j.gendis.2018.01.002.

Kalbfleisch, J. D. and Prentice, R. L., 'Marginal likelihoods based on cox's regression and life model,' Biometrika, 1973, **60**(2), pp. 267–278.

Kalbfleisch, J. D. and Prentice, R. L., *The statistical analysis of failure time data*, John Wiley & Sons, 2011.

Kaplan, E. L. and Meier, P., 'Nonparametric estimation from incomplete observations,' Journal of the American statistical association, 1958, **53**(282), pp. 457–481.

Khaliq, A., Waqas, A., Nisar, Q. A., Haider, S., and Asghar, Z., 'Application of ai and robotics in hospitality sector: A resource gain and resource loss perspective,' Technology in Society, 2022, **68**, p. 101807.

Kilaru, V., Barfield, R. T., Schroeder, J. W., Smith, A. K., and Conneely, K. N., 'MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data,' Epigenetics, Mar 2012, **7**(3), pp. 225–229.

Klein, J. P. and Moeschberger, M. L., *Survival analysis: techniques for censored and truncated data*, volume 1230, Springer, 2003.

Kleinbaum, D. G. and Klein, M., *Survival analysis a self-learning text*, Springer, third edition, 2012.

Kuk, A. Y. and Chen, C.-H., 'A mixture model combining logistic regression with proportional hazards regression,' Biometrika, 1992, **79**(3), pp. 531–541.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., and Ma'ayan, A., 'Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,' Nucleic Acids Res, 07 2016, **44**(W1), pp. W90–97.

Laird, P. W., 'Principles and challenges of genome-wide dna methylation analysis,' Nature Reviews Genetics, 2010, **11**(3), pp. 191–203.

Lao, V. V. and Grady, W. M., 'Epigenetics and colorectal cancer,' Nat Rev Gastroenterol Hepatol, Oct 2011, **8**(12), pp. 686–700.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Sung, K. W. K., Rigoutsos, I., Loring, J., and Wei, C. L., 'Dynamic changes in the human methylome during differentiation,' Genome Res., 2010, **20**(3), pp. 320–331, ISSN 10889051, doi:10.1101/gr.101907.109.

Lee, E. T. and Wang, J., *Statistical methods for survival data analysis*, volume 476, John Wiley & Sons, 2003.

Leek, J. T. and Storey, J. D., 'Capturing heterogeneity in gene expression studies by surrogate variable analysis,' PLoS Genet, Sep 2007, **3**(9), pp. 1724–1735.

Legrand, C., *Advanced survival models*, Chapman and Hall/CRC, 2021.

Lehmann, E. L., *Elements of large-sample theory*, Springer Science & Business Media, 2004.

Li, C.-S. and Lu, M., 'A lack-of-fit test for generalized linear models via single-index techniques,' Computational Statistics, 2018, **33**, pp. 731–756.

Li, C.-S. and Taylor, J. M., 'Smoothing covariate effects in cure models,' Communications in Statistics-Theory and Methods, 2002, **31**(3), pp. 477–493.

Li, D., Xie, Z., Pape, M. L., and Dye, T., 'An evaluation of statistical methods for DNA methylation microarray data analysis,' BMC Bioinformatics, jul 2015, **16**(1), ISSN 14712105, doi:10.1186/s12859-015-0641-x.

Lin, D. Y. and Ying, Z., 'Semiparametric analysis of the additive risk model,' Biometrika, 1994, **81**(1), pp. 61–71.

Liu, A., Jiang, C., Liu, Q., Yin, H., Zhou, H., Ma, H., and Geng, Q., 'The inverted u-shaped association of caffeine intake with serum uric acid in us adults,' The journal of nutrition, health & aging, 2022, **26**(4), pp. 391–399.

López-Cheda, A., Cao, R., Jácome, M. A., and Van Keilegom, I., 'Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models,' Computational Statistics & Data Analysis, 2017, **105**, pp. 144–165.

Lu, W., 'Maximum likelihood estimation in the proportional hazards cure model,' Annals of the Institute of Statistical Mathematics, 2008, **60**, pp. 545–574.

Ma, Y. and He, H., 'Imbalanced learning: foundations, algorithms, and applications,' 2013.

Madadizadeh, F., Ghanbarnejad, A., Ghavami, V., Bandamiri, M. Z., and Mohammadianpanah, M., 'Applying additive hazards models for analyzing survival in patients with colorectal cancer in fars province, southern iran,' Asian Pacific journal of cancer prevention: APJCP, 2017, **18**(4), p. 1077.

Maghbooli, Z., Larijani, B., Emamgholipour, S., Amini, M., Keshtkar, A., and Pasalar, P., 'Aberrant DNA methylation patterns in diabetic nephropathy,' J Diabetes Metab Disord, 2014, **13**, p. 69.

Maksimovic, J., Gordon, L., and Oshlack, A., 'SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips,' Genome Biology, 2012, **13**(6), p. R44, doi:10.1186/gb-2012-13-6-r44.

Maksimovic, J., Oshlack, A., and Phipson, B., 'Gene set enrichment analysis for genome-wide DNA methylation data,' Genome Biol, 06 2021, **22**(1), p. 173.

Maller, R. A. and Zhou, X., *Survival analysis with long-term survivors*, volume 525, Wiley New York, 1996.

Mallik, S., Odom, G. J., Gao, Z., Gomez, L., Chen, X., and Wang, L., 'An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays,' Brief Bioinform, 11 2019, **20**(6), pp. 2224–2235.

McCartney, D. L., Walker, R. M., Morris, S. W., McIntosh, A. M., Porteous, D. J., and Evans, K. L., 'Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationepic beadchip,' Genomics data, 2016, **9**, pp. 22–24.

McCullagh, P. and Nelder, J. A., *Generalized linear models*, Routledge, 2019.

McGregor, K., Bernatsky, S., Colmegna, I., Hudson, M., Pastinen, T., Labbe, A., and Greenwood, C. M., 'An evaluation of methods correcting for cell-type heterogeneity in dna methylation studies,' Genome biology, 2016, **17**(1), pp. 1–17.

McLachlan, G. J. and Krishnan, T., *The EM algorithm and extensions*, John Wiley & Sons, 2007.

Mood, A., Graybill, F., and Boes, D., '(1974), introduction to the theory of statistics,' 1974.

Moran, S., Arribas, C., and Esteller, M., 'Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences,' Epigenomics, 2016, **8**(3), pp. 389–399.

Patilea, V. and Van Keilegom, I., 'A general approach for cure models in survival analysis,' 2020.

Peng, Y. and Dear, K. B., 'A nonparametric mixture model for cure rate estimation,' Biometrics, 2000, **56**(1), pp. 237–243.

Peng, Y. and Yu, B., *Cure Models: Methods, Applications, and Implementation*, Chapman and Hall/CRC, 2021.

Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., V Lord, R., Clark, S. J., and Molloy, P. L., 'De novo identification of differentially methylated regions in the human genome,' Epigenetics Chromatin, 2015, **8**, p. 6.

Piao, Y., Xu, W., Park, K. H., Ryu, K. H., and Xiang, R., 'Comprehensive Evaluation of Differential Methylation Analysis Methods for Bisulfite Sequencing Data,' Int J Environ Res Public Health, 07 2021, **18**(15).

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C., 'A data-driven approach to preprocessing illumina 450k methylation array data,' BMC genomics, 2013, **14**(1), pp. 1–10.

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., and Clark, S. J., 'Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling,' Genome Biol, 10 2016, **17**(1), p. 208.

Price, D., *Survival Models for Heterogeneous Populations with Cure*, Ph.D. thesis, Emory University, 2000.

Procter, M., Chou, L.-S., Tang, W., Jama, M., and Mao, R., 'Molecular diagnosis of prader–willi and angelman syndromes by methylation-specific melting analysis and methylation-specific multiplex ligation-dependent probe amplification,' Clinical chemistry, 2006, **52**(7), pp. 1276–1283.

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., and Jaenisch, R., 'Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a,' Proc. Natl. Acad. Sci. U. S. A., 2000, **97**(10), pp. 5237–5242, ISSN 00278424, doi:10.1073/pnas.97.10.5237.

Resnick, S. I., *A probability path*, Birkhäuser Boston, 1999.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K., 'limma powers differential expression analyses for RNA-sequencing and microarray studies,' Nucleic Acids Res, Apr 2015, **43**(7), p. e47.

Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X., 'Statistical methods for detecting differentially methylated loci and regions,' Frontiers in genetics, 2014, **5**, p. 324.

Rossignol, S., Steunou, V., Chalas, C., Kerjean, A., Rigolet, M., Viegas-Pequignot, E., Jouannet, P., Le Bouc, Y., and Gicquel, C., 'The epigenetic imprinting defect of patients with beckwith—wiedemann syndrome born after assisted reproductive technology is not restricted to the 11p15 region,' Journal of medical genetics, 2006, **43**(12), pp. 902–907.

Roy, N. K., Monisha, J., Padmavathi, G., Lalhruaitluanga, H., Kumar, N. S., Singh, A. K., Bordoloi, D., Baruah, M. N., Ahmed, G. N., Longkumar, I., Arfuso, F., Kumar, A. P., and Kunnumakkara, A. B., 'Isoform-Specific Role of Akt in Oral Squamous Cell Carcinoma,' Biomolecules, 06 2019, **9**(7).

Ruppert, D., Wand, M. P., and Carroll, R. J., *Semiparametric regression*, 12, Cambridge university press, 2003.

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., and Esteller, M., 'Validation of a dna methylation microarray for 450,000 cpg sites in the human genome,' Epigenetics, 2011, **6**(6), pp. 692–702.

Satterthwaite, F. E., 'An approximate distribution of estimates of variance components,' Biometrics, Dec 1946, **2**(6), pp. 110–114.

Shafi, A., Mitrea, C., Nguyen, T., and Draghici, S., 'A survey of the approaches for identifying differential methylation using bisulfite sequencing data,' Brief Bioinform, 09 2018, **19**(5), pp. 737–753.

Shang, S., Liu, M., Zeleniuch-Jacquotte, A., Clendenen, T. V., Krogh, V., Hallmans, G., and Lu, W., 'Partially linear single index cox regression model in nested case-control studies,' Computational statistics & data analysis, 2013, **67**, pp. 199–212.

Shiah, Y. J., Fraser, M., Bristow, R. G., and Boutros, P. C., 'Comparison of pre-processing methods for Infinium HumanMethylation450 BeadChip array,' Bioinformatics, Oct 2017, **33**(20), pp. 3151–3157.

Shu, C., Zhang, X., Aouizerat, B. E., and Xu, K., 'Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells,' Epigenetics Chromatin, 11 2020, **13**(1), p. 51.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, volume 26, CRC Press, 1986.

Smith, M. L., Baggerly, K. A., Bengtsson, H., Ritchie, M. E., and Hansen, K. D., 'illuminaio: An open source idat parsing tool for illumina microarrays,' F1000Research, 2013, **2**.

Smyth, G. K., 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments,' Stat Appl Genet Mol Biol, 2004, **3**, p. Article3.

Sofer, T., Schifano, E. D., Hoppin, J. A., Hou, L., and Baccarelli, A. A., 'A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure,' Bioinformatics, Nov 2013, **29**(22), pp. 2884–2891.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M., 'The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1,' 1949.

Sun, H. and Wang, S., 'Penalized logistic regression for high-dimensional DNA methylation data with case-control studies,' Bioinformatics, May 2012, **28**(10), pp. 1368–1375.

Sun, L., Namboodiri, S., Chen, E., and Sun, S., 'Preliminary Analysis of Within-Sample Co-methylation Patterns in Normal and Cancerous Breast Samples,' Cancer Inform, 2019, **18**, p. 1176935119880516.

Sun, L. and Sun, S., 'Within-sample co-methylation patterns in normal tissues,' BioData Min, 2019, **12**, p. 9.

Sun, S., Dammann, J., Lai, P., and Tian, C., 'Thorough statistical analyses of breast cancer co-methylation patterns,' BMC Genom Data, 04 2022, **23**(1), p. 29.

Susan, J. C., Harrison, J., Paul, C. L., and Frommer, M., 'High sensitivity mapping of methylated cytosines,' Nucleic acids research, 1994, **22**(15), pp. 2990–2997.

Sy, J. P. and Taylor, J. M., 'Estimation in a cox proportional hazards cure model,' Biometrics, 2000, **56**(1), pp. 227–236.

Szyf, M., 'Dna methylation signatures for breast cancer classification and prognosis,' Genome medicine, 2012, **4**(3), pp. 1–12.

Taylor, H. L., 'Physical activity: is it still a risk factor?' Preventive medicine, 1983, **12**(1), pp. 20–24.

Taylor, J. M., 'Semi-parametric estimation in failure time mixture models,' Biometrics, 1995, pp. 899–907.

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S., 'A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data,' Bioinformatics, Jan 2013, **29**(2), pp. 189–196.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 'A promoter-level mammalian expression atlas,' Nature, 2014, **507**(7493), pp. 462–470.

Therneau, T. M., Grambsch, P. M., Therneau, T. M., and Grambsch, P. M., *The cox model*, Springer, 2000.

Touleimat, N. and Tost, J., 'Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation,' Epigenomics, Jun 2012, **4**(3), pp. 325–341.

Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., and Siegmund, K. D., 'Low-level processing of Illumina Infinium DNA Methylation BeadArrays,' Nucleic Acids Res, Apr 2013, **41**(7), p. e90.

Wang, D., Yan, L., Hu, Q., Sucheston, L. E., Higgins, M. J., Ambrosone, C. B., Johnson, C. S., Smiraglia, D. J., and Liu, S., 'Ima: an r package for high-throughput analysis of illumina's 450k infinium methylation data,' Bioinformatics, 2012, **28**(5), pp. 729–730.

Wang, L. and Cao, G., 'Efficient estimation for generalized partially linear single-index models,' 2018.

Wang, T., Guan, W., Lin, J., Boutaoui, N., Canino, G., Luo, J., Celedón, J. C., and Chen, W., 'A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data,' Epigenetics, 2015, **10**(7), pp. 662–669.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Foekens, J. A., 'Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,' Lancet, 2005, **365**(9460), pp. 671–679.

Wang, Z., Wu, X., and Wang, Y., 'A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip,' BMC Bioinformatics, 04 2018, **19**(Suppl 5), p. 115.

Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y. C., 'COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis,' Nucleic Acids Res, Jun 2013a, **41**(11), p. e117.

Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y.-C., 'Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis,' Nucleic acids research, 2013b, **41**(11), pp. e117–e117.

Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R., Dymov, S., Szyf, M., and Meaney, M. J., 'Epigenetic programming by maternal behavior,' Nat Neurosci, Aug 2004, **7**(8), pp. 847–854.

Weisenberger, D., Van Den Berg, D., Pan, F., Berman, B., and Laird, P., 'Comprehensive dna methylation analysis on the illumina infinium assay platform,' Illumina, San Diego, 2008.

Wood, S., 'mgcv: Mixed gam computation vehicle with gcv/aic/reml smoothness estimation,' 2012.

Wood, S. N., 'Thin plate regression splines,' Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2003, **65**(1), pp. 95–114.

Wu, D., Gu, J., and Zhang, M. Q., 'Fastdma: an infinium humanmethylation450 beadchip analyzer,' PloS one, 2013, **8**(9), p. e74275.

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., and Ma'ayan, A., 'Gene Set Knowledge Discovery with Enrichr,' Curr Protoc, Mar 2021, **1**(3), p. e90.

Xu, J. and Peng, Y., 'Nonparametric cure rate estimation with covariates,' Canadian Journal of Statistics, 2014, **42**(1), pp. 1–17.

Yu, Y. and Ruppert, D., 'Penalized spline estimation for partially linear single-index models,' Journal of the American Statistical Association, 2002, **97**(460), pp. 1042–1054.

Yu, Y., Wu, C., and Zhang, Y., 'Penalised spline estimation for generalised partially linear single-index models,' Statistics and Computing, 2017, **27**, pp. 571–582.

Zeng, Z., Gao, Y., Li, J., Zhang, G., Sun, S., Wu, Q., Gong, Y., and Xie, C., 'Violations of proportional hazard assumption in cox regression model of transcriptomic data in tcga pan-cancer cohorts,' Computational and Structural Biotechnology Journal, 2022, **20**, pp. 496–507.

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E., 'Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements,' Genome Biol, Jan 2015, **16**, p. 14.

Zhang, Y., Liu, H., Lv, J., Xiao, X., Zhu, J., Liu, X., Su, J., Li, X., Wu, Q., Wang, F., *et al.*, 'Qdmr: a quantitative method for identification of differentially methylated regions by entropy,' Nucleic acids research, 2011, **39**(9), pp. e58–e58.

Zhang, Y., Wang, S., and Wang, X., 'Data-Driven-Based Approach to Identifying Differentially Methylated Regions Using Modified 1D Ising Model,' Biomed Res. Int., 2018, **2018**, ISSN 23146141, doi:10.1155/2018/1070645.

Zhao, Y., Lee, A. H., Yau, K. K., Burke, V., and McLachlan, G. J., 'A score test for assessing the cured proportion in the long-term survivor mixture model,' Statistics in medicine, 2009, **28**(27), pp. 3454–3466.

SECTION

4. SUMMARY AND CONCLUSIONS

In the first part of this dissertation, methods were developed that increase both statistical power and precision in detecting novel regions in the human genome that are differentially methylated. These methods also have a higher susceptibility of capturing the true length of DMRs. This type of advancement can contribute to the field of precision medicine. Genes that are affected by the methylation process can be detected when the proposed methods are combined with pathway analysis methods, in efforts to develop medicine that targets the genes of interest. While the proposed methods have led to significant progress, there is still a need for additional research to enhance DMR detection, especially for small effect sizes. The selection of an "adaptive" bandwidth could be a crucial factor. Moreover, the choice of kernel may impact the results, suggesting that further exploration into different kernels could be beneficial.

The second part of the dissertation focused on advances in cure survival models research. Specifically, a mixture cure model and a computational algorithm are proposed that flexibly model the probability of cured in the cured sub-population using a generalized partially linear single-index model and models the time to event of the uncured sub-population using the additive hazard model. This method was shown via simulation studies to perform much better with large sample size and is less sensitive to the misspecification of the cure probability model. The use of the additive hazard model offers an alternative to the Cox proportional hazard method. Hence this mixture cure model will be beneficial to fields where large sample sizes are the norm and where one has reason to believe the Cox-proportional hazard is not

appropriate for the uncured sub-population. The initialization approach employed in the EM algorithm warrants additional investigation. Moreover, more research is required to provide guidance to users in selecting the bases for estimating the index function, as this could potentially yield improved results. The proposed method, when compared to the logistic/Cox mixture cure model, tends to require more execution time. Therefore, it could be advantageous to conduct a statistical test before fitting the model to ascertain whether a logistic model for the cure probability would suffice.

APPENDIX

SUPPLEMENTARY FILE I

1. ASYMPTOTIC RESULTS

This section pertains to the proof of the asymptotic results of the proposed estimator for paper I. More specifically it investigates the situation under which $S(x_i)$ is consistent and obeys the central limit theorem (CLT).

Lemma 1. *$\{Y_{1,v}; v \geq 1\}$ be a family of F-distributed random variables with degrees of freedom 1 and v, i.e. $Y \sim F_{1,v}$. Then the family is uniformly tight.*

Proof. $E(Y) = \dfrac{v}{v-2}$ for $v \geq 3$. $\exists\ v_0$ such that $\dfrac{v}{v-2} \leq 2$. $\forall \epsilon > 0$, $\exists\ M_1$ such that $P(Y_{1,1} \leq M_1) > 1 - \frac{\epsilon}{2}$ and $P(Y_{1,2} \leq M_1) > 1 - \frac{\epsilon}{2}$. Hence $Y_{1,1}$ and $Y_{1,2}$ are tight. Now, $\forall v \geq v_0$, $P(Y_{1,v} \geq M) \leq \frac{E(Y_{1,v})}{M} \leq \frac{3}{M}$. For $\frac{3}{M} < \epsilon$, we have that $\forall v \geq v_0$, $P(Y_{1,n} > M_\epsilon) < \epsilon$. Set $M_0 = \max(M_1, M_\epsilon)$ we have $\sup\limits_{v \geq v_0} P(Y_{1,v} < M_0) > 1 - \epsilon$. □

Theorem 1. *Let $\{Y_j,\ j > 1\}$ be independent F-distributed random variables obtained from site-level testing via limma's moderated F-statistic (t^2-statistic) and $\{w_j,\ j > 1\}$ be constants satisfying $\sum_{j=1}^{n} w_j^2 = \mathcal{O}(nw_n^2)$. Further, at CpG site x_i, let $S(x_i) = \sum_{j=1}^{n} w_j(x_i) Y_j$ where*

$$w_j(x_i) = \begin{cases} 1, & j = i \\ t \in (0,1), & j \neq i \end{cases}.$$

Then,

$$\frac{\sum_{j=1}^{n} w_j \left(Y_j - EY\mathbf{1}_{\{|Y| \leq n^2\}}\right)}{w_n n^2} \xrightarrow{p} 0.$$

Proof. To reduce notational complexity, we write $S(x_i)$ as S and $w_j(x_i)$ as w_j. We will take Y to be χ^2 distributed for the basis of our context. For $n > 1$, $S(x_i)$ has $(n-1)$ weighted Y_i's and one unweighted Y_i. Since Y_i's are iid we can write (1) , without loss of generality:

$$\begin{aligned} S &= Y_1 + w_2 Y_2 + \cdots + w_{n-1} Y_{n-1} + w_n Y_n \\ &= Y_1 + \sum_{j=2}^{n} w_j Y_j \\ &= \sum_{j=1}^{n} w_j Y_j \end{aligned} \tag{1}$$

where $\sum_{j=2}^{n} w_j = 1$ and $w_1 = 1$.
Since the Y's are obtained from *limma*, we assume the denominator df is large enough so that $Y \sim \chi^2_1$ (see Section 2 of the appendix for proof). Define $b_n = w_n n^2$ so that $c_n = n^2$ where $0 < w_j \leq 1$.

$$\begin{aligned} \lim_{n\to\infty} nP(Y > n^2) &\leq \frac{nE(Y)}{n^2} && \text{(By Markov Inequality)} \\ &= \frac{E(Y)}{n} \\ &= \frac{1}{n} \to 0 \text{ as } n \to \infty \end{aligned}$$

$\implies nP(Y > n^2) = o(1)$. Now,

$$\begin{aligned} \frac{n^2}{n} &= n \uparrow \text{ for } n > 1 \\ \frac{\sum_{j=1}^{n} w_j^2}{n w_n^2} &\leq \frac{n}{n w_2^2} \\ &= \frac{1}{w_n^2} \end{aligned}$$

$\implies \sum_{j=1}^{n} w_j^2 = \mathcal{O}\left(\frac{1}{w_n^2}\right)$.

By the Lemma 1 (Adler and Rosalsky, 1991) (see Section 2.7) we have that,

$$\sum_{j=1}^{n} w_j^2 E\left(Y^2 \mathbf{1}_{\{|Y| \leq n^2\}}\right) = o(w_n n^2) \tag{2}$$

and by Theorem 1 (Adler and Rosalsky, 1991) (see Section 2.7), we have that

$$\frac{\sum_{j=1}^{n} w_j \left(Y_j - EY\mathbf{1}_{\{|Y| \leq n^2\}}\right)}{w_n n^2} \xrightarrow{p} 0. \tag{3}$$

□

Theorem 2. *Let* $\{Y_n,\ n > 1\}$ *be independent F-distributed random variables obtained from site-level testing via limma's moderated t-statistic and let* $\{w_n,\ n > 1\}$ *be constants satisfying* $0 < w_n < 1$ *and* $\sum_{j=1}^{n} w_j = 1$. *At CpG site* x_i *for observed* y_i, *define* $S_{yi}(x_i) = y_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} w_j Y_j$. *Then*

$$y_i + \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} w_j \left(Y_j - E(Y)\right)}{\sqrt{\sum_{\substack{j=1 \\ j \neq i}}^{n} w_j^2 \sigma_j^2}} \xrightarrow{d} y_i + Z \tag{4}$$

where $Var\left(\sum_{\substack{j=1 \\ j \neq i}}^{n} w_j Y_j\right) = \sum_{\substack{j=1 \\ j \neq i}}^{n} w_j^2 \sigma_j^2$ *and* $Z \sim N(0,1)$

Proof. It suffices to show that $\frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} w_j (Y_j - E(Y))}{\sqrt{\sum_{\substack{j=1 \\ j \neq i}}^{n} w_j^2 \sigma_j^2}} \xrightarrow{d} Z$. Define $U_j = \sum_{\substack{j=1 \\ j \neq i}}^{n} w_j \left(Y_j - E(Y)\right)$ so that $E(U_j) = 0$ and $Var(U_j) = \sum_{\substack{j=1 \\ j \neq i}}^{n} w_j^2 \sigma_j^2 := s_n^2$

Applying Lyapunov's CLT condition (Billingsley, 1986; Resnick, 1999) and taking $\delta = 1$ we only need to show that

$$\frac{\sum_{j=1}^{n} E|U_j|^3}{s_n^3} \to 0 \tag{5}$$

To complete the proof we need the concept of uniform tightness. See Definition 7, Lemma 1 and the proof that shows the family of F distributed random variables is uniformly tight. By Lemma 1, U_j is uniformly bounded by M_0. Then,

$$\frac{\sum_{j=1}^{n} E|U_j|^3}{s_n^3} \leq \frac{\sum_{j=1}^{n} M_0 E|U_j|^2}{s_n^3} = \frac{M_0}{s_n} \to 0 \text{ as } n \to \infty$$

□

2. MISCELLANEOUS RESULT

Corollary 1. *Let $Y \sim F_{(1,\nu)}$. Then as $\nu \to \infty$,*

$$Y \xrightarrow{d} \chi_1^2 \tag{6}$$

where χ_1^2 is a chi-squared distribution with 1 degree of freedom.

Proof. Define $F = \dfrac{U/\mu}{V/\nu}$. If U be a chi-square random variable with μ degrees of freedom, V be a chi-square random variable with ν degrees of freedom and U and V be independent then by definition F is an F-random variable. It only suffices to show that with $\mu = 1$, as $\nu \to \infty$, $\frac{V}{\nu} \xrightarrow{p} 1$. By Chebychev's inequality,

$$\begin{aligned} P\left(\left|\frac{V}{\nu} - 1\right| > \varepsilon\right) &\leq \frac{E\left(\frac{V}{\nu} - 1\right)^2}{\varepsilon^2} \\ &= \frac{E(V-\nu)^2}{\nu^2\varepsilon^2} \\ &= \frac{Var(V)}{\nu^2\varepsilon^2} \\ &= \frac{2}{\nu\varepsilon^2} \to 0 \qquad as \qquad \nu \to \infty. \end{aligned}$$

The result, $Y \xrightarrow{d} \chi_1^2$ follows by Slutsky's theorem. □

3. STEPS TO OBTAIN CHOL DATA FROM THE CANCER GENOME ATLAS PROGRAM (TCGA)

We outline briefly steps to obtaining the Cholangiocarcinoma (CHOL) data set used in our simulation for paper I.

- Use the link (`https://portal.gdc.cancer.gov/repository`) to access the TCGA GDC data portal repository.
- Under the `Cases` tab on the left, select `TCGA` under Program. Select `TCGA-CHOL` from Project section.
- Under the `Files` tab and section Data Format, select `idat`. You should see 90 idat files.
- Download the manifest file and using the GDC transfer tool, download data.

After cleaning, only 18 sample names matched which resulted in 36 idat files.

4. DESCRIPTION OF SIMULATED DATA

We simulated the data for paper I in a manner similar to Peters *et al.* (2015); therefore, we refer the reader to the supplementary text of Peters *et al.* (2015), as the simulation description outlined in this paper is based on their ideas. We outline these below:

1. On the basis of the 450K array containing 485,512 probes, we constructed an empirical methylation data set.

   (a) Methylated and Unmethylated Probes

   Probes were classified as fully methylated or unmethylated using TCGA data (n = 18 samples) on bile duct cancer (CHOL) (see Section 2 of the appendix). Probes with an average beta greater than 0.5 after normalization

using the functional normalization method (Fortin *et al.*, 2014) were classified as fully methylated; otherwise, they were classified as unmethylated. The fully methylated to unmethylated ratio was 55.5% to 44.5%.

(b) Define Candidate Differentially Methylated Regions (DMRs)

The candidate DMRs were defined as genomics regions with probes annotated as "TSS200" or "TSS1500" that were no more than 1000 base pairs apart. We obtained 21,363 candidate DMRs with probe counts ranging from 1 to 88, with a median of 6 probes and an average of 6.55 probes.

2. For each simulated data set, we randomly assigned 5% of the 21,363 (i.e., 1068) candidate DMRs to be true hypermethylated DMRs and 5% to be true hypomethylated DMRs. The remaining 90% of the candidate regions were not true DMRs. We set the true methylation difference to be 0.2 (a large methylation difference) as do Peters *et al.* (2015) and also investigate a small methylation difference equal to 0.09. We simulated random values from a uniform distribution to represent the mode of a beta distribution as follows. We generated two values from the uniform distribution to act as beta modes for each region, one for treatment and one for control: beta1 $\sim$ Uniform(0.01, 0.79) and beta2 = beta1 + 0.2. As a result, beta2 $\sim$ Uniform (0.21, 0.99). For hypermethylated regions, the control samples' base methylation level was set to beta1, and the treatment samples' base methylation level was set to beta2. For hypomethylated regions, this allocation was reversed.

3. For the probes inside the selected DMRs, we simulate $\beta-$values representing the proportion of methylation for control and treatment samples from beta (a, b) with a and b obtained using the mode in step (2) and $a + b + 2 = K = 100$.

This value was selected as a level of variability in the sampling distribution that is consistent with reality (Peters *et al.*, 2015). Given K and the mode, the following R code simulates the random $\beta-$values:

```
r <- mode/(1 - mode)
B <- K/(1+r)
A <- r*B
a <- A + 1
b <- B + 1
beta <- rbeta(a=a, b=b)
```

4. For the probes outside the selected DMRs, we generate $\beta-$values from two beta distributions using the methylated or unmethylated status described in step 1. We sample from
   `rbeta(a = 14, b = 3.12)` and `rbeta(a = 2, b = 11.11)`. These values of a and b were chosen to be reasonably close to Peters *et al.* (2015).

5. The aforementioned steps yielded a data set of 485,512 rows and 20 columns (10 treatment, 10 controls). The entire simulation was repeated 1000 times.

REFERENCES

'Illumina methylation beadchips achieve breadth of coverage using 2 Infinium chemistries,' Technical Report Pub. No. 270-2012-00, Illumina, Inc..,, 2015, techsupport@illumina.com.

Aalen, O., 'A model for nonparametric regression analysis of counting processes,' in 'Mathematical statistics and probability theory,' pp. 1–25, Springer, 1980.

Aalen, O., Borgan, O., and Gjessing, H., *Survival and event history analysis: a process point of view*, Springer Science & Business Media, 2008.

Aalen, O. O., 'A linear regression model for the analysis of life times,' Statistics in medicine, 1989, **8**(8), pp. 907–925.

Adler, A. and Rosalsky, A., 'On the weak law of large numbers for normed weighted sums of iid random variables,' International Journal of Mathematics and Mathematical Sciences, 1991, **14**(1), pp. 191–202.

Amico, M. and Van Keilegom, I., 'Cure models in survival analysis,' Annual Review of Statistics and Its Application, 2018, **5**, pp. 311–342.

Amico, M., Van Keilegom, I., and Legrand, C., 'The single-index/cox mixture cure model,' Biometrics, 2019, **75**(2), pp. 452–462.

Anderssen, R. and Bloomfield, P., 'A time series approach to numerical differentiation,' Technometrics, 1974, **16**(1), pp. 69–75.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A., 'Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays,' Bioinformatics, 2014, **30**(10), pp. 1363–1369, doi:10.1093/bioinformatics/btu049.

Barfield, R. T., Kilaru, V., Smith, A. K., and Conneely, K. N., 'Cpgassoc: an r function for analysis of dna methylation microarray data,' Bioinformatics, 2012, **28**(9), pp. 1280–1281.

Basu, B., Chakraborty, J., Chandra, A., Katarkar, A., Baldevbhai, J. R. K., Dhar Chowdhury, D., Ray, J. G., Chaudhuri, K., and Chatterjee, R., 'Genome-wide DNA methylation profile identified a unique set of differentially methylated immune genes in oral squamous cell carcinoma patients in India,' Clin Epigenetics, 2017, **9**, p. 13.

Benjamini, Y. and Hochberg, Y., 'Controlling the false discovery rate: a practical and powerful approach to multiple testing,' Journal of the Royal statistical society: series B (Methodological), 1995a, **57**(1), pp. 289–300.

Benjamini, Y. and Hochberg, Y., 'Controlling the false discovery rate: a practical and powerful approach to multiple testing,' Journal of the Royal statistical society: series B (Methodological), 1995b, **57**(1), pp. 289–300.

Bennett, S., 'Analysis of survival data by the proportional odds model,' Statistics in medicine, 1983, **2**(2), pp. 273–277.

Berkson, J. and Gage, R. P., 'Survival curve for cancer patients following treatment,' Journal of the American Statistical Association, 1952, **47**(259), pp. 501–515.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., *et al.*, 'High density dna methylation array with single cpg site resolution,' Genomics, 2011, **98**(4), pp. 288–295.

Billingsley, P., *Probability and Measure*, John Wiley and Sons, second edition, 1986.

Billingsley, P., *Probability and measure*, John Wiley & Sons, 1995.

Boag, J. W., 'Maximum likelihood estimates of the proportion of patients cured by cancer therapy,' Journal of the royal statistical society series b-methodological, 1949, doi:10.1111/j.2517-6161.1949.tb00020.x.

Breton-Larrivée, M., Elder, E., and McGraw, S., 'DNA methylation, environmental exposures and early embryo development,' Anim Reprod, Oct 2019, **16**(3), pp. 465–474.

Butcher, L. M. and Beck, S., 'Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data,' Methods, Jan 2015, **72**, pp. 21–28.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P., 'Generalized partially linear single-index models,' Journal of the American Statistical Association, 1997, **92**(438), pp. 477–489.

Casella, G. and Berger, R. L., *Statistical inference*, Cengage Learning, 2021.

Cedar, H., 'Dna methylation and gene activity.' Cell, 1988, **53**(1), pp. 3–4.

Center for Cancer Genomics - National Cancer Institute, 'The Cancer Genome Atlas Program (TCGA),' Retrieved from `https://www.cancer.gov/ccg/research/genome-sequencing/tcga`, n.d., accessed June 8, 2023.

Chen, D. P., Lin, Y. C., and Fann, C. S., 'Methods for identifying differentially methylated regions for sequence- and array-based data,' Brief. Funct. Genomics, 2016, **15**(6), pp. 485–490, ISSN 20412657, doi:10.1093/bfgp/elw018.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A., 'Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool,' BMC Bioinformatics, Apr 2013, **14**, p. 128.

Chen, Y. A., Choufani, S., Grafodatskaya, D., Butcher, D. T., Ferreira, J. C., and Weksberg, R., 'Cross-reactive DNA microarray probes lead to false discovery of autosomal sex-associated DNA methylation,' Am J Hum Genet, Oct 2012, **91**(4), pp. 762–764.

Chen, Y. Q. and Wang, M.-C., 'Analysis of accelerated hazards models,' Journal of the American Statistical Association, 2000, **95**(450), pp. 608–618.

Chiou, S. H., Austin, M. D., Qian, J., and Betensky, R. A., 'Transformation model estimation of survival under dependent truncation and independent censoring,' Statistical methods in medical research, 2019, **28**(12), pp. 3785–3798.

Cox, D. R., 'Regression models and life-tables,' J. Roy. Statist. Soc. Ser. B, 1972, **34**, pp. 187–220, ISSN 0035-9246.

Cox, D. R., 'Partial likelihood,' Biometrika, 1975, **62**(2), pp. 269–276.

Cox, D. R. and Oakes, D., *Analysis of survival data*, Chapman and Hall/CRC, 1984.

Cramér, H., *Mathematical Methods of Statistics (PMS-9), Volume 9*, Princeton university press, 2016.

Crary-Dooley, F. K., Tam, M. E., Dunaway, K. W., Hertz-Picciotto, I., Schmidt, R. J., and LaSalle, J. M., 'A comparison of existing global dna methylation assays to low-coverage whole-genome bisulfite sequencing for epidemiological studies,' Epigenetics, 2017, **12**(3), pp. 206–214.

Craven, P. and Wahba, G., 'Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation,' Numerische mathematik, 1978, **31**(4), pp. 377–403.

DasGupta, A., *Asymptotic theory of statistics and probability*, Springer Science & Business Media, 2008.

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F., 'Evaluation of the Infinium Methylation 450K technology,' Epigenomics, Dec 2011, **3**(6), pp. 771–784.

Dempster, A. P., Laird, N. M., and Rubin, D. B., 'Maximum likelihood from incomplete data via the em algorithm,' Journal of the Royal Statistical Society: Series B (Methodological), 1977, **39**(1), pp. 1–22.

Dirick, L., Claeskens, G., Vasnev, A., and Baesens, B., 'A hierarchical mixture cure model with unobserved heterogeneity for credit risk,' Econometrics and Statistics, 2022, **22**, pp. 39–55.

Du, P., Zhang, X., Huang, C. C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M., 'Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis,' BMC Bioinformatics, Nov 2010, **11**, p. 587.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S., 'DNA methylation profiling of human chromosomes 6, 20 and 22,' Nat Genet, Dec 2006, **38**(12), pp. 1378–1385.

Eden, S. and Cedar, H., 'Role of DNA methylation in the regulation of transcription,' Current opinion in genetics & development, 1994, **4**(2), pp. 255–259.

Efron, B. and Morris, C., 'Empirical bayes on vector observations: An extension of stein's method,' Biometrika, 1972, **59**(2), pp. 335–347.

Ehrlich, M. and Wang, R. Y.-H., '5-methylcytosine in eukaryotic DNA,' Science, 1981, **212**(4501), pp. 1350–1357.

Eilers, P. H. and Marx, B. D., 'Flexible smoothing with b-splines and penalties,' Statistical science, 1996, **11**(2), pp. 89–121.

ENCODE Project Consortium, 'An integrated encyclopedia of dna elements in the human genome,' Nature, 2012, **489**(7414), p. 57, doi:10.1038/nature11247.

Farewell, V. T., 'A model for a binary variable with time-censored observations,' Biometrika, 1977, **64**(1), pp. 43–46.

Farewell, V. T., 'The use of mixture models for the analysis of survival data with long-term survivors,' Biometrics, 1982, pp. 1041–1046.

Felizzi, F., Paracha, N., Pöhlmann, J., and Ray, J., 'Mixture cure models in oncology: a tutorial and practical guidance,' PharmacoEconomics-Open, 2021, **5**, pp. 143–155.

Fernandez, A., O'Leary, C., O'Byrne, K. J., Burgess, J., Richard, D. J., and Suraweera, A., 'Epigenetic mechanisms in dna double strand break repair: A clinical review,' Frontiers in Molecular Biosciences, 2021, **8**, p. 685440.

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., Greenwood, C. M., and Hansen, K. D., 'Functional normalization of 450k methylation array data improves replication in large cancer studies,' Genome Biology, 2014, **15**(12), p. 503, doi:10.1186/s13059-014-0503-2.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L., 'A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands,' Proc Natl Acad Sci U S A, Mar 1992, **89**(5), pp. 1827–1831.

Greenberg, M. V. and Bourc'his, D., 'The diverse roles of DNA methylation in mammalian development and disease,' Nature reviews Molecular cell biology, 2019, **20**(10), pp. 590–607.

Haibo, H. and Yunqian, M., 'Imbalanced learning: foundations, algorithms, and applications,' Wiley-IEEE Press, 2013, **1**, p. 27.

Hanin, L. and Huang, L.-S., 'Identifiability of cure models revisited,' Journal of Multivariate Analysis, 2014, **130**, pp. 261–274.

Hardle, W., Hall, P., and Ichimura, H., 'Optimal smoothing in single-index models,' The annals of Statistics, 1993, **21**(1), pp. 157–178.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., *et al.*, *Nonparametric and semiparametric models*, volume 1, Springer, 2004.

Härdle, W. K. *et al.*, *Smoothing techniques: with implementation in S*, Springer Science & Business Media, 1991.

Heiss, J. A. and Just, A. C., 'Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses,' Clin Epigenetics, 01 2019, **11**(1), p. 15.

Hsu, W.-W., Todem, D., and Kim, K., 'A sup-score test for the cure fraction in mixture models for long-term survivors,' Biometrics, 2016, **72**(4), pp. 1348–1357.

Huffer, F. W. and McKeague, I. W., 'Weighted least squares estimation for aalen's additive risk model,' Journal of the American Statistical Association, 1991, **86**(413), pp. 114–129.

Huster, W. J., Brookmeyer, R., and Self, S. G., 'Modelling paired survival data with covariates,' Biometrics, 1989, pp. 145–156.

Ichiyanagi, T., Ichiyanagi, K., Miyake, M., and Sasaki, H., 'Accumulation and loss of asymmetric non-CpG methylation during male germ-cell development,' Nucleic Acids Res, Jan 2013, **41**(2), pp. 738–745.

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A., 'Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies,' Int J Epidemiol, Feb 2012, **41**(1), pp. 200–209.

James, G., Witten, D., Hastie, T., and Tibshirani, R., *An introduction to statistical learning*, volume 112, Springer, 2013.

Jeong, M., Guzman, A. G., and Goodell, M. A., 'Genome-wide analysis of dna methylation in hematopoietic cells: Dna methylation analysis by wgbs,' Acute Myeloid Leukemia: Methods and Protocols, 2017, pp. 137–149.

Jiang, J., *Large sample techniques for statistics*, Springer Science & Business Media, 2010.

Jin, Z. and Liu, Y., 'DNA methylation in human diseases,' Genes Dis., 2018, **5**(1), pp. 1–8, ISSN 23523042, doi:10.1016/j.gendis.2018.01.002.

Kalbfleisch, J. D. and Prentice, R. L., 'Marginal likelihoods based on cox's regression and life model,' Biometrika, 1973, **60**(2), pp. 267–278.

Kalbfleisch, J. D. and Prentice, R. L., *The statistical analysis of failure time data*, John Wiley & Sons, 2011.

Kaplan, E. L. and Meier, P., 'Nonparametric estimation from incomplete observations,' Journal of the American statistical association, 1958, **53**(282), pp. 457–481.

Khaliq, A., Waqas, A., Nisar, Q. A., Haider, S., and Asghar, Z., 'Application of ai and robotics in hospitality sector: A resource gain and resource loss perspective,' Technology in Society, 2022, **68**, p. 101807.

Kilaru, V., Barfield, R. T., Schroeder, J. W., Smith, A. K., and Conneely, K. N., 'MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data,' Epigenetics, Mar 2012, **7**(3), pp. 225–229.

Klein, J. P. and Moeschberger, M. L., *Survival analysis: techniques for censored and truncated data*, volume 1230, Springer, 2003.

Kleinbaum, D. G. and Klein, M., *Survival analysis a self-learning text*, Springer, third edition, 2012.

Kuk, A. Y. and Chen, C.-H., 'A mixture model combining logistic regression with proportional hazards regression,' Biometrika, 1992, **79**(3), pp. 531–541.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., and Ma'ayan, A., 'Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,' Nucleic Acids Res, 07 2016, **44**(W1), pp. W90–97.

Laird, P. W., 'Principles and challenges of genome-wide dna methylation analysis,' Nature Reviews Genetics, 2010, **11**(3), pp. 191–203.

Lao, V. V. and Grady, W. M., 'Epigenetics and colorectal cancer,' Nat Rev Gastroenterol Hepatol, Oct 2011, **8**(12), pp. 686–700.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Sung, K. W. K., Rigoutsos, I., Loring, J., and Wei, C. L., 'Dynamic changes in the human methylome during differentiation,' Genome Res., 2010, **20**(3), pp. 320–331, ISSN 10889051, doi:10.1101/gr.101907.109.

Lee, E. T. and Wang, J., *Statistical methods for survival data analysis*, volume 476, John Wiley & Sons, 2003.

Leek, J. T. and Storey, J. D., 'Capturing heterogeneity in gene expression studies by surrogate variable analysis,' PLoS Genet, Sep 2007, **3**(9), pp. 1724–1735.

Legrand, C., *Advanced survival models*, Chapman and Hall/CRC, 2021.

Lehmann, E. L., *Elements of large-sample theory*, Springer Science & Business Media, 2004.

Li, C.-S. and Lu, M., 'A lack-of-fit test for generalized linear models via single-index techniques,' Computational Statistics, 2018, **33**, pp. 731–756.

Li, C.-S. and Taylor, J. M., 'Smoothing covariate effects in cure models,' Communications in Statistics-Theory and Methods, 2002, **31**(3), pp. 477–493.

Li, D., Xie, Z., Pape, M. L., and Dye, T., 'An evaluation of statistical methods for DNA methylation microarray data analysis,' BMC Bioinformatics, jul 2015, **16**(1), ISSN 14712105, doi:10.1186/s12859-015-0641-x.

Lin, D. Y. and Ying, Z., 'Semiparametric analysis of the additive risk model,' Biometrika, 1994, **81**(1), pp. 61–71.

Liu, A., Jiang, C., Liu, Q., Yin, H., Zhou, H., Ma, H., and Geng, Q., 'The inverted u-shaped association of caffeine intake with serum uric acid in us adults,' The journal of nutrition, health & aging, 2022, **26**(4), pp. 391–399.

López-Cheda, A., Cao, R., Jácome, M. A., and Van Keilegom, I., 'Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models,' Computational Statistics & Data Analysis, 2017, **105**, pp. 144–165.

Lu, W., 'Maximum likelihood estimation in the proportional hazards cure model,' Annals of the Institute of Statistical Mathematics, 2008, **60**, pp. 545–574.

Ma, Y. and He, H., 'Imbalanced learning: foundations, algorithms, and applications,' 2013.

Madadizadeh, F., Ghanbarnejad, A., Ghavami, V., Bandamiri, M. Z., and Mohammadianpanah, M., 'Applying additive hazards models for analyzing survival in patients with colorectal cancer in fars province, southern iran,' Asian Pacific journal of cancer prevention: APJCP, 2017, **18**(4), p. 1077.

Maghbooli, Z., Larijani, B., Emamgholipour, S., Amini, M., Keshtkar, A., and Pasalar, P., 'Aberrant DNA methylation patterns in diabetic nephropathy,' J Diabetes Metab Disord, 2014, **13**, p. 69.

Maksimovic, J., Gordon, L., and Oshlack, A., 'SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips,' Genome Biology, 2012, **13**(6), p. R44, doi:10.1186/gb-2012-13-6-r44.

Maksimovic, J., Oshlack, A., and Phipson, B., 'Gene set enrichment analysis for genome-wide DNA methylation data,' Genome Biol, 06 2021, **22**(1), p. 173.

Maller, R. A. and Zhou, X., *Survival analysis with long-term survivors*, volume 525, Wiley New York, 1996.

Mallik, S., Odom, G. J., Gao, Z., Gomez, L., Chen, X., and Wang, L., 'An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays,' Brief Bioinform, 11 2019, **20**(6), pp. 2224–2235.

McCartney, D. L., Walker, R. M., Morris, S. W., McIntosh, A. M., Porteous, D. J., and Evans, K. L., 'Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationepic beadchip,' Genomics data, 2016, **9**, pp. 22–24.

McCullagh, P. and Nelder, J. A., *Generalized linear models*, Routledge, 2019.

McGregor, K., Bernatsky, S., Colmegna, I., Hudson, M., Pastinen, T., Labbe, A., and Greenwood, C. M., 'An evaluation of methods correcting for cell-type heterogeneity in dna methylation studies,' Genome biology, 2016, **17**(1), pp. 1–17.

McLachlan, G. J. and Krishnan, T., *The EM algorithm and extensions*, John Wiley & Sons, 2007.

Mood, A., Graybill, F., and Boes, D., '(1974), introduction to the theory of statistics,' 1974.

Moran, S., Arribas, C., and Esteller, M., 'Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences,' Epigenomics, 2016, **8**(3), pp. 389–399.

Patilea, V. and Van Keilegom, I., 'A general approach for cure models in survival analysis,' 2020.

Peng, Y. and Dear, K. B., 'A nonparametric mixture model for cure rate estimation,' Biometrics, 2000, **56**(1), pp. 237–243.

Peng, Y. and Yu, B., *Cure Models: Methods, Applications, and Implementation*, Chapman and Hall/CRC, 2021.

Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., V Lord, R., Clark, S. J., and Molloy, P. L., 'De novo identification of differentially methylated regions in the human genome,' Epigenetics Chromatin, 2015, **8**, p. 6.

Piao, Y., Xu, W., Park, K. H., Ryu, K. H., and Xiang, R., 'Comprehensive Evaluation of Differential Methylation Analysis Methods for Bisulfite Sequencing Data,' Int J Environ Res Public Health, 07 2021, **18**(15).

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., and Schalkwyk, L. C., 'A data-driven approach to preprocessing illumina 450k methylation array data,' BMC genomics, 2013, **14**(1), pp. 1–10.

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., and Clark, S. J., 'Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling,' Genome Biol, 10 2016, **17**(1), p. 208.

Price, D., *Survival Models for Heterogeneous Populations with Cure*, Ph.D. thesis, Emory University, 2000.

Procter, M., Chou, L.-S., Tang, W., Jama, M., and Mao, R., 'Molecular diagnosis of prader–willi and angelman syndromes by methylation-specific melting analysis and methylation-specific multiplex ligation-dependent probe amplification,' Clinical chemistry, 2006, **52**(7), pp. 1276–1283.

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., and Jaenisch, R., 'Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a,' Proc. Natl. Acad. Sci. U. S. A., 2000, **97**(10), pp. 5237–5242, ISSN 00278424, doi:10.1073/pnas.97.10.5237.

Resnick, S. I., *A probability path*, Birkhäuser Boston, 1999.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K., 'limma powers differential expression analyses for RNA-sequencing and microarray studies,' Nucleic Acids Res, Apr 2015, **43**(7), p. e47.

Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X., 'Statistical methods for detecting differentially methylated loci and regions,' Frontiers in genetics, 2014, **5**, p. 324.

Rossignol, S., Steunou, V., Chalas, C., Kerjean, A., Rigolet, M., Viegas-Pequignot, E., Jouannet, P., Le Bouc, Y., and Gicquel, C., 'The epigenetic imprinting defect of patients with beckwith—wiedemann syndrome born after assisted reproductive technology is not restricted to the 11p15 region,' Journal of medical genetics, 2006, **43**(12), pp. 902–907.

Roy, N. K., Monisha, J., Padmavathi, G., Lalhruaitluanga, H., Kumar, N. S., Singh, A. K., Bordoloi, D., Baruah, M. N., Ahmed, G. N., Longkumar, I., Arfuso, F., Kumar, A. P., and Kunnumakkara, A. B., 'Isoform-Specific Role of Akt in Oral Squamous Cell Carcinoma,' Biomolecules, 06 2019, **9**(7).

Ruppert, D., Wand, M. P., and Carroll, R. J., *Semiparametric regression*, 12, Cambridge university press, 2003.

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., and Esteller, M., 'Validation of a dna methylation microarray for 450,000 cpg sites in the human genome,' Epigenetics, 2011, **6**(6), pp. 692–702.

Satterthwaite, F. E., 'An approximate distribution of estimates of variance components,' Biometrics, Dec 1946, **2**(6), pp. 110–114.

Shafi, A., Mitrea, C., Nguyen, T., and Draghici, S., 'A survey of the approaches for identifying differential methylation using bisulfite sequencing data,' Brief Bioinform, 09 2018, **19**(5), pp. 737–753.

Shang, S., Liu, M., Zeleniuch-Jacquotte, A., Clendenen, T. V., Krogh, V., Hallmans, G., and Lu, W., 'Partially linear single index cox regression model in nested case-control studies,' Computational statistics & data analysis, 2013, **67**, pp. 199–212.

Shiah, Y. J., Fraser, M., Bristow, R. G., and Boutros, P. C., 'Comparison of pre-processing methods for Infinium HumanMethylation450 BeadChip array,' Bioinformatics, Oct 2017, **33**(20), pp. 3151–3157.

Shu, C., Zhang, X., Aouizerat, B. E., and Xu, K., 'Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells,' Epigenetics Chromatin, 11 2020, **13**(1), p. 51.

Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, volume 26, CRC Press, 1986.

Smith, M. L., Baggerly, K. A., Bengtsson, H., Ritchie, M. E., and Hansen, K. D., 'illuminaio: An open source idat parsing tool for illumina microarrays,' F1000Research, 2013, **2**.

Smyth, G. K., 'Linear models and empirical bayes methods for assessing differential expression in microarray experiments,' Stat Appl Genet Mol Biol, 2004, **3**, p. Article3.

Sofer, T., Schifano, E. D., Hoppin, J. A., Hou, L., and Baccarelli, A. A., 'A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure,' Bioinformatics, Nov 2013, **29**(22), pp. 2884–2891.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M., 'The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1,' 1949.

Sun, H. and Wang, S., 'Penalized logistic regression for high-dimensional DNA methylation data with case-control studies,' Bioinformatics, May 2012, **28**(10), pp. 1368–1375.

Sun, L., Namboodiri, S., Chen, E., and Sun, S., 'Preliminary Analysis of Within-Sample Co-methylation Patterns in Normal and Cancerous Breast Samples,' Cancer Inform, 2019, **18**, p. 1176935119880516.

Sun, L. and Sun, S., 'Within-sample co-methylation patterns in normal tissues,' BioData Min, 2019, **12**, p. 9.

Sun, S., Dammann, J., Lai, P., and Tian, C., 'Thorough statistical analyses of breast cancer co-methylation patterns,' BMC Genom Data, 04 2022, **23**(1), p. 29.

Susan, J. C., Harrison, J., Paul, C. L., and Frommer, M., 'High sensitivity mapping of methylated cytosines,' Nucleic acids research, 1994, **22**(15), pp. 2990–2997.

Sy, J. P. and Taylor, J. M., 'Estimation in a cox proportional hazards cure model,' Biometrics, 2000, **56**(1), pp. 227–236.

Szyf, M., 'Dna methylation signatures for breast cancer classification and prognosis,' Genome medicine, 2012, **4**(3), pp. 1–12.

Taylor, H. L., 'Physical activity: is it still a risk factor?' Preventive medicine, 1983, **12**(1), pp. 20–24.

Taylor, J. M., 'Semi-parametric estimation in failure time mixture models,' Biometrics, 1995, pp. 899–907.

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S., 'A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data,' Bioinformatics, Jan 2013, **29**(2), pp. 189–196.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 'A promoter-level mammalian expression atlas,' Nature, 2014, **507**(7493), pp. 462–470.

Therneau, T. M., Grambsch, P. M., Therneau, T. M., and Grambsch, P. M., *The cox model*, Springer, 2000.

Touleimat, N. and Tost, J., 'Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation,' Epigenomics, Jun 2012, **4**(3), pp. 325–341.

Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., and Siegmund, K. D., 'Low-level processing of Illumina Infinium DNA Methylation BeadArrays,' Nucleic Acids Res, Apr 2013, **41**(7), p. e90.

Wang, D., Yan, L., Hu, Q., Sucheston, L. E., Higgins, M. J., Ambrosone, C. B., Johnson, C. S., Smiraglia, D. J., and Liu, S., 'Ima: an r package for high-throughput analysis of illumina's 450k infinium methylation data,' Bioinformatics, 2012, **28**(5), pp. 729–730.

Wang, L. and Cao, G., 'Efficient estimation for generalized partially linear single-index models,' 2018.

Wang, T., Guan, W., Lin, J., Boutaoui, N., Canino, G., Luo, J., Celedón, J. C., and Chen, W., 'A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data,' Epigenetics, 2015, **10**(7), pp. 662–669.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Foekens, J. A., 'Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,' Lancet, 2005, **365**(9460), pp. 671–679.

Wang, Z., Wu, X., and Wang, Y., 'A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip,' BMC Bioinformatics, 04 2018, **19**(Suppl 5), p. 115.

Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y. C., 'COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis,' Nucleic Acids Res, Jun 2013a, **41**(11), p. e117.

Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y.-C., 'Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis,' Nucleic acids research, 2013b, **41**(11), pp. e117–e117.

Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R., Dymov, S., Szyf, M., and Meaney, M. J., 'Epigenetic programming by maternal behavior,' Nat Neurosci, Aug 2004, **7**(8), pp. 847–854.

Weisenberger, D., Van Den Berg, D., Pan, F., Berman, B., and Laird, P., 'Comprehensive dna methylation analysis on the illumina infinium assay platform,' Illumina, San Diego, 2008.

Wood, S., 'mgcv: Mixed gam computation vehicle with gcv/aic/reml smoothness estimation,' 2012.

Wood, S. N., 'Thin plate regression splines,' Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2003, **65**(1), pp. 95–114.

Wu, D., Gu, J., and Zhang, M. Q., 'Fastdma: an infinium humanmethylation450 beadchip analyzer,' PloS one, 2013, **8**(9), p. e74275.

Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., and Ma'ayan, A., 'Gene Set Knowledge Discovery with Enrichr,' Curr Protoc, Mar 2021, **1**(3), p. e90.

Xu, J. and Peng, Y., 'Nonparametric cure rate estimation with covariates,' Canadian Journal of Statistics, 2014, **42**(1), pp. 1–17.

Yu, Y. and Ruppert, D., 'Penalized spline estimation for partially linear single-index models,' Journal of the American Statistical Association, 2002, **97**(460), pp. 1042–1054.

Yu, Y., Wu, C., and Zhang, Y., 'Penalised spline estimation for generalised partially linear single-index models,' Statistics and Computing, 2017, **27**, pp. 571–582.

Zeng, Z., Gao, Y., Li, J., Zhang, G., Sun, S., Wu, Q., Gong, Y., and Xie, C., 'Violations of proportional hazard assumption in cox regression model of transcriptomic data in tcga pan-cancer cohorts,' Computational and Structural Biotechnology Journal, 2022, **20**, pp. 496–507.

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E., 'Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements,' Genome Biol, Jan 2015, **16**, p. 14.

Zhang, Y., Liu, H., Lv, J., Xiao, X., Zhu, J., Liu, X., Su, J., Li, X., Wu, Q., Wang, F., *et al.*, 'Qdmr: a quantitative method for identification of differentially methylated regions by entropy,' Nucleic acids research, 2011, **39**(9), pp. e58–e58.

Zhang, Y., Wang, S., and Wang, X., 'Data-Driven-Based Approach to Identifying Differentially Methylated Regions Using Modified 1D Ising Model,' Biomed Res. Int., 2018, **2018**, ISSN 23146141, doi:10.1155/2018/1070645.

Zhao, Y., Lee, A. H., Yau, K. K., Burke, V., and McLachlan, G. J., 'A score test for assessing the cured proportion in the long-term survivor mixture model,' Statistics in medicine, 2009, **28**(27), pp. 3454–3466.

VITA

In Fall of 2013, Daniel Ahmed Alhassan completed his Bachelor of Science (2013) degree in Actuarial Science and later he worked as a Teaching Assistant in the same institution teaching statistics courses. In Fall of 2015, he moved to the United States to begin his master's degree in Mathematics. After receiving his Master of Science (2017) degree, he enrolled at the Missouri University of Science and Technology (Missouri S & T) to earn a Ph.D in Mathematics with statistics emphasis which he received in July 2023. While at Missouri S & T, Daniel was employed as a teaching assistant in the Department of Mathematics and Statistics. During those six years he taught different courses at various math/statistics levels.