Smooth modelling of covariate effects in bisulfite sequencing-derived measures of DNA methylation

Kaiqiong Zhao

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montréal, Québec, Canada
October 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Acknowledgements

The completion of this thesis would never have been possible without the support of many remarkable people.

First, my gratitude to my co-supervisors, Dr. Celia Greenwood and Dr. Karim Oualkacha, cannot be overstated. Their insightful guidance, passion and enthusiasm for research, kindness and caring have helped me grow professionally and personally. Throughout my doctoral studies, they provided encouragement, sound advice, good company, and many good ideas. I could not imagine having had a better supervision team during my Ph.D. studies.

Next, I would like to thank Dr. Lajmi Lakhal-Chaieb and Dr. Aurélie Labbe for their inspiring questions, thought-provoking conversations, and valuable suggestions that helped to improve the methodological work in manuscripts I and II. Special thanks to Dr. Yi Yang for sharing his profound expertise in penalized methods and computational algorithms with me. Without his guidance, I could not have completed the work in manuscript III. I would also like to thank our collaborators, Dr. Marie Hudson, Dr. Sasha Bernatsky, Dr. Inés Colmegna, Dr. Tomi Pastinen, Dr. Tieyuan Zhang, and Dr. Denise Daley, for sharing with us the methylation data used in this thesis and providing their insights into autoimmune diseases and epigenetic events. I am deeply grateful to Dr. Kathleen Klein, who preprocessed the raw binary sequencing files and prepared cleaned data matrices for my analysis. I would also like to thank Dr. Antonio Ciampi for his editorial help in revising manuscript I. My sincerest thanks to my thesis examiners, committee members and protocol examiners, Dr. Kimberly Siegmund, Dr. David Stephens, Dr. Taoufik Bouezmarni, Dr. Alexandra Schmidt, Dr. Shirin Golchi, Dr. Guillaume Bourque, and Dr. Michal Abrahamowicz. Their insightful comments and suggestions have made this work more rigorous.

I would also like to acknowledge the McGill Faculty of Medicine's internal studentship–the Gerald Clavet Fellowship for one year's support, and the Fonds de recherche du Québec–

Santé for offering me a three-year doctoral training award.

I am very fortunate to be part of the Department of Epidemiology, Biostatistics, & Occupational Health (EBOH), which has provided me with a rich learning experience and a highly supportive environment. I wish to express my gratitude to Dr. James Hanley, Dr. Erica Moodie, Dr. Andrea Benedetti and Dr. Alexandra Schmidt. Their courses have helped me discover the charm of Biostatistics. In addition, I want to thank the faculty from the Department of Mathematics and Statistics; many of their courses and seminars have helped build my statistical foundation. I am grateful to Dr. Sahir Bhatnagar for teaching me how to write an R package and to Dr. Yue Li for introducing Rcpp to me. I would also like to thank my Biostatistics peers for providing a stimulating and fun environment to learn and grow, including (but not limited to): Kevin McGregor, Janie Coulombe, Gabrielle Simoneau, Yu Luo, Ting Zhang, Guanbo Wang, and Yi Lian.

I acknowledge the usage of Compute Canada and Lady Davis Institute (LDI) computational facilities, which enabled the simulations and data analyses in this thesis. I want to extend my sincere gratitude to the staff from the Laday Davis Institute and the EBOH department, who have always been friendly and willing to assist without a second thought. I would like to highlight the excellent technical support from Dr. Kathleen Klein and Dr. Vincenzo Forgetta and the administrative support from Ms. Katherine Hayden and Ms. Darin Adra. Special thanks to my colleagues and friends at LDI for their companionship, relaxing talks and abundant laughter: Sirui, Haoyu, Tomoko, Yiheng, Ting, Yixiao, Lai, Tianyuan and Amadou.

Finally, I would like to thank my family, Mom, Dad, Xiaorui, Lirong and Zige. Although we are ten thousand kilometres apart, your constant love and support have always been my motivation. Thank you, Jinwen, for your encouragement and trust along my Ph.D. journey.

Preface & Contribution of Authors

This thesis comprises three original scholarly manuscripts. Kaiqiong Zhao is the primary author on all three manuscripts. The contributions of the authors are as follows:

Manuscript I: Kaiqiong Zhao, Karim Oualkacha, Lajmi Lakhal-Chaieb, Aurélie Labbe, Kathleen Klein, Antonio Ciampi, Marie Hudson, Inés Colmegna, Tomi Pastinen, Tieyuan Zhang, Denise Daley, Celia M.T. Greenwood (2020). A novel statistical method for modelling covariate effects in bisulfite sequencing derived measures of DNA methylation. *Biometrics*. DOI: 10.1111/biom.13307.

KZ: Conceptualization, Data curation, Methodology, Formal analysis, Simulation, Software, Visualization, Writing – original draft, Writing – review and editing

KO: Conceptualization, Methodology, Writing – review and editing, Funding acquisition, Supervision

LL: Conceptualization, Methodology, Writing – review and editing, Funding acquisition, Supervision

AL: Methodology, Writing – review and editing, Funding acquisition, Supervision

KK: Data curation, Data preprocessing

AC: Methodology, Writing – review and editing

MH: Data curation, Funding acquisition

IC: Data curation, Writing – review and editing, Funding acquisition

TP: Data curation, Funding acquisition

TZ: Data curation, Funding acquisition

DD: Data curation, Funding acquisition

CMTG: Conceptualization, Methodology, Writing – review and editing, Funding acquisition, Supervision

Manuscript II: Kaiqiong Zhao, Karim Oualkacha, Lajmi Lakhal-Chaieb, Aurélie Labbe, Kathleen Klein, Sasha Bernatsky, Marie Hudson, Inés Colmegna, Celia M.T. Greenwood (2021+). A hierarchical quasi-binomial varying coefficient mixed model for detecting differentially methylated regions in bisulfite sequencing data. In preparation for the *Annual of Applied Statistics*.

KZ: Conceptualization, Data curation, Methodology, Formal analysis, Simulation, Software, Visualization, Writing – original draft, Writing – review and editing

KO: Conceptualization, Methodology, Writing – review and editing, Funding acquisition, Supervision

LL: Methodology, Writing – review and editing, Funding acquisition, Supervision

AL: Methodology, Writing – review and editing, Funding acquisition, Supervision

KK: Data curation, Data preprocessing

SB: Data curation, Funding acquisition

MH: Data curation, Funding acquisition

IC: Data curation, Funding acquisition

CMTG: Conceptualization, Methodology, Writing – review and editing, Funding acquisition, Supervision

Manuscript III: Kaiqiong Zhao, Yi Yang, Karim Oualkacha, Celia M.T. Greenwood (2021+). A sparse high-dimensional generalized varying coefficient model for identifying genetic variants associated with regional methylation levels. In preparation for *Biostatistics*.

KZ: Conceptualization, Methodology, Simulation, Software, Visualization, Writing – original draft, Writing – review and editing

YY: Conceptualization, Methodology, Writing – review and editing, Supervision

KO: Conceptualization, Methodology, Writing – review and editing, Supervision

CMTG: Conceptualization, Methodology, Writing – review and editing, Supervision

Abstract

DNA methylation is an essential epigenetic modification that regulates gene activity and contributes to tissue differentiation and disease susceptibility. Identifying disease-associated changes in DNA methylation can help us gain a better understanding of disease etiology. Bisulfite sequencing allows the generation of high-throughput methylation profiles at the single-base resolution of DNA. However, optimally modelling and analyzing these sparse and discrete sequencing data is still challenging due to variable read depth, missing data patterns, long-range correlations, data errors, and confounding from cell type mixtures. This thesis consists of three manuscripts about developing methods to better estimate regional association patterns in bisulfite sequencing-derived DNA methylation data, particularly useful for analyzing data from targeted custom capture sequencing libraries.

In the first manuscript, I develop a novel hierarchical varying coefficient regression method called SmOoth ModeliNg of BisUlfite Sequencing (SOMNiBUS), which allows covariate effects to vary smoothly along genomic positions. I build a specialized Expectation-Maximization algorithm, which allows for measurement errors in the outcomes (i.e. methylated counts) and leads to both regional measures of association and pointwise tests and confidence intervals. Simulations show that the proposed method provides accurate estimates of covariate effects and captures the major underlying methylation patterns with excellent power. I also apply this method to analyze data from cell type-separated blood samples taken from rheumatoid arthritis patients and controls.

In the second manuscript, I extend SOMNiBUS to allow the outcomes to exhibit extra-parametric variation by proposing a hierarchical quasi-binomial varying coefficient mixed model. This model allows for both multiplicative and additive dispersion, thereby providing a plausible representation of realistic dispersion trends observed in regional methylation data. I also propose a hybrid Expectation-Solving algorithm to estimate this model, which explicitly accounts for measurement errors in the outcomes and results in a regional association test statistic with a simple F limiting distribution. I demonstrate the theoretical proper-

ties of the resulting estimators, as well as their marginal and conditional interpretations. I also apply the proposed method to two sets of methylation data, both containing subjects sampled from the CARTaGENE biobank (www.cartagene.qc.ca). The two datasets were both designed to compare individuals with high and low anti-citrullinated protein antibody levels, a biomarker associated with rheumatoid arthritis. Results from simulations and data applications show that the new approach provides accurate estimates of covariate effects and detects covariates that influence methylation levels with excellent power.

The third manuscript focuses on developing a sparse high-dimensional varying coefficient model, intending to identify a subset of the genetic variants with local influence on regional methylation levels. To enable variable selection in varying coefficient models, I propose a composite sparse penalty that encourages both sparsity and smoothness for the varying/non-linear covariate effects. I also present an efficient proximal gradient descent algorithm to obtain the penalized estimation of the varying regression coefficients. Extensive simulations are conducted to evaluate the performance of the proposed approach in terms of estimation, prediction and variable selection.

The methods proposed in the first two manuscripts have been implemented in an R Bioconductor package `SOMNiBUS`. The method developed in the third manuscript has been implemented in a prototype R package `sparseSOMNiBUS`, available in Github.

Abrégé

La méthylation de l'ADN est une modification épigénétique essentielle qui régule l'activité des gènes et contribue à la différenciation des tissus et à la susceptibilité aux maladies. L'identification des modifications de la méthylation de l'ADN associées aux maladies peut nous aider à mieux comprendre l'étiologie des maladies. Le séquençage au bisulfite permet de générer des profils de méthylation à haut débit à la résolution d'une seule base d'ADN. Cependant, la modélisation et l'analyse optimales de ces données de séquençage éparses et discrètes restent un défi en raison, notamment, de la variation dans la profondeur de lecture, de données manquantes, des corrélations spatiales entre les régions génomiques adjacentes, des erreurs de données et des facteurs confondants due aux mélanges de types de cellules. Cette thèse consiste en trois manuscrits portant sur le développement de méthodes permettant de mieux estimer les patrons d'association régionaux dans les données de méthylation de l'ADN issues du séquençage au bisulfite, particulièrement utiles pour l'analyse des données provenant de librairies de séquençage de capture personnalisées et ciblées.

Dans le premier manuscrit, je mets au point une nouvelle méthode de régression hiérarchique à coefficients variables appelée SOMNiBUS (de l'anglais, SmOoth ModeliNg of BisUlfite Sequencing), qui permet aux effets des covariables de varier de façon régulière le long des positions génomiques. Je conçois un algorithme spécialisé d'espérance-maximisation, qui tient compte des erreurs de mesure dans la variable réponse (c'est-à-dire les comptes méthylés). Cet algorithme est à la fois capable de détecter des associations à l'échelle d'une région génomique au complet ainsi que produire des tests ponctuels pour chaque position génomique et des intervalles de confiance. Les simulations montrent que la méthode proposée fournit des estimations précises des effets des covariables et capture les principaux patrons de méthylation sous-jacents avec une excellente puissance. J'applique également cette méthode à des données provenant d'échantillons sanguins séparés par type de cellule, prélevés chez des patients atteints de polyarthrite rhumatoïde et des témoins.

Dans le second manuscrit, je réalise une extension de SOMNiBUS pour permettre à la vari-

able réponse de présenter une variation extra-paramétrique en proposant un modèle mixte hiérarchique quasi-binomial à coefficients variables. Ce modèle permet une dispersion à la fois multiplicative et additive, fournissant ainsi une représentation plausible des tendances réalistes de dispersion observées dans les données régionales de méthylation. Je propose également un algorithme hybride basé sur la technique "Expectation-Solving" pour estimer ce modèle. Cet algorithme tient compte explicitement des erreurs de mesure dans la variable réponse et mène à une statistique de test d'association régionale avec une distribution limite, F, standard. Je démontre les propriétés théoriques des estimateurs résultants, ainsi que leurs interprétations marginales et conditionnelles. J'applique également la méthode proposée à deux ensembles de données de méthylation, contenant tous deux des sujets échantillonnés dans la biobanque CARTaGENE (www.cartagene.qc.ca). Les deux ensembles de données ont été conçus pour comparer des individus présentant des niveaux élevés et faibles d'anticorps anti-protéines citrullinés, un biomarqueur associé à la polyarthrite rhumatoïde. Les résultats des simulations et des applications de données montrent que la nouvelle approche fournit des estimations précises des effets des covariables et détecte les covariables qui influencent les niveaux de méthylation avec une excellente puissance.

Le troisième manuscrit se concentre sur le développement d'un modèle éparse à coefficients variables en présence de données de grande dimension, dans le but d'identifier un sous-ensemble de variants génétiques ayant une influence locale sur les niveaux de méthylation à une région sous étude. Pour permettre la sélection des variants dans un tel modèle à coefficients variables, je propose une pénalité éparse composite qui encourage à la fois la sélection des variables importantes et le lissage de leurs effets non linéaires. Je présente également un algorithme efficace de descente de gradient proximal pour obtenir l'estimation pénalisée des coefficients de régression variables. Des simulations approfondies sont réalisées pour évaluer les performances de l'approche proposée en termes d'estimation, de prédiction et de sélection des variables.

Les méthodes proposées dans les deux premiers manuscrits ont été mises en œuvre dans une

librairie R Bioconductor `SOMNiBUS`. La méthode développée dans le troisième manuscrit a été implémenté dans un prototype de la librairie R `sparseSOMNiBUS`, disponible sur Github.

Table of contents

List of Tables

List of Figures

Abbreviations

ACPA	anti-citrullinated protein antibodies
BS	bisulfite sequencing
C	cytosines
CI	confidence intervals
CpG	cytosine-phosphate-guanine
DMCs	differentially methylated cytosines
DMRs	differentially methylated regions
GAM	Generalized additive models
GC	guanine-cytosine
GCV	generalized cross-validation
GLM	generalized linear model
GLMM	generalized linear mixed model
HMM	Hidden Markov models
kb	kilobase pairs
LASSO	least absolute shrinkage and selection operator
MCC-Seq	Methylation capture sequencing
mQTLs	methylation quantitative trait loci
PCR	polymerase chain reaction
P-IRLS	penalized iteratively reweighted least squares
RA	rheumatoid arthritis
REML	restricted maximum likelihood
SCAD	smoothly clipped absolute deviation penalty
SNPs	single nucleotide polymorphisms
SOMNiBUS	SmOoth ModeliNg of BisUlfite Sequencing
T	thymines
TCCBS	Targeted Custom Capture Bisulfite Sequencing

VC	varying coefficient
WGBS	whole-genome bisulfite sequencing

Chapter 1

Introduction

Although genome-wide association studies have provided valuable insights into the genetic basis of a wide range of human diseases (Buniello et al., 2019; MacArthur et al., 2017), there is still a gap between disease heritability and heritability attributable to genetic variation (Ober & Vercelli, 2011). Environmental exposures are suggested to play a crucial role in explaining the "missing" heritability (Maher, 2008; Trerotola et al., 2015; A. I. Young et al., 2018). Plausibly, such exposures, in interaction with genetic predisposition, may lead to epigenetic modification, which alters gene regulation without changing genome sequence (Jaenisch & Bird, 2003). The rapidly evolving field of epigenetics is contributing to our understanding of gene-environment interactions (Barros & Offenbacher, 2009), and providing novel insights into disease etiology (L. Gu et al., 2015; Z. Zhang et al., 2014) and possible therapies (Jones et al., 2019; Tough et al., 2016).

DNA methylation is the most studied epigenetic modification and involves the addition of a methyl group to the DNA, mostly at cytosine-phosphate-guanine (CpG) sites. Aberrant DNA methylation has been linked to a plethora of human diseases, including neurological disorders (Miranda-Morales et al., 2017; J. I. Young et al., 2019; Zulet et al., 2017), autoimmune disorders (Mazzone et al., 2019; Zouali, 2020) and cancer (Kulis & Esteller, 2010;

Locke et al., 2019).

Measuring large-scale DNA methylation at single-nucleotide resolution is possible owing to the development of bisulfite sequencing protocols (Frommer et al., 1992), which can be implemented either genome-wide or in targeted regions. This thesis focuses on data from targeted custom capture sequencing libraries, with methylation levels measured for CpGs in a set of targeted regions.

The sequencing platforms measure the methylation level at a single site as a pair of counts: the number of methylated reads and the total number of reads aligned to the site, i.e. read depth. Many existing methods convert the counts into proportions and model them using continuous distributions, such as normal distribution (Hansen et al., 2012; Korthauer et al., 2018) or beta distribution (Hebestreit et al., 2013). This conversion, however, can lead to information loss, as it fails to distinguish between noisy and accurate measurements and disregards the discrete nature of the data. In addition, optimally modelling and analyzing these discrete sequencing data can be greatly hindered by the many missing values, the possibility of data errors, and the confounding from cell type mixtures (Khavari et al., 2010; McGregor et al., 2016) or genetic variations (Gaunt et al., 2016; Hannon et al., 2018). The principal focus of this thesis is on developing statistical methods to address these challenges for better estimating regional association patterns in bisulfite sequencing-derived DNA methylation data.

The main content of this thesis is comprised of three manuscripts, corresponding to Chapters 3-5. Each manuscript is presented as a standalone piece of literature. Overall, this thesis is structured as follows. Chapter 2 presents a brief literature review, outlining several unique features in bisulfite sequencing-derived DNA methylation measures, as well as the flexible modelling approaches and their extensions used in this thesis. In Chapter 3, I develop a novel hierarchical varying coefficient regression method called SmOoth ModeliNg of BisUlfite Sequencing (SOMNiBUS) for modelling the association between regional methylation patterns

and multiple covariates. I also build a specialized Expectation-Maximization algorithm, which allows for measurement errors in the outcomes (i.e. methylated counts) and leads to both regional measures of association and pointwise tests and confidence intervals. In Chapter 4, I extend SOMNiBUS to allow the outcomes to exhibit extra-parametric variation by proposing a hierarchical quasi-binomial varying coefficient mixed model. This model allows for both multiplicative and additive dispersion, thereby providing a plausible representation of realistic dispersion trends observed in regional methylation data. I also propose a hybrid Expectation-Solving algorithm accompanied by a plug-in estimator for the scale parameter to estimate this model, which explicitly accounts for measurement errors in the outcomes and results in a regional association test statistic with a simple F limiting distribution. Chapter 5 focuses on a high-dimensional extension to the standard SOMNiBUS, intending to identify a subset of the genetic variants with local influence on regional methylation levels. To enable variable selection, I propose a high-dimensional generalized varying coefficient model accompanied by a composite penalty function that encourages both sparsity and smoothness for the varying coefficients. I also present an efficient proximal gradient descent algorithm to estimate such a model. Finally, Chapter 6 summarizes the contributions of the thesis and discusses future research avenues.

Chapter 3 has been published in *Biometrics* (Zhao et al., 2020). Chapters 4 and 5 will be submitted for publication shortly after the submission of the thesis. The methods proposed in Chapters 3 and 4 have been implemented in an R Bioconductor package `SOMNiBUS` (`https://www.bioconductor.org/packages/release/bioc/html/SOMNiBUS.html`). The method developed in Chapter 5 has been implemented in a prototype R package `sparseSOMNiBUS`, available in Github (`https://github.com/kaiqiong/sparseSOMNiBUS`).

Chapter 2

Literature Review

This literature review consists of four sections. Section 2.1 briefly describes how measurements of an individual's methylation profiles can be obtained. Section 2.2 provides a short overview of generalized additive models, emphasizing their smoothing techniques, estimation and inference methods and underlying assumptions. In Section 2.3, I discuss the phenomenon, termed *overdispersion*, arising when the variance in the data exceeds the nominal variance predicted by the presumed model. Section 2.4 presents two sparse penalization approaches, namely the LASSO and the group LASSO, upon which the methodology development in Chapter 5 builds. Additional discussion of literature relevant to each of the manuscript is provided within the introduction and method sections of Chapters 3-5.

2.1 Measuring DNA methylation

DNA methylation is the most studied epigenetic modification and involves the addition of a methyl group to the DNA, mostly at CpG sites. Approximately, there are 28 million CpG dinucleotides in the human genome, of which 60-80% are methylated (Ziller et al., 2013). It has been demonstrated that CpG sites are unevenly distributed over the genome, and their

methylation patterns are associated with genomic contexts. For example, CpG sites located within gene promoters with high guanine-cytosine (GC) frequencies are generally unmethylated (Choy et al., 2010). Such CpG-dense regions are termed as CpG islands. Typically, hypermethylation of CpG islands in promoters can silence gene expression by preventing transcriptional factor binding to DNA (Choy et al., 2010). In contrast, distal regulatory elements (e.g. enhancers) have relatively low CpG density and hypo- to hemimethylated profiles with more inter-individual and inter-tissue variation (Grundberg et al., 2013; Irizarry et al., 2009). Gene body regions are CpG-poor and vastly methylated (Bock et al., 2012); their hypermethylation can correlate with increased gene expression (Ball et al., 2009). Generally speaking, DNA methylation can either activate or repress gene expression, depending on whether the mark inactivates a positive or negative regulatory element (Jones, 1999). Notably, these methylation patterns can also vary substantially between cell types (Khavari et al., 2010; McGregor et al., 2016), environmental exposures (Karabegović et al., 2021; Stenz et al., 2018) and disease conditions (Robertson, 2005; Skvortsova et al., 2019).

2.1.1 Targeted custom capture bisulfite sequencing

In contrast to microarray-based assays, which primarily target CpG-rich regions such as promoters, whole-genome bisulfite sequencing (WGBS) allows a comprehensive characterization of the genome-wide methylation landscape and is the current gold standard for DNA methylation profiling. However, WGBS is not cost-effective for large-scale studies as only 20% or less of CpGs are known to have variable methylation across individuals or tissues (Ziller et al., 2013). To improve efficiency, Allum et al. (2015) have developed Methylation capture sequencing (MCC-Seq), a next-generation sequencing capture approach for interrogating functional (i.e. regulatory active) CpGs in disease-targeted tissues or cells. Its customizable and flexible design allows prior selection of biologically relevant CpG regions and easy elimination of invariable CpG sites across individuals. For example, its blood cell-specific immune

panels cover the majority of human gene promoters, active regulatory regions observed in blood, blood-cell-lineage-specific enhancer regions, CpGs from Illumina Human Methylation 450 Bead Chips, as well as published autoimmune-related single nucleotide polymorphisms (SNPs) and SNPs in their LD regions (Shao et al., 2019). In summary, such a platform produces DNA methylation levels for comprehensive subsets of informative CpGs, thus capturing epigenomic dysregulation at a much lower cost than WGBS.

Studies using MCC-Seq have yielded encouraging results. For example, with MCC-Seq, Allum et al. (2015) have identified novel methylation variation within enhancers that are strongly associated with plasma triglyceride and HDL-cholesterol. Using phased methylation measurements from both MCC-Seq and WGBS, Cheung et al. (2017) have demonstrated the significant utility of MCC-Seq over WGBS and identified genetically regulated methylation loci that reveal novel epigenetic alterations in the human genome. Shao et al. (2019) have detected rheumatoid arthritis-relevant DNA methylation changes in anti-citrullinated protein antibodies (ACPA)-positive asymptomatic individuals using MCC-Seq.

2.1.2 Possible data errors in bisulfite sequencing

During the bisulfite sequencing experiment, sodium bisulfite treatment of DNA converts unmethylated cytosines (C) to uracils, which are subsequently read as thymines (T) after polymerase chain reaction (PCR)-amplification. In contrast, methylated cytosines are kept unmodified. After proper alignment and data processing (Krueger et al., 2012; Wreczycka et al., 2017), the methylation level at a single cytosine can therefore be inferred by counting the number of C-to-T conversions and the sum of Ts and Cs in the aligned reads. Specifically, the number of C (T) reads is referred to as the (un)methylated count, the total number of Cs and Ts is known as the read depth, and the ratio of methylated count over the read depth measures the methylation proportion at each site.

In fact, the observed counts of methylated and unmethylated reads could be contaminated

by errors arising from various steps of the sequencing processes (Adusumalli et al., 2015; Krueger et al., 2012). For example, the bisulfite conversion may be incomplete, whereby not all unmethylated cytosines are converted to thymines. Misinterpreting the insufficient conversion of unmethylated cytosines as methylated can introduce false-positive methylation calls. In contrast, excessive bisulfite treatment can lead to increased incidences of methylated Cs converting to Ts (Laird, 2010; R. Y.-H. Wang et al., 1980), thus resulting in false-negative methylation calls. One way to measure conversion rates is to add spike-in sequences of DNA known in advance to be methylated or unmethylated.

Moreover, the DNA segments after bisulfite treatment do not precisely match the unmodified reference genome, and the numbers of mismatches depend on the underlying methylation status. Such errors can be minimized by using methylation-aware alignment algorithms; see overviews in Krueger et al. (2012); Wreczycka et al. (2017). Furthermore, high throughput sequencing technologies have a non-negligible error rate in base calls, particularly on the ends of sequencing segments, or in genomic regions such as highly-repetitive sequences (DePristo et al., 2011). These miscalled bases can be erroneously counted as C-T conversions and thus bias the methylation level measurements. Trimming off these low-quality base calls before read alignments can lead to not only reduced methylation call errors but also increased mapping efficiency (Krueger et al., 2012). Nevertheless, despite high-quality data-cleaning protocols, there will inevitably be remaining sequencing errors. Also, it is hard to disentangle the errors arising from the pre-treatment steps and the ones from the sequencing step. Therefore, it is preferable to consider overall error rates when analyzing bisulfite sequencing data, such as analytical models introduced in Cheng & Zhu (2013); Lakhal-Chaieb et al. (2017).

2.1.3 Differentially methylated regions

Once the methylation measurements at each cytosine are available, the general task in genome-wide DNA methylation analysis is to identify differentially methylated cytosines

(DMCs) or differentially methylated regions (DMRs) that are associated with phenotypes or traits. This thesis focuses on examining methylation changes at the regional level rather than at each CpG site. The motivation is multi-fold. First, various studies have shown that methylation levels are strongly correlated across the genome. For instance, Eckhardt et al. (2006) have established a significant spatial correlation of comethylation, described as the percentage of CpGs with similar methylation levels, especially for CpGs located within 1 kilobase pairs (kb). By performing an ultra-deep targeted bisulfite sequencing analysis on different tissues from multiple species (human, mouse and zebrafish), Affinito et al. (2020) have shown that closer CpG sites are more likely to share the same methylation status, independent of tissue types and species. Joint modelling of regional methylation levels allows us to borrow information from this local correlation structure, thus coping naturally with missing values or low counts, of which univariate analyses are incapable. Furthermore, many functionally relevant methylation changes have been found in genomic regions rather than individual CpGs, such as CpG islands (Jaenisch & Bird, 2003), genomic blocks (Hansen et al., 2011) or generic 2kb regions (Lister et al., 2009). These synergistic changes in methylation across a region often convey more substantial regulatory influence (Rackham et al., 2017). In addition, region-based analyses are natural choices for targeted bisulfite sequencing data, the data type that is the focus in this thesis, which are measured in a set of predefined regions. Also, the resulting DMRs can be subsequently explored and annotated easily by examining their overlap with other known genomic features to provide context and perspective of the potential methylation events.

2.2 Generalized additive models

2.2.1 Motivation

Due to the stochastic nature of sequencing and alignment, read depth varies substantially across CpG sites and individual samples, which leads to wide-ranging precision for methylation proportions and many missing values. The spatial correlation of methylation between neighbouring CpG sites suggests that the raw methylation measures, especially the ones with low read-depth, can be improved by smoothing. Taking this into account, many existing methods (Hansen et al., 2012; Hebestreit et al., 2013; Korthauer et al., 2018; Lakhal-Chaieb et al., 2017) use kernel smoothing (i.e. local likelihood estimation) to obtain the smoothed methylation proportions for each sample. They then use these smoothed values to test for differential methylation. These two-step approaches are convenient but may suffer from two major drawbacks. First, the smoothed methylation measures are treated as if they are equally precise, but some might be derived from poor-quality samples (e.g. with few measured CpGs). Second, these methods fail to recognize that the smoothed values are estimated quantities and can lead to underestimating sources of variation.

Generalized additive models (GAM) (Hastie & Tibshirani, 1987; Wood, 2017) are flexible regression tools in which a response variable is related to smooth functions of some predictor variables. GAM provides a unified modelling framework that collapses smoothing and testing steps into a single step and allows to estimate covariate (disease status or other phenotypes of interest) effect from the raw regional methylation measures. This flexible regression approach can be easily applied to non-normal response distributions and simultaneously estimate multiple covariate effects.

In Chapters 3, 4 and 5, I use a special type of GAM, a generalized varying coefficient model, to represent the regional methylation measures derived from bisulfite sequencing, and I have developed various extensions and improvements based upon GAM for better-estimating

regional association patterns. The rest of this section outlines some major rationale about GAM, emphasizing its smoothing techniques, estimation and inference methods.

2.2.2 Model

Throughout this literature review chapter, I will use GAM with binomial outcomes for demonstration, though the outlined methodology can be directly carried over to any exponential families.

Specifically, I consider DNA methylation measures over a targeted genomic region from N independent samples. Let m_i be the number of CpG sites for the i^{th} sample, $i = 1, 2, \dots N$. Write t_{ij} for the genomic position (in base pairs) for the i^{th} sample at the j^{th} CpG site, $j = 1, 2, \dots, m_i$. The set of genomic positions captured in different samples do not have to be identical because each sample has an individual profile of covered CpG sites, due to read depth variability. Methylation levels at a site are quantified by the number of methylated reads and the total number of reads. Define X_{ij} as the total number of reads aligned to CpG j from sample i, of which S_{ij} reads are truly methylated. Furthermore, I assume that I have the information on P covariates for the N samples, denoted as $\boldsymbol{Z_i} = (Z_{i1}, Z_{i2}, \dots Z_{iP})$, for $i = 1, 2, \dots N$.

To model the relationship between S_{ij} and sample-level covariates $\boldsymbol{Z_i}$ across the region, a natural choice is a generalized varying coefficient model

$$\begin{aligned} S_{ij} \mid \boldsymbol{Z_i}, X_{ij} &\sim \text{Binomial}(X_{ij}, \pi_{ij}), \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_0(t_{ij}) + \sum_{p=1}^{P} \beta_p(t_{ij}) Z_{ip}. \end{aligned} \tag{2.1}$$

The benefit of using such a varying coefficient model is extensively discussed in Chapters 3, 4 and 5. Here, I focus on the existing methodologies on the estimation and inference for $\beta_p(t)$s.

2.2.3 Smoothness penalty

Let $f(S_{ij} \mid \boldsymbol{\beta}(t))$ be the probability density function for S_{ij}. Here I simply write $\boldsymbol{\beta}(t)$ for a collection of the functional parameters $\{\beta_p(t)\}_{p=0}^{P}$. Assuming that S_{ij} are independent, given $\boldsymbol{Z}_i$ and t_{ij}, the log-likelihood function for model (2.1) is $\ell(\boldsymbol{\beta}(t)) = \sum_{i,j} \log f(S_{ij} \mid \boldsymbol{\beta}(t))$. Without further constraint on the shapes of $\boldsymbol{\beta}(t)$, the solution $\widehat{\boldsymbol{\beta}}(t) = \text{argmax}_{\boldsymbol{\beta}(t)}\, \ell(\boldsymbol{\beta}(t))$ is not unique; any $\boldsymbol{\beta}(t)$ with the same evaluated values on each t_{ij} would yield equal likelihood.

To avoid this identifiability issue, a constraint with more structure on $\boldsymbol{\beta}(t)$ is needed. One commonly used assumption is that $\boldsymbol{\beta}(t)$ are smooth, which is reasonable considering that the regional methylation proportions often exhibit long-range correlation structure (Affinito et al., 2020; Eckhardt et al., 2006). A mathematical characterization of smoothness is the integrated squared second derivative (Reinsch, 1967). Using such a smoothness penalty, the selection of candidate $\widehat{\boldsymbol{\beta}}(t)$ that are smoothest in between the t_{ij} values, can be achieved by optimizing the penalized log likelihood (Green & Silverman, 1994),

$$\widehat{\boldsymbol{\beta}}(t) = \underset{\boldsymbol{\beta}(t)}{\text{argmax}} \left\{ \ell(\boldsymbol{\beta}(t)) - \frac{1}{2} \sum_{p=0}^{P} \lambda_p \int \left(\beta_p''(t)\right)^2 dt \right\}. \tag{2.2}$$

Here, the weights λ_p, i.e. the smoothing parameters, are positive parameters that establish a tradeoff between the closeness of the curve to the data and the smoothness of the fitted curves. Imposing such a penalty is equivalent to putting upper bounds on the values of $\int \left(\beta_p''(t)\right)^2 dt$, with $\lambda_p/2$ playing the role of Lagrange multipliers.

2.2.4 Basis functions

The optimization problem in (2.2) requires searching an infinite-dimensional function space, which is hardly computable. Basis functions allows the use of finite numbers of parameters to yield estimates and inference for infinite-dimensional function parameters. Specifically,

for a single smooth term $\beta(t)$ (here the subscript p is dropped for notational simplicity)

$$\beta(t) = \sum_{l=1}^{K} \theta_l B_l(t), \tag{2.3}$$

where $\{B_l(\cdot)\}_{l=1}^{K}$ denotes the basis functions, and $\boldsymbol{\theta} = (\theta_1, \dots \theta_K)^T \in \mathbb{R}^K$ are the coefficients to be estimated. Each basis function $B_l(t)$ has known fixed form. It generally defines a linear combination among the locations within the function, and thus establishing a specific framework for borrowing strength (Morris, 2015).

Natural cubic splines

In Chapters 3, 4 and 5, I have used the natural cubic splines as basis functions to expand the functional parameters. A detailed description of natural cubic spline basis functions can be found in Section C.1. Specifically, equations (C.1) and (C.2) define the natural cubic spline basis functions with K knots placed at $t_1 \leq t_2 \leq \dots \leq t_K$.

An appealing property of natural cubic splines is that they are the smoothest interpolators. Consider a simple one-dimensional model for smoothing a continuous response variable $\{y_i\}_{i=1}^{n}$ with respect to a predictor variable $\{t_i\}_{i=1}^{n}$. Of all continuous functions that have absolutely continuous first derivatives, the solution

$$\widehat{f}(t) = \underset{f(t)}{\arg\min} \left\{ \sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \lambda \int f''(t)^2 dt \right\}, \tag{2.4}$$

is a natural cubic spline with knots at distinct observed values of t_i; the proof can be found in Green & Silverman (1994). Note that this result holds when substituting the negative log likelihood for the residual sum of squares in (2.4) (Wood, 2017).

Smoothing splines

Such types of splines, which arise from optimizing objective functions with a smoothness penalty and have dimensions as the number of unique observations of the predictor variable, are also called *smoothing splines* (Parker & Rice, 1985; Wahba, 1980). The property in (2.4) suggests any smooth terms in statistical models can be closely approximated using smoothing splines. However, one drawback of smoothing splines is that the number of free parameters is of the same magnitude as the number of data to be smoothed, which will undoubtedly impose severe computation expense, especially when more than one smooth term is present.

(Reduced rank) penalized regression splines

A compromise between retaining the good properties of smoothing splines and computational efficiency is to use reduced rank penalized regression splines (Parker & Rice, 1985; Wahba, 1980). Let $\mathcal{N}$ be the number of unique t_{ij}. Specifically, penalized regression splines have K knots distributed on the complete set of t_{ij}, where $K < \mathcal{N}$. Studies have shown that the basis dimension K can grow rather slowly with sample size to achieve statistical performance asymptotically indistinguishable from full smoothing splines (Claeskens et al., 2009; Hall & Opsomer, 2005; Kauermann et al., 2009), which demonstrates the capacity of penalized regression splines. In Chapters 3, 4 and 5, I have used the natural cubic penalized regression splines with basis dimension less than the number of unique t_{ij}s.

P-splines

An alternative type of spline basis, which is closely related to penalized regression splines, is P-spline (Eilers & Marx, 1996; Eilers et al., 2015; Ruppert et al., 2003). The idea of P-spline is to use B-spline bases (de Boor, 1978) for representing smooth terms, and then

impose a *difference penalty* on the basis coefficients, such as $\sum_l (\theta_{l+1} - 2\theta_l + \theta_{l-1})^2$. Like the derivative-based smoothness penalty in (2.2), this difference penalty quantifies the roughness of the fitted curve. This difference penalty can be written as a quadratic form with respect to basis coefficients, $\boldsymbol{\alpha}$, and thus has connection with (Gaussian) random effects—a good property possessed by smoothness penalty as well (see details in Section 4.3.1).

Beyond spline-type basis functions

Splines are suited to modelling smooth functions. When the underlying additive terms are irregular functions with spikes or abrupt changes, other types of basis expansions, such as Fourier series or wavelets, can be explored instead. Fourier series can effectively model periodic shapes (Bilodeau, 1992). Wavelets are capable of characterizing spatially heterogeneous data (Donoho & Johnstone, 1995; Morris & Carroll, 2006). For these basis functions, the smoothness penalty in Section 2.2.3 is no longer appropriate, and other types of regularization constraints should be imposed instead (Antoniadis & Fan, 2001).

Kernel smoothers (Hastie & Tibshirani, 1987), which estimate a real-valued function as the local weighted average, can also be thought of as spline-type basis functions (Morris, 2015). For example, Silverman (1984) has demonstrated that for independent data, kernel methods and spline methods are asymptotically equivalent.

2.2.5 Estimation and inference

The original model-fitting method for GAM is the backfitting algorithm (Hastie & Tibshirani, 1987, 1993; Yee & Wild, 1996). This algorithm cycles through the smooth components in the model and estimates each of the components by iteratively smoothing the partial residuals with respect to the covariate(s) involved in the smooth term. The partial residuals for the p^{th} smooth term are obtained by subtracting the current estimates of the 'linear' predictor

without the contribution of the p^{th} term from the (linearized transformed) response variable. The backfitting algorithm has the advantage that it allows the component functions to be estimated using various types of smoothing or regression techniques, such as smoothing splines, kernel smoothing, or boosting (Schmid & Hothorn, 2008; Tutz & Binder, 2006). However, it is challenging to integrate the estimation of smoothness degrees into this approach. In practice, users need to set the values of degrees of freedom for each smooth component in the model or select among a modest set of predefined smoothness degrees.

On the other hand, approaches have been developed for integrating the multiple smoothing parameter estimations with flexible regression model-fitting. The first such developments date back to C. Gu (1992); C. Gu & Wahba (1991), who proposed to optimize the generalized cross-validation (GCV) score for estimating smoothing parameters in GAM (represented with smoothing splines). Subsequently, Wood (2000) proposed to use reduced rank penalized regression splines for expanding the smooth terms and provided a much more efficient method for optimizing the GCV criterion. Later on, Wood (2011) demonstrated the capacity of using restricted maximum likelihood (REML) and marginal likelihood maximization to estimate smoothing parameters. This seminal study provides a fast computation algorithm that yields improvement in numerical robustness over the GCV-based methods. The method of Wood (2011) has been used in Chapters 3 and 4 for estimating smooth covariate effects for the complete data (i.e. the methylated counts are measured without error); a detailed description of this approach is also presented in Sections 3.2.3 and 4.3.2.

Furthermore, there is a natural link between GAM with smoothing or penalized splines and generalized linear mixed model (GLMM). After applying the basis expansion $\beta_p(t) = \sum_{l=1}^{K} \theta_{pl} B_l(t)$, the smoothness penalty becomes

$$\sum_{p=0}^{P} \lambda_p \int \left(\beta_p''(t)\right)^2 dt = \sum_{p=0}^{P} \lambda_p \boldsymbol{\theta_p}^T \boldsymbol{A} \boldsymbol{\theta_p}, \tag{2.5}$$

where $\boldsymbol{A}s$ are $K \times K$ positive semidefinite matrices with the (l, l') element $\boldsymbol{A}(l, l') =$

$\int b_l''(t)b_{l'}''(t)dt$. Imposing such smoothness penalty can be also viewed as assuming Gaussian random effects for basis coefficients $\boldsymbol{\theta}_p$. Therefore, the inference for GAM with smoothing or penalized regression splines can employ the existing methods for fitting GLMM. Specifically, the Laplace approximation (Shun & McCullagh, 1995) for GLMM is used in Chapters 3 and 4; see details in Section A.1.3 and Section 4.3.1. Studies (Handayani et al., 2017; Ju et al., 2020) have shown that the Laplace approximation method shows better properties in terms of convergence rate, bias and coverage, compared to other widely used approximation methods for GLMM, including penalized quasi-likelihood (Breslow & Clayton, 1993) and adaptive Gauss-Hermite quadrature (Rabe-Hesketh et al., 2002).

2.3 Overdispersion

In model (2.1), methylated counts are assumed to follow binomial distributions, dependent on the read-depths. Under this assumption, the variance of methylated counts are entirely determined by its mean,

$$\mathbb{V}ar(S_{ij}) = X_{ij}\pi_{ij}(1-\pi_{ij}).$$

This assumption is fairly strict. However, in practice, the data might exhibit greater variability than assumed by this mean-variance relationship, which is known as *overdispersion*. Similarly, *underdispersion* implies that the empirical variance in the data is less than that predicted by the binomial model. Underdispersion can occur when the binary variates that constitute the methylated counts are negatively correlated. For example, underdispersion can arise when the reads obtained at a CpG site originate from distinct cell types.

One way that under- or overdispersion arises is through a violation of the binomial's independence assumption. Model (2.1) assumes that, given t_{ij} and $\boldsymbol{Z}_i$, the binary random variables $\{S_{ijk}\}_{k=1}^{X_{ij}}$ are mutually independent. When $\mathbb{C}or(S_{ijk}, S_{ijk'}) = \rho > 0$ for any

$k \neq k' \in \{1, 2, \ldots, X_{ij}\}$, the variance of S_{ij} becomes

$$\mathbb{V}ar(S_{ij}) = X_{ij}\pi_{ij}(1-\pi_{ij})\left[1+\rho(X_{ij}-1)\right].$$

In this case, $\rho > 0$ leads to overdispersion relative to a binomial model and $\rho < 0$ leads to underdispersion. One way to account for this scenario is to incorporate a multiplicative scale factor, $\phi > 0$, in the variance of response, i.e. assuming

$$\mathbb{V}ar(S_{ij}) = \phi X_{ij}\pi_{ij}(1-\pi_{ij}).$$

The type of dispersion, which can be characterized by a generalized variance function obtained from multiplying the standard variance function by a free parameter, is referred to as *multiplicative* dispersion in Chapter 4. To adequately address the multiplicative dispersion, the regression model can be fitted using quasi-likelihood approaches, with ϕ estimated by generalizations of moment methods (McCullagh & Nelder, 1989a).

Overdispersion can also occur due to complex correlation structures in hierarchical, clustered, or spatial data (Browne et al., 2005; Grueber et al., 2011). Model (2.1) assumes that given t_{ij} and $\boldsymbol{Z}_i$, the methylated counts S_{ij} are independent across samples and positions. Although the smoothing techniques in GAM can implicitly account for the spatial correlations among neighbouring CpGs, there are additional correlations among methylation measurements on the same subject. Therefore, the binomial variation assumed in model (2.1) can be only a tiny part of the overall data variability. This type of overdispersion is referred to as an *additive* dispersion in Chapter 4. One way to address it is to add a subject-level random effect. Studies have demonstrated the use of random effect models for overdispersed count data (Harrison, 2014; Molenberghs et al., 2012).

Mis-specifying the systematic part of the model can also lead to excess variation around the fitted values, which is termed as *apparent* overdispersion (Hilbe, 2011). Possible sources

of apparent overdispersion include missing important covariates, missing interaction terms, inappropriate functional form of the mean or the presence of outliers (Berk & MacDonald, 2008). Residual diagnostics can be used for identifying the sources, and if possible, the model should be amended accordingly to account for such overdispersion.

2.4 Sparsity penalties

The smoothness penalty $\sum_{p=0}^{P} \lambda_p \boldsymbol{\theta_p}^T \boldsymbol{A} \boldsymbol{\theta_p}$ is a generalized Ridge penalty (Hoerl & Kennard, 1970). It places bounds on the square of the (transformed) basis coefficients $\boldsymbol{\theta}_p$ (ℓ_2 penalty) and leads to penalized estimates of the regression coefficients. However, Ridge penalty cannot shrink the coefficients to exactly 0, and might fail to provide parsimonious models (Friedman et al., 2001). This limitation can be overcomed by using sparse penalty functions; one well-known example is the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). This section will present the penalty functions defined in LASSO and one of its generalizations, named group LASSO.

For the varying coefficient model in (2.1), the regression coefficients to be estimated are a collection of basis coefficients. Let $\boldsymbol{\theta}_p \in \mathbb{R}^K$ be the basis coefficients for $\beta_p(t)$, $p = 0, 1, \ldots P$. The regression coefficient, denoted as $\boldsymbol{\theta}$, is the vectorization of $(P+1) \times K$-dimensional coefficient matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P)^T$ by row, i.e. $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\Theta})$.

2.4.1 LASSO

LASSO places constraint on the sum of absolute values (ℓ_1 norm) of the regression coefficients. For model in (2.1), its LASSO estimator is defined by

$$\widehat{\boldsymbol{\theta}}^{\text{lasso}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, \ell(\boldsymbol{\theta}), \text{ subject to } \sum_{j=K+1}^{(P+1)K} |\theta_j| \leq C, \tag{2.6}$$

where θ_j is the j^{th} element of $\boldsymbol{\theta}$. Here, the coefficients for the intercept $\beta_0(t)$ is left out of the sparsity constraint. One can also write constrained optimization problem in (2.6) in the equivalent Lagrangian form

$$\widehat{\boldsymbol{\theta}}^{\text{lasso}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \left\{ \ell(\boldsymbol{\theta}) - \lambda \sum_{j=K+1}^{(P+1)K} |\theta_j| \right\}. \tag{2.7}$$

A one-to-one correspondence can be established between the parameters λ in (2.7) and C in (2.6).

For a sufficiently large λ (i.e. small C), the LASSO method will yield solution $\widehat{\theta}_j = 0$ for $\forall j \in \{K+1, \ldots (P+1)K\}$, leading to an intercept-only model. In general, larger λ leads to less number of nonzero θ_j.

2.4.2 Group LASSO

With basis expansion, there is a natural grouping structure of the coefficient vector $\boldsymbol{\theta}$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_0^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta_P}^T)^T$. Thus, the selection of important varying coefficients $\beta_p(t)$ amounts to the selection of groups of coefficients in the basis expansions. However, the LASSO estimator ignores this grouping structure and can yield less interpretable solutions in this context.

To take into account the group structure in $\boldsymbol{\theta}$, Yuan & Lin (2006) have proposed the group LASSO estimator, defined by

$$\widehat{\boldsymbol{\theta}}^{\text{groupLasso}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \left\{ \ell(\boldsymbol{\theta}) - \lambda \sum_{p=1}^{P} \sqrt{\boldsymbol{\theta}_p^T \boldsymbol{\theta}_p} \right\}. \tag{2.8}$$

The solution from this group LASSO-penalized likelihood will also display a grouping structure; the elements of $\widehat{\boldsymbol{\theta}}_p$ ($p \geq 1$) will be either all zero or all nonzero. Notably, the LASSO method in (2.7) can be viewed as a special case of group LASSO in (2.8) with each coefficient falling in its own group.

Chapter 3

Manuscript I: A novel statistical method for modelling covariate effects in bisulfite sequencing derived measures of DNA methylation

Preamble to Manuscript I: Bisulfite sequencing allows the generation of high-throughput methylation profiles at the single-base resolution of DNA. However, optimally modelling and analyzing these sparse and discrete sequencing data is challenging due to variable read depth, missing data patterns, data errors, and confounding from cell type mixtures. The spatial correlation of methylation between neighbouring CpG sites can be exploited to improve the raw methylation measures. Thus, many existing methods (Hansen et al., 2012; Hebestreit et al., 2013; Korthauer et al., 2018; Lakhal-Chaieb et al., 2017) use kernel smoothing (i.e. local likelihood estimation) to obtain the smoothed methylation proportions for each sample. Typically, they then use these smoothed values to test for differential methylation. These two-step approaches are convenient but could lead to biased uncertainty estimates.

To overcome the limitations and challenges of existing methods, we proposed a unified analysis framework that collapses smoothing and testing steps into a single step and can achieve accurate statistical uncertainty assessment of differential methylation. In addition, our approach simultaneously addresses the discrete nature of the data, regional testing, estimation of multiple covariate effects, adjustment for read-depth variability and experimental errors.

This manuscript was published in *Biometrics* in May 2020, and the article was accompanied by a freely available R Bioconductor package called `SOMNiBUS` (https://www.bioconductor.org/packages/devel/bioc/html/SOMNiBUS.html).

Note that the supporting material for this chapter can be found in Appendix A.

A novel statistical method for modelling covariate effects in bisulfite sequencing derived measures of DNA methylation

Kaiqiong Zhao[1,2], Karim Oualkacha[3], Lajmi Lakhal-Chaieb[4], Aurélie Labbe[5], Kathleen Klein[2], Antonio Ciampi[1,2], Marie Hudson[2,6], Inés Colmegna[6,7], Tomi Pastinen[8], Tieyuan Zhang[9], Denise Daley[10], Celia M.T. Greenwood[1,2,11]

[1]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada*

[2]*Lady Davis Institute, Jewish General Hospital, Montréal, Québec, Canada*

[3] *Département de Mathématiques, Université du Québec à Montrèal, Montreal, QC, Canada*

[4]*Département de Mathématiques et de Statistique, Université Laval, Quebec City, QC, Canada*

[5]*Département des Sciences de la Décision, HEC Montrèal, Montreal, QC, Canada*

[6]*Department of Medicine, McGill University, Montréal, Québec, Canada*

[7]*The Research Institute of the McGill University Health Centre, Montreal, QC, Canada*

[8]*Center for Pediatric Genomic Medicine, Children's Mercy Kansas City, Kansas City, MO, USA*

[9]*Department of Psychiatry, Douglas Mental Health University Institute, McGill University, Montreal, QC, Canada*

[10]*The Centre for Heart Lung Innovation, and Department of Medicine, University of*

British Columbia, Vancouver, BC, Canada

[11] *Gerald Bronfman Department of Oncology, Department of Human Genetics, McGill University, Montreal, QC, Canada.*

Paper published in *Biometrics* (2020). Online ahead of print DOI: 10.1111/biom.13307.

Abstract

Identifying disease-associated changes in DNA methylation can help us gain a better understanding of disease etiology. Bisulfite sequencing allows the generation of high-throughput methylation profiles at single-base resolution of DNA. However, optimally modelling and analyzing these sparse and discrete sequencing data is still very challenging due to variable read depth, missing data patterns, long-range correlations, data errors, and confounding from cell type mixtures. We propose a regression-based hierarchical model that allows covariate effects to vary smoothly along genomic positions and we have built a specialized EM algorithm, which explicitly allows for experimental errors and cell type mixtures, to make inference about smooth covariate effects in the model. Simulations show that the proposed method provides accurate estimates of covariate effects and captures the major underlying methylation patterns with excellent power. We also apply our method to analyze data from rheumatoid arthritis patients and controls. The method has been implemented in R package `SOMNiBUS`.

3.1 Introduction

Heritability is high for a wide range of human diseases (Maurano et al., 2012), but only a portion of it is attributable to additive genetic variation (Ober & Vercelli, 2011). Maher (2008) suggested that environmental exposures play an important role in explaining the "missing" heritability. Plausibly, such exposures, in interaction with genetic predisposition, may lead to epigenetic modification which alters gene regulation without changing genome sequence (Jaenisch & Bird, 2003). For example, differences in epigenetic profiles may explain how risk factors like age (Horvath, 2013) and smoking (Teschendorff et al., 2015) impact disease susceptibility. Consequently, examining how epigentic profiles contribute to disease development and are influenced by environmental factors, can provide novel insights into disease etiology and possible therapies (Feinberg, 2007).

The most-studied epigenetic mark is DNA methylation which primarily occurs at a cytosine-guanine dinucleotide (i.e. CpG site) (Lister et al., 2009). Localized differential methylation is a characteristic feature of many diseases, such as diabetes (Nilsson et al., 2014), Alzheimer's disease (De Jager et al., 2014) and autoimmune disorders (Liu et al., 2013).

Measuring large-scale DNA methylation at single nucleotide resolution is now possible owing to the development of bisulfite sequencing protocols (Frommer et al., 1992), which can be implemented genome-wide or in a set of targeted regions. Targeted Custom Capture Bisulfite Sequencing (TCCBS) platforms produce DNA methylation levels for comprehensive subsets of informative CpGs. Thus, epigenomic dysregulation can be captured at a much lower cost than whole-genome bisulfite sequencing (WGBS). This approach's capacity to detect novel disease associations has been demonstrated (Allum et al., 2015; Li et al., 2015). In this work, we focus on analysis of predefined regions targeted by TCCBS, with the aim to identify differentially methylated regions (DMRs) that are associated with phenotypes or traits.

Methods for extracting interpretable results from the raw methylation data derived from either WGBS or TCCBS are greatly hindered by the variability in read depths, the many missing values and the possibility of data errors. Specifically, due to the stochastic nature of sequencing and alignment, coverage – the total number of reads spanning a CpG site – varies substantially across sites and individual samples, which leads to wide-ranging precision for methylation proportions, and to many missing values. In fact, estimates of DNA methylation are correlated with read depths (Stephens et al., 2016). Furthermore, the observed counts of methylated and unmethylated reads could be contaminated by errors arising from excessive or insufficient bisulfite treatment, and from misalignment of reads or other aspects of the sequencing processes. Studies show that ignoring these errors could bias inference about the associations of interest (Cheng & Zhu, 2013; Lakhal-Chaieb et al., 2017).

Additionally, due to cell type specific differences in methylation levels, variability in cell type mixture proportions has a strong effect on observed levels of methylation from mixed tissue samples. This mixture, as well as factors known to alter methylation levels, such as age (Horvath, 2013), can confound associations of interest. Hence, it is essential to develop methods to adjust methylation signals for multiple covariates.

Moving in this direction, approaches have been proposed for identifying DMRs from bisulfite sequencing data; see overviews in Shafi et al. (2017) and Yu & Sun (2016a). Typically, to account for spatial correlations of methylation between neighboring CpG sites, strategies include Hidden Markov models (HMM) (Shokoohi et al., 2018; Sun & Yu, 2016; Yu & Sun, 2016b), hierarchical models with autoregressive or random walk correlation structures (Korthauer et al., 2018; Rackham et al., 2017), and kernel-based smoothing methods (Hansen et al., 2012; Hebestreit et al., 2013; Lakhal-Chaieb et al., 2017). However, none of these methods meet all the desirable objectives *simultaneously*: regional testing, estimation of multiple covariate effects, adjustment for read depth variability and experimental errors. For example, several of the current HMM-based (Sun & Yu, 2016; Yu & Sun, 2016b) and hi-

erarchical methods (Rackham et al., 2017) only test for differential methylation between two independent groups of samples and do not allow for the adjustment of multiple covariates. Approaches using a binomial mixed model for DNA methylation analysis (Lea et al., 2015; Weissbrod et al., 2017) allow for multiple covariates and can capture sample correlations, but were only designed for single site analysis. BSmooth (Hansen et al., 2012), a kernel-based method, detects differential methylation after converting the methylated and total counts to proportions. However, this conversion could lead to reduced power since it disregards read depth variability and fails to distinguish between noisy and accurate measurements (Rackham et al., 2017). Moreover, most of the existing methods ignore experimental errors. On the other hand, the only approach accounting for data errors, the Smooth Methylation Status Call (SMSC) (Lakhal-Chaieb et al., 2017), is only developed for data from a single cell type.

More importantly, most of the existing methods are of a two-stage nature (Hansen et al., 2012; Hebestreit et al., 2013; Lakhal-Chaieb et al., 2017). Typically, they first smooth the raw methylation data for each sample separately, and then, in the second stage, they estimate covariate effects by modelling the smoothed methylation data. These per-sample smoothing strategies do not take advantage of information contained across samples and fail to fully exploit the fact that samples with similar covariate profiles (eg. disease status, cell type composition or other phenotypes of interest) can be expected to share similar methylation patterns. In addition, separating smoothing and inference steps results in biased uncertainty estimates. In summary, it would be highly desirable to develop a general framework of analysis, which collapses smoothing and testing steps into a single step, and simultaneously addresses regional testing, estimation of multiple covariate effects, adjustment for read depth variability and experimental errors.

In this paper, we propose such a general framework. Our strategy allows information to be shared not only between nearby CpGs, but also across samples, thus providing greater

sensitivity for capturing patterns common to several samples of similar characteristics (rather than one sample).

Specifically, our approach is built on a hierarchical regression model that describes bisulfite sequencing data. We assume, as in Lakhal-Chaieb et al. (2017) and Cheng & Zhu (2013), that the observed read counts arise from an unobserved latent true methylation state compounded by errors. These true methylation counts are then modeled by a binomial distribution, dependent on read depth. Note that the probability parameter of this binomial distribution depends on the sample-level covariates of interest, such as cell-type mixture proportions and the trait of interest, but also nearby methylation information. To capture realistic methylation patterns across regions, we additionally allow baseline methylation levels, covariate effects and adjustment effects to vary smoothly along genomic positions: this is done by using splines. This amounts to borrowing information from the local correlation structures between methylation levels, and allows us to remedy local information gaps due to missingness. This formulation naturally allows for any number of covariates in the model.

This article is organized as follows. Section 3.2 describes the proposed model along with its estimation and inference procedures. A motivating data example from a study of cases with rheumatoid arthritis and controls is described in Section 3.3. Simulation studies evaluating the performance of our proposed method and comparing our type I errors and power to existing methods are summarized in Section 3.4. The paper concludes with a discussion in Section 3.5.

3.2 Method

3.2.1 Notation and data

We consider DNA methylation measures over a targeted genomic region from N independent samples. Let m_i be the number of CpG sites for the i^{th} sample, $i = 1, 2, \ldots N$. We write t_{ij} for the genomic position (in base pairs) for the i^{th} sample at the j^{th} CpG site, $j = 1, 2, \ldots, m_i$. The set of genomic positions captured in different samples do not have to be identical because each sample has an individual profile of covered CpG sites, due to read depth variability. Methylation levels at a site are quantified by the number of methylated reads and the total number of reads. We define X_{ij} as the total number of reads aligned to CpG j from sample i. The tissue samples sent for bisulfite sequencing experiments from most studies will normally be composed of a mixture of cell types. For example, common cell types are, in blood: granulocytes, T cells, B cells, monocytes, neutrophils, and eosinophils; in adipose tissues: adipocyte, preadipocyte, endothelial and mural cells. Thus, the reads obtained at a CpG site are likely to capture contributions from different cell types; the true underlying methylation statuses are probably different across these X_{ij} reads. We denote the *true* methylation status for the k^{th} read obtained at CpG j of sample i as S_{ijk}, where $k = 1, 2, \ldots X_{ij}$. S_{ijk} is binary and we define $S_{ijk} = 1$ if the corresponding read is methylated and $S_{ijk} = 0$ otherwise. In the presence of experimental errors in sequencing or preprocessing, the *observed* methylation status, written as Y_{ijk}, can be distinct from the true underlying information S_{ijk}. We denote $Y_{ijk} = 1$ if the corresponding read is observed as methylated and $Y_{ijk} = 0$ otherwise. We additionally denote the *true* and *observed methylated counts* at CpG j for sample i with $S_{ij} = \sum_{k=1}^{X_{ij}} S_{ijk}$ and $Y_{ij} = \sum_{k=1}^{X_{ij}} Y_{ijk}$, respectively. Furthermore, we assume that we have the information on P covariates for the N samples, denoted as $\boldsymbol{Z_i} = (Z_{1i}, Z_{2i}, \ldots Z_{Pi})$, for $i = 1, 2, \ldots N$.

3.2.2 Model

We built here on concepts introduced in Cheng & Zhu (2013) and Lakhal-Chaieb et al. (2017) to account for experimental errors. We assume that, depending on the true underlying methylation status S_{ijk}, the observed status Y_{ijk} is a Bernoulli variable with parameters p_0 or p_1, i.e.

$$\begin{aligned} p_0 &= \mathbb{P}(Y_{ijk} = 1 \mid S_{ijk} = 0), \\ p_1 &= \mathbb{P}(Y_{ijk} = 1 \mid S_{ijk} = 1). \end{aligned} \tag{3.1}$$

Here, these two parameters capture errors; p_0 is the rate of false methylation calls, and $1 - p_1$ is the rate of false non-methylation calls. These rates are assumed to be constant across all reads and positions. The error parameters p_0 and p_1 can be estimated by looking at raw sequencing data at CpG sites known in advance to be methylated or unmethylated (Wreczycka et al., 2017). We assume hereafter that p_0 and p_1 are known. Implications of such an assumption is discussed later in the Supporting Information Section A.2.2.

We then assume the true methylated counts S_{ij} follows a binomial distribution with a methylation proportion parameter π_{ij} that depends on the sample-level covariates $\boldsymbol{Z_i}$, and on nearby methylation patterns. Specifically,

$$\begin{aligned} S_{ij} \mid \boldsymbol{Z_i}, X_{ij} &\sim \text{Binomial}(X_{ij}, \pi_{ij}), \\ g(\pi_{ij}) &= \beta_0(t_{ij}) + \sum_{p=1}^{P} \beta_p(t_{ij}) Z_{pi}, \end{aligned} \tag{3.2}$$

where $g(\cdot)$ is a logit link function and $\beta_0(t_{ij})$ and $\{\beta_p(t_{ij})\}_{p=1}^{P}$ are functional parameters for the intercept and covariate effects. This amounts to assuming smoothly varying methylation levels and covariate effects on methylation levels across our targeted small genomic regions. In practice, to estimate Model (3.2), the functions $\beta_p(t_{ij})$ can be represented by the

coefficients of a chosen spline bases of rank L_p,

$$\beta_p(t_{ij}) = \sum_{l=1}^{L_p} \alpha_{pl} B_l^{(p)}(t_{ij}), \text{ for } p = 0, 1, \dots P,$$

where $\left\{B_l^{(p)}(\cdot)\right\}_{l=1}^{L_p}$ denotes the spline basis, and $\boldsymbol{\alpha_p} = (\alpha_{p1}, \dots \alpha_{pL_p})^T \in \mathcal{R}^{L_p}$ are the coefficients to be estimated. In this way, model (3.2) becomes a generalized linear model (GLM), $g(\boldsymbol{\pi}) = \mathbb{X}\boldsymbol{\alpha}$, where $\boldsymbol{\pi} = (\pi_{11}, \dots \pi_{1m_1}, \pi_{21}, \dots \pi_{2m_2}, \dots \pi_{Nm_N})^T \in [0,1]^M$ with $M = \sum_{i=1}^N m_i$, $\boldsymbol{\alpha} \in \mathcal{R}^K$ with $K = \sum_{p=0}^P L_p$, and $\mathbb{X}$ is the spanned design matrix of dimension $M \times K$, stacked with elements $B_l^{(p)}(t_{ij}) \times Z_{pi}$; for detailed forms see Supporting Information Appendix A.1.1.

To avoid over-fitting, we penalize departure from smoothness, using penalized regression splines (Parker & Rice, 1985; Wahba, 1980). Specifically, we use a comparatively large number of knots (equivalent to large L_p) and a penalization, quantified by the integrated squared curvature of the splines, is added as an extra term in the log-likelihood function (loss function),

$$\mathcal{L}^{\text{Penalization}} = \sum_{p=0}^{P} \lambda_p \int \left(\beta_p''(t)\right)^2 dt = \sum_{p=0}^{P} \lambda_p \boldsymbol{\alpha_p}^T \boldsymbol{A_p} \boldsymbol{\alpha_p}. \tag{3.3}$$

In equation (3.3), $\boldsymbol{A_p}'s$ are $L_p \times L_p$ positive semidefinite matrices with the (l, l') element $\boldsymbol{A_p}(l, l') = \int {B^{(p)}}_l''(t) {B^{(p)}}_{l'}''(t) dt$; these are fixed quantities given the specified set of basis functions. The weights λ_p, i.e. the smoothing parameters, are positive parameters which establish a tradeoff between the closeness of the curve to the data and the smoothness of the fitted curves. Note that there is one smoothing parameter per covariate in our model. The smoothing process across targeted regions is accomplished by adding the penalization terms in equation (3.3) to the model in equation (3.2).

3.2.3 Estimation

Penalized complete likelihood

If the true methylated counts S_{ij} were available, model (3.2) with penalization (3.3) would be estimated by maximizing the penalized log-likelihood,

$$l^{\text{complete}}(\boldsymbol{S};\boldsymbol{\alpha},\boldsymbol{\lambda}) = l(\boldsymbol{S};\boldsymbol{\alpha}) - \frac{1}{2}\sum_{p=0}^{P}\lambda_p \boldsymbol{\alpha_p^T A_p \alpha_p} = l(\boldsymbol{S;\alpha}) - \frac{1}{2}\boldsymbol{\alpha^T A_\lambda \alpha},$$

where $l(\boldsymbol{S;\alpha}) = \sum_{i=1}^{N}\sum_{j=1}^{m_i}\{S_{ij}\log(\pi_{ij}) + (X_{ij}-S_{ij})\log(1-\pi_{ij})\}$, and $\boldsymbol{A_\lambda}$ is a $K\times K$ positive semidefinite block diagonal matrix of the form $\boldsymbol{A_\lambda} = \text{Diag}\{\lambda_0\boldsymbol{A_0},\lambda_1\boldsymbol{A_1},\ldots,\lambda_P\boldsymbol{A_P}\}$. This is also the complete-data log-likelihood of the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{S}$, i.e. $\log(f(S)) + \log(f(Y \mid S))$; indeed, $f(Y \mid S)$ only depends on the known error rates p_0 and p_1, and bears no information on the parameters of interest.

Smoothed E-M algorithm

In practice, the true methylation data, S_{ij}, are unknown and one only observes Y_{ij}, which is a mixture of binomial counts arising from both the truly methylated and truly unmethylated reads. The EM algorithm (Dempster et al., 1977) allows us to estimate model (3.2) based on the observed data Y_{ij}, by repeatedly replacing a trial estimate $(\boldsymbol{\alpha^\star},\boldsymbol{\lambda^\star})$ by a new $(\boldsymbol{\alpha},\boldsymbol{\lambda})$, which is a maximum of the function

$$Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha^\star}) = \mathbb{E}\left\{l^{\text{complete}}(\boldsymbol{S};\boldsymbol{\alpha},\boldsymbol{\lambda}) \mid \boldsymbol{Y},\boldsymbol{\alpha^\star}\right\} = l(\boldsymbol{\eta^\star;\alpha}) - \frac{1}{2}\boldsymbol{\alpha^T A_\lambda \alpha}. \tag{3.4}$$

E step In equation (3.4) $\boldsymbol{\eta^\star} = (\eta^\star_{11},\ldots,\eta^\star_{1m_1},\eta^\star_{21},\ldots,\eta_{2m_2^\star},\ldots,\eta^\star_{Nm_N})^T \in \mathcal{R}^M$ are conditional expectations of S_{ij} given Y_{ij} evaluated at the trial estimates $(\boldsymbol{\alpha^\star},\boldsymbol{\lambda^\star})$; in our case these

take the form

$$\eta_{ij}^{\star} = \mathbb{E}\left(S_{ij} \mid Y_{ij}; \boldsymbol{\alpha}^{\star}, \boldsymbol{\lambda}^{\star}\right) = \frac{Y_{ij} p_1 \pi_{ij}^{\star}}{p_1 \pi_{ij}^{\star} + p_0(1-\pi_{ij}^{\star})} + \frac{(X_{ij} - Y_{ij})\,(1-p_1)\pi_{ij}^{\star}}{(1-p_1)\pi_{ij}^{\star} + (1-p_0)(1-\pi_{ij}^{\star})}, \tag{3.5}$$

with $\pi_{ij}^{\star} = g^{-1}(\mathbb{X}\boldsymbol{\alpha}^{\star})$, which depends on $\boldsymbol{\lambda}^{\star}$ via the dependence of $\boldsymbol{\alpha}^{\star}$ on $\boldsymbol{\lambda}^{\star}$. Calculating these conditional expectations $\eta_{ij}^{\star}$ from (3.5) constitutes the E step of our algorithm.

M step Each M step involves maximizing the Q function in (3.4) to update $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$. This is a penalized (GLM) likelihood maximization problem with multiple quadratic penalties, previously studied in Wood (2011); Wood & Fasiolo (2017); Wood et al. (2016). Our computational strategy for estimating smoothing parameters $\boldsymbol{\lambda}$ is a nested optimization procedure (Wood, 2011), with an outer iteration for optimizing $\boldsymbol{\lambda}$ and an inner P-IRLS iteration to estimate $\boldsymbol{\alpha}$ given the trial value of $\boldsymbol{\lambda}$ from the outer iteration.

For given values of smoothing parameters $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots \lambda_P)$, a unique maximizer of expression (3.4) is readily computed by penalized iteratively reweighted least squares (P-IRLS); see more details in the Supporting Information Appendix A.1.2. Specifically, the outer iteration involves maximizing a restricted likelihood for smoothing parameters $\boldsymbol{\lambda}$, which is obtained by integrating $\boldsymbol{\alpha}$ out of the joint likelihood for $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$. We rely on the work done by Wood (2011) and use a Laplace approximated restricted likelihood; see more details in the Supporting Information Appendix A.1.3. Since the analytical forms for derivatives and Hessians of this restricted likelihood are also available, the optimization for $\boldsymbol{\lambda}$ in the outer iteration can be readily achieved via Newton's method.

Although the combination is undoubtedly computationally complex, for models with properly defined likelihoods, the nested iterations will guarantee convergence provided that the number of smooth terms increasing at no higher rate than $N^{1/3}$ (Shun & McCullagh, 1995; Wood, 2011; Wood et al., 2016).

E-M iteration We iterate between the E and M steps until convergence to obtain $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\lambda}}$. Given the estimates of basis coefficients $\boldsymbol{\alpha}_p$, for $p = 0, 1, \ldots P$, the functional parameters $\beta_p(t)$ can be thus estimated by $\widehat{\beta_p(t)} = \left\{\boldsymbol{B}^{(p)}(t)\right\}^T \{\widehat{\boldsymbol{\alpha}_p}\}$, where t is a genomic position lying within the range of the input positions $\{t_{ij}\}$, and $\boldsymbol{B}^{(p)}(t) = \left(B_1^{(p)}(t), B_2^{(p)}(t), \ldots B_{L_p}^{(p)}(t)\right)^T \in \mathcal{R}^{L_p}$ is a column vector with nonrandom quantities obtained from evaluating the set of basis functions $\left\{B_l^{(p)}(\cdot)\right\}_l$ at position t.

3.2.4 Inference for smooth covariate effects

To obtain a quantification of the uncertainty accompanying the smoothed EM estimates for the covariate effects $\{\beta_1(t), \beta_2(t), \ldots \beta_P(t)\}$, we additionally estimate their pointwise confidence intervals (CI) in Section 4.3.4, and obtain tests of hypotheses for these effects in Section 4.3.4. This inference is carried out conditional on the values of smoothing parameter $\boldsymbol{\lambda}$; i.e. the uncertainty in estimating λ is not accounted for. The potential bias associated with this assumption is shown to be small; see the pointwise confidence interval coverage in Figure 3.4 and the distribution of region-based p-values under the null in Figure 3.5.

Confidence interval estimation

Analytical derivation for standard errors usually involves calculating the observed Fisher information for parameters $\boldsymbol{\alpha}$ from the marginal log-likelihood for $\boldsymbol{Y}$. However, in this case, a direct calculation of the observed Fisher information is analytically intractable because the observed $\boldsymbol{Y}$ follows a mixture of two binomial distributions. To circumvent this problem, we rely on the work of Louis (1982) and Oakes (1999), which showed that this Fisher information can be calculated solely from the Q function (3.4), without referring to the marginal distribution of $\boldsymbol{Y}$.

Theorem 1. *Under the usual regularity conditions for maximum likelihood, we have the*

following asymptotic results for the estimators $\widehat{\boldsymbol{\alpha}}$ *obtained from the smoothed-EM algorithm,*

$$\sqrt{M}\left(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\right) \xrightarrow{\mathcal{L}} \boldsymbol{MVN}_K(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1}), \quad as\ M \to \infty.$$

Here, K *is the dimension of the spline coefficients* $\boldsymbol{\alpha}$*, and* $\boldsymbol{\mathcal{I}} = \mathbb{E}\left[-\mathcal{H}_{ij}(\boldsymbol{\alpha})\right]$*. Specifically* $\mathcal{H}_{ij}(\boldsymbol{\alpha})$ *has the form*

$$\mathcal{H}_{ij}(\boldsymbol{\alpha}) = \mathbb{X}_{(l,)}^T\left(-X_{ij}w_{ij} + \delta_{ij}w_{ij}\right)\mathbb{X}_{(l,)} - \boldsymbol{A}_{\boldsymbol{\lambda}},$$

where $\mathbb{X}_{(l,)}$ *is the* l^{th} *row of the design matrix* $\mathbb{X}$*, which corresponds to the CpG* j *of sample* i*,* $w_{ij} = \pi_{ij}(1-\pi_{ij})$ *is the element of the weight matrix, and*

$$\delta_{ij} = \frac{Y_{ij}p_1p_0}{\left[p_1\pi_{ij} + p_0(1-\pi_{ij})\right]^2} + \frac{\left(X_{ij}-Y_{ij}\right)(1-p_1)(1-p_0)}{\left[(1-p_1)\pi_{ij} + (1-p_0)(1-\pi_{ij})\right]^2}.$$

The proof of Theorem 1 is given in the Supporting Information Appendix A.1.4. Theorem 1 provides the desired variance-covariance matrix of the EM estimators $\widehat{\boldsymbol{\alpha}}$, which can be estimated using the observed Fisher information

$$\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\boldsymbol{\alpha}}) = \left\{-\mathcal{H}(\widehat{\boldsymbol{\alpha}})\right\}^{-1},$$

where $\mathcal{H}(\widehat{\boldsymbol{\alpha}}) = \sum_{i,j}\mathcal{H}_{ij}(\widehat{\boldsymbol{\alpha}})$. Let $\boldsymbol{V}$ denote this variance estimator and $\boldsymbol{V_p}$ be the diagonal blocks of $\boldsymbol{V}$ corresponding to $\boldsymbol{\alpha}_p$, with dimensions $L_p \times L_p$. Since $\beta(t)$ is a linear combination of coefficients $\boldsymbol{\alpha_p}$, the estimated variance of $\widehat{\beta_p(t)}$ takes the form $\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta_p(t)}) = \left\{\boldsymbol{B}^{(p)}(t)\right\}^T \boldsymbol{V_p}\left\{\boldsymbol{B}^{(p)}(t)\right\}$. Therefore, the confidence interval for $\beta_p(t)$ at significance level ν can be estimated by $\widehat{\beta_p(t)} \pm \mathbb{Z}_{\nu/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta_p(t)})}$, for any t in the range of interest, where $\mathbb{Z}_{\nu/2}$ is $\nu/2$ (upper-tail) quantile of a standard normal distribution.

Hypothesis testing for a regional zero effect

We can also construct a region-wide test of the null hypothesis

$$H_0 : \beta_p(t) = 0, \text{for any } t \text{ in the genomic interval.}$$

This test depends on the association between covariate Z_p and methylation levels across the region, after adjustment for all the other covariates, and the null hypothesis is equivalent to $H_0 : \boldsymbol{\alpha}_p = \mathbf{0}$. A Wald-type statistic can be naturally proposed as

$$T_p = \widehat{\boldsymbol{\alpha}_{\boldsymbol{p}}}^T \left\{\boldsymbol{V}_{\boldsymbol{p}}\right\}^{-1} \widehat{\boldsymbol{\alpha}_{\boldsymbol{p}}},$$

where $\left\{\boldsymbol{V}_{\boldsymbol{p}}\right\}^{-1}$ denotes inverse if $\boldsymbol{V}_{\boldsymbol{p}}$ is nonsigular; for singular $\boldsymbol{V}_p$, the inverse is replaced by the Moore-Penrose inverse $\left\{\boldsymbol{V}_{\boldsymbol{p}}\right\}^{-}$. If $\boldsymbol{\alpha}_{\boldsymbol{p}}$ is a vector of unpenalized coefficients, under the null hypothesis, T_p asymptotically follows a Chi-square distribution with degrees of freedom L_p. In the presence of smoothness penalization, L_p should be replaced by the effective degrees of freedom (EDF), τ_p, which depends on the magnitude of smoothing parameter λ_p and is smaller than L_p. Motivated by the work of Wood (2013a), we define the EDF τ_p as

$$\tau_p = \sum_{l=a_p}^{b_p} \left(2\boldsymbol{F} - \boldsymbol{F}\boldsymbol{F}\right)_{(l,l)}, \text{ for } p = 0, 1, \ldots P, \tag{3.6}$$

where $a_p = \sum_{m=0}^{p-1} L_m + 1$ if $p > 0$ and $a_p = 1$ if $p = 0$, $b_p = \sum_{m=0}^{p} L_m$ for any p, and $(\bullet)_{(l,l)}$ stands for the l^{th} leading diagonal element of a matrix. In (3.6), $\boldsymbol{F}$ is the smoothing matrix of our model, which has the form $\boldsymbol{F} = (\mathbb{X}^T\widehat{\boldsymbol{W}}\mathbb{X} + \boldsymbol{A}_{\widehat{\boldsymbol{\lambda}}})^{-1}\mathbb{X}^T\widehat{\boldsymbol{W}}\mathbb{X}$, where $\widehat{\boldsymbol{W}}$ is the weight matrix whose diagonal is $X_{ij}\widehat{\pi}_{ij}(1-\widehat{\pi}_{ij})$. The definition for EDF in (3.6) stems from the smoothing-bias-corrected estimate for $\boldsymbol{\theta} = \log\left(\boldsymbol{\pi}/(1-\boldsymbol{\pi})\right)$. It takes the form $\widehat{\boldsymbol{\theta}} + (\widehat{\boldsymbol{\theta}} - \boldsymbol{F}\widehat{\boldsymbol{\theta}}) = (2\boldsymbol{F} - \boldsymbol{F}\boldsymbol{F})\widetilde{\boldsymbol{\eta}}$, where $\widetilde{\boldsymbol{\eta}} = \log\left(\widehat{\boldsymbol{\eta}}/(1-\widehat{\boldsymbol{\eta}})\right)$ is the adjusted/pseudo outcome. The detailed derivation for (3.6) can be also found in Wood (2013a). A joint null hypothesis

that evaluates the effects of multiple covariates can be defined in a similar way.

Hereafter we refer the proposed novel method including the region-wide test and the smooth covariate estimation as SOMNiBUS (SmOoth ModeliNg of BisUlfite Sequencing).

3.3 DNA methylation data from a rheumatoid arthritis study

To illustrate our method, we report our analysis on data from a rheumatoid arthritis study (Hudson et al., 2017). DNA methylation profiles of cell-separated blood samples of 22 rheumatoid arthritis (RA) patients and 21 healthy individuals were measured with custom captured targeted bisulfite sequencing. We focus on one targeted region on chromosome 4 near gene *BANK1*, which is known to show cell-type-specific DNA methylation levels (Hillier et al., 2005). Three additional targeted regions from the same dataset are also analyzed in the Supporting Information Section A.3. In this *BANK1* region, DNA methylation levels are available at 123 CpG sites. There are 25 samples from circulating T cells and 18 samples from monocytes. We consider two binary covariates—RA status and cell type—and study their impact on DNA methylation pattern in this region.

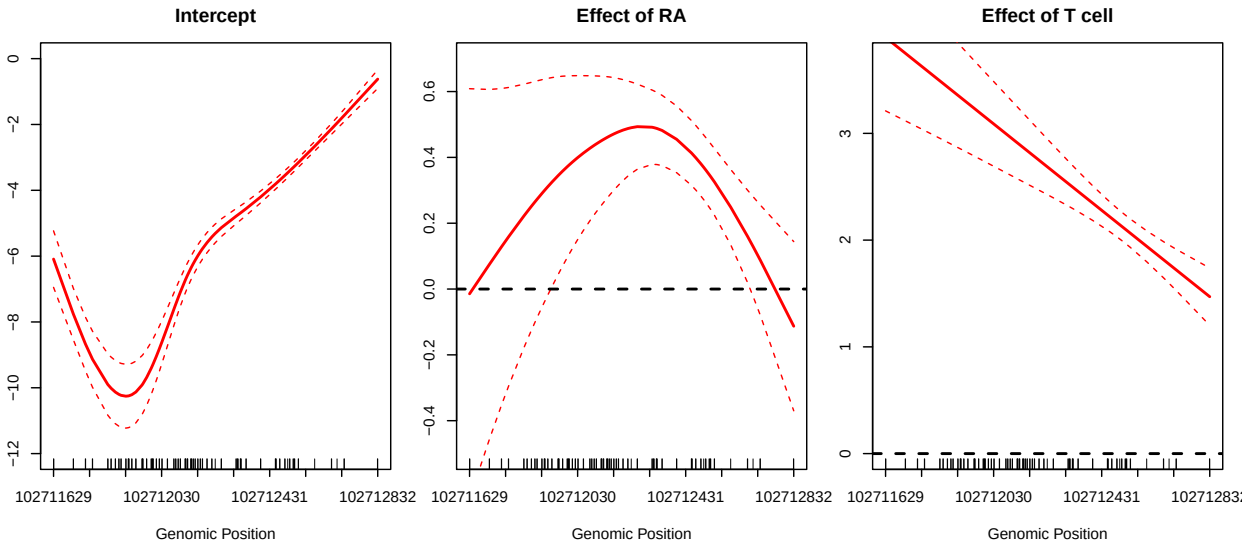
To fit SOMNiBUS, we specified error parameters $p_0 = 0.003$ and $1-p_1 = 0.1$; the value 0.003 was reported by Prochenka et al. (2015) as insufficient conversion rate and 0.1 was estimated as the average excessive conversion rate in our data using the method proposed by Lakhal-Chaieb et al. (2017). We used cubic splines of rank $L_p = 5$ to expand the smooth terms in the model. Figure 3.1 (A) shows the estimated smooth covariate effects on DNA methylation levels in the targeted *BANK1* region. The panel "Intercept" displays the DNA methylation pattern (on the logit scale) for control samples with the monocyte cell type. The panel "Effect of RA" displays the pattern of DNA methylation difference (on the logit scale) between RA samples and control samples with the same cell type. This figure suggests that RA patients

show slightly higher DNA methylation levels in the middle part of the region, compared to controls. The panel “Effect of Tcell" represents the difference of DNA methylation levels (on the logit scale) between T cell samples and monocyte samples with the same disease status. This effect curve, along with the confidence interval bands, clearly shows a highly significant increase of DNA methylation in T cells relative to monocytes. Figure 3.1 (B) displays the predicted DNA methylation proportions in the 4 groups of samples, defined by cell type and RA status. Overall, Figure 3.1 demonstrates the smoothness of the fits, the ability to use multiple covariates simultaneously, and the ease of interpretation of results across the region. Region-wide tests of significance for the 2 covariates are highly significant (Figure 3.1). We also applied five alternative methods, described in Section 4; see Table A.3 in the Supporting Information.

3.4 Simulation study

We conducted simulation to i) demonstrate that the proposed inference of smooth covariate effects is valid, and to ii) compare the performance of our method with five existing methods: `BiSeq` (Hebestreit et al., 2013), `BSmooth` (Hansen et al., 2012), `SMSC` (Lakhal-Chaieb et al., 2017), `dmrseq` (Korthauer et al., 2018) and `GlobalTest` (Goeman et al., 2006), in terms of type I error and power. The first three methods are typical examples of two-stage analytic approaches. In the first stage, kernel smoothing (local likelihood estimation) is applied to the DNA methylation data of each sample separately. In the second stage, the smoothed methylation data are further analyzed. Specifically, `BiSeq` calculates the average of Wald statistics from single-site beta regression models, while `BSmooth` and `SMSC` calculate the sum of t-statistics across loci; these statistics are used to test for differential methylation of a region. In contrast, `dmrseq` and `GlobalTest` are one-stage approaches which fit their models directly to the raw methylation proportions in a region. Specifically, `dmrseq` assesses the strength of the covariate effect using a Wald test statistic within a generalized least

(A)



(B)

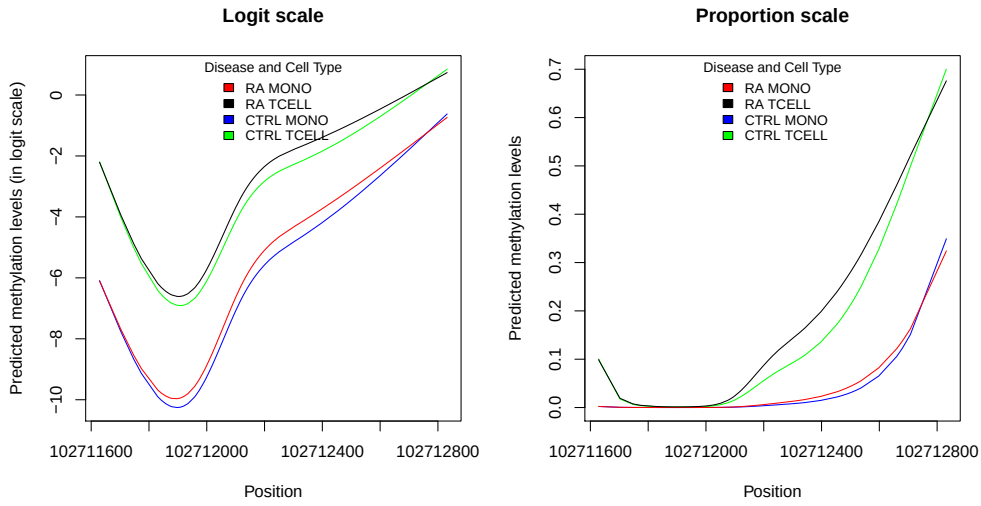


Figure 3.1: (A) The estimates (solid red lines) and 95% pointwise confidence intervals (dashed red lines) of the intercept, the smooth effect of rheumatoid arthritis (RA) and cell type (T cells versus monocytes) on DNA methylation levels. (B) The predicted DNA methylation levels in the logit scale (left) and proportion scale (right) for the 4 groups of samples with different disease and cell type status. The region-based p-values for the effect of RA status and cell type are calculated as $1.11E-16$ and $6.37E-218$, respectively.

square regression model, while `GlobalTest` uses an improved score test in a linear regression model.

Notably, like `SOMNiBUS`, both `GlobalTest` and `BiSeq` are primarily tailored to targeted bisulfite sequencing data with previously identified regions, whereas `BSmooth`, `SMSC` and `dmrseq` are designed for WGBS data. Specifically, `BSmooth` and `SMSC` define DMRs at adjacent CpG sites with absolute t-statistics above a defined threshold. The final product from the original software of `BSmooth` is a list of DMRs that are ranked by the sum of t-statistics; however, `BSmooth` does not provide region-based p-values. To allow comparisons with `SOMNiBUS`, we estimated the empirical regional p-values for `BSmooth` by permuting the values of the covariate of interest 1000 times. When analyzing WGBS data, `dmrseq` first constructs candidate regions based on a user-defined cutoff of the smoothed methylation proportion differences, and then fits a generalized least squares regression model with autoregressive error structure to the transformed methylation proportions. Furthermore, the inference inside `dmrseq` is drawn from permutations – its approximate null distribution is generated by pooling a set of region-level statistics of many candidate regions from all permutations. To better adapt `dmrseq` to a single targeted region: i) we used a small cutoff of methylation differences ($1E-5$) for detecting candidate (sub)regions, which ensures fewer CpG sites to be filtered out; ii) we applied a relatively large number of permutations ($B = 500$) to generate a null distribution of test statistics; iii) we reported the raw p-values without the multiplicity corrections. Note that in some simulations, `dmrseq` reported more than one DMR in the region. Therefore, for a fairer comparison, we calculated the `dmrseq`'s p-value as the minimum over the reported chunks' p-values.

Among the five competitive methods, `dmrseq`, `GlobalTest` and `BiSeq` allow adjustment for multiple covariates. `SMSC` is the only approach accounting for experimental errors; however, it is conceptually restricted to data from a single cell type.

3.4.1 Simulation design

Our simulation design is inspired by the data example described in Section 3.3. Methylation regions of the same size and with the same CpG distribution as the *BANK1* region were simulated under various settings. We first generated the read depth X_{ij} by resampling the read depth values of all positions and samples in the BANK1 data, with replacement. To specify covariates Z_p and their effect curves $\beta_p(t)$, we then considered the following two scenarios.

Scenario 1 – Multiple covariates In this case, $P = 3$ binary covariates Z_1, Z_2, and Z_3 were generated independently for each sample. Z_1 and Z_2 were simulated from Bernoulli distributions with proportions 0.51 and 0.58, which were the proportions of RA and T cell samples in the RA dataset. The functional parameters for intercept and covariate effects, $\beta_0(t)$, $\beta_1(t)$ and $\beta_2(t)$, were specified to have the same shapes as seen in the *BANK1* region (Figure 3.1 (A)). Covariate Z_3 was generated from a Bernoulli distribution with proportion parameter 0.5 and had zero effect on methylation, i.e. $\beta_3(t) = 0$, for all t in the region. The inference results for the effect of the null covariate, Z_3, provide information on type I error.

Scenario 2 – Single covariate We also considered the case of a single binary covariate ($P = 1$), generated from Bernoulli (0.5), with a variety of regional effect curves. The forms of the functional parameters $\beta_0(t)$ and $\beta_1(t)$ were specified to yield methylation proportion parameters $\pi_0(t)$ and $\pi_1(t)$ as depicted in Figure 3.2, where $\pi_0(t)$ and $\pi_1(t)$ denote the methylation parameters for samples with $Z = 0$ and $Z = 1$ at position t. As shown in Figure 3.2, these 14 settings of $\pi_0(t)$ correspond to varying levels of closeness between methylation patterns from the two groups.The corresponding values of $\beta_0(t)$ and $\beta_1(t)$ under these 14 settings are shown in the Supporting Information Figure A.1. We defined the maximum deviation (MD) as the maximum difference between $\pi_1(t)$ and $\pi_0(t)$, for t in the section

indicated by the dashed lines in Figure 3.2, where the curves of π_1 and π_0 mainly differ. Simulation scenario 2 is aimed at investigating the power for detecting DMRs at varying levels of maximum derivations.

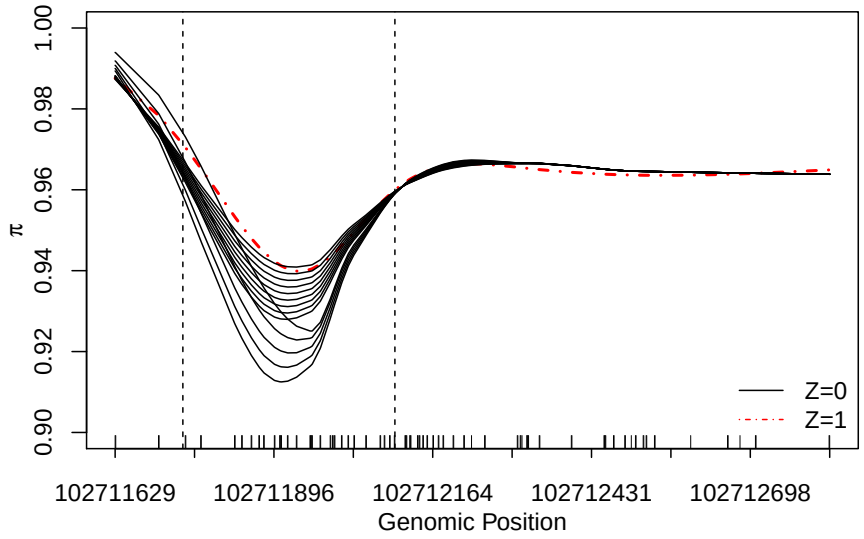


Figure 3.2: The 14 simulation settings of methylation parameters $\pi(t)$ in Scenario 2. Methylation parameters for samples with $Z = 1$ (red dotted-dashed curve) are fixed across settings, whereas the methylation parameters for samples from group $Z = 0$ (black solid lines) vary across simulations corresponding to different degrees of closeness between methylation patterns in the two groups.

Given the values of $\{Z_1, \ldots Z_P\}$ and $\{\beta_p(t), p = 0, 1, \ldots P\}$ under each setting, the true methylation counts S_{ij} were simulated from the model specified in (3.2). We then generated the observed methylated counts Y_{ij} according to equation (3.1), which implies

$$Y_{ij} \mid S_{ij} \sim \text{Binomial}(S_{ij}, p_1) + \text{Binomial}(X_{ij} - S_{ij}, p_0).$$

We considered two settings for error parameters p_0 and p_1: (1) $p_0 = 0.003$ and $1 - p_1 = 0.1$, and (2) $p_0 = 1 - p_1 = 0$.

Under each scenario and setting, we generated data sets with sample sizes $N = 40, 100, 150$ and 400, each 1000 times. We then applied `SOMNiBUS` along with methods `BiSeq`, `dmrseq`,

BSmooth, SMSC and GlobalTest to the simulated data sets. Unless otherwise stated, default settings were used for the five alternative methods. For our approach SOMNiBUS, we used cubic splines with dimension $L_p = 5$ to parameterize the smooth terms of interest. We also assumed that the correct values of error parameters p_0 and p_1 were known, although we conducted sensitivity analyses to this assumption (see Discussion and Supporting Information Section A.2.2). All simulation parameters are summarized in the Supporting Information Table A.1.

3.4.2 Simulation results

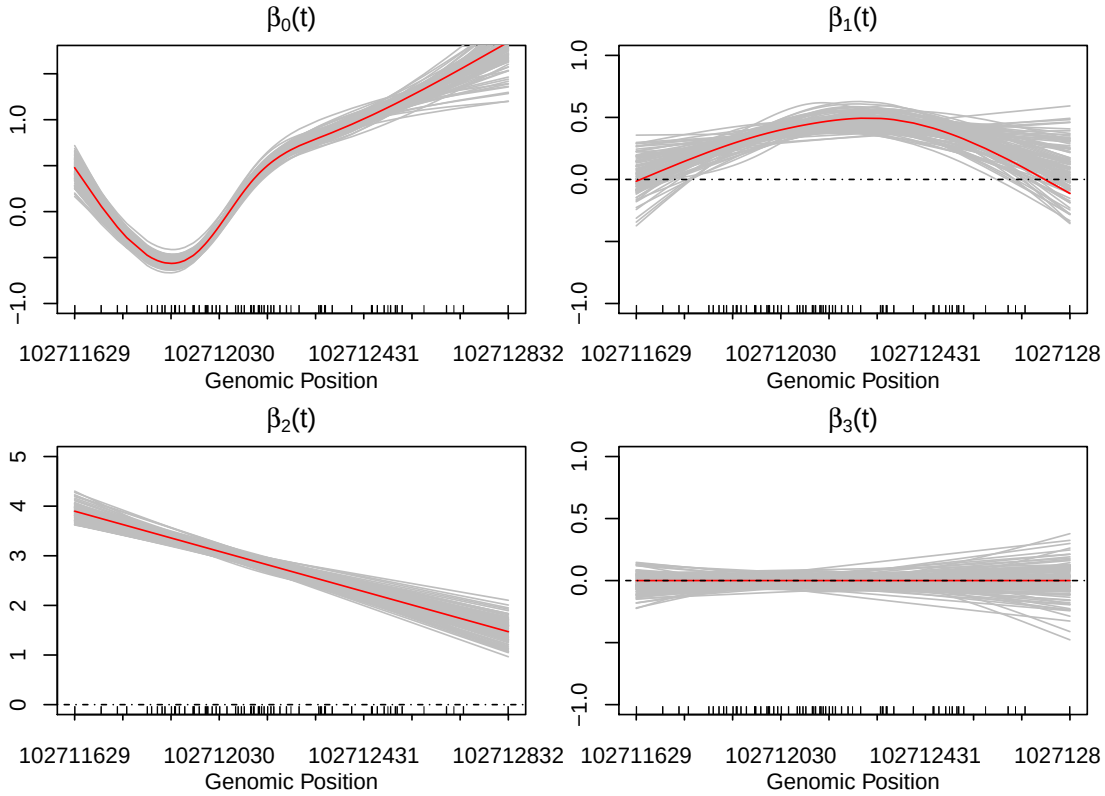


Figure 3.3: Estimates of smooth covariate effects (gray) over the 100 simulations in Scenario 1, using SOMNiBUS. The red curves are the true functional parameters used to generate the data. Data with sample size $N = 40$ were generated with error.

Figure 3.3 presents the estimates of the functional parameters $\beta_0(t), \beta_1(t), \beta_2(t)$ and $\beta_3(t)$ over 100 simulations, obtained from SOMNiBUS; here, data were generated under Scenario 1,

with sample size $N = 40$ and error parameters $p_0 = 0.003$ and $1 - p_1 = 0.1$. It demonstrates that the proposed method provides unbiased curve estimates for all the four functional parameters in the model, and it can correctly capture both linear and nonlinear smooth covariate effects.

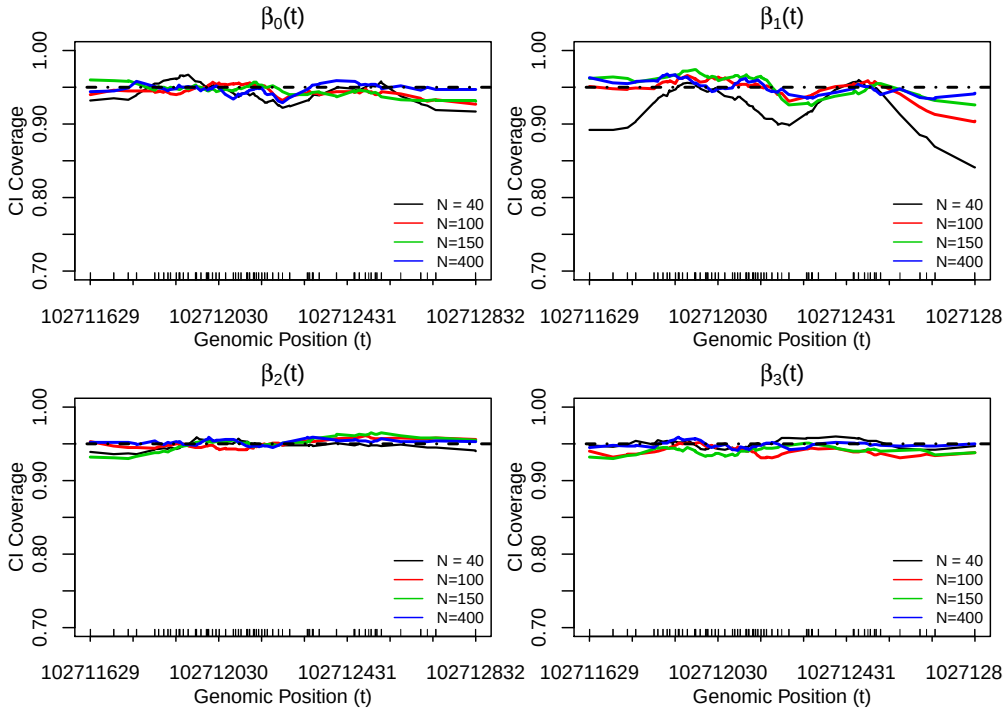


Figure 3.4: Coverage probability of confidence intervals over 1000 simulations under different sample sizes ($N = 40, 100, 150, 400$). Data were generated with error, under simulation Scenario 1.

Figure 3.4 displays the empirical coverage probabilities of CIs over 1000 simulations of Scenario 1. The empirical coverage probabilities are defined as the percentage of simulations where the analytical 95% confidence interval (proposed in Section 3.2.4) covers the true value of the parameter. Overall, the coverage probabilities for $\beta_2(t)$ and $\beta_3(t)$ with linear shapes are closer to the nominal level 95% than the two nonlinear shapes for $\beta_0(t)$ and $\beta_1(t)$. This result can be expected, because nonlinear patterns require more parameters, which leads to less accurate inference results than linear patterns, given the same amount of information. When sample size is 40, the coverages for $\beta_1(t)$ tend to be less than 95%, especially at the boundaries. This may be because $\beta_1(t)$ has a nonlinear shape with relatively small effect

sizes across the region, which poses extra difficulties in estimation compared to the shapes that are away from the null, such as $\beta_0(t)$. In summary, Figure 3.4 shows that the coverages of our 95% confidence intervals attain their nominal values in most of the simulation settings. This suggests that the proposed CI estimation approach quantifies the underlying uncertainty in the smoothed-EM estimates with reasonable accuracy, although it ignores the uncertainty from estimating the smoothing parameters.

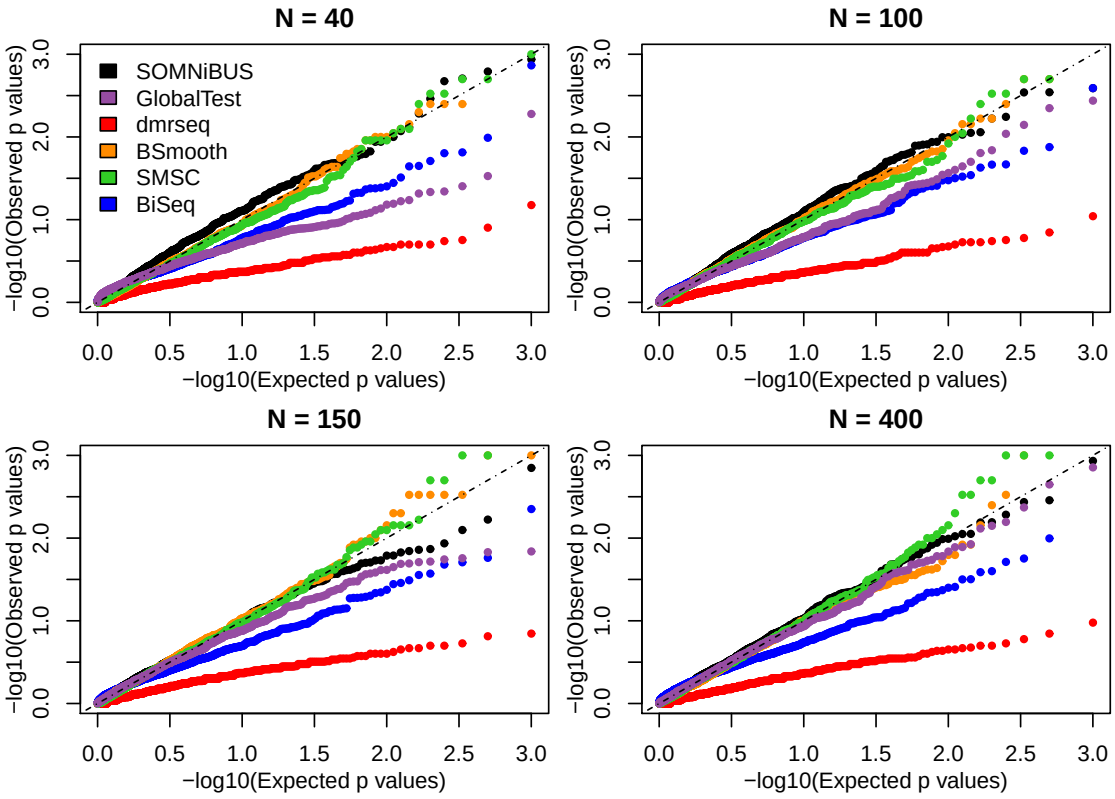


Figure 3.5: Quantile-Quantile (Q-Q) plots of the region-based p-values for the null covariate Z_3, obtained from the six methods, over 1000 simulations. Data were generated without error with a range of sample sizes ($N = 40, 100, 150, 400$), under simulation Scenario 1. Here, the Expected p-values are uniformly distributed numbers, equal to $= (1/1001, 2/1001, \ldots, 1000/1001)$.

Figures 3.5 and 3.6 further demonstrate the performance of the proposed regional test, described in Section 3.2.4. The results of type I error rate and power from our smoothed-EM method are compared to the five existing methods `GlobalTest`, `dmrseq`, `BSmooth`, `SMSC` and `BiSeq`. Figure 3.5 shows the distributions of p-values for the regional effect of the null co-

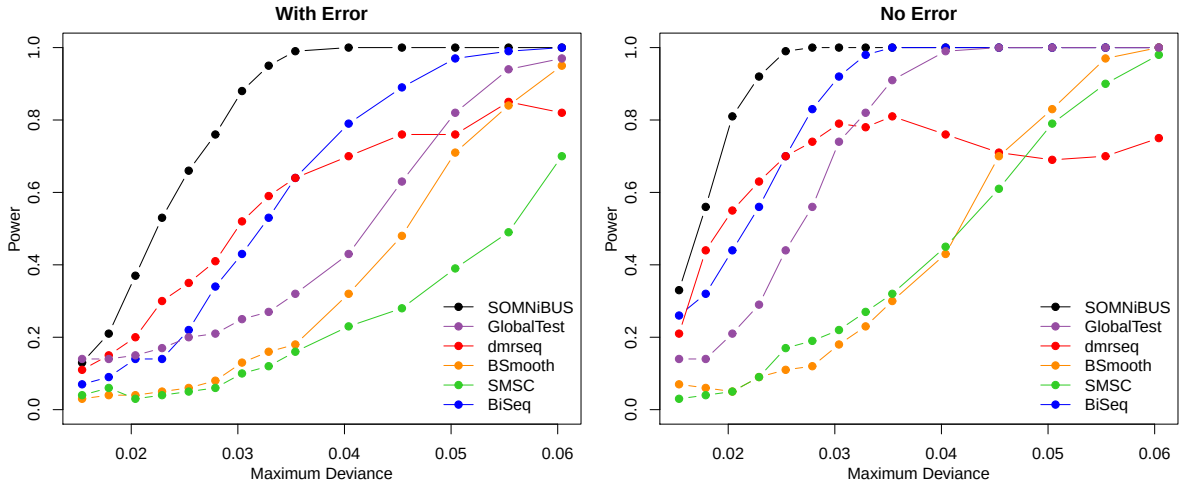


Figure 3.6: Powers to detect DMRs using the six methods for the 14 simulation settings in Scenario 2 under different levels of maximum deviation between $\pi_0(t)$ and $\pi_1(t)$, calculated over 100 simulations. (Sample size $N = 100$).

variate Z_3, obtained from the six methods. Because none of `GlobalTest`, `dmrseq`, `BSmooth` nor `BiSeq` accounts for the presence of experimental errors, for a fair comparison, the simulated data used in Figure 3.5 were generated without error (i.e. $p_0 = 1 - p_1 = 0$). The corresponding results for data generated with error are shown in the Supporting Information Figure A.2. Figure 3.5 shows that the region-based p-values for Z_3, calculated from our smoothed-EM approach (black dots), are uniformly distributed, under all sample sizes considered. In contrast, the distributions of p-values from `dmrseq`, `BiSeq` and `GlobalTest` are biased away from what would be expected under the null. Because the inferences for `BSmooth` and `SMSC` are drawn from permutations, both methods are able to control type I error. Similar results were observed when data were generated with error. The results demonstrate that the distribution of the SOMNiBUS region-based statistics under the null is well calibrated even at a relatively small sample size $N = 40$, indicating the proposed regional zero effect test can correctly control the type I error. Figure 3.6 shows the powers of the six methods for detecting DMRs under the 14 settings of methylation patterns displayed in Figure 3.2. In Figure 3.6, the left panel presents the results obtained from data with error ($p_0 = 0.003$ and $1 - p_1 = 0.1$); the right panel presents results obtained from data without

error ($p_0 = 1 - p_1 = 0$). Figure 3.6 shows that the proposed smoothed-EM method has a higher power than the five alternative methods; this superiority is even more pronounced when the data were generated with error.

In summary, `SOMNiBUS` provides accurate estimates for smooth covariate effects; when compared with the existing methods considered here, `SOMNiBUS` exhibits greater power to detect DMRs, while correctly controlling type I error rates.

3.5 Discussion

Currently, there are no tools for estimating smooth covariate effects for bisulfite sequencing data. In this paper, we propose and evaluate a method, `SOMNiBUS`, that aims to fill this gap. Our contribution is three-fold. First, we develop a novel model to represent the bisulfite sequencing data from multiple samples, which naturally accounts for variable read depth, experimental errors and a mixture of cell types. Second, we provide a formal inference for smooth covariate effects across a region of interest, where outcomes may be contaminated by errors. Third, we construct a region-based statistic with a simple chi-squared limiting distribution for jointly testing multiple coefficients in the presence of penalization. Results from simulations and one real data example show that the new method captures important underlying methylation patterns, provides accurate estimates of covariate effects, and correctly quantifies the underlying uncertainty in the estimates. The method has been implemented in R package `SOMNiBUS`, which will be submitted to CRAN.

Our method assumes that the error parameters p_0 and p_1 are known and do not vary across the region of interest. While it is conceptually feasible to estimate these parameters by an EM-type approach, the added computational burden in the E step would be substantial, because the complete-data likelihood is not linear in the methylated counts. Moreover, there are cases in which these parameters can actually be measured, for example by adding spike-

in sequences of DNA that are known in advance to be methylated or unmethylated into the bisulfite sequencing procedure. The results from the sensitivity analyses (Supplementary Information Figures S3 and S4) show that misspecified error rates can introduce a minor bias in regional p-values; however, this is not likely to affect the power of our tests, as demonstrated in the Supporting Information Table A.2. An extension worth exploring in the future will be to accommodate variations of p_0 and p_1 across genomic positions into our model. For example, the error rates could be modeled to depend on prior annotation information, CG content, or on the experimental quality in the test region.

Another potential limitation of our inference procedures is the treatment of the smoothing parameters as fixed, disregarding the uncertainty in estimating them. However, our simulation results show that both the confidence interval coverage at each site and the type I error rates at the region level, are close to their nominal value; hence, our compromise does not lead to a major efficiency loss. Nevertheless, this uncertainty could be accounted for by adding in our method an approximate correction, as proposed by Kass & Steffey (1989), or considering a full Bayesian inference where one could specify a prior distribution for the smoothing parameters $\boldsymbol{\lambda}$.

There is a substantial computational burden in our estimation algorithm, because the M step includes two inner iteration schemes: P-IRLS for updating smooth covariate effects, and Newton's optimization for updating smoothing parameters. A summary of runtimes for `SOMNiBUS` and the five alternative methods is displayed in the Supporting Information Figure A.5. This figure shows that `SOMNiBUS` requires longer computational times than `GlobalTest`, `BSmooth`, `SMSC` and `BiSeq`, but less than `dmrseq`. Note that our proposed method, `SOMNiBUS`, is capable of estimating the effects of multiple covariates simultaneously, whereas, other methods require repeating the analysis for each covariate, which will multiply the runtimes. Our algorithm could be sped up by transforming the methylation proportions into a continuous-type variable, as in Korthauer et al. (2018), which allows us to replace

the P-IRLS with the ordinary least square, and mitigate any instability in estimation of methylation levels near the boundaries (proportions of zero or one). However, transforming the count outcome into a continuous variable causes extra difficulties in the Expectation step, for which no closed-form exact expression is available.

The proposed approach is tailored to targeted bisulfite sequencing data. Another future direction is to extend our method to WGBS data. This requires first partitioning whole genome into regions or using a sliding window; optimal partitioning or choices of window sizes are challenges to be met. We recommend for the moment that algorithms such as `BSmooth` or `dmrseq` be used to find interesting regions. These regions could then be re-analyzed with `SOMNiBUS` to more comprehensively and simultaneously estimate covariate influences on methylation.

Acknowledgements

This work was supported by a Genome Canada Bioinformatics and Computational Biology 2017 (B/CB) competition grant and the Canadian Institutes of Health Research MOP 130344. K.Z was supported by a Doctoral Training Award from the Fonds de Recherche du Québec - Santé (FRQS), the McGill University Faculty of Medicine's Gerald Clavet Fellowship, as well as a scholarship from Canadian Statistical Sciences Institute - Collaborative Research Team Projects. K.O. acknowledges the Natural Sciences and Engineering Research Council of Canada and Fonds de recherché du Québec-santé grant FRQS-31110. We also acknowledge the Compute Canada Resources for Research Groups (RRG) ID 2514. We would also like to thank the editor, the associate editor, and the referee for their constructive comments that helped improve this manuscript.

Data Availability Statement

The data that support the findings in this paper are available on request from the co-author Dr. Marie Hudson. The data are not publicly available due to privacy or ethical restrictions.

Supporting Information

Web Appendices, Tables, and Figures, referenced in Section 3.2, 3.3, 3.4 and 3.5, are available with this paper at the *Biometrics* website on Wiley Online Library. Codes to replicate the simulation results in the article are deposited in the Github repository `https://github.com/kaiqiong/SOMNiBUS_Simu`. The R package, `SOMNiBUS`, implementing the proposed method is available from Github at `https://github.com/GreenwoodLab/SOMNiBUS`, with a user guide.

Chapter 4

Manuscript II: A hierarchical quasi-binomial varying coefficient mixed model for detecting differentially methylated regions in bisulfite sequencing data

Preamble to Manuscript II: In Chapter 3, I introduced a new method SOMNiBUS for the analysis of regional associations in sequencing-derived DNA methylation data. This method explicitly accounts for measurement errors in the methylated counts and leads to both regional measures of association and pointwise tests and confidence intervals. However, it had an important limitation: its underlying binomial assumption may be overly restrictive and only applicable when data exhibit variability levels similar to those anticipated based on a binomial distribution. The goal of the second manuscript in this thesis is to extend the standard SOMNiBUS to allow the outcomes to exhibit extra-parametric variations.

To achieve that, I propose a hierarchical quasi-binomial varying coefficient mixed model. This model allows for both multiplicative and additive dispersion, thereby providing a plausible representation of realistic dispersion trends observed in regional methylation data. In addition, the new approach can provide reliable inference for differential methylation at the region level. The methodology improvement has been implemented in the R Bioconductor package `SOMNiBUS` (https://www.bioconductor.org/packages/devel/bioc/html/SOMNiBUS.html).

Note that the supporting material for this chapter can be found in Appendix B.

A hierarchical quasi-binomial varying coefficient mixed model for detecting differentially methylated regions in bisulfite sequencing data

Kaiqiong Zhao[1,2], Karim Oualkacha[3], Lajmi Lakhal-Chaieb[4], Aurélie Labbe[5], Kathleen Klein[2], Sasha Bernatsky[6,7], Marie Hudson[2,6], Inés Colmegna[6,7], Celia M.T. Greenwood[1,2,8,9]

[1]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University*

[2]*Lady Davis Institute for Medical Research, Jewish General Hospital*

[3]*Département de Mathématiques, Université du Québec à Montrèal*

[4]*Département de Mathématiques et de Statistique, Université Laval*

[5]*Département des Sciences de la Décision, HEC Montrèal*

[6]*Department of Medicine, McGill University*

[7]*The Research Institute of the McGill University Health Centre*

[8]*Department of Human Genetics, McGill University*

[9]*Gerald Bronfman Department of Oncology, McGill University*

Abstract

Identifying disease-associated changes in DNA methylation can help to gain a better understanding of disease etiology. Bisulfite sequencing technology allows the generation of methylation profiles at single base of DNA. We previously developed a method for estimating smooth covariate effects and identifying differentially methylated regions (DMRs) from bisulfite sequencing data, which copes with experimental errors and variable read depths; this method utilizes the binomial distribution to characterize the variability in the methylated counts. However, bisulfite sequencing data frequently include low-count integers and can exhibit over or under dispersion relative to the binomial distribution. We present a substantial improvement to our previous work by proposing a quasi-likelihood-based regional testing approach which accounts for multiplicative and additive sources of dispersion. We demonstrate the theoretical properties of the resulting tests, as well as their marginal and conditional interpretations. Simulations show that the proposed method provides correct inference for smooth covariate effects and captures the major methylation patterns with excellent power.

4.1 Introduction

Conceptually, the emergence of a disease phenotype is believed to stem from the combined effects of genetic predisposition and environmental exposures (Ober & Vercelli, 2011). A plausible mechanism behind this gene-environment interplay is epigenetic modification, which regulates gene activity through modifications of DNA accessibility. Epigenetics may explain how exposures leave heritable marks on the genome that impact disease susceptibility (Jaenisch & Bird, 2003). Therefore, increased understanding of epigenetic-disease association could lead to novel insights into disease causation and possible therapies (Feinberg, 2007).

The most studied epigenetic mark is DNA methylation, which involves the covalent addition of a methyl group to a cytosine nucleotide. DNA methylation, in the mammalian genomes, occurs predominantly at cytosine-guanine dinucleotides (i.e. CpG sites) (Lister et al., 2009). Methylation of CpG-rich promoters can silence gene expression by preventing transcriptional factor binding to DNA (Choy et al., 2010). More generally, DNA methylation has the potential to activate or repress gene expression, depending on whether the mark inactivates a positive or negative regulatory element (Jones, 1999). Known or suspected drivers behind methylation alterations include genetic variations (McRae et al., 2014), environmental toxins (Hanson & Gluckman, 2008), external stressors (Dolinoy et al., 2007) and aging (Horvath, 2013). There is also evidence that localized abnormal methylation is strongly linked to many diseases, including breast cancer (Hu et al., 2005), autism spectrum disorder (Dunaway et al., 2016), and systemic autoimmune disease (Kato et al., 2005).

High-resolution, large-scale measurement of DNA methylation is now possible with recent advances in bisulfite sequencing (BS-seq) protocol, which is implemented either genome-wide or in targeted regions. Although whole-genome bisulfite sequencing (WGBS) allows a comprehensive characterization of the methylation landscape, it is inefficient for large-scale studies as only 20% or less of CpGs are thought to have variable methylation across individ-

uals or tissues (Ziller et al., 2013). On the other hand, Targeted Custom Capture Bisulfite Sequencing (TCCBS) platform enables a comprehensive yet cost-effective interrogation of functional CpGs in disease-targeted tissues or cells (Allum et al., 2015). This approach has been successfully used to identify novel disease-associated epigenetic variants (Allum et al., 2019; Shao et al., 2019; Ziller et al., 2016). In this work, we aim to improve sensitivity to detect, among all the regions targeted by TCCBS, differentially methylated regions (DMRs) that are associated with phenotypes or traits.

Like other sequencing experiments, the raw data from TCCBS are short sequence reads. After proper alignment and data processing, the methylation level at a single cytosine can be summarized as a pair of counts: the number of methylated reads and the total number of reads covering the site, i.e. read depth. Such data possess several challenges for statistical analysis. Typically, read depth varies drastically across sites and individuals, which leads to measures with wide-ranging precision and many missing values (Sims et al., 2014). Additional statistical challenges are created by the strong spatial correlations observed in methylation levels at neighboring CpG sites (Hansen et al., 2012; Korthauer et al., 2018; Rackham et al., 2017; Shokoohi et al., 2018), as well as the possibility of data errors, arising from excessive or insufficient bisulfite treatment or other aspects of the sequencing processes (Cheng & Zhu, 2013; Lakhal-Chaieb et al., 2017). Furthermore, in addition to the trait of interest (e.g. disease or treatment group), other factors, such as age (Horvath, 2013), batch effects (Leek et al., 2010), or cell-type mixture proportions (for mixed tissue samples) (McGregor et al., 2016) have effects on methylation levels. Hence, it is desirable to adjust methylation signals for multiple covariates simultaneously.

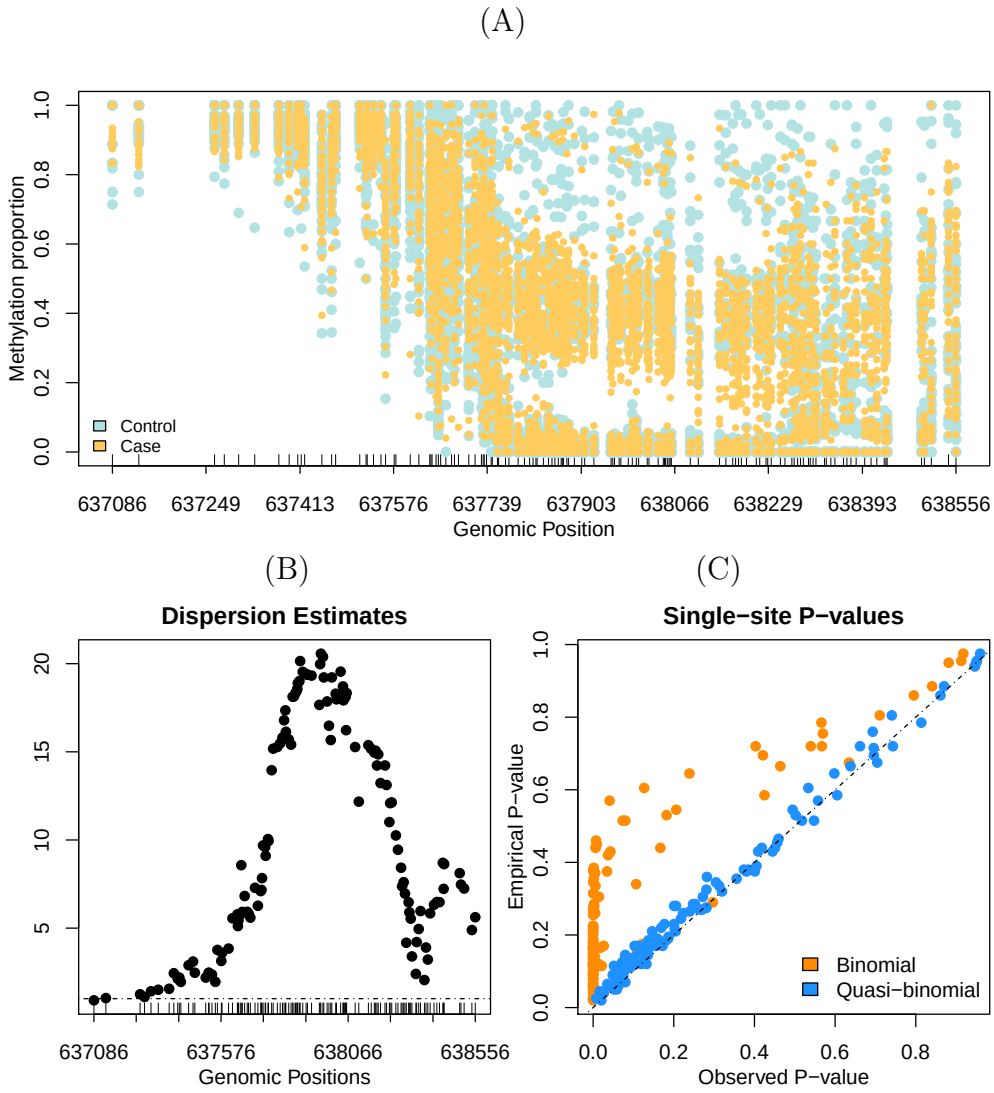


Figure 4.1: *Illustration of observed dispersion in a targeted region that underwent bisulfite sequencing.* (A) Observed methylation proportions in one region for two groups of samples (yellow and blue); data are fully described in Section 4.4. (B) Estimated dispersion for each CpG site from a single-site quasi-binomial GLM. (C) Single-site p-values for methylation difference between the two groups. Horizontal axis are the p-values estimated from either binomial (ignoring dispersion) or quasi-binomial (accounting for dispersion) GLMs. Vertical axis shows the empirical p-values computed from 199 permutations; the empirical p-value is a benchmark for valid statistical tests. (Single-site beta-binomial regression models generate similar dispersion estimate pattern and p-value distribution to quasi-binomial GLM).

To detect truly differentially methylated regions without finding false associations, it is crucial to accurately account for the sources of variability across individuals. We ran into this issue in a recent analysis of methylation profiles and anti-citrullinated protein antibodies

(ACPA). Figure 4.1 (A) illustrates methylation proportions in a targeted region for samples from this study. (A full description of the study, referred to as the ACPA dataset, is in Section 4.4). Clearly, dispersion is much larger between samples in the blue group. In panel (C), it can be seen that p-values testing for methylation differences, assuming a binomial mean-variance relationship are much too small. In contrast, allowing for dispersion through a quasi-binomial model provides p-values in line with null expectation for this region. As such, the restrictive mean-variance relationship implied by a binomial generalized linear model (GLM) may not adequately accommodate the data variability, and thus can lead to inflation of false positives. This is known as over or underdispersion, i.e. data presenting greater or lower variability than assumed by a GLM model.

Moving in this direction, we have developed a SmOoth ModeliNg of BisUlfite Sequencing (SOMNiBUS) method to detect DMRs in targeted bisulfite sequencing data (Zhao et al., 2020). The method provides a general framework of analysis, and simultaneously addresses regional testing, estimation of multiple covariate effects, adjustment for read depth variability and experimental errors. Specifically, Zhao et al. (2020) proposed a hierarchical binomial regression model, which allows covariate effects to vary smoothly along genomic position. A salient feature of SOMNiBUS is its one-stage nature. Several existing methods first smooth methylation data and then, in a second stage, estimate covariate effects based on the smoothed data (Hansen et al., 2012; Hebestreit et al., 2013; Lakhal-Chaieb et al., 2017), and this two-stage framework could lead to biased uncertainty estimates. In contrast, SOMNiBUS collapses smoothing and testing steps into a single step, and achieves accurate statistical uncertainty assessment of DMRs. That said, its underlying binomial assumption may be overly restrictive and is only applicable when data exhibit variability levels that are similar to those anticipated based on a binomial distribution (such as data from inbred animal or cell line experiments). In this work, we propose an extension of SOMNiBUS, which maintains all the good properties of the standard SOMNiBUS, and at the same time explicitly allows the variability in regional methylation counts to exceed or fall short of what

binomial model permits.

The importance of accounting for dispersion in BS-seq data has been well recognized in analysis of single CpG sites. Faced with dispersion in discrete data analysis, one commonly used option is to convert the methylated and total counts to proportions. In this way, testing of differentially methylated single CpG sites can be done via the two sample t-test (Hansen et al., 2012) or beta regression (Hebestreit et al., 2013), both of which allow direct computation of (within-group) sample variation. However, this conversion loses information, since it fails to distinguish between noisy and accurate measurements (Wu et al., 2015), often as a consequence of the stochasticity of read depth, and also disregards the discrete nature of the data (Lea et al., 2015). On the other hand, there are approaches for DNA methylation analysis that directly model counts while accounting for dispersion. These count-based approaches use either *additive* overdispersion models, or *multiplicative* under- or overdispersion models to describe the variation driving the dispersion (Browne et al., 2005). In a multiplicative model, one includes a multiplicative scale factor, i.e. the dispersion parameter, in the variance of the binomial response. Thus, the dispersion inflates or deflates the variance estimates of the covariate effect by the multiplicative factor. Such approaches include the quasi-binomial regression model (Akalin et al., 2012) and the beta-binomial regression model (Dolzhenko & Smith, 2014; Feng et al., 2014; Park et al., 2014; Park & Wu, 2016). In contrast, additive overdispersion methods add a subject-level random effect (RE) to capture the extra-binomial variation among individual observations. Both ABBA (Rackham et al., 2017) and MACAU (Lea et al., 2015), that use binomial mixed effect models fall in this category. An advantage of the multiplicative approach, particularly the quasi-binomial model, is that it naturally allows for both overdispersion and underdispersion, whereas the additive model only allows overdispersion. On the other hand, the additive overdispersion approach links directly with a multilevel model and can be readily extended to analyze data with a hierarchical or clustering structure.

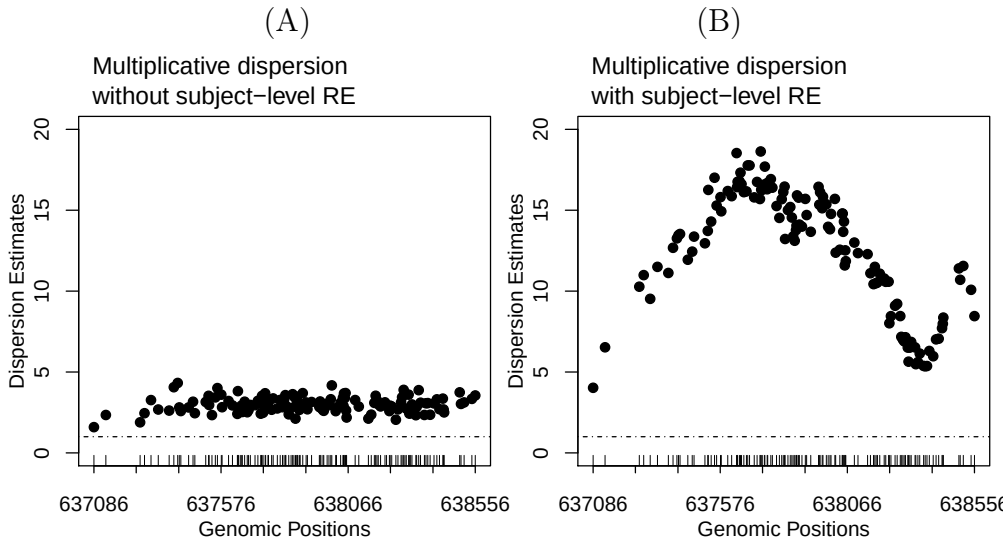


Figure 4.2: *A byproduct of introducing a subject-level RE, on top of a multiplicative dispersion parameter, to a model with smooth covariate effects is a regional dispersion pattern of varying degree.* Estimated dispersion for each CpG site obtained from a single-site quasi-binomial GLM, for two simulated regional methylation datasets: (A) data were simulated from a multiplicative-dispersion-only model ($\phi = 3, \sigma_0^2 = 0$), and (B) data were simulated from a model with both a multiplicative dispersion and a subject-level RE ($\phi = 3, \sigma_0^2 = 3$); see Section 4.2 for detailed model formulations and notation definitions.

The challenge of accounting for dispersion when detecting DMRs is further complicated by several factors. Firstly, even within a small genomic region, different CpG sites may exhibit different levels of dispersion and strong spatial correlation (Figure 4.1 B). Hence, a multiplicative dispersion model with a common dispersion parameter does not adequately capture the dispersion heterogeneity across loci (Figure 4.2 A). In addition, challenges are presented by the complex correlation structure in the regional methylation data. Apart from the spatial correlations among neighboring CpGs, there are additional correlations among methylation measurements on the same subject. Ignoring this within-subject correlation could lead to overestimation of precision and invalid statistical tests (Cui et al., 2016). One means to accommodate such a correlation structure is to add a subject-level RE that can also capture the overdispersion induced by independent variation across different subjects. Furthermore, when modelling discrete data with a hierarchical structure, extra non-structural specific random dispersion can arise, beyond that introduced by the subject-level RE (Breslow &

Clayton, 1993; Molenberghs et al., 2007; Vahabi et al., 2019), and thus, often, parametric distributions with restrictive mean-variance relations poorly describe the outcomes for individual subjects (i.e. the conditional distribution of outcome given the RE) (Ivanova et al., 2014; Molenberghs et al., 2010, 2012). Hence, properly addressing both multiplicative and additive sources of dispersion in methylation data is essential for making reliable inference at the region level.

Table 4.1: List of existing DNA methylation analytical methods and our proposal with their capabilities.

Method	regional	one-stage	count-based	read-depth variability	adjust for confounding	within-subject correlation	non-structural dispersion	varying levels of dispersion across loci	experimental errors
dSOMNiBUS	✓	✓	✓	✓	✓	✓	✓	✓	✓
SOMNiBUS	✓	✓	✓	✓	✓				✓
BSmooth	✓			✓*			✓	✓	
SMSC	✓			✓*			✓	✓	✓
dmrseq	✓	✓		✓	✓	✓	✓		
Biseq	✓			✓*	✓	✓	✓	✓	
GlobalTest	✓	✓			✓	NA†	NA†	NA†	
ABBA	✓	✓	✓	✓		✓	✓	✓	
MACAU		✓	✓	✓	✓	NA‡	✓	✓	

✓*: These three methods are of a two-stage nature. Their smoothing stage indeed accounts for read-depth variability, but their testing stage, which relies on t-test or beta regression, ignores the read-depth variability.

†: GlobalTest treats methylation levels at multiple loci as covariates and trait of interest as outcome. It is not necessary for GlobalTest to account for the three features on covariance structure of methylation across samples and loci.

‡: MACAU is a single-site method and within-subject correlation is irrelevant when analyzing individual sites one at a time.

Given our preliminary exploration of dispersion in the ACPA dataset, we recognized the need for a regional one-stage method of analysis that accommodates both the hierarchically-induced overdispersion (and/or correlation) and the extra unstructured individual dispersion. This desired method should also simultaneously address discrete nature of the data, varying strength of dispersion across a region, estimation of multiple covariate effects, adjustment for read depth variability and experimental errors. However, to the best of our knowledge, none of the existing methods meet all aforementioned objectives (Table 4.1). For example, dmrseq (Korthauer et al., 2018), which fits a generalized least squares regression model with autoregressive error structure to the transformed methylation proportions, accommodates both within-subject correlation and non-structural dispersion, but it assumes a constant dispersion parameter for all loci in a region. Biseq (Hebestreit et al., 2013) is capable of

capturing the covariance structure of regional methylation data (by estimating the variogram of site-specific test statistics). However, this method separates smoothing and inference steps and its final significance assessment does not account for the uncertainty in the smoothing step.

To overcome the limitations and challenges of existing methods, we propose a novel approach for identifying DMRs, dSOMNiBUS (dispersion-adjusted SmOoth ModeliNg of BisUlfite Sequencing). Our strategy directly models raw read counts while accounting for all (known) sources of data variability and varying degree of dispersion across loci, thus providing accurate assessments of regional statistical significance.

Specifically, we propose a quasi-binomial mixed model to describe bisulfite sequencing data, which allows covariate effects to vary smoothly along genomic positions, and specially, captures the extra-binomial variation by the *combination* of a subject-specific RE (i.e an additive overdispersion) and a multiplicative dispersion. The RE term accounts for between-sample heterogeneity, and at the same time enables flexible dispersion patterns in a region (Figure 4.2 B), which is highly plausible in methylation data (Figure 4.1 B). The multiplicative dispersion, on the other hand, explicitly allows the variability in individual subject's methylation levels to exceed or fall short of what binomial distribution assumes, and thus captures the extra dispersion that cannot explained by RE. We also demonstrate their marginal and conditional interpretations.

In addition, our approach accounts for possible data errors in the observed methylated counts. Specifically, we assume that the observed read counts arise from an unobserved latent true methylation state compounded by errors. To estimate such a hierarchical model, we build a hybrid Expectation-Solving (ES) algorithm, which has a special treatment for the multiplicative dispersion parameter and results in a regional association test statistic with a simple F limiting distribution. We have demonstrated the properties of the resulting estimators using both simulation evaluations and data applications.

4.2 A hierarchical quasi-binomial varying coefficient mixed model

4.2.1 Notation and data

We consider DNA methylation measures over a targeted genomic region from N independent samples. Let m_i be the number of CpG sites for the i^{th} sample, $i = 1, 2, \ldots N$. We write t_{ij} for the genomic position (in base pairs) for the i^{th} sample at the j^{th} CpG site, $j = 1, 2, \ldots, m_i$. Methylation levels at a site are quantified by the number of methylated reads and the total number of reads. We define X_{ij} as the total number of reads aligned to CpG j from sample i. We denote the *true* methylation status for the k^{th} read obtained at CpG j of sample i as S_{ijk}, where $k = 1, 2, \ldots X_{ij}$. For a single DNA strand read, S_{ijk} is binary and we define $S_{ijk} = 1$ if the corresponding read is methylated and $S_{ijk} = 0$ otherwise. In the presence of experimental errors, the *observed* methylation status, written as Y_{ijk} can be different from the true underlying information S_{ijk}. We define $Y_{ijk} = 1$ if the corresponding read is observed as methylated and $Y_{ijk} = 0$ otherwise. We additionally denote the *true* and *observed* methylated counts at CpG j for sample i with $S_{ij} = \sum_{k=1}^{X_{ij}} S_{ijk}$, and $Y_{ij} = \sum_{k=1}^{X_{ij}} Y_{ijk}$, respectively. Furthermore, we assume that we have the information on P covariates for the N samples, denoted as $\boldsymbol{Z_i} = (Z_{1i}, Z_{2i}, \ldots Z_{Pi})$, for $i = 1, 2, \ldots N$.

4.2.2 A hierarchical quasi-binomial varying coefficient mixed model

In the presence of experimental errors, the true methylation data, S_{ij} are unknown and one only observes Y_{ij}. We assume the following error mechanism

$$\begin{aligned} P(Y_{ijk} = 1 \mid S_{ijk} = 0) &= p_0 \\ P(Y_{ijk} = 1 \mid S_{ijk} = 1) &= p_1. \end{aligned} \tag{4.1}$$

Here, p_0 is the rate of false methylation calls, and $1 - p_1$ is the rate of false non-methylation calls. These rates are assumed to be constant across all reads and positions. The error parameters p_0 and p_1 can be estimated by looking at raw sequencing data at CpG sites known in advance to be methylated or unmethylated (Wreczycka et al., 2017). We assume hereafter that p_0 and p_1 are known.

We then propose a quasi-binomial varying coefficient mixed effect model to describe the relationship between the true methylated counts, S_{ij} for $j = 1, 2, \ldots m_i$, and the sample-level covariates $\boldsymbol{Z}_i$. Specifically,

$$\begin{aligned} \log \frac{\pi_{ij}}{1-\pi_{ij}} &= \beta_0(t_{ij}) + \sum_{p=1}^{P} \beta_p(t_{ij}) Z_{pi} + u_i, \qquad (4.2) \\ u_i &\overset{iid}{\sim} N(0, \sigma_0^2) \\ \mathbb{V}\text{ar}(S_{ij} \mid u_i) &= \phi X_{ij} \pi_{ij} (1 - \pi_{ij}) \qquad (4.3) \end{aligned}$$

where $\pi_{ij} = \mathbb{E}\left(S_{ij} \mid u_i\right) / X_{ij}$ is the *individual's* methylation proportion (i.e. the conditional mean), $\beta_0(t_{ij})$ and $\{\beta_p(t_{ij})\}_{p=1}^{P}$ are functional parameters for the intercept and covariate effects on π_{ij}, and σ_0^2 is the random effect variance. In this model, we assume the underlying proportion of methylated reads for the i^{th} sample at the j^{th} CpG site, π_{ij}, depends on covariates $\boldsymbol{Z}_i$ and on nearby methylation patterns through a logit link function. In addition, each π_{ij} incorporates a subject-specific random intercept (i.e. an additive overdispersion) u_i that is normally distributed and independent across samples. The inclusion of u_i allows for sample heterogeneity in baseline methylation patterns, and at the same time accounts for the correlation among methylation measurements taken on the same sample. Moreover, we assume the variance of S_{ij} for individual samples to be a product of a multiplicative dispersion parameter ϕ and a known mean-variance function implied by a binomial distribution (i.e. $V(\pi_{ij}) = X_{ij}\pi_{ij}(1 - \pi_{ij})$).

Both the random effects $\boldsymbol{u} = (u_1, u_1, \ldots u_N)^T$ and the multiplicative dispersion parameter ϕ

capture extra-binomial dispersion. However, they address two different aspects of dispersion: $\boldsymbol{u}$ models the variation that is due to independent noise across samples, while ϕ aims to relax the assumption of the conditional distribution of S_{ij} given $\boldsymbol{u}$ such that it is not confined to a binomial distribution. In fact, our model generalizes the binomial-based model in Zhao et al. (2020) by introducing both the additive dispersion term $\boldsymbol{u}$ and multiplicative dispersion term ϕ. Specially, imposing $\phi = 1$ in model (4.2) leads to an additive-dispersion-only model and $\sigma_0^2 = 0$ corresponds to a multiplicative-dispersion-only model. When $\sigma_0^2 = 0$ and $\phi = 1$, our model reduces to the binomial-based model in Zhao et al. (2020).

4.2.3 Marginal interpretations

A key feature of the mixed effect model in (4.2) is that the regression coefficients $\beta_p(t_{ij})$ need to be interpreted conditional on the value of random effect u_i. For example, $\beta_p(t_{ij})$ describes how an *individual's* methylation proportions in a region depend on covariate Z_p. If one desires estimates of such covariate effects on the average population, it is more appropriate to determine the marginal model implied by (4.2). After applying a cumulative Gaussian approximation to the logistic function and taking an expectation over u_i, it can be shown that the marginal mean, π_{ij}^M, has the form

$$\pi_{ij}^M = \mathbb{E}(S_{ij})/X_{ij} \approx g\left(\sum_{p=0}^{P} a\ \beta_p(t_{ij}) Z_{pi}\right), \tag{4.4}$$

where $g(x) = 1/\left(1+\exp(-x)\right)$, $Z_{0i} \equiv 1$, and the constant $a = (1 + c^2\sigma_0^2)^{-1/2}$ with $c = \sqrt{3.41}/\pi$; see detailed derivations in Appendix B.1.1. The approximation in (4.4) is quite accurate with errors ≤ 0.001. Thus, the marginal mean induced by our mixed effect model depends on the covariates Z_p through a logistic link with attenuated regression coefficients $a\beta_p(t_{ij})$. Although the smooth covariate effect parameters $\beta_p(t_{ij})$ have no marginal interpretation, they do have a strong relationship to their marginal counterparts. Hence, the

results from hypothesis testing $H_0 : \beta_p(t_{ij}) = 0$ describe the significance of the covariate effect on both the population-averaged and an individual's DNA methylation levels across a region.

Similarly, the marginal variance of S_{ij} does not coincide with its conditional counterpart as shown in (4.3). Specifically, our mixed effect model implies a marginal variance of S_{ij} defined as

$$\begin{aligned}\mathbb{V}\text{ar}(S_{ij}) \;\approx\; & X_{ij}\pi^{\star}_{ij}(1-\pi^{\star}_{ij})\left\{\phi + \sigma_0^2\left(X_{ij}-\phi\right)\pi^{\star}_{ij}(1-\pi^{\star}_{ij})\right. \\ & \left. + \sigma_0^2/2(1-2\pi^{\star}_{ij})^2\left[1+\sigma_0^2\pi^{\star}_{ij}(1-\pi^{\star}_{ij})(X_{ij}-\phi-1/2)\right]\right\}, \qquad (4.5)\end{aligned}$$

where $\pi^{\star}_{ij} = g\left(\sum_{p=0}^{P}\beta_p(t_{ij})Z_{pi}\right)$; see detailed derivations in Appendix B.1.1. Note that $\pi^{\star}_{ij}$ is the mean methylation proportion when setting random effects u_i to zero and is related to the marginal mean π^{M}_{ij} via $\pi^{\star}_{ij} = g\left(g^{-1}\left(\pi^{M}_{ij}\right)/a\right)$. Equation (4.5) illustrates that, under the dSOMNiBUS model, the marginal variance of methylated counts at a CpG site is approximately the variance of the binomial model multiplied by a dispersion factor $\phi^{\star} = \phi + \sigma_0^2\left(X_{ij}-\phi\right)\pi^{\star}_{ij}(1-\pi^{\star}_{ij}) + \sigma_0^2/2(1-2\pi^{\star}_{ij})^2\left[1+\sigma_0^2\pi^{\star}_{ij}(1-\pi^{\star}_{ij})(X_{ij}-\phi-1/2)\right]$, which depends on the combined effect of ϕ, the multiplicative dispersion for the conditional variance given the RE, and σ_0^2, the variance of the subject-level RE. Notably, the marginal dispersion factor $\phi^{\star}$ also depends on genomic position t_{ij} via the dependence of $\pi^{\star}_{ij}$ on t_{ij}. Consequently, our dSOMNiBUS model in (4.2) naturally allows dispersion levels to vary across loci, whereas a multiplicative-dispersion-only model (i.e. $\sigma_0^2 = 0$) can only accommodate constant dispersion in a region, as illustrated in Figure 4.2. It is also clear from Equation (4.5) that an additive-dispersion-only model (i.e., $\phi = 1$) only allows for overdispersion, and the combination of additive and multiplicative dispersion naturally accounts for both over- and underdispersion.

4.3 Inference

In this section, we present the methodology details on how to make inference about covariate effects $\beta_p(t_{ij})$ and simultaneously estimate the additive and multiplicative dispersion parameters ϕ and σ_0^2 in our smoothed quasi-binomial mixed model (4.2). We start with the case where true methylation counts S_{ij} are available, and determine the complete data marginal quasi-likelihood function in Section 4.3.1. Then we describe the estimating algorithms for the complete and contaminated data in Sections 4.3.2 and 4.3.3, respectively. We additional estimate the pointwise CIs for covariate effects $\beta_p(t_{ij})$ and obtain tests of hypotheses for these effects in Section 4.3.4.

4.3.1 Laplace-approximated marginal quasi-likelihood function

Basis representation

In model (4.2), the function parameters $\beta_p(t_{ij})$ can be represented by the coefficients of chosen spline bases of rank L_p, $\beta_p(t_{ij}) = \sum_{l=1}^{L_p} \alpha_{pl} B_l^{(p)}(t_{ij})$, for $p = 0, 1, \ldots P$. Here $\left\{B_l^{(p)}(\cdot)\right\}_{l=1}^{L_p}$ denotes the spline basis, and $\boldsymbol{\alpha_p} = (\alpha_{p1}, \ldots \alpha_{pL_p})^T \in \mathcal{R}^{L_p}$ are the coefficients to be estimated. In this way, we can write the conditional mean in (4.2) in a compact way as

$$g^{-1}(\boldsymbol{\pi}) = \mathbb{X}^{(B)}\boldsymbol{\alpha} + \mathbb{X}^{(1)}\boldsymbol{u},$$

where $\boldsymbol{\pi} = (\pi_{11}, \ldots \pi_{1m_1}, \pi_{21}, \ldots \pi_{2m_2}, \ldots \pi_{Nm_N})^T \in [0,1]^M$ with $M = \sum_{i=1}^N m_i$, $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T, \ldots \boldsymbol{\alpha}_p^T\right)^T \in \mathcal{R}^K$ with $K = \sum_{p=0}^P L_p$, and $\boldsymbol{u} = (u_1, u_2, \ldots u_N)^T$. $\mathbb{X}^{(B)}$ is the spanned design matrix for $\boldsymbol{\alpha}$ of dimension $M \times K$, stacked with elements $B_l^{(p)}(t_{ij}) \times Z_{pi}$ with $Z_{0i} \equiv 1$. $\mathbb{X}^{(1)}$ is a random effect model matrix of dimension $M \times N$, with element 1 if the corresponding CpG site in the row belongs to the sample in the column, and 0 otherwise. If we write the overall spanned design matrix $\mathbb{X} = \left[\mathbb{X}^{(B)}, \mathbb{X}^{(1)}\right] \in \mathcal{R}^{M\times(K+N)}$ and $\boldsymbol{\mathcal{B}} = (\boldsymbol{\alpha}^T, \boldsymbol{u}^T)^T$, the conditional

mean can be further simplified as

$$g^{-1}(\boldsymbol{\pi}) = \mathbb{X}\boldsymbol{\mathcal{B}}.$$

Smoothness penalty

To impose the assumption that the true covariate effect function is more likely to be smooth than jumpy, we add a smoothness penalty for each $\beta_p(t)$, $p = 0, 1, \ldots P$. The total amount of such penalty is an aggregate from all smooth terms, i.e.

$$\mathcal{L}^{\text{Smooth}} = \sum_{p=0}^{P} \lambda_p \int \left(\beta_p''(t)\right)^2 dt = \sum_{p=0}^{P} \lambda_p \boldsymbol{\alpha_p}^T \boldsymbol{A_p} \boldsymbol{\alpha_p} = \boldsymbol{\alpha}^{\boldsymbol{T}} \boldsymbol{A_\lambda} \boldsymbol{\alpha}, \tag{4.6}$$

where $\boldsymbol{A_p}'s$ are $L_p \times L_p$ positive semidefinite matrices with the (l, l') element $\boldsymbol{A_p}(l, l') = \int {B^{(p)}}_l''(t) {B^{(p)}}_{l'}''(t) dt$, which are fixed quantities given the specified set of bases. The weights λ_p, i.e. the smoothing parameters, are positive parameters which establish a tradeoff between the closeness of the curve to the data and the smoothness of the fitted curves. $\boldsymbol{A}_\lambda$ is a $K \times K$ positive semidefinite block diagonal matrix of the form $\boldsymbol{A_\lambda} = \text{Diag}\left\{\lambda_0 \boldsymbol{A_0}, \lambda_1 \boldsymbol{A_1}, \ldots, \lambda_P \boldsymbol{A_P}\right\}$.

Random-effect view of the smoothness penalty. As justified in Wahba (1983) and Silverman (1985), employing such smoothing penalty (4.6) during fitting is equivalent to imposing random effects for spline coefficients $\boldsymbol{\alpha}$. Specifically, $\boldsymbol{\alpha}$ is assumed to follow a (degenerate) multivariate normal distribution with precision matrix $\boldsymbol{A_\lambda}$,

$$\boldsymbol{\alpha} \sim MVN(\boldsymbol{0}, \boldsymbol{A_\lambda}^-),$$

where $\boldsymbol{A_\lambda}^-$ is the pseudoinverse of $\boldsymbol{A_\lambda}$. From a Bayesian viewpoint, imposing smoothness is equivalent to specifying a prior distribution on function roughness. This random-effect formulation of the smooth curve estimation problem opens up the possibility of estimating $\boldsymbol{\lambda}$ and ϕ using marginal (quasi-)likelihood maximization. In addition, under such a formulation,

it requires no extra effort to estimate the 'actual' RE term $\boldsymbol{u}$ in our model (4.2), once the inference procedure for $\boldsymbol{\alpha}$ is well established. In the rest of inference steps, we treat $\boldsymbol{\alpha}$ as random effects.

Conditional quasi-likelihood function

We first consider specifying the conditional "distribution" of $\boldsymbol{S}$ given the values of REs $\boldsymbol{\mathcal{B}}$. Following the notion of extended quasi-likelihood (McCullagh & Nelder, 1989b, Section 9.6), we define the following conditional quasi-likelihood

$$qL^{(\boldsymbol{S}|\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}},\phi) \propto \exp\left\{-\frac{1}{2\phi}\sum_{i,j} d_{ij}\left(S_{ij},\pi_{ij}\right) - \frac{M}{2}\log\phi\right\}, \tag{4.7}$$

where

$$d_{ij}(S_{ij},\pi_{ij}) = -2\int_{S_{ij}/X_{ij}}^{\pi_{ij}} \frac{S_{ij}-X_{ij}\pi_{ij}}{\pi_{ij}(1-\pi_{ij})}d\pi_{ij}$$

is the quasi-deviance function corresponding to a single observation. It can be easily checked that this quasi-likelihood exhibits the properties of log-likelihood, with respect to $\boldsymbol{\mathcal{B}}$. Such properties approximately hold for the dispersion parameter ϕ, provided that ϕ be small and $\kappa_r = O(\phi^{r-1})$, where κ_r is the rth-order cumulant of $\boldsymbol{S} \mid \boldsymbol{B}$ (Efron, 1986; Jørgensen, 1987; McCullagh & Nelder, 1989b). Let $ql^{(\boldsymbol{S}|\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}},\phi) = \log\left[qL^{(\boldsymbol{S}|\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}},\phi)\right]$ denote the conditional log-quasi-likelihood. It should be noted that the integral inside $ql^{(\boldsymbol{S}|\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}},\phi)$ rarely needs to be evaluated for the estimation of $\boldsymbol{\mathcal{B}}$, because the inference described later only requires the computation of its first and second derivatives, i.e.

$$\frac{\partial ql^{(\boldsymbol{S}|\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}},\phi)}{\partial\boldsymbol{\mathcal{B}}} = \frac{1}{\phi}\mathbb{X}^T\left(\boldsymbol{S}-\boldsymbol{\Lambda}_{\boldsymbol{X}}\boldsymbol{\pi}\right),$$

$$\frac{\partial^2 ql^{(\boldsymbol{S}|\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}},\phi)}{\partial\boldsymbol{\mathcal{B}}\partial\boldsymbol{\mathcal{B}}^T} = -\frac{1}{\phi}\mathbb{X}^T\boldsymbol{W}\mathbb{X},$$

where $\boldsymbol{\Lambda_X} \in \mathbb{R}^{M\times M}$ is the diagonal matrix with values of read-depths, and $\boldsymbol{W}$ is the weight matrix whose diagonal is $X_{ij}\pi_{ij}(1-\pi_{ij})$.

Joint quasi-likelihood functions

For notational simplicity, we write $\boldsymbol{\Theta} = (\boldsymbol{\lambda}, \sigma_0^2)$ for the parameters involved in the covariance structure of random effects $\boldsymbol{\mathcal{B}}$. Combining the conditional 'distribution' $\boldsymbol{S} \mid \boldsymbol{\mathcal{B}}$ with the marginal distribution of $\boldsymbol{\mathcal{B}}$, we obtain the following *joint* log-quasi-likelihood of the observed data $\boldsymbol{S}$ and unobserved random effects $\boldsymbol{\mathcal{B}}$

$$\begin{aligned} q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi, \boldsymbol{\Theta}) &= ql^{(\boldsymbol{S}|\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi) \underbrace{-\frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{A}_{\boldsymbol{\lambda}} \boldsymbol{\alpha} - \frac{1}{2\sigma_0^2}\boldsymbol{u}^T\boldsymbol{u}}_{-\frac{1}{2\phi}\boldsymbol{\mathcal{B}}^T\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\boldsymbol{\mathcal{B}}} \\ &\quad + \underbrace{\frac{1}{2}\log\left\{|\boldsymbol{A}_{\boldsymbol{\lambda}}|_+\right\} + \frac{N}{2}\log\left(1/\sigma_0^2\right)}_{1/2\log\{|\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}/\boldsymbol{\phi}|_+\}}, \end{aligned} \tag{4.8}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}} = \operatorname{diag}\left\{\phi\boldsymbol{A}_{\boldsymbol{\lambda}}, \phi/\sigma_0^2\boldsymbol{I}_N\right\} \in \mathbb{R}^{(K+N)\times(K+N)}$, and $|\bullet|_+$ denotes the generalized determinant of a matrix, i.e. the product of its non-zero eigenvalues. Here we introduce the scaling by ϕ in $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ merely for later convenience, and this allows us to factor out the dispersion parameter ϕ in the penalized quasi-score in (4.12). In such way, the point estimates of random effects $\boldsymbol{\mathcal{B}}$ are independent of the estimate of ϕ.

This joint log-quasi-likelihood is composed of three parts: 1) the outcome 'distribution' depending on $\boldsymbol{\mathcal{B}}$ and ϕ, 2) multiple quadratic penalties for $\boldsymbol{\mathcal{B}}$ depending on regularization parameters $\boldsymbol{\Theta}$, and 3) fixed regularized terms for $\boldsymbol{\Theta}$. Our goals are to estimate the variance component parameters $\boldsymbol{\Theta}$, the dispersion parameter ϕ, and also predict the values of random effects $\boldsymbol{\mathcal{B}}$. When $\phi = 1$, this fits a generalized linear mixed model (GLMM).

Laplace-approximated marginal quasi-likelihood function

A legitimate (quasi-)likelihood is the *marginal* 'density' evaluated at the observed data $\boldsymbol{S}$ only, which is obtained by integrating out random effects $\boldsymbol{\mathcal{B}}$ from the joint quasi-likelihood of $\boldsymbol{S}$ and $\boldsymbol{\mathcal{B}}$,

$$qL^M(\phi, \boldsymbol{\Theta}) = \int \exp\left\{q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi, \boldsymbol{\Theta})\right\} d\boldsymbol{\mathcal{B}}. \tag{4.9}$$

Conceptually, maximizing $qL^M(\phi, \boldsymbol{\Theta})$ yields the maximum quasi-likelihood estimators for $\boldsymbol{\Theta}$, and ϕ. However, the analytical solutions for this high-dimensional integral are not easy to find, and an approximation approach is needed.

As in Wood (2011), we use the Laplace approximation to evaluate the integral inside the marginal quasi-likelihood. Let $\widehat{\boldsymbol{\mathcal{B}}}_{\boldsymbol{\Theta}}$ be the value of $\boldsymbol{\mathcal{B}}$ maximizing the joint quasi-likelihood $q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi, \boldsymbol{\Theta})$ given the values of variance component parameters $\boldsymbol{\Theta}$, i.e.

$$\widehat{\boldsymbol{\mathcal{B}}}_{\boldsymbol{\Theta}} = \text{argmax}\left\{ql^{(\boldsymbol{S}|\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi) - \frac{1}{2\phi}\boldsymbol{\mathcal{B}}^T\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\boldsymbol{\mathcal{B}}\right\}, \tag{4.10}$$

where terms not dependent on $\boldsymbol{\mathcal{B}}$ have been dropped from the joint quasi-likelihood. The objective function in (4.10) is often referred to as the penalized (quasi-)likelihood. A second-order Taylor expansion of $q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi, \boldsymbol{\Theta})$, around $\widehat{\boldsymbol{\mathcal{B}}}$ (the subscript $\boldsymbol{\Theta}$ has been dropped for notational simplicity), gives

$$q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi, \boldsymbol{\Theta}) \approx q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\widehat{\boldsymbol{\mathcal{B}}}, \phi, \boldsymbol{\Theta}) - \frac{1}{2}\left(\boldsymbol{\mathcal{B}} - \widehat{\boldsymbol{\mathcal{B}}}\right)^T \boldsymbol{H}_{\widehat{\boldsymbol{\mathcal{B}}}}\left(\boldsymbol{\mathcal{B}} - \widehat{\boldsymbol{\mathcal{B}}}\right),$$

where $\boldsymbol{H}_{\widehat{\boldsymbol{\mathcal{B}}}} = -\nabla^2_{\boldsymbol{\mathcal{B}}}\, q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\widehat{\boldsymbol{\mathcal{B}}}, \phi, \boldsymbol{\Theta}) = \frac{1}{\phi}\left(\mathbb{X}^T\widehat{\boldsymbol{W}}\mathbb{X} + \boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\right)$. Therefore, the *marginal* quasi-

likelihood in (4.9) can be approximately written as

$$
\begin{aligned}
qL^{M}(\phi, \boldsymbol{\Theta}) &\approx \exp\left\{q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\widehat{\boldsymbol{\mathcal{B}}}, \phi, \boldsymbol{\Theta})\right\} \int \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mathcal{B}}-\widehat{\boldsymbol{\mathcal{B}}}\right)^{T} \boldsymbol{H}_{\widehat{\boldsymbol{\mathcal{B}}}}\left(\boldsymbol{\mathcal{B}}-\widehat{\boldsymbol{\mathcal{B}}}\right)\right\} d\boldsymbol{\mathcal{B}} \\
&\approx \exp\left\{q\ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\widehat{\boldsymbol{\mathcal{B}}}, \phi, \boldsymbol{\Theta})\right\} \frac{\sqrt{2\pi}^{K+N}}{\left|\dfrac{\mathbb{X}^{T}\widehat{\boldsymbol{W}}\mathbb{X}+\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}}{\phi}\right|^{1/2}} \\
&\propto \phi^{-M/2} \exp\left(-\frac{\sum_{i,j}\widehat{d}_{ij}}{2\phi}\right) \exp\left(-\frac{1}{2\phi}\widehat{\boldsymbol{\mathcal{B}}}^{T}\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\widehat{\boldsymbol{\mathcal{B}}}\right) |\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}/\boldsymbol{\phi}|_{+}^{1/2} \left|\frac{\mathbb{X}^{T}\widehat{\boldsymbol{W}}\mathbb{X}+\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}}{\phi}\right|^{-1/2}.
\end{aligned}
\tag{4.11}
$$

In equation (4.11), $\widehat{d}_{ij} = d_{ij}(S_{ij}, \widehat{\pi}_{ij})$, where $\widehat{\pi}_{ij} = g^{-1}(\mathbb{X}_{(l,)}\widehat{\boldsymbol{\mathcal{B}}})$ and l is the row in the model matrix $\mathbb{X}$ corresponding to CpG j for sample i. We denote this Laplace-approximated marginal quasi-likelihood in (4.11) as $qL^{\text{Laplace}}(\phi, \boldsymbol{\Theta}; \widehat{\boldsymbol{\mathcal{B}}})$ and simply write $\text{Laplace}(\phi, \boldsymbol{\Theta}; \widehat{\boldsymbol{\mathcal{B}}}) = \log[qL^{\text{Laplace}}(\phi, \boldsymbol{\Theta}; \widehat{\boldsymbol{\mathcal{B}}})]$, which depends on $\boldsymbol{\Theta}$ via the dependence of $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\mathcal{B}}}$ (and thus $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{d}}$) on $\boldsymbol{\Theta}$.

4.3.2 Estimation algorithm for the complete data

The essence of estimating $\boldsymbol{\Theta}, \boldsymbol{\mathcal{B}}$, and ϕ, is to optimize the Laplace-approximated marginal quasi-likelihood in (4.11). Note that such approximation requires calculating the maximum of the penalized quasi-likelihood in (4.10), $\widehat{\boldsymbol{\mathcal{B}}}$, along with its corresponding Hessian $\boldsymbol{H}_{\widehat{\boldsymbol{\mathcal{B}}}}$, which is only feasible for given values of the penalty parameters $\boldsymbol{\Theta}$. To disentangle the complicated dependence of $\widehat{\boldsymbol{\mathcal{B}}}$ on $\boldsymbol{\Theta}$, we adopt a nested-optimization strategy proposed by Wood (2011). Specifically, the algorithm has an outer iteration for updating $\boldsymbol{\Theta}$ and ϕ, with each iterative step supplementing with an inner iteration to estimate random effects $\boldsymbol{\mathcal{B}}$ corresponding to the current $\boldsymbol{\Theta}$, as summarized in Algorithm 1. This Section proceeds with the detailed description of each step in Algorithm 1.

Algorithm 1: Algorithm to find $(\widehat{\boldsymbol{\mathcal{B}}}, \widehat{\phi}, \widehat{\boldsymbol{\Theta}}) = \text{argmax}_{\boldsymbol{\mathcal{B}},\phi,\boldsymbol{\Theta}}\, \ell^{(\boldsymbol{S},\boldsymbol{\mathcal{B}})}(\boldsymbol{\mathcal{B}}, \phi, \boldsymbol{\Theta})$ using data $\{\boldsymbol{S}, \boldsymbol{Z}, \boldsymbol{Y}\}$

Initialize $\boldsymbol{\Theta}^{(0)}, \phi^{(0)}$; Choose $\varepsilon = 10^{-6}$; Set $s = 0$;

repeat

Step 1. Solve $\boldsymbol{U}(\boldsymbol{\mathcal{B}}; \boldsymbol{\Theta}^{(s)}) = \mathbf{0}$ (4.12) to obtain $\boldsymbol{\mathcal{B}}^{(s)}$;

Step 2. Newton's update for the Laplace-approximated marginal likelihood

$$(\log(\phi), \log(\boldsymbol{\Theta}))^{(s+1)} = (\log(\phi), \log(\boldsymbol{\Theta}))^{(s)} - \left[\nabla^2 \text{Laplace}(\boldsymbol{\mathcal{B}}^{(s)})\right]^{-1} \nabla \text{Laplace}(\boldsymbol{\mathcal{B}}^{(s)});$$

$s \leftarrow s + 1$;

until $\|\boldsymbol{\mathcal{B}}^{(s)} - \boldsymbol{\mathcal{B}}^{(s-1)}\|_2 < \varepsilon$;

Return $\boldsymbol{\Theta}^{(s)}, \boldsymbol{\mathcal{B}}^{(s)}, \phi^{(s)}$;

Step 3: Calculate $\widehat{\phi}_{Fle}$ using $\boldsymbol{\mathcal{B}}^{(s)}$

Inner iteration: estimate $\mathcal{B}$ given the current Θ

Given the estimates of penalty parameters $\boldsymbol{\Theta}$, $\widehat{\boldsymbol{\mathcal{B}}}$ can be computed as the solution to

$$\boldsymbol{U}(\boldsymbol{\mathcal{B}}) = \frac{1}{\phi}\left\{\mathbb{X}^T(\boldsymbol{S} - \boldsymbol{\Lambda}_{\boldsymbol{X}}\boldsymbol{\pi}) - \boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\boldsymbol{\mathcal{B}}\right\} = \mathbf{0}, \tag{4.12}$$

where $\boldsymbol{U}(\boldsymbol{\mathcal{B}})$ is the *quasi-score* for the penalized quasi-likelihood in (4.10) with respect to $\boldsymbol{\mathcal{B}}$. We use the Newton's method to solve these system of nonlinear equations. Specifically we compute the gradient of $\boldsymbol{U}(\boldsymbol{\mathcal{B}})$,

$$\nabla \boldsymbol{U}(\boldsymbol{\mathcal{B}}) = -\frac{\mathbb{X}^T\boldsymbol{W}\mathbb{X} + \boldsymbol{\Sigma}_{\boldsymbol{\Theta}}}{\phi},$$

and a single update from step l to step $l+1$ for $\boldsymbol{\mathcal{B}}$ thus takes the form

$$\boldsymbol{\mathcal{B}}^{(l+1)} = \boldsymbol{\mathcal{B}}^{(l)} + \left(\mathbb{X}^T\boldsymbol{W}\mathbb{X} + \boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\right)^{-1}\left[\mathbb{X}^T\left(\boldsymbol{S} - \boldsymbol{\Lambda}_{\boldsymbol{X}}\boldsymbol{\pi}^{(l)}\right) - \boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\boldsymbol{\mathcal{B}}^{(l)}\right].$$

We then iteratively update $\boldsymbol{\mathcal{B}}$ until convergence, which constitutes iteration Step 1 in Algorithm 1.

Outer iteration: maximize the Laplace-approximated marginal quasi-likelihood

The outer iteration, which aims to maximize the Laplace-approximated marginal quasi-likelihood in (4.11), is also achieved by a Newton's method. Wood (2011) has derived the derivatives and Hessian of Laplace$(\phi, \boldsymbol{\Theta}; \widehat{\boldsymbol{\mathcal{B}}})$ with respect to $\boldsymbol{\rho} = (\log(\boldsymbol{\Theta}), \log(\phi))$, using a mixture of implicit and direct differentiations. We denote these first and second derivatives as $\nabla\text{Laplace}(\boldsymbol{\rho}; \widehat{\boldsymbol{\mathcal{B}}})$ and $\nabla^2\text{Laplace}(\boldsymbol{\rho}; \widehat{\boldsymbol{\mathcal{B}}})$, respectively. Relying on the work of Wood (2011), the maximization in the outer iteration can be readily achieved via

$$\boldsymbol{\rho}^{(s+1)} = \boldsymbol{\rho}^{(s)} - \left[\nabla^2\text{Laplace}\left(\boldsymbol{\rho}^{(s)}; \widehat{\boldsymbol{\mathcal{B}}}^{(s)}\right)\right]^{-1} \nabla\text{Laplace}\left(\boldsymbol{\rho}^{(s)}; \widehat{\boldsymbol{\mathcal{B}}}^{(s)}\right). \tag{4.13}$$

Here, $\widehat{\boldsymbol{\mathcal{B}}}^{(s)}$ are the estimated mean parameters given the current $\boldsymbol{\Theta}^{(s)}$, obtained from the inner iteration in Section 4.3.2. Each update in (4.13) constitutes iteration Step 2 in Algorithm 1. We iterate between the Step 1 and Step 2 until convergence to obtain $\widehat{\boldsymbol{\mathcal{B}}}$, $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\phi}$.

Estimating ϕ using the moment-based estimator

As described in the previous section, the dispersion parameter ϕ can be estimated as part of the outer iteration of the marginal quasi-likelihood maximization. We refer to this estimator as likelihood-based dispersion estimator, denoted as $\widehat{\phi}_{Lik}$.

In generalized linear models, it is common to estimate ϕ by dividing Pearson's lack-of-fit statistic by the residual degrees of freedom, and this is known as the moment-based scale/dispersion estimator. We can apply the similar ideas here. Instead of using $\widehat{\phi}_{Lik}$, we take one step further and estimate ϕ using the final estimate $\widehat{\boldsymbol{\mathcal{B}}}$ (and thus $\widehat{\boldsymbol{\pi}}$). Specifically, Pearson's dispersion estimator can be written as

$$\widehat{\phi}_P = \frac{1}{M - \tau} \sum_{i,j} \left(\frac{S_{ij} - X_{ij}\widehat{\pi}_{ij}}{\sqrt{X_{ij}\widehat{\pi}_{ij}(1 - \widehat{\pi}_{ij})}} \right)^2.$$

Here τ is the effective degrees of freedom (Wood, 2017), defined as

$$\tau = \text{trace}\left(\boldsymbol{F}\right), \text{ with } \boldsymbol{F} = \left(\mathbb{X}^T \widehat{\boldsymbol{W}} \mathbb{X} + \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\Theta}}}\right)^{-1} \mathbb{X}^T \widehat{\boldsymbol{W}} \mathbb{X}. \tag{4.14}$$

However, $\widehat{\phi}_P$ can be unstable at finite sample sizes, especially when a few Pearson residuals are huge (Farrington, 1995; Fletcher, 2012). For example, in our model, $\widehat{\pi}_{ij}$ close to 0 can lead to a huge Pearson residual, even though the deviance $d_{ij}(S_{ij}, \widehat{\pi}_{ij})$ in (4.7) is modest. Therefore, we adopt an improved version of the Pearson estimator, i.e. the Fletcher estimator (Fletcher, 2012), which is designed to mitigate this problem. The Fletcher's dispersion estimator $\widehat{\phi}_{Fle}$ is defined as

$$\widehat{\phi}_{Fle} = \frac{\widehat{\phi}_P}{1+\overline{a}}, \text{ where } a_{ij} = \frac{1-2\widehat{\pi}_{ij}}{X_{ij}\widehat{\pi}_{ij}(1-\widehat{\pi}_{ij})}\left(S_{ij} - X_{ij}\widehat{\pi}_{ij}\right) \text{ and } \overline{a} = \frac{1}{M}\sum_{i,j} a_{ij}.$$

If the mean model is adequate, then approximately we have

$$\frac{(M-\tau)\widehat{\phi}_{Fle}}{\phi} \sim \chi^2_{M-\tau} \tag{4.15}$$

(Fletcher, 2012; McCullagh, 1985). Therefore, $\widehat{\phi}_{Fle}$ provides an unbiased estimator for ϕ, which is also confirmed by simulation results as shown in Supporting Information Figure B.9. In contrast, the estimation using $\widehat{\phi}_{Lik}$ can be considerably biased (Supporting Information Figure B.9). Hence, we calculate the moment-based estimate for the dispersion parameter, which constitutes the Step 3 in Algorithm 1.

4.3.3 Estimating algorithm for the contaminated data

In the presence of experimental errors, the true methylation data, S_{ij} are unknown and one only observes Y_{ij}, which is assumed to be a mixture of binomial counts arising from both the truly methylated and truly unmethylated reads. When S_{ij} is modeled by a parametric

distribution, like in Zhao et al. (2020), the EM algorithm (Dempster et al., 1977) provides accurate estimation of the smooth covariate effects even though the true methylation data are missing. Motivated by the work of Elashoff & Ryan (2004), we propose an extension of the EM algorithm with special treatment for the multiplicative dispersion parameter ϕ, to the case of quasi-likelihood-based analyses.

Expectation-Solving algorithm

Elashoff & Ryan (2004) proposed an extension of the EM algorithm, called Expectation-Solving (ES) algorithm, to accommodate missing (or mis-measured) data when a natural set of estimating equations exists for the complete data setting. Specifically, the E step computes the conditional expectation of the estimating equations given the observed data, and S step solves these expected estimating equations.

To apply the ES algorithm to our case, we need to evaluate the conditional expectation of three sets of estimating equations:

$$\begin{aligned}
\boldsymbol{U}(\boldsymbol{\mathcal{B}}; \boldsymbol{\Theta}^{(s)}, \boldsymbol{S}) &= \frac{1}{\phi}\left[\mathbb{X}^T\left(\boldsymbol{S} - \boldsymbol{\Lambda}_{\boldsymbol{X}}\boldsymbol{\pi}^{(s)}\right) - \boldsymbol{\Sigma}_{\boldsymbol{\Theta}^{(s)}}\boldsymbol{\mathcal{B}}\right] = \mathbf{0} \\
\nabla_{\boldsymbol{\Theta}}\text{Laplace}(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}, \boldsymbol{S}) &= \frac{1}{\phi}\sum_{i,j}\left\{\frac{S_{ij} - X_{ij}\pi_{ij}^{(s)}}{\pi_{ij}^{(s)}(1-\pi_{ij}^{(s)})} \times \frac{d\pi_{ij}^{(s)}}{d\boldsymbol{\Theta}}\right\} + f_1(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}) = \mathbf{0} \\
\nabla_{\phi}\text{Laplace}(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}, \boldsymbol{S}) &= \frac{1}{\phi^2}\sum_{i,j}\int_{S_{ij}/X_{ij}}^{\pi_{ij}^{(s)}} \frac{S_{ij} - X_{ij}\pi_{ij}}{\pi_{ij}(1-\pi_{ij})} d\pi_{ij} + f_2(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}) = \mathbf{0},
\end{aligned}$$

for $\boldsymbol{\mathcal{B}}, \boldsymbol{\Theta}$ and ϕ, respectively. Here, $\boldsymbol{\Theta}^{(s)}$, $\boldsymbol{\mathcal{B}}^{(s)}$, and $\boldsymbol{\pi}^{(s)}$ are estimates from the previous iterations, $f_1(\cdot)$ and $f_2(\cdot)$ denote the components that are independent of $\boldsymbol{S}$.

E step for $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\Theta}$. The estimating equations for $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\Theta}$ are linear in the latent methylated counts $\boldsymbol{S}$, and thus their expectations equal $\boldsymbol{U}(\boldsymbol{\mathcal{B}}; \boldsymbol{\Theta}^{(s)}, \boldsymbol{\eta}^{\star})$ and $\nabla_{\boldsymbol{\Theta}}\text{Laplace}(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}, \boldsymbol{\eta}^{\star})$, respectively. Here, $\boldsymbol{\eta}^{\star} \in \mathcal{R}^M$ are the conditional expectations of $\boldsymbol{S}$ given $\boldsymbol{Y}$ evaluated at the

trial estimates $(\boldsymbol{\mathcal{B}}^\star, \boldsymbol{\Theta}^\star)$, and for our model, take the form

$$\eta_{ij}^\star = \mathbb{E}\left(S_{ij} \mid Y_{ij}; \boldsymbol{\mathcal{B}}^\star, \boldsymbol{\Theta}^\star\right) = \frac{Y_{ij} p_1 \pi_{ij}^\star}{p_1 \pi_{ij}^\star + p_0(1-\pi_{ij}^\star)} + \frac{(X_{ij} - Y_{ij})\,(1-p_1)\pi_{ij}^\star}{(1-p_1)\pi_{ij}^\star + (1-p_0)(1-\pi_{ij}^\star)}, \quad (4.16)$$

where $\pi_{ij}^\star = g^{-1}(\mathbb{X}_{(l,)}\boldsymbol{\mathcal{B}}^\star)$ and l is the row in the model matrix $\mathbb{X}$ corresponding to CpG j for sample i. These expected estimating equations can then be solved using the direct nested iteration method in Algorithm 1.

E step for ϕ. However, the estimating equation for ϕ is not linear in the unknown methylated counts $\boldsymbol{S}$; see details in Appendix B.1.2. Therefore, the closed-form exact expression for $\mathbb{E}_{\boldsymbol{S}|Y;\boldsymbol{\mathcal{B}}^\star,\boldsymbol{\Theta}^\star}(\nabla_\phi \text{Laplace}(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}, \boldsymbol{S}))$ is not available, and the E-S algorithm cannot be readily applied to estimating ϕ from the contaminated data. To circumvent this problem, we propose a direct method to estimate ϕ without undergoing the E-S iteration.

A plug-in estimator for ϕ

Specifically, we estimate ϕ by exploiting its relationship with the dispersion for the observed outcome $\boldsymbol{Y}$, denoted as ϕ_{ij}^Y, which is defined as

$$\phi_{ij}^Y = \frac{\mathbb{V}\text{ar}(Y_{ij} \mid u_i)}{X_{ij}\pi_{ij}^Y(1-\pi_{ij}^Y)}, \text{ with } \pi_{ij}^Y = \mathbb{E}(Y_{ij} \mid u_i) = \pi_{ij}p_1 + (1-\pi_{ij})p_0.$$

Based on our assumed mean-variance relationship (4.3) and error model (4.1), we can express ϕ_{ij}^Y in terms of ϕ, π_{ij} and error parameters p_0 and p_1,

$$\phi_{ij}^Y = 1 + (\phi - 1)\frac{(\pi_{ij}^Y - p_0)(p_1 - \pi_{ij}^Y)}{\pi_{ij}^Y(1-\pi_{ij}^Y)}; \quad (4.17)$$

see detailed derivations in Appendix B.1.2. Although we assume a constant dispersion ϕ for the true outcome $\boldsymbol{S}$, the observed outcome $\boldsymbol{Y}$ implied by our error model, possesses dispersion parameter ϕ_{ij}^Y varying with each CpG site, when $\phi \neq 1$.

Directly running the nested iteration method (Algorithm 1) on the observed data $\{\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}\}$ reports a constant dispersion estimate $\widehat{\phi^Y}$ and $\widehat{\pi}_{ij}^Y$ for all i and j, along with other useful estimates. We assume that $\widehat{\phi^Y}$ is an estimate for the mean of individual dispersions ϕ_{ij}^Y, i.e.

$$\frac{1}{M}\sum_{i,j}\phi_{ij}^Y = 1 + (\phi - 1)\frac{1}{M}\sum_{i,j}\frac{(\pi_{ij}^Y - p_0)(p_1 - \pi_{ij}^Y)}{\pi_{ij}^Y(1-\pi_{ij}^Y)}; \tag{4.18}$$

empirical results show that this is a reasonable assumption, as shown in Supporting Information Figure B.11. We then propose to estimate ϕ by plugging in the error-prone outcome-related estimates $\widehat{\phi^Y}$ and $\widehat{\pi}_{ij}^Y$ to the relation in (4.18):

$$\widehat{\phi} = (\widehat{\phi^Y} - 1)\left[\frac{1}{M}\sum_{i,j}\frac{(\widehat{\pi}_{ij}^Y - p_0)(p_1 - \widehat{\pi}_{ij}^Y)}{\widehat{\pi}_{ij}^Y(1-\widehat{\pi}_{ij}^Y)}\right]^{-1} + 1.$$

A hybrid ES algorithm

We propose a hybrid ES algorithm to estimate our model using the error-prone outcomes $\boldsymbol{Y}$. We first estimate ϕ using the aforementioned plug-in approach and then estimate $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\Theta}$ using ES iterations assuming ϕ is fixed and known; detailed steps are summarized in Algorithm 2. We denote the final estimates from our algorithm as $\widehat{\phi}$, $\widehat{\boldsymbol{\mathcal{B}}}$ and $\widehat{\boldsymbol{\Theta}}$. The components of $\widehat{\boldsymbol{\alpha}}$ inside the vector of $\widehat{\boldsymbol{\mathcal{B}}}$ leads to estimates of the functional parameters $\beta_p(t)$, for $p = 0, 1, \ldots, P$:

$$\widehat{\beta_p(t)} = \left\{\boldsymbol{B}^{(p)}(t)\right\}^T \{\widehat{\boldsymbol{\alpha}_p}\},$$

where t is a genomic position lying within the range of the input positions $\{t_{ij}\}$, and $\boldsymbol{B}^{(p)}(t) = (B_1^{(p)}(t), B_2^{(p)}(t), \ldots B_{L_p}^{(p)}(t))^T \in \mathcal{R}^{L_p}$ is a column vector with nonrandom quantities obtained from evaluating the set of basis functions $\{B_l^{(p)}(\cdot)\}_l$ at position t.

Algorithm 2: A hybrid ES algorithm to estimate the smoothed quasi-binomial mixed model with error-prone outcomes.

Step 1: run Algorithm 1 on $\{\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}\}$; return $\widehat{\boldsymbol{\pi}}^{Y}$, $\widehat{\phi}_Y$, $\widehat{\boldsymbol{B}}$, and $\widehat{\boldsymbol{\Theta}}$;

Step 2: calculate the plug-in estimator $\widehat{\phi}$;

Step 3: E-S iterations with ϕ fixed at $\widehat{\phi}$ to estimate $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\Theta}$; specifically Initialize $\boldsymbol{\Theta}^{(0)} = \widehat{\boldsymbol{\Theta}}, , \boldsymbol{\mathcal{B}}^{(0)} = \widehat{\boldsymbol{B}}$; Choose $\varepsilon = 10^{-6}$; Set $\ell = 0$;

repeat

- E step: $\eta_{ij}^{(\ell)} = \mathbb{E}(S_{ij} \mid Y_{ij}; \boldsymbol{\mathcal{B}}^{(\ell)})$;
- S step: $(\boldsymbol{\mathcal{B}}^{(\ell)}, \boldsymbol{\Theta}^{(\ell)}) = \text{argmax}_{\boldsymbol{\mathcal{B}},\boldsymbol{\Theta}}\, \ell^{(\boldsymbol{\mathcal{B}},\boldsymbol{\Theta})}\left(\boldsymbol{\mathcal{B}}, \boldsymbol{\Theta}; \eta_{ij}^{(\ell)}, \widehat{\phi}\right)$. Specifically **repeat**
  - Solve $\boldsymbol{U}(\boldsymbol{\mathcal{B}}; \boldsymbol{\Theta}^{(s)}; \boldsymbol{\eta}^{(\ell)}) = \boldsymbol{0}$ to obtain $\boldsymbol{\mathcal{B}}^{(s)}$ using data $\eta_{ij}^{(\ell)}$;
  - Newton's update for the Laplace approximated marginal likelihood evaluated at data $\eta_{ij}^{(\ell)}$:

    $$(\log \boldsymbol{\Theta})^{(s+1)} = (\log \boldsymbol{\Theta})^{(s)} - \left[\nabla^2_{\boldsymbol{\Theta}} \text{Laplace}(\boldsymbol{\mathcal{B}}^{(s)})\right]^{-1} \nabla_{\boldsymbol{\Theta}} \text{Laplace}(\boldsymbol{\mathcal{B}}^{(s)});$$

  - $s \leftarrow s + 1$;

  until $\|\boldsymbol{\mathcal{B}}^{(s)} - \boldsymbol{\mathcal{B}}^{(s-1)}\|_2 < \varepsilon$;
- $\ell \leftarrow \ell + 1$;

until $\|\boldsymbol{\mathcal{B}}^{(\ell)} - \boldsymbol{\mathcal{B}}^{(\ell-1)}\|_2 < \varepsilon$;

Return $\boldsymbol{\Theta}^{(\ell)}, \boldsymbol{\mathcal{B}}^{(\ell)}$;

4.3.4 Inference for smooth covariate effects

We then estimate the pointwise confidence intervals (CI) for the smoothed covariate effects $\{\beta_1(t), \beta_2(t), \ldots, \beta_P(t)\}$, and obtain tests of hypotheses for these effects. Note that the inference is carried out conditional on the values of variance component parameters $\boldsymbol{\Theta}$ and dispersion parameter ϕ, i.e. the uncertainty in estimating them is not accounted for.

Estimating the variance of the resulting parameter estimates

As did in Elashoff & Ryan (2004), we can re-express the E step as the solution to the following M-dimensional estimating equation:

$$\boldsymbol{U}^{(2)}(\boldsymbol{S}) \;\; = \;\; \boldsymbol{S} - \widehat{\boldsymbol{\eta}} = \boldsymbol{0},$$

where $\widehat{\boldsymbol{\eta}}$ are the conditional expectations in (4.16) evaluated at the current estimate $\widehat{\boldsymbol{\pi}}$. In this way, the overall ES algorithm can be viewed as solving an expanded set of equations of dimension $K+N+M$, whose first $K+N$ components are $\boldsymbol{U}(\boldsymbol{\mathcal{B}})=\mathbf{0}$ in (4.12) and whose second M components are $\boldsymbol{U}^{(2)}(\boldsymbol{S})=\mathbf{0}$.

Under this formulation, we use the established theory for estimating equations (Heyde & Morton, 1996; Lindsay, 1982; Small et al., 2003), and propose a model-based variance estimator for $\widehat{\boldsymbol{\mathcal{B}}}$. Specifically, under correct specification of the first two moments of $\boldsymbol{S}$, the asymptotic variance of $\widehat{\boldsymbol{\mathcal{B}}}$ can be written as

$$\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\mathcal{B}}})=\left[(-\boldsymbol{D})^{-1}\right]_{(\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{B}})},$$

where $\boldsymbol{D}$ is the first order derivative of the expanded estimating equations for $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{S}$, and $[\,\bullet\,]_{(\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{B}})}$ stands for the matrix block corresponding to $\boldsymbol{\mathcal{B}}$. In our case, $\boldsymbol{D}$ takes the form

$$\boldsymbol{D}=-\begin{bmatrix}\frac{1}{\phi}\mathbb{X}^T\boldsymbol{W}\mathbb{X}+\frac{1}{\phi}\boldsymbol{\Sigma}_{\boldsymbol{\Theta}} & -\frac{1}{\phi}\mathbb{X}^T\\ \boldsymbol{W}_{\boldsymbol{\delta}}\mathbb{X} & -\boldsymbol{I}_M.\end{bmatrix}$$

Here, $\boldsymbol{W}_{\boldsymbol{\delta}}$ is a diagonal matrix with elements $X_{ij}\delta_{ij}$, where

$$\delta_{ij}=\frac{Y_{ij}p_1p_0}{\left[p_1\pi_{ij}+p_0(1-\pi_{ij})\right]^2}+\frac{(X_{ij}-Y_{ij})\,(1-p_1)(1-p_0)}{\left[(1-p_1)\pi_{ij}+(1-p_0)(1-\pi_{ij})\right]^2},$$

and reduces to a zero matrix when $p_0=1-p_1=0$. Then, the asymptotic variance of $\widehat{\boldsymbol{\mathcal{B}}}$ can be simplified as

$$\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\mathcal{B}}})=\left[\mathbb{X}^T(\boldsymbol{W}-\boldsymbol{W}_{\boldsymbol{\delta}})\mathbb{X}+\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\right]^{-1}\phi. \tag{4.19}$$

Therefore, the desired variance estimator of $\widehat{\boldsymbol{\mathcal{B}}}$ can be obtained by plugging in the final estimates $\widehat{\boldsymbol{\mathcal{B}}}$, $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\phi}$ into equation (4.19).

Confidence interval estimation

Let $\widehat{\boldsymbol{V}}$ denote the aforementioned variance estimator and $\widehat{\boldsymbol{V}}_{\boldsymbol{p}}$ be the diagonal blocks of $\widehat{\boldsymbol{V}}$ corresponding to $\boldsymbol{\alpha}_p$, with dimensions $L_p \times L_p$. We then immediately have the estimated variance of $\widehat{\beta_p(t)}$: $\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta_p(t)}) = \left\{\boldsymbol{B}^{(p)}(t)\right\}^T \widehat{\boldsymbol{V}}_{\boldsymbol{p}} \left\{\boldsymbol{B}^{(p)}(t)\right\}$. Therefore, the confidence interval for $\beta_p(t)$ at significance level ν can be approximately estimated by $\widehat{\beta_p(t)} \pm \mathbb{Z}_{\nu/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta_p(t)})}$, for any t in the range of interest, where $\mathbb{Z}_{\nu/2}$ is $\nu/2$ (upper-tail) quantile of a standard normal distribution.

Hypothesis testing for a regional zero effect

We can also construct a region-wide test of the null hypothesis

$$H_0 : \beta_p(t) = 0, \text{for any } t \text{ in the genomic interval.}$$

This test depends on the association between covariate Z_p and methylation levels across the region, after adjustment for all the other covariates, and the null hypothesis is equivalent to $H_0 : \boldsymbol{\alpha}_p = \mathbf{0}$. We propose the following region-based F statistic

$$T_p = \frac{\widehat{\boldsymbol{\alpha}_{\boldsymbol{p}}}^T \left\{\widehat{\boldsymbol{V}}_{\boldsymbol{p}}\right\}^{-1} \widehat{\boldsymbol{\alpha}_{\boldsymbol{p}}}}{\tau_p},$$

where $\{\widehat{\boldsymbol{V}}_{\boldsymbol{p}}\}^{-1}$ denotes inverse if $\widehat{\boldsymbol{V}}_{\boldsymbol{p}}$ is nonsigular; for singular $\widehat{\boldsymbol{V}}_p$, the inverse is replaced by the Moore-Penrose inverse $\{\widehat{\boldsymbol{V}}_{\boldsymbol{p}}\}^{-}$. Here, τ_p is the effective degrees of freedom (EDF) for smooth term $\beta_p(t)$, which depends on the magnitude of smoothing parameter $\boldsymbol{\lambda}$ and random effect variances σ_0^2. Motivated by the work of Wood (2013b), we define the EDF τ_p as

$$\tau_p = \sum_{l=a_p}^{b_p} \left(2\boldsymbol{F} - \boldsymbol{F}\boldsymbol{F}\right)_{(l,l)}, \text{ for } p = 0, 1, \ldots P,$$

where $a_p = \sum_{m=0}^{p-1} L_m + 1$ if $p > 0$ and $a_p = 1$ if $p = 0$, $b_p = \sum_{m=0}^{p} L_m$ for any p, and $(\bullet)_{(l,l)}$ stands for the l^{th} leading diagonal element of a matrix. $\boldsymbol{F}$ is the smoothing matrix of our model, as defined in (4.14), which can be viewed as the matrix mapping the pseudo data to its predicted mean.

Let $\boldsymbol{V}_p = \widehat{\boldsymbol{V}}_p \cdot \phi/\widehat{\phi}$ be the variance estimator for $\boldsymbol{\alpha}_p$ when the dispersion parameter ϕ is known. Zhao et al. (2020) have shown the following asymptotic results under the null

$$\widehat{\boldsymbol{\alpha}_{\boldsymbol{p}}}^T \{\boldsymbol{V}_{\boldsymbol{p}}\}^{-1} \widehat{\boldsymbol{\alpha}_{\boldsymbol{p}}} \sim \chi^2_{\tau_p}.$$

Combining with the property of moment-based dispersion estimator in (4.15), we can conclude that, under the null hypothesis, T_p asymptotically follows a F distribution with degrees of freedom τ_p and $M - \tau$, i.e. $T_p \sim F_{\tau_p, M-\tau}$.

4.4 Illustration of performance of dSOMNiBUS in the ACPA dataset

We first apply our approach to targeted bisulfite sequencing data from a rheumatoid arthritis study (Shao et al., 2019). Participants were sampled from the CARTaGENE cohort (`https://www.cartagene.qc.ca/`), a population-based cohort including 43,000 general population subjects aged 40 to 69 years in Quebec, Canada. The study aims to investigate association between DNA methylation and the levels of anti-citrullinated protein antibodies (ACPA), a marker of rheumatoid arthritis (RA) risk that often presents prior to any clinical manifestations (Forslind et al., 2004).

Firstly, the serum ACPA levels were measured for a randomly sampled 3600 individuals from the CARTaGENE cohort, based upon which individuals were classified as either ACPA positive or ACPA negative. Then, the whole blood samples of the ACPA positive individuals,

and a selected subset of age-sex-and-smoking-status-matched ACPA negative individuals were sent for Targeted Custom Capture Bisulfite Sequencing. Specifically, the sequencing used blood cell-specific immune panels that cover the majority of human gene promoters, active regulatory regions observed in blood, blood-cell-lineage-specific enhancer regions and CpGs from Illumina Human Methylation 450 Bead Chips. Cell type proportions in the blood samples were also measured at the time of the sampling (Shao et al., 2019).

Using this sampling approach, two batches of data, referred to as data 1 and data 2, were collected in 2017 and 2019, respectively. Notably, the classification criteria for ACPA status are slightly different between data 1 and 2. When sampling data 1, subjects with serum ACPA levels greater than 20 optical density (OD) units were called as ACPA postive and samples with ACPA levels less than 20 OD were defined as ACPA negative. After data cleaning, data 1 consisted of 69 ACPA positive subjects and 68 ACPA negative subjects. In contrast, the sampling of data 2 was based on more extreme cutoffs for ACPA levels, and resulted in 60 ACPA positive subjects (ACPA levels $\geq$ 60 OD) and 60 ACPA negative subjects (ACPA levels $<$ 20 OD). This change in decision is reflected in the different distributions of serum ACPA levels between data 1 and 2, as shown in Supporting Information Figure B.1. Average sequence read depths in targeted regions were 5 and 35 in data 1 and 2, respectively (Supporting Information Figure B.2), due to improvement in the sequencing protocols implemented between the two experiments.

In this article, we restricted our attention to regions with at least 50 CpG sites. In addition, we excluded regions with more than 95% CpGs having median read depth 0 or having median methylation proportion as 0. Overall, we analyzed 10,759 regions in dataset 1 and 12,983 regions in dataset 2. We excluded the samples who reported a diagnosis of RA before the CARTaGENE study started. Subjects with missing information on cell type proportions were also removed from our analysis. Supplementary Table S1 presents the sample characteristics in data 1 and 2.

We apply our approach to both data 1 and 2, with the aim to identify the differentially methylated regions that show association with ACPA, after adjustment for age, sex, smoking status and cell type composition. Specifically, we assumed no data errors in the datasets ($p_0 = 1 - p_1 = 0$). We used natural cubic splines to expand the smooth terms in the model, and its rank L_p was approximate as the number of CpGs in a region divided by 10 for $\beta_0(t)$, and divided by 20 for $\beta_p(t), p \geq 1$. Since we place the knots at the empirical quantiles of t_{ij}, this choice of L_p guarantees that approximately 20 CpGs are available for interpolation on each interval between two consecutive knots of $\beta_p(t), p \geq 1$. The intercept $\beta_0(t)$ generally has a more flexible shape than the covariate effects $\beta_p(t), p \geq 1$, and is therefore assigned a larger rank L_0.

4.4.1 Both additive and multiplicative dispersion is present in the data

Figure 4.3 presents the distribution of estimated multiplicative dispersion ϕ and additive dispersion σ_0^2 for all test regions in dataset 1 and 2. Overall, widespread overdispersion is observed; 98.5% regions show multiplicative dispersion ϕ greater than 1 and 51.2% regions show additive dispersion σ_0^2 greater than 0.05. The Pearson correlation coefficient between the estimated ϕ and σ_0^2 is -0.015. There exist 49.8% regions with both multiplicative dispersion $\phi > 1$ and additive dispersion $\sigma_0^2 > 0.05$.

4.4.2 Ignoring either type of dispersion leads to inflated type I errors

Figure 4.4 shows quantile-quantile (QQ) plots for the regional p-values for the effect of ACPA on the 292 regions of Chromosome 18 in the two datasets. Detailed inference steps are given in Section 4.3. The results are compared among four different approaches: (1) dSOM-

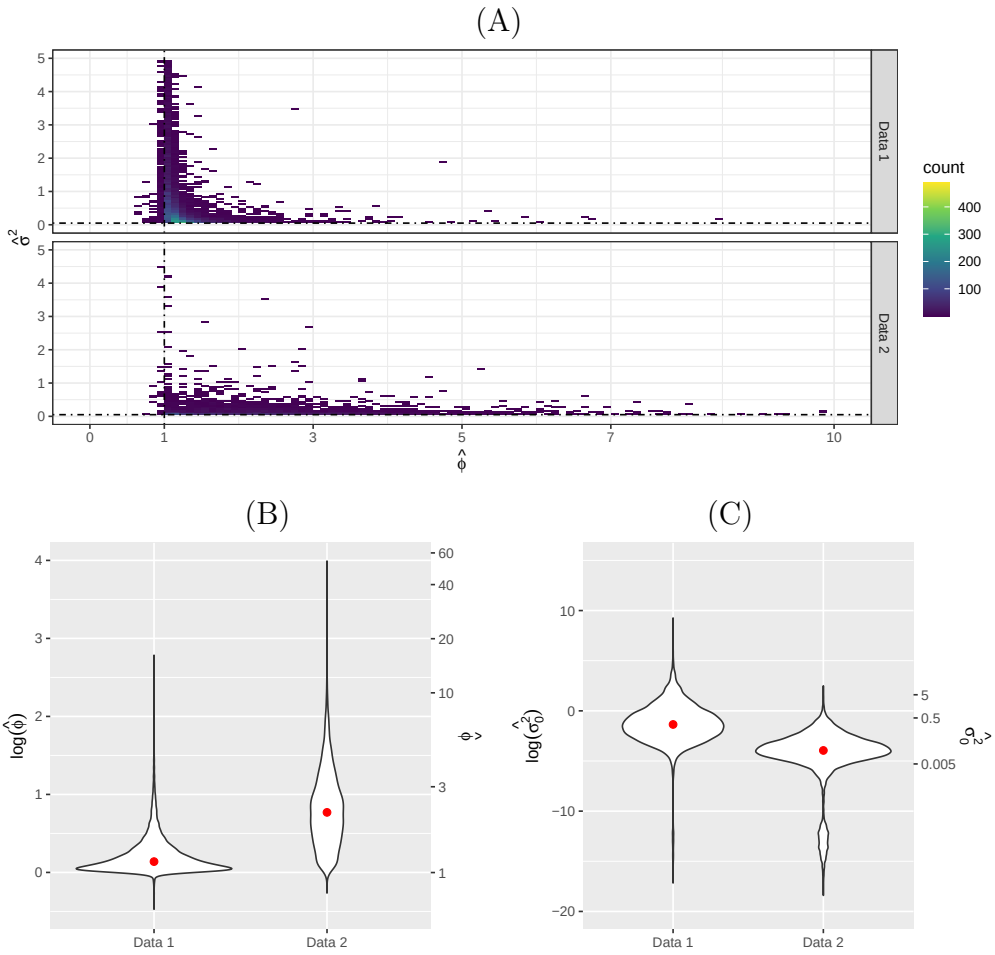


Figure 4.3: Distribution of the estimated multiplicative dispersion parameter ϕ and additive dispersion parameter σ_0^2, for all test regions in dataset 1 and 2. Panel (A) shows the 2-dimensional histogram for $\widehat{\phi}$ and $\widehat{\sigma}_0^2$, where the color intensity represents the number of regions with a particular combination of values of $\widehat{\phi}$ and $\widehat{\sigma}_0^2$. Panels (B) and (C) show the rotated kernel density plots (i.e. violin plots) for $\widehat{\phi}$ and $\widehat{\sigma}_0^2$ (in a natural logarithmic scale), separately.

NiBUS which models both the multiplicative and additive dispersion, (2) the multiplicative-dispersion-only model, (3) the additive-dispersion-only model, and (4) the standard SOM-NiBUS which ignores any extra-binomial variation. Figure 4.4 reveals that, when ignoring either type of dispersion, the distribution of regional p-values is biased away from what would be expected under the null. The inclusion of both multiplicative and additive dispersion is important for correct type I error control.

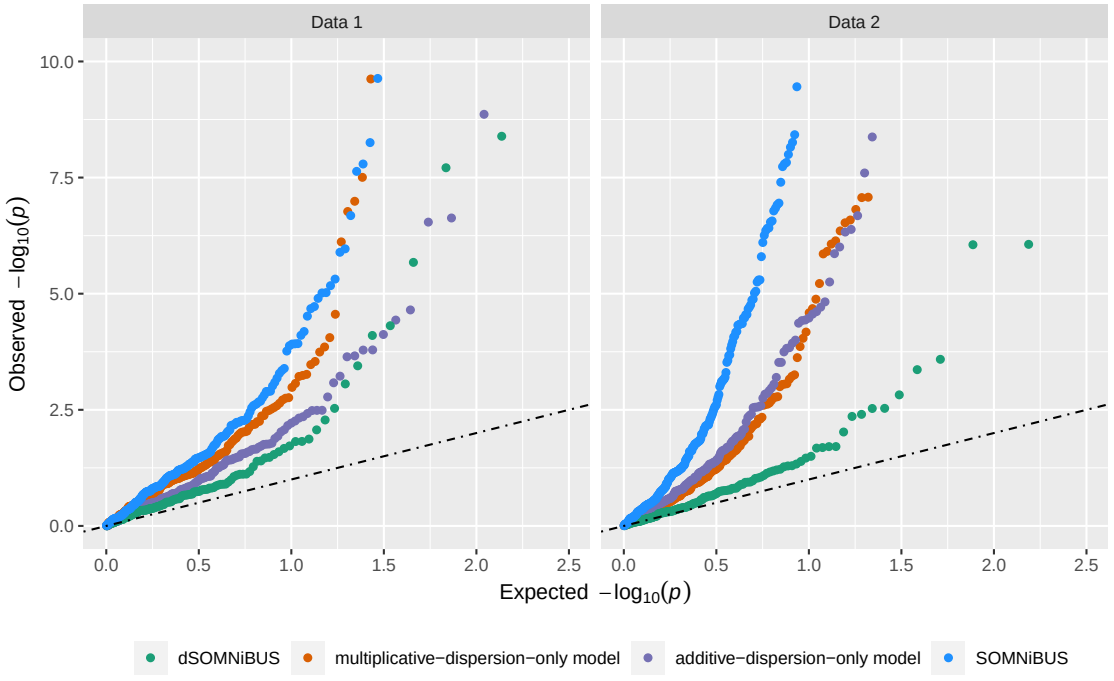


Figure 4.4: QQ plot for regional p-values, obtained from models addressing different types of dispersion.

4.4.3 Our inference procedure provides well-calibrated p-values

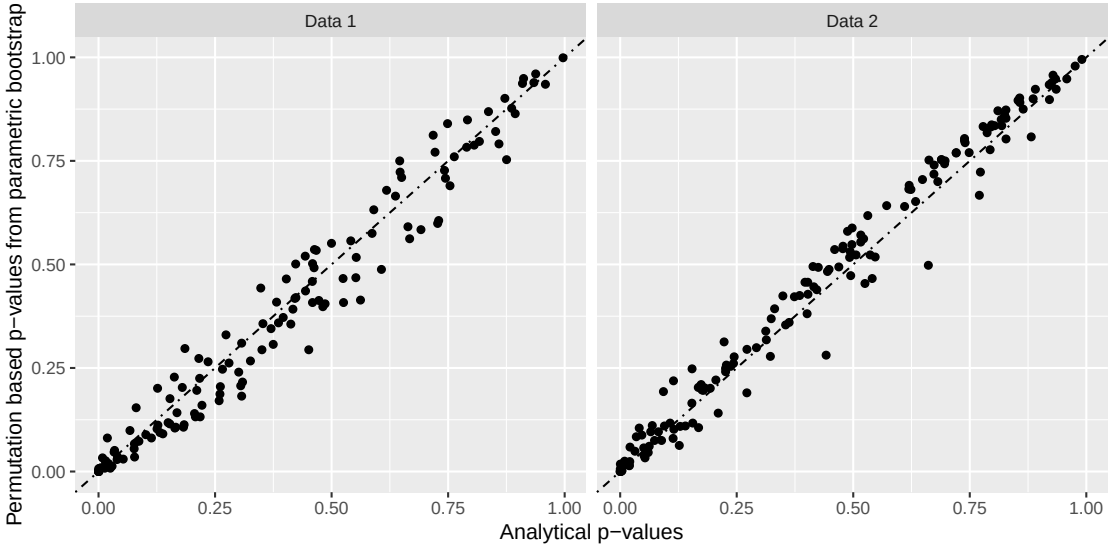


Figure 4.5: Comparison between the observed regional p values from our approach and the permulation-based p values from parametric bootstrap.

To test DMRs, we propose a region-based statistic with a F limiting distribution; see details

in Section 4.3.4. To test the validity of our inference, we compare our regional p-values to bootstrap-based p-values, whose null distribution is constructed by parametric bootstraps (Davison & Hinkley, 1997) and does not rely on any distributional assumptions. Figure 4.5 shows the distributions of bootstrap-based and our analytical p-values for the targeted regions on chromosome 18, demonstrating that our inference method generates p-values in line with the bootstrap-based results. Thus, dSOMNiBUS provides accurate tests for DMRs without requiring extensive computational time.

4.5 Simulation study

We conducted simulations to assess the proposed inference of smooth covariate effects, and to compare the performance of our method with five existing methods: `BiSeq` (Hebestreit et al., 2013), `BSmooth` (Hansen et al., 2012), `SMSC` (Lakhal-Chaieb et al., 2017), `dmrseq` (Korthauer et al., 2018) and `GlobalTest` (Goeman et al., 2006), in terms of type I error and power. Detailed descriptions of these five methods are given in Supplementary Section 3.2. We also made special modifications for the implementations of `BSmooth`, `SMSC` and `dmrseq`, which are primarily designed for WGBS data, to make them as appropriate as possible for targeted regions. see details in Supporting Information Section B.2.1.

4.5.1 Simulation design

We adopt similar simulation parameters as described in Zhao et al. (2020), and simulated methylation regions with 123 CpG sites under various settings. We first generated the vector of read depth for each sample, $(X_{i1}, \ldots, X_{i123})$, by adding 123 independent Bernoulli random variables (with proportion 0.5) to a pre-specified regional read-depth pattern (Supporting Information Figure B.3). In this way, the spatial correlation of read depth observed in real data was well preserved in the simulated data. The rest of simulation parameters were

defined in Table 4.2.

Table 4.2: Simulation settings for the functional parameters $\beta_p(t)$, sample size N, error parameters p_0 and p_1, multiplicative parameter ϕ and RE variances σ_0^2.

Simulation parameters	Possible values
$\beta_p(t)$	Scenario 1: three covariates: $Z_1 \sim Bernoulli(0.51)$, $Z_2 \sim Bernoulli(0.58)$ and $Z_3 \sim Bernoulli(0.5)$ with effects $\beta_1(t), \beta_2(t)$ and $\beta_3(t)$ and intercept $\beta_0(t)$, shown in the red curves in Figure 4.7. Here, Z_3 is the null covariate with effect $\beta_3(t) \equiv 0$. Scenario 2: one covariate: $Z \sim Bernoulli(0.5)$ with 15 different settings of $(\beta_0(t), \beta_1(t))$, which yield methylation proportion parameters as depicted in Figure 4.6.
N	100
(p_0, p_1)	$(0.003, 0.9)^\dagger$ or $(0, 1)$
ϕ	$(1, 3)$
σ_0^2	$(0, 1, 3, 9)$, and the corresponding subject-specific RE $u_i \overset{i.i.d}{\sim} N(0, \sigma_0^2)$ for $i = 1, 2, \ldots N$

† the value 0.003 was reported by Prochenka et al. (2015) as insufficient Bisulfite conversion rate and 0.1 was estimated as the average excessive conversion rate from a (single-cell-type) bisulfite dataset in Hudson et al. (2017) using the method SMSC (Lakhal-Chaieb et al., 2017).

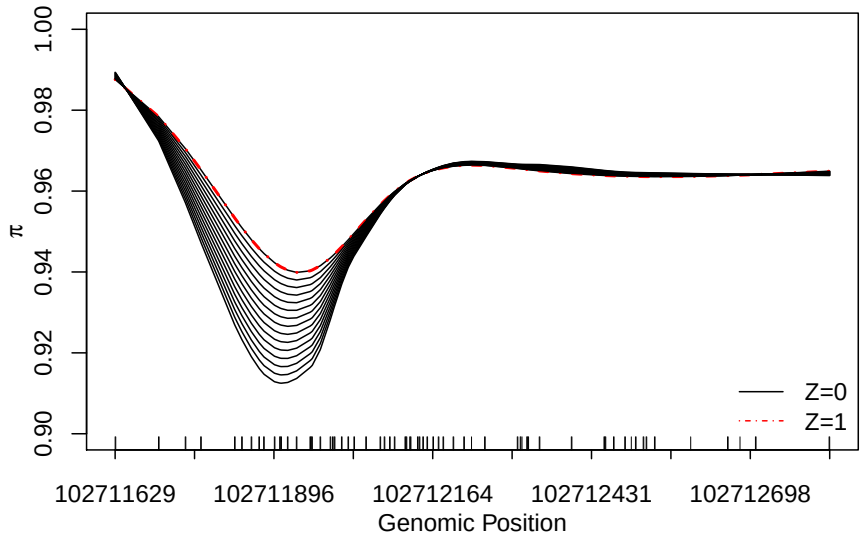


Figure 4.6: The 15 simulation settings of methylation parameters $\pi_0(t)$ and $\pi_1(t)$ in Scenario 2. Here, $\pi_0(t)$ and $\pi_1(t)$ denote the methylation parameters for samples with $Z = 0$ and $Z = 1$ at position t, respectively. Under this scenario, $\pi_1(t)$ (red dotted-dashed curve) is fixed across settings, whereas $\pi_0(t)$s (black solid lines) vary across settings corresponding to different degrees of closeness between methylation patterns in the two groups.

Simulate dispersed-binomial counts. Given the values of $\{Z_1, \ldots Z_P\}$, $\{\beta_p(t), p = 0, 1, \ldots P\}$ and $\{u_i, i = 1, 2, \ldots N\}$ under each setting, the individual's methylation pro-

portion, π_{ij}, can be readily calculated from the mean model in (4.2). We then generated the true methylation counts S_{ij} from a beta-binomial distribution with proportion parameter $\mu = \pi_{ij}$, correlation parameter $\rho = \dfrac{\phi - 1}{X_{ij} - 1}$, and size parameter $n = X_{ij}$. Specifically, S_{ij} were drawn from the following probability mass function

$$P(S_{ij} = k \mid \mu, \rho, n) = \binom{n}{k} \frac{B(k + \alpha, n - k + \beta)}{B(\alpha, \beta)}$$

where $\alpha = \mu(1 - \rho)/\rho, \beta = (1 - \mu)(1 - \rho)(1 - \mu)/\rho$, and $B(\cdot, \cdot)$ is the beta function. The variance of S_{ij} can be thus derived as

$$\mathbb{V}\text{ar}(S_{ij}) = [1 + (n - 1)\rho]\,[n\mu(1 - \mu)] = \phi X_{ij}\pi_{ij}(1 - \pi_{ij}),$$

which coincides with our assumed mean-variance relationship in (4.3). We then generated the observed methylated counts Y_{ij} according to the error model in (4.1), which implies

$$Y_{ij} \mid S_{ij} \sim \text{Binomial}(S_{ij}, p_1) + \text{Binomial}(X_{ij} - S_{ij}, p_0).$$

Under each scenario and setting, we generated data sets with sample sizes $N = 100$, each 1000 times. We then applied `dSOMNiBUS` along with methods `BiSeq`, `dmrseq`, `BSmooth`, `SMSC` and `GlobalTest` to the simulated data sets. For our approach `dSOMNiBUS`, we used cubic splines with dimension $L_p = 5$ to parameterize the smooth terms of interest. We also assumed that the correct values of error parameters p_0 and p_1 were known.

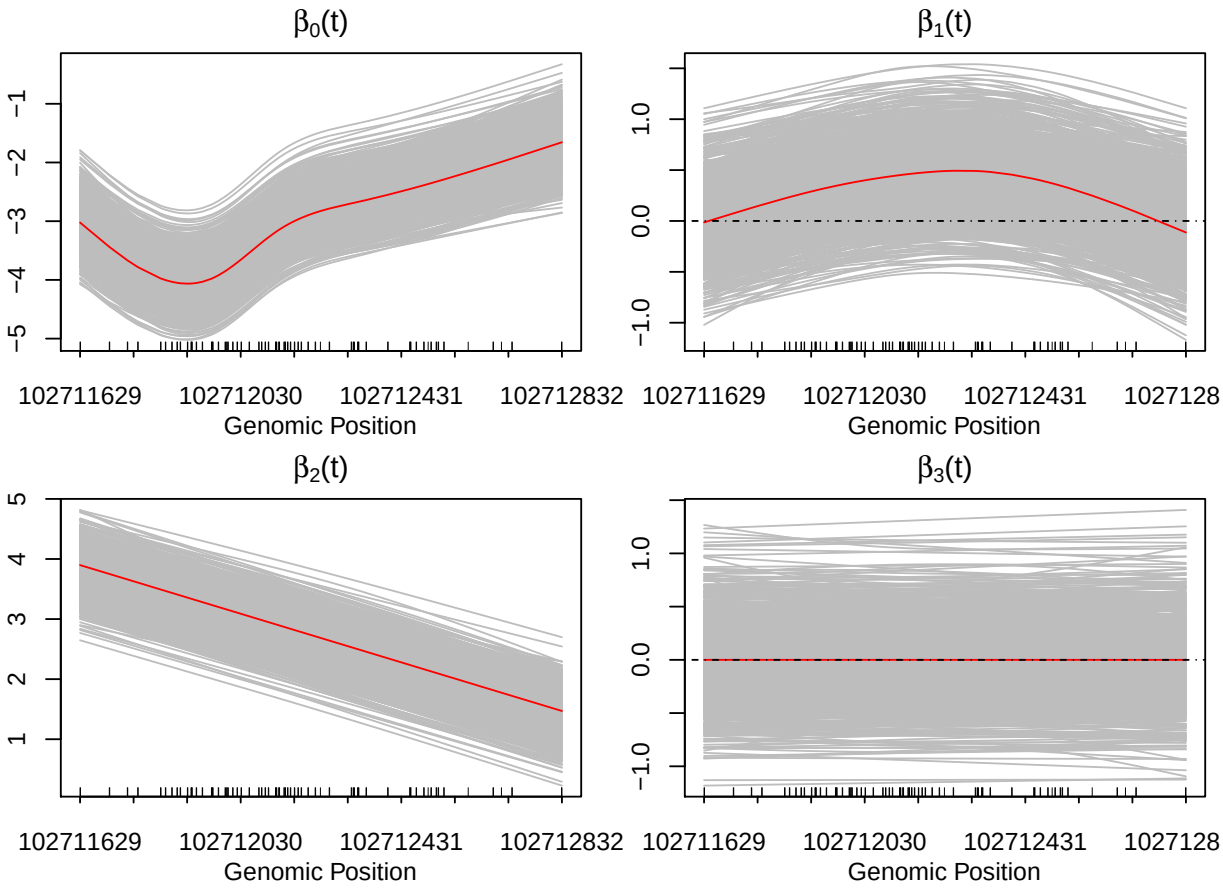


Figure 4.7: Estimates of smooth covariate effects (gray) over the 1000 simulations in Scenario 1, using dSOMNiBUS. The red curves are the true functional parameters used to generate the data. Data were generated with error using $\phi = 3$ and $\sigma_0^2 = 3$.

4.5.2 Simulation results

dSOMNiBUS provides accurate inference for smooth covariate effects

Figure 4.7 presents the estimates of the functional parameters $\beta_0(t), \beta_1(t), \beta_2(t)$ and $\beta_3(t)$ over 1000 simulations, obtained from dSOMNiBUS; here, data were generated under Scenario 1, with multiplicative dispersion parameter $\phi = 3$, RE variance $\sigma_0^2 = 3$, and error parameters $p_0 = 0.003$ and $1 - p_1 = 0.1$. Figure 4.7 demonstrates that the proposed method provides unbiased curve estimates for smooth covariate effects when the regional methylation counts exhibit extra-parametric variation and are measured with errors.

Figure 4.8 and 4.9 demonstrate the performance of the proposed pointwise confidence in-

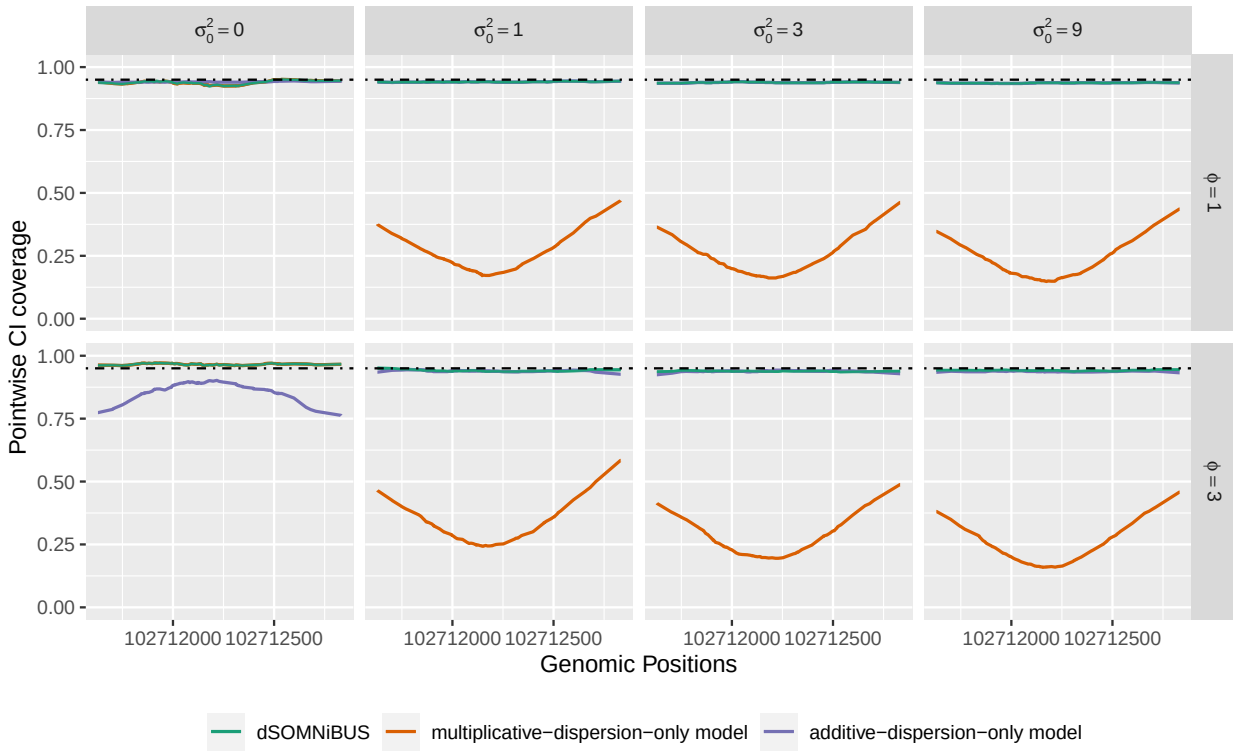


Figure 4.8: Empirical coverage probability of the analytical 95% CIs for $\beta_3(t)$ over 1000 simulations, under different vales of ϕ and σ_0^2. The empirical coverage probabilities are defined as the percentage of simulations where the analytical CIs cover the true value of $\beta_3(t)$. Data were generated with error, under simulation Scenario 1. The results from dSOMNiBUS (green) and the additive-dispersion-only model (purple) are indistinguishable in all settings but $\sigma_0^2 = 0$ and $\phi = 3$ and dSOMNiBUS (green) and the multiplicative-dispersion-only model (orange) are indistinguishable when $\sigma_0^2 = 0$.

terval (CI) estimates (Section 4.3.4) and regional test (Section 4.3.4), respectively. The results from dSOMNiBUS are compared to the multiplicative-dispersion-only model and the additive-dispersion-only model. Figure 4.8 displays the empirical coverage probabilities of the analytical 95% CIs for $\beta_3(t)$, under different settings of ϕ and σ_0^2. Figure 4.9 shows the QQ plots for the regional p-values when the null hypothesis $H_0 : \beta_3(t) = 0$ is correct. The results show that ignoring the presence of additive dispersion (i.e. the multiplicative-dispersion-only model) leads to substantial estimation bias, poor CI coverage probabilities and highly inflated type I errors. Although the additive-dispersion-only model provides relatively accurate pointwise CIs, the distributions of its regional p-values are biased away from

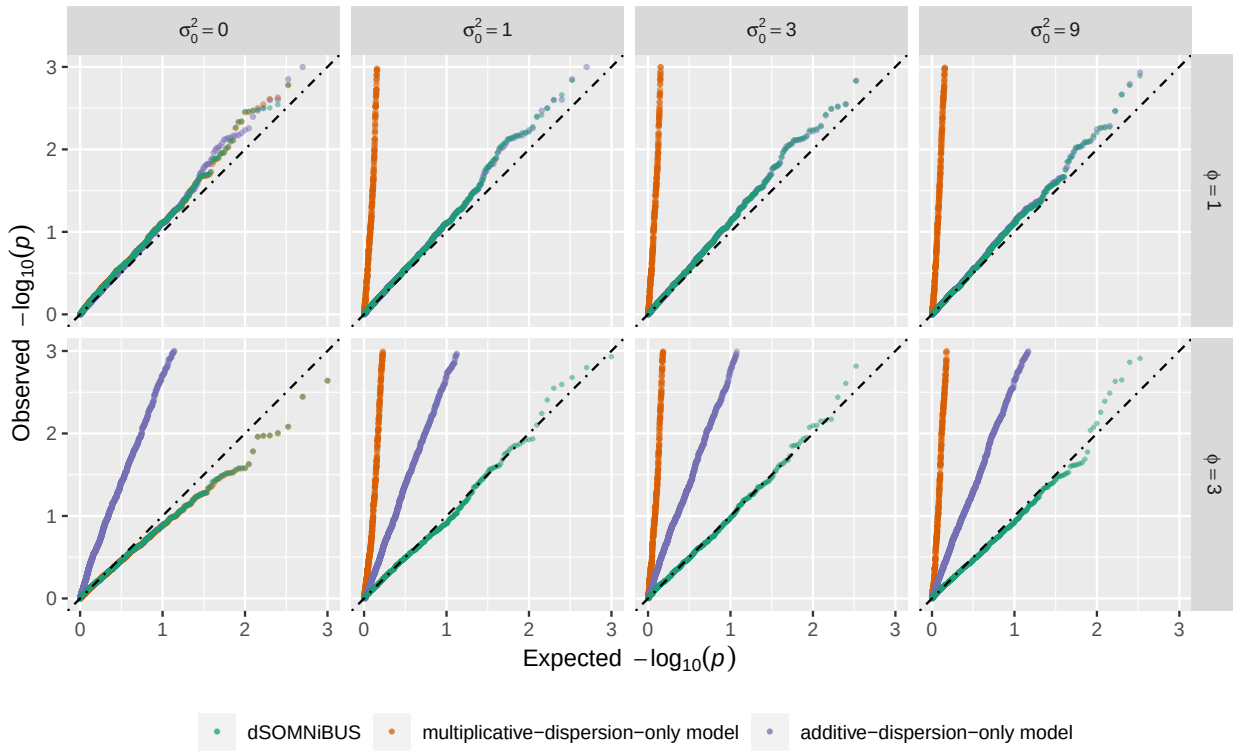


Figure 4.9: QQ plot for regional p-values for the test $H_0 : \beta_3(t) = 0$, obtained from dSOMNiBUS, the multiplicative-dispersion-only model and the additive-dispersion-only model. Data were simulated with error, under simulation Scenario 1. When $\phi = 1$, the results from dSOMNiBUS (green) and the additive-dispersion-only model (purple) are indistinguishable. When $\sigma_0^2 = 0$, the lines for the multiplicative-dispersion-only model (orange) and dSOMNiBUS (green) are indistinguishable.

what would be expected under the null, when multiplicative dispersion $\phi > 1$. Overall, dSOMNiBUS provides pointwise CIs attaining their nominal levels, and region-based statistics whose distribution under the null is well calibrated, regardless of the types and degrees of dispersion that data exhibit. Similar results were observed when data were generated without error (Supplementary Figures S5 and S6).

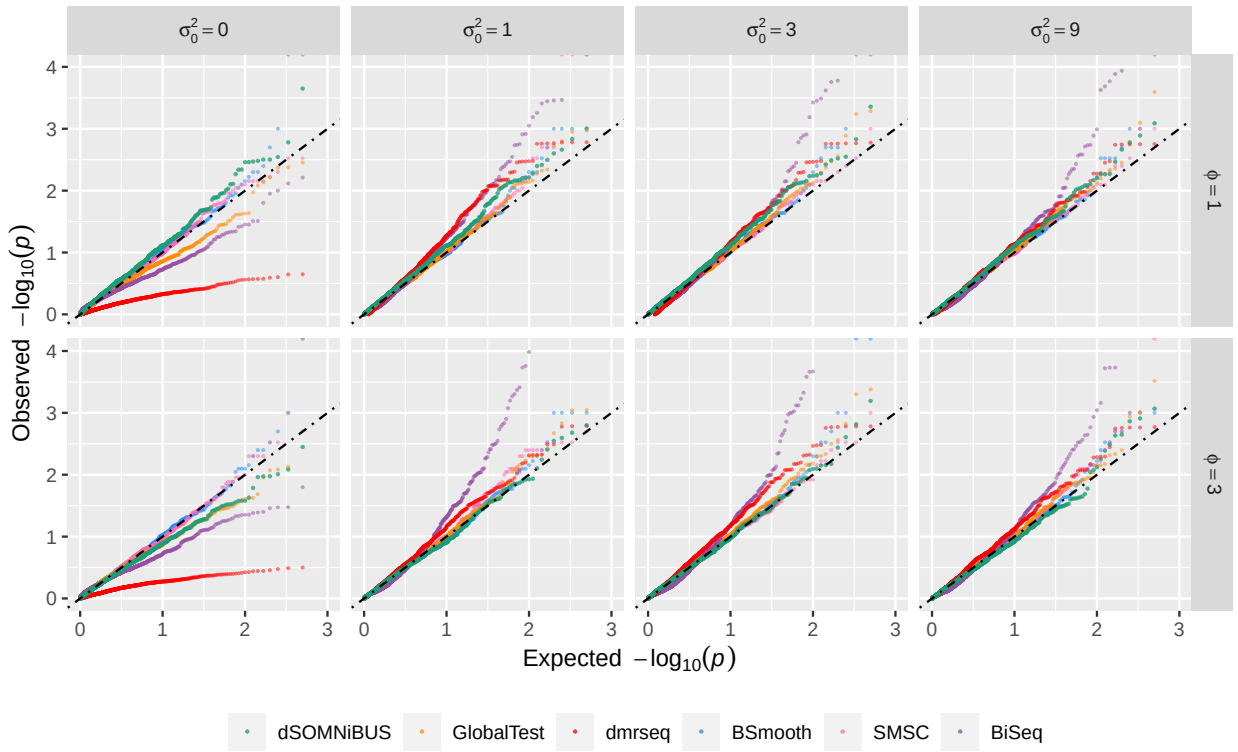


Figure 4.10: QQ plot for regional p-values for the test $H_0 : \beta_3(t) = 0$, obtained from dSOMNiBUS, GlobalTest, dmrseq, BSmooth, SMSC, and BiSeq. Data were simulated with error, under simulation Scenario 1.

dSOMNiBUS exhibits greater power to detect DMRs while correctly controlling type I error rates

Figures 4.10 and 4.11 further demonstrate the performance of the proposed regional test, when compared with the existing methods `GlobalTest`, `dmrseq`, `BSmooth`, `SMSC`, and `BiSeq`. Here, data were simulated with error parameters $p_0 = 0.003$ and $1 - p_1 = 0.1$. Figure 4.10 shows the distributions of p-values for the regional effect of the null covariate Z_3. Because we estimated the empirical regional p-values for `BSmooth` and `SMSC` by permutations, both methods are able to control type I errors, under all settings of ϕ and σ_0^2. Both `BiSeq` and `dmrseq` show deflated type I error rate when $\sigma_0^2 = 0$ and inflated type I error rate when $\sigma_0^2 > 0$. The distributions of p-values from `GlobalTest` are well calibrated when the within subject correlation $\sigma_0^2 > 0$, but are slightly biased away from the uniform distribution when

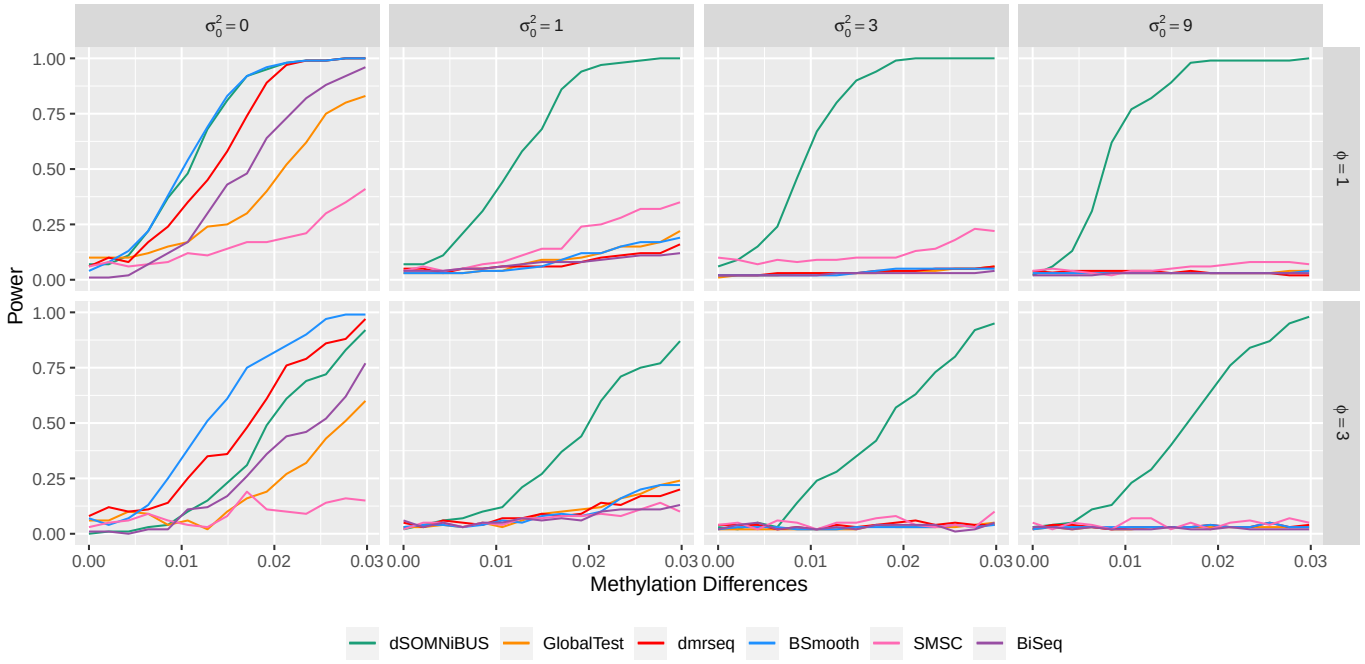


Figure 4.11: Powers to detect DMRs using the six methods for the 15 simulation settings in Scenario 2 under different levels of maximum methylation differences between $\pi_0(t)$ and $\pi_1(t)$ in the region, calculated over 100 simulations.

$\sigma_0^2 = 0$. When $\sigma_0^2 = 0$ and $\phi = 3$, dSOMNiBUS provides slightly conservative type I errors; this bias vanishes when the data were generated without error (Supplementary Figures S7). Figure 4.11 shows the powers of the six methods for detecting DMRs under the 15 settings of methylation patterns displayed in Figure 4.6. Here, methylation difference is defined as the maximum difference between $\pi_1(t)$ and $\pi_0(t)$ in the region. When data exhibit neither additive nor multiplicative dispersion, dSOMNiBUS and `BSmooth` provide the highest power, followed by `dmrseq`, `BiSeq`, `GlobalTest`, and `SMSC`. When $\sigma_0^2 = 0$ and $\phi = 3$, `BSmooth` and `dmrseq` are more powerful than other methods. When there are correlations among methylation measurements on the same subject, i.e. $\sigma_0^2 > 0$, dSOMNiBUS clearly outperforms the five alternative methods; this superiority remains when the data were generated without error (Supplementary Figures S8). In summary, dSOMNiBUS exhibits greater power to detect DMRs, while correctly controlling type I error rates, especially when the regional methylation counts exhibit (additive) extra-binomial variation.

4.6 Discussion

We have proposed and evaluated a novel method, called dSOMNiBUS, for estimating smooth covariate effects for BS-seq data. We demonstrate that our model, which incorporates both multiplicative and additive sources of data dispersion, provides a plausible representation of realistic dispersion trends in regional methylation data. In addition, dSOMNiBUS simultaneously accounts for experimental errors, estimation of multiple covariate effects, and flexible dispersion patterns in a region. Also, we provide a formal inference for smooth covariate effects and construct a region-based statistic for the test of DMRs, where outcomes might be contaminated by errors and/or exhibit extra-parametric variations. Results from simulations and real data applications show that the new method captures important underlying methylation patterns with excellent power, provides accurate estimates of covariate effects, and correctly quantifies the underlying uncertainty in the estimates. The method has been implemented in the R package `SOMNiBUS`, which has been submitted to R Bioconductor.

Our model captures dispersion in the regional count data via the combination of a subject-specific RE and a multiplicative dispersion. The latter aims to capture the extra random dispersion beyond that introduced by the subject-to-subject variation. An alternative way to add multiplicative despersion might be to add locus-specific REs. Such model would avoid the problem of estimating ϕ, but would result in substantially increased number of REs, in which case our Laplace approximation is unlikely to provide well-founded inference (Shun & McCullagh, 1995). In addition, such a model only captures overdispersion. In contrast, our quasi-binomial mixed effect model provides an adequate representation of any kind of dispersion without much increase in computational complexity.

An extension worth exploring in the future is to model the dispersion parameter ϕ as a function of covariates. For example, the methylation variation across cancer samples has been found to be higher than for normal samples (Hansen et al., 2011; Schoofs et al., 2013). Identification of such disease-associated methylation variation changes might provide further

insights into the biological mechanisms. This extension would also allow modelling of the hypothesis that some individuals are more sensitive to their environment (Meaney & Szyf, 2005).

Our proposed methods can also be applied to other types of next-generation sequencing data. For example, allele-specific gene expression (ASE) measured from RNA-seq data are quantified by the numbers of reads originating from the two alleles for that site (J. Fan et al., 2020). Such data share a similar structure to bisulfite sequencing data and could be analyzed by dSOMNiBUS. From the methodology point of view, our proposal of combining quasi-likelihood with random effects can be generally applied to any type of count data for a more comprehensive representation of dispersion.

Chapter 5

Manuscript III: A sparse high-dimensional generalized varying coefficient model for identifying genetic variants associated with regional methylation levels

Preamble to Manuscript III: In Chapter 4, I introduced a flexible quasi-binomial mixed model to account for the excess variation (relative to a binomial model) observed in the methylated counts in a region. This remedy addresses overdispersion by reformulating the stochastic component of the binomial model. In practice, the problem of overdispersion can be also caused by an error in the systematic part of the regression model, such as missing crucial covariates in the conditional mean. If one indeed has the measurements for these crucial covariates, one natural remedy for overdispersion is to include them in the regression model.

Studies have shown that genetic variants or SNPs can massively influence methylation variations (Gaunt et al., 2016; Hannon et al., 2018). Another solution for overdispersion is to include SNPs as covariates in the SOMNiBUS model developed in Chapter 3. However, there are hundreds or thousands of SNPs surrounding or within a methylation region and all of them are candidate contributing factors to methylation. In addition, our sample sizes tend to be small due to the cost of sequencing and the challenges associated with obtaining samples. In such a high-dimensional setting, the statistical methods in Chapter 3 show important limitations (Chouldechova & Hastie, 2015; J. Fan et al., 2014).

Therefore, the goal of the third manuscript in this thesis is to extend the standard SOMNiBUS for high-dimensional settings. The new approach automatically selects important variables among an extensive collection of covariates and can be applied to identifying the subset of genetic variants associated with regional methylation levels. This method has been implemented in a prototype R package `sparseSOMNIBUS` (`https://github.com/kaiqiong/sparseSOMNiBUS`).

Note that the supporting material for this chapter can be found in Appendix C.

A sparse high-dimensional generalized varying coefficient model for identifying genetic variants associated with regional methylation levels

Kaiqiong Zhao[1,2], Yi Yang[3], Karim Oualkacha[4], Celia M.T. Greenwood[1,2,5,6]

[1]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada*

[2]*Lady Davis Institute, Jewish General Hospital, Montréal, Québec, Canada*

[3]*Department of Mathematics and Statistics, Montréal, Québec, Canada*

[4]*Département de Mathématiques, Université du Québec à Montrèal*

[5]*Department of Human Genetics, McGill University*

[6]*Gerald Bronfman Department of Oncology, McGill University*

Abstract

Varying coefficient models offer the flexibility to learn the dynamic changes of regression coefficients. Despite their good interpretability and diverse applications, in high-dimensional settings, existing estimation methods for such models have important limitations. For example, we routinely encounter the need for variable selection when faced with a large collection of covariates with nonlinear/varying effects on outcomes, and no ideal solutions exist. One illustration of this situation could be identifying a subset of genetic variants with local influence on methylation levels in a regulatory region. To address this problem, we propose a composite sparse penalty that encourages both sparsity and smoothness for the varying coefficients. We present an efficient proximal gradient descent algorithm to obtain the penalized estimation of the varying regression coefficients in the model. A comprehensive simulation study has been conducted to evaluate the performance of our approach in terms of estimation, prediction and selection accuracy. We show that the inclusion of smoothness control yields much better results than having the sparsity-regularization only.

5.1 Introduction

DNA methylation is an essential epigenetic modification that regulates gene activity and contributes to tissue differentiation and disease susceptibility. It primarily occurs at a cytosine-guanine dinucleotide (i.e. CpG site) and involves the covalent addition of a methyl group to a cytosine. Notably, DNA methylation variation has a vital genetic component (Gaunt et al., 2016; Hannon et al., 2018). Loci harbouring genetic variants that influence methylation levels are called methylation quantitative trait loci (mQTLs). Identifying mQTLs can provide important insight into the underlying molecular events within multiple human tissues and thus enhance our understanding of the genetic basis of disease development (Taylor et al., 2019). However, mQTLs may also confound the association between methylation levels and phenotype of interest (Hannon et al., 2016; Van Dongen et al., 2016). Therefore, it is essential to identify mQTLs and adjust for their effects when testing the methylation signals. Analyses that identify genetic effects on DNA methylation levels is usually referred to as mQTL mapping.

Recent advances in bisulfite sequencing (BS) technology have enabled high-resolution large-scale measurements of DNA methylation. Such sequencing platforms measure the methylation level at a single site as a pair of counts: the number of methylated reads and the total number of reads aligned to the site, i.e. read depth. Studies performing mQTL mapping using BS data have yielded encouraging results (Banovich et al., 2014; Cheung et al., 2017; Schmitz et al., 2013). However, existing mQTL approaches have identified the genetic loci or single nucleotide polymorphisms (SNPs) associated with each CpG site separately (Y. Fan et al., 2019; Zhou & Stephens, 2014) and ignore the spatial correlation structure of methylation at neighbouring CpG sites. In practice, researchers are often interested in exploring the genetic contribution to localized methylation patterns within a functional genomic region or a regulatory element rather than at individual sites (Gutierrez-Arcelus et al., 2015). We have recently proposed a novel varying coefficient (VC) model, SOMNiBUS (Zhao et

al., 2021, 2020), to analyze regional BS-derived methylation data, enabling comprehensive and simultaneous estimates of covariate effects which are smoothly varying along genomic positions. It adopts the fitting framework for generalized additive models proposed by Wood (2011)—first expanding the varying coefficients using spline-type basis functions and then maximizing the penalized likelihood with quadratic smoothness penalties for the basis coefficients. This penalty function is quantified by the integrated squared second derivatives of the varying coefficients, summed over all covariates. Specially, the smoothness/penalty parameters, which determine the appropriate amount of smoothness of individual varying coefficients (i.e. function complexity), are estimated by restricted maximum likelihood. This method allows us to borrow information from the local correlation structures and offers good interpretability.

A natural solution for regional mQTL analysis is to include SNPs as covariates in such a VC model. However, we routinely face hundreds or thousands of candidate SNPs within or near a regulatory region and sample sizes tend to be small due to the cost of sequencing and the challenges associated with obtaining samples. Many traditional statistical methods, including Wood (2011), face significant challenges when estimating the varying coefficients in such a high-dimensional setting (Chouldechova & Hastie, 2015; J. Fan et al., 2014). In addition, only a small subset of the candidate SNPs is expected to influence the methylation patterns in a region of interest. In contrast, the traditional VC models using quadratic smoothness penalties cannot provide sparse solutions for the varying coefficients, and are thus unsuitable for regional mQTL mapping. In this paper, we propose a novel sparse high-dimensional varying coefficient model, which automatically selects important variables among an extensive collection of covariates with varying/nonlinear coefficients and can therefore be seamlessly applied to regional mQTL mapping.

There has been extensive literature on using sparse penalized regression methods to enable variable selection in high-dimensional VC models. The proposed methods mainly differ in

their choices of penalty functions. Major classes of the method are based upon LASSO (Lin & Zhang, 2006), group LASSO (Barber et al., 2017; Gertheiss et al., 2013; Huang et al., 2010; Meier et al., 2009; Ravikumar et al., 2009; H. Wang & Xia, 2009; Wei et al., 2011), group smoothly clipped absolute deviation penalty (SCAD) (Noh & Park, 2010; L. Wang et al., 2007, 2008) and L0-penalization (Xue & Qu, 2012). Notably, on top of the regularization imposed by these sparse penalty functions, estimations of nonparametric models (e.g. VC models) inevitably require regularization for function complexity, i.e. the smoothness of the nonparametric component. The smoothness regularization is even crucial if we use an unnecessarily large number of basis functions to expand the functional coefficients. However, most of the existing sparse nonparametric regression methods fail to address the smoothness regularization adequately. For example, Huang et al. (2010); Ravikumar et al. (2009); Xue & Qu (2012) control the smoothness by employing fixed numbers of truncation dimensions. But it is not always feasible to perceive the appropriate complexity levels for the functional parameters of interest, making their methods less straightforward to use in practical applications. For more refined control of smoothness, one can start with a comparatively large number of basis functions and then impose the quadratic smoothness penalty in the estimation, as in penalized regression splines or smoothing splines. In this way, the exact value of the basis dimension, which sets an upper limit on the function complexity, becomes less critical for the final fitted model, although this strategy alone does not meet our objective for sparsity.

Additionally, some methods disentangle the two regularization tasks—sparsity and smoothness. For example, L. Wang et al. (2008) treat the number of basis functions as the tuning parameter for smoothness, and separately use the usual shrinkage parameter to control sparsity. They propose to tune both regularization parameters using the generalized cross-validation criterion. H. Wang & Xia (2009) first select the optimal bandwidth (i.e. the smoothness parameter for the local polynomial nonparametric fitting method) using cross-validation assuming no sparsity penalization is present, and then separately tune the shrinkage param-

eter under the selected bandwidth. However, it is desirable to have a unified penalization method that simultaneously controls the overall sparsity of the model and the smoothness of the nonzero functional coefficients.

In this paper, we propose such a unified framework for estimating high-dimensional generalized varying coefficient models. Our strategy combines the appealing features of smoothing splines and sparse penalties, thus providing sparse varying coefficient estimates that are less dependent on basis dimensions.

Specifically, we propose a sparse high-dimensional binomial varying coefficient model for regional mQTL mapping. Here, we model the regional methylation counts by a binomial distribution, dependent on read depth. Effects of each candidate SNP are modelled as functional coefficients varying along genomic positions. To encourage both sparsity and smoothness for the varying coefficients, we propose a composite sparse penalty, which is inspired from the penalty function developed in high-dimensional additive models (Meier et al., 2009). This penalty function incorporates two tuning parameters, separately controlling the overall model complexity and smoothness of the nonzero functional coefficients. We then develop an efficient proximal gradient descent algorithm to obtain the penalized estimation of the varying regression coefficients, where the tuning parameters are chosen via cross-validation. Our unified estimating procedure can simultaneously select important mQTLs and estimate their corresponding varying effects across a methylation region of interest. An R package called `sparseSOMNiBUS` that implements our method is freely available on GitHub and it provides a routine to fit high-dimensional varying coefficient models for non-binary binomial outcomes.

The remainder of the article is organized as follows. We describe the data and present the proposed sparse high-dimensional binomial varying coefficient model in Section 5.2. In Section 5.3, we provide a detailed description of our estimating algorithm. Section 5.4 contains an adaptive penalized estimation method for our model. Simulation experiments evaluating

the performance of our method are summarized in Section 5.5. The paper concludes with a discussion in Section 5.6.

5.2 High-dimensional binomial varying coefficient models

5.2.1 Notation and data

We consider DNA methylation measures over a genomic region from N independent samples. Let m_i be the number of CpG sites for the i-th sample, $i = 1, 2, \dots N$. Let t_{ij} be the genomic position (in base pairs) for the i-th sample at the j-th CpG site, $j = 1, 2, \dots, m_i$. Methylation levels at a site are quantified by the number of methylated reads and the total number of reads. We define X_{ij} as the total number of reads aligned to CpG j from sample i and S_{ij} as the methylated counts at CpG j for sample i. Furthermore, we assume that we have the genotype information on P candidate SNPs for the N samples, denoted as $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots Z_{iP}) \in \mathbb{R}^P$, for $i = 1, 2, \dots N$. For mQTL analysis, one typical choice of the candidate SNPs would be loci located within and 20kb up- and downstream of the methylation region. Hence, the number of candidate SNPs is usually greater than sample size N.

5.2.2 Model

We assume the methylated counts S_{ij} follows a binomial distribution with a methylation proportion parameter π_{ij} that depends on the genetic variants $\boldsymbol{Z}_i$, and nearby methylation

patterns. Specifically,

$$\begin{aligned} S_{ij} \mid \mathbf{Z}_i, X_{ij} &\sim \text{Binomial}(X_{ij}, \pi_{ij}), \\ g(\pi_{ij}) &= \beta_0(t_{ij}) + \sum_{p=1}^{P} \beta_p(t_{ij}) Z_{ip}, \end{aligned} \tag{5.1}$$

where $g(x) = \log(x/(1-x))$ is a logit link function, $\pi_{ij} = \mathbb{E}(S_{ij})/X_{ij}$ is the methylation proportion for CpG j from sample i, and $\beta_0(t_{ij})$ is the intercept term . Here $\{\beta_p(t_{ij})\}_{p=1}^{P}$ are functional parameters for the genetic effects. This amounts to assuming smoothly varying methylation levels and genetic effects on methylation levels across our targeted small genomic regions. Sensitivity analysis to explore implications of such a smoothness assumption is conducted and discussed later in Section 5.5.

We express each function coefficient $\beta_p(t_{ij})$ in terms of natural cubic spline functions. Without loss of generality, we use the same expansion dimension K for all the functional coefficients in (5.1), i.e.

$$\beta_p(t) = \boldsymbol{\theta}_p^T \mathbf{B}(t) = \sum_{k=1}^{K} \theta_{p,k} b_k(t), \text{ for } p = 0, \dots P,$$

where $\mathbf{B}(t) = (b_1(t), \dots, b_K(t))^T$ consists of K natural cubic basis functions $b_k(t) : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{\theta}_p = (\theta_{p,1}, \dots \theta_{p,K})^T$ is a vector of coefficients with $\theta_{p,k}$ being the coefficient for the k-th basis of the p-th covariate. Specific expressions for basis functions $b_k(t)$ can be found in equation (C.5). We use a comparatively large number of basis functions for the expansion to capture varying coefficients with possibly high complexity.

Let $\boldsymbol{\theta}$ be the parameter vector to be estimated and specifically it is the vectorization of $(P+1) \times K$-dimensional coefficient matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_P)^T$ by row, i.e. $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\Theta})$. We can then write the mean function in (5.1) in a compact way,

$$g^{-1}(\boldsymbol{\pi}) = \mathbb{X}\boldsymbol{\theta},$$

where $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1m_1}, \pi_{21}, \ldots, \pi_{2m_2}, \ldots, \pi_{Nm_N})^T \in [0,1]^M$ with $M = \sum_{i=1}^N m_i$. $\mathbb{X} = [\mathbb{X}_0 \mid \mathbb{X}_1 \mid \ldots \mid \mathbb{X}_P]$ is the spanned design matrix of dimension $M \times (P+1)K$, where $\mathbb{X}_p$ is the $M \times K$ design matrix for the pth covariate, whose kth column is stacked with elements $b_k(t_{ij}) \times Z_{pi}$ where $Z_{0i} \equiv 1$.

5.2.3 The sparsity-smoothness penalty

In practice, only a small subset of the candidate SNPs is expected to influence the methylation patterns in the test region. It is therefore desirable to produce functional estimators that are sparse, i.e. $\widehat{\beta}_p(t) = 0$ for some $p \in \{1, \ldots, P\}$. At the same time, we would like to avoid too rough estimators for those nonzero $\beta_p(t)$ that can arise from a large K. To this end, we consider a composite sparse penalty that simultaneously controls the sparsity and smoothness for the varying coefficients. This proposed penalty is inspired from the so-called sparsity-smoothness penalty (SSP) which was first introduced by Meier et al. (2009) for variable selection in high-dimensional additive models. Specifically, we define the penalty function as

$$\mathcal{L}^{\text{SSP}}(\boldsymbol{\theta}) = \lambda \sum_{p=1}^{P} \sqrt{(1-\alpha) J_1\left(\beta_p(t)\right) + \alpha J_2\left(\beta_p(t)\right)}, \tag{5.2}$$

where

$$J_1(\beta_p(t)) = \|\beta_p(t)\|_2^2 = \int \left(\beta_p(t)\right)^2 dt$$

quantifies the L2-norm of the functional coefficients $\beta_p(t)$, and

$$J_2(\beta_p(t)) = M^2 \int \left(\beta_p''(t)\right)^2 dt$$

controls the smoothness of $\beta_p(t)$. The squared root over both J_1 and J_2 enables the sparsity of $\beta_p(t)$ at the function level. Notably the definition in (5.2) is slightly different from the

original proposal in Meier et al. (2009) and they used an empirical L2-norm of $\beta_p(t)$, i.e.

$$J_1(\beta_p(t))^{\text{Meier}} = \frac{1}{M}\boldsymbol{\theta}_p^T \boldsymbol{B}^T \boldsymbol{B}\boldsymbol{\theta}_p,$$

where $\boldsymbol{B} = (\boldsymbol{B}(t_{11}), \ldots, \boldsymbol{B}(t_{1m_i}), \boldsymbol{B}(t_{21}), \ldots, \boldsymbol{B}(t_{2m_i}), \ldots, \boldsymbol{B}(t_{Nm_N}))^T$ is the basis expansion matrix of dimension $M \times K$.

The amount of penalization in (5.2) is jointly controlled by two tuning parameters, $\lambda \geq 0$ and $0 \leq \alpha < 1$. Specifically, λ controls the overall model complexity, and α separately controls the smoothness of the functional estimators. When $\alpha = 0$, no smoothness constraint would be imposed, and as α approaches to 1, smoother estimates would be favoured. Here, we add the scaling constant M^2 in $J_2(\beta_p(t))$ merely for convenience when specifying candidate values for α in cross-validation.

Plugging in the basis expansion for $\beta_p(t)$, we can equivalently write

$$J_1\left(\beta_p(t)\right) = \boldsymbol{\theta}_p^T \boldsymbol{\Omega}^{(1)} \boldsymbol{\theta}_p \text{ and } J_2\left(\beta_p(t)\right) = \boldsymbol{\theta}_p^T \boldsymbol{\Omega}^{(2)} \boldsymbol{\theta}_p,$$

where $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$ are two $K \times K$ matrices with the (k, k')-th element $[\boldsymbol{\Omega}^{(1)}]_{k,k'} = \int b_k(t) b_{k'}(t) dt$, and $[\boldsymbol{\Omega}^{(2)}]_{k,k'} = M^2 \int b_k''(t) b_{k'}''(t) dt$, for $k, k' \in \{1, \ldots, K\}$, respectively. Notably, the sparsity-penalty matrix $\boldsymbol{\Omega}^{(1)}$ and the (unscaled) smoothness-penalty matrix $\boldsymbol{\Omega}^{(2)}/M^2$ have fixed quantities given the specified set of basis functions and do not vary with covariates $\boldsymbol{Z}_i$ or outcomes $\{S_{ij}, X_{ij}\}$. We have derived the closed-form expression for $\boldsymbol{\Omega}^{(1)}$ when using natural cubic spline basis functions with K knots placed at $t_1, t_2, \ldots t_K$; see Theorem 3 in Appendix C.1. The smoothness-penalty matrix $\boldsymbol{\Omega}^{(2)}/M^2$ appears in the regularization for many traditional smoothing spline type methods (Parker & Rice, 1985; Wahba, 1980; Wahba et al., 1995; Wood, 2011) and can be directly calculated from existing R packages like `mgcv` (Wood, 2017).

We can then rewrite the sparsity-smoothness penalty (5.2) in a more compact way,

$$\mathcal{L}^{\text{SSP}}(\boldsymbol{\theta}) = \lambda \sum_{p=1}^{P} \sqrt{\boldsymbol{\theta}_p^T \boldsymbol{H}_\alpha \boldsymbol{\theta}_p} \tag{5.3}$$

where $\mathbf{H}_\alpha = (1-\alpha)\boldsymbol{\Omega}^{(1)} + \alpha\boldsymbol{\Omega}^{(2)}$ and this is a general group lasso penalty (Yuan & Lin, 2006) for any fixed α.

Relations with other regularization methods The composite sparsity-smoothness penalty function in (5.3) encompasses a class of regularization methods for high-dimensional generalized additive models. When fixing $\alpha = 0$, the SSP penalty is closely related to the SpAM (Ravikumar et al., 2009) and the method of Wei et al. (2011). This formulation decouples the choice of smoother complexity from the sparsity constraint, and its estimation accuracy can be sensitive to the choice of basis dimensions (see Figure 5.3). When $\boldsymbol{H}_\alpha$ equals an identity matrix, the SSP penalty is reduced to an ordinary group Lasso problem and is related to the method of Huang et al. (2010). In this case, the sparsity penalty is imposed directly on the basis coefficients $\boldsymbol{\theta}_p$, other than the entire functional component $\beta_p(t)$.

5.3 Computational algorithm

Model (5.1) with penalization in (5.3) would be estimated by optimizing

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \ell(\boldsymbol{\theta}) + \lambda \sum_{p=1}^{P} \sqrt{\boldsymbol{\theta}_p^T \boldsymbol{H}_\alpha \boldsymbol{\theta}_p} \right\}, \tag{5.4}$$

where $\ell(\boldsymbol{\theta})$ is the twice negative log likelihood for model (5.1) and takes the form

$$\ell(\boldsymbol{\theta}) = -2 \sum_{i=1}^{N} \sum_{j=1}^{m_i} \{ S_{ij} \log(\pi_{ij}) + (X_{ij} - S_{ij}) \log(1 - \pi_{ij}) \}. \tag{5.5}$$

We refer to $\widehat{\boldsymbol{\theta}}$ in (5.4) as the SSP estimator.

5.3.1 Proximal gradient descent algorithm

In this section, we present a proximal gradient descent algorithm for solving the optimization problem in (5.4). Proximal gradient descent algorithm (Nesterov, 2013; Parikh & Boyd, 2014) is commonly used to optimize composite objective function that can be written as a sum of two terms: one is differentiable (e.g. $\ell(\boldsymbol{\theta})$) and another is a general closed convex function that can be nonsmooth (e.g. $\lambda \sum_{p=1}^{P} \sqrt{\boldsymbol{\theta}_p^T \boldsymbol{H}_\alpha \boldsymbol{\theta}_p}$). It consists of a gradient descent step followed by a proximal mapping step. For many important penalty functions, such as lasso and (genera)l group lasso penalty, the proximal mapping can be computed analytically, thus improving computational efficiency.

To solve the optimization problem in (5.4), we first decompose $\boldsymbol{H}_\alpha = \boldsymbol{L}_\alpha^T \boldsymbol{L}_\alpha$, where $\boldsymbol{L}_\alpha$ is an upper triangular matrix with positive diagonal entries. Define $\widetilde{\boldsymbol{\theta}}_p = \boldsymbol{L}_\alpha \boldsymbol{\theta}_p$ and $\widetilde{\mathbb{X}}_p = \mathbb{X}_p \boldsymbol{L}_\alpha^{-1}$. We can thus rewrite the optimization problem in (5.4) simply as

$$\widehat{\widetilde{\boldsymbol{\theta}}} = \arg\min_{\widetilde{\boldsymbol{\theta}}} \left\{ \ell(\widetilde{\boldsymbol{\theta}}) + \lambda \sum_{p=1}^{P} \sqrt{\widetilde{\boldsymbol{\theta}}_p^T \widetilde{\boldsymbol{\theta}}_p} \right\},$$

where $\ell(\widetilde{\boldsymbol{\theta}})$ is defined in (5.5) with $\boldsymbol{\pi} = [1 + \exp(\widetilde{\mathbb{X}}\widetilde{\boldsymbol{\theta}})]^{-1}$. We will first use proximal gradient descent to find $\widehat{\widetilde{\boldsymbol{\theta}}}_p$, and then obtain $\widehat{\boldsymbol{\theta}}_p = \boldsymbol{L}_\alpha^{-1} \widehat{\widetilde{\boldsymbol{\theta}}}_p$, for $p = 0, 1, \ldots P$.

After initialization of $\widetilde{\boldsymbol{\theta}}^{(0)}$, at the s-th iteration we update $\widetilde{\boldsymbol{\theta}}^{(s)}$ by the following updating formula

$$\widetilde{\boldsymbol{\theta}}^{(s)} \longleftarrow \text{prox}_{t_s} \left[\widetilde{\boldsymbol{\theta}}^{(s-1)} - t_s \nabla \ell(\widetilde{\boldsymbol{\theta}}^{(s-1)}) \right], \tag{5.6}$$

for $s = 1, 2, 3, \ldots$, until the convergence of $\widetilde{\boldsymbol{\theta}}$. In (5.6) the *proximal operator* $\text{prox}_f : \mathbb{R}^{PK} \to$

$\mathbb{R}^{PK}$ is defined as

$$\text{prox}_t(\mathbf{u}) = \arg\min_{\widetilde{\boldsymbol{\theta}}} \left(\frac{1}{2t} \|\mathbf{u} - \widetilde{\boldsymbol{\theta}}\|_2^2 + \lambda \sum_{p=1}^{P} \sqrt{\widetilde{\boldsymbol{\theta}}_p^T \widetilde{\boldsymbol{\theta}}_p} \right).$$

This optimization problem has an analytical solution that can be efficiently computed. Specifically, it is easy to show that

$$[\text{prox}_t(\mathbf{u})]_p = \left(1 - \frac{t\lambda}{\sqrt{\mathbf{u}_p^T \mathbf{u}_p}} \right)_+ \mathbf{u}_p, \quad \text{for } p = 1, \ldots, P$$

where $[\text{prox}_t(\mathbf{u})]_p \in \mathbb{R}^K$ is the sub-vector corresponding to the p-th group of $\text{prox}_t(\mathbf{u})$. Therefore, we have $\text{prox}_t(\mathbf{u}) = ([\text{prox}_t(\mathbf{u})]_1^T, \ldots, [\text{prox}_t(\mathbf{u})]_p^T)^T$.

Backtracing line search for the step size To guarantee convergence, we determine the step size t_s at each iteration s in (5.6) using backtracking line search. Define the generalized gradient

$$G_t(\widetilde{\boldsymbol{\theta}}) = \frac{1}{t} \left[\widetilde{\boldsymbol{\theta}} - \text{prox}_t(\widetilde{\boldsymbol{\theta}} - t\nabla \ell(\widetilde{\boldsymbol{\theta}})) \right].$$

Using the notation of $G_t(\widetilde{\boldsymbol{\theta}})$, the update in (5.6) can be equivalently written as $\widetilde{\boldsymbol{\theta}}^{(s)} = \widetilde{\boldsymbol{\theta}}^{(s-1)} - t_s G_{t_s}(\widetilde{\boldsymbol{\theta}}^{(s-1)})$. The backtracing line search works as follows. We first initialize $t = t_{\text{init}} > 0$ and repeatedly shrink t with $t \leftarrow \delta t$ for some pre-specified $0 < \delta < 1$ until

$$\ell\left(\widetilde{\boldsymbol{\theta}}^{(s-1)} - tG_t(\widetilde{\boldsymbol{\theta}}^{(s-1)}) \right) \leq \ell(\widetilde{\boldsymbol{\theta}}^{(s-1)}) - t\nabla \ell(\widetilde{\boldsymbol{\theta}}^{(s-1)})^T G_t(\widetilde{\boldsymbol{\theta}}^{(s-1)}) + \frac{t}{2} \|G_t(\widetilde{\boldsymbol{\theta}}^{(s-1)})\|_2^2. \tag{5.7}$$

Once (5.7) is satisfied by some t, we set $t_s \leftarrow t$ and update $\boldsymbol{\theta}^{(s)}$ using (5.6) with this chosen step size. The proposed overall estimating algorithm is summarized in Algorithm 3.

Algorithm 3: Proximal gradient algorithm with backtracking line search.

Initialize $\widetilde{\boldsymbol{\theta}}^{(0)} = \mathbf{0}$; Choose some $0 < \delta < 1$; Choose $\varepsilon = 10^{-6}$; Set $s = 0$;
repeat
  $s \leftarrow s + 1$;
  Initialize $t = t_{\text{init}}$;
  repeat
    $t \leftarrow \delta t$;
  until $\ell(\widetilde{\boldsymbol{\theta}}^{(s-1)} - tG_t(\widetilde{\boldsymbol{\theta}}^{(s-1)})) \leq \ell(\widetilde{\boldsymbol{\theta}}^{(s-1)}) - t\nabla\ell(\widetilde{\boldsymbol{\theta}}^{(s-1)})^T G_t(\widetilde{\boldsymbol{\theta}}^{(s-1)}) + \frac{t}{2}\|G_t(\widetilde{\boldsymbol{\theta}}^{(s-1)})\|_2^2$;
  Set $t_s = t$;
  Update $\widetilde{\boldsymbol{\theta}}^{(s)} \longleftarrow \text{prox}_{t_s}\left[\widetilde{\boldsymbol{\theta}}^{(s-1)} - t_s\nabla\ell(\widetilde{\boldsymbol{\theta}}^{(s-1)})\right]$ as defined in (5.6);
until $\|\widetilde{\boldsymbol{\theta}}^{(s)} - \widetilde{\boldsymbol{\theta}}^{(s-1)}\|_2 < \varepsilon$;
Return $\widetilde{\boldsymbol{\theta}}^{(s)}$;

5.3.2 Choosing the tuning parameters

The algorithm in the previous section computes the estimates for $\boldsymbol{\theta}$ for given values of tuning parameters λ and α. We use cross-validation (CV) to select the values of λ and α by minimizing the averaged prediction errors in the validation sets, called mean CV errors. In our case, the prediction error in the validation set for the o^{th} CV fold, $\mathcal{V}^o$, is quantified by the mean deviance $\frac{1}{M_o}\sum_{i,j\in\mathcal{V}^o}\{-2\left[S_{ij}\log(\widehat{\pi}_{ij}) + (X_{ij} - S_{ij})\log(1-\widehat{\pi}_{ij})\right]\}$, where M_o is the total number of observations in $\mathcal{V}^o$. In our package `sparseSOMNiBUS`, we also allow users to select the value of λ based on the "one-standard-error" rule (1-SE-rule) (Friedman et al., 2010). Specifically, we select the largest value of λ such that the mean CV error is within 1 SE of the minimum. This strategy generally favors parsimonious models.

For a given value of α, we can derive the smallest λ that gives the entire effect vector $\widehat{\boldsymbol{\theta}}_{\mathbf{1}} = \widehat{\boldsymbol{\theta}}_{\mathbf{2}} = \ldots = \widehat{\boldsymbol{\theta}}_{\boldsymbol{P}} = \mathbf{0}$ in our optimization problem (5.4). Such value is the so-called λ_{max} in the regularization path for λ. We first derive the expression of λ_{max}, then present a computationally efficient way for computing $\widehat{\boldsymbol{\theta}}$s under a sequence of λ.

Derive λ_{max}

The derivation of λ_{max} relies on calculating the optimality conditions for the nonlinear programming problem in (5.4). Such conditions test whether a solution is optimal and are also called Karush-Kuhn-Tucker (KKT) conditions. Write $f(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \lambda \sum_{p=1}^{P} \sqrt{\boldsymbol{\theta}_p^T \mathbf{H}_\alpha \boldsymbol{\theta}_p}$ for our objective function. Its optimality condition simply states that $\boldsymbol{\theta}^\star$ is a minimizer of $f(\boldsymbol{\theta})$ if and only if 0 is a subgradient of $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^\star$.

The (sub)gradient for the differential part $\ell(\boldsymbol{\theta})$ in $f(\boldsymbol{\theta})$ is

$$\nabla \ell(\boldsymbol{\theta}) = -2 \left[\mathbb{X}^T (\boldsymbol{S} - \boldsymbol{\Lambda_X \pi}) \right],$$

where $\boldsymbol{S} \in \mathbb{R}^M$ is the vector concatenating S_{ij}, and $\boldsymbol{\Lambda_X} \in \mathbb{R}^{M \times M}$ is the diagonal matrix with values of read-depths X_{ij}. The subgradient for the non-differential part $h(\boldsymbol{\theta_p}) = \sqrt{\boldsymbol{\theta_p}^T \boldsymbol{H}_\alpha \boldsymbol{\theta_p}}$ can be shown as

$$\partial h(\boldsymbol{\theta_p}) = \begin{cases} \dfrac{\boldsymbol{H}_\alpha \boldsymbol{\theta_p}}{\sqrt{\boldsymbol{\theta_p}^T \boldsymbol{H}_\alpha \boldsymbol{\theta_p}}}, & \boldsymbol{\theta_p} \neq \mathbf{0} \\ \left\{ \boldsymbol{g} \in \mathbb{R}^K : \sqrt{\boldsymbol{g}^T \boldsymbol{H}_\alpha^{-1} \boldsymbol{g}} \leq 1 \right\}, & \boldsymbol{\theta_p} = \mathbf{0}; \end{cases}$$

see detailed derivations in Appendix C.2. Therefore, the KKT conditions for a solution $\boldsymbol{\theta}$ to be optimal for our nonlinear programing problem in (5.4) are

$$\begin{cases} \boldsymbol{a}_p = \lambda \dfrac{\boldsymbol{H}_\alpha \boldsymbol{\theta_p}}{\sqrt{\boldsymbol{\theta_p}^T \boldsymbol{H}_\alpha \boldsymbol{\theta_p}}}, & \text{if } \boldsymbol{\theta_p} \neq \mathbf{0} \\ \sqrt{\boldsymbol{a}_p^T \boldsymbol{H}_\alpha^{-1} \boldsymbol{a}_p} \leq \lambda, & \text{if } \boldsymbol{\theta_p} = \mathbf{0}, \end{cases} \tag{5.8}$$

where $\boldsymbol{a}_p = 2 \left[\mathbb{X}_p^T (\boldsymbol{S} - \boldsymbol{\Lambda_X \pi}) \right] \in \mathbb{R}^K$ denotes the sub-vector of $-\nabla \ell(\boldsymbol{\theta})$ corresponding to $\boldsymbol{\theta}_p$ and $p = 1, 2, \ldots P$. The condition for $p = 0$ is simply $\boldsymbol{a}_0 = \mathbf{0}$.

Considering the case when the optimal minimizer for (5.4) is a vector of $\boldsymbol{\theta}$ with components

$\boldsymbol{\theta_1} = \boldsymbol{\theta_2} = \ldots = \boldsymbol{\theta_P} = \mathbf{0}$, the optimality conditions in (5.8) imply that

$$\lambda \geq \sqrt{\boldsymbol{b}_p^T \boldsymbol{H}_\alpha^{-1} \boldsymbol{b}_p}, \text{ for } \forall p \in \{1, 2, \ldots, P\}.$$

Here, $\boldsymbol{b}_p = 2\left[\mathbb{X}_p^T(\boldsymbol{S} - \boldsymbol{\Lambda_X}\boldsymbol{\pi}_0)\right]$ is the sub-vector of $-\nabla\ell(\boldsymbol{\theta})$ corresponding to $\boldsymbol{\theta}_p$ evaluating from an intercept-only model, where $\boldsymbol{\pi}_0 \in \mathbb{R}^M$ has elements $[1+\exp(-\beta_0(t_{ij}))]^{-1}$. Therefore, we show that the smallest λ that gives $\boldsymbol{\theta_1} = \boldsymbol{\theta_2} = \ldots = \boldsymbol{\theta_P} = \mathbf{0}$ takes the form

$$\lambda_{max} = \max_{p \in \{1,2,\ldots P\}} \left\{ \sqrt{\boldsymbol{b}_p^T \boldsymbol{H}_\alpha^{-1} \boldsymbol{b}_p} \right\}.$$

The warm start strategy

For a given α, we construct a sequence of L values for λ decreasing from λ_{max} to $\tau\lambda_{max}$ on the log scale, where τ is a small constant. The defaults in our package are set to $L = 100$, $\tau = 0.01$ if $M < (P+1)K$, and $\tau = 0.001$ if $M \geq (P+1)K$, following Friedman et al. (2010). We then fit a sequence of models from λ_{max} to $\tau\lambda_{max}$ using the warm start strategy (Friedman et al., 2007). That is, the solution for the lth λ is used as the initial value for the $(l+1)$th λ. This strategy provides a good initialization for the optimization problem at a new λ and leads to considerable computational speedups.

5.4 The adaptive sparsity-smoothness penalty

Similar to the adaptive LASSO (Zou, 2006), we can introduce weights to allow for different amounts of penalties for individual functional components in the model. Specifically, we define the adaptive sparsity-smoothness penalty function as

$$\mathcal{L}^{\text{SSP,adp}}(\boldsymbol{\theta}) = \lambda \sum_{p=1}^{P} \sqrt{w_{1,p}(1-\alpha) J_1\left(\beta_p(t)\right) + w_{2,p}\alpha J_2\left(\beta_p(t)\right)}, \tag{5.9}$$

where $w_{1,p}$ and $w_{2,p}$ are data-adaptive weights (Meier et al., 2009). A typical choice for the weights would be

$$w_{1,p} = \frac{1}{\sqrt{J_1\left(\widehat{\beta}_{p,int}(t)\right)}} \text{ and } w_{2,p} = \frac{1}{\sqrt{J_2\left(\widehat{\beta}_{p,int}(t)\right)}},$$

where $\widehat{\beta}_{p,int}(t)$ is the ordinary SSP estimator. We then compute the estimator for $\boldsymbol{\theta}$ similarly as described in Section 5.3. We refer to the estimator obtained from this adaptive approach as adaptive SSP estimator.

5.5 Simulation study

We conducted simulations to assess the finite-sample properties of our proposed estimator. In addition to the general SSP estimator that involves both sparsity and smoothness penalties, we consider its two special cases — SSP0 estimator, which involves no smoothness penalty (i.e. $\alpha = 0$), and group LASSO (gLASSO) estimator obtained by fixing $\boldsymbol{H}_\alpha = \boldsymbol{I}$. We compared their performances with the method implemented in `mgcv` (Wood, 2011), which is commonly used for fitting generalized additive models (GAM) but imposes no sparsity constraints. We also applied the adaptive SSP and the SSP with 1-SE-rule to some of our simulation examples.

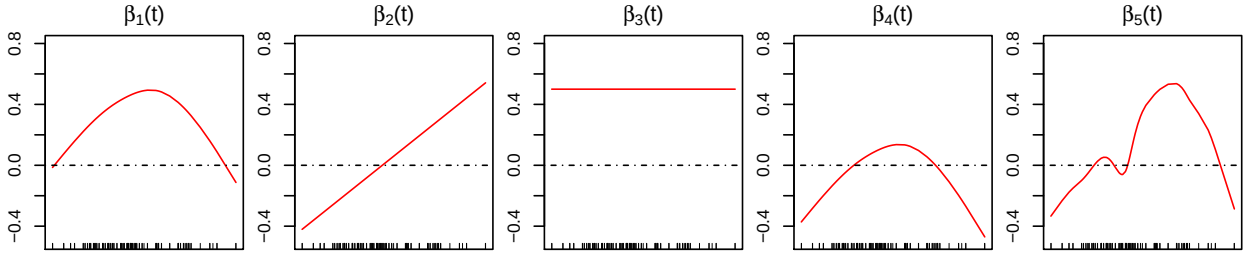
5.5.1 Simulation design

Our simulation design was inspired by a methylation region described in the data example in Zhao et al. (2020). We simulated methylation regions of the same size (123 CpGs) and with the same CpG distribution as the *BANK1* region in Zhao et al. (2020). We considered four simulation examples with various settings for the functional parameters $\beta_p(t)$, sample size N, total number of candidate SNPs P, and the number of true mQTLs P_{true}, as summarized

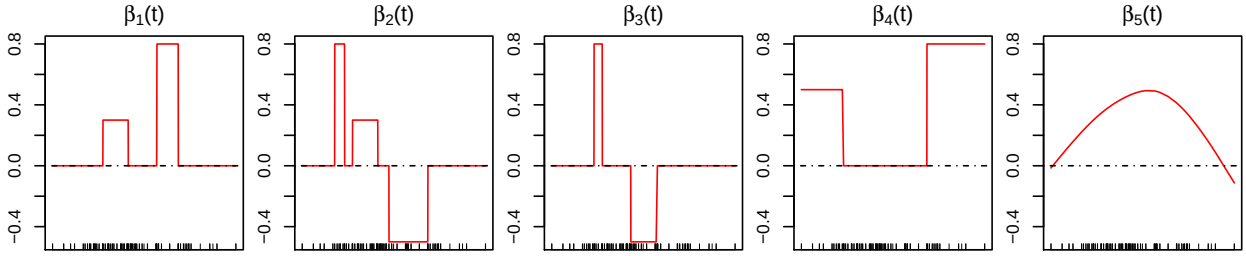
in Table 5.1.

Table 5.1: The shapes of the nonzero $\beta_p(t)$s associated with covariates Z_1 to Z_5 in our four simulation examples. $\beta_p(t) = 0$ for all remaining covariates except for the illustrated ones.

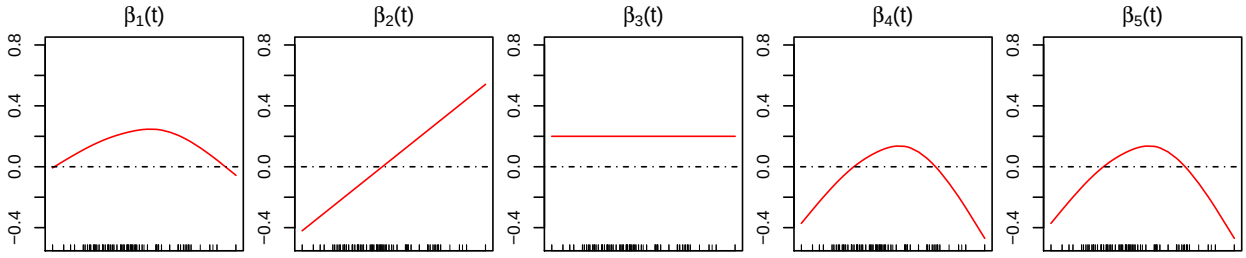
Example 1: (smooth, $N = 50$, $P_{true} = 5$, $P = 100$ or 1000)



Example 2: (nonsmooth, $N = 50$, $P_{true} = 5$, $P = 100$ or 1000)



Example 3: (smooth, $N = 20$, $P_{true} = 5$, $P = 50, 100, 150, 200$ or 1000);
Similar to Example 1 but with smaller effect sizes for Z_1, Z_3 and Z_5 and smaller sample sizes.



Example 4: (smooth, $N = 20$, $P_{true} = 10$, $P = 50, 100, 150, 200$ or 1000)
The same $\beta_p(t)$ for $p = 1, \ldots 5$ as Example 3 and $\beta_p(t) = \beta_{p-5}(t)$ for $p = 6, 7, \ldots 10$.

We first simulated the minor allele frequencies for each candidate SNP independently from a uniform distribution, i.e. $f_p \sim \text{Uniform}(0.1, 0.5)$, for $p = 1, 2, \ldots P$. We then generated genotype Z_p from the truncated multivariate normal distribution with correlation matrix $\mathbf{\Sigma} \in \mathcal{R}^{P \times P}$ and appropriate thresholding for the mean such that

$$P(Z_p = 0) = (1 - f_p)^2, P(Z_p = 2) = f_p^2, \text{ and } P(Z_p = 1) = 2f_p(1 - f_p).$$

We specified $\boldsymbol{\Sigma}$ as a block diagonal matrix, consisting of sub-matrices $\boldsymbol{\Sigma}^{\text{sub}} \in \mathcal{R}^{20\times 20}$ of the form $\boldsymbol{\Sigma}^{\text{sub}} = (1-\rho)\boldsymbol{I}+\rho\mathbf{1}$, where $\boldsymbol{I} \in \mathcal{R}^{20\times 20}$ is an identify matrix, $\mathbf{1} \in \mathcal{R}^{20\times 20}$ is a matrix with all elements as 1. Here ρ is the correlation coefficient and we explored the settings $\rho = 0, 0.3$ or 0.7, corresponding to no, moderate and strong dependence among SNPs. To simulate realistic read depths X_{ij}, we first extracted a spatially correlated read-depth pattern from the real data, denoted as $f^X(t)$, by fitting a cubic spline to the median read-depth across positions. We then generated the read depth X_{ij} by adding Bernoulli random variables (with proportion 0.5) to $f^X(t)$. Given the values of $\{\boldsymbol{Z}, \boldsymbol{X}\}$ and $\{\beta_p(t), p = 0, 1, \ldots, P\}$ under each example and setting, we simulated the methylated counts S_{ij} from the model in (5.1). In addition, an independent test set of the same size was generated for model validation purposes. A total of $R = 100$ simulation runs were used.

When fitting the sparsity-based approaches (i.e. SSP, SSP0 and group LASSO), we used 5-fold cross-validation to select the values of tuning parameters. Specifically, we specified a grid of λ of size 100, using the strategies described in Section 5.3.2, and used a grid of α of size 12, $\alpha = (0, 0.1, 0.2, 0.3, 0.4, 0.5,\ 0.6, 0.7, 0.8, 0.9, 0.95, 0.99)$, when applying the SSP method. Natural cubic splines of rank $K = 10$ were used to expand the functional parameters in the model for all the approaches, unless otherwise stated.

Performance measures

We compared the performances of SSP, SSP0, group LASSO and GAM in terms of their estimation, prediction and variable selection accuracy. Note that the GAM method (implemented in `mgcv`) uses quadratic smoothness penalties and cannot provide sparse solutions. Thus, the variable selection performances were only compared among the sparsity-based approaches.

Estimation To compare the estimation accuracies, we used the Monte Carlo estimates of the integrated mean squared error (IMSE), along with the integrated squared bias (IBIAS^2), and integrated variance (IVAR) for each $\beta_p(t)$. Specifically, let $\left\{\widehat{\beta}^{(r)}(t), r = 1, \ldots, R\right\}$ be the estimates for $\beta(t)$ over the R simulation runs, where the subscript p is dropped for notational simplicity. We define the simulation-based mean estimates for $\beta(t)$ as $\widehat{E}(t) = \frac{1}{R}\sum_{r=1}^{R}\widehat{\beta}^{(r)}(t)$. Thus, we can calculate the three estimation measures

$$\text{IBIAS}^2 = \sum_t \left\{\left[\widehat{E}(t) - \beta(t)\right]^2\right\},\ \text{IVAR} = \sum_t \left\{\frac{1}{R}\sum_{r=1}^{R}\left[\widehat{\beta}^{(r)}(t) - \widehat{E}(t)\right]^2\right\},$$

$$\text{and IMSE} = \sum_t \left\{\frac{1}{R}\sum_{r=1}^{R}\left[\widehat{\beta}^{(r)}(t) - \beta(t)\right]^2\right\},$$

for all the functional coefficients $\beta_p(t)$ in the model.

Prediction We used the hold-out test sets to calculate four prediction measures—deviance error, root mean squared error (RMSE), and correlation between the predicted and observed proportions in the raw and transformed scales (shortly denoted as CorRaw and CorTrans, respectively). The deviance error is defined as $\frac{1}{M}\sum_{i,j\in\text{test}}[\ell(\widehat{\pi}_{ij}; S_{ij}, X_{ij}) - \ell(\pi_{ij}; S_{ij}, X_{ij})]$, where $\ell(\pi; S_{ij}, X_{ij}) = -2\left[S_{ij}\log(\pi) + (X_{ij} - S_{ij})\log(1-\pi)\right]$, $\widehat{\pi}_{ij}$ and π_{ij} are the predicted and true mean for the jth CpGs from the ith sample in the test set, respectively. We define the RMSE as $\left\{\frac{1}{M}\sum_{i,j\in\text{test}}[h(\widehat{\pi}_{ij}) - h(S_{ij}/X_{ij})]^2\right\}^{0.5}$, where $h(\pi) = \arcsin(2\pi - 1)$ is a variance stabilizing transformation of binomial variables (Korthauer et al., 2018). Similarly, CorRaw is calculated as the sample correlation between $\widehat{\pi}_{ij}$ and S_{ij}/X_{ij}, and CorTran is the correlation between $h(\widehat{\pi}_{ij})$ and $h(S_{ij}/X_{ij})$. We reported the mean and standard deviation (SD) of these four measures over all simulation runs to compare different methods.

Selection We used the number of true positives (TP) and false positives (FP) at each simulation run for evaluating the variable selection performances.

5.5.2 Simulation results

A catalogue of the results on the three evaluation themes under different simulation examples and settings is shown in Supporting Table C.3. We first compare the performance of SSP with the two sparsity-only methods SSP0 and gLASSO and discuss the role of smoothness control in Section 5.5.2. We then demonstrate the importance of sparsity control in Section 5.5.2. The results from two extensions of ordinary SSP are shown in Section 5.5.2.

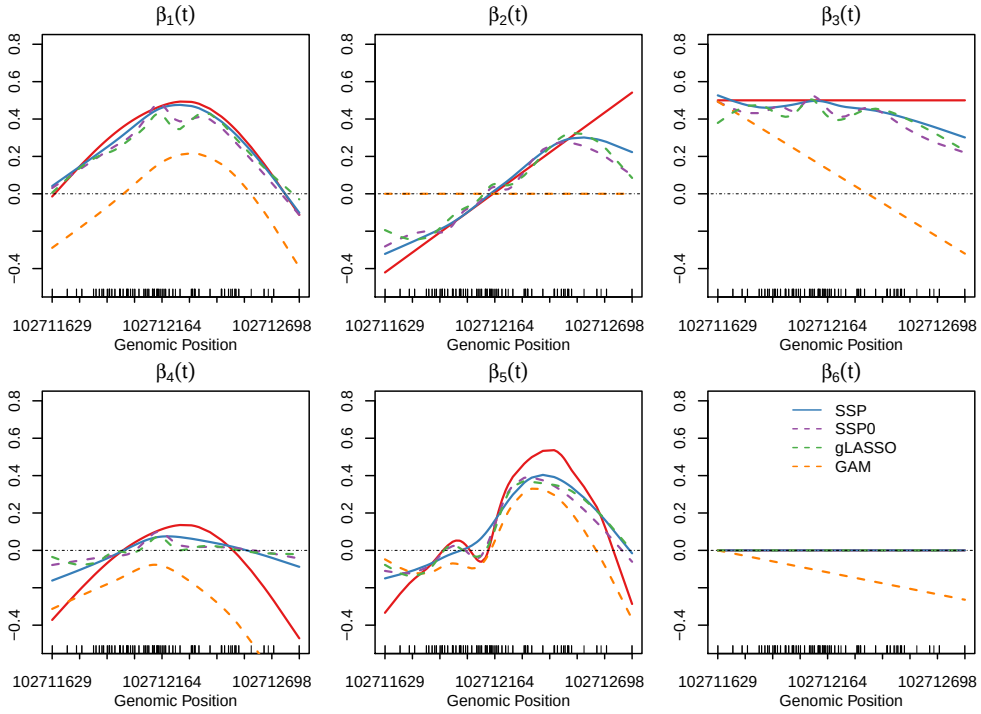


Figure 5.1: Estimates of the first 6 varying coefficients of one simulation run of Example 1 ($P = 100, \rho = 0$), using the SSP, SSP0, group LASSO and GAM approaches. The red curves are the true $\beta_p(t)$ used to generate the data. The results over 100 simulation runs are shown in Supporting Figures C.1-C.4.

The role of the smoothness control in SSP

Figure 5.1 displays the estimated functions from one simulation run of Example 1 ($P = 100, \rho = 0$). It clearly shows that when the true underlying function is smooth, estimates from SSP0 and gLASSO are too wiggly compared to the truth. The estimation plots over 100

simulation runs are shown in Supporting Figures C.1-C.3, and the values of the corresponding IBIAS^2, IVAR and IMSE are given in Table 5.2. The results confirm that for this example, adding the smoothness control reduces both estimation bias and variance compared to the methods that only control the sparsity (i.e SSP0 and gLASSO), which is consistent when $\rho > 0$ (Supporting Table C.4 and C.5) and $P = 1000$ (Supporting Table C.14). In this case, compared to SSP0 and gLASSO, SSP also shows smaller prediction errors (top 2 panels in Table 5.3 and Supporting Table C.8), slightly increased numbers of TPs and decreased numbers of FPs (top 2 panels in Table 5.4). This superiority remains for Examples 3 and 4 (smaller sample sizes and effect sizes than Example 1) under various settings of P and P_{true}; see estimation results in Supporting Tables C.16-C.18, prediction results in Supporting Table C.15 and selection results in Table 5.5.

Table 5.2: Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) of the first 10 varying coefficients of Example 1 ($P = 100, \rho = 0$), using SSP, SSP0, group LASSO and GAM.

	IBIAS^2				IVAR				IMSE			
	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
$\beta_1(t)$	0.260	0.497	0.589	4.681	0.182	0.284	0.320	3.745	0.441	0.781	0.910	8.426
$\beta_2(t)$	0.671	1.289	1.165	1.052	0.183	0.241	0.232	3.570	0.854	1.530	1.397	4.623
$\beta_3(t)$	0.567	1.114	1.081	12.110	0.287	0.439	0.467	8.101	0.855	1.554	1.548	20.211
$\beta_4(t)$	0.473	0.740	0.917	0.066	0.154	0.183	0.158	2.432	0.627	0.922	1.075	2.498
$\beta_5(t)$	0.951	0.961	1.026	0.473	0.298	0.368	0.365	2.174	1.249	1.329	1.391	2.647
$\beta_6(t)$	4.2e-05	1.6e-04	8.7e-05	3.5e-02	1.3e-02	1.4e-02	1.3e-02	1.800	1.3e-02	1.4e-02	1.3e-02	1.835
$\beta_7(t)$	2.0e-04	2.4e-04	7.7e-05	1.1e-02	6.8e-03	6.4e-03	5.3e-03	2.451	7.0e-03	6.6e-03	5.4e-03	2.462
$\beta_8(t)$	1.1e-04	1.6e-04	1.1e-04	7.6e-03	9.5e-03	9.4e-03	8.1e-03	2.014	9.6e-03	9.6e-03	8.2e-03	2.022
$\beta_9(t)$	1.2e-04	1.1e-04	1.3e-04	4.4e-02	7.2e-03	8.5e-03	7.2e-03	1.735	7.3e-03	8.6e-03	7.4e-03	1.779
$\beta_{10}(t)$	1.2e-04	1.1e-04	9.0e-05	1.3e-02	7.1e-03	7.7e-03	6.8e-03	2.228	7.2e-03	7.8e-03	6.9e-03	2.241
$^{\dagger}\sum_1^{100}$	2.931	4.613	4.787	21.429	1.939	2.464	2.401	225.656	4.870	7.077	7.188	247.085

†: sum of the corresponding estimation measures across all varying coefficients in the model

When the true underlying functions are nonsmooth (for Example 2), the estimation results are similar for all the three methods, SSP, SSP0 and gLASSO, as shown in Figure 5.2 and Supporting Table C.13. In this case, the benefit of adding the smooth control is minimal. All the three approaches show considerable bias in estimating the nonsmooth $\beta_p(t)$s and greater prediction errors (see Table 5.3), compared to their performances for Examples 1. This result could have been expected because splines are suited to modelling smooth functions and are less ideal for irregular functions with spikes or abrupt changes. Nevertheless, their variable

Table 5.3: Average values of the deviance error and RMSE over 100 simulations for simulation examples 1 and 2. Standard deviations are given in parentheses.

	Deviance				RMSE			
ρ	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
Example 1 (smooth, $P_{true} = 5, P = 100$)								
0	0.026(0.010)	0.037(0.015)	0.038(0.016)	1.527(0.930)	0.412(0.007)	0.414(0.009)	0.415(0.009)	0.582(0.582)
0.3	0.027(0.014)	0.037(0.018)	0.039(0.019)	1.573(1.135)	0.411(0.007)	0.414(0.008)	0.415(0.008)	0.581(0.581)
0.7	0.026(0.013)	0.036(0.021)	0.040(0.024)	1.559(1.077)	0.409(0.007)	0.412(0.008)	0.413(0.008)	0.575(0.575)
Example 1 (smooth, $P_{true} = 5, P = 1000$)								
0	0.049(0.027)	0.064(0.035)	0.068(0.038)	NA†	0.417(0.009)	0.420(0.011)	0.421(0.011)	NA
0.3	0.043(0.031)	0.054(0.035)	0.058(0.036)	NA	0.416(0.010)	0.418(0.011)	0.419(0.011)	NA
Example 2 (nonsmooth, $P_{true} = 5, P = 100$)								
0	0.175(0.039)	0.176(0.039)	0.165(0.037)	0.977(0.581)	0.425(0.010)	0.425(0.010)	0.424(0.010)	0.515(0.515)
0.3	0.158(0.041)	0.160(0.042)	0.152(0.040)	1.235(0.615)	0.424(0.011)	0.425(0.011)	0.423(0.010)	0.545(0.545)
Example 2 (nonsmooth, $P_{true} = 5, P = 1000$)								
0	0.204(0.066)	0.205(0.066)	0.196(0.066)	NA	0.430(0.013)	0.431(0.013)	0.430(0.013)	NA

† GAM involves no sparsity regularizations and cannot estimate a model with 1000 smooth components for $N = 50$.

Table 5.4: Average values of the number of TP and FP for simulation examples 1 and 2. Standard deviations are given in parentheses.

	TP			FP		
ρ	SSP	SSP0	gLASSO	SSP	SSP0	gLASSO
Example 1 (smooth, $P_{true} = 5, P = 100$)						
0	5.00(0.00)	4.97(0.17)	4.97(0.17)	38.17(8.12)	42.66(7.41)	40.78(7.06)
0.3	4.97(0.23)	4.94(0.28)	4.94(0.28)	34.68(7.84)	37.89(6.89)	36.97(6.72)
0.7	4.95(0.22)	4.89(0.35)	4.88(0.36)	31.86(8.89)	35.73(7.39)	33.22(6.70)
Example 1 (smooth, $P_{true} = 5, P = 1000$)						
0	4.84(0.37)	4.80(0.41)	4.77(0.45)	82.51(16.07)	91.97(14.83)	90.57(15.12)
0.3	4.91(0.28)	4.82(0.46)	4.80(0.50)	77.49(18.30)	83.24(14.41)	83.81(16.57)
Example 2 (nonsmooth, $P_{true} = 5, P = 100$)						
0	5.00(0.00)	5.00(0.00)	5.00(0.00)	47.39(8.18)	46.94(8.15)	40.12(7.13)
0.3	5.00(0.00)	5.00(0.00)	5.00(0.00)	46.68(6.67)	45.67(6.87)	38.96(7.17)
Example 2 (nonsmooth, $P_{true} = 5, P = 1000$)						
0	4.98(0.14)	4.98(0.14)	4.98(0.14)	105.73(18.71)	101.63(18.00)	88.05(14.95)

Table 5.5: Average values of the number of TP and FP for simulation examples 3 and 4 ($N = 20$). Standard deviations are given in parentheses.

		TP			FP		
P	P_{true}	SSP	SSP0	gLASSO	SSP	SSP0	gLASSO
50	5	4.60(0.64)	4.42(0.76)	4.40(0.77)	16.48(4.79)	18.16(4.61)	16.77(4.24)
	10	8.54(1.09)	8.24(1.22)	8.23(1.25)	20.38(4.02)	21.14(3.40)	20.43(3.11)
100	5	4.22(0.84)	3.96(0.86)	3.91(0.93)	23.28(6.93)	24.38(6.76)	23.57(6.70)
	10	7.50(1.11)	7.28(1.20)	7.04(1.23)	30.66(5.32)	31.76(5.43)	30.85(6.23)
150	5	3.96(0.78)	3.72(0.81)	3.72(0.80)	27.56(6.00)	29.16(6.48)	28.13(6.80)
	10	6.62(1.28)	6.36(1.26)	6.36(1.17)	35.98(5.49)	37.94(5.35)	36.06(5.25)
200	5	3.82(0.87)	3.68(0.89)	3.62(0.90)	29.80(7.35)	30.84(7.07)	29.66(6.79)
	10	6.18(1.35)	5.80(1.34)	5.55(1.38)	39.80(7.33)	41.36(7.52)	39.34(7.35)
1000	5	2.42(1.20)	2.34(1.14)	2.29(1.10)	45.28(11.28)	44.78(8.58)	44.84(8.64)
	10	3.66(1.33)	3.36(1.44)	3.13(1.50)	60.64(10.74)	60.26(10.19)	55.64(13.45)

selection performances are less compromised. For example, on average, gLASSO identifies 4.98 TPs and 88.05 FPs when fitting the nonsmooth example with $P_{true=5}$ and $P = 1000$, which are similar to the corresponding results for the smooth example (Table 5.4). We also observe that gLASSO shows slightly smaller prediction errors and reduced numbers of FPs than SSP and SSP0, for the nonsmooth Example 2.

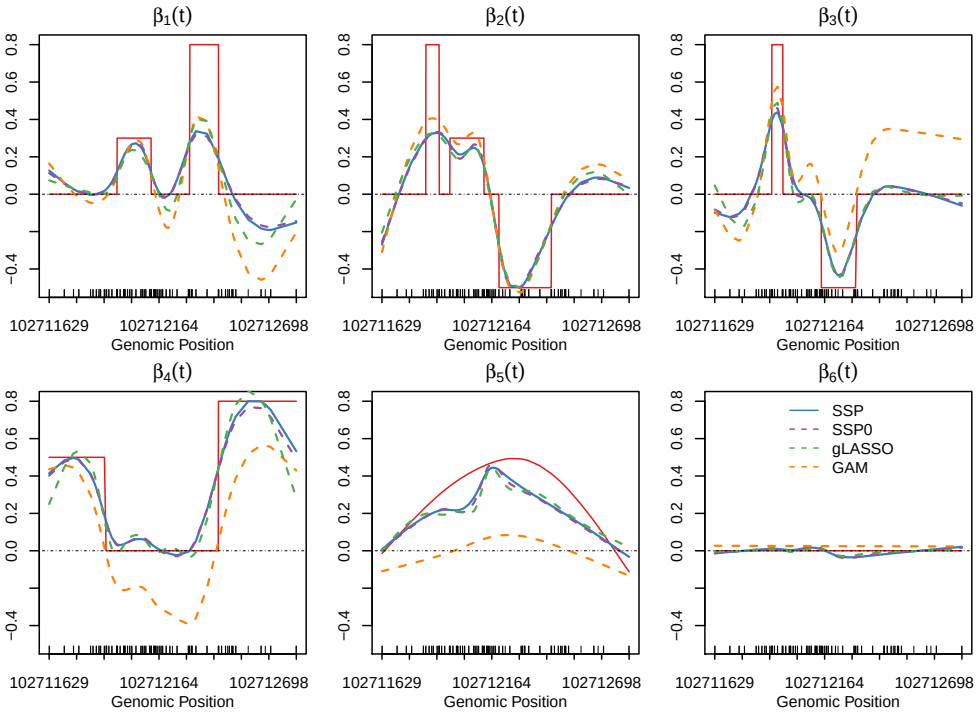


Figure 5.2: Estimates of the first 6 varying coefficients of one simulation run of Example 2 ($P = 100, \rho = 0$), using the SSP, SSP0, group LASSO and GAM approaches. The red curves are the true $\beta_p(t)$ used to generate the data. The results over 100 simulation runs are shown in Supporting Figures C.5-C.8.

Figure 5.3 further demonstrates the role of the smoothness control when the underlying functions are smooth. We compare the results of SSP, SSP0 and gLASSO when using a relatively large number of basis function, $K = 30$, to expand $\beta_p(t)$s. For illustration purposes, we have also plotted the performance measures based on $K = 10$. The results show that the performances of SSP0 and gLASSO deteriorate when using an unnecessarily large number of basis functions; they show increased values of estimation errors, deviance errors, and FP numbers compared to the results from $K = 10$. In contrast, the SSP method

that imposes smoothness penalty is less sensitive to the exact value of the basis dimension and generates almost identical performance measures for $K = 10$ and $K = 30$. In practice, many basis functions are necessary to capture potentially complex functional relationships, such as genetic effects on relatively large methylation regions. SSP can handle this situation and produce smooth estimates by using a more refined control of the smoothness.

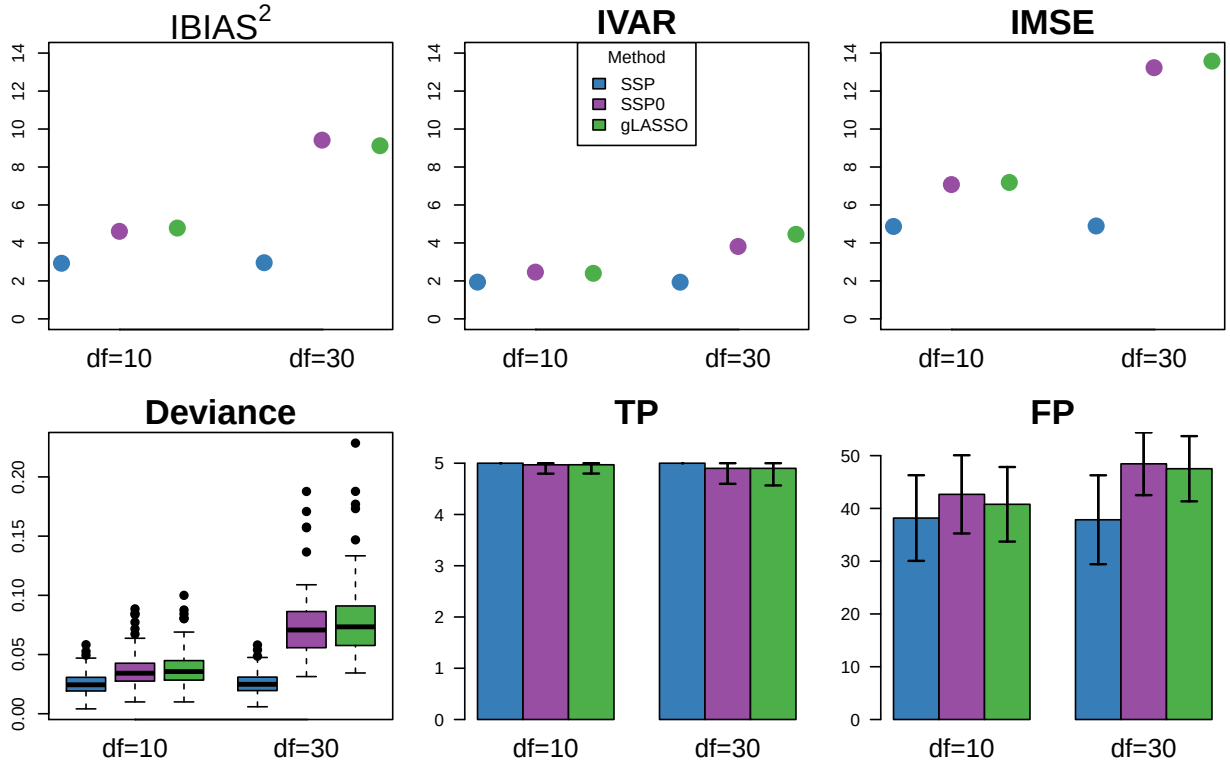


Figure 5.3: Performance measures of SSP, SSP0 and gLASSO when using 10 or 30 basis functions to expand $\beta_p(t)$, labeled as "df=10" and "df=30". Data were generated from Example 1 ($P_{true} = 5, P = 100, \rho = 0$). The top three panels show the values of IBIAS^2, IVAR and IMSE aggregated from all the 100 varying coefficients in the model. The bottom left panel displays the distribution of deviance errors. The "TP" and "FP" panels display the mean values of TP and FP numbers, as well as their SD (indicated by the error bar), over 100 simulation runs.

Sparsity-based methods outperform GAM

Now, we compare the sparsity-based methods SSP, SSP0 and gLASSO with GAM to demonstrate the role of sparsity regularization in estimating high-dimensional VC models. Table

5.2 clearly shows that without the sparsity constraint, GAM has substantially greater estimation variance, which is consistent for the nonsmooth Example 2 (Supporting Table C.13) and Examples 3 and 4 with sample size 20 (Supporting Table C.18). In addition, overall, GAM displays a more significant estimation bias than SSP. Such differences are less pronounced when the actual functions are nonsmooth (Supporting Table C.13) and P is smaller (Supporting Table C.18). GAM also shows worse prediction performance than sparsity-based methods, as shown in Table 5.3, Supporting Tables C.8 and C.15. Notably, GAM cannot estimate models with P that are much larger than N and fails to fit the models with $P = 1000$. In contrast, the sparsity-based methods maintain reasonable prediction accuracies as P increases, as indicated by the deviance errors in Supporting Table C.15. In addition, GAM with quadratic smoothness penalties does not shrink regression coefficients to 0 and thus cannot enable variable selections.

Two extensions of SSP substantially improve variable selection accuracy

Figure 5.4 presents the results obtained from two types of extensions of SSP: the SSP using the 1-SE-rule for choosing λ, as described in Section 5.3.2 and the adaptive SSP, as described in Section 5.4. We also applied these extensions to the two special cases of SSP—the SSP0 and gLASSO approaches. The exact performance measures can be found in Supporting Tables C.6-C.7 and Tables C.9-C.12. The results show that the adaptive approaches generally outperform the ordinary counterparts, and they show reduced estimation bias, prediction errors, and the number of FPs. The estimation variance and the numbers of TPs are similar between the adaptive and ordinary approaches. As expected, the 1-SE-rule can substantially reduce the number of FPs at the cost of increased estimation and prediction errors. The number of TPs using 1-SE-rule is slightly decreased compared to the ordinary approaches.

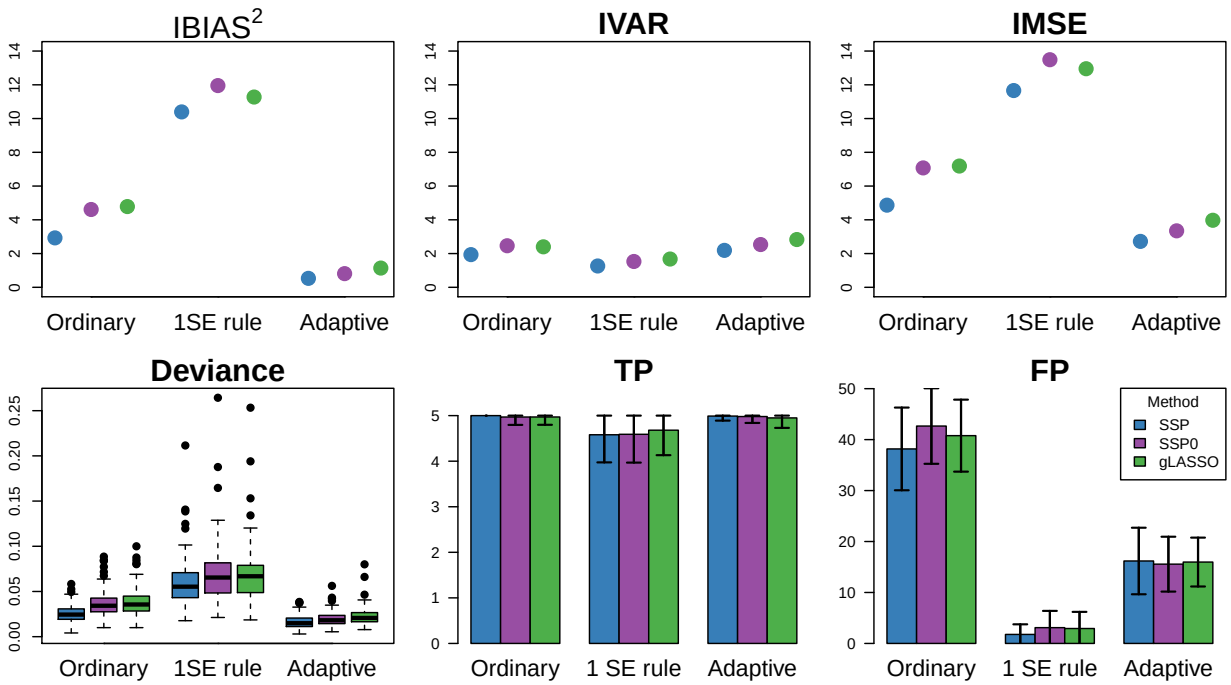


Figure 5.4: Performance measures using the ordinary, 1SE rule and adaptive version of SSP, SSP0 and gLASSO. Data were generated from Example 1 ($P_{true} = 5, P = 100, \rho = 0$). The top three panels show the values of IBIAS^2, IVAR and IMSE aggregated from all the 100 varying coefficients in the model. The bottom left panel displays the distribution of deviance errors. The "TP" and "FP" panels display the mean values of TP and FP numbers, as well as their SDs (indicated by the error bars), over 100 simulation runs.

5.6 Discussion

We have proposed a sparse high-dimensional generalized varying coefficient model for identifying genetic variants associated with regional methylation levels. With different regularization for the sparsity and the smoothness of the functional coefficients, our approach can simultaneously select important mQTLs and estimate their corresponding genetic effects across a methylation region of interest. Furthermore, we present a computationally efficient proximal gradient descent algorithm to estimate the model. A comprehensive simulation study has been conducted to evaluate the performance of our approach in terms of estimation, prediction and selection accuracy. We demonstrate that the inclusion of smoothness control yields much better results than having the sparsity-regularization only if the underlying effects are smooth. When the underlying effects are irregular functions with spikes or

abrupt changes, one can use other types of basis functions, such as Fourier series or wavelets, to achieve higher estimation and prediction accuracy. In addition, we show that combining the sparsity and smoothness regularization provides sparse varying coefficient estimates that are less dependent on basis dimensions.

On the other hand, we have shown that our approach is well suited for high-dimensional cases where the number of covariates is much larger than the sample size, thereby significantly outperforming the traditional smoothing spline-based approach, GAM. Furthermore, using an adaptive version of our penalty function, we can achieve notable additional gains in estimation, prediction and selection accuracy. We have also implemented the 1-SE-rule for selecting the shrinkage parameter λ, which acknowledges that the deviance/risk obtained from cross-validation is subject to estimation errors. We show that this strategy substantially improves variable selection accuracy.

The method has been implemented in R package `sparseSOMNiBUS` (https://github.com/kaiqiong/sparseSOMNiBUS). This tool fills the gap in the existing software for fitting penalized regression models for *non-binary* binomial outcomes. Moreover, our code has options to specify a class of penalty functions, including the general sparsity-smoothness penalty (SSP), the sparsity-only penalty (SSP0), and the simple group LASSO penalty, thereby providing users more flexibility.

Our model assumes that the observed counts of methylated reads represent the true underlying methylation status. However, errors arising from excessive or insufficient bisulfite treatment or other aspects of the sequencing processes can contaminate the observed data. This contamination is unlikely to affect the variable selection results in that covariates with zero effect on the true outcomes are not predictive of the mismeasured outcomes either. Nevertheless, ignoring the measurement error can bias the estimation for the nonzero varying effects on the true methylation status. An extension worth exploring in the future will be to accommodate mismeasured outcomes into our high-dimensional model. Using the error

model in Zhao et al. (2021, 2020), the required developments incorporate theories of adding sparsity penalties to hierarchical binomial regression models whose outcome is dependent on an unobserved latent variable.

Another potential restriction of our method is the distributional assumptions for the outcomes. It could be helpful to set up an additional set of simulations assuming over-dispersed data, such as methylated counts from a beta-binomial distribution, and to see how this affects the estimation, variable selection and predictive performances of our approach. Moving in this direction, another topic worth exploring in the future would be variable selection for quasi-likelihood-based regression models. Such an approach would automatically relax the distributional assumptions for penalized regression models. Furthermore, due to the equivalence between smoothness penalty and (Gaussian) random effects (Silverman, 1985; Wahba, 1983), the idea of adding a square root to the quadratic penalty can be generally applied to random effect selection for mixed effect models.

Chapter 6

Conclusion

6.1 Summary

This thesis presents a body of work that addresses regional association estimation and selection in bisulfite sequencing-derived DNA methylation data. Particularly, the datasets concerned in this thesis are from targeted custom capture sequencing platforms, which produce DNA methylation levels for CpGs in a set of predefined regions. This thesis consists of three original scholarly manuscripts, presented in Chapters 3, 4 and 5, about developing novel methods for better analyzing these targeted methylation regions.

In Chapter 3, I propose a novel framework for the estimation of covariate effects as smooth functions varying along genomic positions within a region of interest. Here, the regional methylation counts are modelled by a binomial distribution, dependent on read depth. This estimation framework, called SOMNiBUS, simultaneously addresses the discrete nature of the data, the possibility of experimental errors, and the estimation of multiple covariate effects. In addition, SOMNiBUS provides a formal inference for both regional and pointwise tests of differential methylation. Simulation results show that SOMNiBUS provides accurate estimates of covariate effects and has greater power to detect differentially methylated regions

than existing methods. However, one main limitation of SOMNiBUS is that its underlying binomial assumption may be overly restrictive. One sign of violation of such an assumption is overdispersion, arising when data exhibit greater variability than those anticipated based on a binomial regression model. Thus, in the next chapter, I pursue an extension to the standard SOMNiBUS to account for potential overdispersion.

In Chapter 4, I propose a hierarchical quasi-binomial varying coefficient mixed model, called dSOMNiBUS, to allow the outcomes to exhibit extra-binomial variation. This model accommodates both multiplicative and additive dispersion, thereby providing a plausible representation of realistic dispersion trends observed in regional methylation data. Like SOMNiBUS, dSOMNiBUS assumes the observed read counts arise from an unobserved latent true methylation state compounded by errors. To estimate such a hierarchical model, I build a hybrid Expectation-Solving algorithm and propose a special plug-in estimator for the multiplicative dispersion parameter. The properties of the resulting estimators are evaluated using both simulations and data applications. Results show that dSOMNiBUS can provide reliable inference for differential methylation at the regional level, regardless of the types and degrees of overdispersion that data exhibit. The R package implementing both the standard SOMNiBUS and its extension dSOMNiBUS, has been published in R Bioconductor (https://www.bioconductor.org/packages/release/bioc/html/SOMNiBUS.html).

Finally, in Chapter 5, I pursue a high-dimensional extension to the standard SOMNiBUS. The problem concerned here is identifying a subset of the genetic variants with local influence on regional methylation levels, i.e. identifying methylation quantitative trait loci (mQTLs). Such analyses are challenging because one routinely faces hundreds or thousands of candidate SNPs within or surrounding a methylation region and sample sizes tend to be small due to the cost of sequencing. To address this problem, I propose a high-dimensional generalized varying coefficient model accompanied by a composite penalty function that encourages both sparsity and smoothness for the varying coefficients. I also present an

efficient proximal gradient descent algorithm to estimate such a high-dimensional model. Finally, a comprehensive simulation study is conducted to evaluate the performance of the proposed approach in terms of estimation, prediction and variable selection. Results show that this new approach can simultaneously select important mQTLs and estimate their corresponding varying effects across a methylation region with excellent accuracy. A prototype R package for this method, named `sparseSOMNiBUS`, is available in Github (https://github.com/kaiqiong/sparseSOMNiBUS).

6.2 Future work

There are many potential areas for future development to advance the work done in the three main chapters.

The first extension worth exploring will be to accommodate more rich correlation structures for the residual errors in our regression models. Regional methylation measurements can be viewed as functional data, whose observation units are functions or curves defined across a targeted region. This extension amounts to exploring different ways to capturing within-function correlations. For the current work in this thesis, such correlations are accommodated through basis functions and smoothness regularization; this is equivalent to assuming that the within-function covariance has a fixed structure up to multiple constants, i.e. smoothing parameters. On top of that, the method in manuscript II (Chapter 4) adds a curve-level random effect to capture curve-to-curve deviations, leading to a compound symmetry correlation structure for the residual errors. Nevertheless, other types of correlation structures could be explored for the residual errors, such as assuming continuous autoregressive correlation structures or adding curve-level random effects that are functions depending on genomic positions. For models with such complex correlation structures, Bayesian methods are preferred for estimation and inference. More generally, in the functional regression

context, it would be helpful to carefully study the benefit of accounting for complex within-function correlation structures beyond the use of basis functions, and the consequences of ignoring it. In addition, it would help to develop innovative inference or estimation procedures that are less dependent on the correct specification of covariance structures underlying the functional regression models in Chapters 3-5.

All of the methods in Chapters 3-5 are built on the assumption that samples are independent. In practice, one might encounter data sets with correlated samples, such as methylation levels for individuals belonging to the same family, or methylation levels for the same individual measured at different ages. Correlations can also be expected when multiple tissues are sampled on the same individual. These inter-sample (or inter-function) correlations must be appropriately taken into account in the analysis to obtain accurate estimates of the statistical significance of associations. Therefore, another area for future development will be to extend the methods in Chapters 4-5 to allow for correlated samples. One solution would be incorporating additional sets of random effects to capture the between-function correlations induced by the multilevel designs or longitudinally sampled functional observations.

The methods proposed in this thesis are tailored to targeted bisulfite sequencing data. Another future direction is to extend these methods to whole-genome bisulfite sequencing (WGBS) data. This development requires first segmenting the whole genome into regions or using sliding windows. The optimal segmentation definitions or choices of window sizes are challenges to be faced.

Furthermore, in this thesis, I consider the genome as a linear sequence of nucleotides and model covariate effects on methylation as functions defined on that linear sequence, i.e. one-dimensional (1D) functions. In fact, inside the nucleus, the genome does not exist as a linear entity but has a three-dimensional (3D) structure. Payne et al. (2021) have recently developed an in situ genome-sequencing technique that allows simultaneous sequencing and imaging of the genome and provides direct information on 3D genomic coordinates in single

cells. As the technology evolves, more refined information on the 3D spatial localization of DNA in intact samples will be expected. Therefore, one promising direction for future work will be developing statistical methods to estimate functional parameters defined on higher dimensional domains, such as on 3D spaces.

6.3 Concluding remarks

Combining DNA bisulfite treatment with high-throughput sequencing technologies has opened new avenues for understanding the role of DNA methylation in disease development. However, extracting interpretable results from raw sequencing data is challenging. This thesis has provided novel analytical tools to estimate and test association patterns in bisulfite sequencing data. Furthermore, the methods developed in these three manuscripts complement the existing statistical literature on the flexible modelling for mismeasured (and overdispersed) binomial outcomes or high-dimensional covariates. This work could be of great value considering the massive popularity of DNA methylation studies in the last decades.

Appendix A

Supporting Information for Chapter 3

This Supporting Information includes detailed derivations, proofs, additional simulation and data application results, and software and data guidance.

A.1 Detailed derivations and proofs

A.1.1 Appendix A: the form of the spanned design matrix

The design matrix $\underset{[M\times K]}{\mathbb{X}} = \left(\boldsymbol{B}^{(Z_0)}, \boldsymbol{B}^{(Z_1)}, \ldots \boldsymbol{B}^{(Z_P)}\right)$ consists of the blocks

$$\underset{(M\times L_p)}{\boldsymbol{B}^{(Z_p)}} = \begin{pmatrix} B_1^{(p)}(t_{11}) \times Z_{p1} & \ldots & B_{L_p}^{(p)}(t_{11}) \times Z_{p1} \\ \vdots & & \vdots \\ B_1^{(p)}(t_{1m_1}) \times Z_{p1} & \ldots & B_{L_p}^{(p)}(t_{1m1}) \times Z_{p1} \\ B_1^{(p)}(t_{21}) \times Z_{p2} & \ldots & B_{L_p}^{(p)}(t_{21}) \times Z_{p2} \\ \vdots & & \vdots \\ B_1^{(p)}(t_{2m_2}) \times Z_{p2} & \ldots & B_{L_p}^{(p)}(t_{2m2}) \times Z_{p2} \\ \vdots & & \vdots \\ B_1^{(p)}(t_{Nm_N}) \times Z_{pN} & \ldots & B_{L_p}^{(p)}(t_{NmN}) \times Z_{pN} \end{pmatrix} \text{ for } p = 0, 1, \ldots P,$$

where $Z_{0i} \equiv 1$ for $i = 1, 2 \ldots N$.

A.1.2 Appendix B: the P-IRLS step given the values of smoothing parameters

One update in the P-IRLS estimation from step r to step $r+1$ is

$$\boldsymbol{\alpha}^{(r+1)} = (\mathbb{X}^T \boldsymbol{W}^{(r)} \mathbb{X} + \boldsymbol{A}_{\boldsymbol{\lambda}})^{-1} \mathbb{X}^T \boldsymbol{W}^{(r)} \widetilde{\boldsymbol{S}}^{(r)},$$

where $\boldsymbol{W}^{(r)} = \text{Diag}\{w_{11}, \ldots w_{1m_1}, w_{21}, \ldots w_{2m_2}, \ldots w_{Nm_N}\} \in \mathcal{R}^{M\times M}$ with $w_{ij} = \pi_{ij}^{(r)}(1-\pi_{ij}^{(r)})$ is the weight matrix, and $\widetilde{\boldsymbol{S}}^{(r)} = \left(\widetilde{S}_{11}^{(r)}, \ldots \widetilde{S}_{1m_1}^{(r)}, \widetilde{S}_{21}^{(r)}, \ldots \widetilde{S}_{2m_2}^{(r)}, \ldots \widetilde{S}_{Nm_N}^{(r)}\right) \in \mathcal{R}^M$ with $\widetilde{S}_{ij}^{(r)} = g\left(\pi_{ij}^{(r)}\right) + g'\left(\pi_{ij}^{(r)}\right) \quad \left(\eta_{ij}^{\star} - \pi_{ij}^{(r)}\right)$ is the vector of adjusted response (also called pseudo response) variables.

A.1.3 Appendix C: Laplace approximated restrictive log-likelihood

In the outer optimization, $\boldsymbol{\lambda}$ is estimated by maximizing the Laplace approximated restricted likelihood (Wood, 2011), denoted by $l_r(\boldsymbol{\lambda})$,

$$2l_r(\boldsymbol{\lambda}) = 2l(\widehat{\boldsymbol{\alpha}_{\boldsymbol{\lambda}}}) + \log\left(|\boldsymbol{A}_{\boldsymbol{\lambda}}|\right) - \widehat{\boldsymbol{\alpha}_{\boldsymbol{\lambda}}}^{\boldsymbol{T}} \boldsymbol{A}_{\boldsymbol{\lambda}} \widehat{\boldsymbol{\alpha}_{\boldsymbol{\lambda}}} - \log\left(|\boldsymbol{H} + \boldsymbol{A}_{\boldsymbol{\lambda}}|\right) + M_A log(2\pi)$$

with $\boldsymbol{H} = -\boldsymbol{\partial^2 l(\alpha)/\partial\alpha\partial\alpha^T} = \mathbb{X}^{\boldsymbol{T}}\boldsymbol{\Lambda_X W}\mathbb{X}$. Here, $l(\alpha)$ is the log-likelihood derived from the binomial distribution as defined in the main manuscript, and $\boldsymbol{\Lambda_X} = \text{Diag}\{X_{11}, \ldots, X_{1m_1}, X_{21}, \ldots, X_{2m_2}, \ldots, X_{Nm_N}\}$ is the diagonal matrix with values of read-depths. $\boldsymbol{H}$ depends on the vector $\boldsymbol{\lambda}$ via the dependence of $\boldsymbol{A_\lambda}$ and $\widehat{\boldsymbol{\alpha}}$ on $\boldsymbol{\alpha}$, and M_A is the dimension of the null space of $\boldsymbol{A_\lambda}$.

A.1.4 Appendix D: Proof of Theorem 1

The proof of Theorem 1 is based on Lemmas 1 and 2. Lemma 1 shows the second derivatives of the conditional log-likelihood $Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha^\star})$, and Lemma 2 obtains the Hessian matrix of the marginal log-likelihood of $\boldsymbol{Y}$.

Lemma 1. *The second derivative of the conditional log-likelihood function $Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha^\star})$ with respect to $\boldsymbol{\alpha}$ is*

$$\frac{\partial^2 Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha^\star})}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha^T}} = -\mathbb{X}^T \boldsymbol{\Lambda_X W}\mathbb{X} - \boldsymbol{A_\lambda}, \tag{A.1}$$

where $\boldsymbol{W} = Diag\{w_{11}, \ldots w_{1m_1}, w_{21}, \ldots, w_{2m_2}, \ldots w_{Nm_N}\} \in \mathcal{R}^{M\times M}$ is the weight matrix with element $w_{ij} = \pi_{ij}(1-\pi_{ij})$, and $\boldsymbol{\Lambda_X} = Diag\{X_{11}, \ldots, X_{1m_1}, X_{21}, \ldots, X_{2m_2}, \ldots, X_{Nm_N}\}$ is the diagonal matrix with values of read-depths. The mixed second derivatives of $Q(\alpha \mid \alpha^\star)$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha^\star}$ are

$$\frac{\partial^2 Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha^\star})}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha^{\star T}}} = \mathbb{X}^T \boldsymbol{\Lambda^\star_\eta W^\star}\mathbb{X} \tag{A.2}$$

where $\boldsymbol{W}^\star$ *is the weight matrix evaluated at* $\pi^\star$*, which is the current iteration estimates and* $\boldsymbol{\Lambda}^\star_{\boldsymbol{\eta}}$ *is a diagonal matrix with diagonal elements* $\delta^\star_{ij}$*, defined as,*

$$\delta^\star_{ij} = \frac{Y_{ij}p_1p_0}{\left[p_1\pi^\star_{ij} + p_0(1-\pi^\star_{ij})\right]^2} + \frac{(X_{ij}-Y_{ij})\,(1-p_1)(1-p_0)}{\left((1-p_1)\pi^\star_{ij} + (1-p_0)(1-\pi^\star_{ij})\right)^2}. \tag{A.3}$$

Proof. The Q function takes the form

$$Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^\star) = \sum_{i=1}^{N}\sum_{j=1}^{m_i}\left\{\eta^\star_{ij}\theta_{ij} - X_{ij}\log(1+e^{\theta_{ij}})\right\} - \frac{1}{2}\boldsymbol{\alpha}^{\boldsymbol{T}}\boldsymbol{A}_{\boldsymbol{\lambda}}\boldsymbol{\alpha},$$

where $\theta_{ij} = \log\left(\pi_{ij}/\left(1-\pi_{ij}\right)\right)$. The first term is the binomial log-likelihood function evaluated at $\eta^\star(\boldsymbol{\alpha}^\star)$, the conditional expectations of the true outcome S_{ij}.

We derive the first and second derivatives of $Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^\star)$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^\star$. First, it is easy to show that

$$\frac{\partial Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^\star)}{\partial \boldsymbol{\alpha}} = \sum_{(i,j)}\left\{\left[\eta^\star_{ij} - X_{ij}\pi_{ij}\right]\left[(\mathbb{X})_{(l,\cdot)}\right]^T\right\} - \begin{Bmatrix} \lambda_0\boldsymbol{A_0}\boldsymbol{\alpha_0} \\ \lambda_1\boldsymbol{A_1}\boldsymbol{\alpha_1} \\ \cdots \\ \lambda_P\boldsymbol{A_P}\boldsymbol{\alpha_P} \end{Bmatrix}. \tag{A.4}$$

Here we use $(\mathbb{X})_{(l,\cdot)}$ to denote the l^{th} row of the design matrix, which is the row corresponding to the CpG j of sample i.

Differentiation of equation (A.4) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^\star$ yields respectively

$$\left(\frac{\partial^2 Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^\star)}{\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha}^{\boldsymbol{T}}}\right)_{(m,m')} = \sum_{(i,j)}\left\{-X_{ij}\pi_{ij}(1-\pi_{ij})\,(\mathbb{X})_{(l,m)}\,(\mathbb{X})_{(l,m')}\right\} - \lambda_{\widetilde{p}}\,(\boldsymbol{A_p})_{(k,\widetilde{k})}\,\mathcal{I}_{(m,m')}, \tag{A.5}$$

$$\left(\frac{\partial^2 Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^\star)}{\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha}^{\boldsymbol{\star T}}}\right)_{(m,m')} = \sum_{(i,j)}\left\{\frac{\partial\eta^\star_{ij}}{\partial\pi^\star_{ij}}\pi^\star_{ij}(1-\pi^\star_{ij})\,(\mathbb{X})_{(l,m)}\,(\mathbb{X})_{(l,m')}\right\}, \tag{A.6}$$

for $m, m' = 1, 2, \ldots K$. In the above formulas, $(\bullet)_{(m,m')}$ represents the (m,m') entry of

a matrix. $\mathcal{I}_{(m,m')} = 1$ if α_m and $\alpha_{m'}$ are the basis coefficients for the same functional parameter $\beta_p(t)$, and $\mathcal{I}_{(m,m')} = 0$ otherwise. For the pairs (m, m') that satisfy $\mathcal{I}_{(m,m')} = 1$, we use k and $\widetilde{k}$ to denote the index of the bases associated with coefficients α_m and $\alpha_{m'}$; in other words, α_m and $\alpha_{m'}$ are the k^{th} and $\widetilde{k}^{\text{th}}$ basis coefficients in the linear expansion that are used to represent functional parameter $\beta_p(t)$. In addition, the $\partial\eta^\star_{ij}/\partial\pi^\star_{ij}$ in the formula (A.6) equals to $\delta^\star_{ij}$, as defined in (A.3). The values of δ_{ij} reduce to 0 when error parameters $p_0 = 1 - p_1 = 0$.

Finally, we rewrite the expressions in (A.5) and (A.6) in a compact way using matrices $\boldsymbol{\Lambda_X}, \boldsymbol{W}, \boldsymbol{\Lambda^\star_\eta}$, and obtain the expressions in (A.1) and (A.2). □

Lemma 2. *The Hessian matrix of the marginal log-likelihood of Y has the form*

$$\mathcal{H}(\boldsymbol{\alpha}) = \mathbb{X}^T(-\boldsymbol{\Lambda}_X + \boldsymbol{\Lambda}_\eta)\boldsymbol{W}\mathbb{X} - \boldsymbol{A_\lambda},$$

where $\boldsymbol{\Lambda}_\eta$ is a diagonal matrix with elements δ_{ij}, which is of the similar form as $\delta^\star_{ij}$ in (A.3) but replacing $\pi^\star_{ij}$ with π_{ij}.

Proof. Due to the presence of the latent methylation state S_{ij}, the observed counts Y_{ij} follow a mixture of two binomial distributions. A direct calculation of the observed Fisher information (Hessian matrix) from this marginal distribution is analytically intractable. However, Oakes (1999) showed that, although the marginal log-likelihood itself is not expressible, its observed Fisher information, can be expressed in terms of the Q function (i.e. the conditional expectation of the log-likelihood of S_{ij} given the observed data Y_{ij}) and its derivatives. Specifically, we rely on the work done by Oakes (1999) and calculate the Hessian matrix of the marginal log-likelihood of Y for parameter $\boldsymbol{\alpha}$, $\mathcal{H}(\boldsymbol{\alpha})$, as the sum of two second derivatives of the Q function,

$$\mathcal{H}(\boldsymbol{\alpha}) = \left\{ \frac{\partial^2 Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha^\star})}{\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha^T}} + \frac{\partial^2 Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha^\star})}{\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha^{\star T}}} \right\}\Bigg|_{\boldsymbol{\alpha^\star}=\boldsymbol{\alpha}}.$$

Using the results in Lemma 1, it can be easily shown that the Hessian matrix $\mathcal{H}(\boldsymbol{\alpha})$ of the marginal log-likelihood of Y is

$$\mathcal{H}(\boldsymbol{\alpha}) = \mathbb{X}^T(-\boldsymbol{\Lambda}_X + \boldsymbol{\Lambda}_\eta)\boldsymbol{W}\mathbb{X} - \boldsymbol{A}_{\boldsymbol{\lambda}}.$$

The diagonal matrix $\boldsymbol{\Lambda}_{\boldsymbol{\eta}}$ will be equal to 0 when error parameters $p_0 = 1 - p_1 = 0$, which corresponds to the case with no experimental error present in the data. □

Theorem 2. *Under the usual regularity conditions for maximum likelihood, we have the following asymptotic results for the estimators $\widehat{\boldsymbol{\alpha}}$ obtained from the smoothed-EM algorithm,*

$$\sqrt{M}\,(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{\mathcal{L}} \boldsymbol{MVN}_K(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1}), \quad \text{as } M \to \infty.$$

Here, K is the dimension of the spline coefficients $\boldsymbol{\alpha}$, and $\boldsymbol{\mathcal{I}} = \mathbb{E}\left[-\mathcal{H}_{ij}(\boldsymbol{\alpha})\right]$. Specifically $\mathcal{H}_{ij}(\boldsymbol{\alpha})$ has the form

$$\mathcal{H}_{ij}(\boldsymbol{\alpha}) = \mathbb{X}_{(l,)}^T\left(-X_{ij}w_{ij} + \delta_{ij}w_{ij}\right)\mathbb{X}_{(l,)} - \boldsymbol{A}_{\boldsymbol{\lambda}}, \tag{A.7}$$

where $\mathbb{X}_{(l,)}$ is the l^{th} row of the design matrix $\mathbb{X}$, which corresponds to the CpG j of sample i, and $w_{ij} = \pi_{ij}(1 - \pi_{ij})$ is the element of the weight matrix.

Proof. Based on the results in Lemma 2, we can show that the Hessian matrix calculated from the individual contribution from observation i at position j, $\mathcal{H}_{ij}(\boldsymbol{\alpha})$, can be expressed as in equation (A.7).

Hence, the asymptotic result follows from the fact that smoothed-EM estimate $\widehat{\boldsymbol{\alpha}}$ is a MLE of $\boldsymbol{\alpha}$ for the marginal distribution of $\boldsymbol{Y}$ (Dempster et al., 1977), and $\mathcal{H}_{ij}(\boldsymbol{\alpha})$ is the Hessian matrix of $\boldsymbol{\alpha}$ for the marginal distribution of $\boldsymbol{Y}_{ij}$ (Oakes, 1999).

□

A.2 Additional simulation results

In this section, we present additional Figures and Tables referenced in Sections 4 and 5 in the main manuscript.

A.2.1 Simulation settings and additional results for Type I Error assessment

Figure A.1 displays the 14 simulation settings of functional parameters $\beta_0(t)$ and $\beta_1(t)$ in Scenario 2. Each pairs of $\beta_0(t)$ and $\beta_1(t)$ correspond to the 14 settings for $\pi_0(t)$ and $\pi_1(t)$ shown in Figure 2 in the main manuscript (the black solid lines). Once we fixed the shapes of $\pi_0(t)$ and $\pi_1(t)$ (in Figure 2 in the main manuscript), $\beta_0(t)$ and $\beta_1(t)$ have the forms

$$\begin{aligned} \beta_0(t) &= \log \frac{\pi_0(t)}{1-\pi_0(t)} \\ \beta_1(t) &= \log \frac{\pi_1(t)}{1-\pi_1(t)} - \beta_0(t). \end{aligned}$$

Table A.1: Simulation settings outlined in Section 4.1 in the main manuscript, for the functional parameters $\beta_p(t)$, sample size N, and error parameters p_0 and p_1.

Simulation parameters	Possibilities
$\beta_p(t)$	Scenario 1: three covariates: $Z_1 \sim Bernoulli(0.51)$, $Z_2 \sim Bernoulli(0.58)$ and $Z_3 \sim Bernoulli(0.5)$ with effects $\beta_1(t), \beta_2(t)$ and $\beta_3(t)$ and intercept $\beta_0(t)$, shown in the red curves in Figure 1 of the main manuscript.
	Scenario 2: one covariate $Z \sim Bernoulli(0.5)$ with 14 different settings of $(\beta_0(t), \beta_1(t))$, as shown in Figure A.1 in the Supporting Information.
N	(40, 100, 150, 400)
(p_0, p_1)	$p_0 = 0.003; p_1 = 0.9$

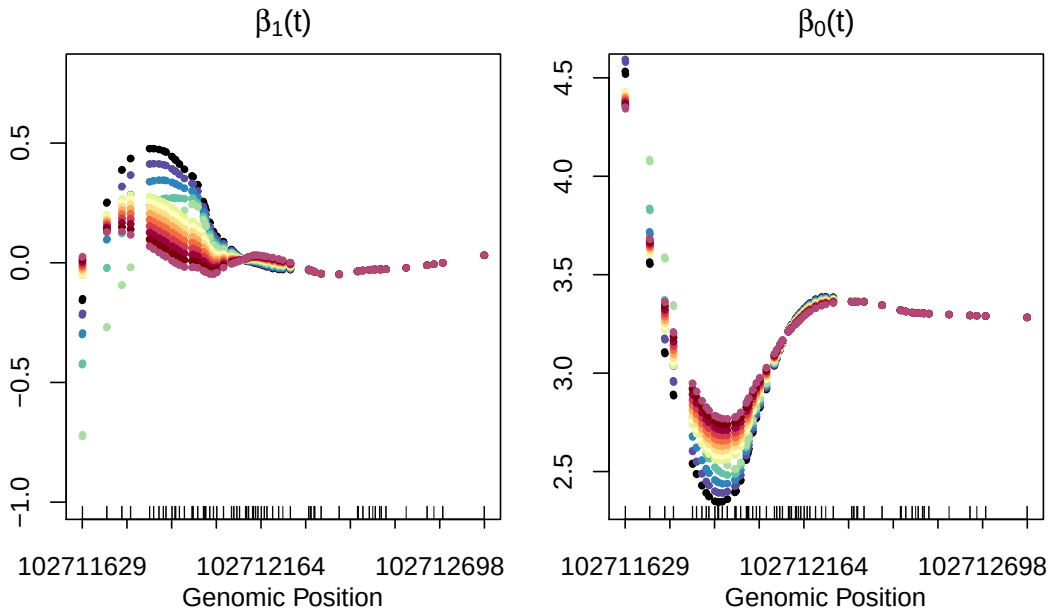


Figure A.1: The 14 simulation settings of functional parameters $\beta_0(t)$ and $\beta_1(t)$ in Scenario 2, which correspond to the 14 settings for $\pi_0(t)$ shown in Figure 2 in the main manuscript.

Table A.1 summarizes the simulation settings outlined in Section 4.1 in the main manuscript.

Figure 3.5 shows the distribution of p-values for the regional effect of the null covariate Z_3 when data were generated with error.

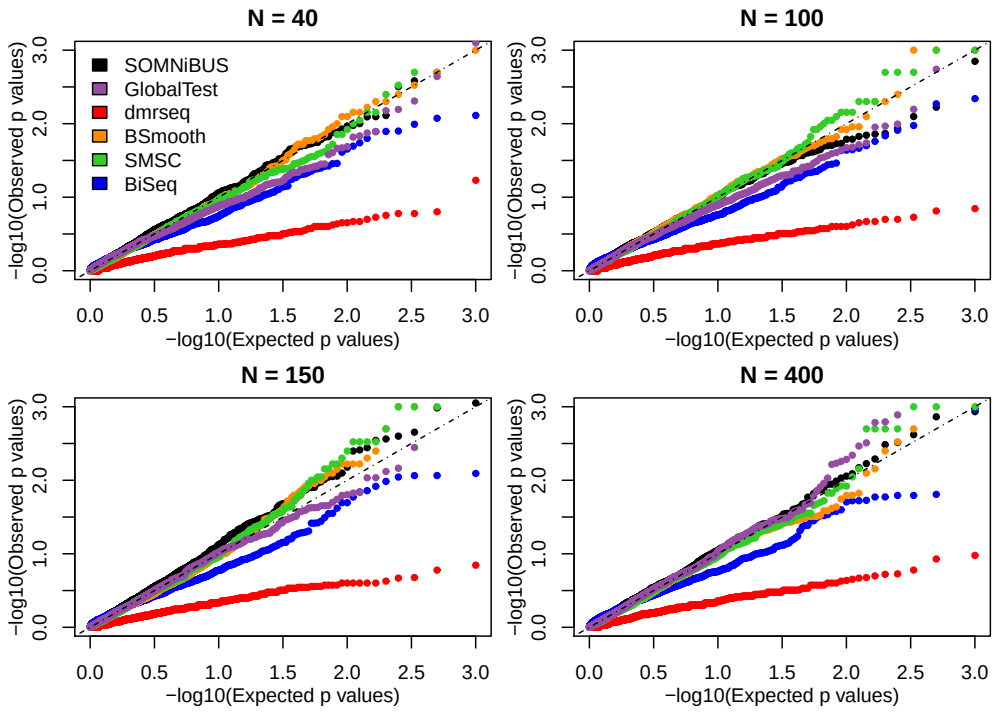


Figure A.2: Quantile-Quantile (Q-Q) plots of the region-based p-values for the null covariate Z_3, obtained from the six methods, over 1000 simulations. Data were generated **with error** with a range of sample sizes ($N = 40, 100, 150, 400$), under simulation Scenario 1. Here, the Expected p-values are uniformly distributed numbers, equal to $= (1/1001, 2/1001, \ldots, 1000/1001)$ and both axes are transformed with $-log10(p)$.

A.2.2 Sensitivity to Bisulfite Sequencing Error Parameters

We explored additional simulation scenarios where the error parameters p_0 and p_1 were misspecified. Specifically, the data were generated subject to errors $p_0 = 0.003$ and $1 - p_1 = 0.1$ but analyses were conducted using a grid of values for p_0 and p_1, constructed from $p_0 = (0, 0.003, 0.005, 0.1, 0.2)$ and $p_1 = (0.88, 0.89, 0.9, 0.95, 1)$. We considered the 14 settings of Scenario 2 that were described in Section 4.1 and graphed in Figure 2 in the main manuscript. These results are shown in columns named S1-S14 in Table A.2. We also included one simulation with a null covariate effect and with varying error parameters, and these results are shown in a column named S0 in Table A.2. These 15 settings S0-S14 correspond to increasing levels of differences between methylation patterns from two groups, i.e. with increasing maximum deviation (MD) between the methylation levels of $Z = 0$ and $Z = 1$.

The powers to detect DMRs for different configurations of p_0 and p_1 under each simulation setting (S0-S14) were given in Table A.2 (note that the power under S0 is the type I error rate). The actual region-based p-values from the 100 simulations for setting S1 with small methylation differences, and setting S14 with large methylation differences, were displayed in Figure A.3 and Figure A.4, respectively. In Figures A.3 and A.4, the region-based p-values using the (mis)specified p_0 and p_1 (vertical axis) were plotted against the ones using the correct p_0 and p_1 (horizontal axis).

Figure A.3 and Figure A.4 show that misspecified error rates can lead to minor differences in regional p-values from the ones with correctly-specified error rates. This difference tends to be greater when the effect size of the covariate of interest is large and when the bias in the error parameters are big. Despite the differences in the actual regional p-values, the powers under various misspecified error rates are shown to be similar to the case with known error rates, as demonstrated in Table A.2. In addition, when the error rates are specified with strong bias, the EM algorithm will not converge. For example, for the simulation scenarios considered in Table A.2, the analyses using $p_1 \leq 0.88$ failed to converge. This also provides a sign of error misspecification.

Table A.2: Powers to detect DMRs using SOMNiBUS when the error parameters p_0 and p_1 were specified differently, under the 14 settings as shown in Figure 2 in the main manuscript (S1-S14) and 1 setting under Null (S0). The powers were calculated over 100 simulations and the data were generated based on the error parameters $p_0 = 0.003$ and $p_1 = 0.9$ (in gray shade), and sample size $N = 100$.

p_1	p_0	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
	0	0.04	0.13	0.24	0.38	0.59	0.71	0.85	0.94	0.97	0.99	1	1	1	1	1
	0.003	0.04	0.13	0.24	0.38	0.59	0.71	0.85	0.94	0.97	0.99	1	1	1	1	1
0.88	0.005	0.04	0.13	0.24	0.38	0.59	0.71	0.85	0.94	0.97	0.99	1	1	1	1	1
	0.1	0.04	0.13	0.24	0.38	0.59	0.7	0.85	0.95	0.98	1	1	1	1	1	1
	0.2	0.04	0.13	0.24	0.38	0.58	0.7	0.85	0.95	0.98	1	1	1	1	1	1
	0	0.04	0.12	0.21	0.39	0.55	0.67	0.81	0.9	0.96	0.98	1	1	1	1	1
	0.003	0.04	0.12	0.21	0.39	0.55	0.67	0.81	0.9	0.96	0.98	1	1	1	1	1
0.89	0.005	0.04	0.12	0.21	0.39	0.55	0.67	0.81	0.9	0.96	0.98	1	1	1	1	1
	0.1	0.04	0.12	0.21	0.39	0.55	0.67	0.81	0.9	0.96	0.98	1	1	1	1	1
	0.2	0.04	0.12	0.21	0.38	0.55	0.67	0.81	0.9	0.96	0.98	1	1	1	1	1
	0	0.04	0.14	0.22	0.37	0.53	0.65	0.77	0.87	0.94	0.99	1	1	1	1	1
	0.003	0.04	0.14	0.22	0.37	0.53	0.65	0.77	0.87	0.94	0.99	1	1	1	1	1
0.9	0.005	0.04	0.14	0.22	0.37	0.53	0.65	0.77	0.87	0.94	0.99	1	1	1	1	1
	0.1	0.04	0.14	0.23	0.37	0.53	0.65	0.77	0.87	0.94	0.99	1	1	1	1	1
	0.2	0.04	0.14	0.23	0.37	0.52	0.65	0.77	0.87	0.93	0.99	1	1	1	1	1
	0	0.06	0.12	0.2	0.31	0.44	0.58	0.7	0.78	0.87	0.94	0.97	1	1	1	1
	0.003	0.06	0.12	0.2	0.31	0.43	0.58	0.7	0.78	0.87	0.94	0.97	1	1	1	1
0.95	0.005	0.06	0.12	0.2	0.31	0.43	0.58	0.7	0.78	0.87	0.94	0.97	1	1	1	1
	0.1	0.06	0.12	0.2	0.3	0.44	0.59	0.68	0.77	0.88	0.95	0.98	1	1	1	1
	0.2	0.06	0.11	0.2	0.32	0.43	0.59	0.68	0.78	0.87	0.95	0.99	1	1	1	1
	0	0.06	0.13	0.18	0.29	0.42	0.54	0.67	0.75	0.87	0.91	0.97	1	1	1	1
	0.003	0.06	0.13	0.18	0.29	0.42	0.54	0.67	0.75	0.87	0.91	0.97	1	1	1	1
1	0.005	0.06	0.13	0.18	0.29	0.42	0.54	0.67	0.75	0.87	0.91	0.97	1	1	1	1
	0.1	0.06	0.13	0.18	0.29	0.42	0.54	0.65	0.75	0.87	0.91	0.98	1	1	1	1
	0.2	0.06	0.13	0.18	0.27	0.42	0.52	0.66	0.75	0.86	0.91	0.97	1	1	1	1

A.2.3 Runtime Comparison

Figure A.5 shows the runtime when fitting a single covariate using the methods under investigation. For dmrseq, we used three different numbers of permutations for comparison (10, 100 and 500). "SOMNiBUS No Error" refers to assuming no sequencing errors in SOMNiBUS, which reduces the full model to a pure generalized additive model. Figure A.5 shows that `SOMNiBUS` requires longer computational times than `GlobalTest`, `BSmooth`, `SMSC` and `BiSeq`, but less than `dmrseq`. Note that our proposed method, `SOMNiBUS`, is capable of estimating the effects of multiple covariates simultaneously, whereas, other methods require repeating the analysis for each covariate, which will multiply the runtime by the number of covariates.

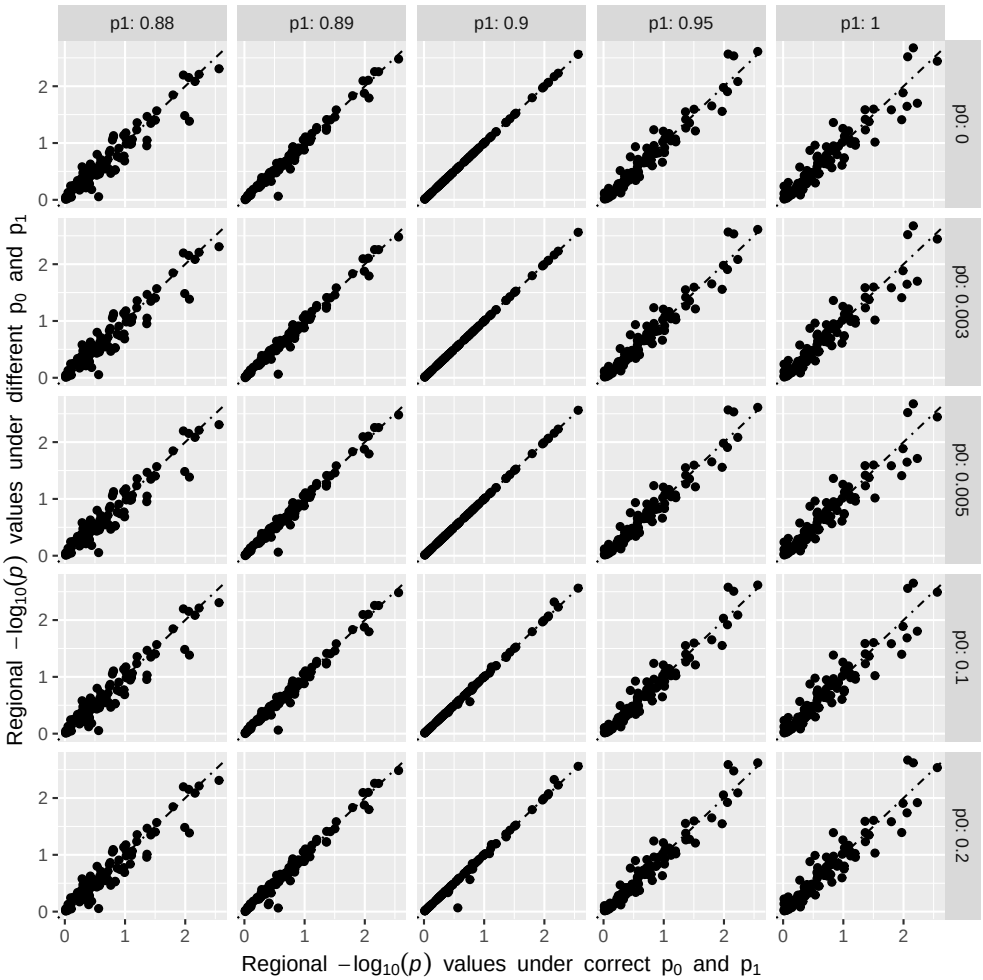


Figure A.3: Scatter plots of the region-based p-values using the specified p_0 and p_1 (vertical axis) compared to the region-based p-values using the correct p_0 and p_1 (horizontal axis), for various settings of p_0 and p_1 specified in the facet labels, under 100 simulations. Here, data were simulated under S1 where MD between the methylation curves in two groups is 0.01 – small effect size.

A.3 Additional data application results

In addition to the *BANK1* region (Orozco et al., 2009), described in Section 3 in the main manuscript, we considered three more regions which overlap with genes *BLK*, *HLA-DRB* and *PTPN22*. These genes have been known associated with risk of rheumatoid arthritis (RA) (Balsa et al., 2010; Hinks et al., 2006; H. Zhang et al., 2012). We applied our method `SOMNiBUS`, along with the five alternative methods—`BiSeq`, `BSmooth`, `SMSC`, `dmrseq` and

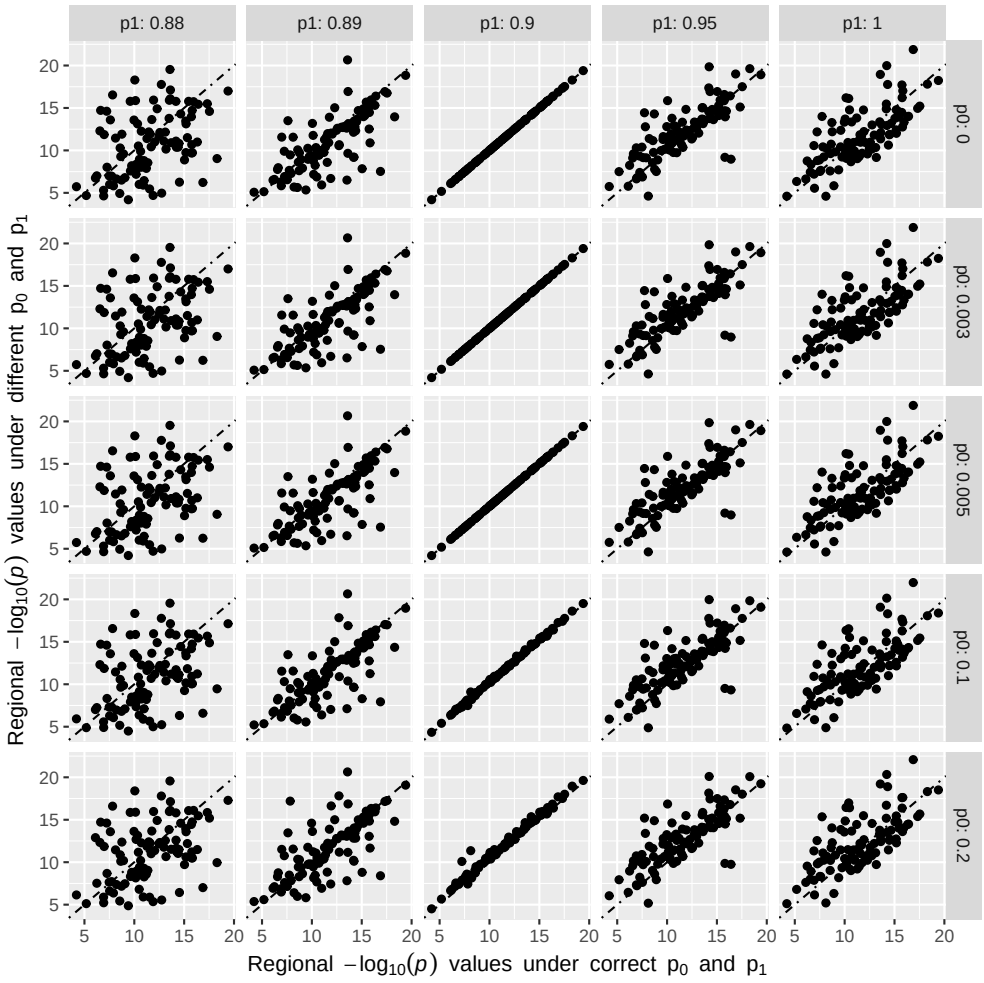


Figure A.4: Scatter plots of the region-based p-values using the specified p_0 and p_1 (vertical axis) compared to the region-based p-values using the correct p_0 and p_1 (horizontal axis), for various settings of p_0 and p_1 specified in the facet labels, under 100 simulations. Here, data were simulated under S14 where MD between the methylation curves in two groups is 0.06 - large effect size.

`GlobalTest`—to each targeted region of interest. Table 3.3 presents the region-based p-values for covariate effects on the four methylation regions. This table shows that `SOMNiBUS` reports smaller regional p-values, and exhibits an improved power to detect these RA-related methylation regions, as compared to the alternative methods.

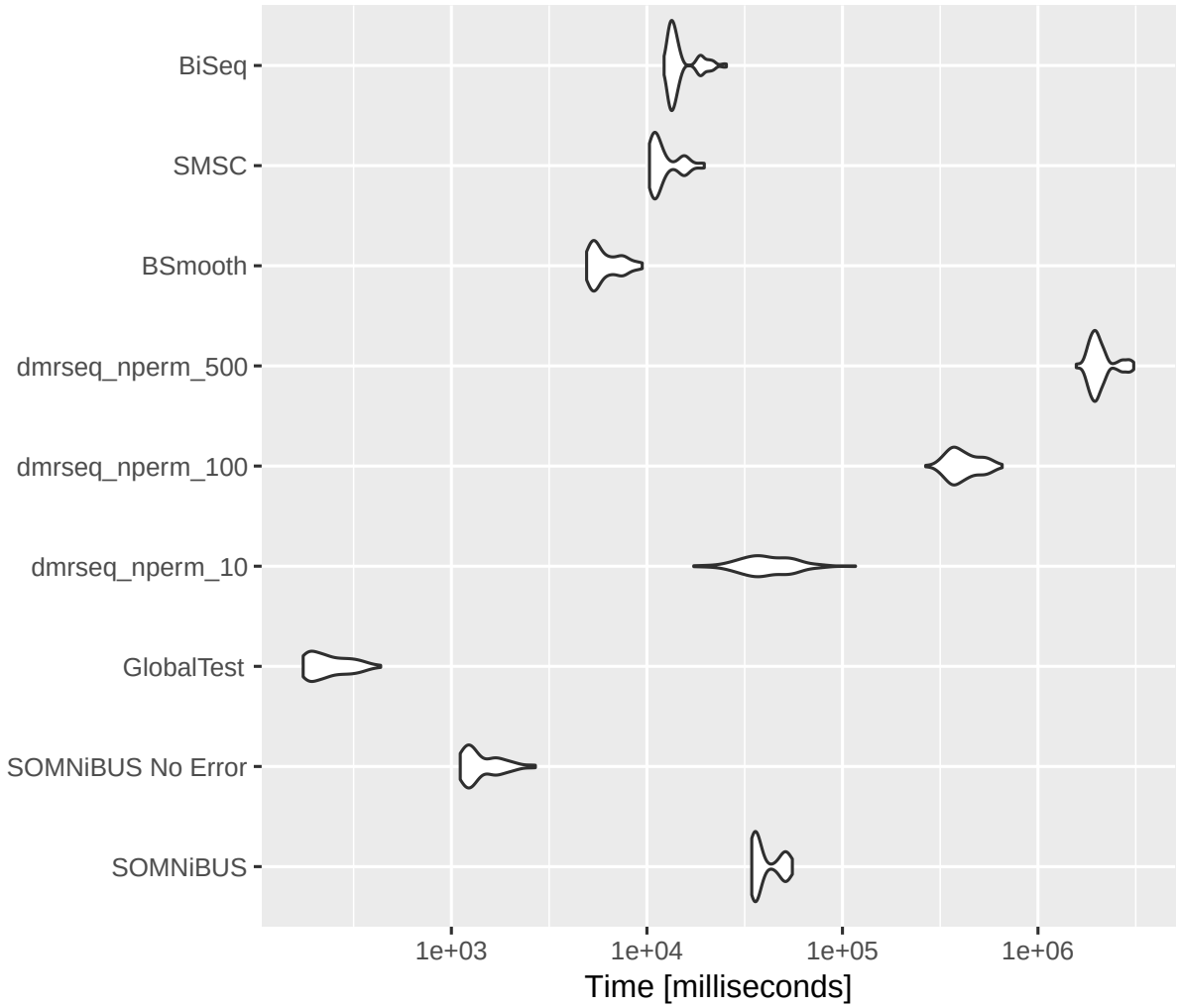


Figure A.5: Summary of runtime under 100 replications. Time axis is presented on the log scale. Data were generated from the S1 of Scenario 2 (with small maximum deviance among the 14 settings in Figure 1) and subject to error $p_0 = 0.003$ and $p_1 = 0.9$. (Sample size N = 100)

Table A.3: Region-based p-values for covariates "RA" and "Tcell" on four different regions using the 6 methods under investigation. Because `dmrseq` is designed for WGBS data, for some regions, `dmrseq` reported more than one DMR in the region. Thus, we calculated the `dmrseq`'s pvalue as the minimum over the reported chunks's p-values. The individual p-values for the small chunks within a region are displayed in the last column. The positions of the four regions are given in the left panel of the table.

	CHR	Start	End	nCpG		SOMNiBUS	GlobalTest	BSmooth	SMSC	BiSeq	dmrseq	(chunk p-values from dmrseq)
BANK1	4	102711629	102712832	123	Tcell	5.61E-217	1.48E-15	0.000	0.000	0.000	0.001	(0.001)
					RA	1.10E-08	0.112	0.714	0.700	0.020	0.161	(0.161, 0.318, 0.718)
BLK	8	11350054	11356772	161	Tcell	1.20E-35	0.130	0.158	0.439	0.112	0.001	(0.001, 0.054, 0.251, 0.288)
					RA	7.72E-42	0.924	0.584	0.978	0.529	0.347	(0.347, 0.602, 0.859, 0.979)
HLA_DRB	6	32546614	32557009	61	Tcell	8.89E-250	2.15E-35	0.000	0.000	0.000	3.63E-04	(3.63E-04, 3.63E-04,3.63E-04, 3.63E-04)
					RA	0.029	0.540	0.643	0.966	0.414	0.593	(0.593, 0.618, 0.651, 0.701, 0.714)
PTPN22	1	114353981	114355828	257	Tcell	1.01E-83	5.29E-17	0.000	0.009	0.330	0.001	(0.001, 0.523, 0.54)
					RA	0.045	0.002	0.360	0.136	0.413	0.062	(0.062, 0.693, 0.829, 0.917)

A.4 Software and data

R-package for SOMNiBUS routine: R-package SOMNiBUS contains code to perform the methods described in the article. (GNU zipped tar file) (https://github.com/kaiqiong/SOMNiBUS)

SOMNiBUS Vignette: A user guide of how to use SOMNiBUS package. The vignette also contains the codes for replicating the data example results in this article. (Rmd and HTML files) (https://github.com/kaiqiong/SOMNiBUS/tree/master/vignettes)

Simulation Codes: Codes to replicate the simulation results in the article are deposited in the Github repository https://github.com/kaiqiong/SOMNiBUS_Simu.

Appendix B

Supporting Information for Chapter 4

This Supporting Information includes detailed derivations, proofs, additional simulation and data application results.

B.1 Detailed derivations and proofs

B.1.1 Appendix A: Marginal interpretations for dSOMNiBUS

Marginal mean

The latent variable representation of the logistic mixed effect model in (4.2) is

$$\begin{aligned} S^{\star}_{ijk} &= \eta_{ij} + \epsilon_{ij} + u_i \\ S_{ijk} &= \begin{cases} 1, & \text{if } S^{\star}_{ijk} \geq 0 \\ 0, & \text{if } S^{\star}_{ijk} < 0 \end{cases} \end{aligned}$$

where $S^{\star}_{ijk}$ is the unobserved latent variable, $\eta_{ij} = \sum_{p=0}^{P} \beta_p(t_{ij}) Z_{pi}$ is the linear predictor calculated from all the fixed effect, ϵ_{ij} are iid error terms following a logistic distribution, and u_i is the subject-specific random effect as defined in Section 4.2. In addition, the error term ϵ_{ij} and RE u_i are mutually independent. Specifically, the cumulative distribution function (cdf) for ϵ takes the form $g(x) = 1/(1 + \exp(-x))$. The calculation of marginal mean $\pi^{M}_{ij} = \mathbb{P}(\eta_{ij} + \epsilon_{ij} + u_i \geq 0)$ requires integration over the joint distribution of ϵ_{ij} and u_i, which has no closed-form solution. Instead, we can approximate the logistic cdf $g(x)$ by a normal cdf (Johnson et al., 1995, p. 119), which will lead to a more analytically tractable solution. Specifically, we have

$$g(x) \approx \Phi(cx), \text{ with } c = \sqrt{3.41}/\pi,$$

where $\Phi(x)$ is the cdf of the standard normal distribution. For any x value, the maximum absolute difference of this approximation is 0.00948.

Therefore, we can approximately view ϵ_{ij} as a normal random variable, $\epsilon_{ij} \sim N(0, 1/c^2)$. Since ϵ_{ij} and u_i are independent, we have $\epsilon_{ij} + u_i \sim N(0, 1/c^2 + \sigma_0^2)$. The marginal mean

can be thus derived as

$$\begin{aligned}\pi_{ij}^{M} &= \mathbb{P}(\epsilon_{ij}+u_i \geq -\eta_{ij}) = \mathbb{P}\left(\frac{\epsilon_{ij}+u_i}{\sqrt{1/c^2+\sigma_0^2}} \geq \frac{-\eta_{ij}}{\sqrt{1/c^2+\sigma_0^2}}\right) \\ &\approx \Phi\left(\frac{\eta_{ij}}{\sqrt{1/c^2+\sigma_0^2}}\right) \approx g\left(\frac{\eta_{ij}}{\sqrt{1+c^2\sigma_0^2}}.\right)\end{aligned}$$

Marginal variance

We will use the mixed effect model formulation in (4.2) to derive the marginal variance. Using the law of total variance, the marginal variance of S_{ij} is the sum of two parts:

$$\begin{aligned}\mathbb{V}ar(S_{ij}) &= \mathbb{E}\left\{\mathbb{V}ar(S_{ij} \mid u_i)\right\} + \mathbb{V}ar\left\{\mathbb{E}(S_{ij} \mid u_i)\right\} \\ &= \phi X_{ij}\mathbb{E}\left\{\pi_{ij}(1-\pi_{ij})\right\} + X_{ij}^2\mathbb{V}ar\left(\pi_{ij}\right), \qquad \text{(B.1)}\end{aligned}$$

where $\pi_{ij} = g(\eta_{ij}+u_i)$ is the conditional mean dependent on u_i. The exact closed-form formula does not exist for either $\mathbb{E}\left(\pi_{ij}\right)$ or $\mathbb{V}ar\left(\pi_{ij}\right)$. Nevertheless, we can work on the second-order Taylor expansion of π_{ij} around $u_i = 0$, i.e. $\pi_{ij} = g\left(\eta_{ij}+u_i\right) \approx g\left(\eta_{ij}\right) + g'\left(\eta_{ij}\right)u_i + g''\left(\eta_{ij}\right)u_i^2/2$. Thus, we have $\mathbb{E}(\pi_{ij}) \approx g(\eta_{ij}) + g''\left(\eta_{ij}\right)\sigma_0^2/2$,

$$\begin{aligned}\mathbb{V}ar\left(\pi_{ij}\right) &\approx \mathbb{E}\left\{\left[g'(\eta_{ij})u_i + \frac{g''(\eta_{ij})}{2}\left(u_i^2-\sigma_0^2\right)\right]^2\right\} \\ &= \sigma_0^2\left[g'(\eta_{ij})\right]^2 + \frac{\sigma_0^4}{2}\left[g''(\eta_{ij})\right]^2,\end{aligned}$$

and $\mathbb{E}\left(\pi_{ij}^2\right) \approx \sigma_0^2\left[g'(\eta_{ij})\right]^2 + \frac{\sigma_0^4}{2}\left[g''(\eta_{ij})\right]^2 + \left[g(\eta_{ij}) + \frac{g''t(\eta_{ij})}{2}\sigma_0^2\right]^2$. Substituting the above approximations into (B.1) yields the results in equation (4.5).

B.1.2 Appendix B: Estimate ϕ from the contaminated data

No exact expression available for the E step for ϕ

Once evaluated the integral in the quasi-deviance $d_{ij}(S_{ij}, \pi_{ij})$ (4.7), the estimating equation for ϕ takes the form

$$\begin{aligned}\nabla_\phi \text{Laplace}(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}, \boldsymbol{S}) &= \frac{1}{\phi^2}\sum_{i,j}\int_{S_{ij}/X_{ij}}^{\pi_{ij}^{(s)}} \frac{S_{ij} - X_{ij}\pi_{ij}}{\pi_{ij}(1-\pi_{ij})} d\pi_{ij} + f_2(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}) \\ &= \frac{1}{\phi^2}\sum_{i,j}\Big\{(X_{ij} - S_{ij})\log(1-\pi_{ij}^{(s)}) + S_{ij}\log(\pi_{ij}^{(s)}) \\ &\quad -(X_{ij} - S_{ij})\log(1 - S_{ij}/X_{ij}) - S_{ij}\log(S_{ij}/X_{ij})\Big\} + f_2(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}).\end{aligned}$$

This estimating equation is not linear in terms of the unknown methylated counts $\boldsymbol{S}$. Thus, replacing S_{ij} by $\eta_{ij}^\star = \mathbb{E}\left(S_{ij} \mid Y_{ij}; \boldsymbol{\mathcal{B}}^\star, \boldsymbol{\Theta}^\star\right)$ does not necessarily provide an accurate estimate for $\mathbb{E}_{\boldsymbol{S}|Y;\boldsymbol{\Theta}^\star,\boldsymbol{\mathcal{B}}^\star}(\nabla_\phi \text{Laplace}(\boldsymbol{\Theta}, \phi; \boldsymbol{\mathcal{B}}^{(s)}, \boldsymbol{S}))$, and the exact expression for this expectation is not readily available from the first two moments of the distribution of S_{ij}.

The relation between ϕ_{ij}^Y and ϕ

All the expectation and variance in this section are conditional on the values of random effects u_i. For notational simplicity, we drop u_i from all the derivations in this section.

The variance of Y_{ij} depends on its mean π_{ij}^Y as well as the joint probability $\mathbb{P}(Y_{ijk} = 1, Y_{ijk'} = 1)$, i.e. observing methylated signals at both the k^{th} and k'^{th} reads, where $k, k' = 1, 2, \ldots X_{ij}$

and $k \neq k'$:

$$\begin{aligned}
\mathbb{V}\text{ar}(Y_{ij}) &= \mathbb{E}(Y_{ij}^2) - [\mathbb{E}(Y_{ij})]^2 = \mathbb{E}\left\{\left(\sum_{k=1}^{X_{ij}} Y_{ijk}\right)^2\right\} - X_{ij}^2(\pi_{ij}^Y)^2 \\
&= \sum_{k=1}^{X_{ij}} \mathbb{E}(Y_{ijk}^2) + 2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1} \mathbb{E}(Y_{ijk}Y_{ijk'}) - X_{ij}^2(\pi_{ij}^Y)^2 \\
&= X_{ij}\pi_{ij}^Y - X_{ij}^2(\pi_{ij}^Y)^2 + 2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1} \mathbb{P}(Y_{ijk}=1, Y_{ijk'}=1). \qquad \text{(B.2)}
\end{aligned}$$

By the law of total probability, we have

$$\mathbb{P}(Y_{ijk} = Y_{ijk'} = 1) = \sum_{s_1=0}^{1}\sum_{s_2=0}^{1} \mathbb{P}(S_{ijk}=s_1, S_{ijk'}=s_2)\mathbb{P}(Y_{ijk}=Y_{ijk'}=1 \mid S_{ijk}=s_1, S_{ijk'}=s_2).$$

Joint distribution of the bivariate outcomes $(S_{ijk}, S_{ijk'})$**.** Note that, under our assumed mean-variance relationship in (4.3), S_{ijk} and $S_{ijk'}$ are not necessarily independent. Define $a_{ijkk\prime} = \mathbb{P}(S_{ijk}=1, S_{ijk'}=1)$. The joint probability mass function of $(S_{ijk}, S_{ijk'})$ can be thus written as

$$\begin{aligned}
\mathbb{P}(S_{ijk}=1, S_{ijk'}=1) &= a_{ijkk\prime} \\
\mathbb{P}(S_{ijk}=1, S_{ijk'}=0) &= \pi_{ij} - a_{ijkk\prime} \\
\mathbb{P}(S_{ijk}=0, S_{ijk'}=1) &= \pi_{ij} - a_{ijkk\prime} \\
\mathbb{P}(S_{ijk}=0, S_{ijk'}=0) &= 1 - 2\pi_{ij} + a_{ijkk\prime}.
\end{aligned}$$

We now can write the probability of observing two methylated reads as

$$\mathbb{P}(Y_{ijk}=Y_{ijk'}=1) = p_0^2(1-2\pi_{ij}+a_{ijkk\prime}) + 2p_0p_1(\pi_{ij}-a_{ijkk\prime}) + p_1^2 a_{ijkk\prime}.$$

Here, we assume that given the true methylation states S_{ijk} and $S_{ijk'}$, the observed methylation states Y_{ijk} and $Y_{ijk'}$ are independent.

Derive the values of $a_{ijkk\prime}$. From first principle, we can express the variance of $S_{ij} = \sum_{k=1}^{X_{ij}} S_{ijk}$,

$$\begin{aligned}
\mathbb{V}\text{ar}(S_{ij}) &= \sum_{k=1}^{X_{ij}} \mathbb{V}\text{ar}(S_{ijk}) + 2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1} \mathbb{C}\text{ov}(S_{ijk}, S_{ijk'}) \\
&= X_{ij}\pi_{ij}(1-\pi_{ij}) + 2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1} \mathbb{E}(S_{ijk}S_{ijk'}) - 2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1} \mathbb{E}(S_{ijk})\mathbb{E}(S_{ijk'}) \\
&= X_{ij}\pi_{ij}(1-\pi_{ij}) + 2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1} \mathbb{P}(S_{ijk}=1, S_{ijk'}=1) - X_{ij}(X_{ij}-1)\pi_{ij}^2 \\
&= X_{ij}\pi_{ij}(1-\pi_{ij}) + 2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1} a_{ijkk\prime} - X_{ij}(X_{ij}-1)\pi_{ij}^2.
\end{aligned}$$

On the other hand, we have $\mathbb{V}\text{ar}(S_{ij}) = \phi X_{ij}\pi_{ij}(1-\pi_{ij})$. Equating these two quantities gives

$$2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1} a_{ijkk\prime} = (\phi-1)X_{ij}\pi_{ij}(1-\pi_{ij}) + X_{ij}(X_{ij}-1)\pi_{ij}^2$$

Derive $\mathbb{V}\mathbf{ar}(Y_{ij})$ and ϕ_Y. Now, we can plug the expression of $\mathbb{P}(Y_{ijk} = Y_{ijk'} = 1)$ in (B.2) and write $\mathbb{V}\text{ar}(Y_{ij})$ in terms of ϕ

$$\begin{aligned}
\mathbb{V}\text{ar}(Y_{ij}) &= X_{ij}\pi_{ij}^Y - X_{ij}^2(\pi_{ij}^Y)^2 + 2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1}\left[p_0^2(1-2\pi_{ij}+a_{ijkk\prime}) + 2p_0p_1(\pi_{ij}-a_{ijkk\prime}) + p_1^2 a_{ijkk\prime}\right] \\
&= X_{ij}\pi_{ij}^Y - X_{ij}^2(\pi_{ij}^Y)^2 + X_{ij}(X_{ij}-1)\left\{p_0^2(1-2\pi_{ij}) + 2p_0p_1\pi_{ij}\right\} + 2(p_0-p_1)^2\sum_{k=1}^{X_{ij}}\sum_{k'=1}^{k-1} a_{ijkk\prime}. \\
&= X_{ij}\pi_{ij}^Y - X_{ij}^2(\pi_{ij}^Y)^2 + X_{ij}(X_{ij}-1)\left\{p_0^2(1-2\pi_{ij}) + 2p_0p_1\pi_{ij}\right\} \\
&\quad +(p_0-p_1)^2\left\{(\phi-1)X_{ij}\pi_{ij}(1-\pi_{ij}) + X_{ij}(X_{ij}-1)\pi_{ij}^2\right\} \\
&= X_{ij}\pi_{ij}^Y(1-\pi_{ij}^Y) + (p_0-p_1)^2(\phi-1)X_{ij}\pi_{ij}(1-\pi_{ij})
\end{aligned}$$

The multiplicative dispersion parameter for the mis-measured outcome Y is thus

$$\phi_{ij}^Y = \frac{\mathbb{V}\text{ar}(Y_{ij})}{X_{ij}\pi_{ij}^Y(1-\pi_{ij}^Y)} = 1 + (\phi-1)\frac{\pi_{ij}(1-\pi_{ij})}{\pi_{ij}^Y(1-\pi_{ij}^Y)}(p_0-p_1)^2.$$

Plugging in $\pi_{ij} = \dfrac{\pi_{ij}^{Y} - p_0}{p_1 - p_0}$ leads to the relation in (4.17).

B.2 Additional methods and materials

B.2.1 Existing methods used in the simulation

We compared the performance of our method with five existing methods: `BiSeq` (Hebestreit et al., 2013), `BSmooth` (Hansen et al., 2012), `SMSC` (Lakhal-Chaieb et al., 2017), `dmrseq` (Korthauer et al., 2018) and `GlobalTest` (Goeman et al., 2006), in terms of type I error and power. `BSmooth`, `SMSC`, `Biseq` are typical examples of two-stage analytic approaches. In the first stage, kernel smoothing (local likelihood estimation) is applied to the methylation data of each sample separately. In the second stage, the smoothed methylation data are further analyzed. Specifically, `BiSeq` calculates the average of Wald statistics from single-site beta regression models, while `BSmooth` and `SMSC` calculate the sum of t-statistics across loci; these statistics are used to test for differential methylation of a region. In contrast, `dmrseq` and `GlobalTest` are one-stage approaches which fit their models directly to the raw methylation proportions in a region. Specifically, `dmrseq` assesses the strength of the covariate effect using a Wald test statistic within a generalized least square regression model, while `GlobalTest` uses an improved score test in a linear regression model.

Notably, like `SOMNiBUS`, both `GlobalTest` and `BiSeq` are primarily tailored to targeted bisulfite sequencing data with previously identified regions, whereas `BSmooth`, `SMSC` and `dmrseq` are designed for WGBS data. Specifically, `BSmooth` and `SMSC` define DMRs at adjacent CpG sites with absolute t-statistics above a defined threshold. The final product from the original software of `BSmooth` is a list of DMRs that are ranked by the sum of t-statistics; however, `BSmooth` does not provide region-based p-values. To allow comparisons with `SOMNiBUS`, we estimated the empirical regional p-values for `BSmooth` by permuting the values of the covari-

ate of interest 1000 times. When analyzing WGBS data, `dmrseq` first constructs candidate regions based on a user-defined cutoff of the smoothed methylation proportion differences, and then fits a generalized least squares regression model with autoregressive error structure to the transformed methylation proportions. Furthermore, the inference inside `dmrseq` is drawn from permutations – its approximate null distribution is generated by pooling a set of region-level statistics of many candidate regions from all permutations. To better adapt `dmrseq` to a single targeted region: i) we used a small cutoff of methylation differences (10^{-5}) for detecting candidate (sub)regions, which ensures that most CpGs are retained; ii) we applied a relatively large number of permutations ($B = 500$) to generate a null distribution of test statistics; iii) we reported the raw p-values without the multiplicity corrections. Note that in some simulations, `dmrseq` reported more than one DMR in the region. Therefore, for a fairer comparison, we calculated the `dmrseq`'s p-value as the minimum over the reported chunks' p-values. Among the five competitive methods, `dmrseq`, `GlobalTest` and `BiSeq` allow adjustment for multiple covariates. `SMSC` is the only approach accounting for experimental errors; however, it is conceptually restricted to data from a single cell type.

B.3 Additional data example results

	data 1 (N =116)	data 2 (N = 102)
ACPA Positives	55	48
ACPA Negatives	61	54
Number of targeted regions (with at least 50 CpGs)	10,759	12,985

Table B.1: Sample characteristics in dataset 1 and 2.

B.4 Additional simulation results

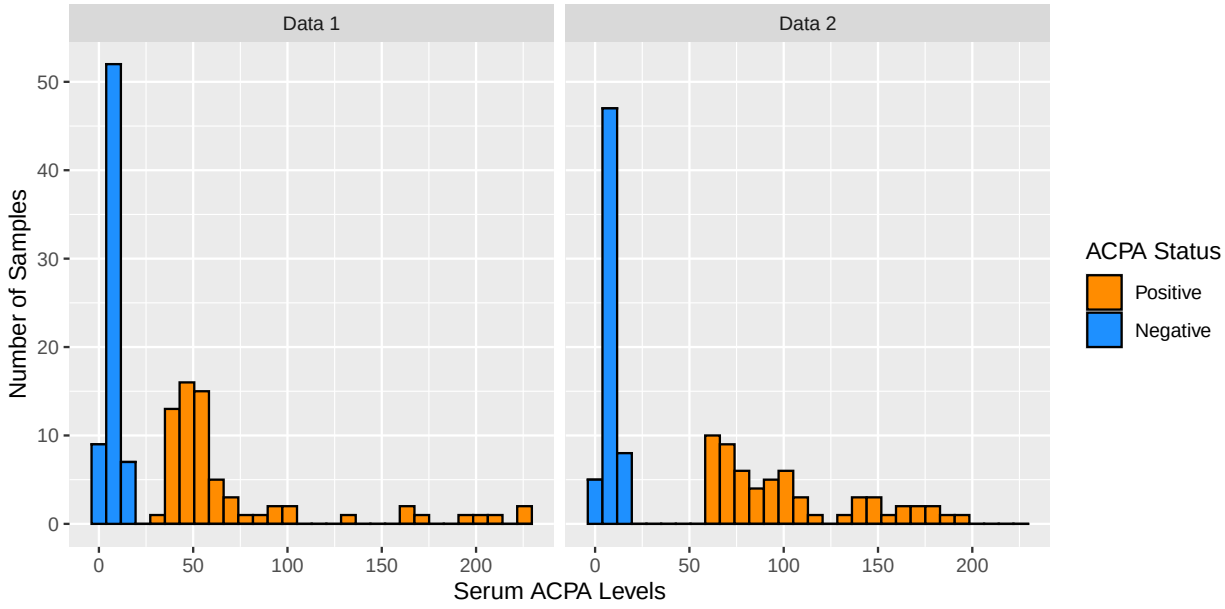


Figure B.1: Distribution of ACPA levels in dataset 1 and 2.

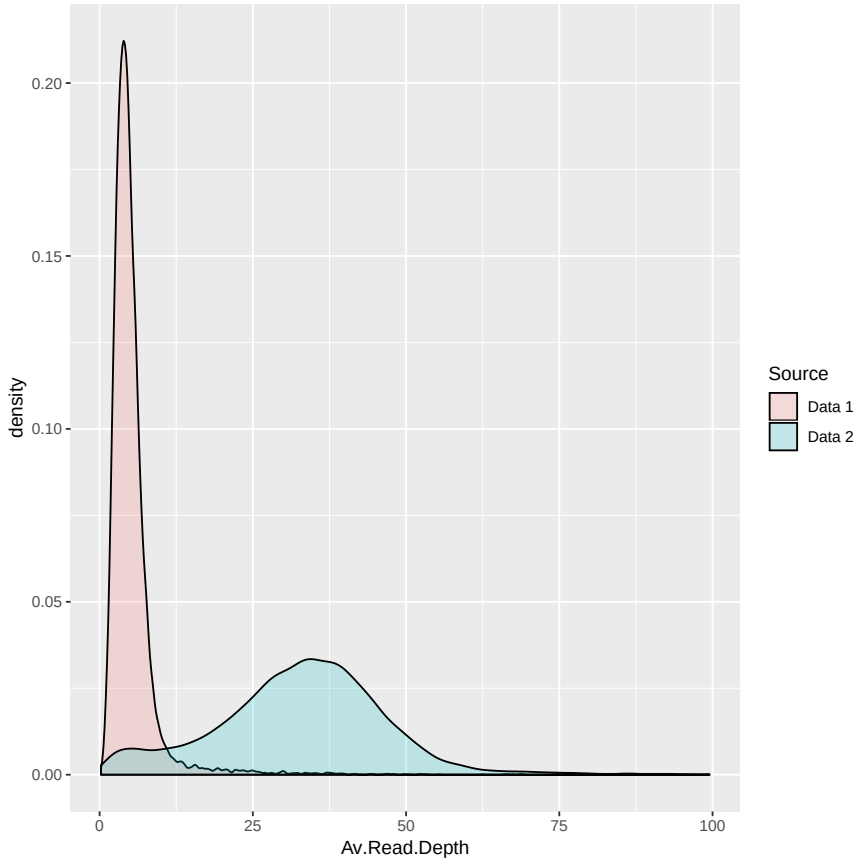


Figure B.2: Distribution of average read depth in all targeted regions in data 1 and data 2

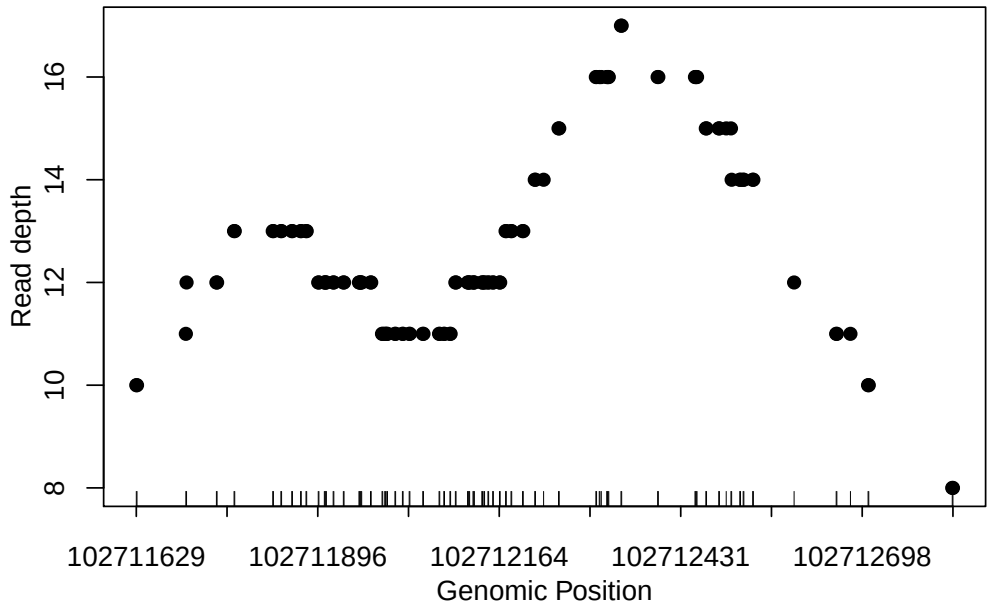


Figure B.3: The read-depth pattern used in the simulation. Median read depths were calculated for one targeted region that underwent bisulfite sequencing in a dataset described in (Zhao et al., 2020, Section 3). For the simulation, we then fit a cubic spline with 10 knots to the median read depths.

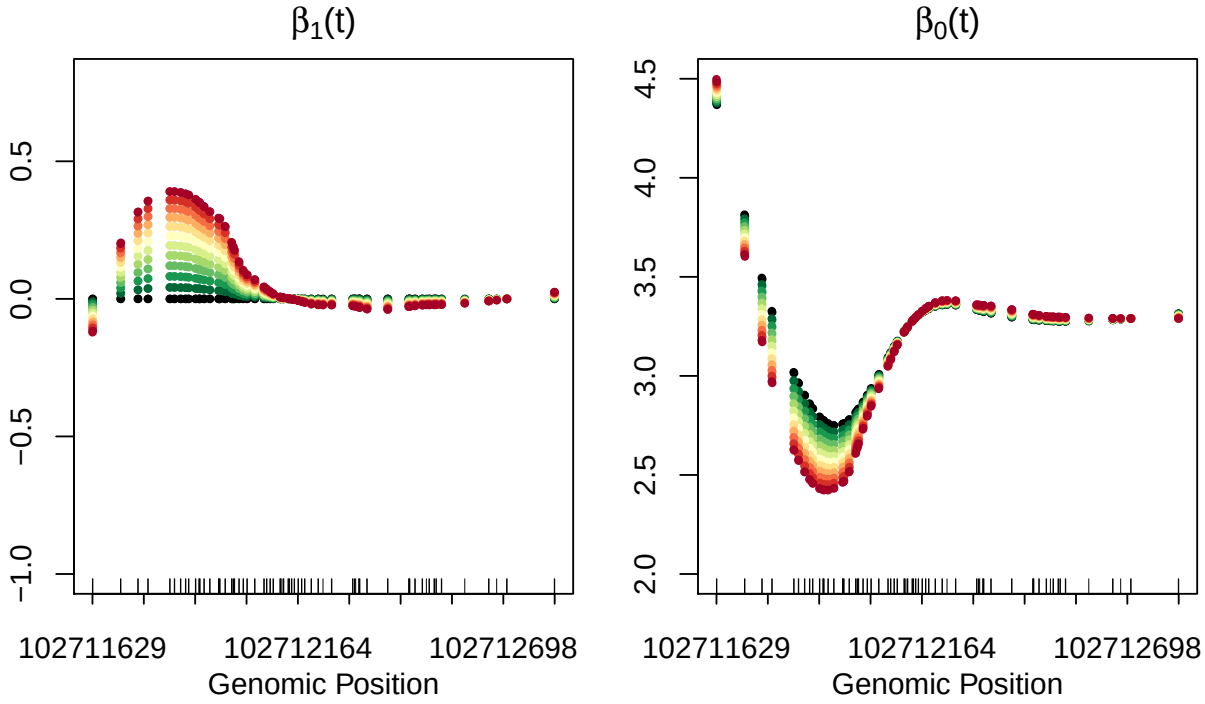


Figure B.4: The 15 simulation settings of functional parameters $\beta_0(t)$ and $\beta_1(t)$ in Scenario 2, which correspond to the 15 settings for $\pi_0(t)$ shown in Figure 6 in the main manuscript.

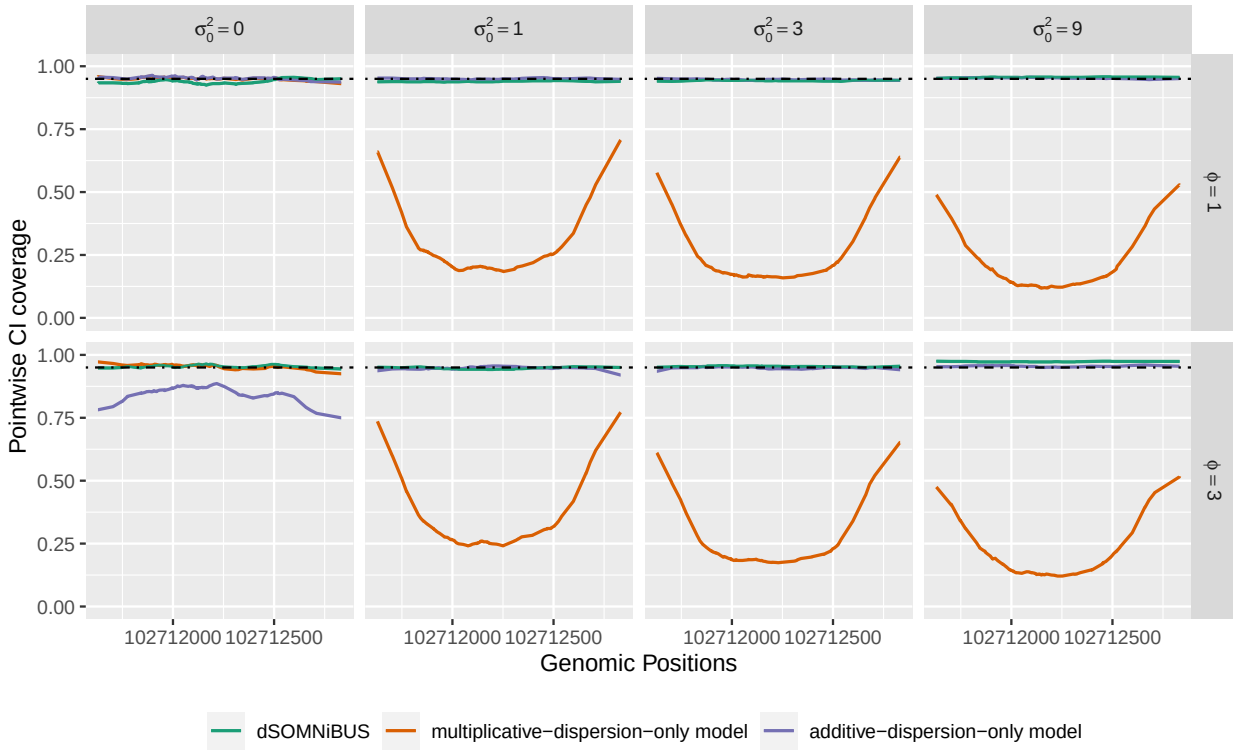


Figure B.5: Empirical coverage probability of the analytical 95% pointwise CIs for $\beta_3(t)$ over 1000 simulations, under different vales of ϕ and σ_0^2. The empirical coverage probabilities are defined as the percentage of simulations where the analytical CIs cover the true value of $\beta_3(t)$. Data were generated without error, under simulation Scenario 1. The results from dSOMNiBUS and the additive-dispersion-only model are indistinguishable in all settings but $\sigma_0^2 = 0$ and $\phi = 3$.

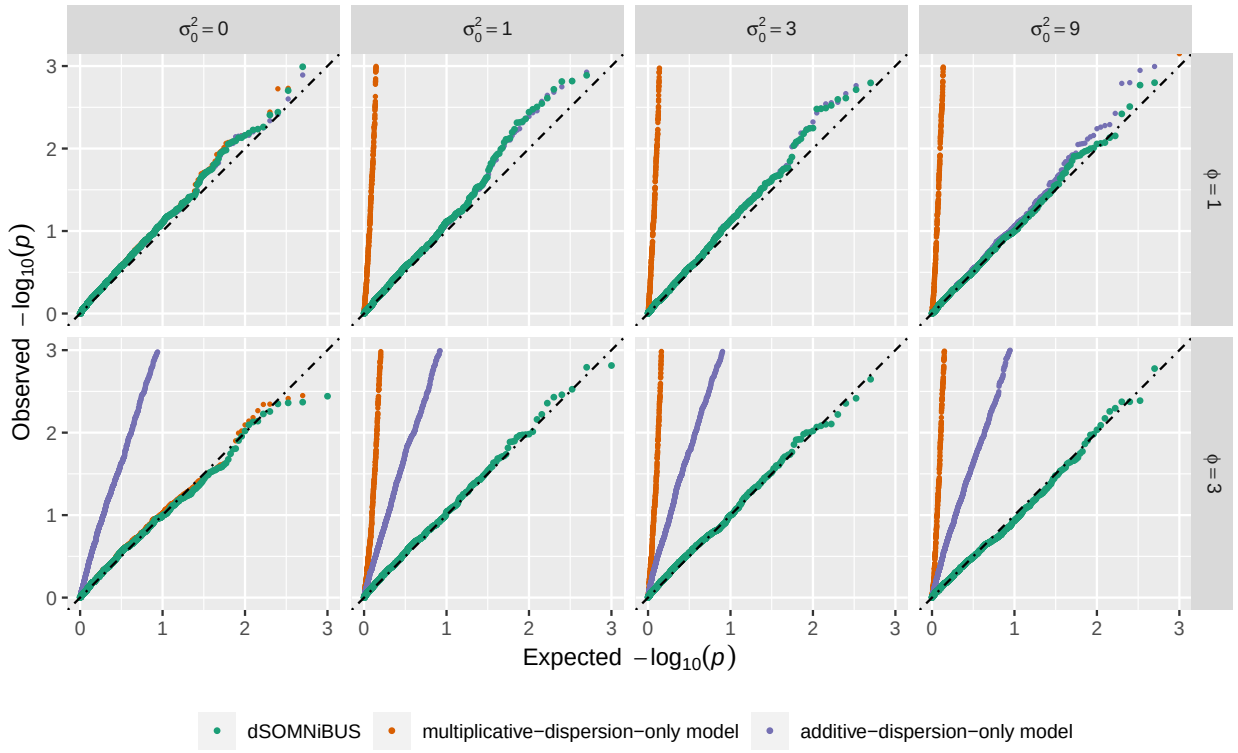


Figure B.6: QQ plot for regional p-values for the test $H_0 : \beta_3(t) = 0$, obtained from dSOMNiBUS, the multiplicative-dispersion-only model and the additive-dispersion-only model. Data were simulated without error, under simulation Scenario 1. When $\phi = 1$, the results from dSOMNiBUS and the additive-dispersion-only model are indistinguishable. When $\sigma_0^2 = 0$, the lines for the multiplicative-dispersion-only model and dSOMNiBUS are indistinguishable.

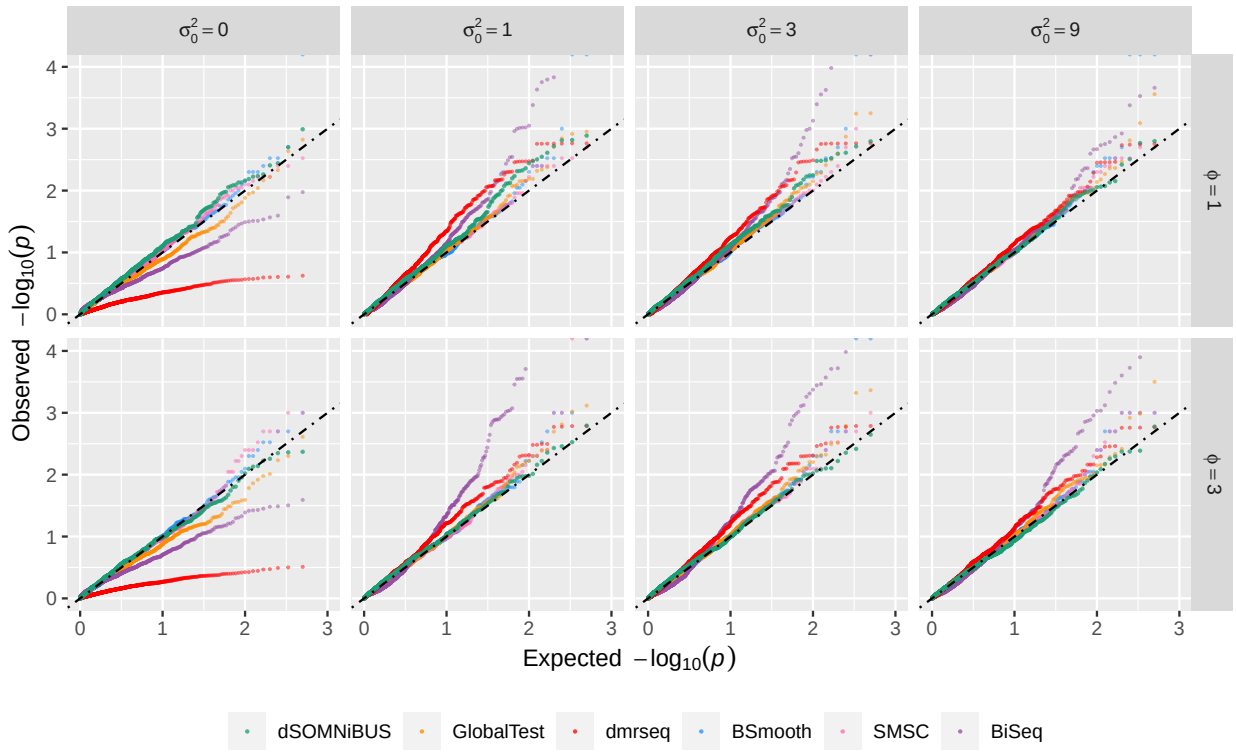


Figure B.7: QQ plot for regional p-values for the test $H_0 : \beta_3(t) = 0$, obtained from dSOMNiBUS, GlobalTest, dmrseq, BSmooth, SMSC, and BiSeq. Data were simulated without error, under simulation Scenario 1, and ϕ was estimated using the moment-based estimator.

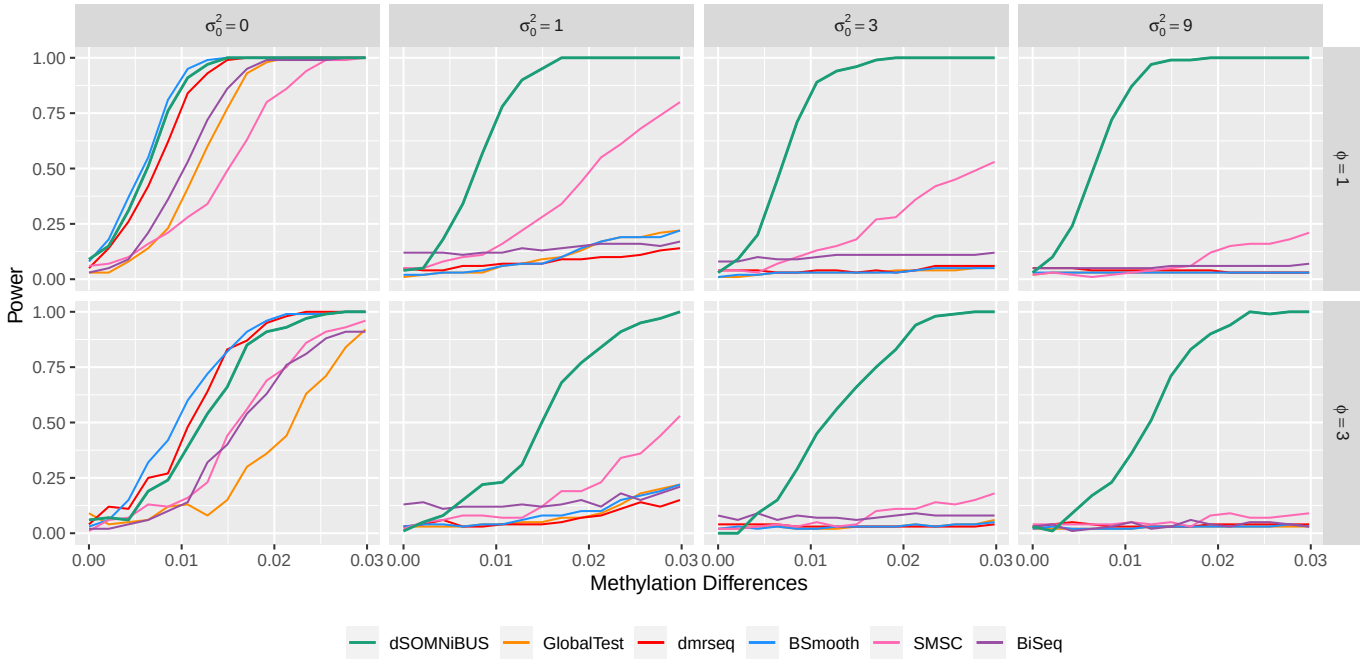


Figure B.8: Powers to detect DMRs using the six methods for the 14 simulation settings in Scenario 2 under different levels of maximum methylation differences between $\pi_0(t)$ and $\pi_1(t)$ in the region, calculated over 100 simulations. Data were simulated without error, under simulation Scenario 1, and ϕ was estimated using the moment-based estimator.

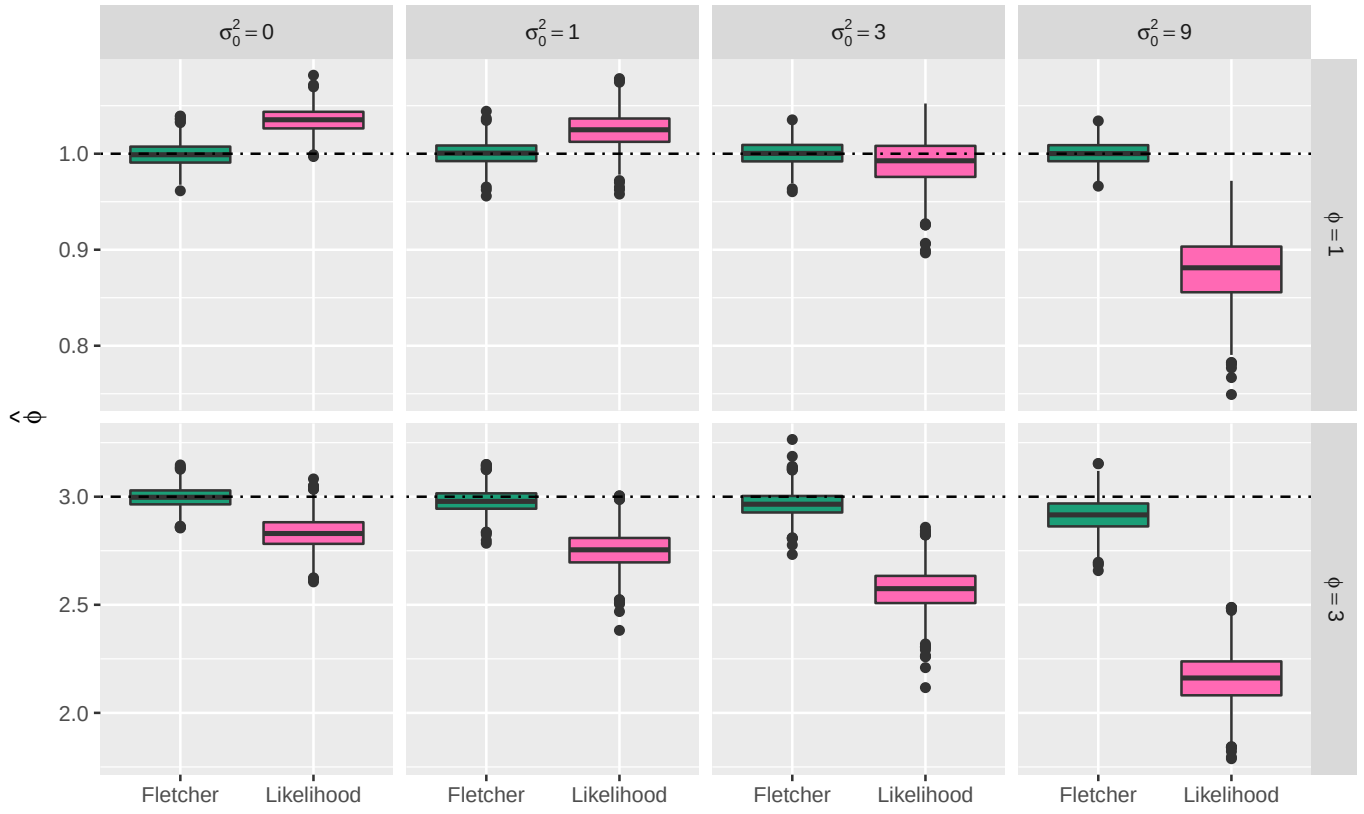


Figure B.9: Moment-based ($\widehat{\phi}_{Fle}$) and likelihood-based ($\widehat{\phi}_{Lik}$) estimates of the multiplicative dispersion parameter ϕ. Data were simulated without error, under simulation Scenario 1. There is less bias in $\widehat{\phi}_{Fle}$ than $\widehat{\phi}_{Lik}$.

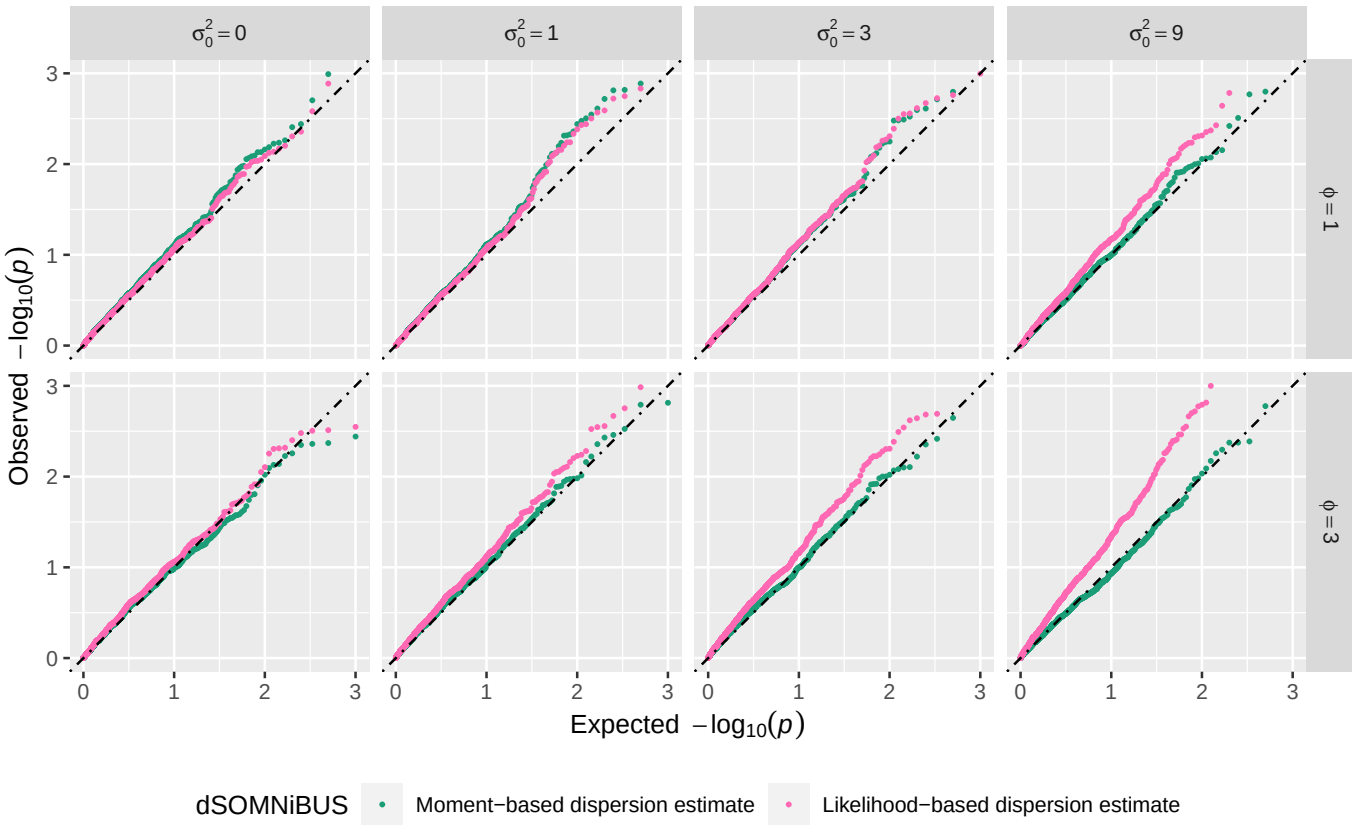


Figure B.10: QQ plot for regional p-values for the test $H_0 : \beta_3(t) = 0$, obtained from dSOMNiBUS using the moment-based dispersion estimator $\widehat{\phi}_{Fle}$ and the likelihood-based dispersion estimator $\widehat{\phi}_{Lik}$. Data were simulated without error, under simulation Scenario 1.

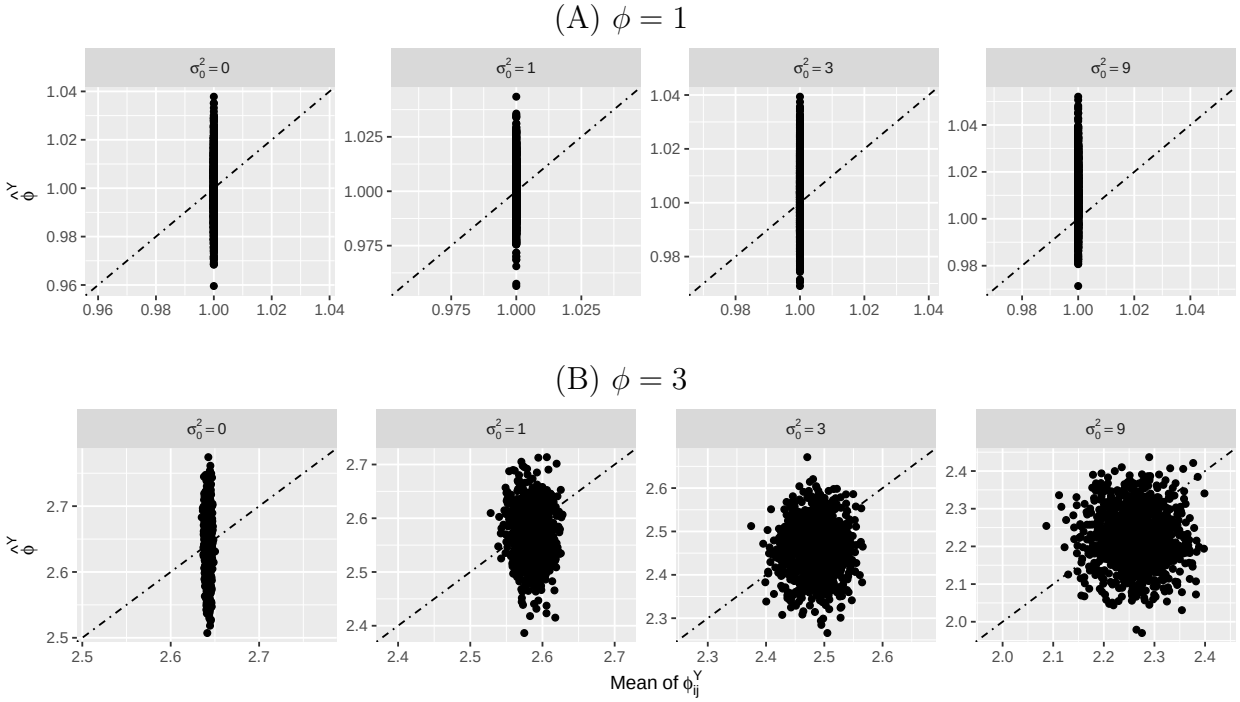


Figure B.11: Scatter plots of the estimated constant dispersion $\widehat{\phi}^Y$ and the mean of the truth of individual dispersion ϕ_{ij}^Y. Data were generated with errors $p_0 = 0.003$ and $1 - p_1 = 0.1$, and $\phi = 1$ (A) or $\phi = 3$ (B). Here $\widehat{\phi}^Y$ denotes the estimated dispersion parameter when ignoring the presence of error, and individual ϕ_{ij}^Ys are calculated from equation (4.17) using the true π_{ij}, ϕ, p_0 and p_1 that were used to simulate the data. $\widehat{\phi}^Y$ can be roughly viewed as an estimate of the average of individual dispersion ϕ_{ij}^Y.

Appendix C

Supporting Information for Chapter 5

This Supporting Information includes detailed derivations, proofs, and additional simulation results.

C.1 Natural cubic spline and its sparsity-penalty matrix $\mathbf{\Omega^{(1)}}$

Splines are polynomial pieces jointed at certain values (i.e. knots). Cubic spline is the most commonly used one, which is represented by piecewise cubic polynomial with continuous first and second derivatives at the knots. To avoid erratic behaviors (high variance) of cubic fit near the boundaries, a natural cubic spline adds additional constrains that the function is linear beyond the two end-points. In this work, we use natural cubic regression spline to represent the functional parameters $\beta_p(t_{ij})$. Without loss of generality, we drop the subscript p from the notations here and consider defining a natural cubic spline function $\beta(t)$, with K given knots, $t_1, t_2, \ldots t_K$.

There are many equivalent bases definitions that can be used to expand the cubic spline

$\beta(t)$. We adopt the basis used in the R package `mgcv` (Wood, 2017), where the spline is parameterized in terms of its values at the knots. One advantage of this basis definition is that it does not require any re-scaling of the predictor variable t. Let $\theta_j = \beta(t_j)$ and $\delta_j = \beta''(t_j)$, $j = 1, 2, \ldots K$. Then the spline $\beta(t)$ can be written as

$$\beta(t) = a_j^-(t)\theta_j + a_j^+(t)\theta_{j+1} + c_j^-(t)\delta_j + c_j^+(t)\delta_{j+1}, \text{ if } t_j \leq t \leq t_{j+1}. \tag{C.1}$$

Let $h_j = t_{j+1} - t_j$, the 'basis' functions a_j^-, a_j^+, c_j^- and c_j^+ in (C.1) are defined as

$$\begin{aligned} a_j^-(t) &= \frac{t_{j+1} - t}{h_j}, \quad c_j^-(t) = \frac{1}{6}\left[\frac{(t_{j+1} - t)^3}{h_j} - h_j(t_{j+1} - t)\right], \\ a_j^+(t) &= \frac{t - t_j}{h_j}, \quad c_j^+(t) = \frac{1}{6}\left[\frac{(t - t_j)^3}{h_j} - h_j(t - t_j)\right]. \end{aligned} \tag{C.2}$$

As explained in Section 2.3, a core part in the sparsity-penalty is the squared L2-norm of the function $\beta(t)$, which can be written as a quadratic form in terms of $\boldsymbol{\theta}$ and penalty matrix $\boldsymbol{\Omega^{(1)}}$,

$$\int_{t_1}^{t_K} (\beta(t))^2 \, dt = \boldsymbol{\theta}^T \boldsymbol{\Omega^{(1)}} \boldsymbol{\theta}.$$

In the rest of this section, $\boldsymbol{\Omega^{(1)}}$ is derived under the basis representation in (C.1).

C.1.1 Relation between spline values $\boldsymbol{\theta}$ and their second derivatives $\boldsymbol{\delta}$

The conditions that the spline should be continuous to second derivatives, at each interior knots t_j, and have zero second derivative at t_1 and t_K imply a deterministic relation between function values $\boldsymbol{\theta} = (\theta_1, \ldots \theta_K)$ and second derivatives $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$,

$$\boldsymbol{\delta} = \boldsymbol{F}\boldsymbol{\theta}. \tag{C.3}$$

The mapping matrix $\boldsymbol{F} \in \mathcal{R}^{K\times K}$ takes the form

$$\boldsymbol{F} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{B^{-1}D} \\ \mathbf{0} \end{bmatrix}, \tag{C.4}$$

where matrices $\boldsymbol{B} \in \mathcal{R}^{(K-2)\times(K-2)}$ and $\boldsymbol{D} \in \mathcal{R}^{(K-2)\times K}$ have non-zero elements,

$$\begin{aligned} D_{i,i} &= \frac{1}{h_i}, \qquad D_{i,i+1} = -\frac{1}{h_i} - \frac{1}{h_{i+1}}, \qquad D_{i,i+2} = \frac{1}{h_{i+1}} \\ B_{i,i} &= \frac{h_i + h_{i+1}}{3}, \qquad i = 1, \dots k-2 \\ B_{i,i+1} &= \frac{h_{i+1}}{6} \qquad B_{i+1,i} = \frac{h_{i+1}}{6} \qquad i = 1, \dots k-3. \end{aligned}$$

The detailed derivation for (C.4) can be found in (Wood, 2017, Section 5.3.1). Thus, the expansion in (C.1) can be rewritten entirely in terms of $\boldsymbol{\theta}$ as

$$\beta(t) = a_j^-(t)\theta_j + a_j^+(t)\theta_{j+1} + c_j^-(t)\boldsymbol{F_j\theta} + c_j^+(t)\boldsymbol{F_{j+1}\theta}, \text{ if } t_j \le t \le t_{j+1},$$

where $\boldsymbol{F_j}$ is the j^{th} row of matrix $\boldsymbol{F}$. The expansion can be further expressed in a more compact way, $\beta(t) = \sum_{i=1}^{K} b_i(t)\theta_i$, where basis functions $b_i(t)$ are

$$b_i(t) = \begin{cases} a_{i-1}^+(t) + c_{i-1}^-(t)F_{i-1,i} + c_{i-1}^+(t)F_{i,i} & \text{if} \quad t_{i-1} \le t \le t_i \\ a_i^-(t) + c_i^-(t)F_{i,i} + c_i^+F_{i+1,i} & \text{if} \quad t_i \le t \le t_{i+1} \\ c_k^-(t)F_{k,i} + c_k^+F_{k+1,i} & \text{if} \quad t_k \le t \le t_{k+1}, \text{ and } k \ne i \text{ or } i-1. \end{cases} \tag{C.5}$$

Writing $\boldsymbol{b}(t)$ as the vector with i$^{\text{th}}$ element $b_i(t)$, it is easy to show that

$$\int_{t_1}^{t_K} (\beta(t))^2\, dt = \boldsymbol{\theta}^T \left\{ \int_{t_1}^{t_K} \boldsymbol{b}(t)\boldsymbol{b}(t)^T dt \right\} \boldsymbol{\theta},$$

which immediately implies $\boldsymbol{\Omega^{(1)}} = \int_{t_1}^{t_K} \boldsymbol{b}(t)\boldsymbol{b}(t)^T dt$. However, it is quite complicated, although possible, to evaluate this integral analytically, because $b_i(t)$ takes non-zero values in each of the intervals between two knots. To mitigate the problem, instead of working on the expansion solely in terms of $\boldsymbol{\theta}$, we seek to calculate the integral based on the expansion in terms of both $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$, i.e. the expression in (C.1). Then, we can determine the sparsity penalty matrix $\boldsymbol{\Omega^{(1)}}$ by transforming $\boldsymbol{\delta}$ back to $\boldsymbol{\theta}$.

C.1.2 L2-norm of a natural cubic spline

Theorem 3. *Suppose $\beta(t)$ is a natural cubic spline function, with K knots at $t_1, t_2, \ldots t_K$, and basis expansion defined in (C.1). We have the following result for the L2-norm of $\beta(t)$,*

$$\|\beta(t)\|_2^2 = \int_{t_1}^{t_K} (\beta(t))^2 \, dt = \boldsymbol{\theta}^T \boldsymbol{\Omega^{(1)}} \boldsymbol{\theta}.$$

Here $\boldsymbol{\theta}$ is a vector with basis coefficients for $\theta_j = \beta(t_j)$, for $j = 1, 2, \ldots K$ and $\boldsymbol{\Omega^{(1)}}$ takes the form

$$\boldsymbol{\Omega^{(1)}} = \boldsymbol{A}_{11} + \boldsymbol{F}^T \boldsymbol{A_{12}}^T + \boldsymbol{A}_{12}\boldsymbol{F} + \boldsymbol{F}^T \boldsymbol{A}_{22} \boldsymbol{F}.$$

Specifically, $\boldsymbol{A}_{11}, \boldsymbol{A}_{12}, \boldsymbol{A}_{22} \in \mathbb{R}^{K \times K}$ are tri-diagonal matrices with elements defined in Table C.1, and $\boldsymbol{F}$ is the matrix mapping the function values at the knots onto their second derivatives, as given in (C.4).

Proof. The basis expansion of $\beta(t)$ in (C.1) can be re-expressed as

$$\beta(t) = \sum_{i=1}^{K} d_i(t)\theta_i + \sum_{i=1}^{K} e_i(t)\delta_i,$$

where the sets of 'basis' functions $d_i(t)$ and $e_i(t)$ are defined in the Table C.2

Evaluating the integral involving $d_i(t)$ and $e_i(t)$ is much easier compared to evaluating $b_i(t)$,

Table C.1: Elements in the tri-diagonal matrices $\boldsymbol{A}_{11}, \boldsymbol{A}_{12}, \boldsymbol{A}_{22} \in \mathbb{R}^K$, which are used to define the L2-norm of natural cubic spline. $h_i = t_{i+1} - t_i$.

	$(1,1)$	(i,i), $i = 2,3,\ldots K-1)$	(K,K)	$(i,i-1)$ and $(i-1,i)$
$\boldsymbol{A}_{11}$	$\frac{h_1}{3}$	$\frac{h_{i-1}}{3} + \frac{h_i}{3}$	$\frac{h_{K-1}}{3}$	$\frac{h_{i-1}}{6}$
$\boldsymbol{A}_{12}$	$-\frac{h_1^3}{45}$	$-\frac{h_{i-1}^3}{45} - \frac{h_i^3}{45}$	$-\frac{h_{K-1}^3}{45}$	$-\frac{7}{360}h_{i-1}^3$
$\boldsymbol{A}_{22}$	$\frac{4}{315}h_1^5$	$\frac{4}{315}\left(h_{i-1}^5 + h_i^5\right)$	$\frac{4}{315}h_{K-1}^5$	$\frac{31}{15120}h_{i-1}^5$

Table C.2: Definitions of basis functions $d_i(t)$ and $e_i(t)$ used to define a natural cubic regression spline $\beta(t)$.

	$i = 1$	$i = 2,3,\ldots K-1$	$i = K$
$d_i(t) =$	$a_1^-(t)\mathbb{1}(t_1 \le t \le t_2)$	$a_{i-1}^+(t)$ if $t_{i-1} \le t \le t_i$ $a_i^-(t)$ if $t_i \le t \le t_{i+1}$ 0 otherwise	$a_{K-1}^+(t)\mathbb{1}(t_{K-1} \le t \le t_K)$
$e_i(t) =$	$c_1^-(t)\mathbb{1}(t_1 \le t \le t_2)$	$c_{i-1}^+(t)$ if $t_{i-1} \le t \le t_i$ $c_i^-(t)$ if $t_i \le t \le t_{i+1}$ 0 otherwise	$c_{K-1}^+(t)\mathbb{1}(t_{K-1} \le t \le t_K)$

because $d_i(t)$ and $e_i(t)$ are non-zero over no more than 2 consecutive intervals.

Concatenate the coefficients vectors $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ into a $2K-$ dimensional vector $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\delta})^T$ and define $\boldsymbol{q}(t)$ as the basis vector joining $d_i(t)$ and $e_i(t)$, $\boldsymbol{q}(t) = (d_1(t), \ldots d_K(t), e_1(t), \ldots e_K(t))^T$. We can thus rewrite the L2-norm of $\beta(t)$ as

$$\int_{t_1}^{t_K} \left(\beta(t)\right)^2 dt = \boldsymbol{\alpha}^T \int_{t_1}^{t_K} \boldsymbol{q}(t)\boldsymbol{q}(t)^T dt \; \boldsymbol{\alpha}. \tag{C.6}$$

It is clear that $\int_{t_1}^{t_K} \boldsymbol{q}(t)\boldsymbol{q}(t)^T dt$ is symmetric, by construction, and consists of four blocks,

$\boldsymbol{A}_{11}, \boldsymbol{A}_{12}, \boldsymbol{A}_{12}^T$ and $\boldsymbol{A}_{22}$, as defined below,

$$\int_{t_1}^{t_K} \boldsymbol{q}(t)\boldsymbol{q}(t)^T dt = \begin{bmatrix} \int_{t_1}^{t_K} \boldsymbol{d}(t)\boldsymbol{d}(t)^T dt & \int_{t_1}^{t_K} \boldsymbol{d}(t)\boldsymbol{e}(t)^T dt \\ \int_{t_1}^{t_K} \boldsymbol{e}(t)\boldsymbol{d}(t)^T dt & \int_{t_1}^{t_K} \boldsymbol{e}(t)\boldsymbol{e}(t)^T dt \end{bmatrix} := \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{12}^T & \boldsymbol{A}_{22} \end{bmatrix}.$$

Calculating $\boldsymbol{A}_{11}$ $\boldsymbol{A}_{11}$ is tri-diagonal because each $d_i(t)$ is non-zero over only 2 intervals. The i^{th} leading diagonal element, for $i = 2, \ldots K-1$, is given by

$$\begin{aligned} [\boldsymbol{A}_{11}]_{i,i} = \int_{t_1}^{t_K} d_i(t)^2 dt &= \int_{t_{i-1}}^{t_i} \left(\frac{t - t_{i-1}}{h_{i-1}}\right)^2 dt + \int_{t_i}^{t_{i+1}} \left(\frac{t_{i+1} - t}{h_i}\right)^2 dt \\ &= \frac{(t - t_{i-1})^3}{3h_{i-1}^2}\Big|_{t_{i-1}}^{t_i} - \frac{(t_{i+1} - t)^3}{3h_i^2}\Big|_{t_i}^{t_{i+1}} = \frac{h_{i-1}}{3} + \frac{h_i}{3}. \end{aligned}$$

The first and last leading diagonal elements are

$$[\boldsymbol{A}_{11}]_{1,1} = \frac{h_1}{3} \text{ and } [\boldsymbol{A}_{11}]_{K,K} = \frac{h_{K-1}}{3}.$$

Similarly, the off-diagonal elements $[\boldsymbol{A}_{11}]_{(i-1,i)}$ and $[\boldsymbol{A}_{11}]_{(i,i-1)}$, where $i = 2, \ldots K$, are given by

$$\int_{t_1}^{t_K} d_i(t) d_{i-1}(t) dt = \int_{t_{i-1}}^{t_i} a_{i-1}^+(t) a_{i-1}^-(t) dt = \int_{t_{i-1}}^{t_i} \left(\frac{t - t_{i-1}}{h_{i-1}}\right)\left(\frac{t_i - t}{h_{i-1}}\right) dt = \frac{h_{i-1}}{6}.$$

Calculating $\boldsymbol{A}_{12}$ $\boldsymbol{A}_{12}$ is also tri-diagonal because both $d_i(t)$ and $e_i(t)$ are non-zero over only 2 intervals. The i^{th} leading diagonal element, for $i = 2, \ldots K-1$, is given by

$$\begin{aligned}
[\boldsymbol{A}_{12}]_{i,i} = \int_{t_1}^{t_K} d_i(t)e_i(t)dt &= \int_{t_{i-1}}^{t_i} a_{i-1}^{+}(t)c_{i-1}^{+}(t)dt + \int_{t_i}^{t_{i+1}} a_i^{-}(t)c_i^{-}(t)dt \\
&= \int_{t_{i-1}}^{t_i} \left(\frac{t-t_{i-1}}{h_{i-1}}\right)\frac{1}{6}\left[\frac{(t-t_{i-1})^3}{h_{i-1}} - h_{i-1}(t-t_{i-1})\right]dt \\
&\quad + \int_{t_i}^{t_{i+1}} \left(\frac{t_{i+1}-t}{h_i}\right)\frac{1}{6}\left[\frac{(t_{i+1}-t)^3}{h_i} - h_i(t_{i+1}-t)\right]dt \\
&= -\frac{h_{i-1}^3}{45} - \frac{h_i^3}{45}
\end{aligned}$$

The first and last leading diagonal elements of $\boldsymbol{A}_{12}$ are given by

$$[\boldsymbol{A}_{12}]_{1,1} = -\frac{h_1^3}{45} \text{ and } [\boldsymbol{A}_{12}]_{K,K} = -\frac{h_{K-1}^3}{45}.$$

Similarly, the off-diagonal elements $[\boldsymbol{A}_{12}]_{(i-1,i)}$ and $[\boldsymbol{A}_{12}]_{(i,i-1)}$, where $i = 2, \ldots K$, where $i = 2, \ldots K$ can be obtained as

$$\begin{aligned}
[\boldsymbol{A}_{12}]_{(i,i-1)} &= \int_{t_1}^{t_K} d_i(t)e_{i-1}(t)dt = \int_{t_{i-1}}^{t_i} a_{i-1}^{+}(t)c_{i-1}^{-}(t)dt + \\
&= \int_{t_{i-1}}^{t_i} \left(\frac{t-t_{i-1}}{h_{i-1}}\right)\frac{1}{6}\left[\frac{(t_i-t)^3}{h_{i-1}} - h_{i-1}(t_i-t)\right]dt \\
&= \int_{t_{i-1}}^{t_i} \frac{(t-t_{i-1})(t_i-t)^3}{6h_{i-1}^2}dt - \int_{t_{i-1}}^{t_i} \frac{(t-t_{i-1})(t_i-t)}{6}dt \\
&= \frac{h_{i-1}^3}{120} - \frac{h_{i-1}^3}{36} = -\frac{7}{360}h_{i-1}^3
\end{aligned}$$

$$
\begin{aligned}
[\boldsymbol{A}_{12}]_{(i-1,i)} &= \int_{t_1}^{t_K} d_{i-1}(t)e_i(t)dt = \int_{t_{i-1}}^{t_i} a_{i-1}^{-}(t)c_{i-1}^{+}(t)dt + \\
&= \int_{t_{i-1}}^{t_i} \left(\frac{t_i - t}{h_{i-1}}\right) \frac{1}{6}\left[\frac{(t-t_{i-1})^3}{h_{i-1}} - h_{i-1}(t-t_{i-1})\right] dt \\
&= \int_{t_{i-1}}^{t_i} \frac{(t-t_{i-1})^3\,(t_i-t)}{6h_{i-1}^2}dt - \int_{t_{i-1}}^{t_i} \frac{(t-t_{i-1})\,(t_i-t)}{6}dt \\
&= \frac{h_{i-1}^3}{120} - \frac{h_{i-1}^3}{36} = -\frac{7}{360}h_{i-1}^3
\end{aligned}
$$

Calculating $\boldsymbol{A}_{22}$ The i^{th} leading diagonal element of $\boldsymbol{A}_{22}$, for $i = 2, \ldots K-1$, is given by

$$
\begin{aligned}
[\boldsymbol{A}_{22}]_{i,i} = \int_{t_1}^{t_K} e_i(t)^2 dt &= \int_{t_{i-1}}^{t_i} c_{i-1}^{+}(t)^2 dt + \int_{t_i}^{t_{i+1}} c_i^{-}(t)^2 dt \\
&= \int_{t_{i-1}}^{t_i} \left\{\frac{1}{6}\left[\frac{(t-t_{i-1})^3}{h_{i-1}} - h_{i-1}(t-t_{i-1})\right]\right\}^2 dt \\
&\quad + \int_{t_i}^{t_{i+1}} \left\{\frac{1}{6}\left[\frac{(t_{i+1}-t)^3}{h_i} - h_i(t_{i+1}-t)\right]\right\}^2 dt \\
&= \frac{4}{315}\left(h_{i-1}^5 + h_i^5\right)
\end{aligned}
$$

The first and last leading diagonal elements of $\boldsymbol{A}_{12}$ are given by

$$
[\boldsymbol{A}_{12}]_{1,1} = \frac{4}{315}h_1^5 \text{ and } [\boldsymbol{A}_{12}]_{K,K} = \frac{4}{315}h_{K-1}^5.
$$

Similarly, the off-diagonal elements $[\boldsymbol{A}_{22}]_{(i-1,i)}$ and $[\boldsymbol{A}_{22}]_{(i,i-1)}$, where $i = 2, \dots K$, where $i = 2, \dots K$ can be obtained as

$$\begin{aligned}
[\boldsymbol{A}_{22}]_{(i,i-1)} &= [\boldsymbol{A}_{22}]_{(i-1,i)} = \int_{t_1}^{t_K} e_i(t) e_{i-1}(t) dt \\
&= \int_{t_{i-1}}^{t_i} c_{i-1}^{+}(t) c_{i-1}^{-}(t) dt + \\
&= \int_{t_{i-1}}^{t_i} \frac{1}{6} \left[\frac{(t - t_{i-1})^3}{h_{i-1}} - h_{i-1}(t - t_{i-1}) \right] \frac{1}{6} \left[\frac{(t_i - t)^3}{h_{i-1}} - h_{i-1}(t_i - t) \right] dt \\
&= \frac{31}{15120} h_{i-1}^5
\end{aligned}$$

Re-express the L2-norm of $\beta(t)$ Using the relation between $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$, $\boldsymbol{\delta} = \boldsymbol{F}\boldsymbol{\theta}$, we can re-express the L2-norm in (C.6) as

$$\begin{aligned}
\int_{t_1}^{t_K} (\beta(t))^2 \, dt &= (\boldsymbol{\theta}^T, \boldsymbol{\theta}^T \boldsymbol{F}^T) \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{12}^T & \boldsymbol{A}_{22} \end{bmatrix} (\boldsymbol{\theta}, \boldsymbol{F}\boldsymbol{\theta}) \\
&= \boldsymbol{\theta}^T \left(\boldsymbol{A}_{11} + \boldsymbol{F}^T \boldsymbol{A}_{\mathbf{12}}{}^T + \boldsymbol{A}_{12}\boldsymbol{F} + \boldsymbol{F}^T \boldsymbol{A}_{22} \boldsymbol{F} \right) \boldsymbol{\theta}.
\end{aligned}$$

Therefore, the sparsity-penalty matrix $\boldsymbol{\Omega}^{(\mathbf{1})}$ is given by $\boldsymbol{A}_{11} + \boldsymbol{F}^T \boldsymbol{A}_{\mathbf{12}}{}^T + \boldsymbol{A}_{12}\boldsymbol{F} + \boldsymbol{F}^T \boldsymbol{A}_{22} \boldsymbol{F}$. This completes the proof. □

C.1.3 Natural cubic spline and its smoothness-penalty matrix $\boldsymbol{\Omega}^{(\mathbf{2})}$

As derived in (Wood, 2017, Section 5.3.1), the unscaled smoothness-penalty matrix takes the form

$$\boldsymbol{\Omega}^{(\mathbf{2})}/M^2 = \boldsymbol{D}^{\boldsymbol{T}} \boldsymbol{B}^{-\mathbf{1}} \boldsymbol{D},$$

which can be readily extracted using `smooth.construct.cr()` from `mgcv` package.

C.2 Subgradient of $h(\boldsymbol{\theta}) = \sqrt{\boldsymbol{\theta}^T \boldsymbol{H} \boldsymbol{\theta}}$, $h : \mathbb{R}^K \to \mathbb{R}$

When $\boldsymbol{\theta} \neq \mathbf{0}$, the subgradient of $h(\boldsymbol{\theta})$ coincides with its gradient, and we have

$$\partial h(\boldsymbol{\theta}) = \frac{\boldsymbol{H}\boldsymbol{\theta}}{\sqrt{\boldsymbol{\theta}^T \boldsymbol{H} \boldsymbol{\theta}}}, \quad \text{for } \boldsymbol{\theta} \neq \mathbf{0}.$$

By the definition, a subgradient of $h(\cdot)$ at $\mathbf{0}$ is any $\boldsymbol{g} \in \mathbb{R}^K$ such that,

$$h(\boldsymbol{\theta}) \geq h(\mathbf{0}) + \boldsymbol{g}^{\boldsymbol{T}} \boldsymbol{\theta}, \ \forall \boldsymbol{\theta} \in \mathbb{R}^K,$$

which implies

$$\sqrt{\boldsymbol{\theta}^T \boldsymbol{H} \boldsymbol{\theta}} \geq \boldsymbol{g}^{\boldsymbol{T}} \boldsymbol{\theta} \ \forall \boldsymbol{\theta} \in \mathbb{R}^K.$$

Write the Cholesky decomposition of positive semidefinite matrix $\boldsymbol{H}$ as $\boldsymbol{H} = \boldsymbol{L}^T \boldsymbol{L}$, where $\boldsymbol{L}$ is an upper triangular matrix with positive diagonal entries. Define $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{\theta} \in \mathbb{R}^K$, then we have

$$\boldsymbol{\theta}^T \boldsymbol{H} \boldsymbol{\theta} = \boldsymbol{\theta}^T \boldsymbol{L}^{\boldsymbol{T}} \boldsymbol{L} \boldsymbol{\theta} = \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{x} = \|\boldsymbol{x}\|_2^2.$$

Thus, the set of subgradients $\partial h(\mathbf{0})$ can be rewritten as

$$\left\{ \boldsymbol{g} \in \mathbb{R}^K : \frac{\boldsymbol{g}^{\boldsymbol{T}} \boldsymbol{\theta}}{\|\boldsymbol{x}\|_2} \leq 1 \ \forall \boldsymbol{\theta} \in \mathbb{R}^K. \right\}.$$

We can also rewrite the inner product $\boldsymbol{g}^{\boldsymbol{T}} \boldsymbol{\theta}$ in terms of $\boldsymbol{x}$,

$$\boldsymbol{g}^{\boldsymbol{T}} \boldsymbol{\theta} = \boldsymbol{g}^{\boldsymbol{T}} \boldsymbol{L}^{-1} \boldsymbol{x} = \left[\left(\boldsymbol{L}^{-1} \right)^T \boldsymbol{g} \right]^T \boldsymbol{x}.$$

By the Cauchy-Schwarz inequality, we have

$$\left[\left(\boldsymbol{L}^{-1} \right)^T \boldsymbol{g} \right]^T \boldsymbol{x} \leq \| \left(\boldsymbol{L}^{-1} \right)^T \boldsymbol{g} \|_2 \|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{g}^{\boldsymbol{T}} \boldsymbol{L}^{-\mathbf{1}} \left(\boldsymbol{L}^{-1} \right)^T \boldsymbol{g}} \cdot \|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{g}^{\boldsymbol{T}} \boldsymbol{H}^{-\mathbf{1}} \boldsymbol{g}} \cdot \|\boldsymbol{x}\|_2,$$

for $\forall \boldsymbol{x} \in \mathbb{R}^K$. In other words, we have derived the upper bound of $\frac{\boldsymbol{g^T\theta}}{\sqrt{\boldsymbol{\theta}^T\boldsymbol{H\theta}}}$ for any $\boldsymbol{\theta}$ as $\sqrt{\boldsymbol{g^T H^{-1} g}}$. Therefore, the subgradient of $h(\boldsymbol{\theta})$ at $\mathbf{0}$ is the set

$$\partial h(\mathbf{0}) = \left\{ \boldsymbol{g} \in \mathbb{R}^K : \sqrt{\boldsymbol{g^T H^{-1} g}} \leq 1 \right\}.$$

C.3 Additional simulation results

Table C.3: A catalogue of all the simulation results

Evaluation themes	Examples & Settings	Results
Methods: SSP, SSP0, gLASSO, and GAM		
Estimation	Example 1 ($P = 100, \rho = 0$)	Table 5.2; Figures C.1-C.4
	Example 1 ($P = 100, \rho = 0.3$)	Table C.4
	Example 1 ($P = 100, \rho = 0.7$)	Table C.5
	Example 1 ($P = 1000$)	Table C.14
	Example 2	Table C.13; Figures C.5-C.8
	Examples 3 and 4	Table C.16; Table C.17; Table C.18
Prediction	Examples 1 and 2	Table 5.3; Table C.8
	Examples 3 and 4	Table C.15
Selection	Examples 1 and 2	Table 5.4
	Examples 3 and 4	Table 5.5
Methods: ^{a}SSP, ^{a}SSP0, ^{a}gLASSO — the adaptive versions		
Estimation	Example 1 ($P = 100, \rho = 0$)	Table C.6
Prediction	Example 1 ($P = 100, \rho = 0$)	Table C.9
Selection	Example 1 ($P = 100, \rho = 0$)	Table C.11
Methods: SSP^{1SE}, SSP0^{1SE}, gLASSO^{1SE} — with the 1-SE-rule		
Estimation	Example 1 ($P = 100, \rho = 0$)	Table C.7
Prediction	Example 1 ($P = 100, \rho = 0$)	Table C.10
Selection	Example 1 ($P = 100, \rho = 0$)	Table C.12

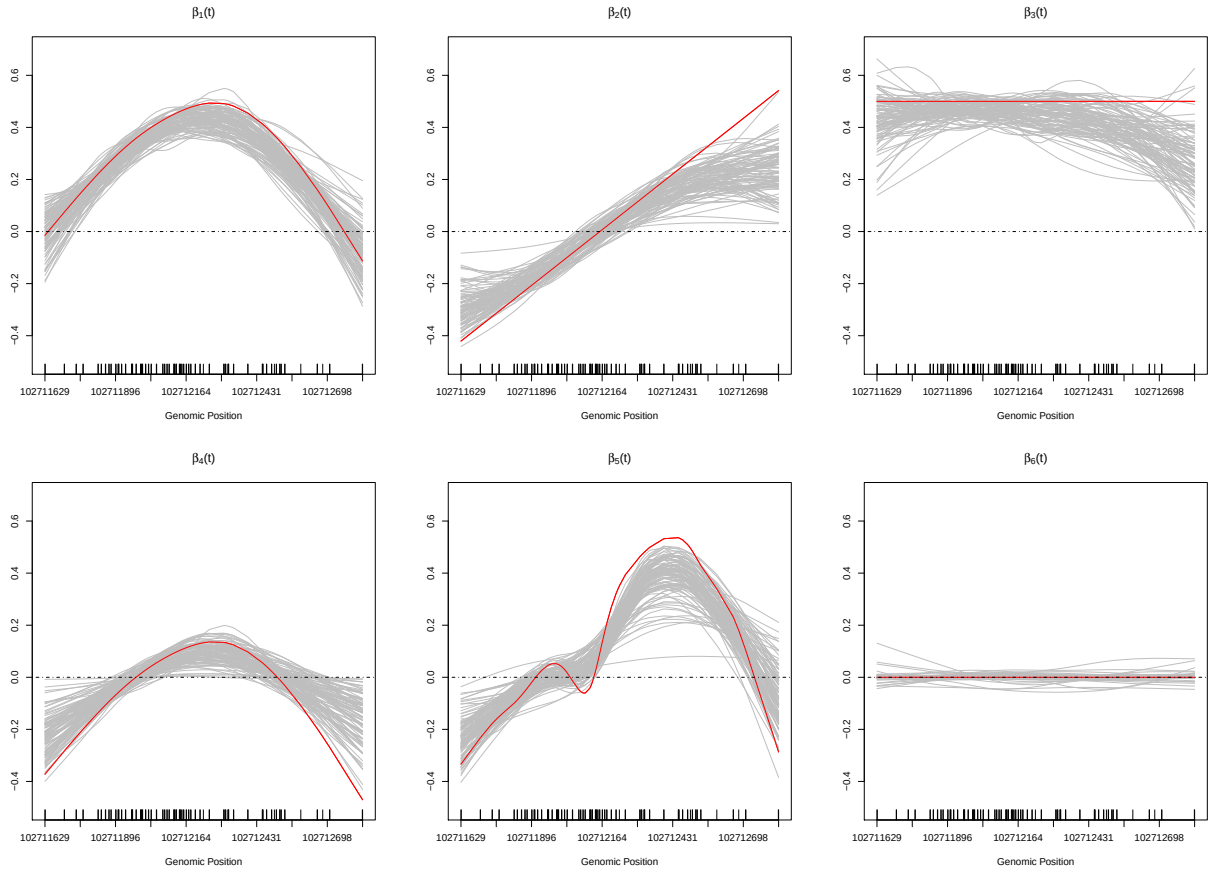


Figure C.1: **SSP** estimates of the first 6 varying coefficients (gray) in Example 1 ($P = 100, \rho = 0$) over 100 simulation runs. The red curves are the truth.

Table C.4: Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) of the first 10 varying coefficients of **Example 1** ($P = 100, \rho = 0.3$), using SSP, SSP0, group LASSO and GAM.

	IBIAS^2				IVAR				IMSE			
	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
$\beta_1(t)$	0.223	0.421	0.554	3.245	0.243	0.374	0.460	5.150	0.466	0.795	1.014	8.395
$\beta_2(t)$	0.744	1.280	1.264	1.425	0.264	0.318	0.320	3.561	1.008	1.597	1.584	4.986
$\beta_3(t)$	0.532	0.987	0.980	7.943	0.273	0.410	0.449	9.100	0.805	1.396	1.429	17.043
$\beta_4(t)$	0.679	0.905	0.977	0.459	0.176	0.203	0.210	2.745	0.856	1.108	1.187	3.205
$\beta_5(t)$	0.900	0.798	0.882	0.344	0.354	0.422	0.428	3.491	1.254	1.220	1.310	3.835
$\beta_6(t)$	1.0e-03	1.3e-03	1.4e-03	1.2e-02	1.5e-02	1.6e-02	1.8e-02	1.654	1.6e-02	1.8e-02	1.9e-02	1.665
$\beta_7(t)$	9.3e-04	1.0e-03	9.6e-04	4.3e-02	9.4e-03	1.2e-02	1.1e-02	2.646	1.0e-02	1.3e-02	1.2e-02	2.689
$\beta_8(t)$	1.7e-03	1.7e-03	1.5e-03	2.8e-02	1.4e-02	1.7e-02	1.8e-02	1.851	1.6e-02	1.8e-02	1.9e-02	1.879
$\beta_9(t)$	3.7e-04	5.0e-04	6.2e-04	2.2e-02	8.0e-03	1.1e-02	1.1e-02	2.611	8.3e-03	1.2e-02	1.1e-02	2.633
$\beta_{10}(t)$	1.4e-03	1.6e-03	1.4e-03	2.0e-02	1.6e-02	1.8e-02	1.8e-02	3.177	1.7e-02	1.9e-02	1.9e-02	3.197
$^\dagger\sum_1^{100}$	3.097	4.412	4.676	16.374	2.031	2.536	2.600	251.761	5.128	6.948	7.276	268.135

†: sum of the estimation measures across the 100 varying coefficients

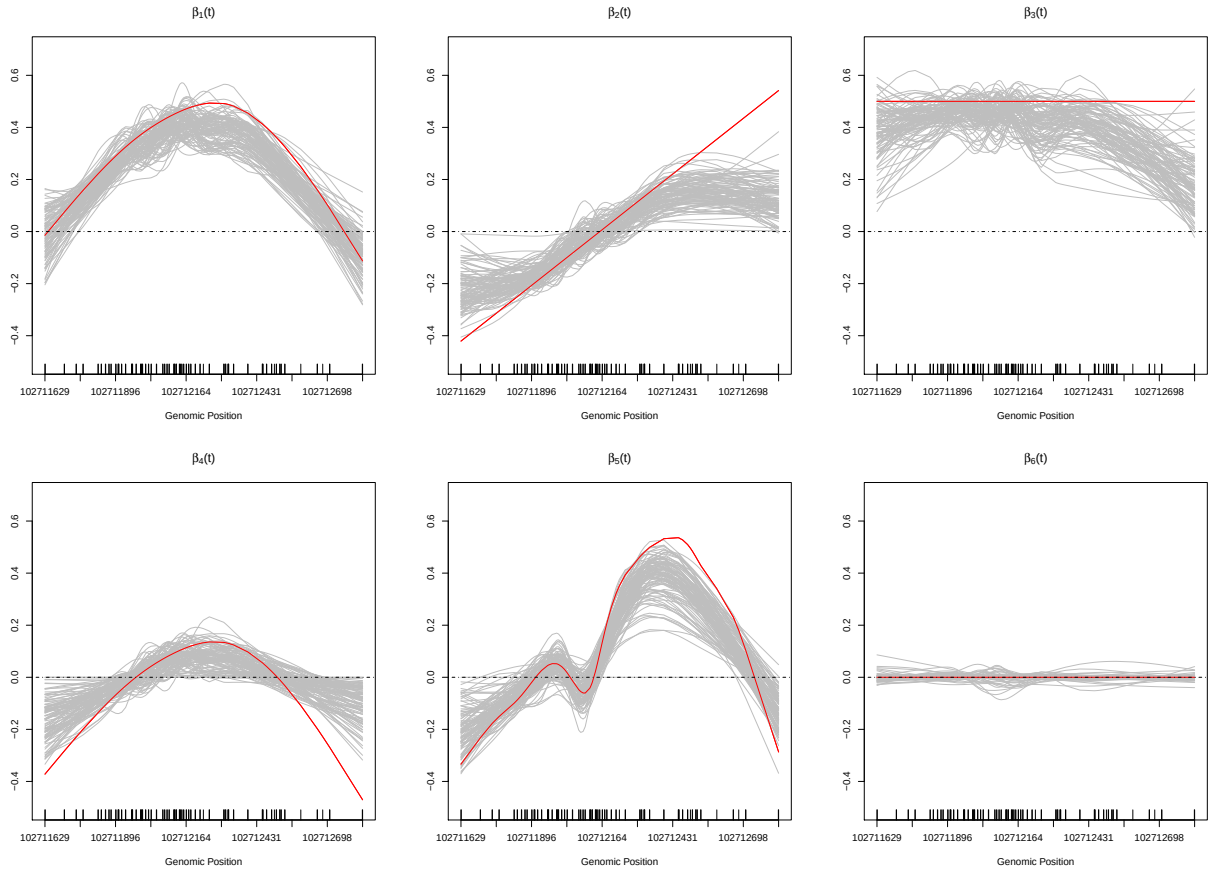


Figure C.2: **SSP0** estimates of the first 6 varying coefficients (gray) in Example 1 ($P = 100, \rho = 0$) over 100 simulation runs. The red curves are the truth.

Table C.5: Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) of the first 10 varying coefficients of **Example 1** ($P = 100, \rho = 0.7$), using SSP, SSP0, group LASSO and GAM.

	IBIAS^2				IVAR				IMSE			
	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
$\beta_1(t)$	0.258	0.489	0.691	4.223	0.403	0.625	0.756	6.143	0.661	1.115	1.447	10.367
$\beta_2(t)$	0.916	1.521	1.528	1.192	0.305	0.372	0.383	4.842	1.221	1.894	1.911	6.034
$\beta_3(t)$	0.710	1.269	1.317	7.269	0.347	0.529	0.572	11.026	1.056	1.799	1.890	18.295
$\beta_4(t)$	0.917	1.109	1.108	0.755	0.204	0.223	0.231	4.967	1.121	1.332	1.340	5.722
$\beta_5(t)$	1.177	0.968	1.137	0.333	0.548	0.632	0.631	4.914	1.725	1.601	1.768	5.248
$\beta_6(t)$	3.6e-03	4.7e-03	4.2e-03	1.1e-01	1.9e-02	2.5e-02	2.5e-02	4.389	2.3e-02	2.9e-02	2.9e-02	4.497
$\beta_7(t)$	2.5e-03	3.1e-03	2.8e-03	2.8e-02	1.5e-02	2.1e-02	2.0e-02	4.244	1.7e-02	2.4e-02	2.3e-02	4.273
$\beta_8(t)$	4.3e-03	5.1e-03	3.9e-03	9.5e-02	2.5e-02	3.2e-02	2.6e-02	2.102	2.9e-02	3.7e-02	3.0e-02	2.197
$\beta_9(t)$	1.2e-03	2.4e-03	2.8e-03	1.7e-01	1.7e-02	2.3e-02	2.0e-02	3.504	1.9e-02	2.5e-02	2.3e-02	3.674
$\beta_{10}(t)$	3.1e-03	4.1e-03	3.9e-03	7.0e-02	2.1e-02	2.5e-02	2.3e-02	2.432	2.4e-02	2.9e-02	2.7e-02	2.501
$^\dagger\sum_1^{100}$	4.022	5.415	5.840	17.318	2.661	3.342	3.438	359.066	6.684	8.757	9.278	376.385

†: sum of the estimation measures across the 100 varying coefficients

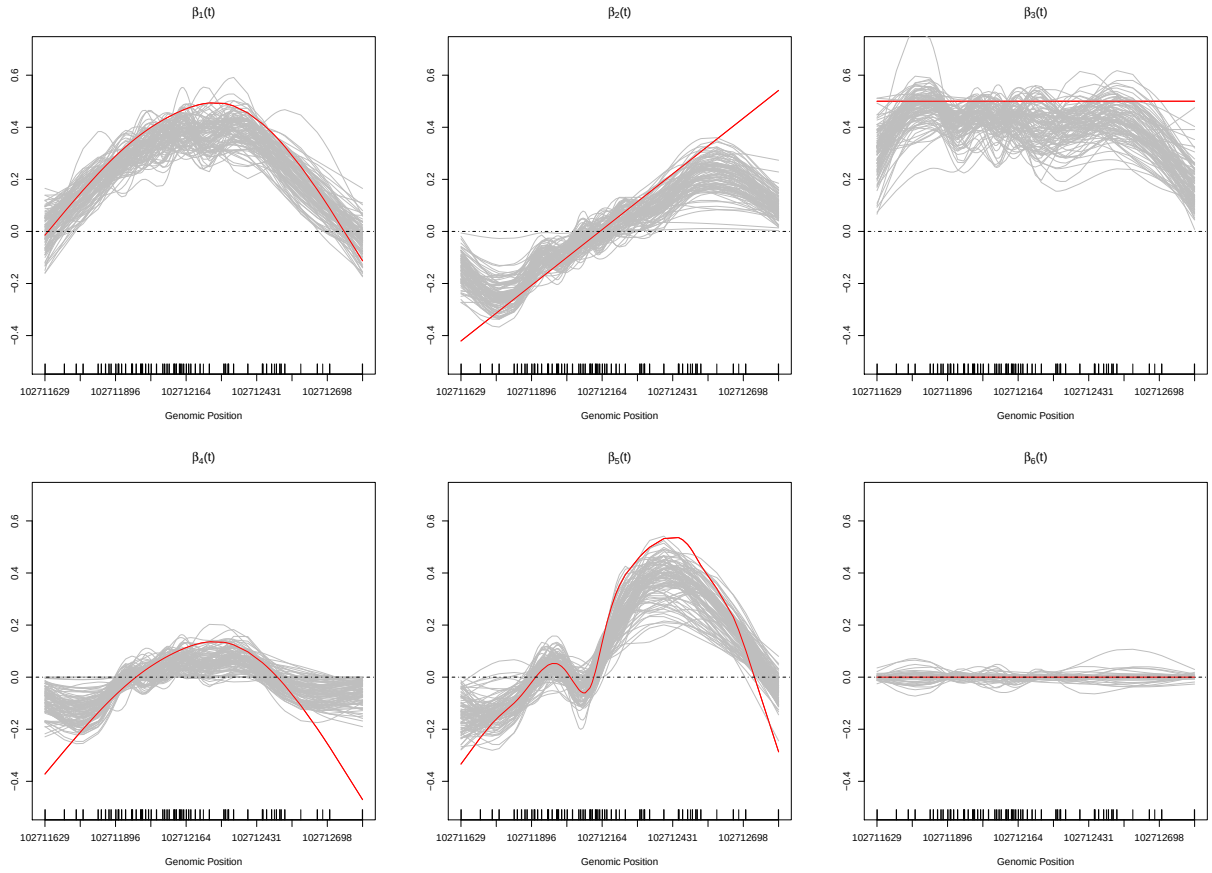


Figure C.3: **Group LASSO** estimates of the first 6 varying coefficients (gray) in Example 1 ($P = 100, \rho = 0$) over 100 simulation runs. The red curves are the truth.

Table C.6: Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) of the first 10 varying coefficients of Example 1 ($P = 100, \rho = 0$), using the **adaptive** SSP, SSP0, and group LASSO.

	IBIAS^2				IVAR				IMSE			
	[a]SSP	[a]SSP0	[a]SSP-fix$\alpha^{\ddagger}$	[a]gLASSO	[a]SSP	[a]SSP0	[a]SSP-fixα	[a]gLASSO	[a]SSP	[a]SSP0	[a]SSP-fixα	[a]gLASSO
$\beta_1(t)$	0.022	0.033	0.023	0.029	0.248	0.371	0.239	0.412	0.270	0.404	0.262	0.442
$\beta_2(t)$	0.090	0.264	0.078	0.377	0.245	0.358	0.226	0.355	0.335	0.623	0.304	0.732
$\beta_3(t)$	0.037	0.085	0.035	0.080	0.299	0.453	0.283	0.486	0.336	0.539	0.318	0.566
$\beta_4(t)$	0.161	0.297	0.165	0.467	0.246	0.325	0.247	0.323	0.407	0.622	0.412	0.790
$\beta_5(t)$	0.214	0.126	0.257	0.180	0.341	0.384	0.377	0.439	0.555	0.510	0.634	0.619
$\beta_6(t)$	1.1e-04	1.6e-04	1.1e-04	1.4e-04	1.6e-02	1.4e-02	1.6e-02	1.4e-02	1.7e-02	1.5e-02	1.6e-02	1.4e-02
$\beta_7(t)$	9.6e-05	9.5e-05	1.0e-04	3.6e-05	5.3e-03	3.8e-03	5.1e-03	5.0e-03	5.4e-03	3.9e-03	5.2e-03	5.0e-03
$\beta_8(t)$	8.3e-05	1.4e-04	9.4e-05	1.5e-04	1.0e-02	1.0e-02	1.1e-02	9.9e-03	1.1e-02	1.0e-02	1.1e-02	1.0e-02
$\beta_9(t)$	4.9e-05	3.9e-05	4.6e-05	1.1e-04	8.0e-03	5.1e-03	7.5e-03	8.4e-03	8.1e-03	5.2e-03	7.5e-03	8.5e-03
$\beta_{10}(t)$	7.9e-05	1.7e-05	9.7e-05	7.1e-05	8.4e-03	2.8e-03	8.7e-03	5.9e-03	8.5e-03	2.8e-03	8.8e-03	6.0e-03
$^{\dagger}\sum_1^{100}$	0.533	0.812	0.565	1.142	2.191	2.536	2.177	2.833	2.723	3.348	2.742	3.975

†: sum of the estimation measures across the 100 varying coefficients.

‡: adaptive SSP with fix α. Here we use the α selected by the ordinary SSP method and only tune the values of λ.

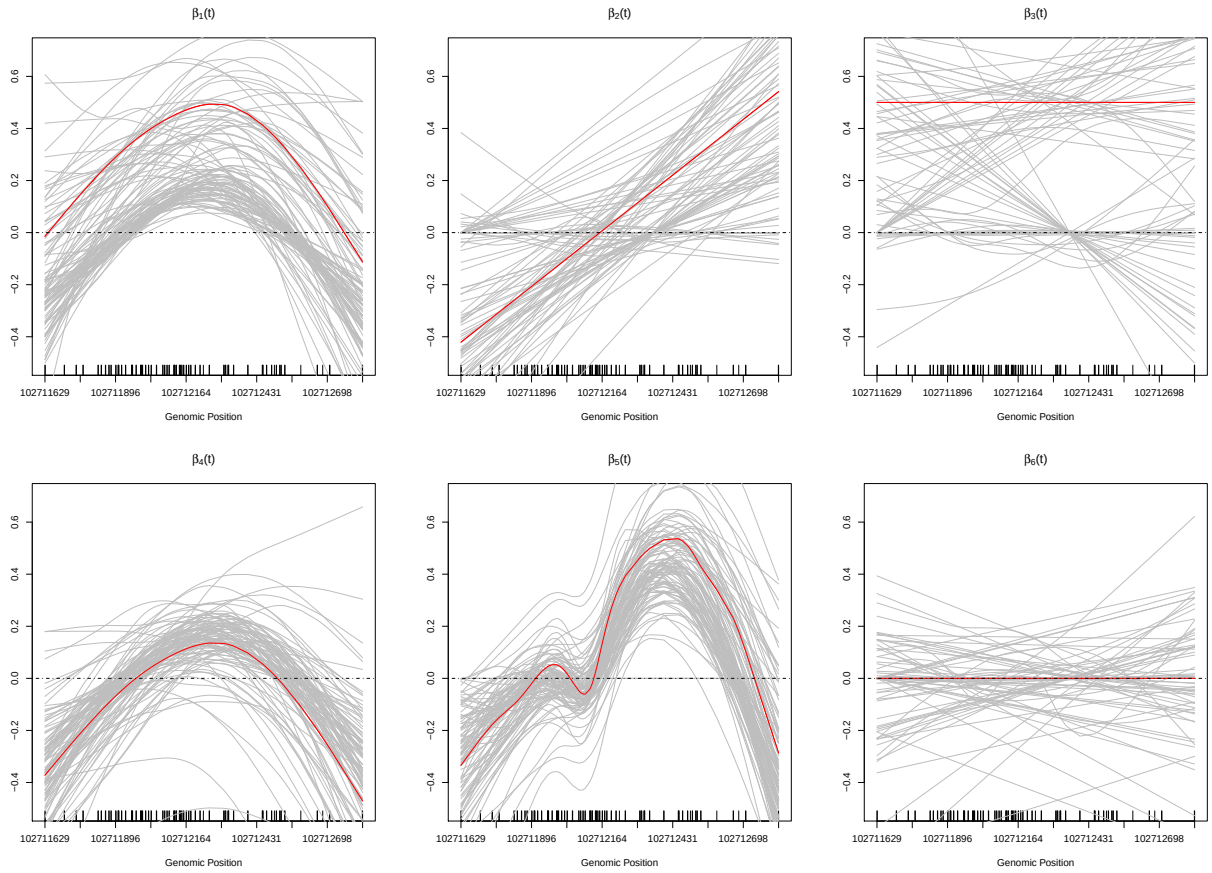


Figure C.4: **GAM** estimates of the first 6 varying coefficients (gray) in Example 1 ($P = 100, \rho = 0$) over 100 simulation runs. The red curves are the truth.

Table C.7: Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) of the first 10 varying coefficients of Example 1 ($P = 100, \rho = 0$), using **the 1 SE rule** for SSP, SSP0, and group LASSO.

	IBIAS2			IVAR			IMSE		
	SSP^{1SE}	SSP0^{1SE}	gLASSO^{1SE}	SSP^{1SE}	SSP0^{1SE}	gLASSO^{1SE}	SSP^{1SE}	SSP0^{1SE}	gLASSO^{1SE}
$\beta_1(t)$	1.142	1.593	1.872	0.195	0.278	0.348	1.337	1.871	2.220
$\beta_2(t)$	2.356	2.881	2.358	0.199	0.208	0.213	2.556	3.089	2.570
$\beta_3(t)$	2.198	3.019	2.829	0.374	0.512	0.575	2.571	3.531	3.404
$\beta_4(t)$	1.513	1.584	1.562	0.080	0.080	0.079	1.593	1.663	1.641
$\beta_5(t)$	3.186	2.879	2.656	0.390	0.419	0.423	3.576	3.298	3.079
$\beta_6(t)$	4.1e-07	6.4e-07	7.5e-06	4.5e-04	2.2e-04	7.4e-04	4.5e-04	2.2e-04	7.5e-04
$\beta_7(t)$	4.7e-07	5.9e-06	1.3e-06	4.7e-05	5.9e-04	1.3e-04	4.7e-05	5.9e-04	1.3e-04
$\beta_8(t)$	2.6e-08	1.0e-06	1.4e-06	5.1e-06	8.6e-05	1.6e-04	5.1e-06	8.7e-05	1.7e-04
$\beta_9(t)$	3.1e-11	4.0e-07	4.3e-08	3.1e-09	1.4e-04	4.3e-06	3.1e-09	1.4e-04	4.3e-06
$\beta_{10}(t)$	0.0e+00	4.3e-07	5.2e-07	0.0e+00	4.1e-05	2.2e-04	0.0e+00	4.2e-05	2.2e-04
$\sum_1^{100}$	10.396	11.956	11.276	1.265	1.532	1.678	11.661	13.489	12.954

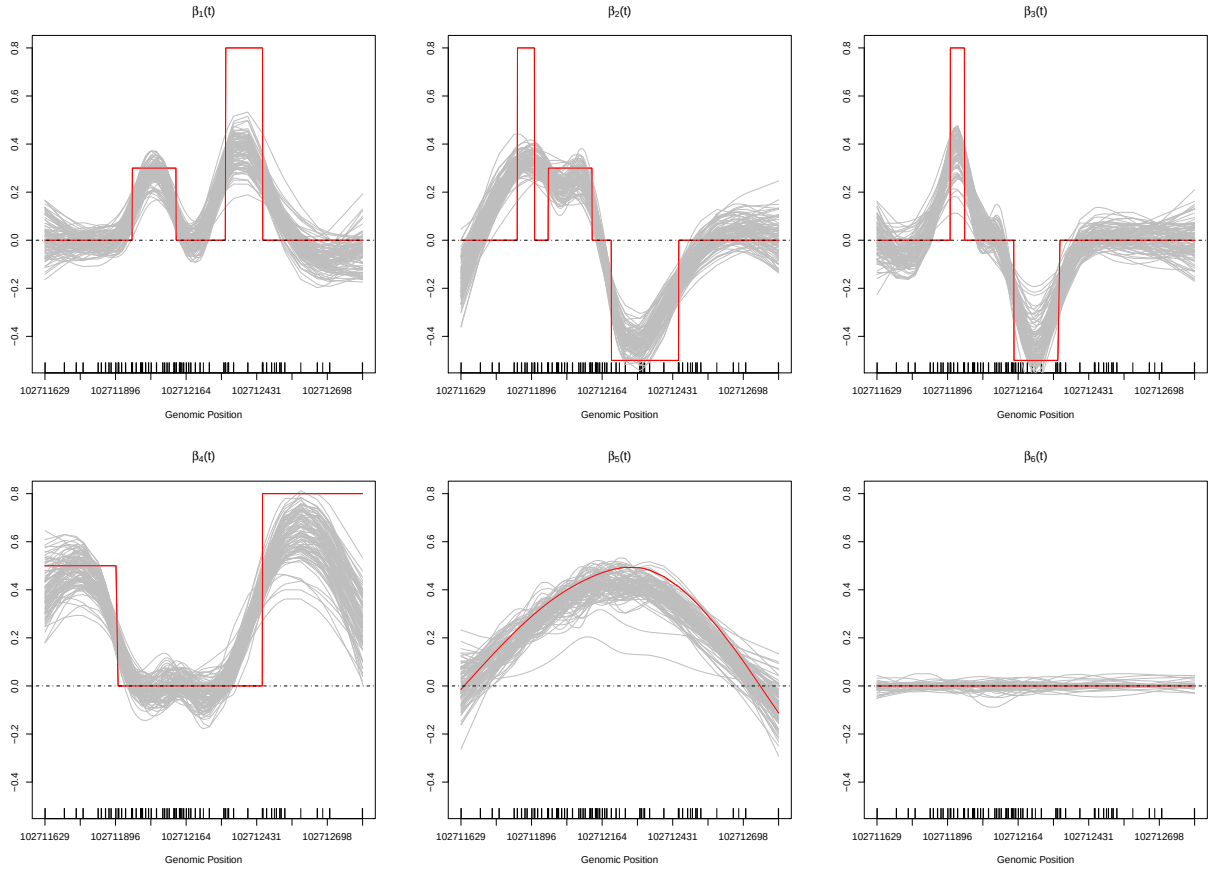


Figure C.5: **SSP** estimates of the first 6 varying coefficients (gray) in Example 2 ($P = 100, \rho = 0$) over 100 simulation runs. The red curves are the truth.

Table C.8: Average values of the CorRaw and CorTrans over 100 simulations for simulation **Examples 1 and 2**. Standard deviations are given in parentheses.

	CorRaw				CorTrans			
ρ	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
Example 1 (smooth, $P_{true} = 5, P = 100$)								
0	0.790(0.020)	0.788(0.020)	0.787(0.020)	0.552(0.111)	0.769(0.020)	0.766(0.020)	0.765(0.020)	0.542(0.109)
0.3	0.789(0.018)	0.787(0.018)	0.786(0.018)	0.528(0.145)	0.773(0.019)	0.771(0.019)	0.770(0.019)	0.526(0.144)
0.7	0.789(0.017)	0.788(0.017)	0.787(0.017)	0.525(0.147)	0.780(0.017)	0.778(0.018)	0.778(0.017)	0.533(0.144)
Example 1 (smooth, $P_{true} = 5, P = 1000$)								
0	0.788(0.019)	0.785(0.020)	0.783(0.020)	NA	0.766(0.019)	0.763(0.020)	0.761(0.021)	NA
0.3	0.789(0.015)	0.787(0.016)	0.786(0.015)	NA	0.773(0.016)	0.770(0.016)	0.769(0.016)	NA
Example 2 (nonsmooth, $P_{true} = 5, P = 100$)								
0	0.655(0.019)	0.655(0.019)	0.658(0.019)	0.467(0.082)	0.658(0.023)	0.658(0.023)	0.661(0.021)	0.479(0.080)
0.3	0.669(0.019)	0.668(0.019)	0.670(0.020)	0.418(0.105)	0.669(0.022)	0.669(0.022)	0.670(0.023)	0.428(0.106)
Example 2 (nonsmooth, $P_{true} = 5, P = 1000$)								
0	0.645(0.024)	0.645(0.024)	0.648(0.025)	NA	0.649(0.026)	0.648(0.026)	0.651(0.027)	NA

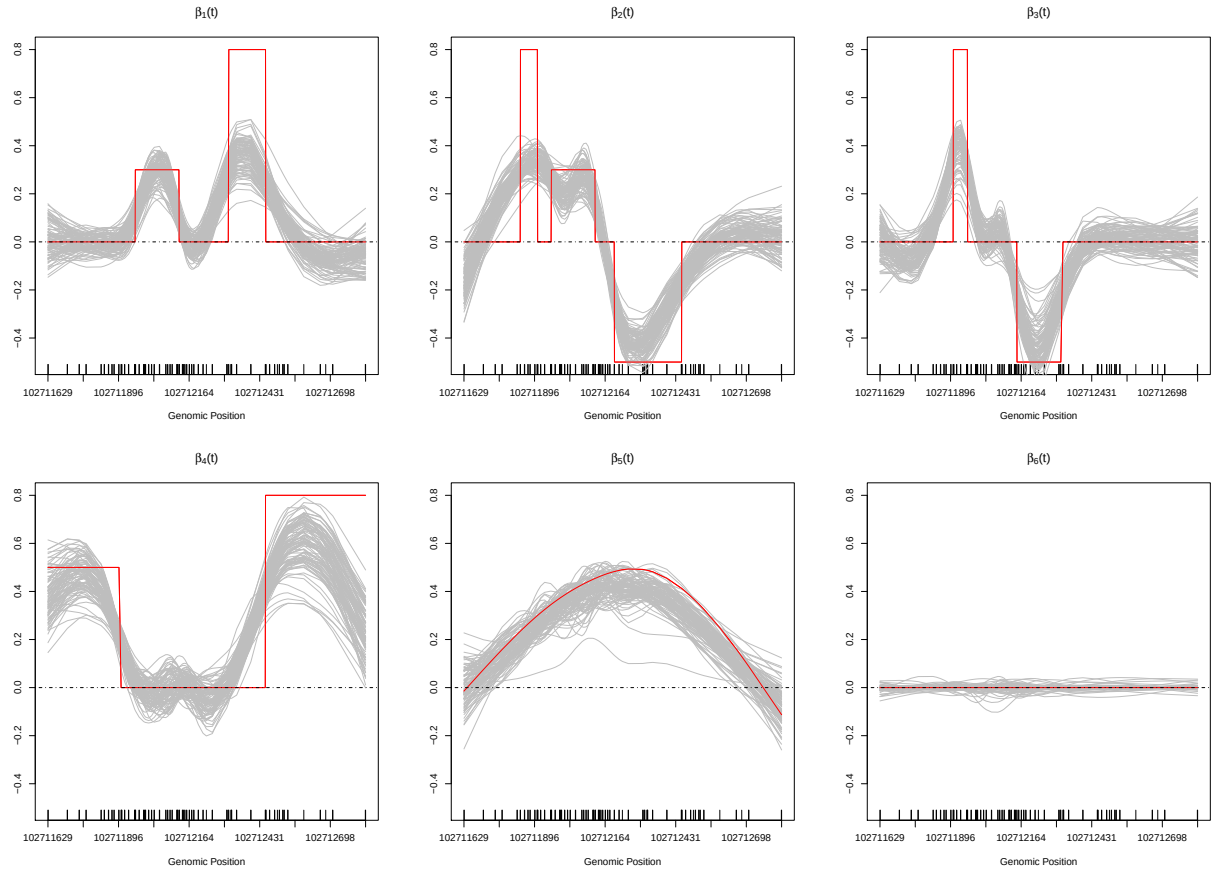


Figure C.6: **SSP0** estimates of the first 6 varying coefficients (gray) in Example 2 ($P = 100, \rho = 0$) over 100 simulation runs. The red curves are the truth.

Table C.9: Average values of the deviance errors, RMSE CorRaw and CorTrans over 100 simulations using the **adaptive** SSP, SSP0 and gLASSO. Standard deviations are given in parentheses.

Example 1 (smooth, $P_{true} = 5, P = 100, \rho = 0$)								
	Deviance				RMSE			
ρ	^{a}SSP	^{a}SSP0	^{a}SSP-fixα	^{a}gLASSO	^{a}SSP	^{a}SSP0	^{a}SSP-fixα	^{a}gLASSO
0	0.016(0.007)	0.020(0.008)	0.016(0.008)	0.023(0.010)	0.408(0.006)	0.409(0.007)	0.408(0.408)	0.410(0.007)
	CorRaw				CorTrans			
	^{a}SSP	^{a}SSP0	^{a}SSP-fixα	^{a}gLASSO	^{a}SSP	^{a}SSP0	^{a}SSP-fixα	^{a}gLASSO
0	0.791(0.019)	0.790(0.019)	0.791(0.019)	0.790(0.019)	0.771(0.020)	0.770(0.020)	0.771(0.020)	0.769(0.020)

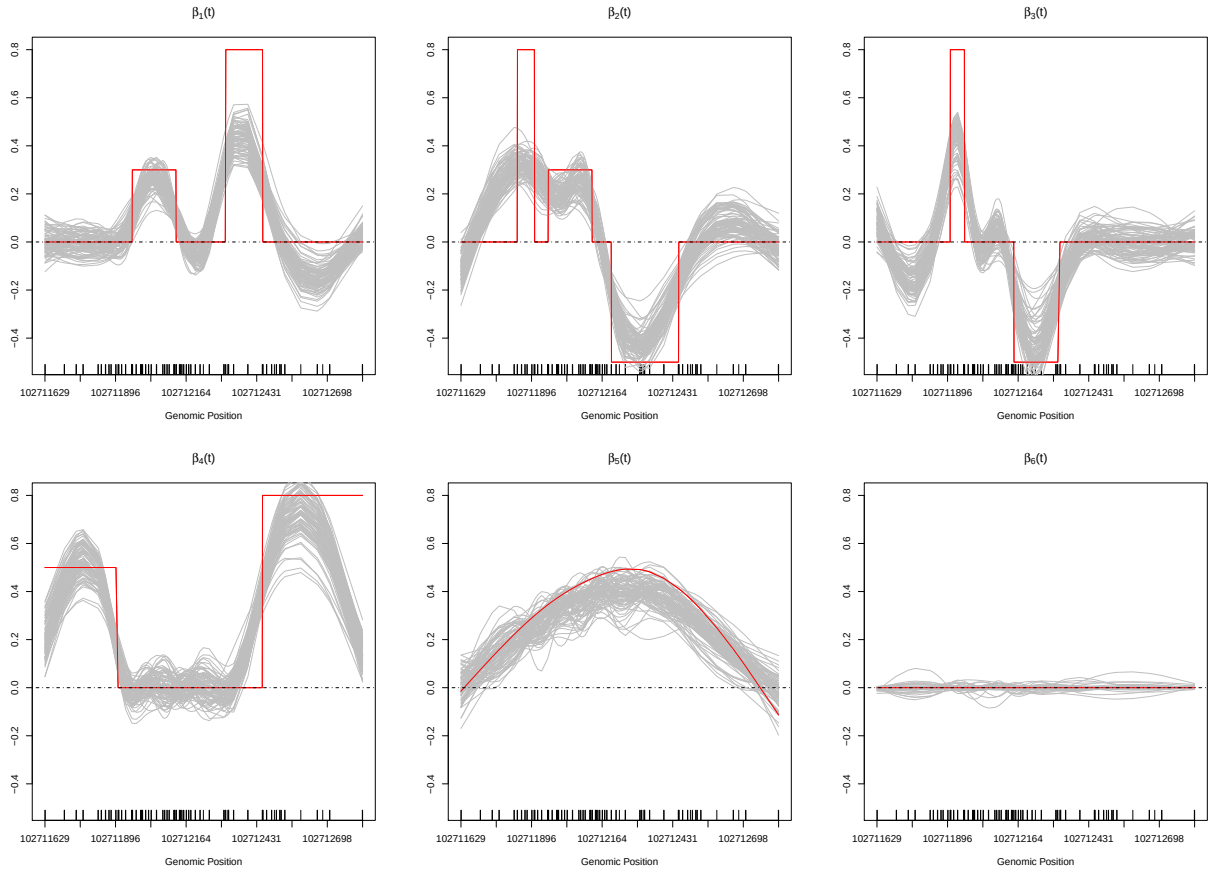


Figure C.7: **Group LASSO** estimates of the first 6 varying coefficients (gray) in Example 2 ($P = 100, \rho = 0$) over 100 simulation runs. The red curves are the truth.

Table C.10: Average values of the deviance errors, RMSE CorRaw and CorTrans over 100 simulations using the **the 1 SE rule** for SSP, SSP0 and gLASSO. Standard deviations are given in parentheses.

Example 1 (smooth, $P_{true} = 5, P = 100, \rho = 0$)						
	Deviance			RMSE		
ρ	SSP^{1SE}	SSP0^{1SE}	gLASSO^{1SE}	SSP^{1SE}	SSP0^{1SE}	gLASSO^{1SE}
0	0.060(0.028)	0.070(0.033)	0.070(0.032)	0.420(0.011)	0.422(0.012)	0.422(0.012)
	CorRaw			CorTrans		
ρ	SSP^{1SE}	SSP0^{1SE}	gLASSO^{1SE}	SSP^{1SE}	SSP0^{1SE}	gLASSO^{1SE}
0	0.785(0.020)	0.783(0.021)	0.783(0.021)	0.762(0.021)	0.761(0.021)	0.760(0.021)

Table C.11: Average values of the number of TP and FP for simulation examples 1 and 2, using the **adaptive** SSP, SSP0, and gLASSO. Standard deviations are given in parentheses.

Example 1 (smooth, $P_{true} = 5, P = 100, \rho = 0$)								
	TP				FP			
ρ	^{a}SSP	^{a}SSP0	^{a}SSP-fixα	^{a}gLASSO	^{a}SSP	^{a}SSP0	^{a}SSP-fixα	^{a}gLASSO
0	4.99(0.10)	4.98(0.14)	4.99(0.10)	4.95(0.22)	16.18(6.53)	15.56(5.38)	16.38(6.72)	15.97(4.79)

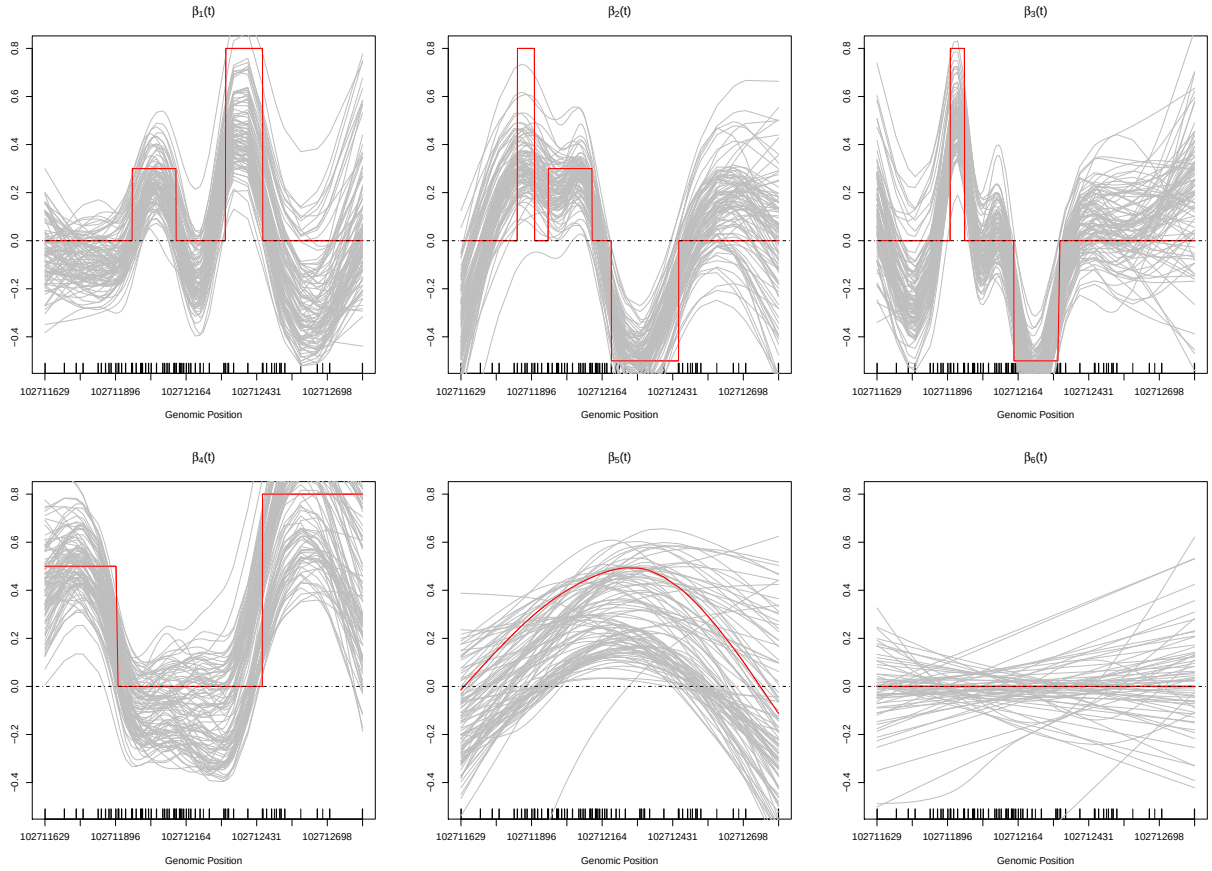


Figure C.8: **GAM** estimates of the first 6 varying coefficients (gray) in Example 2 ($P = 100, \rho = 0$) over 100 simulation runs. The red curves are the truth.

Table C.12: Average values of the number of TP and FP for simulation examples 1 and 2, using the **1 SE rule** for SSP, SSP0, and gLASSO. Standard deviations are given in parentheses.

Example 1 (smooth, $P_{true} = 5, P = 100, \rho = 0$)						
	TP			FP		
ρ	SSP^{1SE}	SSP0^{1SE}	gLASSO^{1SE}	SSP^{1SE}	SSP0^{1SE}	gLASSO^{1SE}
0	4.58(0.61)	4.59(0.62)	4.68(0.55)	1.77(1.98)	3.10(3.29)	2.93(3.26)

Table C.13: Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) of the first 10 varying coefficients of **Example 2 (non smooth)**, using SSP, SSP0, group LASSO and GAM.

		IBIAS^2				IVAR				IMSE			
ρ		SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
Example 2 (nonsmooth, $P = 100$)													
0	$\beta_1(t)$	3.641	3.698	3.353	4.095	0.235	0.234	0.208	1.839	3.876	3.931	3.562	5.934
	$\beta_2(t)$	4.874	4.900	4.956	4.949	0.235	0.242	0.264	1.553	5.109	5.142	5.220	6.501
	$\beta_3(t)$	4.460	4.248	4.030	3.794	0.315	0.301	0.321	1.741	4.775	4.548	4.352	5.535
	$\beta_4(t)$	3.089	3.463	2.969	2.734	0.352	0.382	0.302	3.495	3.441	3.845	3.271	6.230
	$\beta_5(t)$	0.385	0.438	0.534	3.977	0.296	0.316	0.294	3.626	0.682	0.754	0.828	7.603
	$\beta_6(t)$	1.7e-04	1.4e-04	6.3e-05	3.8e-02	1.4e-02	1.4e-02	9.6e-03	1.641	1.4e-02	1.4e-02	9.7e-03	1.679
	$\beta_7(t)$	7.2e-05	1.2e-04	1.4e-04	3.7e-02	8.6e-03	8.9e-03	8.4e-03	0.903	8.6e-03	9.1e-03	8.5e-03	0.940
	$\beta_8(t)$	1.6e-04	2.3e-04	2.1e-04	1.5e-02	1.3e-02	1.4e-02	1.1e-02	1.659	1.3e-02	1.4e-02	1.1e-02	1.674
	$\beta_9(t)$	1.1e-04	1.2e-04	1.3e-04	4.4e-02	9.2e-03	1.0e-02	5.9e-03	0.902	9.3e-03	1.0e-02	6.1e-03	0.946
	$\beta_{10}(t)$	4.6e-05	6.9e-05	4.7e-05	1.6e-03	1.1e-02	1.1e-02	5.9e-03	1.214	1.1e-02	1.1e-02	6.0e-03	1.215
	$\sum_1^{100}$	16.460	16.757	15.851	21.327	2.470	2.525	2.104	119.279	18.930	19.282	17.955	140.606
0.3	$\beta_1(t)$	3.905	3.964	3.481	3.519	0.281	0.285	0.290	3.552	4.186	4.249	3.771	7.072
	$\beta_2(t)$	5.033	5.081	5.141	4.807	0.327	0.329	0.357	2.140	5.359	5.410	5.498	6.947
	$\beta_3(t)$	4.400	4.226	3.899	3.772	0.328	0.313	0.298	2.112	4.728	4.540	4.197	5.883
	$\beta_4(t)$	3.594	4.033	3.080	3.168	0.558	0.599	0.418	4.366	4.152	4.632	3.498	7.534
	$\beta_5(t)$	0.415	0.471	0.638	3.394	0.328	0.354	0.386	4.819	0.743	0.825	1.025	8.212
	$\beta_6(t)$	6.0e-04	6.1e-04	5.5e-04	5.3e-03	9.9e-03	9.5e-03	8.0e-03	1.799	1.0e-02	1.0e-02	8.6e-03	1.804
	$\beta_7(t)$	4.1e-04	4.2e-04	4.6e-04	4.4e-03	1.3e-02	1.2e-02	1.2e-02	1.700	1.3e-02	1.3e-02	1.2e-02	1.705
	$\beta_8(t)$	5.1e-04	4.1e-04	3.3e-04	5.0e-03	1.2e-02	1.1e-02	8.4e-03	1.338	1.3e-02	1.2e-02	8.7e-03	1.343
	$\beta_9(t)$	3.5e-04	3.9e-04	4.4e-04	6.1e-02	1.2e-02	1.3e-02	9.9e-03	1.513	1.2e-02	1.3e-02	1.0e-02	1.573
	$\beta_{10}(t)$	1.2e-03	1.2e-03	9.6e-04	1.5e-02	1.4e-02	1.4e-02	1.3e-02	1.797	1.5e-02	1.5e-02	1.4e-02	1.812
	$\sum_1^{100}$	17.368	17.797	16.255	20.825	2.940	2.974	2.522	175.293	20.308	20.770	18.777	196.118
Example 2 (nonsmooth, $P = 1000$)													
0	$\beta_1(t)$	4.283	4.358	4.001	NA	0.308	0.306	0.310	NA	4.591	4.664	4.311	NA
	$\beta_2(t)$	5.448	5.476	5.601	NA	0.357	0.366	0.392	NA	5.805	5.842	5.993	NA
	$\beta_3(t)$	4.975	4.752	4.464	NA	0.319	0.313	0.325	NA	5.293	5.066	4.789	NA
	$\beta_4(t)$	4.653	5.201	4.098	NA	0.585	0.603	0.539	NA	5.238	5.804	4.637	NA
	$\beta_5(t)$	1.035	1.112	1.250	NA	0.421	0.437	0.470	NA	1.456	1.550	1.720	NA
	$\beta_6(t)$	6.5e-06	6.1e-06	5.2e-06	NA	9.5e-04	8.1e-04	7.8e-04	NA	9.5e-04	8.1e-04	7.8e-04	NA
	$\beta_7(t)$	7.4e-06	5.5e-06	5.7e-06	NA	1.5e-03	1.7e-03	1.1e-03	NA	1.6e-03	1.7e-03	1.1e-03	NA
	$\beta_8(t)$	5.5e-06	3.3e-06	1.4e-06	NA	6.8e-04	7.2e-04	2.4e-04	NA	6.8e-04	7.2e-04	2.4e-04	NA
	$\beta_9(t)$	1.1e-05	1.0e-05	1.4e-05	NA	1.4e-03	1.3e-03	1.2e-03	NA	1.4e-03	1.3e-03	1.2e-03	NA
	$\beta_{10}(t)$	1.7e-05	1.7e-05	8.5e-06	NA	1.2e-03	1.4e-03	1.4e-03	NA	1.2e-03	1.4e-03	1.4e-03	NA
	$\sum_1^{1000}$	20.404	20.909	19.423	NA	2.910	2.940	2.784	NA	23.314	23.850	22.206	NA

Table C.14: Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) of the first 10 varying coefficients of **Example 1** ($P = 1000$), using SSP, SSP0, and group LASSO.

		IBIAS^2			IVAR			IMSE		
ρ		SSP	SSP0	gLASSO	SSP	SSP0	gLASSO	SSP	SSP0	gLASSO
	$\beta_1(t)$	0.947	1.362	1.560	0.409	0.539	0.641	1.356	1.901	2.201
	$\beta_2(t)$	1.356	1.960	1.845	0.284	0.292	0.310	1.641	2.252	2.156
	$\beta_3(t)$	1.345	2.013	2.113	0.586	0.745	0.802	1.931	2.757	2.915
	$\beta_4(t)$	0.966	1.241	1.365	0.134	0.116	0.104	1.100	1.357	1.469
	$\beta_5(t)$	1.898	2.233	2.261	0.343	0.407	0.401	2.241	2.640	2.662
0	$\beta_6(t)$	8.7e-07	3.8e-06	2.6e-06	2.4e-04	5.1e-04	4.7e-04	2.4e-04	5.1e-04	4.8e-04
	$\beta_7(t)$	2.1e-06	1.4e-05	1.4e-05	9.7e-04	2.0e-03	1.4e-03	9.7e-04	2.0e-03	1.5e-03
	$\beta_8(t)$	5.3e-06	1.1e-05	1.1e-05	8.8e-04	1.0e-03	1.5e-03	8.9e-04	1.0e-03	1.5e-03
	$\beta_9(t)$	8.1e-06	7.4e-06	2.1e-05	6.6e-04	4.7e-04	7.0e-04	6.7e-04	4.8e-04	7.2e-04
	$\beta_{10}(t)$	9.6e-06	8.8e-06	5.7e-06	1.2e-03	7.1e-04	4.9e-04	1.3e-03	7.2e-04	4.9e-04
	$\sum_1^{1000}$	6.523	8.819	9.156	2.623	3.070	3.197	9.146	11.889	12.353
	$\beta_1(t)$	0.636	0.908	1.165	0.487	0.623	0.733	1.123	1.531	1.899
	$\beta_2(t)$	1.461	2.014	1.908	0.332	0.351	0.357	1.793	2.365	2.265
	$\beta_3(t)$	1.070	1.590	1.641	0.354	0.479	0.530	1.424	2.070	2.172
	$\beta_4(t)$	0.893	1.093	1.141	0.149	0.157	0.149	1.042	1.249	1.290
	$\beta_5(t)$	1.291	1.368	1.356	0.487	0.554	0.537	1.778	1.922	1.893
0.3	$\beta_6(t)$	2.5e-04	3.7e-04	3.9e-04	2.8e-03	4.2e-03	4.7e-03	3.1e-03	4.5e-03	5.1e-03
	$\beta_7(t)$	3.4e-04	3.4e-04	2.1e-04	3.8e-03	4.5e-03	4.2e-03	4.1e-03	4.9e-03	4.4e-03
	$\beta_8(t)$	2.5e-04	4.2e-04	5.3e-04	7.6e-03	1.0e-02	1.4e-02	7.8e-03	1.1e-02	1.4e-02
	$\beta_9(t)$	2.0e-04	2.6e-04	2.8e-04	3.6e-03	4.8e-03	4.6e-03	3.8e-03	5.1e-03	4.9e-03
	$\beta_{10}(t)$	5.5e-04	7.5e-04	9.2e-04	6.3e-03	8.3e-03	9.6e-03	6.8e-03	9.1e-03	1.0e-02
	$\sum_1^{1000}$	5.360	6.984	7.224	2.493	2.950	3.100	7.853	9.934	10.324

Table C.15: Average values of the deviance errors, RMSE, CorRaw and CorTrans over 100 simulations for simulation **Examples 3 and 4**. Standard deviations are given in parentheses.

		Deviance				RMSE			
P	P_{true}	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
50	5	0.049(0.021)	0.064(0.025)	0.063(0.023)	0.634(0.405)	0.405(0.009)	0.408(0.010)	0.409(0.010)	0.482(0.046)
	10	0.148(0.047)	0.169(0.047)	0.162(0.041)	0.690(0.345)	0.429(0.013)	0.433(0.014)	0.431(0.013)	0.495(0.038)
100	5	0.071(0.032)	0.080(0.030)	0.077(0.023)	0.593(0.243)	0.412(0.010)	0.414(0.010)	0.414(0.010)	0.483(0.030)
	10	0.190(0.069)	0.204(0.068)	0.194(0.050)	0.915(0.671)	0.438(0.016)	0.440(0.016)	0.439(0.014)	0.520(0.063)
150	5	0.072(0.030)	0.084(0.032)	0.083(0.028)	0.596(0.346)	0.413(0.011)	0.415(0.011)	0.414(0.010)	0.483(0.038)
	10	0.192(0.068)	0.202(0.063)	0.192(0.047)	0.918(0.401)	0.439(0.018)	0.441(0.018)	0.439(0.016)	0.525(0.046)
200	5	0.087(0.040)	0.098(0.042)	0.092(0.035)	0.598(0.276)	0.415(0.012)	0.417(0.013)	0.416(0.011)	0.484(0.034)
	10	0.236(0.104)	0.242(0.092)	0.227(0.070)	0.804(0.416)	0.444(0.021)	0.445(0.020)	0.443(0.018)	0.511(0.046)
1000	5	0.108(0.038)	0.115(0.036)	0.113(0.032)	NA	0.415(0.010)	0.417(0.009)	0.417(0.009)	NA
	10	0.272(0.076)	0.272(0.073)	0.274(0.073)	NA	0.447(0.018)	0.447(0.017)	0.447(0.017)	NA
		CorRaw				CorTrans			
P	P_{true}	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
50	5	0.757(0.023)	0.753(0.023)	0.753(0.022)	0.634(0.079)	0.729(0.022)	0.725(0.022)	0.725(0.021)	0.608(0.079)
	10	0.710(0.039)	0.703(0.039)	0.703(0.039)	0.569(0.087)	0.677(0.038)	0.669(0.038)	0.670(0.038)	0.542(0.084)
100	5	0.749(0.026)	0.747(0.026)	0.748(0.025)	0.616(0.067)	0.720(0.026)	0.717(0.025)	0.718(0.025)	0.588(0.065)
	10	0.700(0.035)	0.695(0.036)	0.696(0.034)	0.521(0.136)	0.665(0.035)	0.659(0.035)	0.661(0.033)	0.493(0.131)
150	5	0.751(0.024)	0.749(0.023)	0.748(0.023)	0.633(0.051)	0.723(0.023)	0.720(0.022)	0.720(0.022)	0.606(0.050)
	10	0.702(0.036)	0.699(0.035)	0.698(0.036)	0.513(0.113)	0.668(0.034)	0.664(0.033)	0.663(0.034)	0.487(0.108)
200	5	0.751(0.025)	0.749(0.025)	0.749(0.025)	0.630(0.065)	0.722(0.025)	0.719(0.024)	0.719(0.024)	0.601(0.063)
	10	0.695(0.048)	0.692(0.048)	0.692(0.047)	0.533(0.102)	0.660(0.047)	0.657(0.046)	0.656(0.046)	0.505(0.098)
1000	5	0.747(0.027)	0.745(0.027)	0.745(0.027)	NA	0.717(0.026)	0.716(0.026)	0.716(0.026)	NA
	10	0.683(0.047)	0.682(0.046)	0.679(0.045)	NA	0.647(0.045)	0.646(0.044)	0.643(0.043)	NA

Table C.16: Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) of the first 5 varying coefficients of **Examples 3 and 4** ($N = 20, P = 50, 100$), using SSP, SSP0, group LASSO and GAM.

		IBIAS^2				IVAR				IMSE			
P	P_{true}	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
50	5	0.568	0.870	1.090	2.426	3.0e-01	3.9e-01	4.2e-01	1.6e+00	0.868	1.256	1.512	4.048
50	5	1.480	2.165	2.083	1.359	5.8e-01	5.8e-01	5.9e-01	3.6e+00	2.057	2.744	2.675	4.977
50	5	0.584	0.946	1.091	1.709	4.2e-01	5.1e-01	5.4e-01	2.8e+00	1.005	1.451	1.627	4.477
50	5	0.921	1.169	1.290	0.308	2.9e-01	2.7e-01	2.3e-01	1.7e+00	1.211	1.441	1.523	2.017
50	5	0.864	1.152	1.241	0.160	2.9e-01	2.7e-01	2.3e-01	2.2e+00	1.156	1.419	1.470	2.387
50	10	0.829	1.058	1.174	1.662	6.7e-01	7.0e-01	7.0e-01	2.3e+00	1.499	1.760	1.872	3.986
50	10	1.530	2.066	2.075	2.016	7.5e-01	7.6e-01	7.0e-01	3.1e+00	2.283	2.828	2.775	5.083
50	10	0.974	1.212	1.317	1.216	7.4e-01	8.2e-01	8.2e-01	2.4e+00	1.712	2.028	2.138	3.650
50	10	1.047	1.191	1.301	0.319	3.6e-01	3.2e-01	2.7e-01	1.7e+00	1.403	1.514	1.568	2.033
50	10	0.974	1.172	1.316	0.206	4.4e-01	3.9e-01	3.4e-01	1.9e+00	1.417	1.565	1.654	2.071
100	5	1.200	1.402	1.554	3.159	4.4e-01	4.3e-01	4.3e-01	6.9e-01	1.644	1.832	1.984	3.848
100	5	1.563	2.280	2.189	3.557	5.6e-01	5.1e-01	5.2e-01	1.5e+00	2.121	2.790	2.706	5.025
100	5	1.045	1.392	1.500	3.480	5.1e-01	5.3e-01	5.5e-01	9.2e-01	1.551	1.920	2.049	4.403
100	5	1.129	1.412	1.513	0.281	2.0e-01	1.6e-01	1.5e-01	1.5e+00	1.330	1.567	1.665	1.815
100	5	1.289	1.515	1.603	0.474	1.7e-01	1.5e-01	1.2e-01	8.9e-01	1.464	1.661	1.719	1.361
100	10	1.878	2.048	2.203	2.964	4.7e-01	4.5e-01	4.3e-01	9.2e-01	2.349	2.493	2.631	3.880
100	10	1.460	1.955	1.998	3.284	7.6e-01	7.4e-01	6.8e-01	1.5e+00	2.215	2.697	2.677	4.798
100	10	1.656	1.904	2.071	2.517	6.2e-01	5.9e-01	5.7e-01	2.2e+00	2.272	2.499	2.640	4.766
100	10	1.329	1.519	1.621	0.647	2.5e-01	2.0e-01	1.8e-01	1.5e+00	1.580	1.715	1.801	2.117
100	10	1.391	1.542	1.634	0.461	2.3e-01	1.9e-01	1.5e-01	1.3e+00	1.619	1.734	1.788	1.728

Table C.17: Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) of the first 5 varying coefficients of **Examples 3 and 4 (**$N = 20, P = 150, 200, 1000$), using SSP, SSP0, group LASSO and GAM.

		IBIAS^2				IVAR				IMSE			
P	P_{true}	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
150	5	1.326	1.638	1.849	3.720	3.6e-01	3.9e-01	3.9e-01	4.2e-01	1.684	2.027	2.238	4.144
150	5	1.818	2.566	2.501	3.287	4.3e-01	4.1e-01	4.2e-01	1.3e+00	2.249	2.981	2.922	4.610
150	5	1.314	1.678	1.817	3.623	5.6e-01	5.3e-01	5.4e-01	6.8e-01	1.879	2.212	2.352	4.299
150	5	1.464	1.647	1.726	0.614	1.5e-01	1.1e-01	8.7e-02	8.7e-01	1.611	1.760	1.813	1.483
150	5	1.266	1.531	1.585	0.460	2.0e-01	1.5e-01	1.3e-01	7.9e-01	1.467	1.678	1.717	1.249
150	10	2.311	2.473	2.531	3.117	3.8e-01	3.9e-01	4.1e-01	9.9e-01	2.695	2.863	2.940	4.106
150	10	1.853	2.460	2.514	3.619	6.1e-01	5.5e-01	5.0e-01	1.7e+00	2.467	3.013	3.016	5.317
150	10	1.887	2.068	2.164	3.644	6.2e-01	6.2e-01	6.1e-01	9.8e-01	2.507	2.691	2.772	4.628
150	10	1.434	1.572	1.677	0.879	2.3e-01	2.0e-01	1.6e-01	1.1e+00	1.668	1.775	1.832	2.007
150	10	1.445	1.616	1.703	0.764	1.8e-01	1.6e-01	1.4e-01	1.2e+00	1.626	1.774	1.847	1.945
200	5	1.552	1.757	1.917	4.151	5.6e-01	5.4e-01	5.0e-01	2.3e-01	2.110	2.292	2.417	4.381
200	5	1.670	2.438	2.290	4.455	4.3e-01	4.1e-01	4.3e-01	8.1e-01	2.102	2.851	2.718	5.265
200	5	1.629	2.030	2.115	4.433	4.1e-01	4.0e-01	4.1e-01	3.2e-01	2.035	2.426	2.525	4.751
200	5	1.494	1.665	1.735	0.957	1.6e-01	1.3e-01	1.2e-01	6.4e-01	1.657	1.793	1.851	1.601
200	5	1.536	1.727	1.751	0.705	9.5e-02	6.7e-02	6.8e-02	9.0e-01	1.631	1.793	1.820	1.604
200	10	2.147	2.276	2.390	3.570	5.8e-01	5.3e-01	4.9e-01	5.7e-01	2.726	2.803	2.876	4.137
200	10	2.153	2.664	2.685	4.331	6.2e-01	5.6e-01	5.4e-01	9.2e-01	2.777	3.229	3.221	5.248
200	10	2.832	3.002	3.092	4.917	4.4e-01	4.3e-01	4.2e-01	2.4e-02	3.275	3.433	3.511	4.941
200	10	1.648	1.738	1.799	1.139	2.2e-01	1.8e-01	1.5e-01	7.2e-01	1.868	1.921	1.948	1.864
200	10	1.613	1.721	1.798	1.064	1.3e-01	1.0e-01	8.6e-02	5.9e-01	1.745	1.824	1.884	1.659
1000	5	3.157	3.334	3.448	NA	2.3e-01	2.0e-01	1.7e-01	NA	3.385	3.532	3.618	NA
1000	5	2.676	3.262	3.189	NA	4.4e-01	4.0e-01	4.0e-01	NA	3.114	3.659	3.584	NA
1000	5	3.028	3.184	3.320	NA	3.7e-01	3.5e-01	3.1e-01	NA	3.403	3.530	3.627	NA
1000	5	1.862	1.953	1.997	NA	5.3e-02	4.2e-02	3.3e-02	NA	1.915	1.995	2.031	NA
1000	5	1.862	1.945	1.972	NA	6.2e-02	4.6e-02	4.0e-02	NA	1.925	1.992	2.012	NA
1000	10	3.936	3.961	3.986	NA	1.8e-01	1.6e-01	1.6e-01	NA	4.114	4.123	4.149	NA
1000	10	3.338	3.622	3.830	NA	4.9e-01	4.5e-01	4.1e-01	NA	3.825	4.069	4.236	NA
1000	10	3.642	3.750	3.885	NA	2.8e-01	2.5e-01	2.3e-01	NA	3.921	3.998	4.114	NA
1000	10	1.949	1.999	2.018	NA	3.5e-02	2.4e-02	2.3e-02	NA	1.984	2.023	2.041	NA
1000	10	1.934	1.955	1.976	NA	7.6e-02	7.9e-02	6.8e-02	NA	2.010	2.034	2.044	NA

Table C.18: The **aggregated** Integrated Squared Bias (IBIAS^2), Integrated Variance (IVAR) and Integrated Mean Square Error (IMSE) across all the varying coefficents of **Examples 3 and 4 (**$N = 20$), using SSP, SSP0, group LASSO and GAM.

		IBIAS^2				IVAR				IMSE			
P	P_{true}	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM	SSP	SSP0	gLASSO	GAM
50	5	4.447	6.330	6.819	8.742	3.141	3.278	3.161	67.968	7.588	9.608	9.979	76.710
	10	12.162	14.919	15.893	13.916	8.636	8.631	7.924	80.401	20.798	23.550	23.817	94.317
100	5	6.256	8.031	8.385	13.132	3.274	3.124	3.041	56.516	9.530	11.155	11.426	69.648
	10	16.753	19.262	20.084	22.470	8.684	8.165	7.609	82.141	25.436	27.428	27.693	104.610
150	5	7.221	9.090	9.506	13.391	3.181	3.028	2.998	52.189	10.402	12.118	12.504	65.580
	10	19.276	21.556	22.272	27.436	8.313	7.857	7.274	76.277	27.589	29.413	29.547	103.712
200	5	7.909	9.644	9.833	16.301	3.009	2.835	2.690	54.774	10.918	12.479	12.523	71.075
	10	21.991	24.006	24.711	31.823	7.784	7.308	6.770	69.310	29.776	31.314	31.481	101.133
1000	5	12.617	13.707	13.954	NA	2.761	2.521	2.369	NA	15.379	16.228	16.323	NA
	10	30.592	31.621	32.434	NA	5.945	5.494	4.860	NA	36.537	37.116	37.295	NA

References

Adusumalli, S., Mohd Omar, M. F., Soong, R., & Benoukraf, T. (2015). Methodological aspects of whole-genome bisulfite sequencing analysis. *Briefings in bioinformatics*, *16*(3), 369–379.

Affinito, O., Palumbo, D., Fierro, A., Cuomo, M., De Riso, G., Monticelli, A., ... Cocozza, S. (2020). Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics*, *112*(1), 144–150.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., & Mason, C. E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome biology*, *13*(10), 1–9.

Allum, F., Hedman, A. K., Shao, X., Cheung, W. A., Vijay, J., Guénard, F., ... Grundberg, E. (2019). Dissecting features of epigenetic variants underlying cardiometabolic risk using full-resolution epigenome profiling in regulatory elements. *Nature communications*, *10*(1), 1–13.

Allum, F., Shao, X., Guénard, F., Simon, M.-M., Busche, S., Caron, M., ... Grundberg, E. (2015). Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nature communications*, *6*, 7211.

Antoniadis, A., & Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, *96*(455), 939–967.

Ball, M. P., Li, J. B., Gao, Y., Lee, J.-H., LeProust, E. M., Park, I.-H., ... Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology*, *27*(4), 361–368.

Balsa, A., Cabezón, A., Orozco, G., Cobo, T., Miranda-Carus, E., López-Nevot, M. Á., ... Pascual-Salcedo, D. (2010). Influence of HLA DRB1 alleles in the susceptibility of rheumatoid arthritis and the regulation of antibodies against citrullinated proteins and rheumatoid factor. *Arthritis research & therapy*, *12*(2), R62.

Banovich, N. E., Lan, X., McVicker, G., Van de Geijn, B., Degner, J. F., Blischak, J. D., ... Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*, *10*(9), e1004663.

Barber, R. F., Reimherr, M., & Schill, T. (2017). The function-on-scalar LASSO with applications to longitudinal GWAS. *Electronic Journal of Statistics*, *11*(1), 1351–1389.

Barros, S. P., & Offenbacher, S. (2009). Epigenetics: connecting environment and genotype to phenotype and disease. *Journal of dental research*, *88*(5), 400–408.

Berk, R., & MacDonald, J. M. (2008). Overdispersion and Poisson regression. *Journal of Quantitative Criminology*, *24*(3), 269–284.

Bilodeau, M. (1992). Fourier smoother and additive models. *Canadian Journal of Statistics*, *20*(3), 257–269.

Bock, C., Beerman, I., Lien, W.-H., Smith, Z. D., Gu, H., Boyle, P., ... Meissner, A. (2012). DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Molecular cell*, *47*(4), 633–647.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, *88*(421), 9–25.

Browne, W. J., Subramanian, S. V., Jones, K., & Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(3), 599–613.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., ... Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, *47*(D1), D1005–D1012.

Cheng, L., & Zhu, Y. (2013). A classification approach for DNA methylation profiling with bisulfite next-generation sequencing data. *Bioinformatics*, *30*(2), 172–179.

Cheung, W. A., Shao, X., Morin, A., Siroux, V., Kwan, T., Ge, B., ... Grundberg, E. (2017). Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. *Genome biology*, *18*(1), 1–21.

Chouldechova, A., & Hastie, T. (2015). Generalized additive model selection. *arXiv preprint arXiv:1506.03850*.

Choy, M.-K., Movassagh, M., Goh, H.-G., Bennett, M. R., Down, T. A., & Foo, R. S. (2010). Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. *BMC genomics*, *11*(1), 519.

Claeskens, G., Krivobokova, T., & Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, *96*(3), 529–544.

Cui, S., Ji, T., Li, J., Cheng, J., & Qiu, J. (2016). What if we ignore the random effects when analyzing RNA-seq data in a multifactor experiment. *Statistical applications in genetics and molecular biology*, *15*(2), 87–105.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (No. 1). Cambridge university press.

de Boor, C. (1978). *A practical guide to splines*. Springer, New York.

De Jager, P. L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L. C., Yu, L., ... Bennett, D. A. (2014). Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature neuroscience*, *17*(9), 1156.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 1–38.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, *43*(5), 491.

Dolinoy, D. C., Huang, D., & Jirtle, R. L. (2007). Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proceedings of the National Academy of Sciences*, *104*(32), 13056–13061.

Dolzhenko, E., & Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC bioinformatics*, *15*(1), 215.

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*(432), 1200–1224.

Dunaway, K. W., Islam, M. S., Coulson, R. L., Lopez, S. J., Ciernia, A. V., Chu, R. G., ... LaSalle, J. M. (2016). Cumulative impact of polychlorinated biphenyl and large chromosomal duplications on DNA methylation, chromatin, and expression of autism candidate genes. *Cell reports*, *17*(11), 3035–3048.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., ... Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics*, *38*(12), 1378–1385.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, *81*(395), 709–721.

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

Eilers, P. H., Marx, B. D., & Durbán, M. (2015). Twenty years of P-splines. *SORT: statistics and operations research transactions*, *39*(2), 0149–186.

Elashoff, M., & Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, *13*(1), 48–65.

Fan, J., Hu, J., Xue, C., Zhang, H., Susztak, K., Reilly, M. P., ... Li, M. (2020). ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genetics*, *16*(5), e1008786.

Fan, J., Ma, Y., & Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, *109*(507), 1270–1284.

Fan, Y., Vilgalys, T. P., Sun, S., Peng, Q., Tung, J., & Zhou, X. (2019). IMAGE: high-powered detection of genetic effects on DNA methylation using integrated methylation QTL mapping and allele-specific analysis. *Genome biology*, *20*(1), 1–18.

Farrington, C. (1995). Pearson statistics, goodness of fit, and overdispersion in generalised linear models. In *Statistical modelling* (pp. 109–116). Springer.

Feinberg, A. P. (2007). Phenotypic plasticity and the epigenetics of human disease. *Nature*, *447*(7143), 433.

Feng, H., Conneely, K. N., & Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, *42*(8), e69–e69.

Fletcher, D. (2012). Estimating overdispersion when fitting a generalized linear model to sparse data. *Biometrika*, *99*(1), 230–237.

Forslind, K., Ahlmén, M., Eberhardt, K., Hafström, I., & Svensson, B. (2004). Prediction of radiological outcome in early rheumatoid arthritis in clinical practice: role of antibodies to citrullinated peptides (anti-CCP). *Annals of the rheumatic diseases*, *63*(9), 1090–1095.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of applied statistics*, *1*(2), 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., ... Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, *89*(5), 1827–1831.

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., ... Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome biology*, *17*(1), 1–14.

Gertheiss, J., Maity, A., & Staicu, A.-M. (2013). Variable selection in generalized functional linear models. *Stat*, *2*(1), 86–101.

Goeman, J. J., Van De Geer, S. A., & Van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(3), 477–493.

Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. Chapman & Hall.

Grueber, C. E., Nakagawa, S., Laws, R. J., & Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: challenges and solutions. *Journal of evolutionary biology*, *24*(4), 699–711.

Grundberg, E., Meduri, E., Sandling, J. K., Hedman, A. K., Keildson, S., Buil, A., ... Deloukas, P. (2013). Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *The American Journal of Human Genetics*, *93*(5), 876–890.

Gu, C. (1992). Cross-validating non-Gaussian data. *Journal of Computational and Graphical Statistics*, *1*(2), 169–179.

Gu, C., & Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computing*, *12*(2), 383–398.

Gu, L., Frommel, S. C., Oakes, C. C., Simon, R., Grupp, K., Gerig, C. Y., ... Santoro, R. (2015). BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer and its overexpression predicts disease recurrence. *Nature genetics*, *47*(1), 22–30.

Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., Montgomery, S. B., Buil, A., Yurovsky, A., ... Dermitzakis, E. T. (2015). Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet*, *11*(1), e1004958.

Hall, P., & Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, *92*(1), 105–118.

Handayani, D., Notodiputro, K. A., Sadik, K., & Kurnia, A. (2017). A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (glmm). In *Aip conference proceedings* (Vol. 1827, p. 020033).

Hannon, E., Knox, O., Sugden, K., Burrage, J., Wong, C. C., Belsky, D. W., ... Mill, J. (2018). Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS genetics*, *14*(8), e1007544.

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., ... Mill, J. (2016). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature neuroscience*, *19*(1), 48–54.

Hansen, K. D., Langmead, B., & Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, *13*(10), R83.

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., ... Feinberg, A. P. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, *43*(8), 768.

Hanson, M. A., & Gluckman, P. D. (2008). Developmental origins of health and disease: new insights. *Basic & clinical pharmacology & toxicology*, *102*(2), 90–93.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616.

Hastie, T., & Tibshirani, R. (1987). Generalized Additive Models. *Statistical Science*, *1*(3), 297–318.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(4), 757–779.

Hebestreit, K., Dugas, M., & Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, *29*(13), 1647–1653.

Heyde, C., & Morton, R. (1996). Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(2), 317–327.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press, Cambridge.

Hillier, L. W., Graves, T. A., Fulton, R. S., Fulton, L. A., Pepin, K. H., Minx, P., ... Wilson, R. K. (2005). Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature*, *434*(7034), 724.

Hinks, A., Worthington, J., & Thomson, W. (2006). The association of PTPN22 with rheumatoid arthritis and juvenile idiopathic arthritis. *Rheumatology*, *45*(4), 365–368.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome biology*, *14*(10), 3156.

Hu, M., Yao, J., Cai, L., Bachman, K. E., Van Den Brûle, F., Velculescu, V., & Polyak, K. (2005). Distinct epigenetic changes in the stromal cells of breast cancers. *Nature genetics*, *37*(8), 899–905.

Huang, J., Horowitz, J. L., & Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics*, *38*(4), 2282.

Hudson, M., Bernatsky, S., Colmegna, I., Lora, M., Pastinen, T., Klein Oros, K., & Greenwood, C. M. (2017). Novel insights into systemic autoimmune rheumatic diseases using shared molecular signatures and an integrative analysis. *Epigenetics*, *12*(6), 433–440.

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., ... Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*, *41*(2), 178–186.

Ivanova, A., Molenberghs, G., & Verbeke, G. (2014). A model for overdispersed hierarchical ordinal data. *Statistical Modelling*, *14*(5), 399–415.

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, *33*, 245.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions*. John Wiley & Sons, Ltd.

Jones, P. A. (1999). The DNA methylation paradox. *Trends in Genetics*, *15*(1), 34–37.

Jones, P. A., Ohtani, H., Chakravarthy, A., & De Carvalho, D. D. (2019). Epigenetic therapy in immune-oncology. *Nature Reviews Cancer*, *19*(3), 151–161.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *49*(2), 127–145.

Ju, K., Lin, L., Chu, H., Cheng, L.-L., & Xu, C. (2020). Laplace approximation, penalized quasi-likelihood, and adaptive gauss–hermite quadrature for generalized linear mixed models: towards meta-analysis of binary outcome with sparse data. *BMC Medical Research Methodology*, *20*(1), 1–11.

Karabegović, I., Portilla-Fernandez, E., Li, Y., Ma, J., Maas, S. C., Sun, D., ... Ghanbari, M. (2021). Epigenome-wide association meta-analysis of DNA methylation with coffee and tea consumption. *Nature communications*, *12*(1), 1–13.

Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, *84*(407), 717–726.

Kato, T., Iwamoto, K., Kakiuchi, C., Kuratomi, G., & Okazaki, Y. (2005). Genetic or epigenetic difference causing discordance between monozygotic twins as a clue to molecular basis of mental disorders. *Molecular psychiatry*, *10*(7), 622–630.

Kauermann, G., Krivobokova, T., & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 487–503.

Khavari, D. A., Sen, G. L., & Rinn, J. L. (2010). DNA methylation and epigenetic control of cellular differentiation. *Cell cycle*, *9*(19), 3880–3883.

Korthauer, K., Chakraborty, S., Benjamini, Y., & Irizarry, R. A. (2018). Detection and accurate False Discovery Rate control of differentially methylated regions from Whole Genome Bisulfite Sequencing. *Biostatistics*.

Krueger, F., Kreck, B., Franke, A., & Andrews, S. R. (2012). DNA methylome analysis using short bisulfite sequencing data. *Nature methods*, *9*(2), 145.

Kulis, M., & Esteller, M. (2010). DNA methylation and cancer. *Advances in genetics*, *70*, 27–56.

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

Lakhal-Chaieb, L., Greenwood, C. M., Ouhourane, M., Zhao, K., Abdous, B., & Oualkacha, K. (2017). A smoothed EM-algorithm for DNA methylation profiles from sequencing-based methods in cell lines or for a single cell type. *Statistical applications in genetics and molecular biology*, *16*(5-6), 333–347.

Lea, A. J., Tung, J., & Zhou, X. (2015). A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS genetics*, *11*(11), e1005650.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*(10), 733–739.

Li, Q., Suzuki, M., Wendt, J., Patterson, N., Eichten, S. R., Hermanson, P. J., ... Greally, J. M. (2015). Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. *Nucleic acids research*, *43*(12), e81–e81.

Lin, Y., & Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, *34*(5), 2272–2297.

Lindsay, B. (1982). Conditional score functions: some optimality results. *Biometrika*, *69*(3), 503–512.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., ... Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *nature*, *462*(7271), 315–322.

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., ... Feinberg, A. P. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology*, *31*(2), 142.

Locke, W. J., Guanzon, D., Ma, C., Liew, Y. J., Duesing, K. R., Fung, K. Y., & Ross, J. P. (2019). DNA methylation cancer biomarkers: translation to the clinic. *Frontiers in genetics*, *10*, 1150.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 226–233.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., ... Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research*, *45*(D1), D896–D901.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News*, *456*(7218), 18–21.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., ... Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 1222794.

Mazzone, R., Zwergel, C., Artico, M., Taurone, S., Ralli, M., Greco, A., & Mai, A. (2019). The emerging role of epigenetics in human autoimmune disorders. *Clinical epigenetics*, *11*(1), 1–15.

McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models. *International Statistical Review/Revue Internationale de Statistique*, 61–67.

McCullagh, P., & Nelder, J. A. (1989a). *Generalized linear models, 2nd den*. Chapman and Hall, London.

McCullagh, P., & Nelder, J. A. (1989b). Generalized Linear Models 2nd Edition Chapman and Hall. *London, UK*.

McGregor, K., Bernatsky, S., Colmegna, I., Hudson, M., Pastinen, T., Labbe, A., & Greenwood, C. M. (2016). An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome biology*, *17*(1), 84.

McRae, A. F., Powell, J. E., Henders, A. K., Bowdler, L., Hemani, G., Shah, S., ... Montgomery, G. W. (2014). Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome biology*, *15*(5), R73.

Meaney, M. J., & Szyf, M. (2005). Environmental programming of stress responses through DNA methylation: life at the interface between a dynamic environment and a fixed genome. *Dialogues in clinical neuroscience*, *7*(2), 103.

Meier, L., Van de Geer, S., & Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, *37*(6B), 3779-3821.

Miranda-Morales, E., Meier, K., Sandoval-Carrillo, A., Salas-Pacheco, J., Vázquez-Cárdenas, P., & Arias-Carrión, O. (2017). Implications of DNA methylation in Parkinson's disease. *Frontiers in molecular neuroscience*, *10*, 225.

Molenberghs, G., Verbeke, G., & Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis*, *13*(4), 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C. G., & Vieira, A. M. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical science*, *25*(3), 325–347.

Molenberghs, G., Verbeke, G., Iddi, S., & Demétrio, C. G. (2012). A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis*, *111*, 94–109.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, *2*, 321–359.

Morris, J. S., & Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 179–199.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, *140*(1), 125–161.

Nilsson, E., Jansson, P. A., Perfilyev, A., Volkov, P., Pedersen, M., Svensson, M. K., ... Ling, C. (2014). Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes*, *63*(9), 2962–2976.

Noh, H. S., & Park, B. U. (2010). Sparse varying coefficient models for longitudinal data. *Statistica Sinica*, 1183–1202.

Oakes, D. (1999). Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 479–482.

Ober, C., & Vercelli, D. (2011). Gene–environment interactions in human disease: nuisance or opportunity? *Trends in genetics*, *27*(3), 107–115.

Orozco, G., Abelson, A.-K., González-Gay, M. A., Balsa, A., Pascual-Salcedo, D., García, A., ... Martín, J. (2009). Study of functional variants of the BANK1 gene in rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, *60*(2), 372–379.

Parikh, N., & Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in optimization*, *1*(3), 127–239.

Park, Y., Figueroa, M. E., Rozek, L. S., & Sartor, M. A. (2014). MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, *30*(17), 2414–2422.

Park, Y., & Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, *32*(10), 1446–1453.

Parker, R., & Rice, J. (1985). Discussion on "some aspects of the spline smoothing approach to non-parametric regression curve fitting" (by B. W. Silverman). *Journal of the Royal Statistical Society. Series B (methodological)*, 40–42.

Payne, A. C., Chiang, Z. D., Reginato, P. L., Mangiameli, S. M., Murray, E. M., Yao, C.-C., ... Chen, F. (2021). In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science*, *371*(6532).

Prochenka, A., Pokarowski, P., Gasperowicz, P., Kosińska, J., Stawiński, P., Zbieć-Piekarska, R., ... Płoski, R. (2015). A cautionary note on using binary calls for analysis of DNA methylation. *Bioinformatics*, *31*(9), 1519–1520.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, *2*(1), 1–21.

Rackham, O. J., Langley, S. R., Oates, T., Vradi, E., Harmston, N., Srivastava, P. K., ... Petretto, E. (2017). A Bayesian Approach for Analysis of Whole-Genome Bisulphite Sequencing Data Identifies Disease-Associated Changes in DNA Methylation. *Genetics*, genetics–116.

Ravikumar, P., Lafferty, J., Liu, H., & Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(5), 1009–1030.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik*, *10*(3), 177–183.

Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics*, *6*(8), 597–610.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression* (No. 12). Cambridge university press.

Schmid, M., & Hothorn, T. (2008). Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis*, *53*(2), 298–311.

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., ... Ecker, J. R. (2013). Patterns of population epigenomic diversity. *Nature*, *495*(7440), 193–198.

Schoofs, T., Rohde, C., Hebestreit, K., Klein, H.-U., Göllner, S., Schulze, I., ... Müller-Tidow, C. (2013). DNA methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. *Blood, The Journal of the American Society of Hematology*, *121*(1), 178–187.

Shafi, A., Mitrea, C., Nguyen, T., & Draghici, S. (2017). A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in bioinformatics*.

Shao, X., Hudson, M., Colmegna, I., Greenwood, C. M., Fritzler, M. J., Awadalla, P., ... Bernatsky, S. (2019). Rheumatoid arthritis-relevant DNA methylation changes identified in ACPA-positive asymptomatic individuals using methylome capture sequencing. *Clinical epigenetics*, *11*(1), 110.

Shokoohi, F., Stephens, D. A., Bourque, G., Pastinen, T., Greenwood, C. M., & Labbe, A. (2018). A hidden markov model for identifying differentially methylated sites in bisulfite sequencing data. *Biometrics*.

Shun, Z., & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(4), 749–760.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *The annals of Statistics*, 898–916.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, *47*(1), 1–21.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–132.

Skvortsova, K., Stirzaker, C., & Taberlay, P. (2019). The DNA methylation landscape in cancer. *Essays in biochemistry*, *63*(6), 797–811.

Small, C. G., Christopher, G., & Wang, J. (2003). *Numerical methods for nonlinear estimating equations* (Vol. 29). Oxford University Press on Demand.

Stenz, L., Schechter, D. S., Serpa, S. R., & Paoloni-Giacobino, A. (2018). Intergenerational transmission of DNA methylation signatures associated with early life stress. *Current genomics*, *19*(8), 665–675.

Stephens, D. A., Shokoohi, F., & Aurélie, L. (2016). Hidden Markov models for identifying differentially methylated regions. 44th Annual Meeting of the Statistical Society of Canada.

Sun, S., & Yu, X. (2016). HMM-Fisher: identifying differential methylation using a hidden Markov model and Fisher's exact test. *Statistical applications in genetics and molecular biology*, *15*(1), 55–67.

Taylor, D. L., Jackson, A. U., Narisu, N., Hemani, G., Erdos, M. R., Chines, P. S., ... Collins, F. S. (2019). Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proceedings of the National Academy of Sciences*, *116*(22), 10883–10888.

Teschendorff, A. E., Yang, Z., Wong, A., Pipinikas, C. P., Jiao, Y., Jones, A., ... Widschwendter, M. (2015). Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. *JAMA oncology*, *1*(4), 476–485.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tough, D. F., Tak, P. P., Tarakhovsky, A., & Prinjha, R. K. (2016). Epigenetic drug discovery: breaking through the immune barrier. *Nature Reviews Drug Discovery*, *15*(12), 835–853.

Trerotola, M., Relli, V., Simeone, P., & Alberti, S. (2015). Epigenetic inheritance and the missing heritability. *Human genomics*, *9*(1), 1–12.

Tutz, G., & Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, *62*(4), 961–971.

Vahabi, N., Kazemnejad, A., & Datta, S. (2019). A joint overdispersed marginalized random-effects model for analyzing two or more longitudinal ordinal responses. *Statistical Methods in Medical Research*, *28*(1), 50–69.

Van Dongen, J., Nivard, M. G., Willemsen, G., Hottenga, J.-J., Helmer, Q., Dolan, C. V., ... Boomsma, D. I. (2016). Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature communications*, *7*(1), 1–13.

Wahba, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation theory III*, *2*.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society: Series B (Methodological)*, *45*(1), 133–150.

Wahba, G., Wang, Y., Gu, C., Klein, R., & Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy: the 1994 Neyman Memorial Lecture. *The Annals of Statistics*, *23*(6), 1865–1895.

Wang, H., & Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, *104*(486), 747–757.

Wang, L., Chen, G., & Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, *23*(12), 1486–1494.

Wang, L., Li, H., & Huang, J. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, *103*(484), 1556–1569.

Wang, R. Y.-H., Gehrke, C. W., & Ehrlich, M. (1980). Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Research*, *8*(20), 4777–4790.

Wei, F., Huang, J., & Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, *21*(4), 1515.

Weissbrod, O., Rahmani, E., Schweiger, R., Rosset, S., & Halperin, E. (2017). Association testing of bisulfite-sequencing methylation data via a Laplace approximation. *Bioinformatics*, *33*(14), i325–i332.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(2), 413–428.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36.

Wood, S. N. (2013a). On p-values for smooth components of an extended generalized additive model. *Biometrika*, *100*(1), 221–228.

Wood, S. N. (2013b). On p-values for smooth components of an extended generalized additive model. *Biometrika*, *100*(1), 221–228.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Wood, S. N., & Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*, *73*(4), 1071–1081.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563.

Wreczycka, K., Gosdschan, A., Yusuf, D., Gruening, B., Assenov, Y., & Akalin, A. (2017). Strategies for analyzing bisulfite sequencing data. *Journal of biotechnology*, *261*, 105–115.

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., ... Conneely, K. N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic acids research*, *43*(21), e141–e141.

Xue, L., & Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research*.

Yee, T. W., & Wild, C. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(3), 481–493.

Young, A. I., Wauthier, F. L., & Donnelly, P. (2018). Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nature genetics*, *50*(11), 1608–1614.

Young, J. I., Sivasankaran, S. K., Wang, L., Ali, A., Mehta, A., Davis, D. A., ... Vance, J. M. (2019). Genome-wide brain DNA methylation analysis suggests epigenetic reprogramming in Parkinson disease. *Neurology Genetics*, *5*(4).

Yu, X., & Sun, S. (2016a). Comparing five statistical methods of differential methylation identification using bisulfite sequencing data. *Statistical applications in genetics and molecular biology*, *15*(2), 173–191.

Yu, X., & Sun, S. (2016b). HMM-DM: identifying differentially methylated regions using a hidden Markov model. *Statistical applications in genetics and molecular biology*, *15*(1), 69–81.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zhang, H., Wang, L., Huang, Y., Zhuang, C., Zhao, G., Liu, R., & Wang, Y. (2012). Influence of BLK polymorphisms on the risk of rheumatoid arthritis. *Molecular biology reports*, *39*(11), 9965–9970.

Zhang, Z., Liu, D., Murugan, A. K., Liu, Z., & Xing, M. (2014). Histone deacetylation of NIS promoter underlies BRAF V600E-promoted NIS silencing in thyroid cancer. *Endocrine-related cancer*, *21*(2), 161.

Zhao, K., Oualkacha, K., Lakhal-Chaieb, L., Labbe, A., Klein, K., Bernatsky, S., ... Greenwood, C. M. (2021). A hierarchical quasi-binomial varying coefficient mixed model for detecting differentially methylated regions in bisulfite sequencing data. *arXiv preprint arXiv:2101.07374*.

Zhao, K., Oualkacha, K., Lakhal-Chaieb, L., Labbe, A., Klein, K., Ciampi, A., ... Greenwood, C. M. (2020). A novel statistical method for modeling covariate effects in bisulfite sequencing derived measures of DNA methylation. *Biometrics*.

Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, *11*(4), 407–409.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., ... Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, *500*(7463), 477–481.

Ziller, M. J., Stamenova, E. K., Gu, H., Gnirke, A., & Meissner, A. (2016). Targeted bisulfite sequencing of the dynamic DNA methylome. *Epigenetics & chromatin*, *9*(1), 1–9.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zouali, M. (2020). DNA methylation signatures of autoimmune diseases in human B lymphocytes. *Clinical Immunology*, 108622.

Zulet, M. I., Fontes, L. P., Blanco, T. A., Bescos, F. L., & Iriarte, M. M. (2017). Epigenetic changes in neurology: Dna methylation in multiple sclerosis. *Neurología (English Edition)*, *32*(7), 463–468.