January 2015

On New Approaches for Variable Selection under Single Index Model and DNA Methylation Status Calling

Longjie Cheng
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Longjie Cheng

Entitled

On New Approaches for Variable Selection under Single Index Model and DNA Methylation Status Calling

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Michael Yu Zhu
Chair

Rebecca W. Doerge

Jun Xie

Mary Ellen Bock

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Michael Yu Zhu

Approved by: Jun Xie 11/24/2015

Head of the Departmental Graduate Program Date

ON NEW APPROACHES FOR VARIABLE SELECTION UNDER SINGLE INDEX MODEL AND DNA METHYLATION STATUS CALLING

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Longjie Cheng

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2015

Purdue University

West Lafayette, Indiana

To my family.

ACKNOWLEDGMENTS

First of all, I would like to express my most sincere gratitute towards my advisor, Professor Michael Yu Zhu. He not only gives me considerable guidance and encouragement on my research, but also gives me valuable advice on my career and life.

I would also like to thank Professor Rebecca W. Doerge, Professor Jun Xie, and Professor Mary Ellen Bock, for their time serving as my committee members. I thank them for their valuable suggestions and inputs on my thesis.

I wish to express my appreciation to Professor Peng Zeng at Auburn Unversity for his suggestions on my research.

It would be remiss of me if I forget to thank the faculty members, the staff, and the fellow students at Department of Statistics. They help me in various ways, and make my time at Purdue a great experience. I also would like to thank the members in Professor Zhu's research group for their helpful suggestions on my research and presentation. I always enjoy the group meeting, and learn a great deal from them.

I must not forget to thank my friends, who continuously understand, support and tolerate me. They bring so much joy to my life. I am grateful that I have them to share the happiness and stress in my life.

Finally, I would like to express my wholehearted gratitute to my family for their unparalleled and unconditional love. I will be no place near where I am without them. I dedicate this work, and all of my future achievements to them.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES ix

ABSTRACT x

1 Variable Selection for High-dimensional Single Index Model 1

1.1 Introduction 1

1.1.1 Single Index Model 1

1.1.2 Variable Selection Methods for SIM 6

1.1.3 Review of the SICA penalty functions 8

1.1.4 Review of B-Splines 10

1.2 BS-SIM: A Spline Estimation and Regularization Method for Single Index Model 12

1.3 Implementation for BS-SIM 15

1.3.1 Coordinate Descent Algorithm for BS-SIM 15

1.3.2 Tuning Parameter Selection for BS-SIM 18

1.4 Theoretical Properties for BS-SIM 21

1.4.1 Estimation Consistency 21

1.4.2 Intuition and Notations for Selection Consistency 21

1.4.3 Selection Consistency 24

1.5 Simulation Studies 28

1.5.1 Performance of the proposed method for small p 29

1.5.2 Performance of the proposed method compared to that of the unpenalized estimator 30

1.5.3 Performance of the proposed method for several choices of a 33

1.5.4 Performance of the proposed method for moderate p 35

Page

1.5.5 Performance of the proposed method for large p 36
1.5.6 Evaluation of the Irrepresentable Conditions 38
1.5.7 Comparison of CV, logGIC and GIC under the violation of the sparsity assumption 42
1.6 Real Data Application 44
1.6.1 Skin Cutaneous Melanoma Data 44
1.6.2 Analysis on Skin Cutaneous Melanoma Data with BS-SIM . 45
1.7 Linearly Constrained Single Index Model 47
1.7.1 Single Index Model with Linear Constraints 47
1.7.2 Coordinate Descent Algorithm for Linearly Constrained Single Index Model 48
1.8 Proofs 52
1.8.1 Regularity Conditions 52
1.8.2 Proof of Theorem 1.4.1 53
1.8.3 Proof of Theorem 1.4.2 54
1.8.4 Proof of Theorem 1.4.3 57
1.8.5 Proof of Corollary 1.4.4 59
2 DNA Methylation Status Quantification for Bisulphite-sequencing Data . 60
2.1 Introduction 60
2.1.1 Introduction to DNA Methylation 60
2.1.2 Review of Bisulphite-sequencing Experiment 63
2.1.3 Review of Quantification Methods for Bisulphite-sequencing Data 65
2.1.4 Review of False Discovery Rate Controlling Procedures . . . 66
2.2 Methods 68
2.2.1 Mixture of Binomial Model 68
2.2.2 Classification based Methylation Status Calling Procedure . 73
2.2.3 Performance Assessment of the MSC Procedure 74
2.2.4 Methylation Status Calling Procedure with FDR control . . 76

Page

2.2.5 EM Algorithm for Computing the Parameters 79
2.3 Simulation Results 80
2.3.1 Performance of MSC and FMSC 80
2.3.2 Estimation of Correct Allocation Rates 83
2.3.3 Choice of Null Hypothesis in FDR control 83
2.3.4 Estimation of FDR and FNDR with Memberships 84
2.4 Real Data Application 86
2.4.1 Performance of MSC and FMSC 86
2.4.2 Comparison of MSC and FMSC with Existing Methods . . . 88
2.4.3 Coverage Distribution 91
2.5 Recent Development on DNA Methylation Analysis and FDR Controlling Procedures 93
2.5.1 DNA methylation status quantification for Bisulphite-sequencing data 93
2.5.2 Sequencing-based DNA Methylation Profiling Approaches . 95
2.5.3 FDR Controlling Procedures for Discrete Tests 96
3 Future Work 97
3.1 Future Research Topics for Variable Selection under Single Index Model 97
3.2 Future Research Topics for DNA Methylation Status Calling 98
REFERENCES 99
VITA 106

LIST OF TABLES

Table		Page
1.1	Comparison between our methods to the existing methods in low dimensional scenario: Model 1 with COR1.	31
1.2	Comparison between our methods to the existing methods in low dimensional scenario: Model 3 with COR1.	32
1.3	Comparison between the penalized estimator and the unpenalized estimator.	33
1.4	Comparison between the LASSO and the SICA penalties with various choices of a for moderate p.	34
1.5	Comparison between the proposed methods and the other existing methods in moderate dimensional scenario: Setting 1.	36
1.6	Comparison between the proposed methods and the other existing methods in moderate dimensional scenario: Setting 2.	37
1.7	Comparison between the proposed methods and the other existing methods in moderate dimensional scenario: Setting 3.	37
1.8	Performance of BS-SIM with $a = 0.1$ under several settings in high dimensional scenario.	39
1.9	Average percentages of times that the true model can be selected with various choices of a.	42
1.10	Summary on $\bar{\eta}_\infty$ for both Identifiability Constraints.	43
1.11	Performance comparison of CV, logGIC and GIC when $q = 15$.	44
2.1	Four possible outcomes from multiple testing procedures.	68
2.2	Possible outcomes from the MSC procedure and the FMSC procedure.	74
2.3	Estimation of the overall correct allocation rate and correct allocation rates for the two subgroups.	83
2.4	Chromosome by Chromosome Results with MSC for the MethylC-Seq data from [66].	87
2.5	Contingency Table for Chromosome-wise and Genome-wide Evaluation	88

Table Page

2.6 Assessment of Genome-wide Analysis by MSC and FMSC at three levels. 88

2.7 Comparison of whole-genome results from the MSC procedure and those from the procedure used by [66] for all covered CpG sites. 89

2.8 Typical examples of sites that the MSC procedure declares to be unmethylated but the procedure used by [66] declares otherwise 90

2.9 Third platform validation of the methylation calls for those sites that MSC and the procedure used by [66] disagree on 91

2.10 Maximum Likelihood Estimates of $\hat{v}$ and $\hat{r}$ for Chromosome 1. 92

LIST OF FIGURES

Figure		Page
1.1	Illustration of SICA for several choices of a.	9
1.2	An illustration of the B-spline basis functions of order 4 with 9 equally-spaced interior knots on [0,1].	11
1.3	The percentages that the proposed BL-SIM method and the proposed BS-SIM method with $a = 2$ select the true model versus $\bar{\eta}_\infty$ for both Identifiability Constraints.	41
1.4	The plot of the fitted regression function and the observed log survival time vesus the estimated index for the Skin Cutaneous Melanoma data.	46
2.1	Workflow for MethylC-Seq experiment.	64
2.2	(a) The box plots display FDRs for IBT at level 0.1, the MSC procedure, and the FMSC procedure with FDR level 0.1, 0.05 and 0.01 from left to right. (b) The box plots display FNDRs for these methods in the same order.	82
2.3	Comparison of different choices of null hypothesis. Left: Proportion of methylated sites that are allocated to unmethylated group among those allocated to unmethylated group. Right: Proportion of unmethylated sites that are allocated to methylated group among those allocated to methylated group.	84
2.4	Estimation of FDR and FNDR for $c = (0.5, 0.4, 0.6)$	85
2.5	Histogram of the coverage with blue dots indicating fitted probabilities.	92

ABSTRACT

Cheng, Longjie PhD, Purdue University, December 2015. On New Approaches for Variable Selection under Single Index Model and DNA Methylation Status Calling . Major Professor: Yu Zhu.

This thesis consists of two main components: a regularization based variable selection method for the single index model and a novel classification based method for DNA methylation status calling for bisulphite-sequencing data.

The single index model is an intuitive extension of the linear regression model. It has become increasingly popular due to its flexibility in modeling. Similar to the linear regression model, the set of predictors for the single index model can contain a large number of irrelevant variables. Therefore, it is important to select the relevant variables when fitting the single index model. However, the problem of variable selection for high-dimensional single index model is not well settled in the literature. In the first part of this thesis, we combine the idea of applying cubic B-splines for estimating the single index model with the idea of using the family of the smooth integration of counting and absolute deviation (SICA) penalty functions for variable selection. Based on this combination, a new method is proposed to simultaneously perform parameter estimation and model selection for the single index model. This method is referred to as the B-spline and SICA method for the single index model, or in short, BS-SIM. Since LASSO is a limiting case of SICA, the proposed BS-SIM framework can also be applied if one prefers LASSO. A coordinate descent algorithm is developed to efficiently implement BS-SIM. Moreover, we develop the regularity conditions under which BS-SIM can consistently estimate the parameter and select the true model. Simulations with various settings and a real data analysis are conducted to demonstrate the estimation accuracy, selection consistency and computational efficiency of

BS-SIM. In addition, we also briefly discuss the problem of estimating the single index model with our framework when linear equality and inequality constraints are imposed.

With the advent of high-throughput sequencing technology, bisulphite-sequencing based DNA methylation profiling methods have emerged as the most promising approaches due to their single-base resolution and genome-wide coverage. Nevertheless, statistical analysis methods for analyzing this type of methylation data are not well developed. Although the most widely used proportion based estimation method is simple and intuitive, it is not statistically adequate in dealing with the various sources of noise in bisulphite-sequencing data. Furthermore, it is not biologically satisfactory in applications that require binary methylation status calls. In the second part of this thesis, we consider the problem of DNA methylation status calling. A mixture of Binomial model is used to characterize bisulphite-sequencing data, and based on the model, we propose to use a classification based procedure, called the Methylation Status Calling (MSC) procedure, to make binary methylation status calls. The MSC procedure is optimal in terms of maximizing the overall correct allocation rate, and the FDR and FNDR of MSC can be estimated. In order to control FDR at any given level, we further develop a FDR-controlled MSC (FMSC) procedure, which combines a local false discovery rate ($Lfdr$) based adaptive procedure with the MSC procedure. Both simulation study and real data application are carried out to examine the performance of the proposed procedures. It is shown in our simulation study that the estimates of FDR and FNDR of the MSC procedure are appropriate. Simulation study also demonstrates that the FMSC procedure is valid in controlling FDR at a prespecified level and is more powerful than the individual Binomial testing procedure. In the real data application, the MSC procedure exhibits an estimated FDR of 0.1426 and an estimated FNDR of 0.0067. The overall correct allocation rate is more than 0.97. These results suggest the effectiveness of the proposed procedures.

1. VARIABLE SELECTION FOR HIGH-DIMENSIONAL SINGLE INDEX MODEL

In this chapter, we focus on the problem of variable selection for single index model. We start with a review of four crucial concepts in Section 1.1. Section 1.2 describes the proposed framework, BS-SIM. Section 1.3 explains the implemention aspects of BS-SIM. In Section 1.4, the theoretical properties of BS-SIM, including estimation consistency and selection consistency, are demonstrated. Section 1.5 and Section 1.6 display the performance of BS-SIM under intensive simulation studies and a real data example. Section 1.7 discusses the problem of variable selection for single index model under linear constraints. The regularity conditions and the proofs to the theoretical properties are given in Section 1.8.

1.1 Introduction

1.1.1 Single Index Model

The linear regression model is the most commonly used approach to model the relationship between a univariate scalar response Y and a p-dimensional predictor X. It assumes the impact of the predictor X on the response Y is modeled through

$$Y = X^T\beta + \varepsilon,$$

where ^{T} indicates the transpose of a matrix, β is a vector of length p, and ε denotes the random error term. The linear regression model is intuitive and easy to interpret. However, the assumption that the relationship between the set of the predictor and the response is linear is not always satisfied.

To make the model more flexible, the single index model (SIM) takes the following form

$$Y = f(X^T\theta_0) + \varepsilon, \tag{1.1}$$

where θ_0 is a vector of length p, ϵ is an independent random error term with mean 0 and finite variance, and f is an unknown smooth function. The one-dimensional projection $X^T\theta_0$ is referred to as the index, and thus entails the name, single index model.

The single index model is a semi-parametric model. It includes a parametric part, θ. Meanwhile, it has two nonparametric components, the unknown link function f and the unknown distribution of the error term. SIM suggests that the information in X about Y is completely contained in the projection $X^T\theta_0$, whereas the exact relationship between the projection and the response f is unknown but one-dimensional. By this specification, the single index model is an intuitive generalization to many parametric models, such as the linear regression model and the generalized linear model (GLM) [1]. On the other hand, it has the advantage of being able to avoid the curse of dimensionality frequently encountered by the nonparametric methodology [2]. Due to these advantages, the single index model has applications in a wide range of fields, such as economics [3]. A number of methods have been proposed to estimate the true index θ_0 in the literature. In what follows, we will briefly review several popular ones among them.

Ichimura's Estimator

The estimation of the single index model was first studied by Ichimura [3]. In his Ph.D. thesis, he proposed to replace f with a leave-one-out kernel estimator as follows.

$$\hat{f}_{-i}(X_i^T\theta) = \frac{\sum_{j\neq i} k\left((X_j - X_i)^T\theta/h\right) y_i}{\sum_{j\neq i} k\left((X_j - X_i)^T\theta/h\right)}.$$

Then he relied on a least-squares methodology to obtain an estimator of θ_0 given below,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \hat{f}_{-i}(X_i^T \theta) \right)^2 .$$

It is shown in [3] that under certain regularity conditions, the above estimator can achieve consistency and asymptotic normality. Nevertheless, this estimator has the disadvantage of being difficult to be computed in practice.

Average Derivative Estimation based Methods

The Average Derivative Estimation (ADE) method was first introduced by Härdle and Stoker [4]. It relies on an intrinsic property of the single index model that θ_0 is proportional to the gradient $\partial f / \partial X$, that is, $\delta \equiv E\left(\frac{df}{d(X_0^\theta)}\right)\theta_0 \equiv \gamma\theta_0$. Let $g(X)$ be the marginal density of X, and $z \equiv -\partial \ln g / \partial x = -g'/g$. Then by some calculations, we have $\delta = E[z(X)y]$. Härdle and Stoker [4] further proposed to use a kernel density estimator $\hat{g}(X)$ to estimate $g(X)$, and as a result, we have $\hat{z}(X) = -\hat{g}'(X)/\hat{g}(X)$. Subsequently, the ADE estimator of δ is defined as

$$\hat{\delta} = n^{-1} \sum_{i=1}^{n} \hat{z}(X_i) y_i.$$

Härdle and Stoker [4] also suggested using a trimming technique to stabilize the above estimator. It was shown in [4] that under mild conditions, the above estimator enjoys good statistical properties such as consistency and asymptotic normality. Several modified ADE methods have been proposed later, including the density-weighted ADE method [5], the structure adaptive approach [6] and the out-product of gradients method [7]. Horowitz and Härdle [8] also proposed a generalized ADE based estimator that can work for discrete covariates. A major drawback of the ADE-based estimators is that most of them use high dimensional kernels in estimation, and thus suffer from the curse of dimensionality. Consequently, they do not perform well in estimation even when the dimension p is moderate. Another drawback for this category of methods is that the conditions for them to be $\sqrt{n}$-consistent are quite restrictive.

Minimum Average Variance Estimation Method

The Minimum Average Variance Estimation (MAVE) method by Xia *et al.* [9] is originally proposed as a dimension-reduction method. When the dimension to be reduced to is set to 1, the MAVE method leads to an estimator for SIM. By [9], we have θ_0 is the solution of

$$\theta_0 = \underset{\theta}{\text{argmin}}[\text{E}(Y - E(Y|X^T\theta))^2] = \underset{\theta}{\text{argmin}}\, E(\sigma^2_\theta(X^T\theta)),$$

where

$$\sigma^2_\theta(X^T\theta) = E((Y - E(Y|X^T\theta))^2|X^T\theta).$$

Xia *et al.* [9] proposed to use a local linear expansion to estimate $\sigma^2_\theta(X^T\theta)$ as follows.

$$\widehat{\sigma}^2_\theta(\theta^T X) = \min_{a,b}\left(\sum_{i=1}^{n}[Y_i - (a + b(x_i - x)^T\theta)]^2\omega_{i0}\right),$$

where ω_{i0} denotes some weights that sum up to 1. Subsequently, $\widehat{E}\left(\sigma^2_\theta(X^T\theta)\right) \approx \frac{1}{n}\sum_{i=1}^{n}\widehat{\sigma}^2_\theta(X^T\theta)$. The MAVE estimator can be obtained below

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\text{argmin}}\,\widehat{E}\left(\sigma^2_\theta(X^T\theta)\right) \\ &= \underset{\theta,a_j,b_j}{\text{argmin}}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i - (a_j + b_j(x_i - x_j)^T\theta)]^2\omega_{ij}\right),\end{aligned}$$

where ω_{ij} are some weights that sum up to 1 for each j. As noted by [9], a natural choice for the weights is to use the p-dimensional kernel. As a result, the above estimator would also suffer from the curse of dimensionality. To overcome this, Xia *et al.* [9] also proposed the refined MAVE (rMAVE) method by replacing the high dimensional kernel with a lower dimensional projection kernel. However, the computational complexity of MAVE and rMAVE still grows rapidly with the sample size n, and they can become unstable when p increases. From the theoretical property perspective, it is shown that under certain conditions, the MAVE estimator for SIM enjoys good statistical properties such as $\sqrt{n}$-consistency and asymptotically normality.

Inverse Regression based Methods

This category of methods is originally intended for the purpose of sufficient dimension reduction. The Inverse Regression based method was first introduced by Li [10]. It relies on regressing the predictors $\mathbf{x}$ on the response y, which is different from the traditional methods that regress y on $\mathbf{x}$. The rationale behind this approach is that under certain conditions, the standardized inverse regression curve falls into the linear space defined by the standardized effective dimension reduction directions. Based upon this, Li [10] described an algorithm, called Sliced Inverse Regression (SIR), to estimate the effective dimension reduction directions. Li [10] further developed the asymptotic properties of SIR under assumptions on the distribution of the predictors $\mathbf{x}$. Due to the virtue of regressing $\mathbf{x}$ on the univariate response y, SIR is very efficient in terms of computation, and it becomes considerably popular since it was proposed. Besides SIR, other popular Inverse Regression based methods include the sliced average variance estimator (SAVE) [11] and directional regression [12].

Single-Index Prediction Estimator

Recently, Wang and Yang [13] proposed the Single-Index Prediction (SIP) estimator. In their work, cubic B-splines were used to obtain an estimator $\hat{f}_\theta$ for each fixed θ. Then the empirical risk function $\hat{R}(\theta)$ can be defined as

$$\hat{R}(\theta) = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{f}_\theta(X_i^T \theta)]^2.$$

Subsequently, the SIP estimator of θ_0 is defined as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \hat{R}(\theta).$$

They showed that under mild conditions, the SIP estimator achieves $\sqrt{n}$-consistency and asymptotic normality. The application of the cubic B-splines circumvents the drawbacks suffered by high dimensional kernels, and as expected, simulation studies showed that SIP is considerably faster than MAVE, especially in the high dimensional case.

1.1.2 Variable Selection Methods for SIM

In practice, when the dimensionality p is large, the set of predictors can contain a large number of irrelevant variables. For the high-dimensional scenario, it is usually computationally inefficient to estimate the single index model with the whole collection of predictors. Moreover, even if an estimator is obtained with all of the predictors, it is difficult to interpret the results. Thus, for interpretability and computational efficiency purpose, it is important to perform variable selection when fitting the high-dimensional single index model. Various traditional variable selection methods have been extended to the single index model; for example, AIC [14] and cross-validation [15]. However, these methods suffer from the same drawbacks as the ones encountered in the linear regression model. They are intensive in terms of computation, and sometimes unstable. Furthermore, it is infeasible to develop the large sample properties for the resulting estimators.

Tibshirani [16] introduced the least absolute shrinkage and selection operator (LASSO) as a regularization method for simultaneous parameter estimation and variable selection in the linear models. LASSO has gained huge popularity since it was proposed, due to its succinctness and computational efficiency. Zhao and Yu [17] studied the sufficient and almost necessary condition, namely the Irrepresentable Condition, under which LASSO can consistently select the true model. There are various extensions or variants of LASSO proposed in the literature; see SCAD [18], adaptive LASSO [19], and the Dantzig selector [20] among others. Several attempts have been made to incorporate LASSO or its variants into the single index model, and we will briefly review some of them below.

Sparse MAVE

Recall the empirical risk function of MAVE for SIM is given by

$$R(\theta) = \sum_{j=1}^{n} \sum_{i=1}^{n} [Y_i - (a_j + b_j(x_i - x_j)^T\theta)]^2 \omega_{ij}.$$

Wang and Yin [21] proposed sparse MAVE or sMAVE. The idea is to add the LASSO penalty on θ to the above risk function, and achieve automatic variable selection.

SIM-LASSO

Zeng *et al.* [22] introduced the SIM-LASSO method which adds a L_1 penalty on $b_j\theta$ to the above MAVE objective function $R(\theta)$. The reason for including b_j in the penalty is that when $f(X^T\theta_0)$ is relatively flat at $x_j^T\theta_0$, the corresponding derivative b_j is close to zero and is not informative about θ_0. Therefore, not only can SIM-LASSO shrink some components of θ to zero, it also is able to shrink some b_j's to zero and exclude the data points that do not contain much information about θ_0 in the estimation procedure. Another nice property of SIM-LASSO is that its target function is invariant when b and θ are scaled by a constant and its reciprocal. This property makes developing an implementation algorithm much more convenient.

SIM-Bridge

Wang *et al.* [23] proposed SIM-Bridge which combines the bridge penalty [24] on θ with the above MAVE objective function $R(\theta)$. By using a concave penalty function, simulation studies suggest that SIM-Bridge is better at controlling the number of false positives than the two preceding LASSO based methods. Nevertheless, all of the three methods mentioned so far combine some penalty function with MAVE, thus they inherit the drawbacks of MAVE. They are computationally inefficient for increasing sample size and become unstable when the dimensionality is high.

SIM-SCAD

Peng and Huang [25] proposed a nonconcave penalized least squares method for variable selection in the single index model, called SIM-SCAD. In SIM-SCAD, a local linear approximation strategy is used to obtain an estimate of f, denoted as $\hat{f}$, at a

current estimate of θ_0, $\hat{\theta}_0$. Subsequently, Peng and Huang [25] proposed to rely on the following quadratic optimization problem to achieve simultaneous estimation and selection for the single index model.

$$\min_{\theta} \sum_{i=1}^{n} [Y_i - \hat{f}(X_i^T \hat{\theta}_0) - \hat{f}'(X_i^T \hat{\theta}_0)(X_i^T \theta - X_i^T \hat{\theta}_0)]^2 + n \sum_{j=1}^{p} p_\lambda(|\theta_j|),$$

where $p_\lambda(|\theta|)$ represents the SCAD penalty [18]. Peng and Huang [25] further showed that under certain regularity conditions, the SIM-SCAD estimator possesses the oracle properties [18].

1.1.3 Review of the SICA penalty functions

As mentioned in Lv and Fan [26], Nikolova [27] first studied a family of L_1 transformed penalty functions, whose form is given by

$$\rho(t) = \frac{bt}{1+bt},$$

where $t \in [0, \infty)$ and $b > 0$. Lv and Fan [26] considered a modified version of the above penalty function, and studied the following family of penalty functions.

$$\rho_a(t) = \left(\frac{t}{a+t}\right) I(t \neq 0) + \left(\frac{a}{a+t}\right) t, \qquad t \in [0, \infty),$$

where I denotes the indicator function, and $t \in [0, \infty)$. It follows that

$$\rho_0(t) = \lim_{a \to 0+} \rho_a(t) = I(t \neq 0), \quad \text{and} \quad \rho_\infty(t) = \lim_{a \to \infty} \rho_a(t) = t.$$

As noted by [26], this family of penalty functions forms a smooth homotopy between the L_0 and L_1 penalities, and thus is referred to as smooth integration of counting and absolute deviation (SICA) penalty functions. By the above equations, it can be seen that SICA includes LASSO as a limiting case.

The SICA penalty functions consist of a family of concave functions, and a is a tuning parameter that controls the maximum concavity. Figure 1.1 shows the shape of the SICA penalty functions on [-2,2] for the following sequence of a, $a = (0, 0.1, 1, 2, 5, \infty)$.

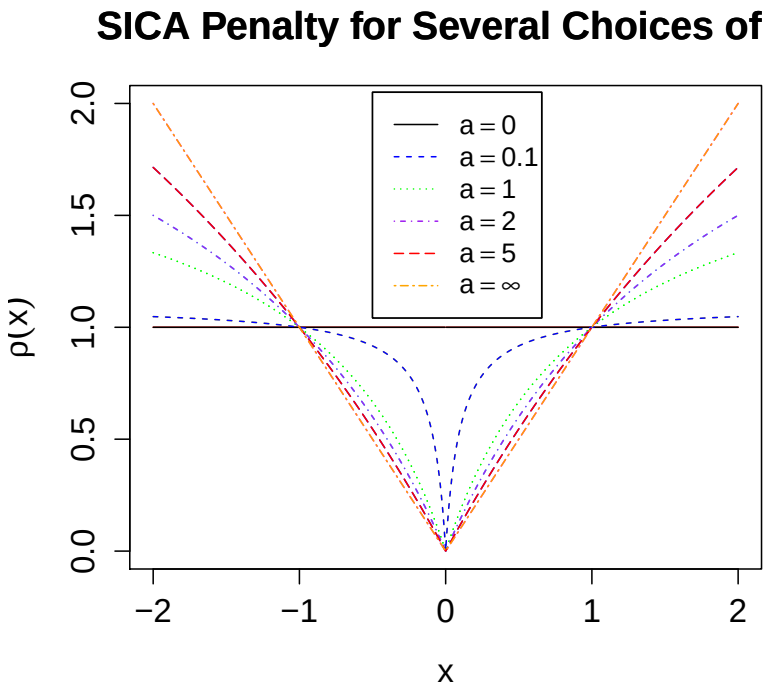


Fig. 1.1. Illustration of SICA for several choices of a.

Based on SICA, Lv and Fan [26] studied the problems of sparsity recovery and variable selection under the linear model. For the variable selection problem with SICA, they obtained the conditions on the design matrix under which the resulting SICA estimator can recover the true model. They showed that these conditions on the design matrix are more restrictive when a increases, and eventually converge to the Irrepresentable Condition developed by Zhao and Yu [17] for LASSO, as $a \to \infty$. This property suggests that under certain conditions, applying SICA with a finite a is more likely to select the true model than applying LASSO. And this may make the SICA penalty more appealing in cases where the Irrepresentable Condition does not hold and LASSO is not consistent in variable selection.

1.1.4 Review of B-Splines

A m-order spline function $f(x)$ is a piecewise polynomial function of order m. The places that these polynomial pieces meet are called knots, or knot sequence. The lowest order for a spline function is 1. A spline function of order 1 is a piecewise constant function, and a spline function of order 2 is a piecewise linear function, and so on. The smoothness of a spline function is largely decided by its order. A spline function of order m has up to $m-2$ order continuous derivatives. In practice, the most commonly used spline functions are order 4 splines, that is, cubic splines.

Spline functions of a given order and a given knot sequence can be represented as a linear combination of the spline basis functions. There are several equivalent forms of the basis functions, including the truncated power basis and the B-spline basis [28]. In what follows, we will introduce the B-spline basis functions. Without loss of generality, we assume the domain for x is $[0,1]$, and assume the sequence of the interior knots are $T = (t_1, t_2, \cdots, t_N)$ with $0 \leqslant t_1 \leqslant t_2 \leqslant \cdots \leqslant t_N \leqslant 1$, where N denotes the number of interior knots. Before we proceed to define the B-spline basis, we need first augment the knot sequence. The additional knots we need are outside of or on the boundary of the domain of x. Let the augmented knot sequence be $S = (s_1, s_2, \cdots, s_{N+2m})$, and they satisfy the following.

1. $s_1 \leqslant s_2 \leqslant \cdots \leqslant s_m \leqslant 0$;
2. $s_{i+m} = t_i$, for $i = 1, 2, \cdots, N$;
3. $1 \leqslant s_{m+N+1} \leqslant \cdots \leqslant s_{N+2m}$.

The additional knots are defined merely for computational convenience purpose. Their locations are arbitrary, as long as they satisfy the three conditions above.

Let $\mathbf{B}_q = (B_{q,1}, B_{q,2}, \ldots, B_{q,N+q})^T$ be the collection of B-spline basis functions for spline functions of order $q \leqslant m$ with knot sequence S. $\mathbf{B}$ can be defined recursively as follows [28].

$$B_{1,i}(x) = \begin{cases} 1, & x \in [s_i, s_{i+1}); \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, 2, \cdots, n + 2m - 1$. And

$$B_{q,i}(x) = \frac{x - s_i}{s_{i+m-1} - s_i} B_{q-1,i} + \frac{s_{i+1} - x}{s_{i+m} - s_{i+1}} B_{q-1,i+1},$$

for $i = 1, 2, \cdots, n + 2m - q$. Note that for $q < m$, not all of the augmented knots are needed in computing the basis functions. By the above construction, the B-spline basis functions of any order can be computed for a given knot sequence [29]. Figure 1.2 illustrates the B-spline basis functions of order 4 with 9 equally-spaced interior knots on [0,1].

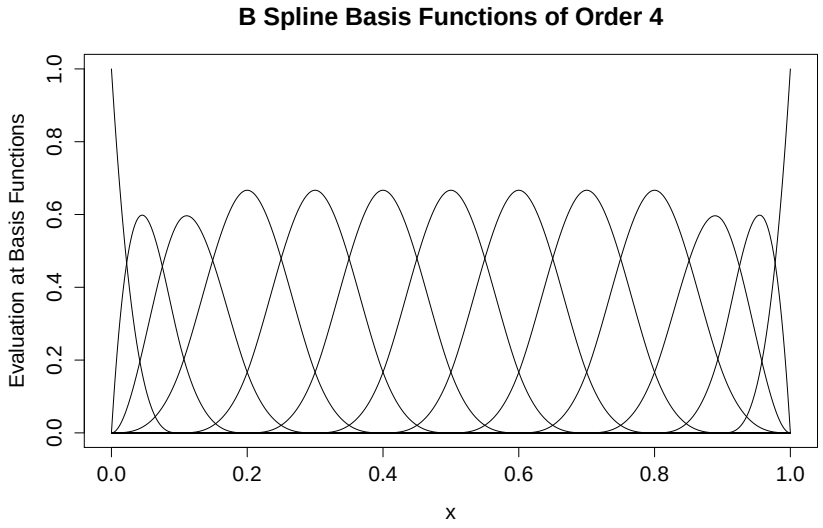


Fig. 1.2. An illustration of the B-spline basis functions of order 4 with 9 equally-spaced interior knots on [0,1].

As a conclusion to this subsection, we will make several remarks on B-splines. First, each basis function of order q is nonzero in only up to q subintervals. Moreover,

at a given $x \in [0,1]$, only q basis functions are nonzero. Due to the orthogonality property, B-spline basis provides the most convenience and efficiency in computing, especially when N is large. Therefore, it is the most widely-used basis in practice. As for the placement of the knot sequence, the most convenient choices include locating the knots at the quantiles of x and spreading the knots with equal space between two adjacent knots. There exists more sophisticated choices, such as those proposed by [30] and [31]. One thing that needs to be careful about here is that when there exists replicates in the knot sequence, the B-spline functions defined by the resulting basis functions will have one less continuous derivative at the corresponding replicated knot.

1.2 BS-SIM: A Spline Estimation and Regularization Method for Single Index Model

Suppose a random sample of n observations is generated from the single index model

$$y_i = f(x_i^T \theta_0) + \varepsilon_i,$$

$i = 1, 2, \cdots, n$, where $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,p})^T$ is the true index, and ε_i's are i.i.d random variables with mean 0 and a common variance σ^2. Let $\mathbf{Y} = (y_1, \cdots, y_n)^T$ denote the $n \times 1$ reponse vector, and $\mathbf{X} = (x_1, x_2, \cdots, x_n)^T$ be the $n \times p$ matrix with x_i representing its i-th row. The true index θ_0 is only identifiable up to a scale constant without further constraint. In the literature, there are two popularly used identifiability constraints:

1. *Identifiability Constraint 1:* $\theta_{0,1} = 1$;

2. *Identifiability Constraint 2:* $\|\theta_0\|_2 = 1$ and $\theta_{0,1} > 0$.

In this thesis, we consider any general and feasible constraint on the scale of θ_0. For example, other than the two popular identifiability constraints, $\sum_{i=1}^{p} \theta_{0,i} = 1$ can also be used. Here we work with the nontrivial case that there is at least one non-zero component in θ_0. Thus, for any constraint, it is important to first identify one

component $\theta_{0,k}$ that is non-zero. This component $\theta_{0,k}$ can be assumed as known from prior knowledge, or identified by methods such as marginal correlation. Without loss of generality, we assume $k = 1$. Although a large number of general identifiability constraints can be used, in Section 1.5, we show with simulation studies that different constraints can have different impacts on the performance of the used method in various aspects.

Suppose one specifies the following identifiability constraint: $\mathcal{C}(\theta) = 1$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$, and $\mathcal{C}$ is an explicit function on the scale of θ. Then θ_1 can be expressed as a function of the remaining components, that is, $\theta_1 = \mathcal{C}_1(\theta_2, \theta_3, \ldots, \theta_p)$. Let $\phi = (\theta_2, \theta_3, \ldots, \theta_p)^T$ be the $(p-1)$-dimensional sub-vector of θ by excluding the first component, and let $t_\theta = X^T\theta$. Let ϕ_0 denote the last $(p-1)$ components of θ_0. Let Φ be the space for ϕ. With an appropriate identifiability constraint imposed, ϕ and θ have a one-to-one association. Then the goal of inference under the single index model is to estimate ϕ_0 (and thus θ_0) and the true link function f.

For a given θ, let $t_\theta^i = x_i^T\theta$ be the projected data onto the direction of θ, $i = 1, 2, \ldots, n$. Let $t_\theta(\min) = \min_i t_\theta^i$ and $t_\theta(\max) = \max_i t_\theta^i$. The interval $[t_\theta(\min), t_\theta(\max)]$ is partitioned into $(N+1)$ subintervals. Let T_N be the sequence of the N interior knots that separate the subintervals. Let $B_4 = (B_{4,1}, B_{4,2}, \ldots, B_{4,N+4})^T$ be the cubic B-spline basis functions on $[t_\theta(\min), t_\theta(\max)]$ with knots T_N. As mentioned in Section 1.1.4, the explicit form of B_4 can be derived recursively. Here we slightly abuse the notations in the sense that θ and T_N are omitted in the representation of the basis functions. The evaluations of the basis functions on the projected data points are denoted as $\mathbf{B}_\theta$. That is, $\mathbf{B}_\theta = (B_4(t_\theta^1), \ldots, B_4(t_\theta^n))^T$, where $B_4(t)$ denotes the evaluation of the cubic B-spline basis functions at t.

The cubic B-spline estimator of f is defined as $\hat{f}_\theta(\cdot) = \hat{\alpha}^T B_4(\cdot)$, where $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_{N+4})^T$, and can be obtained by solving the following least-squares problem

$$\min_{\alpha \in \mathbb{R}^{N+4}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \alpha^T B_4(t_\theta^i) \right)^2 .$$

It immediately follows that $\hat{\alpha} = (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \mathbf{Y}$. Note that $\hat{f}_\theta(\cdot)$ depends on θ. Wang and Yang [13] further proposed to use the following least-squares method to estimate θ_0

$$\hat{\theta}_{\text{un}} = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_\theta(t_\theta^i))^2,$$

where $\hat{\theta}_{\text{un}}$ denotes the unpenalized estimator of θ_0, and $\Theta = \{\theta : \|\theta\|_2^2 = 1, \theta_1 > 0\}$. As discussed previously, the dimension p can be high in practice, and the set of predictors can include a large number of irrelevant variables. Therefore, it is of interest to produce a sparse estimator of θ_0, and thus achieve automatic variable selection. This motivates us to utilize the spline estimator $\hat{f}_\theta(\cdot)$ for f described above, coupled with the regularized least squares method for estimating θ_0 to achieve efficient and simultaneous parameter estimation and variable selection.

Since $\theta_{0,1}$ is assumed to be non-zero, we penalize ϕ instead of θ. We further use the family of the SICA penalty functions. That leads us to the following objective function $R(\phi; \lambda)$.

$$R(\phi; \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_\theta(t_\theta^i) \right)^2 + \lambda \sum_{j=1}^{p-1} \rho_a(|\phi_j|),$$

where $\hat{f}_\theta$ is the cubic B-spline estimator of f for a given θ, λ is a tuning parameter, and $\rho_a(u)$ denotes the SICA penalty function with the following form

$$\rho_a(u) = \left(\frac{u}{a+u} \right) I(u \neq 0) + \left(\frac{a}{a+u} \right) u, \qquad u \in [0, \infty).$$

For simplicity, we do not include a in the notation of R, and write $R(\phi; \lambda)$ as $R(\phi)$ when there is no confusion. For a fixed λ, we define the following estimator of ϕ_0,

$$\hat{\phi} = \underset{\phi \in \Phi}{\operatorname{argmin}} R(\phi), \tag{1.2}$$

The corresponding estimator for θ_0 is denoted as $\hat{\theta}$, and is referred to as the BS-SIM estimator.

Recall that the SICA family of penalty functions provides a smooth homotopy between the L_0 and L_1 penalties, and we have

$$\rho_0(u) = \lim_{a \to 0+} \rho_a(u) = I(u \neq 0), \quad \text{and} \quad \rho_\infty(u) = \lim_{a \to \infty} \rho_a(u) = u.$$

That means, the LASSO penalty is the limiting case of the SICA penalty. In some applications, the LASSO penalty can also be of interest, and the estimator based on LASSO is defined separately below. We denote the objective function when $a = \infty$ as $R_L(\phi; \lambda)$. That is,

$$R_L(\phi; \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_\theta(t_\theta^i) \right)^2 + \lambda \|\phi\|_1,$$

where $\| \cdot \|_1$ denotes the L_1 norm. We write it as $R_L(\phi)$ when there is no confusion. For a fixed λ, we define the following estimator of ϕ_0,

$$\hat{\phi}^L = \underset{\phi \in \Phi}{\operatorname{argmin}} \, R_L(\phi), \tag{1.3}$$

and the corresponding estimator for θ_0 is denoted as $\hat{\theta}^L$. We refer to $\hat{\theta}^L$ as the BL-SIM estimator. It can be expected that the BS-SIM estimator can converge to the BL-SIM estimator as a approaches ∞.

1.3 Implementation for BS-SIM

1.3.1 Coordinate Descent Algorithm for BS-SIM

For ease of representation, we define $H(\phi) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_\theta(t_\theta^i))^2$. Then the objective function $R(\phi)$ can be expressed as $R(\phi) = H(\phi) + \lambda \sum_{j=1}^{p-1} \rho_a(|\phi_j|)$. Next, we develop a coordinate descent algorithm to find $\hat{\phi}$ (or $\hat{\phi}^L$) for any given λ on a dense grid.

Since $H(\phi)$ is a complicated function of ϕ, we further use a local quadratic approximation strategy to iteratively solve Problem (1.2). Let $H^{(1)}(\cdot) = \partial H(\cdot)/\partial \phi$ and $H^{(2)}(\cdot) = \frac{\partial^2 H(\phi)}{\partial \phi \partial \phi^T}(\cdot)$, which are the gradient and Hessian matrix of H, respectively. Then, given a current estimate $\hat{\phi}^{(0)}$, the quadratic approximation to $H(\phi)$ at $\phi^{(0)}$ is given as follows.

$$\begin{aligned} H(\phi) &\approx H(\phi^{(0)}) + (\phi - \phi^{(0)})^T H^{(1)}(\phi^{(0)}) + \frac{1}{2}(\phi - \phi^{(0)})^T H^{(2)}(\phi^{(0)})(\phi - \phi^{(0)}) \\ &= \frac{1}{2}\phi^T H^{(2)}(\phi^{(0)})\phi - \phi^T \left(H^{(2)}(\phi^{(0)})\phi^{(0)} - H^{(1)}(\phi^{(0)}) \right) + \text{constant}. \end{aligned} \tag{1.4}$$

In addition, we use a local approximation to the SICA penalty function suggested by [26] as follows.

$$\sum_{j=1}^{p-1} \rho_a(|\phi_j|) = \sum_{j=1}^{p-1} [\rho_a(|\phi_j^{(0)}|) + \rho_a'(|\phi_j^{(0)}|)(|\phi_j| - |\phi_j^{(0)}|)], \tag{1.5}$$

where $\phi^{(0)} = (\phi_1^{(0)}, \phi_2^{(0)}, \cdots, \phi_{p-1}^{(0)})^T$.

These two approximations entail that for a given $\phi^{(0)}$, Problem (1.2) can be approximated by

$$\min_{\phi \in \Phi} \frac{1}{2} \phi^T H^{(2)}(\phi^{(0)}) \phi - \phi^T \left(H^{(2)}(\phi^{(0)}) \phi^{(0)} - H^{(1)}(\phi^{(0)}) \right) + \lambda \sum_{j=1}^{p-1} w_j |\phi_j|, \tag{1.6}$$

where $w_j = \rho_a'(|\phi_j^{(0)}|)$ for $j = 1, 2, \cdots, p-1$. To solve Problem (1.6), we cyclically update each component of ϕ while holding the other components fixed. That means, for $j = 1, 2, \cdots, p-1$, we solve the following univariate problem

$$\min_{\phi_j} \frac{1}{2} h_{jj} \phi_j^2 + \left(\sum_{k=1, k \neq j}^{p-1} h_{jk} \phi_k - \beta_j \right) \phi_j + \lambda w_j |\phi_j| + \text{constant}, \tag{1.7}$$

where h_{kl} denotes the component in the kth row and the lth column of $H^{(2)}(\phi^{(0)})$, and β_j denotes the jth element of $H^{(2)}(\phi^{(0)})\phi^{(0)} - H^{(1)}(\phi^{(0)})$. Notice that Problem (1.7) is essentially a univariate LASSO problem, and the solution can be written down explicitly as

$$\phi_j = \text{sign}(a_j) \frac{(|a_j| - \lambda w_j)_+}{h_{jj}} = \begin{cases} (a_j - \lambda w_j)/h_{jj}, & \text{if } a_j > \lambda w_j; \\ (a_j + \lambda w_j)/h_{jj}, & \text{if } a_j < -\lambda w_j; \\ 0, & \text{otherwise.} \end{cases} \tag{1.8}$$

where $a_j = \beta_j - \sum_{k \neq j} h_{jk} \phi_k$. We repeatedly iterate through j and update the estimate of ϕ_0, until some convergence criterion is met.

When implementing Algorithm 1, there are two issues that require further attention. First, during the sth cycle of j, linear search method is applied [32]. We start with $\hat{\phi}^{(s)}$, and obtain a tentative update $\hat{\phi}^{(s)}$. Before setting $\hat{\phi}^{(s+1)}$ as the most current estimate of ϕ_0, we need to check that the objective function R is indeed decreasing.

If it is not, the step $\delta = \hat{\phi}^{(s+1)} - \hat{\phi}^{(s)}$ is repeatedly multiplied by 0.8, until the amount of movement along the direction δ that can result in a decrease in R is obtained. Here, 0.8 is chosen for the purpose of convenience, and may not be optimal. A more sophisticated choice can be further explored; see the previously mentioned reference on line search. The other issue faced during the implementation is that the optimization over ϕ should be carried out in the space Φ. However, the algorithm described above does not consider any constraint on the space over which the optimization is executed. For some identifiability constraints, such as the *Identifiability Constraint 1* mentioned earlier, Φ is actually $\mathbb{R}^{p-1}$; for other identifiability constraints, such as the *Identifiability Constraint 2* in the previous section, Φ is a constrained subspace of $\mathbb{R}^{p-1}$. In the former case, no adjustment is needed; in the latter case, there requires an additional step that ensures that the updated $\hat{\phi}$ is in the constrained space Φ. For instance, it needs to be checked that the updated $\hat{\phi}$ satisfies $\|\hat{\phi}\|_2 < 1$, for *Identifiability Constraint 2*. If it does not, the step δ needs to be shortened such that $\hat{\phi}$ falls within Φ. Algorithm 1 outlines the search for $\hat{\phi}$ at a given λ in more detail. Problem (1.3) can be solved in a similar fashion. The only difference is that for Problem (1.3), there is no need to use the local linear approximation to the penalty function. Therefore, the algorithm of searching for $\hat{\phi}^L$ is not separately displayed.

Algorithm 1 *Coordinate Descent Algorithm for BS-SIM*

For any λ,

1. Initialize ϕ to be $\hat{\phi}^{(0)}$ and let $s = 0$.

2. Given $\hat{\phi}^{(s)} = (\hat{\phi}_1^{(s)}, \hat{\phi}_2^{(s)}, \ldots, \hat{\phi}_{p-1}^{(s)})^T$, calculate the quadratic approximation (1.4) to $H(\phi)$ and the linear approximation (1.5) to $p_\lambda(\phi)$.

3. For $j = 1, 2, \ldots, p-1$, update $\hat{\phi}_j$ by the following formulars:

$$\phi_j = \text{sign}(a_j)\frac{(|a_j| - \lambda w_j)_+}{h_{jj}} = \begin{cases} (a_j - \lambda w_j)/h_{jj}, & \text{if } a_j > \lambda w_j; \\ (a_j + \lambda w_j)/h_{jj}, & \text{if } a_j < -\lambda w_j; \\ 0, & \text{otherwise.} \end{cases}$$

   If needed, check whether ϕ is within Φ. If it is not, adjust it to fall within Φ.

4. After one cycle of j, a tentative update $\hat{\phi}^{(s+1)}$ and the corresponding $R(\hat{\phi}^{(s+1)})$ are obtained. If $R(\hat{\phi}^{(s+1)}) > R(\hat{\phi}^{(s)})$, calculate $\delta = \hat{\phi}^{(s+1)} - \hat{\phi}^{(s)}$, and check the objective function for

$$\hat{\phi}^{(s+1)} = \hat{\phi}^{(s)} + (0.8)^k \delta,$$

   for $k = 1, 2, \ldots$ until $R(\hat{\phi}^{(s+1)})$ is smaller than $R(\hat{\phi}^{(s)})$.

5. Calculate $\Delta = R(\hat{\phi}^{(s)}) - R(\hat{\phi}^{(s+1)})$. If Δ is below a prespecified threshold, then stop and set $\hat{\phi} = \hat{\phi}^{(s+1)}$ and calculate the corresponding $\hat{\theta}$; otherwise, set $s = s + 1$ and go back to Step 2.

1.3.2 Tuning Parameter Selection for BS-SIM

For regularization-based approaches, it is crucial to choose the tuning parameters, namely λ and a in our case. We start with the discussion of the selection of λ. We consider two types of methods for determining λ. The first one is m-fold cross-validation, denoted as CV hereafter. In CV, the sample is randomly partitioned into m subsamples of equal size. Among these m folds, $m - 1$ of them are treated as the

training set, and the remaining one is treated as the validation set. At each given candidate value for λ, the proposed approach is applied to the training set, and a fitting is obtained. Subsequently, the test set is used to assess the predictive accuracy of the obtained model. The residual sum of squares can be used as the assessment. This process is repeated m times until each fold of the sample is used as the test set exactly once. For a given λ, the m results on the assessment are then averaged. The value of λ that yields the smallest average is regarded as optimal.

The second type is the Bayesian Information Criterion (BIC) and its variants [33]. For variable selection under the linear model $Y = X^T\beta + \epsilon$, we examine the following four BIC-based criteria (1.9)-(1.12).

$$\text{BIC} = \text{RSS}_\lambda/n + d\sigma^2\log(n)/n, \tag{1.9}$$

$$\text{logBIC} = \log(\text{RSS}_\lambda/n) + d\log(n)/n, \tag{1.10}$$

$$\text{GIC} = \text{RSS}_\lambda/n + d\sigma^2 k_n/n, \tag{1.11}$$

$$\text{logGIC} = \log(\text{RSS}_\lambda/n) + dk_n/n, \tag{1.12}$$

where RSS_λ denotes the residual sum of squares at a given λ, σ^2 denotes the error variance, and d is the size of the identified model at a given λ. Furthermore, for criteria (1.11) and (1.12), k_n represents the additional penalty imposed on the size of the model. In practice, σ^2 is rarely known. On the other hand, according to Shao [34], under certain conditions, the BIC defined in (1.9) has the same asymptotic behavior as the one defined in (1.10). Thus, it is more convenient to rely on logBIC in (1.10) to select the tuning parameter λ. It has been previously proved that, when the number of predictors p is fixed as the number of observations n grows, one can identify the true model with probability tending to 1 in the linear models by using the logBIC criteria [35]. Nevertheless, when p diverges, the logBIC criterion (1.10) tends to yield a model that contains many irrelevant predictors. Several adjustments have been proposed in the literature to circumvent this issue [35–37]. The common approach these adjustments take is to place more penalty on the model complexity d. This idea naturally leads us to consider the GIC criterion in (1.11) and the logGIC criterion in

(1.12). It is clear that GIC and logGIC include BIC and logBIC as a special case, respectively. Thus, GIC and logGIC can be regarded as the unified criteria to achieve the selection of λ for any p, and they can be extended to models other than the linear regression models. It is also worth noting that GIC involves σ^2. When σ^2 is unknown, there are various ways to obtain an estimate $\hat{\sigma}^2$ and replace σ^2 with $\hat{\sigma}^2$ in GIC. We will elaborate on it in the next paragraph.

In order to choose a proper type of method for determining λ under our framework, we carry out extensive simulation studies under both the linear model and the single index model. We try different settings of p and the size of the true model. In the simulation studies, we use $\hat{\sigma}^2 = \mathrm{RSS}_0/(n-p)$ when $n > p$, and $\hat{\sigma}^2 = \mathrm{RSS}_{\lambda_{cv}}/(n-d_{cv})$ otherwise, where λ_{cv} denotes the value of λ selected by CV, and d_{cv} denotes the size of the model selected by CV. For all settings, CV generally leads to an overfitted model. When the true model is sparse, logGIC with an appropriate k_n performs the best in terms of identifying the true model for any p. GIC is a close second. As the number of relevant variables grows, the performance of GIC surpasses that of logGIC, and GIC becomes the most preferable. For the moderately sparse scenario, logGIC starts to break down as p increases. When the size of the true model is large, logGIC fails to work in the sense that it leads to either a very large model, or a very small model. Meanwhile, GIC can still produce significant improvement over CV when p is not large. When p also becomes large, the problem itself becomes too difficult that all of the methods rarely perform satisfactorily.

Based upon these observations, we propose the following rule of thumb principle for the selection of λ under our framework. When sparsity of the true model is assumed, we use logGIC; when the size of the true model is relatively large, we use GIC. An example illustrating the breakdown of logGIC and the advantage of using GIC under the violation of the sparsity assumption is given in Section 1.5.7.

As for the selection of a, it can generally be accomplished by m-fold cross-validation. Since the focus of this work is to study the properties of $\hat{\theta}$ and $\hat{\theta}^L$, we do not intensively examine the selection of a.

1.4 Theoretical Properties for BS-SIM

1.4.1 Estimation Consistency

To begin with, we show that, under mild conditions, $\hat{\theta}$ is consistent in terms of estimation, and can achieve the optimal $\sqrt{n}$ rate for a well-selected λ. Moreover, as a special case, $\hat{\theta}^L$ share the same property on parameter estimation.

Theorem 1.4.1 *Suppose Conditions (A1)-(A3) in Section 1.8 hold.*
(a) If $\lambda = O(n^{-1/2})$, there exists a local minimum $\hat{\phi}$ of $R(\phi)$, such that $\hat{\phi}$ is $\sqrt{n}$-consistent. Consequenly, the BS-SIM estimator $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ_0;
(b) If $\lambda = O(n^{-1/2+\delta})$ for some $\delta \in (0, 1/2)$, there exists a local minimum $\hat{\phi}$ of $R(\phi)$, such that $\|\hat{\phi} - \phi_0\|_2 = O_p(n^{-1/2+\delta})$. As a result, $\|\hat{\theta} - \theta_0\|_2 = O_p(n^{-1/2+\delta})$;
(c) As a special case, the BL-SIM estimator $\hat{\theta}^L$ possesses the above properties.

Theorem 1.4.1 is expected and standard. Part (b) of Theorem 1.4.1 also facilitates the derivations on the selection consistency given below.

1.4.2 Intuition and Notations for Selection Consistency

Observe that if no identifiability constraint is imposed, we have $f(t_\theta) - f(t_{\theta_0}) \approx D'_\theta(t_{\theta_0})(\theta - \theta_0)$, where $D'_\theta(t_{\theta_0}) = \left(\frac{\partial f(t_{\theta_0})}{\partial \theta_1}, \frac{\partial f(t_{\theta_0})}{\partial \theta_2}, \ldots, \frac{\partial f(t_{\theta_0})}{\partial \theta_p}\right)$. By simple calculations, we obtain

$$\frac{\partial f(t^i_{\theta_0})}{\partial \theta_j} = h(t^i_{\theta_0}) X_{ij} \triangleq g_{ij},$$

where $h(t^i_{\theta_0}) = f'|_{t=t^i_{\theta_0}}$ for $j = 1, 2, \ldots, p$, and $i = 1, 2, \ldots, n$. Let

$$F = \begin{pmatrix} \frac{\partial f(t^1_{\theta_0})}{\partial \theta_1}, & \frac{\partial f(t^1_{\theta_0})}{\partial \theta_2}, & \cdots, & \frac{\partial f(t^1_{\theta_0})}{\partial \theta_p} \\ \frac{\partial f(t^2_{\theta_0})}{\partial \theta_1}, & \frac{\partial f(t^2_{\theta_0})}{\partial \theta_2}, & \cdots, & \frac{\partial f(t^2_{\theta_0})}{\partial \theta_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(t^n_{\theta_0})}{\partial \theta_1}, & \frac{\partial f(t^n_{\theta_0})}{\partial \theta_2}, & \cdots, & \frac{\partial f(t^n_{\theta_0})}{\partial \theta_p} \end{pmatrix}_{n \times p} = (g_{ij})_{i=1,2,\ldots,n; j=1,2,\ldots,p}.$$

By the definition of g_{ij}, it is apparent that F is a weighted design matrix. That is, F is computed by multiplying row i of X with the corresponding derivative of f at

$t^i_{\theta_0}$, $h(t^i_{\theta_0})$, for $i = 1, 2, \cdots, n$. When f is flat at $t^i_{\theta_0}$, this data point does not contain much information on θ_0, and the weight placed on row i is small; on the other hand, when f is steep at $t^i_{\theta_0}$, this data point is informative, and the corresponding row is scaled with a larger weight. In the special case of the linear models, F reduces to X.

However, θ_0 is not free of identifiability constraint, and only the last $p-1$ elements of θ_0 are of interest. Consequently, we consider

$$F_0 = \begin{pmatrix} \frac{\partial f(t^1_{\theta_0})}{\partial \theta_2}, & \frac{\partial f(t^1_{\theta_0})}{\partial \theta_3}, & \cdots, & \frac{\partial f(t^1_{\theta_0})}{\partial \theta_p} \\ \frac{\partial f(t^2_{\theta_0})}{\partial \theta_2}, & \frac{\partial f(t^2_{\theta_0})}{\partial \theta_3}, & \cdots, & \frac{\partial f(t^2_{\theta_0})}{\partial \theta_p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(t^n_{\theta_0})}{\partial \theta_2}, & \frac{\partial f(t^n_{\theta_0})}{\partial \theta_3}, & \cdots, & \frac{\partial f(t^n_{\theta_0})}{\partial \theta_p} \end{pmatrix}_{n\times(p-1)}.$$

Here, F_0 depends on the design X, the true link function f, and the true index θ_0. To some extent, F_0 can be treated as the design matrix in the single index models, and it can play a crucial role in the subsequent analysis. For a given identifiability constraint, we can express θ_1 as a function of the rest $(p-1)$ components of θ, that is $\theta_1 = \mathcal{C}_1(\theta_2, \ldots, \theta_p)$. Let J be the corresponding Jacobian matrix for θ_0, that is,

$$J = \begin{pmatrix} \frac{\partial \mathcal{C}_1(\phi_0)}{\partial \theta_2}, & \frac{\partial \mathcal{C}_1(\phi_0)}{\partial \theta_3}, & \cdots & \frac{\partial \mathcal{C}_1(\phi_0)}{\partial \theta_p} \\ 1, & 0, & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0, & 0, & \cdots & 1 \end{pmatrix}_{p\times(p-1)}.$$

And it follows that $F_0 = FJ$. For simplicity, here we omit the dependence of J on the identifiability constraint in the notation. The forms of F_0 for the two popular identifiability constraints are illustrated below. Notice that F_0 is essentially a scaled and adjusted version of the design matrix X.

Identifiability Constraint 1: $\theta_{0,1} = 1$.

In this case, $\mathcal{C}_1(\theta_2, \ldots, \theta_p) \equiv 1$. Thus,

$$J = \begin{pmatrix} 0, & 0, & \cdots & 0 \\ 1, & 0, & \cdots & 0 \\ 0, & 1, & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0, & 0, & \cdots & 1 \end{pmatrix},$$

and

$$F_0 = \begin{pmatrix} g_{12}, & g_{13}, & \cdots & g_{1p} \\ g_{22}, & g_{23}, & \cdots & g_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ g_{n2}, & g_{n3}, & \cdots & g_{np} \end{pmatrix},$$

which is actually a sub-matrix of F.

Identifiability Constraint 2: $\|\theta_0\|_2 = 1$ and $\theta_{0,1} > 0$.

This yields that $\mathcal{C}_1(\theta_2, \ldots, \theta_p) = \sqrt{1 - \theta_2^2 - \cdots - \theta_p^2}$. Thus,

$$J = \begin{pmatrix} -\frac{\theta_{0,2}}{\theta_{0,1}}, & -\frac{\theta_{0,3}}{\theta_{0,1}}, & \cdots & -\frac{\theta_{0,p}}{\theta_{0,1}} \\ 1, & 0, & \cdots & 0 \\ 0, & 1, & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0, & 0, & \cdots & 1 \end{pmatrix},$$

and

$$F_0 = \begin{pmatrix} g_{12} - \frac{\theta_{0,2}}{\theta_{0,1}} g_{11}, & g_{13} - \frac{\theta_{0,3}}{\theta_{0,1}} g_{11}, & \cdots & g_{1p} - \frac{\theta_{0,p}}{\theta_{0,1}} g_{11} \\ g_{22} - \frac{\theta_{0,2}}{\theta_{0,1}} g_{21}, & g_{23} - \frac{\theta_{0,3}}{\theta_{0,1}} g_{21}, & \cdots & g_{2p} - \frac{\theta_{0,p}}{\theta_{0,1}} g_{21} \\ \cdots & \cdots & \cdots & \cdots \\ g_{n2} - \frac{\theta_{0,2}}{\theta_{0,1}} g_{n1}, & g_{n3} - \frac{\theta_{0,3}}{\theta_{0,1}} g_{n1}, & \cdots & g_{np} - \frac{\theta_{0,p}}{\theta_{0,1}} g_{n1} \end{pmatrix}.$$

Without the loss of generality, let $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,q}, \theta_{0,q+1}, \ldots, \theta_{0,p})^T$ where $\theta_{0,j} \neq 0$ for $j = 1, 2, \ldots, q$ and $\theta_{0,j} = 0$ for $j = q+1, q+2, \ldots, p$. Let $\mathcal{A}_1 =$

$\{2, 3, \ldots, q\}$ and $\mathcal{A}_2 = \{q+1, q+2, \ldots, p\}$. For any ϕ, we also decompose it into two sub-vectors as follows $\phi(1) = (\theta_2, \theta_3, \ldots, \theta_q)^T$, and $\phi(2) = (\theta_{q+1}, \ldots, \theta_p)^T$. Let $C_0 = \frac{1}{n} F_0^T F_0$. Let $F_0(1)$ and $F_0(2)$ be the first $q-1$ and the last $p-q$ columns of F_0. Let $C_0(11) = \frac{1}{n} F_0^T(1) F_0(1)$, $C_0(21) = \frac{1}{n} F_0^T(2) F_0(1)$, $C_0(12) = \frac{1}{n} F_0^T(1) F_0(2)$ and $C_0(22) = \frac{1}{n} F_0^T(2) F_0(2)$. Then we can decompose C_0 into the following four blocks

$$C_0 = \begin{pmatrix} C_0(11) & C_0(12) \\ C_0(21) & C_0(22) \end{pmatrix}.$$

In the following subsections, we also rely on this decomposition to formulate the results on the selection consistency of the proposed estimators.

1.4.3 Selection Consistency

As detailed earlier, we use the cubic spline function to estimate the true link function f. For any θ, let $\Gamma(\theta)$ be the cubic spline space defined according to Section 1.2. We denote the projection matrix onto $\Gamma(\theta)$ as $\mathbf{P}_\theta = \mathbf{B}_\theta (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T$. Thus,

$$\hat{f}_\theta = \left(\hat{f}_\theta(t_{\theta,1}), \ldots, \hat{f}_\theta(t_{\theta,n}) \right)^T = \mathbf{P}_\theta \mathbf{Y}.$$

Consequently, we have

$$\mathbb{E}\left(\hat{f}_\theta(t_\theta^i) \right) = \mathbf{P}_\theta f(t_{\theta_0}^i) \triangleq \bar{f}_\theta(t_\theta^i),$$

for $i = 1, 2, \cdots, n$. Then, for any given θ, we can similarly define $\bar{F}_\theta$ and $\bar{C}_\theta$ as

$$\bar{F}_\theta = \left(\frac{\partial \bar{f}_\theta(t_\theta^i)}{\partial \theta_j} \right)_{i=1,2,\ldots,n; j=2,3,\ldots,p},$$

and $\bar{C}_\theta = \frac{1}{n} \bar{F}_\theta^T \bar{F}_\theta$. For succinctness, we write $\bar{F}_{\theta_0}$ and $\bar{C}_{\theta_0}$ as $\bar{F}_0$ and $\bar{C}_0$. Different from F_0, $\bar{F}_0$ not only depends on X, f and θ_0, it also relies on the spline approximation of the link function. We decompose $\bar{C}_0$ into four blocks in the same way we decompose C_0. With the notations introduced above, we can impose the following crucial conditions on $\bar{C}_0$ to establish the selection consistency of BS-SIM.

Condition 1 (Irrepresentable Conditions for BS-SIM) *$\bar{C}_0$ satisfies that*

$$\begin{aligned} \|\bar{C}_0^{-1}(11)\|_\infty &\leqslant \bar{L}_1, \\ \|\bar{C}_0(21)\bar{C}_0^{-1}(11)\|_\infty &\leqslant \bar{L}_2, \end{aligned}$$

where $\bar{L}_1 \in (0,\infty)$, $\bar{L}_2 \in \left(0, \bar{L}\frac{\rho'(0+)}{\rho'(b_0-\lambda\bar{L}_3)}\right)$ for some $\bar{L}$ and $\bar{L}_3 \in (0,\infty)$, and $b_0 = \min_{j\in\mathcal{A}_1} |\theta_{0,j}|$.

Note that $\bar{C}_0$ is related to the spline estimator of f, and thus it depends on the number and the location of the knots. That means the conditions given above are not free of the sample size n. On the other hand, $\bar{F}_0$ is a scaled and adjusted version of the design matrix X. Hence, the Irrepresentable Conditions for BS-SIM are similar to the conditions by [26] in the sense that the above conditions replace the design matrix X in [26] with $\bar{F}_0$. With the Irrepresentable Conditions for BS-SIM, we are ready to state our theorem next.

Theorem 1.4.2 *Assume the Irrepresentable Conditions for BS-SIM hold, and the regularity conditions (A1)-(A3) in Section 1.8 are satisfied. Then for $\lambda = O(n^{c-2/5})$, with some $c \in (0, 2/5)$, there exists a local minimum $\hat{\phi}$ of $R(\phi)$ such that*

$$P\left(sign(\hat{\phi}) = sign(\phi_0)\right) = 1 - o(e^{-n^c}), \text{ as } n \to \infty,$$

where sign(s) is the sign function that equals 1 when s is positive, equals -1 when s is negative, and equals 0 when $s = 0$.

Theorem 1.4.2 characterizes the behaviour of BS-SIM in recovering the true model. It suggests that, if the Irrepresentable Conditions for BS-SIM hold, then the probability that BS-SIM is able to identify the true model converges to 1 exponentially. It can be easily shown that $\rho'(0+) = 1 + a^{-1}$. As noted by [26], the conditions for SICA to identify the true model in the linear regression becomes less restrictive as a decreases, at the sacrifice of computational convenience. This statement also holds in the context of the single index model. That means, with smaller a, the Irrepresentable

Conditions for BS-SIM are less restrictive, but it is harder to find $\hat{\phi}$. As pointed out earlier, LASSO is a limiting case of the SICA penalty. Therefore, it is expected that the BL-SIM estimator $\hat{\phi}^L$ would possess the similar properties as given in Theorem 1.4.2. To present the properties for $\hat{\phi}^L$, we start with the following assumption on $\bar{C}_0$.

Condition 2 (Irrepresentable Condition for BL-SIM) *There exists a positive constant vector $\bar{\eta}$, such that the following inequality holds component-wise*

$$\left|\bar{C}_0(21)\bar{C}_0^{-1}(11)sign(\phi_0(1))\right| \leqslant \mathbb{1}_{p-q} - \bar{\eta},$$

where $\mathbb{1}_{p-q}$ denotes a vector of 1's of length $p-q$.

Again, the Irrepresentable Condition for BL-SIM resembles the Irrepresentable Condition in [17], and the major difference is that the Irrepresentable Condition for BL-SIM replaces X with $\bar{F}_0$.

Theorem 1.4.3 *Assume the Irrepresentable Condition for BL-SIM holds, and the regularity conditions (A1)-(A3) in Section 1.8 are satisfied. Then for $\lambda = O(n^{c-2/5})$, with some $c \in (0, 2/5)$, there exists a local minimum $\hat{\phi}^L$ of $R_L(\phi)$ such that*

$$P\left(sign(\hat{\phi}^L) = sign(\phi_0)\right) = 1 - o(e^{-n^c}).$$

Theorem 1.4.3 demonstrates that with the Irrepresentable Condition for BL-SIM imposed, the probability that BL-SIM selects the true model approaches 1 exponentially. Consistent with the monotonicity of the restrictiveness of the conditions, the Irrepresentable Condition for BL-SIM is more restrictive than the Irrepresentable Conditions for BS-SIM with finite a. This observation is also in line with that in the linear regression scenario, and it implies that BS-SIM may be able to recover the true model when BL-SIM fails.

Recall that the conditions presented previously rely on the sample size n. In what follows, we show that if $\bar{C}_0$ satisfies certain regularity condition, the selection consis-

tency of the proposed methods can be achieved under conditions that are independent of n. From [13], we have

$$\sup_{j=2,3,\ldots,p}\sup_{\theta:\|\theta\|_2=1}\max_i\left|\frac{\partial}{\partial\theta_j}(\bar{f}_\theta - f)(t_\theta^i)\right| = O\left(h^3\right),$$

where $h = 1/(N+1)$ is the bandwith for the cubic B-spline functions. This means that $(\bar{F}_0)_i \to (F_0)_i$, as $n \to \infty$, for any i, and $(\cdot)_i$ denotes the ith row of a matrix. Based on this result, the following regularity condition can be imposed,

$$\bar{C}_0 \to C, \text{ as } n \to \infty,$$

for some matrix C free of n. We decompose C into four blocks in the same way we decompose C_0. Next, we show that if the Irrepresentable Conditions on C are imposed, the proposed methods can consistently select the true variables.

Condition 3 (Limiting Irrepresentable Conditions for BS-SIM) *C satisfies that*

$$\begin{aligned}\|C^{-1}(11)\|_\infty &\leqslant L_1,\\ \|C(21)C^{-1}(11)\|_\infty &\leqslant L_2,\end{aligned}$$

where $L_1 \in (0,\infty)$, and $L_2 \in \left(0, L\frac{\rho'(0+)}{\rho'(b_0-\lambda L_3)}\right)$ for some L and $L_3 \in (0,\infty)$.

Condition 4 (Limiting Irrepresentable Condition for BL-SIM) *There exists a positive constant vector η, such that the following inequality holds component-wise*

$$\left|C(21)C^{-1}(11)sign(\phi_0(1))\right| \leqslant \mathbb{1}_{p-q} - \eta,$$

where $\mathbb{1}_{p-q}$ denotes a vector of 1's of length $p-q$.

Corollary 1.4.4 *(a) Assume that λ satisfies that $\lambda \sim n^{c-2/5}$, for some $c \in (0, 2/5)$, and the Limiting Irrepresentable Conditions for BS-SIM hold. Under regularity conditions (A1)-(A3) in Section 1.8, there exists a local minimum $\hat{\phi}$ of $R(\phi)$ such that*

$$P\left(sign(\hat{\phi}) = sign(\phi_0)\right) = 1 - o(e^{-n^c}).$$

(b) Assume that λ satisfies that $\lambda \sim n^{c-2/5}$, for some $c \in (0, 2/5)$, and the Limiting Irrepresentable Condition for BL-SIM holds. Under regularity conditions (A1)-(A3) in Section 1.8, there exists a local minimum $\hat{\phi}^L$ of $R_L(\phi)$ such that

$$P\left(sign(\hat{\phi}^L) = sign(\phi_0)\right) = 1 - o(e^{-n^c}).$$

Corollary 1.4.4 suggests that under the corresponding Limiting Irrepresentable Conditions, BS-SIM and BL-SIM can consistently recover the true model. On the other hand, same as the statements given in the last subsection, the Limiting Irrepresentable Conditions for BS-SIM become less restrictive as a decreases. As a result, the Limiting Irrepresentable Condition for BL-SIM is more restrictive than those for BS-SIM with finite a. The proofs of the theorems and the corollaries can be found in Section 1.8.

1.5 Simulation Studies

In this section, we present the results from seven simulation studies. We demonstrate that the proposed regularization approach used is indeed beneficial in several aspects. We also look at the impact of the tuning parameter a on the performance of the resulting estimator, and point out a reasonable choice of a in practice. Subsequently, we compare the performance of the proposed methods to other existing methods for small to large p. The last simulation example is concerned about the impact that the Irrepresentable Condition has on our proposed method's ability of recovering the true model. For the purpose of succinctness, we use V1 and V2 to denote the *Identifiability Constraint 1* and *Identifiability Constraint 2* in this section, respectively. For the link function, we consider the following three models:

1. $Y = X^T\theta_0 + 4\sqrt{|X^T\theta_0 + 1|} + \varepsilon$;
2. $Y = 1 + 2(X^T\theta_0 + 3)\log(3|X^T\theta_0| + 1) + \varepsilon$;
3. $Y = (X^T\theta_0)^2 + \varepsilon$.

The models above are refered to as Model 1, Model 2, and Model 3, respectively. Furthermore, let Σ be a p-by-p matrix with the diagonal elements equal 1 and the off-diagonal element in kth row and lth column equal ρ_{kl}. Each x_i is sampled from $N(\mathbf{0}, \Sigma)$. The errors ε_i's are independently sampled from $N(0, 1)$. We examine the following three forms of Σ:

1. (No correlation) $\rho_{kl} = 0$, for $k \neq l$;
2. (Constant correlation) $\rho_{kl} = 0.3$, for $k \neq l$;
3. (Decaying correlation) $\rho_{kl} = 0.5^{|k-l|}$, for $k \neq l$.

We denote these three types of correlation structure as COR1, COR2, COR3, respectively.

For the first four examples, four metrics are used to assess the performance of an estimator, which are Angle, False Positive Rate (FPR), Ture Positive Rate (TPR) and Computing Time (Time), respectively. Angle is defined as Angle $= \arccos(\theta_0^T \hat{\theta})$, where θ_0 is the true index and $\hat{\theta}$ is an estimate, and they are standardized to have unit norm. FPR is defined as the ratio of the number of falsely identified predictors to the total number of identified predictor. TPR is the ratio of the number of correctly identified predictors to the total number of true relevant predictors. Finally, Time is the average time (in seconds) needed to obtain the estimate for one data set. In Examples 2-4, we search the best estimate on a dense grid of λ, and thus, Time represents the total amount of time consumed to find the estimate on the whole grid and yield the final estimate. On the other hand, in Example 1, Time refers to the amount of time used to find the estimate for a particular λ. In the tables presented in this section, the best performance on each metric is highlighted.

1.5.1 Performance of the proposed method for small p

In this section, we will study the performance of the proposed estimators for a small dimension $p = 20$. The other settings are $q = 4$, $n = 100$, and $\theta_0 =$

$(2.0, -1.0, 0.5, 1.0, 0, \cdots, 0)^T$. The first model setting we consider is Model 1 with COR1. Three choices for the number of interior knots N are used, which are N =2, 3, and 4. In this low dimensional case, we will rely on the logBIC criterion to choose λ.

The purpose of this example is two fold. First, it examines the performance of the proposed methods for different choices of N. On the other hand, it compares the performance of our methods to the existing methods in the low dimensional scenario. The comparison results on the four assessments are shown in Table 1.1. In terms of estimation accuracy, all of the methods applied perform well. The two methods that rely on a concave penalty, the proposed BS-SIM method and SIM-Bridge, outperform the rest in controlling FPR. In the computational efficiency aspect, the proposed methods are more efficient than the MAVE based methods in this example.
Given the results in Table 1.1, we will fix the number of knots at $N = 2$ for $n = 100$. We shall use one more example to compare the performance of the aforementioned methods for small p. The same values for p, q, n and θ_0 are used, and the model setting is changed to Model 3 with COR1. The comparison results are shown in Table 1.2. As explained in Example 3 in Section 4 of the main article, the MAVE based methods do not perform well for the quadratic link. The proposed BS-SIM and BL-SIM methods are more preferable in terms of parameter estimation and computational efficiency under this setting. In terms of selection consistency, the proposed BS-SIM method is also among the best.

1.5.2 Performance of the proposed method compared to that of the unpenalized estimator

This example compares the performance of the proposed estimator to that of the unpenalized estimator. We consider a moderate dimension $p = 70$ with $q = 8$ and $\theta_0 = (2.0, -1.0, 0.5, 1.0, -1.5, 1.0, -0.3, 1.2, 0, \cdots, 0)^T$. 100 samples of size $n = 100$ are generated from Model 1 with COR1. The coordinate descent algorithm described

Model 1, COR1, $p = 20$					
Method	N	Angle	FPR	TPR	Time
BS-SIM-V1	2	2.590 (1.269)	0.057	1	3.66
	3	2.904 (2.235)	0.076	1	3.98
	4	2.455 (1.277)	0.077	1	4.16
BS-SIM-V2	2	2.540 (1.214)	0.050	1	**3.30**
	3	2.697 (1.343)	0.072	1	3.58
	4	**2.406** (1.223)	0.065	1	3.98
BL-SIM-V1	2	4.714 (1.490)	0.262	1	6.54
	3	4.913 (2.505)	0.295	1	7.57
	4	4.658 (3.210)	0.340	1	8.11
BL-SIM-V2	2	3.945 (1.454)	0.192	1	6.48
	3	4.052 (1.431)	0.186	1	7.46
	4	3.677 (1.443)	0.215	1	8.33
SIM-LASSO-V2		3.891 (1.365)	0.446	1	18.25
SMAVE-V2		5.313 (2.210)	0.093	1	39.72
SIM-Bridge-V2		2.512 (1.334)	**0.026**	1	59.27

Table 1.1.
Comparison between our methods to the existing methods in low dimensional scenario: Model 1 with COR1.

in Section 1.3.1 is used to implement BS-SIM with $a = 0.1$. The tuning parameter λ is chosen by three criteria, denoted as logBIC, logGIC1, and logGIC2, respectively. They correspond to three choices of k_n for logGIC defined in Section 1.3.2, which are $k_n^0 = \log(n)$, $k_n^1 = \log\log n \log p$, and $k_n^2 = \log p\sqrt{\log n}$, respectively. Our method with $\lambda = 0$ is also applied to obtain the unpenalized estimate for θ_0. In this example, only V2 is used.

Model 3, COR1, $p = 20$				
Method	Angle	FPR	TPR	Time
BS-SIM-V1	0.907 (0.433)	0.044	1.000	18.40
BS-SIM-V2	**0.902** (0.424)	0.040	1.000	**9.98**
BL-SIM-V1	1.936 (0.642)	0.289	1.000	25.95
BL-SIM-V2	1.531 (0.537)	0.216	1.000	21.65
SIM-LASSO-V2	3.716 (1.707)	0.209	1.000	26.44
SMAVE-V2	20.470 (32.846)	0.507	0.980	48.35
SIM-Bridge-V2	2.785 (10.489)	**0.016**	0.985	64.97

Table 1.2.
Comparison between our methods to the existing methods in low dimensional scenario: Model 3 with COR1.

Table 1.3 shows the comparison results on the four aforementioned assessments. In terms of estimation accuracy and computing efficiency, both the BL-SIM estimators and the BS-SIM estimators are considerably better than the unpenalized estimator. It is a strong sign that the proposed regularization approach substantially helps with efficiently providing a more accurate estimator. Comparing the two proposed estimators, the BS-SIM estimators slightly outperform the BL-SIM estimators in estimation. In terms of the performance on variable selection consistency, the BS-SIM estimators are dramatically better. More specifically, the BL-SIM estimators have a more than 3-fold higher average FPR, indicating applying LASSO is more likely to lead to an overfitted model. In the computational efficiency aspect, BS-SIM is slightly faster than BL-SIM. As for the comparison among the three BS-SIM estimators, the estimator using logBIC has a noticeably higher average FPR than the estimators with λ chosen by logGIC1 and logGIC2. Since the number of predictors is not that small ($p = 70$) in this example, this observation on FPR is consistent with the fact that logBIC yields a overfitted model when the dimension p increases. The performance

of the two penalized estimators with λ chosen by logGIC1 and logGIC2 are similar in terms of the four metrics.

Model 1, COR1, $p = 70$					
Method	Selection of λ	Angle	FPR	TPR	Time
BS-SIM-V2	logBIC	4.836 (2.309)	0.124	0.984	0.781
	logGIC1	4.529 (1.868)	0.050	0.976	0.701
	logGIC2	**4.526** (1.976)	**0.015**	0.968	0.610
BL-SIM-V2	logBIC	7.010 (5.421)	0.466	**0.995**	0.796
	logGIC1	6.828 (4.178)	0.457	**0.995**	0.577
	logGIC2	6.228 (3.114)	0.428	0.975	**0.453**
Unpenalized	$\lambda = 0$	50.350 (7.587)	NA	NA	12.749

Table 1.3.
Comparison between the penalized estimator and the unpenalized estimator.

1.5.3 Performance of the proposed method for several choices of a

This example examines the performance of the proposed estimator for several choices of a. 100 samples of size 100 are simulated from Model 2 with COR1. The other settings are $p = 50$, $q = 8$, and $\theta_0 = (2.0, -1.0, 0.5, 1.0, -1.5, 1.0, -0.3, 1.2, 0, \cdots, 0)^T$. BL-SIM and BS-SIM with several choices of a are applied, and their performance on the four assessments introduced previously is compared. We rely on both logBIC and logGIC2 defined in Example 1 to choose the tuning parameter λ, and only use V2 in this example.

The comparison results are shown in Table 1.4. It can be observed that as a increases, both Angle and FPR decrease first, then increase. Furthermore, when a

continues to increase, the performance of the BS-SIM estimator approaches that of the BL-SIM estimator. In theory, the performance of the BS-SIM estimator in terms of variable selection should improve when a decreases. Nevertheless, the pattern shown in Table 1.4 implies that there exists certain computational difficulty in finding a consistent estimate when a is extremely small. On the other hand, the BS-SIM estimator with $a = 0.1$ outforms the rest in terms of selection consistency. When it comes to estimation accuracy, the performance of the BS-SIM estimator with $a = 0.1$ is also satisfactory. Therefore, we recommend to use $a = 0.1$ in practice. For the remaining examples, we fix a at 0.1, unless otherwise specified.

Model 2, COR1, $p = 50$					
Method	Selection of λ	Angle	FPR	TPR	Time
BS-SIM-V2	logBIC	1.392 (0.578)	0.075	1	**26.00**
($a = 0.01$)	logGIC2	1.160 (0.440)	0.013	1	
BS-SIM-V2	logBIC	1.178 (0.396)	0.029	1	32.78
($a = 0.05$)	logGIC2	**1.122** (0.381)	0.005	1	
BS-SIM-V2	logBIC	1.197 (0.399)	0.029	1	38.65
($a = 0.10$)	logGIC2	1.164 (0.397)	**0.004**	1	
BS-SIM-V2	logBIC	1.503 (0.468)	0.140	1	77.70
($a = 0.50$)	logGIC2	1.504 (0.474)	0.132	1	
BS-SIM-V2	logBIC	1.639 (0.472)	0.384	1	103.97
($a = 1.00$)	logGIC2	1.630 (0.470)	0.383	1	
BL-SIM-V2	logBIC	1.938 (0.557)	0.417	1	103.63
($a = \infty$)	logGIC2	1.925 (0.541)	0.413	1	

Table 1.4.
Comparison between the LASSO and the SICA penalties with various choices of a for moderate p.

1.5.4 Performance of the proposed method for moderate p

This example illustrates the performance of the proposed estimator for moderate p. We focus on the comparison between our method and other existing methods. In this example, we implement the proposed BS-SIM method with $a = 0.1$, and the proposed BL-SIM method, as well as the SIM-LASSO method proposed by [22], the SMAVE method proposed by [21], and the MAVE method coupled with the Bridge penalty, proposed by [23]. The last method is denoted as SIM-Bridge hereafter. For SIM-LASSO, the tuning parameter is chosen by 10-fold cross-validation, and for SMAVE and SIM-Bridge, the tuning parameter is selected based on BIC, as suggested in the original papers. Moreover, all of these three methods only use V2. In this example, we let p be moderate and vary it from 50 to 70. 100 data sets of size 100 are simulated from the following settings:

1. Setting 1: Model 1, COR2, and $p = 50$;
2. Setting 2: Model 2, COR3, and $p = 70$;
3. Setting 3: Model 3, COR1, and $p = 50$.

Note that Model 3 is the most difficult one, thus its dimensionality is set to 50. Under each setting, let $q = 8$, and $\theta_0 = (2, -1, 1, -0.5, 0, -1.5, 1.0, -0.3, 1.2, \cdots, 0)^T$. In this example, logGIC2 is used to choose λ. The comparison results are given in Tables 1.5 - 1.7

For both Setting 1 and Setting 2, the BS-SIM estimators outperform the rest in terms of both estimation accuracy and selection consistency. They are followed by the SIM-Bridge estimator in terms of selection performance. The other three methods do not produce satisfactory performance on variable selection, as they tend to result in overfitted models. In the computational efficiency aspect, the proposed BS-SIM method is also among the best. For Setting 3, the quadratic link function is used. Since X_i's are generated from a multivariate normal distribution, they concentrate around 0. However, the MAVE based methods rely on local linear expansion,

thus they do not perform well around the origin, and break down for this quadratic link function. Hence, only the results from the proposed methods are presented for this setting. It can be observed that the proposed BS-SIM method exhibits acceptable performance in each aspect, and considerably outperforms the proposed BL-SIM method. Lastly, it is also worth pointing out that satisfactory performance can be maintained for the proposed methods under other combinations of model setting and correlation structure.

Model 1, COR2, $p = 50$				
Method	Angle	FPR	TPR	Time
BS-SIM-V1	4.866 (2.850)	0.019	0.963	34.463
BS-SIM-V2	**4.819** (2.749)	**0.017**	0.963	**25.262**
BL-SIM-V1	13.269 (3.956)	0.347	0.963	64.532
BL-SIM-V2	8.626 (3.121)	0.160	0.968	52.522
SIM-LASSO-V2	7.476 (2.085)	0.552	**0.990**	56.845
SMAVE-V2	12.493 (9.445)	0.316	0.898	39.747
SIM-Bridge-V2	7.686 (4.434)	0.058	0.901	102.349

Table 1.5.
Comparison between the proposed methods and the other existing methods in moderate dimensional scenario: Setting 1.

1.5.5 Performance of the proposed method for large p

This example demonstrates the performance of the proposed estimator for large p. In this example, two choices of the dimension, $p = 200$ and $p = 400$, are examined. The other settings are $q = 10$, $n = 100$ and $\theta_0 = (2, -1, 0.5, 1, -1.5, 1.2, -0.8, 0.6, 1, -1, 0, 0, \cdots, 0)^T$. For $p = 200$, the results under all of the three aforementioned correlation structures

Model 2, COR3, $p = 70$				
Method	Angle	FPR	TPR	Time
BS-SIM-V1	**2.250** (0.959)	**0.012**	0.999	146.390
BS-SIM-V2	2.429 (0.951)	0.025	0.999	169.485
BL-SIM-V1	7.569 (2.160)	0.694	0.994	519.186
BL-SIM-V2	5.060 (1.673)	0.728	**1.000**	494.885
SIM-LASSO-V2	6.602 (1.920)	0.684	0.993	212.528
SMAVE-V2	9.275 (4.629)	0.784	0.995	**65.740**
SIM-Bridge-V2	6.775 (4.114)	0.094	0.906	166.558

Table 1.6.
Comparison between the proposed methods and the other existing methods in moderate dimensional scenario: Setting 2.

Model 3, COR1, $p = 50$				
Method	Angle	FPR	TPR	Time
BS-SIM-V1	10.003 (21.004)	0.147	0.956	466.565
BS-SIM-V2	**9.346** (19.750)	**0.142**	0.965	**218.957**
BL-SIM-V1	22.328 (27.810)	0.644	**0.979**	1037.898
BL-SIM-V2	35.757 (29.855)	0.705	**0.979**	413.221

Table 1.7.
Comparison between the proposed methods and the other existing methods in moderate dimensional scenario: Setting 3.

are exhibited; for $p = 400$, the proposed method cannot produce acceptable results when there exists correlation among the predictors. Nevertheless, with more data points, the proposed BS-SIM method can still handle this high dimensional scenario with correlation among the predictors. However, we exclusively focus on COR1 and

$n = 100$ for $p = 400$ here. The proposed BL-SIM method suffers greatly from over-selection and is too time-consuming in the large p scenario, and SIM-LASSO and SMAVE break down in this example. Therefore, only the results from the proposed BS-SIM method and SIM-Bridge are presented. Since V1 poses no restriction on the magnitude of ϕ, the estimation with V1 becomes noticeably more unstable, and slower for some models, as p increases. Therefore, it is recommended to use V2 when p is large. Based on our simulation studies, V1 and V2 lead to comparable results under Model 1; whereas for Model 2, V2 is much more preferable. As for the choice of k_n, it is recommended to use $k_n^3 = \log p \log n$.

Table 1.8 shows the results on the four metrics. In terms of estimation accuracy and selection consistency for Model 1 and $p = 200$, the proposed BS-SIM method yields reasonably accurate estimates, while SIM-Bridge does not perform well under all of the three correlation structures. For Model 2 and $p = 200$, comparable results on selection consistency are obtained. However, the proposed BS-SIM method produces more accurate estimate than SIM-Bridge, especially under COR3. When $p = 400$, SIM-Bridge fails, while the proposed BS-SIM method can still yield satisfactory results. In terms of computational capacity, for the proposed BS-SIM method, it takes about 20 minutes on average to complete one run for $p = 200$, and takes less than two hours for $p = 400$. Considering that this amount of time encompasses the search for the optimal λ on a dense grid, this computational efficiency is still acceptable. Moreover, the proposed BS-SIM method is noticeably more efficient than SIM-Bridge in this example.

1.5.6 Evaluation of the Irrepresentable Conditions

This example focuses on the impact of the Irrepresentable Conditions. In this example, let $n = 200$, $p = 30$, $q = 6$ $N = 5$ and $\theta_0 = (2.0, -1.0, 0.5, 1.0, 0.3, -0.7, 0, \cdots, 0)^T$, and we exclusively focus on Model 1. It is clear that, for a given combination of design

p	Model	COR	Method	Angle	FPR	TPR	Time
200	1	1	BS-SIM-V1	**5.355** (5.188)	**0.001**	**0.972**	**490.5**
			BS-SIM-V2	7.086 (8.536)	0.004	0.945	858.1
			SIM-Bridge-V2	30.585 (12.085)	0.292	0.662	2216.0
	1	2	BS-SIM-V1	**8.696** (9.836)	**0.007**	**0.919**	**870.0**
			BS-SIM-V2	10.552 (11.822)	0.021	0.894	894.8
			SIM-Bridge-V2	35.201 (11.743)	0.317	0.585	2196.0
	1	3	BS-SIM-V1	16.904 (16.423)	**0.013**	0.792	**498.6**
			BS-SIM-V2	**15.974** (17.906)	0.024	**0.808**	707.4
			SIM-Bridge-V2	47.550 (11.925)	0.438	0.381	2222.0
	2	1	BS-SIM-V2	**2.124** (0.644)	0.137	**1.000**	**1823.6**
			SIM-Bridge	3.617(2.258)	**0.041**	0.991	1841.0
	2	2	BS-SIM-V2	**2.231** (0.680)	0.039	**1.000**	**1510.9**
			SIM-Bridge-V2	4.365 (3.029)	**0.034**	0.984	2262.0
	2	3	BS-SIM-V2	**2.724** (1.497)	**0.057**	**0.999**	**1786.1**
			SIM-Bridge-V2	12.435 (9.140)	0.227	0.898	2415.0
400	1	1	BS-SIM-V1	17.533 (16.648)	0.060	0.775	2296.7
			BS-SIM-V2	**12.837** (15.665)	**0.035**	**0.855**	**1991.8**
	2	1	BS-SIM-V2	2.508 (2.258)	0.213	1.000	6519.0

Table 1.8.
Performance of BS-SIM with $a = 0.1$ under several settings in high dimensional scenario.

matrix X, link function f and true index θ_0, the Irrepresentable Conditions depend on the choice of a and the Identifiability Constraint used. The following sequence of a, $a = (0.05, 0.1, 0.3, 0.5, 1.0, 2.0, 5.0)$, as well as $a = \infty$, are examined, and V1 and V2 are applied.

The simulation scheme is as follows. A covariance matrix Σ is first generated from Wishart(p,p), as done in [17]. Then we generate a sample of 100 observations of X from N(0,Σ), and standardize them. 100 normalized designs are generated in this way. Next, for each generated design, we run the following simulation 100 times. During each simulation, n copies of ε_i are sampled from $N(0, 0.3^2)$, and y_i's are calculated according to Model 1. Subsequently, the proposed method with the various choices of a specified above are applied, and the percentage of times that the applied method can identify the true model along the solution path is recorded.

Since it is difficult to quantify the Irrepresentable Conditions for BS-SIM, we compute

$$\bar{\eta}_\infty \;\; = \;\; 1 - \|\bar{C}_0(21)\bar{C}_0^{-1}(11)\text{sign}(\phi_0(1))\|_\infty,$$

associated with the Irrepresentable Condition for BL-SIM for each design, instead. The sign of $\bar{\eta}_\infty$ indicates whether the Irrepresentable Condition for BL-SIM holds. That is, if $\bar{\eta}_\infty > 0$, the Irrepresentable Condition for BL-SIM holds; otherwise, it fails to hold. Considering the fact that the Irrepresentable Conditions for BS-SIM are more relaxed than that for BL-SIM, $\bar{\eta}_\infty$ also implies how strongly the Irrepresentable Conditions for BS-SIM satisfy or fail, to some extent. $\bar{\eta}_\infty$ is computed for each generated design according to each Identifiability Constraint. The summary can be found in Table 1.9.

We first look at how the magnitude of $\bar{\eta}_\infty$ affects the performance of the proposed BL-SIM method in selecting the true model. On the two top graphs in Figure 1.3, the percentage of times that the true model can be identified by the proposed BL-SIM method is plotted against the corresponding $\bar{\eta}_\infty$, for the two Identifiability Constraints separately. It can be observed that the percentage increases as $\bar{\eta}_\infty$ increases, for both Identifiability Constraints. The increase is the sharpest around 0, as expected. On

the two bottom graphs in Figure 1.3, the percentage of times of achieving selection consistency for the proposed BS-SIM method with $a = 2$ is plotted against $\bar{\eta}_\infty$, for the two Identifiability Constraints separately. It is obvious that the percentage for BS-SIM is larger than that for BL-SIM at any $\bar{\eta}_\infty$ for both constraints. It is consistent with our expectation that BS-SIM with finite a should perform better in terms of variable selection than BL-SIM.

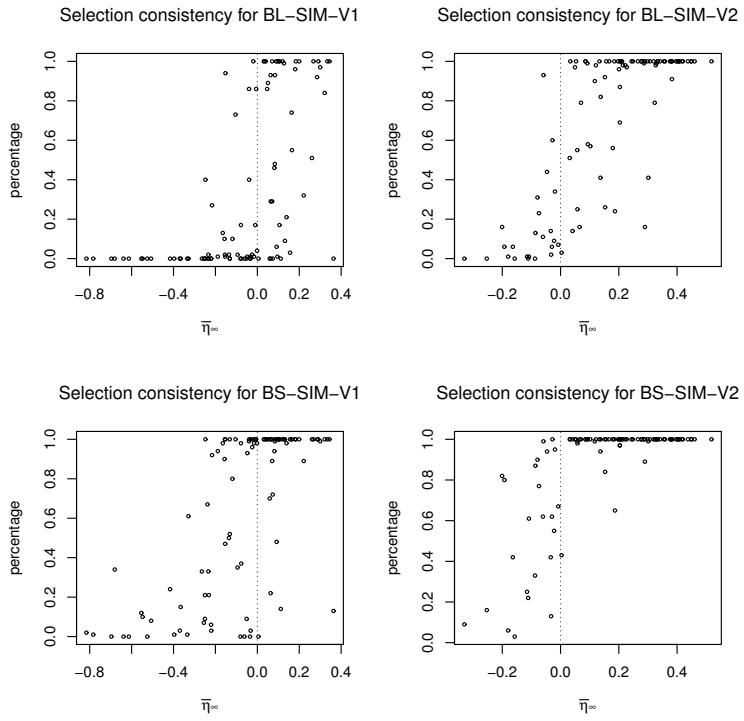


Fig. 1.3. The percentages that the proposed BL-SIM method and the proposed BS-SIM method with $a = 2$ select the true model versus $\bar{\eta}_\infty$ for both Identifiability Constraints.

Next, we examine how a affects the proposed method in terms of selection consistency in more detail. The average percentages of times that the true model can be selected with various choices of a are shown in Table 1.9. In theory, the Irrepresentable Conditions become more restrictive when a increases. Thus, it is expected that it is less likely to choose the true model when a increases. However, as indicated

in Table 1.9, when a gets larger, the percentage of runs that the true model can be identified increases slightly first, then decreases; and when a continues to increase, the percentage for the BS-SIM estimator approaches that for the BL-SIM estimator. This particular pattern for the performance of BS-SIM implies that for extremely small a, it is computationally slightly more difficult to find a consistent estimator, although the Irrepresentable Conditions are relaxed. These observations on the impact of a are in line with those stated in [26].

The results in Table 1.9 also cast light on the role that the Identifiability Constraint plays. In most cases shown in Table 1.9, using V2 leads to a higher chance of recovering the true model. The difference of the chances becomes larger as a increases. This observation is consistent with the observation on the relative magnitude on $\bar{\eta}_\infty$, as shown in 1.10. Among the 100 designs generated above, 92% of them have larger $\bar{\eta}_\infty$ for V2. It is probably due to the fact that the Irrepresentable Conditions for V2 contains more information than those for V1.

a	0.05	0.10	0.30	0.50	1.00	2.00	5.00	∞
V1	0.9995	1.0000	1.0000	0.9924	0.8741	0.6548	0.4665	0.3382
V2	0.9990	0.9999	0.9997	0.9956	0.9562	0.8786	0.7850	0.6909

Table 1.9.
Average percentages of times that the true model can be selected with various choices of a.

1.5.7 Comparison of CV, logGIC and GIC under the violation of the sparsity assumption

In this section, we will illustrate the performance of the three tuning parameter selection methods under the violation of the sparsity assumption. A setting similar to the real data setting is applied. That is, $n = 100$ and $p = 180$. Let $q = 15$ and

	V1			V2		
	min.	mean	max.	min.	mean	max.
$\bar{\eta}_\infty$	-1.177	-0.080	0.436	-0.330	0.160	0.518

Table 1.10.
Summary on $\bar{\eta}_\infty$ for both Identifiability Constraints.

$\theta_0 = (0.5, 0.5, \cdots, 0.5, 0, \cdots, 0)^T$. 100 random samples of size n are generated from Model 1 with COR1. The BS-SIM method with $a = 0.1$ and $N = 2$ are applied to each sample. Each time, CV, logGIC, and GIC are used to determine the best λ and yield the estimates for θ_0. The performance of the obtained estimates on Angle, FPR and TPR is recorded. In addition, we also use the average size of the selected model and the proportion of times that the correct model is selected to assess the performance of each tuning parameter selection method. These two metrics are denoted as Size and CorrectModel in the subsequent table. In this example, four choices of k_n for both logGIC and GIC are considered, which are log$n$, loglog$p$log$n$, log$p\sqrt{\log n}$ and log$p$log$n$.

The comparison result is shown in Table 1.11. For logGIC, using loglog$p$log$n$ and log$p\sqrt{\log n}$ produce exactly the same results as those produced by using log$n$, thus only one of them is displayed. Table 3 suggests that for logGIC there is a lack of an appropriate value for k_n. That means with a small k_n, logGIC always leads to the full model, while with a large k_n, it frequently selects a extremely small model. This pattern becomes more evident when q continues to increase. Meanwhile, GIC can steadily yield reasonable results, and is obviously more advantageous than CV. With an appropriate k_n, the frequency that GIC can identify the true model is twice of that for CV. Moreover, the model obtained by using GIC with an appropriate k_n is more accurate and smaller in size, and also has lower FPR and higher TPR than

Method	k_n	Angle	FPR	TPR	Size	CorrectModel
CV	NA	35.18	0.34	0.85	33.14	0.14
logGIC	$\log(n)$	41.49	0.92	1.00	180.00	0.00
logGIC	$\log(n)\log(p)$	47.81	0.01	0.43	8.05	0.31
GIC	$\log(n)$	30.65	0.40	0.90	37.10	0.29
GIC	$\log(n)\log\log(p)$	28.58	0.31	0.89	31.56	0.30
GIC	$\sqrt{\log(n)}\log(p)$	28.23	0.27	0.87	29.20	0.30
GIC	$\log(n)\log(p)$	30.65	0.20	0.81	25.00	0.27

Table 1.11.
Performance comparison of CV, logGIC and GIC when $q = 15$.

that obtained by using CV. In conclusion, we believe it is beneficial to use GIC to conduct the tuning parameter selection in practice when the true model is not sparse.

1.6 Real Data Application

1.6.1 Skin Cutaneous Melanoma Data

Melanoma is a type of cancer that starts with a certain type of skin cell called melanocyte. There are more than 70 thousands people diagnosed with skin cutaneous melanoma in U.S. each year. While it is not the most prevalent type of skin cancer, skin cutaneous melanoma is believed to be the most aggressive. It can occur in all types of skins, and spread widely to other organs of the body. This type of cancer has a number of potential risk factors. However, it is most likely to be caused by intensive exposure to ultraviolet radiation. Early-stage cutaneous melanomas can often be treated with surgery effectively, while more advanced ones need other treatments or a combination of treatments, such as immunotherapy, chemotherapy, and radiation therapy.

In this real data analysis, we aim to study how the protein expression levels influence the survival time of the patients who suffer from skin cutaneous melanoma. We download the relevant data from The Cancer Genome Atlas (TCGA) data portal. There are two sets of files that we use: clinic dataset and protein expression datasets. On the clinical data, there are in total 433 patients. Their demographic information, tumor status, vital status and survival time are recorded. On a separate set of files, the expression levels of 181 proteins are measured for 207 patients using the M.D. Anderson Reverse Phase Protein Array Core platform. We combine these two files, and based on our goal, we only retain those patients that failed to survive and had protein expression level measured for further analysis. After this pre-processing, we have 94 patients, and the expression levels of 181 proteins. The expression levels are subsequently standardized and used as the predictors. The survival time is taken logarithm, and treated as the response.

1.6.2 Analysis on Skin Cutaneous Melanoma Data with BS-SIM

We apply the proposed BS-SIM method with $a = 0.1$ and 2 interior knots to the aforementioned processed data. Since we speculate there exists a relatively large number of relevant proteins, the GIC criterion introduced at the end of Section 1.3.2 with $k_n = log(n)loglog(p)$ is used to determine the optimal tuning parameter λ. We also try the logGIC defined in Section 1.3.2 with various choices of k_n. However, it fails to effectively yield a reasonable model. This behaviour of the logGIC criterion also to some extent confirms that the number of relevant proteins is relatively large.

Based on the combination mentioned in the last paragraph, we are able to select 30 proteins, which are P21-R-V, 4E-BP1-pT37-T46-R-V, ACC1-R-E, Beclin-G-C, Dvl3-R-V, Notch1-R-V, p27-pT157-R-C, p53-R-E, Paxillin-R-C, PEA15-R-V, PTEN-R-V, Smad1-R-V, Smad4-M-V, Src-pY527-R-V, Syk-M-V, Tuberin-R-E, YB-1-pS102-R-V, FoxM1-R-V, MYH11-R-V, RBM15-R-V, Rictor-R-C, SCD1-M-V, TAZ-R-V, TSC1-

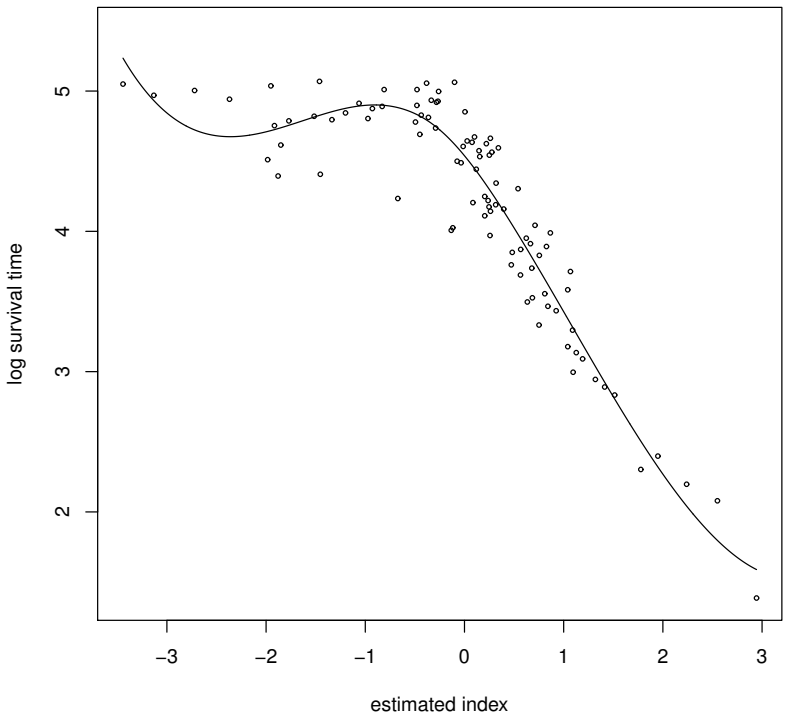


Fig. 1.4. The plot of the fitted regression function and the observed log survival time vesus the estimated index for the Skin Cutaneous Melanoma data.

R-C, Tuberin-pT1462-R-V, VHL-M-C, 53BP1-R-E, c-Jun-pS73-R-V, Caveolin-1-R-V and Rb-pS807-S811-R-V. The estimated projection direction is given below.

$$\begin{aligned}\text{Projection direction} = {} & P21 + 0.63 \cdot \text{4E-BP1} + 0.24 \cdot \text{53BP1} + 0.54 \cdot \text{ACC1} + 0.76 \cdot \text{Beclin} \\ & -0.21 \cdot \text{c-Jun} + 0.25 \cdot \text{Caveolin} - 0.24 \cdot \text{Dvl3} + 0.69 \cdot \text{Notch1} + 0.56 \cdot \text{p27} - 0.37 \cdot \text{p53} \\ & +0.59 \cdot \text{Paxillin} + 0.24 \cdot \text{PEA15} + 0.31 \cdot \text{PTEN} - 0.19 \cdot \text{Rb} - 0.23 \cdot \text{Smad1} + 0.89 \cdot \text{Smad4} \\ & +0.46 \cdot \text{Src} + 0.58 \cdot \text{Syk} + 0.56 \cdot \text{Tuberin} + 0.17 \cdot \text{YB} - 0.25 \cdot \text{FoxM1} - 0.42 \cdot \text{MYH11} \\ & +0.64 \cdot \text{RBM15} + 0.10 \cdot \text{Rictor} - 0.41 \cdot \text{SCD1} - 0.57 \cdot \text{TAZ} - 0.74 \cdot \text{TSC1} - 0.49 \cdot \text{Tuberin} \\ & +0.53 \cdot \text{VHL}.\end{aligned}$$

The final fitted regression function is plotted against the estimated direction in Figure 1.4.

Out of these detected proteins, the irregular expression of the p21, p27, p53, PTEN, TAZ, Notch1, Caveolin, 53BP1, TSC1, Rb and Tuberin proteins have been shown to be related to the survival or occurence of the Skin Cutaneous Melanoma [38–42]. This partially demonstrates the effectiveness of BS-SIM.

1.7 Linearly Constrained Single Index Model

1.7.1 Single Index Model with Linear Constraints

In many applications, prior information about the magnitude of the effects of the predictors on the response is available. Incorporating this information into the estimation procedure can bring considerable value and lead to more accurate results. The problem of variable selection for the linear model under linear constraints has been studied in the literature; see [43, 44] among others. Since the single index model is a intuitive generalization of the linear model, it is also of interest to study how to conduct variable selection for the single index model under linear constraints. Recall that the single index model requires an identifiability constraint that is imposed on the scale of θ. One distinct difference between the linear model and the single index model under linear constraints is that the identifiability constraint used in the

single index model can be important. For some equality constraints, they are not affected by what identifiability constraint is being used. These equality constraints are imposed on the relative scale of the components of θ_0. For instance, $\theta_2 = \theta_3 + \theta_4$ is not influenced by the identifiability constraint applied. However, for most linear constraints, equality or inequality, they need to be scaled according to the identifiability constraint. To name a few among them, $\theta_2 > 0$ and $\theta_2 + \theta_3 = 2$. These linear constraints correspond to the true index θ_0 under certain scale. In this section, for the purpose of convenience, we will use Identifiability Constraint 1 defined in the previous Section 1.2, which is $\theta_1 = 1$. Since it is also a linear constraint, Identifiability Constraint 1 can greatly facilitate the algorithm that we are going to introduce next.

1.7.2 Coordinate Descent Algorithm for Linearly Constrained Single Index Model

Recall that the objective function for BS-SIM, $R(\phi; \lambda)$, is written as

$$R(\phi; \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_\theta(t_\theta^i) \right)^2 + \lambda \sum_{j=1}^{p-1} \rho_a(|\phi_j|),$$

where $\hat{f}_\theta$ is the cubic B-spline estimator of f, λ is a tuning parameter, and $\rho_a(u)$ denotes the SICA penalty functions. For variable selection of the linearly constrained single index model (LC-SIM) problem, we still use the same framework. Thus, LC-SIM problem can be formulated as the following optimization problem.

$$\min_{\phi} R(\phi; \lambda), \qquad \text{subject to } C\phi \geq d, \quad E\phi = f. \tag{1.13}$$

Here, both linear equality constraints $E\phi = f$ and linear inequality constraints $C\phi \geq d$ are considered. Furthermore, E is a $l \times (p-1)$ matrix, where l is the number of equality constraints; C is a $m \times (p-1)$ matrix, where m is the number of inequality constraints; d and f are vectors of length m and l, respectively.

Adopting the approach proposed by Rosset and Zhu [45], we introduce the slacker variables ϕ_k^+ and ϕ_k^- such that $\phi_k = \phi_k^+ - \phi_k^-$, $\phi_k^+ \geqslant 0$ and $\phi_k^- \geqslant 0$,for $k = 1, 2, \cdots, p-1$.

Let $\phi^+ = (\phi_1^+, \phi_2^+, \cdots, \phi_{p-1}^+)^T$, and $\phi^- = (\phi_1^-, \phi_2^-, \cdots, \phi_{p-1}^-)^T$. Then we solve for ϕ at any given λ on a dense grid. Same as in Section 1.3, we still use the local quadratic approximation strategy 1.14 to H and the local approximation 1.15 to the SICA penalty function, at the current estimate $\phi^{(s)}$.

$$H(\phi) \approx \frac{1}{2}\phi^T H^{(2)}(\phi^{(s)})\phi - \phi^T \left(H^{(2)}(\phi^{(s)})\phi^{(s)} - H^{(1)}(\phi^{(s)})\right) + \text{constant}, \tag{1.14}$$

$$\rho_a(|\phi_j|) = \rho_a(|\phi_j^{(s)}|) + \rho_a'(|\phi_j^{(s)}|)(|\phi_j| - |\phi_j^{(s)}|), \text{ for } j = 1, 2, \cdots, p-1. \tag{1.15}$$

These two approximations entail that at the current $\phi^{(s)}$, Problem 1.13 can be approximated by

$$\begin{aligned} \min_{\phi^+, \phi^-} \frac{1}{2}(\phi_j^+ - \phi_j^-)^T H^{(2)}(\phi^{(s)})(\phi_j^+ - \phi_j^-) - (\phi_j^+ - \phi_j^-)^T \left(H^{(2)}(\phi^{(s)})\phi^{(s)} - H^{(1)}(\phi^{(s)})\right) \\ + \lambda \sum_{j=1}^{p-1} w_j^{(s)}(\phi_j^+ + \phi_j^-), \\ \text{subject to } \phi_j^+ \geq 0,\ \phi_j^- \geq 0 \text{ for } k = 1, \ldots, p-1, \\ C(\phi^+ - \phi^-) \geq d \text{ and } E(\phi^+ - \phi^-) = f, \end{aligned} \tag{1.16}$$

where $w_j^{(s)} = \rho_a'(|\phi_j^{(s)}|)$ for $j = 1, 2, \cdots, p-1$. Let $S = H^{(2)}(\phi^{(s)})$, and $\beta = H^{(2)}(\phi^{(s)})\phi^{(s)} - H^{(1)}(\phi^{(s)})$.

To work out a solution to Problem 1.16, we first figure out the KTT conditions as follows.

$$S\phi - \beta + \lambda w - v^+ - C^T\gamma - E^T h = 0, \tag{1.17}$$

$$-S\phi + \beta + \lambda w - v^- + C^T\gamma + E^T h = 0, \tag{1.18}$$

$$v_k^+ \phi_k^+ = 0, \tag{1.19}$$

$$v_k^- \phi_k^- = 0, \tag{1.20}$$

$$\gamma_s(C_s\phi - d_s) = 0, \tag{1.21}$$

$$E_t\phi - f_t = 0, \tag{1.22}$$

$$\phi^+ \geq 0, \quad \phi^- \geq 0, \quad v_k^+ \geq 0, \quad v_k^- \geq 0, \quad \gamma_s \geq 0, \tag{1.23}$$

for $k = 1, 2, \cdots, p-1$, $s = 1, 2, \cdots, m$, and $t = 1, 2, \cdots, l$. Here, v^+ and v^- are vectors of length $p-1$; γ is a vector of length m; h is a vector of l; and all of them

are Lagrange multipliers. Moreover, C_s and E_t denote the s-th and t-th rows of C and E.

For $\phi^{(s)}$, denote the current active set as

$$A = \{i : \phi_i^{(s)} \neq 0\} \subset \{1, \ldots, p-1\},$$

and define the current active set for inequality constraints as

$$\mathcal{B} = \{j : C_j \phi^{(s)} = d_j\} \subset \{1, \ldots, k\}.$$

Define the complement of A and $\mathcal{B}$ as $A^c = \{i : \{1, \ldots, p-1\} \setminus A\}$ and $\mathcal{B}^c = \{i : \{1, \ldots, k\} \setminus \mathcal{B}\}$. Let I, J, and K be sets of integers. Let Z_{IJ} denote the submatrix of Z whose row and column are indexed by I and J, respectively. This notation is not to be confused with C_s and E_t above. When there is only one letter in the subscript, it means we are subsetting certain rows of a matrix; when there are two letters, it indicates we are subsetting both certain rows and columns of a matrix. Let W_K be a subvector of W whose entries are formed by the set K. Then the KTT conditions for the candidate ϕ's subject to the current active set A and the current active set for inequality constraints $\mathcal{B}$ are given by Equations 1.24 through 1.30.

$$S_{AA}^{(s)} \phi_A - \beta_A^{(s)} - C_{\mathcal{B}A}^T \gamma_{\mathcal{B}} - E_A^T h = -\lambda w_A^{(s)} \cdot \text{sign}(\phi_A^{(s)}), \tag{1.24}$$

$$S_{\mathcal{B}A}^{(s)} \phi_A - \beta_{A^c}^{(s)} - C_{\mathcal{B}A^c}^T \gamma_{\mathcal{B}} - E_{A^c}^T h + \lambda w_{A^c}^{(s)} \geqslant 0, \tag{1.25}$$

$$-S_{\mathcal{B}A}^{(s)} \phi_A + \beta_{A^c}^{(s)} + C_{\mathcal{B}A^c}^T \gamma_{\mathcal{B}} + E_{A^c}^T h + \lambda w_{A^c}^{(s)} \geqslant 0, \tag{1.26}$$

$$E_A \phi_A = f, \tag{1.27}$$

$$C_{\mathcal{B}A} \phi_A = d_{\mathcal{B}}, \tag{1.28}$$

$$\gamma_{\mathcal{B}} \geqslant 0, \tag{1.29}$$

$$C_{\mathcal{B}^c A} \phi_A > d_{\mathcal{B}^c}, \tag{1.30}$$

where $\cdot$ means componentwise multiplication, $S^{(s)} = H^{(2)}(\phi^{(s)})$, and $\beta^{(s)} = H^{(2)}(\phi^{(s)})\phi^{(s)} - H^{(1)}(\phi^{(s)})$. For convenience, we define the following notations.

$$G_{\mathcal{B}A} = \begin{pmatrix} C_{\mathcal{B}A} \\ E_A \end{pmatrix}, \qquad g_{\mathcal{B}} = \begin{pmatrix} d_{\mathcal{B}} \\ f \end{pmatrix}, \text{ and } \alpha_{\mathcal{B}} = \begin{pmatrix} \gamma_{\mathcal{B}} \\ h \end{pmatrix}.$$

Then some calculation on Equations 1.24 through 1.30 entails

$$\phi_A^{(s+1)} = S_{AA}^{-1}\big[\beta_A - \lambda w_A \cdot \mathrm{sign}(\phi_A^{(s)}) + G'_{\mathcal{B}A}(G_{\mathcal{B}A}S_{AA}^{-1}G'_{\mathcal{B}A})^{-1}\big(g_{\mathcal{B}} - G_{\mathcal{B}A}S_{AA}^{-1}\beta_A + G_{\mathcal{B}A}S_{AA}^{-1}\lambda w_A \cdot \mathrm{sign}(\phi_A^{(s)})\big)\big], \quad (1.31)$$

$$\alpha_{\mathcal{B}} = (G_{\mathcal{B}A}S_{AA}^{-1}G'_{\mathcal{B}A})^{-1}[g_{\mathcal{B}} + G_{\mathcal{B}A}S_{AA}^{-1}\lambda w_A \cdot \mathrm{sign}(\phi_A^{(s)}) - G_{\mathcal{B}A}S_{AA}^{-1}\beta_A]. \quad (1.32)$$

Here w, β and S depend the current the estimate $\phi^{(s)}$ and are updated during each iteration. However, for simplicity, we drop the notation $^{(s)}$ in these three terms on the above and the following equations when there is no confusion. Subsequently, we also have

$$v_{A^c}^{+} = \lambda w_{A^c} + S_{A^cA}\phi_A^{(s+1)} - \beta_{A^c} - G_{\mathcal{B}A^c}^{T}\alpha_{\mathcal{B}}, \quad (1.33)$$

$$v_{A^c}^{-} = \lambda w_{A^c} - S_{A^cA}\phi_A^{(s+1)} + \beta_{A^c} + G_{\mathcal{B}A^c}^{T}\alpha_{\mathcal{B}}. \quad (1.34)$$

After calculating the updates through Equations 1.31 through 1.34, we need to check if A and $\mathcal{B}$ should be updated accordingly. If they do, we update them, and proceed to update the KTT conditions with the new active sets; if they remain the same, we are ready to start the next iteration of updating ϕ. The detailed algorithm can be found in Algorithm 2.

For the tuning parameter selection of λ under LC-SIM, one clear choice is to use logGIC or GIC defined in Section 1.3.2.

Algorithm 2 *Coordinate Descent Algorithm for BS-SIM with LC*

For a given λ,

1. Initialize ϕ to be $\hat{\phi}^{(0)}$ and let $s = 0$.

2. Given $\hat{\phi}^{(s)} = (\hat{\phi}_1^{(s)}, \hat{\phi}_2^{(s)}, \ldots, \hat{\phi}_{p-1}^{(s)})^T$, obtain A and $\mathcal{B}$. Calculate the quadratic approximation 1.14 to $H(\phi)$ and the linear approximation 1.15 to $p_\lambda(\phi)$. Compute $w^{(s)}$, $\beta^{(s)}$ and $S^{(s)}$.

3. Use Equations 1.31 to 1.34 to yield $\hat{\phi}_A^{(s+1)}$, α_B, $v_{A^c}^+$ and $v_{A^c}^-$. Set $\hat{\phi}_{A^c}^{(s+1)} = 0$. Check if any of the following four happen.

   (a) $\hat{\phi}_i^{(s+1)}$ equals 0 for some $i \in A$.

   (b) v_j^+ or v_j^- equals 0 for some $j \in A^c$.

   (c) $\gamma_k = 0$ for some $k \in \mathcal{B}$.

   (d) $C_l \hat{\phi}^{(s+1)} = d_l$ for some $k \in \mathcal{B}^c$.

   If any of the above four happens, update A and $\mathcal{B}$.

4. Repeat Step 2 and Step 3 until some convergence criterion is met.

1.8 Proofs

1.8.1 Regularity Conditions

(A1) The link function f has continuous and bounded second order derivative.

(A2) Let $R^*(\phi) = E[Y - f(X^T\theta)]^2$ be the population risk function. Define $H^{*(2)}(\phi) = \frac{\partial^2 R^*(\phi)}{\partial\phi\partial\phi^T}$ as the Hessian matrix of $R^*(\phi)$. $H^{*(2)}(\phi_0)$ is positive definite, and its smallest eigenvalue is $\rho(\min)$, for some $\rho(\min) > 0$.

(A3) The number of interior knots N satisfies $N \sim n^{1/5}$.

1.8.2 Proof of Theorem 1.4.1

We will first show that if $\lambda = O(n^{-1/2})$, there exists a local minimum $\hat{\phi}$ of $R(\phi)$ that is $\sqrt{n}$-consistent. To prove the $\sqrt{n}$-consistency, it is sufficient to show for any γ, there exists a large enough D such that

$$P\left(\sup_{\|\phi-\phi_0\|_2=Dn^{-1/2}} R(\phi) > R(\phi_0)\right) \geq 1-\gamma. \tag{1.35}$$

Notice that

$$\begin{aligned} R(\phi)-R(\phi_0) &= H(\phi)+\lambda\sum_{j=1}^{p-1}\rho_a(|\phi_j|)-H(\phi_0)-\lambda\sum_{j=1}^{p-1}\rho_a(|\phi_{0,j}|) \\ &= \left(H^{(1)}(\phi_0)\right)^T(\phi-\phi_0)+\frac{1}{2}(\phi-\phi_0)^T H^{(2)}(\phi_0)(\phi-\phi_0) \\ &\quad +\lambda\sum_{j=1}^{p-1}[\rho_a(|\phi_j|)-\rho_a(|\phi_{0,j}|)]+o(\|\phi-\phi_0\|_2^2) \\ &\triangleq I_1+I_2+I_3+o(\|\phi-\phi_0\|_2^2) \end{aligned}$$

By Lemma A.15 of [13], we have

$$\left|H^{*(2)}(\phi_0)-H^{(2)}(\phi_0)\right| = o_p(1).$$

This together with (A2) lead to

$$I_2 \geq [\rho(\min)+o(1)]\|\theta_{-1}-\theta_{0,-1}\|_2^2/2 = D^2[\rho(\min)+o(1)]/2n.$$

On the other hand, by Cauchy-Schwarz Inequality, we have

$$|I_3| \leqslant (a+1)\lambda\sqrt{p-1}\|\theta_{-1}-\theta_{0,-1}\|_2/a = (a+1)D\lambda\sqrt{p-1}/a\sqrt{n}.$$

Thus, if $\lambda = O(n^{-1/2})$, then for large enough D, I_2 can dominate I_3. From [13], we have that $H^{(1)}(\phi_0) = O_p(n^{-1/2})$, therefore, $I_1 = D\cdot O_p(n^{-1})$. This means, with large enough D, I_1 is also dominated by I_2. As a result, (1) holds. And it implies that there exists a local minimum of $R(\phi)$ in the ball $\{\phi : \|\phi-\phi_0\|_2 \leqslant Dn^{-1/2}\}$. Therefore, there exists a local minimum $\hat{\phi}$ of $R(\phi)$ that is $\sqrt{n}$-consistent.

With similar arguments given above, we can show that only $\lambda = o(1)$ is needed for the estimation consistency of $\hat{\phi}$. When $\lambda = O(n^{-1/2})$, $\hat{\phi}$ is $\sqrt{n}$-consistent; when $\lambda = O(n^{-1/2+\delta})$ for some $\delta \in (0, 1/2)$, it can be shown $\|\hat{\phi} - \phi_0\|_2 = O(n^{-1/2+\delta})$. The proof for $\hat{\phi}^L$ is very similar to that given above, and is not separately displayed.

1.8.3 Proof of Theorem 1.4.2

By its definition, $\hat{\phi}$ is a local minimum of $R(\phi)$. Define $\mu = \phi - \phi_0$ for any ϕ, and $\hat{\mu} = \hat{\phi} - \phi_0$. Define

$$V(\mu) = H(\phi) - H(\phi_0) + \lambda \sum_{j=1}^{p-1} \rho_a(|\phi_j|).$$

Then $\hat{\mu}$ is a local minimum of $V(\mu)$.

It follows that

$$\begin{aligned} V(\mu) &= H(\phi) - H(\phi_0) + \lambda \sum_{j=1}^{p-1} \rho_a(|\phi_j|) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_\theta(t_\theta^i) \right)^2 - \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}_{\theta_0}(t_{\theta_0}^i) \right)^2 + \lambda \sum_{j=1}^{p-1} \rho_a(|\phi_j|) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_\theta(t_\theta^i) - \hat{f}_{\theta_0}(t_{\theta_0}^i) \right)^2 - \frac{2}{n} \sum_{i=1}^{n} \varepsilon_i \left(\hat{f}_\theta(t_\theta^i) - \hat{f}_{\theta_0}(t_{\theta_0}^i) \right) \\ &\quad + \frac{2}{n} \sum_{i=1}^{n} \left(\hat{f}_\theta(t_\theta^i) - \hat{f}_{\theta_0}(t_{\theta_0}^i) \right) \left(\hat{f}_{\theta_0}(t_{\theta_0}^i) - f(t_{\theta_0}^i) \right) + \lambda \sum_{j=1}^{p-1} \rho_a(|\phi_j|). \end{aligned}$$

Notice that

$$\hat{f}_\theta(t_\theta^i) - \hat{f}_{\theta_0}(t_{\theta_0}^i) = \hat{f}'_{\theta^*}(t_{\theta^*}^i)(\phi - \phi_0) = \hat{f}'_{\theta^*}(t_{\theta^*}^i)\mu,$$

where

$$\hat{f}'_{\theta^*}(t_{\theta^*}^i) = \left(\frac{\partial \hat{f}_{\theta^*}(t_{\theta^*}^i)}{\partial \theta_2}, \frac{\partial \hat{f}_{\theta^*}(t_{\theta^*}^i)}{\partial \theta_3}, \ldots, \frac{\partial \hat{f}_{\theta^*}(t_{\theta^*}^i)}{\partial \theta_p} \right), \forall i$$

and θ^* is between θ and θ_0. Here we slightly abuse the notation, and ignore the fact that θ^* may differ for each x_i.

It follows that,

$$V(\mu) = \frac{1}{n}\mu^T \left(\hat{F}^*\right)^T \hat{F}^*\mu - \frac{2}{n}\varepsilon^T \hat{F}^*\mu + \frac{2}{n}\left(\hat{f}_{\theta_0} - f\right)^T \hat{F}^*\mu + \lambda\sum_{j=1}^{p-1}\rho_a(|\phi_j|),$$

where $\left(\hat{f}_{\theta_0} - f\right)^T = \left(\hat{f}_{\theta_0}(t^1_{\theta_0}) - f(t^1_{\theta_0}), \ldots, \hat{f}_{\theta_0}(t^n_{\theta_0}) - f(t^n_{\theta_0})\right)$, $\varepsilon^T = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$, and for simplicity, $\hat{F}^*$ and $\hat{C}^*$ represent $\hat{F}_{\theta^*}$ and $\hat{C}_{\theta^*}$, repectively.

Let $G(\mu) = \frac{1}{n}\mu^T \left(\hat{F}^*\right)^T \hat{F}^*\mu - \frac{2}{n}\varepsilon^T \hat{F}^*\mu + \frac{2}{n}\left(\hat{f}_{\theta_0} - f\right)^T \hat{F}^*\mu$. With the above notations, we have

$$\begin{aligned}
\frac{\partial G(\mu)}{\partial \mu} &= 2\hat{C}^*\mu - \frac{2}{n}\left(\hat{F}^*\right)^T \varepsilon + \frac{2}{n}\left(\hat{F}^*\right)^T \left(\hat{f}_{\theta_0} - f\right) \\
&= 2\bar{C}_0\mu - \frac{2}{n}\bar{F}_0^T\varepsilon + \frac{2}{n}\left(\hat{F}^*\right)^T \left(\hat{f}_{\theta_0} - f\right) + 2\left(\hat{C}^* - \bar{C}_0\right)\mu - \frac{2}{n}\left(\hat{F}^* - \bar{F}_0\right)^T \varepsilon \\
&\triangleq 2\bar{C}_0\mu - \frac{2}{n}\bar{F}_0^T\varepsilon + T_1 + T_2 + T_3,
\end{aligned}$$

From Theorem 1, we can obtain that, if λ satisfies that $\lambda \sim n^{c-2/5}$ for some $c \in (0, 2/5)$, there exists a local minimum of $R(\phi)$ in the ball $\{\phi : \|\phi - \phi_0\|_2 \leqslant Dn^{c-2/5}\}$. Hence, in this proof, we only focus on μ such that $\|\mu\| = O_p(n^{c-2/5})$.

From [13], we have

$$\begin{aligned}
\sup_{\theta:\|\theta\|_2=1} \max_i \left|\hat{f}_\theta(t^i_\theta) - f_\theta(t^i_\theta)\right| &= O_p\left((nh)^{-1/2}\log n + h^4\right); \\
\sup_{j=2,3,\ldots,p} \sup_{\theta:\|\theta\|_2=1} \max_i \left|\frac{\partial}{\partial\theta_j}\{\hat{f}_\theta(t^i_\theta) - f_\theta(t^i_\theta)\}\right| &= O_p\left((nh^3)^{-1/2}\log n + h^3\right); \\
\sup_{j=2,3,\ldots,p} \sup_{\theta:\|\theta\|_2=1} \max_i \left|\frac{\partial}{\partial\theta_j}\{\hat{f}_\theta(t^i_\theta) - \bar{f}_\theta(t^i_\theta)\}\right| &= O_p\left((nh^3)^{-1/2}\log n\right).
\end{aligned}$$

These along with (A3) lead to $T_1 = O_p\left(n^{-2/5}\log n\right)$ componentwise. On the other hand,

$$\sup_{j=2,3,\ldots,p} \sup_{\theta:\|\theta-\theta_0\|_2=O(n^{c-2/5})} \max_i \left|\frac{\partial}{\partial\theta_j}\{\hat{f}_\theta(t^i_\theta) - \hat{f}_{\theta_0}(t^i_{\theta_0})\}\right| = O_p\left(n^{c-2/5}\right).$$

Then,

$$\begin{aligned}
\left|\frac{\partial}{\partial\theta_j}\hat{f}_{\theta^*}(t^i_{\theta^*}) - \frac{\partial}{\partial\theta_j}\bar{f}_{\theta_0}(t^i_{\theta_0})\right| &\leqslant \left|\frac{\partial}{\partial\theta_j}\{\hat{f}_{\theta^*}(t^i_{\theta^*}) - \hat{f}_{\theta_0}(t^i_{\theta_0})\}\right| + \left|\frac{\partial}{\partial\theta_j}\{\hat{f}_{\theta_0}(t^i_{\theta_0}) - \bar{f}_{\theta_0}(t^i_{\theta_0})\}\right| \\
&= O_p\left(n^{-1/5}\log n + n^{c-2/5}\right).
\end{aligned}$$

Thus, $|\hat{F}^* - \bar{F}_0| = O_p\left(n^{-1/5}\log n + n^{c-2/5}\right)$, and $|\hat{C}^* - \bar{C}_0| = O_p\left(n^{-1/5}\log n + n^{c-2/5}\right)$, entry-wise. It follows that,

$$T_2 = O_p\left((n^{-1/5}\log n + n^{c-2/5}) \times n^{c-2/5}\right),$$

componentwise. Furthermore, by Corollary 8.3 of [46],

$$T_3 = O_p\left((n^{-1/5}\log n + n^{c-2/5}) \times n^{-1/2}\log n\right),$$

componentwise. Then for $\lambda = O(n^{c-2/5})$, with some $c \in (0, 2/5)$, λ can dominate T_1, T_2, and T_3.

Let $\bar{W} = \bar{F}_0^T\varepsilon/\sqrt{n}$. We decompose μ and $\bar{W}$ into two sub-vectors that are formed by $\mathcal{A}_1$ and $\mathcal{A}_2$, that is, $\bar{W}^T = (\bar{W}(1), \bar{W}(2))^T$, and $\mu^T = (\phi_1 - \phi_{1,0}, \ldots, \phi_{q-1} - \phi_{0,q-1}, \phi_q, \ldots, \phi_{p-1}) = (\mu(1), \mu(2))^T$.

By the KTT conditions, if there exists $\hat{\mu}$ that satisfies the following

$$2\bar{C}_0(11)\hat{\mu}(1) - \frac{2}{\sqrt{n}}\bar{W}(1) + \lambda\rho_a'(|\hat{\phi}(1)|) \times \text{sign}(\hat{\phi}(1)) = 0; \tag{1.36}$$

$$|\hat{\mu}(1)| < |\phi_0(1)|; \tag{1.37}$$

$$-\lambda\rho_a'(0+)\mathbb{1}_{p-q} \leqslant 2\bar{C}_0(21)\hat{\mu}(1) - \frac{2}{\sqrt{n}}\bar{W}(2) \leqslant \lambda\rho_a'(0+)\mathbb{1}_{p-q}, \tag{1.38}$$

there exists a local minimum of $R(\phi)$, $\hat{\phi}$, such that $\text{sign}(\hat{\phi}(1)) = \text{sign}(\phi_0(1))$ and $\text{sign}(\hat{\phi}(2)) = 0$. Here, $\times$ denotes component-wise multiplication. After some simplification, we have that the existence of $\hat{\mu}$ is implied by

$$\begin{aligned} |\bar{C}_0^{-1}(11)\bar{W}(1)| &< \sqrt{n}\left(|\phi_0(1)| - \frac{\lambda}{2}|\bar{C}_0^{-1}(11)\rho_a'(|\hat{\phi}(1)|) \times \text{sign}(\hat{\phi}(1))|\right) \\ |\bar{C}_0(21)\bar{C}_0^{-1}(11)\bar{W}(1) - \bar{W}(2)| &\leqslant \frac{\sqrt{n}\lambda}{2}\left(\rho_a'(0+)\mathbb{1}_{p-q} - |\bar{C}_0(21)\bar{C}_0^{-1}(11)\rho_a'(|\hat{\phi}(1)|) \times \text{sign}(\hat{\phi}(1))|\right). \end{aligned}$$

By Theorem 1.4.1, we know that there exists $\bar{L}_3$ such that $\|\hat{\mu}(1)\| \leqslant \lambda\bar{L}_3$. Then,

$$|\hat{\phi}(1)| \geqslant (b_0 - \lambda\bar{L}_3)\mathbb{1}_{p-q}.$$

Since ρ' is monotonically decreasing, we have

$$\|\rho'(|\hat{\phi}(1)|) \times \text{sign}(\hat{\phi}(1))\|_\infty \leqslant \rho'(b_0 - \lambda\bar{L}_3).$$

Subsequently, we can obtain

$$\begin{aligned}\|\bar{C}_0^{-1}(11)\rho_a'(|\hat{\phi}(1)|)\times\text{sign}(\hat{\phi}(1))\|_\infty &\leqslant \|\bar{C}_0^{-1}(11)\|_\infty\rho'(b_0-\lambda\bar{L}_3)\leqslant L_1\rho'(b_0-\lambda\bar{L}_3);\\ \|\bar{C}_0(21)\bar{C}_0^{-1}(11)\rho_a'(|\hat{\phi}(1)|)\times\text{sign}(\hat{\phi}(1))\|_\infty &\leqslant \|\bar{C}_0(21)\bar{C}_0^{-1}(11)\|_\infty\rho'(b_0-\lambda\bar{L}_3).\end{aligned}$$

Let

$$\begin{aligned}A &= \{|\bar{C}_0^{-1}(11)\bar{W}(1)|<\sqrt{n}(|\phi_0(1)|-\frac{\lambda}{2}\bar{L}_1\rho'(b_0-\lambda\bar{L}_3))\},\\ B &= \{|\bar{C}_0(21)\bar{C}_0^{-1}(11)\bar{W}(1)-\bar{W}(2)|\leqslant\frac{\sqrt{n}\lambda}{2}\bar{L}_4\},\end{aligned}$$

for some $\bar{L}_4>0$. Then we have, if the Irrepresentable Conditions for BS-SIM hold,

$$P\left(\text{sign}(\hat{\phi})=\text{sign}(\phi_0)\right)\geq P(A\cap B),$$

whereas

$$1-P(A\cap B)\leqslant\sum_{i=1}^{q-1}P\left(|z_i|\geq\sqrt{n}(|\phi_{0,i}|-\frac{\lambda}{2}\bar{L}_1\rho'(b_0-\lambda\bar{L}_3))\right)+\sum_{i=1}^{p-q}P\left(|q_i|\geq\frac{\sqrt{n}\lambda}{2}\bar{L}_4\right),$$

where $z=(z_1,z_2,\ldots,z_{q-1})^T=\bar{C}_0^{-1}(11)\bar{W}(1)$ and $q=(q_1,q_2,\ldots,q_{p-q})^T=\bar{C}_0(21)\bar{C}_0^{-1}(11)\bar{W}(1)-\bar{W}(2)$.

By the definition of $\bar{W}$, we have

$$\begin{aligned}\bar{C}_0^{-1}(11)\bar{W}(1) &\to_d N(0,\sigma^2\bar{C}_0^{-1}(11));\\ \bar{C}_0(21)\bar{C}_0^{-1}(11)\bar{W}(1)-\bar{W}(2) &\to_d N(0,\sigma^2(\bar{C}_0(22)-\bar{C}_0(21)\bar{C}_0^{-1}(11)\bar{C}_0(12))).\end{aligned}$$

Since λ satisfies that $\lambda\sim n^{c-2/5}$, we have

$$\sum_{i=1}^{q-1}P\left(|z_i|\geq\sqrt{n}(|\phi_{0,i}|-\frac{\lambda}{2}L_1\rho'(b_0-\lambda\bar{L}_2))\right)+\sum_{i=1}^{p-q}P\left(|q_i|\geq\frac{\sqrt{n}\lambda}{2}\bar{L}_4\right)=o(e^{-n^c}).$$

As a result, Theorem 1.4.2 follows.

1.8.4 Proof of Theorem 1.4.3

Let $\hat{\mu}^L=\hat{\phi}^L-\phi_0$. Define

$$V_L(\mu)=H(\phi)-H(\phi_0)+\lambda\|\mu+\phi\|_1.$$

Then $\hat{\mu}^L$ is a local minimum of $V_L(\mu)$.

With the arguments used in the proof of Theorem 1.4.2, It follows that

$$\begin{aligned} V_L(\mu) &= H(\phi) - H(\phi_0) + \lambda\|\mu + \phi_0\|_1 \\ &= \frac{1}{n}\mu^T\left(\hat{F}^*\right)^T\hat{F}^*\mu - \frac{2}{n}\varepsilon^T\hat{F}^*\mu + \frac{2}{n}\left(\hat{f}_{\theta_0} - f_{\theta_0}\right)^T\hat{F}^*\mu + \lambda\|\mu + \phi_0\|_1 \\ &= G(\mu) + \lambda\|\mu + \phi_0\|_1. \end{aligned}$$

By the argument given in the proof of Theorem 1.4.2, if there exists $\hat{\mu}^L$ that satisfies the following

$$\begin{gathered} 2\bar{C}_0(11)\hat{\mu}^L(1) - \frac{2}{\sqrt{n}}\bar{W}(1) + \lambda\text{sign}(\phi_0(1)) = 0; \\ |\hat{\mu}^L(1)| < |\phi_0(1)|; \\ -\lambda\mathbb{1}_{p-q} \leqslant 2\bar{C}_0(21)\hat{\mu}^L(1) - \frac{2}{\sqrt{n}}\bar{W}(2) \leqslant \lambda\mathbb{1}_{p-q}, \end{gathered}$$

there exists a local minimum of $R_L(\phi)$, $\hat{\phi}^L$, such that $\text{sign}(\hat{\phi}^L(1)) = \text{sign}(\phi_0(1))$ and $\text{sign}(\hat{\phi}^L(2)) = 0$. After some simplification, we have that the existence of $\hat{\mu}^L$ is implied by

$$\begin{aligned} |\bar{C}_0^{-1}(11)\bar{W}(1)| &< \sqrt{n}\left(|\phi_0(1)| - \frac{\lambda}{2}|\bar{C}_0^{-1}(11)\text{sign}(\phi_0(1))|\right); \\ |\bar{C}_0(21)C_0^{-1}(11)\bar{W}(1) - \bar{W}(2)| &\leqslant \frac{\sqrt{n}\lambda}{2}\left(\mathbb{1} - |\bar{C}_0(21)\bar{C}_0^{-1}(11)\text{sign}(\phi_0(1))|\right). \end{aligned}$$

Let

$$\begin{aligned} A_1 &= \{|\bar{C}_0^{-1}(11)\bar{W}(1)| < \sqrt{n}(|\phi_0(1)| - \frac{\lambda}{2}|\bar{C}_0^{-1}(11)\text{sign}(\phi_0(1))|)\}, \\ B_1 &= \{|\bar{C}_0(21)\bar{C}_0^{-1}(11)\bar{W}(1) - \bar{W}(2)| \leqslant \frac{\sqrt{n}\lambda}{2}\bar{\eta}\}. \end{aligned}$$

Subsequently, we have, if the Irrepresentable Condition for BL-SIM holds,

$$P\left(\text{sign}(\hat{\phi}^L) = \text{sign}(\phi_0)\right) \geq P(A_1 \cap B_1),$$

whereas

$$1 - P(A_1 \cap B_1) \leqslant \sum_{i=1}^{q-1} P\left(|z_i| \geq \sqrt{n}(|\phi_0(1)| - \frac{\lambda}{2}|b_i|)\right) + \sum_{i=1}^{p-q} P\left(|q_i| \geq \frac{\sqrt{n}\lambda}{2}\bar{\eta}\right),$$

where $b = (b_1, b_2, \ldots, b_{q-1})^T = \bar{C}_0^{-1}(11)\text{sign}(\phi_0(1))$.

Since λ satisfies that $\lambda \sim n^{c-2/5}$, we have

$$\sum_{i=1}^{q-1} P\left(|z_i| \geq \sqrt{n}(|\phi_0(1)| - \frac{\lambda}{2}|b_i|)\right) + \sum_{i=1}^{p-q} P\left(|q_i| \geq \frac{\sqrt{n}\lambda}{2}\bar{\eta}\right) = o(e^{-n^c}).$$

As a result, Theorem 1.4.3 follows.

1.8.5 Proof of Corollary 1.4.4

The proof of Corollary 1.4.4 can be derived by using similar arguments applied in the proofs of Theorem 1.4.2 and Theorem 1.4.3.

2. DNA METHYLATION STATUS QUANTIFICATION FOR BISULPHITE-SEQUENCING DATA

In this chapter, we focus on DNA methylation status calling for bisulphite-sequencing data. The first section reviews several key concepts for DNA methylation profiling and quantification. Section 2.2 describes the proposed approaches. Section 2.3 and Section 2.4 demonstrate its performance on simulated datasets and a real data, respectively. The contents in Sections 2.2 through 2.4 are based on our work published in 2012 [47]. The last section of this chapter summarizes the recent development in DNA methylation analysis and false discovery rate control for discrete tests since our work was published.

2.1 Introduction

2.1.1 Introduction to DNA Methylation

Epigenetic modification is defined as heritable changes in chromosome without altering the DNA sequence [48]. There are three components in epigenetic modifications: DNA methylation, histone modification and non-coding RNAs [49]. DNA methylation refers to the addition of a methyl group to some C-5 positions of DNA sequences. It plays a crucial role in a variety of biological processes, including cell development, imprinting and X-chromosome inactivation. It is prevalent at CpG positions with 60-90% of all CpGs being methylated in mammals, whereas it is much less frequent at non-CpG sites with only less than 3% of non-CpGs being found to be methylated. Unmethylated CpGs tend to cluster in small regions of DNA sequences called CpG islands, most of which coincide with promoter regions of many genes [50]. The link between abnormal DNA methylation pattern and cancer is two-fold [51]. First, a

global hypomethylation is associated with genomic instability and is a common characteristic of cancer cells. Second, hypermethylation of CpG islands located at gene promoters results in suppression of gene expression and is conventionally observed in cancer cells. Therefore, it is desirable to reveal both genome-wide and promoter-specific DNA methylation patterns of a cell.

Various methods for genome-wide DNA methylation detection have been developed in the past 20 years. They can mainly be classified into three categories, which are methylation-sensitive enzyme based methods, enrichment based methods, and bisulphite conversion based methods [50]. A quick review of them are given below.

Methylation-sensitive Enzyme Based Methods

The basis of this type of methods is that genomic DNAs are fragmented by a methylation-sensitive restriction enzyme differentially, according to the DNA methylation status. To be more specific, a methylation-sensitive restriction enzyme is used to cut CpG sites that are unmethylated, while it is blocked by methylated sites. As a result, the DNA methylation profile can be inferred based on the cutting pattern, provided by the subsequent read-out approaches. There are several choices for the restriction enzymes, such as *Hpa*II and *Sma*I. One major drawback of this class of methods is that each restriction enzyme has a specific recognition sequence. That means, each enzyme can only cut or be block by CpG sites in one specific sequence. For example, *Hpa*II can only work with CpG sites in CCGG sequences. Thus, the DNA methylation profile of CpG sites in other contexts can not be detected by using *Hpa*II. This restriction makes this type of methods unfavorably when a whole-genome profile of DNA methylation is desired.

Enrichment Based Methods

This type of methods relies on immunoprecipitating DNA fragments containing methylated CpG with specific antibodies or methyl-binding proteins to differentiate methy-

lated sites from unmethylated sites. This is a more direct method to profile DNA methylation status, compared to the enzyme based methods. The advantage of this type is that it is of low cost and can provide whole-genome coverage. However, it has its own limitations. Since the antibodies or proteins would pull down the DNA fragments containing methylated sites, the exact sites that are methylated can be any CpG sites within the fragment. Thus, for this type of method, the DNA methylation status is often called as a whole fragment, rather than on a site-by-site basis. In other words, the enrichment based methods can only provide a moderate resolution.

Bisulphite Conversion Based Methods

Sodium bisulphite can convert unmethylated Cytosine into Urasil whereus methylated Cytosines are not affected by it. Since bisulphite-induced Urasil is subsequently replaced by Thymine during amplification, the sodium bisulphite treatment leads to a methylation specific single nucleotide polymorphism (SNP) at unmethylated sites, and thus methylated sites can be distinguished from unmethylated sites by some subsequent readout method. This finding revolutionized the way how DNA methylation status was profiled in 1990s [50, 52, 53], since this property of sodium bisulphite suggests that this collection of methods may be able to provide DNA methylation profiles at single base resolution with whole genome coverage.

After applying one of the above methods on the genomic DNAs, one read-out method is needed. In the past, array-based techniques, such as microarray technology, were the leading platforms to be combined with methods from all three above categories to survey DNA methylation status. To name one array-based methylation profiling method from each category, methylated CpG island amplification with array hybridization (MCAM) uses a pair of enzymes *Sma*I and *Xma*I with array hybridization [54]; MeDIP relies on a methylation specific antibody coupled with hybridization [55]; Illumina Infinium assay for DNA methylation analysis, or HumanMethy-

lation27 DNA Analysis BeadChip, is a bisulphite conversion based approach [56]. Although the application of these array-based platforms enables comprehensive DNA methylation profiling at economical cost, they can only interrogate C sites at given regions with moderate resolution [50].

In recent years, the rapid development of Next Generation Sequencing (NGS) technology enables read-out methods with high coverage and high resolution [57]. NGS has been incorporated into all three categories of methods for genome-wide methylation profiling. Again, we list one NGS based method from each category here: *Hpa*II tiny fragment enrichment by ligation-mediated PCR (HELP), or HELP-Seq [58]; MeDIP-Seq [59]; and MethylC-Seq [60]. Despite the fact that NGS-equiped methods have relative advantages over array-based methods, those methods from the first two categories are still subject to the same weaknesses they had when coupled with array-based techniques. More specifically, methylation-sensitive enzyme based methods equipped with NGS technology remain restricted to the recognition sites of the particular enzymes used; and enrichment based methods equipped with NGS technology do not overcome the disadvantage of moderate resolution. On the other hand, bisulphite conversion based methods coupled with NGS technology, designated as bisulphite-sequencing methods, have emerged as the most promising methods since they generate whole-genome DNA methylation profiles at single-base resolution. Among all bisulphite-sequencing methods, MethylC-Seq and reduced representation bisulfite sequencing (RRBS) are the two most popularly used methods, which we will introduce in the next subsection [60, 61].

2.1.2 Review of Bisulphite-sequencing Experiment

In MethylC-Seq, genomic DNAs are first sonicated into smaller fragments. After going through end-repair and adapter ligation, these fragments are treated with bisulphite. As introduced in the previous subsecion, the bisulphite treatment converts unmethylated Cytosines into Urasils and leaves methylated Cytosines unchanged. Subsequent

PCR amplification process further replaces Urasils with Thymines. These PCR amplified fragments are then subject to standard sequencing technology to produce short sequencing reads, which are mapped back to the reference genome. Thus the unmethylated Cytosines are distinguishable from methylated Cytosines by examining sequencing reads [50,60]. The workflow for MethylC-Seq experiment is given in Figure 2.1.

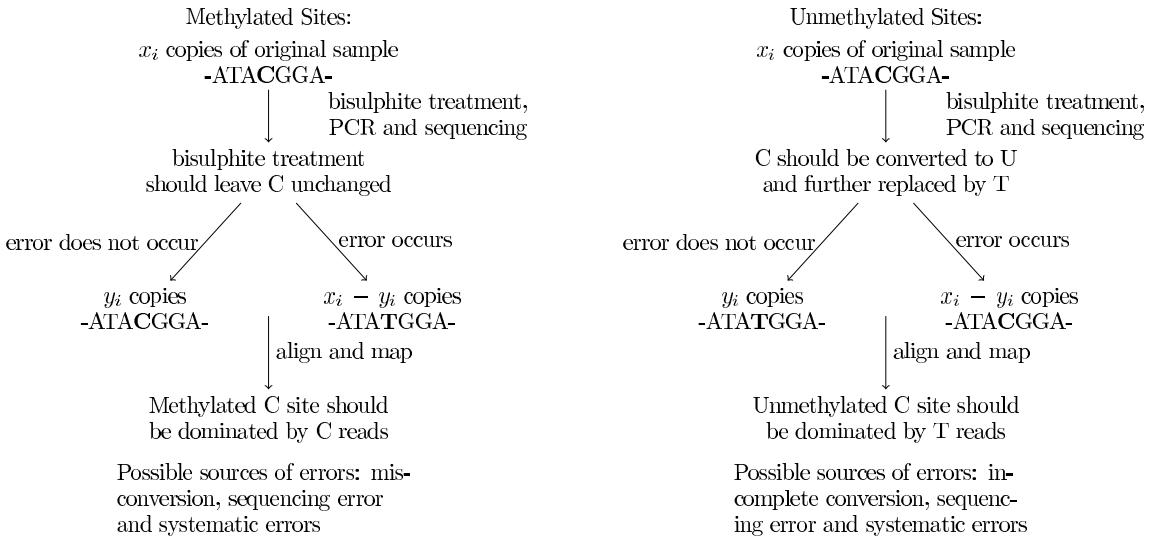


Fig. 2.1. Workflow for MethylC-Seq experiment.

RRBS utilizes the same mechanism as MethylC-Seq. The major difference between RRBS and MethylC-Seq occurs in the first step, that is, the way genomic DNAs are fragmented. In RRBS, genomic DNAs are digested with MspI, an enzyme which cuts all CCGG sites [61]. These two methods have their relative advantages and disadvantages, which make them suitable for different research purposes. By the way genomic DNAs are digested in RRBS, CpG regions are substantially enriched in DNA fragments after size selection. Thus RRBS is more preferable than MethylC-Seq when the research is targeting regions with high density of CpG sites, such as CpG islands. On the other hand, because of its theoretical capacity of capturing methylation information from each C position in the whole genome, MethylC-Seq has

become the golden standard for genome-wide DNA methylation analysis. As reported in Harris *et al.* [62], when these two methods are applied to biological replicates of human embryonic stem cells, MethylC-Seq covers 95% of all CpGs, whereas RRBS shows a genome-wide CpG coverage of only 12%.

2.1.3 Review of Quantification Methods for Bisulphite-sequencing Data

In the data generated by MethylC-Seq or RRBS, ideally there are only C reads or T reads for each covered C position of interest, depending on the methylation status. In other words, if a C position is methylated, then there should be only C reads at that site in the data; whereas if a C position is unmethylated, then there should be only T reads. However, due to various sources of noise, in the real data generated by these two methods, there are both C reads and T reads for most of the target C sites. For instance, the process of bisulphite conversion needs to be carried out under specific experimental conditions [63]. Failure to meet any of those conditions would lead to incomplete conversion, which further results in C reads at unmethylated C positions. Moreover, as a typical and inevitable result of applying NGS technology, there will be sequencing errors in the data, which means a small proportion of C reads will be miscalled to be T reads and vice versa. Because there are both C reads and T reads in the data, it is not straightforward to infer the true methylation status. The aim is then to make methylation call for each target C position based on the number of C reads and the number of T reads it receives, which becomes an interesting statistical problem.

In some studies concerning DNA methylation analysis, researchers use the ratio of C count to the total number of reads received at a site to quantify the methylation level at that site ([62]; [64]; [65]). Note that the total number of reads received at a site is also referred to as coverage or sequencing depth. While this quantification approach has the virtue of being simple and straightforward, it does not use proper inference to deal with the noise in the data, and thus it is not statistically satisfactory.

In other studies, researchers aim to make binary methylation calls for C positions of interest. There exist a few such approaches in the literature. In Harris *et al.* [62], the proportion of C count at each CpG site was calculated and binary methylation call was made for each site with various choices of cutoff for the proportion. However, those choices were not statistically justified. A more sophisticated method was used in Lister *et al.* [66], which applied a multiple testing procedure to identify methylated Cytosines. In the MethylC-Seq experiment conducted by Lister *et al.* [66], an unmethylated Lambda DNA was spiked with the target genomic DNAs before sonication and was used to estimate the error rate at which a C count occurs at an unmethylated C position. We denote the resulting estimate as $\hat{p}_1^{Lis}$. Then the following hypothesis was tested for each C site in the whole genome simultaneously to detect methylated sites with false discovery rate level 0.01.

$$H_{i0} : p = \hat{p}_1^{Lis} \quad vs \quad H_{ia} : p > \hat{p}_1^{Lis}.$$

As will be shown in Section 2.3.3, the above procedure used by [66] is conservative in detecting unmethylated Cytosines due to the underestimation of error rate and the choice of null hypothesis.

2.1.4 Review of False Discovery Rate Controlling Procedures

Due to the rapid development of sequencing technologies, it is often the case that a large number of hypothesis are tested simultaneously. For instance, in a RNA-seq experiment, we may want to test if thousands of genes are differentially expressed at the same time. False discovery rate (FDR), originally proposed by Benjamini and Hochberg [67], is the most popularly used error measurement in multiple testing problems. Its prevalence owes largely to the fact that it is considerably less conservative than the conventional family-wise error rates. To describe FDR, consider the possible outcomes from multiple testing procedures given in Table 2.1. Benjamini and Hochberg [67] gave the following definition of FDR:

$$FDR = E\{\frac{M_{01}}{V}|V > 0\}P(V > 0).$$

In their seminal paper [67], they also proposed a linear step-up approach for controling FDR. They also proved that this step-up procedure can control FDR at a prespecified level for independent test statistics. However, since the BH approach does not consider the unknown proportion of true nulls, it usually leads to conservative results. There are various procedures proposed later that consider this information. For example, Storey [68] proposed to estimate the proportion of true nulls as

$$\text{proportion} = \frac{\#\{\text{p-values} > \lambda\}}{(1-\lambda)M},$$

for some well-selected λ. They also discussed how to choose λ. For more detail, we refer to [68–70]. Since this procedure incorporates the estimation of the proportion of true nulls, it is without surprise that this procedure is more powerful than the procedure proposed by [67].

By the nature of hypothesis testing, the p-values from multiple testing problems can be viewed to follow a mixture of two distributions. Then its cumulative distribution function (CDF) can be written down as

$$F = \pi F_0 + (1-\pi)F_1,$$

where F_0 and F_1 denote the CDF under the null hypothesis and the alternative hypothesis respectively, and π stands for the unknown proportion of true nulls. It is well known that under several assumptions, F_0 is the standard uniform distribution. Then FDR for a given cutoff t can be calculated as

$$FDR(t) = \frac{\pi F_0(t)}{\pi F_0(t) + (1-\pi)F_1(t)} = \frac{\pi t}{\pi t + (1-\pi)F_1(t)}.$$

Efron *et al.* [71] proposed a related concept, local FDR (lfdr), that relies on the density function instead of CDF and is given by

$$lfdr(t) = \frac{\pi}{\pi + (1-\pi)f_1(t)}.$$

Efron *et al.* [71] also showed the linkage between lfdr and FDR. This mixture of two components setup leads to another category of methods for estimating and controling

	Fails to reject H_0	Reject H_0	Total
Null is true	M_{00}	M_{01}	M_0
Alternative is true	M_{10}	M_{11}	M_1
Total	U	V	M

Table 2.1.
Four possible outcomes from multiple testing procedures.

FDR or lfdr. They focus on modeling the distribution of the p-values under the alternative hypothesis; see [71–74] among others.

Most of the procedures mentioned so far are intended for independent and continuous statistics. Although they can also be applied to discrete p-values, they are usually conservative in such scenarios. To overcome the conservativeness induced by discreteness, there are also several FDR controlling procedures developed for discrete tests. Tarone [75] proposed a modified Bonferroni method to control familywise error rate for discrete data. Gilbert [76] combined this method with the original FDR controlling procedure [67] to account for the discreteness in the data. Pounds and Cheng [77] developed two estimators for the true null proportion, as well as a smoothing method to stablize the resulting lfdr estimator, for discrete tests. Heyse [78] proposed to replace p-values in [67] with mid p-values to mitigate the effect of discreteness. All of the mentioned methods have been shown to provide certain improvements in power over the methods that do not take discreteness into consideration.

2.2 Methods

2.2.1 Mixture of Binomial Model

As discussed in Section 2.1.2, MethylC-Seq experiment can roughly cover 95% of all CpGs. Those sites that do not receive any C read and T read are referred to as

uncovered sites and will be excluded from methylation calling analysis. Suppose we consider M covered sites. These M sites can be the collection of all covered sites from a specific DNA segment of interest, a whole chromosome, or even the whole genome. For site i among these M sites, let X_i denote the total number of reads including both C and T reads, and Y_i denote the number of C reads alone. Note $Y_i \leqslant X_i$. Let S_i be the indicator of the unobserved methylation status of site i, with $S_i = 0$ indicating site i is methylated and $S_i = 1$ indicating site i is unmethylated. If there is no error in the experiment, then $X_i = Y_i$ when site i is methylated; and $Y_i = 0$ when site i is unmethylated. In other words, there are only C reads for methylated sites and no C read for unmethylated sites. However, MethylC-Seq experiments are subject to both experimental errors and systematic errors. Thus, there are both C reads and T reads for most sites, or equivalently, $Y_i < X_i$ for most methylated sites and $Y_i > 0$ for most unmethylated sites.

There exist three main causes for experimental errors in MethylC-Seq experiments. First, incomplete conversion of unmethylated Cytosine to Urasil during bisulphite treatment results in C reads at unmethylated sites. In other words, the failure to convert unmethylated Cytosine to Urasil causes $Y_i > 0$ for unmethylated sites. We assume this non-conversion rate is e_{ic}, that is, the probability that an unmethylated Cytosine fails to convert to Thymine. Second, over-treatment with bisulphite can lead to conversion of methylated Cytosine to Thymine [50]. Suppose the miscoversion rate, or equivalently, the probability that a methylated Cytosine converts to Thymine is e_{mc}. Third, sequencing errors can potentially impact both methylated sites and unmethylated sites. For methylated sites, Cytosines can be miscalled to be Thymines and thus $Y_i < X_i$; and for unmethylated sites, bisulphite-converted Thymines can be mistakenly read out as Cytosines and thus $Y_i > 0$. Suppose the probability that a T read is miscalled to be a C read is e_{tc} and the probability that a C read is miscalled to be a T read is e_{ct}. Experimental errors are unavoidable due to the random nature of sequencing technology and has to be incorporated in the model. On the other hand, systematic errors in bisulphite data can be identified and thus eliminated

by carefully conducted data processing procedures. For MethylC-Seq experiment, deamination of methylated Cytosine to Thymine during cell development and those single-nucleotide polymorphisms that a Cytosine in the reference genome varies to a Thymine in the sample DNA lead to systematic errors. Nevertheless, they can be detected by examining the nucleotide on the opposite strand of the C sites and thus can be eliminated from MethylC-Seq data [50]. When they are not removed from the data, let e_{sys} denote the systematic error rate. For a more detailed review of potential sources of noise in bisulphite-sequencing data, see Krueger *et al.* [79].

Let p_1 stand for the overall error rate for obtaining C reads at unmethylated sites caused by incomplete conversion, sequencing error, and systematic errors. Similarly, let $1-p_0$ denote the overall error rate for obtaining T reads at methylated sites caused by misconversion, sequencing error, and systematic errors. It is clear that p_1 depends on e_{ic}, e_{tc} and e_{sys}, and p_0 depends on e_{mc}, e_{ct} and e_{sys}. The dependence of p_0 and p_1 on the various types of individual errors can be greatly simplified if the following three assumptions are imposed. First, there are no systematic errors in the data, that is, $e_{sys} = 0$. Second, the two types of sequencing errors occur equally likely, which implies $e_{tc} = e_{ct}$. Third, the sample is not overtreated with bisulphite, or equivalently, $e_{mc} = 0$. Under these three assumptions, we postulate the relationship between the overall error rates and individual ones to be $p_0 = 1 - e_{tc} = 1 - e_{ct}$ and $p_1 = e_{ic} + e_{tc}$. Under the postulated relationship, if we can identify the overall error rates $1-p_0$ and p_1, the individual error rates e_{tc}, e_{ct} and e_{ic} can also be identified. When any of the three assumptions mentioned above fails to satisfy, further information is needed to identify the various types of individual errors. Nevertheless, the overall error rates $1 - p_0$ and p_1 can still be estimated and methylation calling can be made by the procedure we will describe next. Due to this reason, we shall use p_0 and p_1 in the rest of the paper.

Based on the discussion above, we propose the following Binomial models as the conditional distribution of the C count at site i given the coverage X_i and methylation status S_i:

$$Y_i|(X_i = x, S_i = 0) \sim Bin(x, p_0);$$
$$Y_i|(X_i = x, S_i = 1) \sim Bin(x, p_1).$$

Here one important premise is that all the M sites of interest share the same error rates $1-p_0$ and p_1. This assumption is commonly used in the literature on methylation analysis; see Lister *et al.* [66] and Wu *et al.* [80].

Furthermore, suppose the proportion of methylated sites among these M sites is π, that is, $P(S_i = 0) = \pi$ for any randomly selected site i. Then conditional on the sequencing depth at one site, the corresponding C count follows a mixture of two Binomial distributions:

$$Y_i|(X_i = x) \sim \pi Bin(x, p_0) + (1 - \pi)Bin(x, p_1). \tag{2.1}$$

Even though MethylC-Seq data contain diverse types of errors, they are still assumed to carry information regarding the underlying methylation status in the sense that most methylated sites are dominated by C reads and most unmethylated sites are dominated by T reads. Therefore, it is reasonable to assume that p_1 and p_0 should satisfy $p_1 << p_0$. This assumption assures the identifiability of p_1 and p_0 and guarantees the validity of our procedure.

Suppose the coverages at the sites, i.e. X_i's, are independent and identically distributed with the same probability mass function (pmf) $f(x)$. For convenience, denote the pmf of the conditional distribution of C count of site i given $X_i = x$ defined in (1) as $g(y|x)$. For fixed i, the pmf of the joint distribution of (X_i, Y_i), denoted as $h(x, y)$, is given by $h(x, y) = g(y|x)f(x)$. Let $\phi = (p_0, p_1, \pi)$. Noticing that $f(x)$ does not involve ϕ, therefore we only need to use $g(y|x)$ for estimating ϕ.

Let $\mathbf{y} = (y_1, y_2, ..., y_M)$ be the observed C counts and $\mathbf{x} = (x_1, x_2, ..., x_M)$ the observed coverages. Then under the assumption that y_i is from a mixture of two

Binomial distributions given x_i, the log-likelihood function of ϕ can be written as follows.

$$l(\phi|\mathbf{x},\mathbf{y}) = \sum_{i=1}^{M} ln\left\{g(y_i|x_i)\right\} = \sum_{i=1}^{M} ln\left\{\pi g_{i0} + (1-\pi)g_{i1}\right\},$$

where g_{i0} and g_{i1} are the pmf's of $Bin(x_i, p_0)$ and $Bin(x_i, p_1)$ for each i, respectively. The maximum likelihood estimate (MLE) of ϕ can be obtained by applying the well-established EM algorithm. However, our goal here is beyond estimating ϕ. What we want to achieve is to classify each site i to be either methylated or unmethylated on the basis of an adequate estimate of ϕ. Recall that for each i, S_i is an indicator of the true methylation status of site i with values equal to 0 or 1. Therefore, our goal is essentially to identify the value of S_i for each i.

Let θ_{i0} and θ_{i1} denote the posterior probabilities that $S_i = 0$ and $S_i = 1$, respectively, given $\mathbf{x}$, $\mathbf{y}$ and ϕ. The expressions of θ_{i0} and θ_{i1} are

$$\begin{aligned} \theta_{i0} &= P(S_i = 0|y_i, x_i, \phi) = \frac{\pi g_{i0}(y_i)}{\pi g_{i0}(y_i) + (1-\pi)g_{i1}(y_i)}; \\ \theta_{i1} &= P(S_i = 1|y_i, x_i, \phi) = 1 - \theta_{i0}. \end{aligned} \tag{2.2}$$

Here θ_{i0} and θ_{i1} indicate how likely site i is methylated ($S_i = 0$) and unmethylated ($S_i = 1$), respectively, given the observed data and ϕ. Note that $\theta_{ir}(r = 0, 1)$ will play a role in the EM algorithm for computing the MLE of ϕ. In addition to facilitating the estimation of ϕ, θ_{ir} also play a key role in the methylation status calling procedure we will develop later.

Then the EM algorithm for computing the MLE of ϕ can be developed as follows. We start off with an initial estimate of ϕ, and then compute the initial values of θ_{ir} given the initial values of ϕ. After the initial step, ϕ and θ_{ir} are iteratively updated. Conditional on the current values of θ_{ir}, we update ϕ by

$$\hat{p}_0 = \frac{\sum_{i=1}^{M} \hat{\theta}_{i0} y_i}{\sum_{i=1}^{M} \hat{\theta}_{i0} x_i}; \quad \hat{p}_1 = \frac{\sum_{i=1}^{M} \hat{\theta}_{i1} y_i}{\sum_{i=1}^{M} \hat{\theta}_{i1} x_i}; \quad \hat{\pi} = \frac{\sum_{i=1}^{M} \hat{\theta}_{i0}}{M}. \tag{2.3}$$

This new estimate of ϕ is then substituted back into (2.2) to yield new values of θ_{ir}. These two steps are repeated until certain convergence criterion is met. In our simulation study and real data application, the convergence criterion is that the

change of the log-likelihood function between two consecutive steps is below some prespecified value. A discussion of the convergence properties of EM algorithms can be found in [81]. The derivation of (2.3) is given in Section 2.2.5. Let $\hat{\phi}$ be the MLE of ϕ obtained from the EM algorithm and $\hat{\theta}_{ir}$ the estimate of θ_{ir} by plugging $\hat{\phi}$ into (2.2). We shall call $\hat{\theta}_{ir}$ memberships hereafter.

2.2.2 Classification based Methylation Status Calling Procedure

After the EM algorithm in the last subsection converges, we obtain the estimates $\hat{\phi}$ as well as the memberships $\hat{\theta}_{ir}$. The memberships can be further used to determine the methylation status of each site. We propose to use the following rule to make methylation status calling. *For $i = 1, 2, ..., M$, site i is called to be methylated if $\hat{\theta}_{i0} > \hat{\theta}_{i1}$; otherwise, it is called to be methylated.* We shall refer to this classification procedure as the Methylation Status Calling (MSC) procedure. Based on the Bayes rule, the MSC procedure is optimal in terms of maximizing overall correct allocation rate [82].

As mentioned in Section 2.1.3, sometimes researchers are interested in quantifying the methylation levels due to the heterogeneity of cell types or contamination during cell preparation. When the experiment is conducted on a mixture of different types of cells, it is valuable to directly use the membership $\hat{\theta}_{i0}$ to quantify the methylation level of each C position. Note that in this case, the interpretation of the overall error rates $1 - p_0$ and p_1 is slightly different. More specifically, not only do they stand for the various types of noises caused by the bisulphite-sequencing experiment, they also represent the extent of cell type contamination. Since we are using a MethylC-Seq data of H9 human embryonic stem cells in our real data application, we will focus on the binary methylation status calling in our paper.

Table 2.2.
Possible outcomes from the MSC procedure and the FMSC procedure.

	Classified as methylated (Fails to reject H_0)	Classified as unmethylated (Reject H_0)	Total
Group 0	M_{00}	M_{01}	M_0
Group 1	M_{10}	M_{11}	M_1
Total	U	V	M

2.2.3 Performance Assessment of the MSC Procedure

We use individual and overall correct allocation rates to assess the performance of our proposed MSC procedure. Let Group 0 and Group 1 consist of all methylated sites and all unmethylated sites, respectively. Let M_0 and M_1 be the total number of methylated and unmethylated sites in the sample, respectively. Let M_{ij} be the number of sites that are from Group i and allocated to Group j by the MSC procedure, for $i = 0, 1$, and $j = 0, 1$. Let the total number of sites that are classified to Group 0 be U and let the total number of sites that are classified to Group 1 be V. The four possible outcomes from the proposed MSC procedure are listed in Table 2.2 with their corresponding frequencies.

The correct allocation rate for methylated sites (i.e., Group 0), denoted as P_0, is defined as the proportion of sites that are methylated and correctly allocated to Group 0 among methylated sites; similarly, the correct allocation rate for unmethylated sites (i.e., Group 1), denoted as P_1, is defined as the proportion of sites that are unmethylated and correctly allocated to Group 1 among unmethylated sites. The overall correct allocation rate, denoted as P, is defined as the proportion of correctly classified sites for both groups. Given Table 2.2, the correct allocation rates can be computed by $P_0 = \frac{M_{00}}{M_0}$, $P_1 = \frac{M_{11}}{M_1}$, and $P = \frac{M_{00}+M_{11}}{M_0+M_1}$. Note that the quantities on the right hand side of these equations are unknown. Following Basford and McLachlan

[83], they can be estimated by $\hat{M}_0 = M\hat{\pi}$, $\hat{M}_1 = M(1-\hat{\pi})$, $\hat{M}_{00} = \hat{\theta}_{k0} I(\hat{\theta}_{k0} > \hat{\theta}_{k1})$ and $\hat{M}_{11} = \hat{\theta}_{k1} I(\hat{\theta}_{k0} \leqslant \hat{\theta}_{k1})$, where $I(A)$ is an indicator of event A, such that $I(A)$ equals 1 if A is true and equals 0 otherwise. Thus P_0, P_1 and P can be estimated as follows.

$$\begin{aligned}
\hat{P}_0 &= \frac{1}{M\hat{\pi}} \sum_{k=1}^{M} \left\{ \hat{\theta}_{k0} I(\hat{\theta}_{k0} > \hat{\theta}_{k1}) \right\}; \\
\hat{P}_1 &= \frac{1}{M(1-\hat{\pi})} \sum_{k=1}^{M} \left\{ \hat{\theta}_{k1} I(\hat{\theta}_{k0} \leqslant \hat{\theta}_{k1}) \right\}; \\
\hat{P} &= \frac{1}{M} \sum_{k=1}^{M} \left\{ \hat{\theta}_{k0} I(\hat{\theta}_{k0} > \hat{\theta}_{k1}) + \hat{\theta}_{k1} I(\hat{\theta}_{k0} \leqslant \hat{\theta}_{k1}) \right\}.
\end{aligned}$$

As stated in [83], $\hat{P}_0 - P_0$, $\hat{P}_1 - P_1$ and $\hat{P} - P$ converge to 0 in probability when M goes to infinity. Therefore, $\hat{P}$, $\hat{P}_0$ and $\hat{P}_1$ can be used to assess the performance of the MSC procedure. Basford and McLachlan [83] also proposed two versions of bootstrap based methods to reduce the bias in estimating these correct allocation rates with $\hat{P}$, $\hat{P}_0$ and $\hat{P}_1$. However, we will not elaborate on the bias correction methods here. The reason is that, based on the simulation results reported in Section 2.3.2, the bias of the estimated correct allocation rates for our model is hardly noticeable.

The classification of two groups can also be viewed as a multiple testing problem once one of the groups is specified as the null [69]. For our proposed MSC procedure, if we designate one group (e.g., methylated group) to be the null, then the FDR and FNDR can also be defined. Although the MSC procedure is optimal based on the Bayes rule, it is not ascertained that it has control over FDR, which is the most widely used criterion in multiple testing context. In the next subsection, we will view our classification approach from a multiple testing perspective. We will first show how to estimate the resulting FDR and FNDR for the MSC procedure. Then motivated by the concern that a FDR level other than the estimated FDR may be needed, we will develop a FDR-controlled MSC procedure.

2.2.4 Methylation Status Calling Procedure with FDR control

We consider the following multiple testing problem after obtaining the estimated parameter $\hat{\phi}$ from Section 2.2.1 :

$$H_{i0} : p = \hat{p}_0 \quad vs \quad H_{ia} : p = \hat{p}_1,$$

where $i = 1, 2, ..., M$. As mentioned in Section 2.1.3, Lister *et al.* [66] also applied a multiple testing procedure to quantify DNA methylation status. Unlike the procedure used by [66], our procedure does not need to borrow information from the unmethylated Lambda DNA, instead, it can directly estimate p_1 as well as p_0 from the data.

Since $\hat{\theta}_{i1} = 1 - \hat{\theta}_{i0}$ for any i, only $\hat{\theta}_{i0}$ are used as the test statistic and they are referred to as null memberships hereafter. It is clear that the proposed classification rule is equivalent to the testing rule that rejects H_{i0} if $\hat{\theta}_{i0} \leqslant 0.5$. The four possible outcomes from the MSC procedure given in Table 2.2 can be viewed as the four possible outcomes from the multiple testing perspective. And the frequencies for the outcomes from the above multiple testing rule are exactly the same as those for the outcomes from the MSC procedure.

By the definitions of FDR and FNDR, we have $FDR = E\left[\frac{M_{01}}{V}\right]$ and $FNDR = E\left[\frac{M_{10}}{U}\right]$. For the MSC procedure, $U = \#\left\{\hat{\theta}_{k0} > \hat{\theta}_{k1}\right\}$ and $V = \#\left\{\hat{\theta}_{k0} \leqslant \hat{\theta}_{k1}\right\}$. Furthermore, based on the discussion in Section 2.2.3, we have $\hat{M}_{01} = \hat{M}_0 - \hat{M}_{00} = \sum_{k=1}^{M} \hat{\theta}_{k0} I(\hat{\theta}_{k0} \leqslant \hat{\theta}_{k1})$ and $\hat{M}_{10} = \hat{M}_1 - \hat{M}_{11} = \sum_{k=1}^{M} \hat{\theta}_{k1} I(\hat{\theta}_{k0} > \hat{\theta}_{k1})$. Therefore, FDR and FNDR for the MSC procedure can be estimated as follows.

$$\widehat{FDR} = \frac{\hat{M}_{01}}{V} = \frac{\sum_{k=1}^{M} \hat{\theta}_{k0} I(\hat{\theta}_{k0} \leqslant \hat{\theta}_{k1})}{\sum_{k=1}^{M} I(\hat{\theta}_{k0} \leqslant \hat{\theta}_{k1})}; \tag{2.4}$$

$$\widehat{FNDR} = \frac{\hat{M}_{10}}{U} = \frac{\sum_{k=1}^{M} \hat{\theta}_{k1} I(\hat{\theta}_{k0} > \hat{\theta}_{k1})}{\sum_{k=1}^{M} I(\hat{\theta}_{k0} > \hat{\theta}_{k1})}. \tag{2.5}$$

Although FDR and FNDR can be estimated for the MSC procedure, this procedure cannot control FDR at an arbitrary level. In practice, it can be a concern, especially when the estimated FDR exceeds an acceptable level. Therefore, it is desirable to

incorporate a FDR-controlling component into the MSC procedure. We shall investigate such a method next.

Notice that for the MSC procedure, the cutoff in the decision rule for rejecting the null hypothesis is 0.5. One way to control FDR is to adjust this cutoff according to the desirable FDR level. Suppose the prespecified FDR level is α. Then the goal here is to find a suitable cutoff c for null memberships such that the decision rule that rejects H_0 if

$$\hat{\theta}_{i0} \leqslant c, \quad i = 1, 2, ..., M. \tag{2.6}$$

will have an FDR below α.

We follow an adaptive procedure developed by Sun and Cai [84] to achieve the goal. In their original paper, Sun and Cai [84] aimed to find a multiple-testing procedure that is more efficient than the conventional p-value based procedures. They first developed a $Lfdr$-based procedure for marginal FDR control and showed it is optimal in the sense that it controls marginal FDR at level α with the smallest marginal FNDR. Then they proposed a data-dependent adaptive procedure based on estimated $Lfdr$ and proved that it asymptotically attains the performance of the optimal procedure. It was also demonstrated with numerical results that their adaptive procedure outperforms the conventional p-value based procedures when marginal FDR is controlled at the same level. For our problem, recall that for site i, g_{i0} and g_{i1} are the probability mass functions of $Bin(x_i, p_0)$ and $Bin(x_i, p_1)$, respectively. Note that the local false discovery rate of site i is given by

$$Lfdr_i = P(S_i = 0|y_i, x_i, \phi) = \frac{\pi g_{i0}(y_i)}{\pi g_{i0}(y_i) + (1-\pi) g_{i1}(y_i)}.$$

Therefore the null membership $\hat{\theta}_{i0}$ of site i is also an estimate of $Lfdr$. With this estimated $Lfdr$ of each site, the adaptive procedure proposed by [84] can be incorporated into the MSC procedure.

Since $Fdr(z)$ is the average of $Lfdr(Z)$ for $Z \leqslant z$ [72], the FDR of the decision rule 2.6 can be estimated by

$$\widehat{FDR}(c) = \frac{\sum_{i=1}^{M} \hat{\theta}_{i0} I(\hat{\theta}_{i0} \leqslant c)}{\sum_{i=1}^{M} I(\hat{\theta}_{i0} \leqslant c)}; \tag{2.7}$$

$$\widehat{FNDR}(c) = \frac{\sum_{i=1}^{M}(1-\hat{\theta}_{i0})I(\hat{\theta}_{i0} > c)}{\sum_{i=1}^{M} I(\hat{\theta}_{i0} > c)}. \tag{2.8}$$

When $c = 0.5$, which is the cutoff used by the MSC procedure, the resulting FDR and FNDR can be estimated by: $\widehat{FDR}(0.5) = \sum_{i=1}^{M} \hat{\theta}_{i0} I(\hat{\theta}_{i0} \leqslant 0.5) / \sum_{i=1}^{M} I(\hat{\theta}_{i0} \leqslant 0.5)$ and $\widehat{FNDR}(0.5) = \sum_{i=1}^{M}(1-\hat{\theta}_{i0})I(\hat{\theta}_{i0} > 0.5) / \sum_{i=1}^{M} I(\hat{\theta}_{i0} > 0.5)$. These two estimates are exactly the same as $\widehat{FDR}$ and $\widehat{FNDR}$ given in (2.4) and (2.5) because $\hat{\theta}_{i1} + \hat{\theta}_{i0} = 1$. Therefore $\widehat{FDR}(c)$ and $\widehat{FNDR}(c)$ given in (2.7) and (2.8) are extensions of $\widehat{FDR}$ and $\widehat{FNDR}$ to the general decision rule 2.6. Simulation results given in Section 2.3.4 provides compelling evidence that the estimators in (2.7) and (2.8) are accurate in estimating the true FDR and FNDR.

Suppose the desirable FDR level is α. We apply the method developed by [84] to choose the cutoff c so that the resulting classification procedure will have its FDR controlled at α. The procedure is described as follows.

1. Sort the null memberships in ascending order as $\hat{\theta}_{i_1 0}$,$\hat{\theta}_{i_2 0}$,...,$\hat{\theta}_{i_M 0}$.
2. Find $l = max\{j : \sum_{k=1}^{j} \hat{\theta}_{i_k 0}/j \leqslant \alpha\}$.
3. Then let $c = \hat{\theta}_{i_l 0}$ and all $H_{i_j 0}$ with $j \leqslant l$ are rejected.
4. Site i_j is called to be methylated if $j \leqslant l$; otherwise, it is called to be unmethylated.

We shall refer to this procedure as the FDR-controlled Methylation Status Calling Procedure at level α, or in short, the FMSC procedure at level α. Based on (2.7), the resulting FDR for the FMSC procedure can be estimated by $\widehat{FDR} = \sum_{k=1}^{l} \hat{\theta}_{i_k 0}/l$.

As mentioned in Section 2.2.3, Lister *et al.* [66] used $H_{i0} : p = \hat{p}_1$ as the null hypothesis, that is, the null hypothesis assumes that site i is unmethylated. In contrast, the null hypothesis we use here is $H_{i0} : p = \hat{p}_0$, or equivalently, it assumes that site i is methylated. Considering the fact that methylation is more prevalent in the sense that more than 60% of all CpG sites are expected to be methylated, it is more appropriate to assume the site is methylated in the null hypothesis instead of the

other way around. Assuming the site is unmethylated in the null hypothesis leads to the consequence that a significantly higher proportion of the claimed unmethylated sites are indeed methylated. Therefore, in terms of detecting unmethylated sites, our choice of null hypothesis produces more accurate results than the choice by [66]. See Section 2.3.3 for more detail.

Sun and Cai [84] showed that under several assumptions, the $Lfdr$-based adaptive method asymptotically attains the performance of the optimal method that controls marginal FDR at level α with the smallest marginal FNDR. Despite the discreteness and heterogeneity of the tests used for methylation status calling, our simulation study in Section 2.3.1 shows the incorporation of this adaptive procedure into the MSC procedure leads to satisfactory results. Therefore, we believe the FMSC procedure is adequate in making methylation status calls when controlling FDR at a given level is of interest. When the interest is to control FNDR at a given level, an adaptive procedure similar to FMSC can be developed.

2.2.5 EM Algorithm for Computing the Parameters

The log-likelihood function of ϕ is given by:

$$l(\phi|\mathbf{x},\mathbf{y}) = \sum_{i=1}^{M} ln\left\{\pi g_{i0} + (1-\pi)g_{i1}\right\}.$$

Then EM algorithm for ϕ can be developed as follows.

E-step: Compute memberships

$$\begin{aligned} \theta_{i0} &= P(S_i = 0|y_i) = \frac{\pi g_{i0}(y_i)}{\pi g_{i0}(y_i) + (1-\pi)g_{i1}(y_i)}; \\ \theta_{i1} &= 1 - \theta_{i0}. \end{aligned}$$

M-step: Update the estimates

$$\begin{aligned}
Q(\phi, \phi^{(t)}) &= E_{P(S|\mathbf{x},\mathbf{y},\phi^{(t)})}[lnP(\mathbf{x}, \mathbf{y}, S)|\phi] \\
&= \sum_{i=1}^{M}\sum_{c=0}^{1} P(S_i = c|y_i, x_i, \phi^{(t)}) \cdot lnP(x_i, y_i, S_i = c|\phi) \\
&= \sum_{i=1}^{M}\sum_{c=0}^{1} P(S_i = c|y_i, x_i, \phi^{(t)}) \cdot [ln(g_c(y_i)) + lnP(S_i = c)] \\
&= \sum_{i=1}^{M}[\theta_{i1}(ln(g_1(y_i)) + lnP(S_i = 1)) + \theta_{i0}(ln(g_0(y_i)) + lnP(S_i = 0))] \\
&= \sum_{i=1}^{M}[\theta_{i1} ln\mu + \theta_{i0} ln(1-\mu)] + \sum_{i=1}^{M}\theta_{i0} ln(g_0(y_i)) + \sum_{i=1}^{M} ln(g_1(y_i))
\end{aligned}$$

Differentiating $Q(\phi, \phi^{(t)})$ with respect to each component of ϕ yields the following results.

$$\begin{aligned}
\hat{\pi} &= \frac{\sum_{i=1}^{M}\theta_{i1}}{M}; \\
\hat{p}_0 &= \frac{\sum_{i=1}^{M}\theta_{i0}y_i}{\sum_{i=1}^{M}\theta_{i0}x_i}; \\
\hat{p}_1 &= \frac{\sum_{i=1}^{M}\theta_{i1}y_i}{\sum_{i=1}^{M}\theta_{i1}x_i}.
\end{aligned}$$

2.3 Simulation Results

2.3.1 Performance of MSC and FMSC

In this subsection, simulation results illustrating the behavior of our proposed procedures are presented. To carry out simulation study, we first use MethylC-Seq data of all CpG sites on Chromosome 1 of H9 human embryonic stem cells from [66] to fit a coverage distribution $\hat{f}$ (see Section 2.4.3 for more detail). Then we apply the mixture of Binomial model to the same data to obtain $\hat{\phi} = (\hat{p}_0, \hat{p}_1, \hat{\pi})$ (see Section 2.4.1). The total number of CpG sites in the simulation study is $M = 1000$. The general scheme of our simulation study is described as follows.

Step 1: Draw a random sample of M observations from $\hat{f}$ and use them as the coverage for M CpG sites. Let the simulated coverage of these M sites be $Z = (z_1, z_2, ..., z_M)$. For each of the M sites, generate its methylation status independently from $Bernoulli(\hat{\pi})$. Simulate C count for each site according to its methylation status and coverage. If the status for site i is methylated, the corresponding C count is generated from Bin(z_i,$\hat{p}_0$); otherwise, it is generated from Bin(z_i,$\hat{p}_1$). Denote the generated C counts as $R = (r_1, r_2, ..., r_M)$.

Step 2: Apply the mixture of Binomial model to R and Z, obtain $\tilde{\phi} = (\tilde{p}_0, \tilde{p}_1, \tilde{\pi})$, compute the memberships, and make methylation status call for each site using the MSC procedure.

Step 3: For $i = 1, 2, ..., M$, compute the p-value, denoted as q_i, for testing $H_{i0} : p = \tilde{p}_0 \quad vs \quad H_{ia} : p = \tilde{p}_1$ using exact Binomial test, which is, $q_i = \sum_{k=0}^{r_i} \begin{pmatrix} z_i \\ k \end{pmatrix} (\tilde{p}_0)^k (1-\tilde{p}_0)^{z_i-k}$. After obtaining the p-values, we apply the FDR-controlling procedure proposed by [67] at level $\alpha = 0.1$ to make methylation status calls. We shall refer this procedure to as the individual Binomial testing (IBT) procedure.

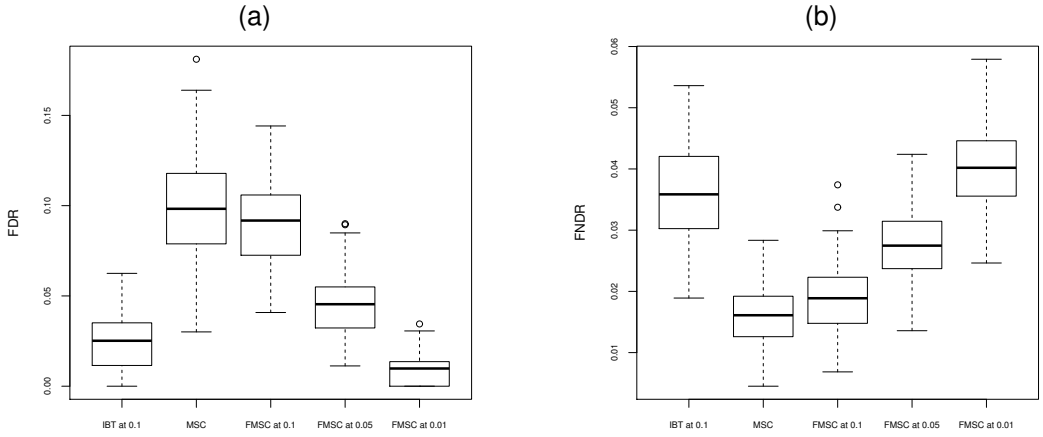
Step 4: Use the FMSC procedure described in Section 2.2.4 to control FDR at three levels $\alpha = (0.1, 0.05, 0.01)$ separately. Note that in the simulation study, for each site, five methylation status calls are made based on three different methods, which are the MSC procedure, the IBT procedure at level 0.1, and the FMSC procedure with three different choices of FDR level. By comparing these calls to the true methylation status, performances of these three procedures can be compared in terms of FDR and FNDR.

The comparison results based on $N = 100$ repeated simulations are displayed in Figure 2.2. Several observations can be made from the two plots in Figure 2.2. First, the median FDRs for MSC and FMSC at level 0.1 is around 0.1, and the corresponding median FNDRs are around 0.018 and 0.020, respectively. It shows that the MSC procedure produces similar FDR and FNDR results as the FMSC

procedure at level 0.1. Second, the FDRs for the FMSC procedures at all three levels are well controlled. Third, the FNDR for the FMSC procedure at level 0.1 is notably smaller than the FNDR for the IBT procedure at level 0.1. It is caused by the fact that the IBT procedure at level 0.1 overcontrols FDR in the sense that the median FDR is only around 0.05. As a result, the FNDR for IBT is compromised. It suggests that the FMSC procedure is more powerful than the IBT procedure when their FDRs are controlled at the same level.

In the simulation study, we also applied other FDR-controlling procedures to q_i's. They include the q-value method [70] and procedures proposed by [68], [76] and [78]. The results are insensitive to the type of procedure used. Hence only the results from the FDR-controlling procedure proposed by [67] are shown here.

Fig. 2.2. (a) The box plots display FDRs for IBT at level 0.1, the MSC procedure, and the FMSC procedure with FDR level 0.1, 0.05 and 0.01 from left to right. (b) The box plots display FNDRs for these methods in the same order.



2.3.2 Estimation of Correct Allocation Rates

In Section 2.2.3, we follow the method proposed by Basford and McLachlan (1985) to estimate the correct allocation rates for MSC. Here, we also apply both parametric version and semiparametric version of bootstrap procedures proposed by them to correct the bias. Table 2.3 reports the simulation results. The simulation setting is described in Step 1 and Step 2 of the simulation scheme given in Section 2.3.1. It can be observed that the estimates for the whole population and the methylated group are more accurate than that for the unmethylated group. However, both of the bootstrap procedures do not help much in providing more accurate estimates.

			Parametric Version		Semiparametric Version	
Population	True Rate	Estimated Rate	Estimate of Bias	Corrected Rate	Estimate of Bias	Corrected Rate
overall	0.9750	0.9684	-0.0002	0.9686	0.0005	0.9678
methylated	0.9841	0.9822	0.0002	0.9819	-0.0003	0.9824
unmethylated	0.9068	0.8684	-0.0029	0.8713	0.0056	0.8628

Table 2.3.
Estimation of the overall correct allocation rate and correct allocation rates for the two subgroups.

2.3.3 Choice of Null Hypothesis in FDR control

In this subsection, we compare FDR and FNDR results for different choices of null hypothesis. 100 repeated simulations are carried out. In each simulation, the setting is the same as that in Section 2.3.1. The individual Binomial testing (IBT) procedure for two choices of null hypothesis and the MSC procedure are applied for each simulated data. To make the results comparable, the FDR for the IBT procedure with

the null hypothesis used by [66] is controlled at level 0.04. The comparison results are shown in Figure 2.3. The proportion of methylated sites that are allocated to unmethylated group among those allocated to unmethylated group is around 0.1 for both MSC and the IBT procedure for our choice of null hypothesis, while this proportion for the IBT procedure for the null hypothesis used by [66] is rarely smaller than 0.3. It illustrates that the MSC procedure outperforms the other two, and using methylated group as the null hypothesis provides significantly better accuracy in terms of detecting unmethylated sites.

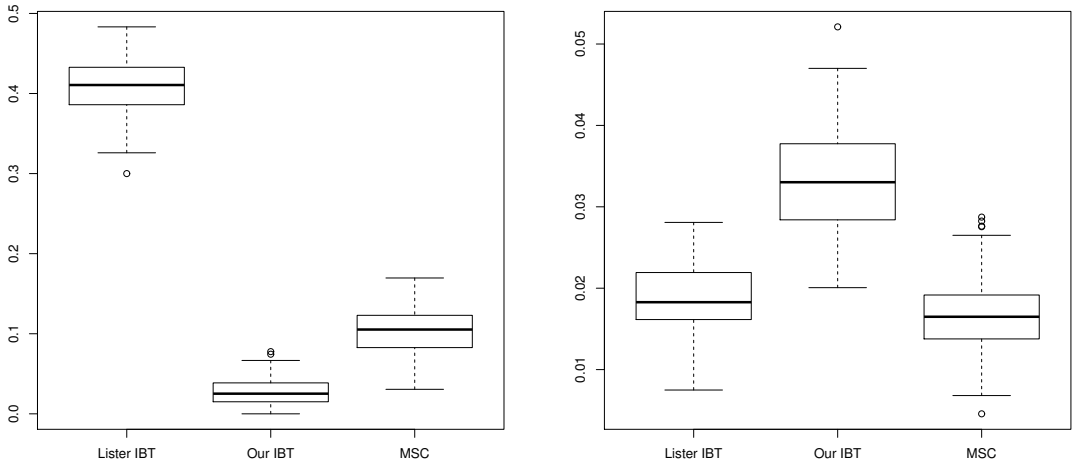


Fig. 2.3. Comparison of different choices of null hypothesis. Left: Proportion of methylated sites that are allocated to unmethylated group among those allocated to unmethylated group. Right: Proportion of unmethylated sites that are allocated to methylated group among those allocated to methylated group.

2.3.4 Estimation of FDR and FNDR with Memberships

In Section 2.2.4, we show how to estimate the true FDR and FNDR for any cutoff c. In this subsection, we report simulation results on the performances of these estimates

for three choices of c, which are 0.5, 0.4 and 0.6. For each choice of c, 100 repeated simulations are carried out, and in each simulation, the setting is the same as that in Section 2.3.1. Figure 2.4 displays the results for the above three choices of c. It can be observed that both of the estimates are accurate for all three choices of c.

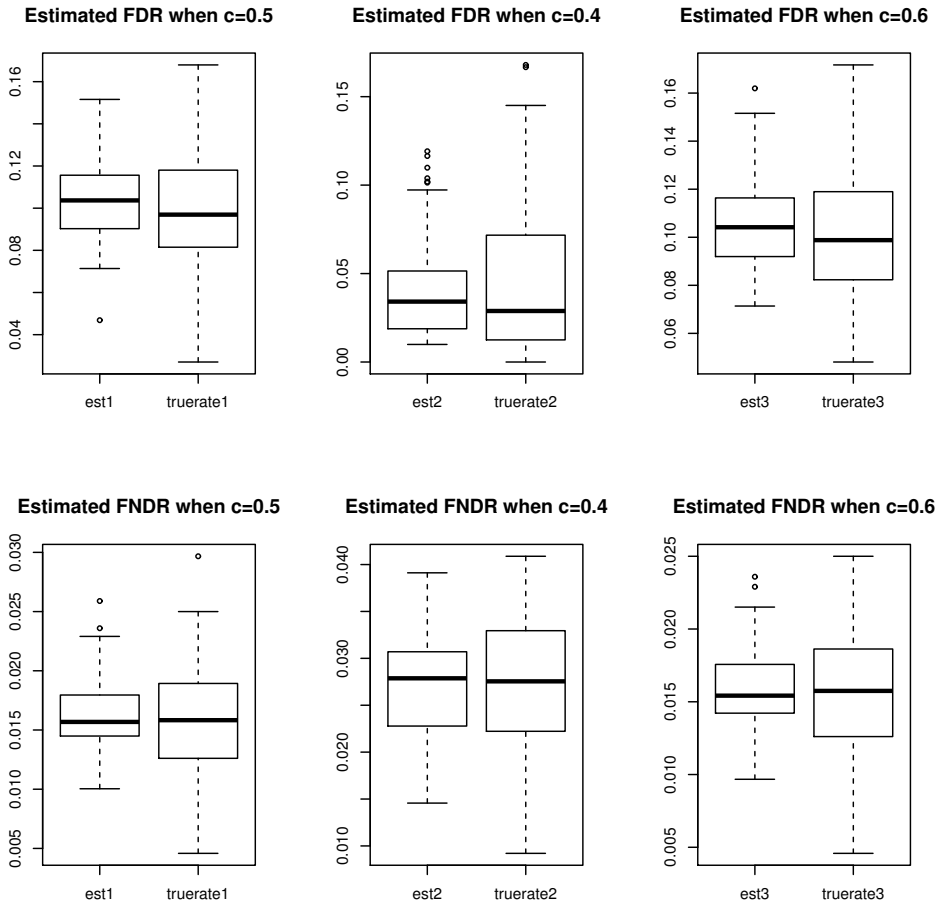


Fig. 2.4. Estimation of FDR and FNDR for $c = (0.5, 0.4, 0.6)$

2.4 Real Data Application

2.4.1 Performance of MSC and FMSC

The MSC and FMSC procedures are applied to a real MethylC-Seq data of H9 human embryonic stem cells from [66]. Three FDR levels, 0.1, 0.05 and 0.01, are considered for FMSC. We first apply the procedures genome-wide. The resulting estimate for $\phi = (p_0, p_1, \pi)$ is $\hat{\phi} = (0.9102, 0.1088, 0.8920)$. The MSC procedure is also applied to the same MethylC-Seq data chromosome-wise. The results are given in Table 2.4. In the genome-wide evaluation with MSC, 42,987,496 out of 48,795,269 CpG sites are called to be methylated. For the chromosome-wise evaluation, a total of 43,097,321 CpG sites are declared to be methylated. The difference is approximately 109 thousands, which account for less than 0.3% of all covered CpG sites. The detailed comparison results are given in Table 2.5. This high concordance suggests the consistency of the MSC procedure.

Correct allocation rates, estimated FDR and estimated FNDR for genome-wide analysis by MSC and FMSC at three FDR levels are also calculated. The results are given in Table 2.6. For MSC, the correct allocation rates for the overall population and the methylated group are 0.9771 and 0.9810, respectively, while the rate for the unmethylated group is 0.9450. As for FDR and FNDR, the estimates for MSC are 0.1426 and 0.0067, respectively. For FMSC, as the FDR level decreases, the correct allocation rate for the overall population decreases slightly and the rate for the methylated group increases slightly, whereas the correct allocation rate for the unmethylated group is influenced more dramatically. It decreases from 0.9450 to 0.6395 as the FDR level decreases from 0.1 to 0.01. For FMSC at any of the three FDR levels, the resulting FDR is well controlled. And as expected, the estimated FNDR increases as the FDR level decreases. Based on these results, the performances of MSC and FMSC are acceptable.

Chromo-some	Estimate of ϕ			Number of covered CpG sites	Number of called methylated CpG sites
	$\hat{p}_1$	$\hat{p}_0$	$\hat{\pi}$		
1	0.9087	0.1024	0.8794	3,975,983	3,452,088
2	0.9126	0.0985	0.9027	3,782,466	3,369,491
3	0.9124	0.0954	0.9075	2,881,395	2,579,804
4	0.9143	0.1160	0.9117	2,550,805	2,357,090
5	0.9126	0.0982	0.9038	2,625,905	2,342,521
6	0.9120	0.0895	0.8938	2,609,387	2,301,334
7	0.9115	0.1179	0.8973	2,690,318	2,386,312
8	0.9120	0.1116	0.9052	2,286,755	2,044,208
9	0.9100	0.1114	0.8881	2,034,256	1,783,988
10	0.9111	0.1054	0.8977	2,369,400	2,099,966
11	0.9087	0.0999	0.8800	2,290,997	1,989,217
12	0.9120	0.0952	0.8907	2,274,053	1,999,301
13	0.9152	0.1050	0.9141	1,418,219	1,311,929
14	0.9100	0.1002	0.8857	1,521,511	1,330,347
15	0.9091	0.1077	0.8836	1,482,079	1,293,617
16	0.9113	0.1018	0.8857	1,899,471	1,660,913
17	0.9118	0.0852	0.8565	2,038,882	1,723,421
18	0.9126	0.1228	0.9089	1,203,933	1,081,334
19	0.9051	0.0970	0.8340	1,887,982	1,555,059
20	0.9071	0.1076	0.8852	1,299,389	1,135,158
21	0.9114	0.1835	0.9008	666,612	596,338
22	0.9104	0.1076	0.8838	1,008,544	879,711
X	0.8984	0.3440	0.8815	1,983,863	1,815,991
Y	0.7738	0.1954	0.6713	13,064	8,183

Table 2.4.
Chromosome by Chromosome Results with MSC for the MethylC-Seq data from [66].

Chromosome-wise \ Genome-wide	Methylated sites	Unmethylated sites	Total
Methylated sites	42,958,932	138,389	43,097,321
Unmethylated sites	28,564	5,669,384	5,697,948
Total	42,987,496	5,807,773	48,795,269

Table 2.5.
Contingency Table for Chromosome-wise and Genome-wide Evaluation

Procedure	Correct Allocation Rate			Estimated FDR	Estimated FNDR
	Overall Population	Methylated Group	Unmethylated Group		
the MSC procedure	0.9771	0.9810	0.9450	0.1426	0.0067
FMSC at level 0.10	0.9758	0.9883	0.8727	0.1000	0.0154
FMSC at level 0.05	0.9744	0.9949	0.8057	0.0500	0.0231
FMSC at level 0.01	0.9604	0.9992	0.6395	0.0100	0.0418

Table 2.6.
Assessment of Genome-wide Analysis by MSC and FMSC at three levels.

2.4.2 Comparison of MSC and FMSC with Existing Methods

Next, the whole-genome results from the MSC procedure are compared to those from the procedure used by [66]. The comparison results are shown in Table 2.7.

Table 2.7.
Comparison of whole-genome results from the MSC procedure and those from the procedure used by [66] for all covered CpG sites.

Our method / Lister	Methylated sites	Unmethylated sites	Total
Methylated sites	42,987,456	1,175,275	44,162,731
Unmethylated sites	40	4,632,498	4,632,538
Total	42,987,496	5,807,773	48,795,269

Table 2.7 shows that these two procedures agree with each other on the methylation status calls of 47,619,954 CpG sites, which account for more than 97% of all covered CpG sites. For the sites that these two procedures make different methylation status calls, they disagree in two directions. There are only 40 CpG sites that our MSC procedure declares to be methylated but the procedure used by [66] declares to be unmethylated; and we refer to this type of disagreement as the first direction. There are roughly 1.17 million CpG sites that are called to be unmethylated by the MSC procedure but called to be methylated by the procedure used by [66]; and we refer to this type of disagreement as the second direction.

Since there are only 40 CpG sites in the first direction but 1.17 million sites in the second direction, we will focus on the second direction in the subsequent analysis. A typical example in the second direction is that for a site with coverage 60 and C count 6, MSC declares it to be unmethylated whereas the procedure used by [66] declares it to be methylated. Several other typical cases are shown in Table 2.8. As mentioned in the last paragraph, the null hypothesis for [66] is that the site is unmethylated, therefore, p-value is computed as p-value $= \sum_{k=y_i}^{x_i} \begin{pmatrix} x_i \\ k \end{pmatrix} (\hat{p}_1^{Lis})^k (1-\hat{p}_1^{Lis})^{x_i-k}$. Because Lister *et al.* [66] used an extremely small $\hat{p}_1^{Lis}$, which is less than 0.01, the resulting p-value relies heavily on the C count in the sense that it decays to zero exponentially

with increasing C count, regardless of coverage x_i and $\hat{\pi}$. Therefore, the C count threshold for declaring one site to be methylated based on the multiple testing procedure used by [66] is generally low, even for sites with high coverage. However, for the MSC procedure, the null membership primarily depends on the proportion of C count at one site instead of C count alone. The cutoff for the proportion is around a half for all sites, which is intuitively more reasonable. Thus, the difference in the cutoff values for these two procedures becomes more evident when coverage increases. This difference is essentially caused by the underestimation of p_1 in the procedure used by [66], and it demonstrates that the procedure used by [66] lacks power in terms of detecting unmethylated sites, especially for sites with moderate to high coverage. Therefore, we believe the MSC procedure makes more accurate methylation status calls for this type of disagreement.

		MSC: Unmethylated	Lister's: Methylated
C count	Coverage	Null membership	P-value
3	10	5.105499e-04	1.461094e-05
3	35	6.193205e-29	7.257909e-04
3	60	7.508808e-54	3.458198e-03
4	80	7.274099e-72	7.303491e-04
5	95	6.780415e-85	1.246015e-04
10	116	3.147355e-96	4.919773e-10
11	156	3.556955e-134	5.842736e-10

Table 2.8.
Typical examples of sites that the MSC procedure declares to be unmethylated but the procedure used by [66] declares otherwise

As a final evaluation, the methylation calls for those sites that MSC and the procedure used by [66] disagree on are compared to the results obtained from Infinium Human Methylation 450K BeadChip. The Human Methylation 450K data used here

is first analyzed by Merling *et al.* [85]. For those roughly 1.17 million sites that MSC and the procedure used by [66] disagree on, 27,637 sites are covered by Human Methylation 450K BeadChip. We use 0.5 as the cutoff value to dichotomize the beta values in Human Methylation 450K BeadChip data to make binary methylation calls, and compare the calls to those obtained from MSC and the procedure used by [66]. The comparison result is given in Table 2.9.

Table 2.9.
Third platform validation of the methylation calls for those sites that MSC and the procedure used by [66] disagree on

Procedure	Number of sites that agree with the third platform	Number of sites that disagree with the third platform
MSC	18,090	9,547
Lister's	9,547	18,090

Table 2.9 shows that for nearly two thirds of the 27,637 target sites, the methylation calls made by MSC are consistent with the calls made by Human Methylation 450K BeadChip. This suggests the calls made by the MSC procedure are more likely to be correct than those obtained by the procedure used by [66].

2.4.3 Coverage Distribution

When fitting the coverage distribution, we find that the zero-truncated Negative Binomial model provides a satisfactory fit. The pmf of the model is given as follows.

$$f(x; r, v) = \frac{\Gamma(x + r)v^r(1 - v)^x}{(1 - v^r)\Gamma(r)x!}.$$

Figure 2.5 and Table 2.10 display coverage data, fitted probabilities, estimated parameters for the MethylC-Seq data on Chromosome 1 from [66].

$\hat{v}$	$\hat{r}$
0.1133268	2.7033616

Table 2.10.
Maximum Likelihood Estimates of $\hat{v}$ and $\hat{r}$ for Chromosome 1.

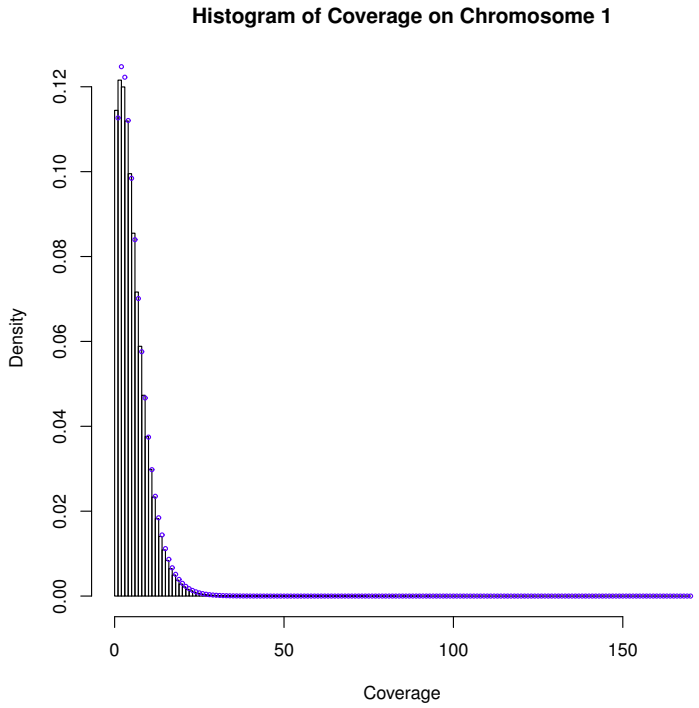


Fig. 2.5. Histogram of the coverage with blue dots indicating fitted probabilities.

2.5 Recent Development on DNA Methylation Analysis and FDR Controlling Procedures

2.5.1 DNA methylation status quantification for Bisulphite-sequencing data

As pointed out in Baumann and Doerge [86], DNA methylation patterns can be dramatically heterogeneous between different annotated regions, which implies that it is of value to bring annotation information into DNA methylation analysis. Baumann and Doerge [87] incorporated the genome annotation information when analyzing bisulphite-sequencing data, and proposed two differential methylation detection approaches, referred to as Methylation Analysis using Genome Information (MAGI). In the first approach, each Cytosine in a given annotated region was tested for differential methylation with Fisher's exact test, and false discovery rate (FDR) was controlled at a pre-specified level within the region. This procedure was employed for each genome annotation region of interest. Subsequently, if the proportion of differentially methylated sites within an annotated region exceeded a certain threshold, the region was called to be differentially methylated. This approach is referred to as MAGI_{C} approach. In the second approach, the observed proportion of methylated reads was first used to make the binary DNA methylation status call for each site: if the proportion exceeded a certain threshold, the site was called to be methylated; otherwise, it was called to be unmethylated. In Baumann and Doerge [87], the threshold was chosen to be the mean of the two cluster centroids obtained by using the k-means clustering on each chromosome and strand. Then, the site level methylation status can be aggregated to the region level, and Fisher's exact test with FDR controlled at a pre-specified level can be conducted to detect differentially methylated genomic regions. This approach is referred to as MAGI_{G} approach. These two approaches can work for both unreplicated and replicated data, and they can provide a gain in statistical power, compared with existing differential methylation detection methods.

Zheng *et al.* [88] focused on the DNA methylation status predictions for CpG sites within CpG island (CGI). They considered a large collection of features, including CGI-related attributes, DNA composition patterns, distributions of the transcription factor binding sites, histone modification marks, and gene functions. These features went through feature selection, and support vector machine was applied on the selected features to quantify DNA methylation status. It was demonstrated that their predictive models perform well for different cell types. It was also shown that histone modification information makes a significant contribution to the prediction of DNA methylation status.

Zhang *et al.* [89] used a random forrest (RF) classifier to make binary DNA methylation status predictions at CpG sites. To build the classifier, they relied on 124 features that can be grouped into four classes: information on neighboring sites, genomic position, DNA sequence properties, and regulatory elements. They showed their method achieves high accuracy for both genome-wide and CGI-specific DNA methylation status predictions. Moreover, the contribution of each feature can be evaluated to identify genomic features related to the occurence of DNA methylation. They also compared the performance of the RF classifier with other popular classification algorithms, such as k-nearest neighbors classifier and logistic regression. It was demonstrated that the RF classifier has higher prediction accuracy and larger area under the receiver operating characteristic curve (AUC).

Prochenka *et al.* [90] pointed out that sometimes it may not be appropriate to use our proposed approach [47] to make binary methylation status calls. The main reason is that when there exists a mixture of DNA molecules, the methylation status at a given C site can be heterogeneous. We acknowledge that for some studies, the goal is to obtain continuous methylation status quantification, instead of binary ones. In such studies, the memberships derived in our framework [47] can serve as the continuous alternative to DNA methylation status quantification.

2.5.2 Sequencing-based DNA Methylation Profiling Approaches

In the past few years, a number of new approaches have been developed for DNA methylation profiling. As noted by Plongthongkum *et al.* [91], these new approaches mainly aim to improve the current ones in four aspects: sample input, throughput, accuracy and cost. To name a few among them, the dRRBS method [92] utilized a pair of enzymes to fragment the sample DNAs. By doing this, an increased coverage in both high-CG and low-CG regions can be achieved. As a result, higher throughput and more accurate DNA methylation profiles can be yielded. Using barcoded adapters, mRRBS [93] was able to process more samples in parallel, compared to the original RRBS. As a result, mRRBS achieves increased throughput at lower cost. Meanwhile, several attemps have been made to reduce the amount of sample DNAs required, for both whole-genome and CpG-specific methylation detection experiments. They include LCM-RRBS [94], single-cell RRBS [95], Tn5mC-seq [96] and PBAT [97] among others.

It has been discovered recently that hydroxymethylation is another important epigenetic modification to the carbon-5 position of the Cytosine. Traditional bisulphite sequencing methods cannot distinguish between 5mC and 5hmC, as they both are read as Cytosines after bisulphite treatment. To discriminate them effectively, oxidative bisulfite sequencing (oxBs-seq) has been developed, where 5hmC was converted to Uracil by oxidation and subsequent bisulphite treatment [98]. Thus, 5mC and 5hmC became distinguishable at single-base resolution by comparing sequencing reads from oxBs-seq and bisulphite-sequencing experiments. Other methods that can discriminate 5mC and 5hmC include RRHP [99] and TAB-Seq [100]. A detailed review of recent development in DNA modification analysis can be found in Plongthongkum *et al.* [91].

The development of third-generation sequencing technologies, such as single molecule real time (SMRT) sequencing [101], induces a new category of DNA methylation profiling methods. Flusberg *et al.* [102] introduced a direct DNA methylation detection

method during SMRT sequencing, which did not need to apply bisulphite treatment; Yang *et al.* [103] coupled bisulphite conversion with SMRT sequencing. As pointed out by Plongthongkum *et al.* [91], the effectiveness of these third-generation sequencing based approaches still remains to be demonstrated.

2.5.3 FDR Controlling Procedures for Discrete Tests

As noted by Chen and Doerge [104], the existing FDR procedures usually suffer from conservativeness when applied to discrete and heterogeneous tests. The reason is two-fold. First, unlike the continuous scenario, the null distribution for the p-values from discrete and heterogeneous tests is dominated by the standard Uniform distribution. Second, the existing estimators for the true proportion of nulls have an upward bias. Taking these two factors into consideration, Chen and Doerge [105] proposed a generalized estimator for the true null proportion, a new divergence function, and a novel grouping strategy for discrete and heterogeneous p-values based on the proposed divergence. In addition to them, a novel FDR controlling procedure under discrete and heterogeneous tests was introduced. This procedure can also be applied to continuous tests, and is easy to implement as it does not require resampling. It was shown with empirical studies that the proposed procedure is more powerful than other existing procedures for the three widely used discrete tests, which are Binomial test, Fisher's exact test and Exact negative binomial test.

Heller and Gur [106] developed a FDR controlling procedure based on Benjamini and Liu [107] by using the mid p-values instead of the original p-values, to mitigate the effect of discreteness. This procedure is similar to the one proposed by Heyse [78] where the mid p-values were combined with the procedure proposed by Benjamini and Hochberg [67]. In simulation studies, it was shown that the mid p-value based procedures are more powerful than the original ones.

3. FUTURE WORK

3.1 Future Research Topics for Variable Selection under Single Index Model

In this subsection, we will briefly discuss several future work directions for the proposed BS-SIM framework. To begin with, we would like to extend our framework to multi-index model

$$Y = f(X^T B) + \varepsilon,$$

where B is a $p \times d$ matrix. One intuitive way here is to use d additive one-dimensional spline functions. This approach has the advantage of conducting univariate estimation for each projection direction, and can possibly avoid the curse of dimensionality. However, it sacrifaces certain flexibility for modeling. Another way for estimating multi-index model with our framework is to apply multivariate splines. This approach is more flexible than additive spline approach. Nevertheless, the number of multivariate spline basis functions grows exponentially with the number of projection directions d [29], and thus this approach may suffer from the curse of dimensionality. We also would like to extend our framework to discrete response. This extension can benefit a number of areas, such as marketing and risk management, where binary response is frequently expected. Secondly, the cubic B-splines are applied in Chapter 1, and the number of knots used is determined by the rule of thumb. It has been previously proved that $N \sim n^{1/5}$ is the optimal rate for the number of knots in terms of minimizing the mean integrated squared error for nonparametric spline estimation. In BS-SIM, we rely on this result to accomplish the selection for N. Nevertheless, we show in Section 1.4 that the number of knots N also plays a crucial role in characterizing BS-SIM's ability in selecting the true variables. It would be of interest to study the optimal choice for N in the sense that it can lead to the best performance on

selection consistency. A related research direction is to look into the location of the knots. In Section 1.1.4, we briefly mention several previous work on the placement for the knots in spline estimation. It is also worthwhile to examine how to incorporate these approaches into BS-SIM. Next, we will implement the algorithm for the linearly constrained single index model developed in Section 1.7.2, and examine its performance with simulation studies. Furthermore, it is also worthwhile to investigate the theoretical properties of the resulting estimator.

3.2 Future Research Topics for DNA Methylation Status Calling

Our work in DNA methylation status calling also points to several future research directions. It is interesting to study how to combine the newly developed methods with our framework to make more precise DNA methylation calls or differential DNA methylation detection. For instance, due to the discreteness and heterogeneity of the p-values in our framework, the FDR procedure proposed in Chen and Doerge [105] can be applied and lead to a more powerful FDR controling procedure. On the other hand, Bowtie was used to align the read sequences by Lister *et al.* [66]. This may lead to a bias towards the reference allele. How to correct this bias is worth exploring. Possible solutions include the methods proposed by Wu *et al.* [108] and Yuan *et al.* [109].

REFERENCES

REFERENCES

[1] P. McCullagh and J. Nelder, *Generalized Linear Models.* Chapman and Hall/CRC, 1989.

[2] L. Wasserman, *All of Nonparametric Statistics.* Springer, 2006.

[3] H. Ichimura, "Semiparametric least squares (sls) and weighted sls estimation of single-index models," *Journal of Econometrics*, vol. 58, pp. 71–120, 1993.

[4] W. Härdle and T. M. Stoker, "Investigating smooth multiple regression by the method of average derivatives." *Journal of the American Statistical Association*, vol. 84, pp. 986–995, 1989.

[5] J. L. Powell, J. H. Stock, and T. M. Stoker, "Semiparametric estimation of index coefficients." *Econometrica*, vol. 57, pp. 1403–1430, 1989.

[6] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny, "Structure adaptive approach for dimension reduction," pp. 1537–1566, 2001.

[7] Y. Xia, "Asymptotic distribution for two estimators of the single-index model." *Econometric Theory*, vol. 22, pp. 1112–1137, 2006.

[8] J. L. Horowitz and W. Härdle, "Direct semiparametric estimation of single-index models with discrete covariates." *Journal of the American Statistical Association*, vol. 91, pp. 1632–1640., 1996.

[9] Y. Xia, H. Tong, W. K. Li, and L. Zhu, "An adaptive estimation of dimension reduction space (with discussion)." *Journal of the Royal Statistical Society, Series B*, vol. 64, pp. 363–410, 2002.

[10] K. C. Li, "Sliced inverse regression for dimension reduction (with discussion)." *Journal of the American Statistical Association*, vol. 86, pp. 316–342., 1991.

[11] R. D. Cook and B. Li, "Dimension reduction for the conditional mean in regression." *Annals of Statistics*, vol. 30, pp. 455–474, 2002.

[12] B. Li and S. Wang, "On directional regression for dimension reduction," *Journal of the American Statistical Association*, vol. 102, p. 997, 2007.

[13] L. Wang and L. Yang, "Spline estimation of single-index models." *Statistica Sinica*, vol. 19, pp. 765–783, 2009.

[14] P. Naik and C.-L. TSAI, "Single-index model selections." *Biometrika*, vol. 88, pp. 821–832, 2001.

[15] K. E. and X. Y.C., "Variable selction for the single-index model." *Biometrika*, vol. 94, pp. 217–229, 2007.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.

[17] P. Zhao and B. Yu, "On model selection consistency of lasso." *Journal of Machine Learning Research*, pp. 2541–2563, 2006.

[18] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, vol. 96, pp. 1348–1360, 2001.

[19] H. Zou, "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[20] E. J. Candes and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n (with discussion)." *The Annals of Statistics*, vol. 35, pp. 2313–2404, 2007.

[21] Q. Wang and X. Yin, "A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave." *Computational Statistics and Data Analysis*, vol. 52, pp. 4512–4520, 2008.

[22] P. Zeng, T. He, and Y. Zhu, "A lasso-type approach for estimation and variable selection in single index models." *Journal of Computational and Graphical Statistics*, vol. 21, pp. 92–109, 2012.

[23] T. Wang, P. Xu, and L. Zhu, "Penalized minimum average variance estimation." *Statistica Sinica*, vol. 22, pp. 543–569, 2013.

[24] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools (with discussion)." *Technometrics*, pp. 109–148, 1993.

[25] H. Peng and T. Huang, "Penalized least squares for single index models." *Journal of Statistical Planning and Inference*, vol. 141, pp. 1362–1379, 2011.

[26] J. Lv and Y. Fan, "A unified approach to model selection and sparse recovery using regularized least squares." *The Annals of Statistics*, vol. 37, no. 6, 2009.

[27] M. Nikolova, "Local strong homogeneity of a regularized estimator." *SIAM J. Appl. Math.*, no. 61, pp. 633–658, 2000.

[28] C. de Boor, *A Practical Guide to Splines.* New York: Springer-Verlag, 2001.

[29] H. T. Friedman, J. and R. Tibshirani, *The elements of statistical learning.* New York: Springer series in statistics, 2001.

[30] S. Zhou and X. Shen, "Spatially adaptive regression splines and accurate knot selection schemes." *Journal of the American Statistical Association*, vol. 96, pp. 247–259, 2001.

[31] M. Osborne, B. Presnell, and T. B.A., "Knot selection for regression splines via the lasso." *Computing Science and Statistics*, vol. 30, pp. 44–49, 1998.

[32] J. Nocedal and S. Wright, *Numerical Optimization.* New York: Springer-Verlag, 2006.

[33] G. Schwarz, "Estimating the dimension of a model." *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[34] J. Shao, "An asymptotic theory for linear model selection (with discussion)." *Statistica Sinica*, vol. 7, pp. 221–264, 1997.

[35] H. Wang, B. Li, and C. Leng, "Shrinkage tuning parameter selection with a diverging number of parameters." *Journal of the Royal Statistical Society, Series B*, vol. 71, pp. 671–683, 2009.

[36] Y. Fan and C. Tang, "Tuning parameter selection in high dimensional penalized likelihood." *Journal of the Royal Statistical Society, Series B*, vol. 75, pp. 531–552, 2013.

[37] J. Chen and Z. Chen, "Extended bayesian information criteria for model selection with large model space." *Biometrika*, vol. 95, pp. 759–771, 2008.

[38] M. Ming and Y. He, "Pten: new insights into its regulation and function in skin cancer." *Journal of Investigative Dermatology*, vol. 129, pp. 2109–2112, 2009.

[39] S. Piccolo, M. Cordenonsi, and S. Dupont, "Molecular pathways: Yap and taz take the center stage in organ growth and tumorigenesis." *Clinical Cancer Research*, vol. 19, pp. 4925–4930, 2013.

[40] A. Roesch and et al., "Overexpression and hyperphosphorylation of retinoblastoma protein in the progression of malignant melanoma." *Modern Pathology*, vol. 18, pp. 565–572, 2005.

[41] M. Lu and et al., "Restoring p53 function in human melanoma cells by inhibiting mdm2 and cyclin b1/cdk1-phosphorylated nuclear iaspp." *Cancer Cell*, vol. 23, pp. 618–633, 2013.

[42] Z. Liu and et al., "Notch1 signaling promotes primary melanoma progression by activating mitogen-activated protein kinase/phosphatidylinositol 3-kinase-akt pathways and up-regulating n-cadherin expression." *Cancer Research*, vol. 66, pp. 4182–4190, 2006.

[43] T. He, "Lasso and general l_1 regularized regression under linear equality and inequality constraints." *Ph.D. thesis, Purdue University*, 2011.

[44] G. James, C. Paulson, and P. Rusmevichientong, "The constrained lasso." *Technical report, University of Southern California*, 2012.

[45] S. Rosset and J. Zhu, "Piecewise linear regularized solution paths." *The Annals of Statistics*, vol. 35, no. 3, pp. 1012–1030, 2007.

[46] S. van de Geer, *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

[47] L. Cheng and Y. Zhu, "A classification approach for dna methylation profiling with bisulfite next-generation sequencing data." *Bioinformatics*, vol. 30, 2014.

[48] S. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard, "An operational definition of epigenetics." *Genes Dev.*, vol. 23, 2009.

[49] R. Feil and M. Fraga, "Epigenetics and the environment: emerging patterns and implications." *Nature Reiew. Genetics*, vol. 13, pp. 97–109, 2011.

[50] P. Laird, "Principles and challenges of genomewide dna methylation analysis." *Nature Reviews Genetics*, vol. 11, pp. 191–203, 2010.

[51] K. Robertson, "Dna methylation and human disease." *Nature Reviews Genetics*, vol. 6, pp. 597–610, 2005.

[52] H. Hayatsu, "Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for dna methylation analysis a personal account." *Proc. Jpn Acad. Ser. B Phys. Biol. Sci.*, vol. 84, pp. 321–330, 2008.

[53] M. Frommer and et al., "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands." *Proc. Natl Acad. Sci. USA*, vol. 89, pp. 1827–1831, 1992.

[54] M. R. Estecio and et al., "High-throughput methylation profiling by mca coupled to cpg island microarray." *Genome Research*, vol. 17, pp. 1529–1536, 2007.

[55] M. Weber and et al., "Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome." *Nature Genetics*, vol. 39, pp. 457–466, 2007.

[56] M. Bibikova and et al., "Genome-wide dna methylation profiling using infinium assay." *Epigenomics 1*, pp. 177–200, 2009.

[57] M. Metzker., "Sequencing technologies - the next generation." *Nature Reviews Genetics*, vol. 11, 2010.

[58] M. Oda and et al., "High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers." *Nucleic Acids Research*, vol. 37, pp. 3829–3839, 2009.

[59] T. A. Down and et al., "A bayesian deconvolution strategy for immunoprecipitation-based dna methylome analysis." *Nature Biotechnology*, vol. 26, pp. 779–785, 2008.

[60] R. Lister and et al., "Human dna methylomes at base resolution show widespread epigenomic differences." *Nature*, vol. 462, pp. 315–322, 2009.

[61] A. Meissner and et al., "Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis." *Nucleic Acids Research.*, vol. 33, pp. 5868–5877, 2005.

[62] R. Harris and et al., "Comparison of sequencing-based methods to profile dna methylation and identification of monoallelic epigenetic modifications." *Nature Biotechnology*, vol. 28, pp. 1097–1105, 2010.

[63] Z. Smith and et al., "High-throughput bisulfite sequencing in mammalian genomes." *Methods*, vol. 48, pp. 226–232, 2009.

[64] C. Bock and et al., "Quantitative comparison of genome-wide dna methylation mapping technologies." *Nature Biotechnology*, vol. 28, pp. 1106–1114, 2010.

[65] H. Gu and et al., "Genome-scale dna methylation mapping of clinical samples at single-nucleotide resolution." *Nature Methods*, vol. 7, pp. 133–136, 2010.

[66] R. Lister and et al., "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells." *Nature*, vol. 471, pp. 68–73, 2011.

[67] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society, Series B*, vol. 57, pp. 289–300, 1995.

[68] J. D. Storey, "A direct approach to false discovery rates." *Journal of the Royal Statistical Society, Series B*, vol. 64, pp. 479–498, 2002.

[69] ——, "The positive false discovery rate: a bayesian interpretation and the q-value." *Annals of Statistics*, vol. 31, pp. 2013–2035, 2003.

[70] J. D. Storey and R. Tibshirani, "Statistical significance for genome-wide studies." *Proceedings of the National Academy of Sciences*, vol. 100, pp. 9440–9445, 2003.

[71] B. Efron, R. Tibshirani, J. Storey, and V. Tusher, "Empirical bayes analysis of a microarray experiment." *Journal of the American Statistical Association*, vol. 96, pp. 1151–1160, 2001.

[72] B. Efron, "Size, power and false discovery rates." *The Annals of Statistics*, vol. 35, pp. 1351–1377, 2007.

[73] J. Liao, Y. Lin, Z. Selvanayagam, and J. Weichung, "A mixture model for estimating the local false discovery rate in dna microarray analysis." *Bioinformatics*, vol. 20, pp. 2694–2701, 2004.

[74] D. B. Allison, G. L. Gadbury, M. Heo, J. R. Fernandez, C. K. Lee, T. A. Prolla, and R. Weindruch, "A mixture model approach for the analysis of microarray gene expression data." *Computational Statistics and Data Analysis*, vol. 35, pp. 1–20, 2002.

[75] R. E. Tarone, "A modified bonferroni method for discrete data." *Biometrics*, vol. 46, p. 515, 1990.

[76] P. Gilbert, "A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeciency virus genetics." *Journal of the Royal Statistical Society, Series C*, vol. 54, pp. 143–158, 2005.

[77] S. Pounds and C. Cheng, "Robust estimation of the false discovery rate." *Bioinformatics*, vol. 22, 2006.

[78] J. Heyse, "A false discovery rate procedure for categorical data." *In Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, pp. 43–58, 2011.

[79] F. Krueger and et al., "Dna methylome analysis using short bisulfite sequencing data." *Nature Methods*, vol. 9, pp. 145–151, 2012.

[80] G. Wu and et al., "Statistical quantification of methylation levels by next-generation sequencing." *PLoS ONE*, vol. 6, 2011.

[81] C. Wu, "On the convergence properties of the em algorithm." *Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[82]

[83] K. Basford and G. McLachlan, "Estimation of allocation rates in a cluster analysis context." *Journal of the American Statistical Association*, vol. 80, pp. 286–293, 1985.

[84] W. Sun and T. Cai, "Oracle and adaptive compound decision rules for false discovery rate control." *Journal of the American Statistical Association*, vol. 102, pp. 901–912, 2007.

[85] R. Merling and et al., "Transgene-free ipscs generated from small volume peripheral blood nonmobilized cd34+ cells." *Blood*, vol. 121, pp. 98–107, 2013.

[86] D. Baumann and R. Doerge, "Issues in testing dna methylation using next-generation sequencing." *The Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture. Manhattan, KS*, p. 1, 2011.

[87] ——, "Magi: methylation analysis using genome information." *Epigenetics*, vol. 9, 2014.

[88] H. Zheng, H. Wu, J. Li, and S.-W. Jiang, "Cpgimethpred: computational model for predicting methylation status of cpg islands in human genome." *BMC Medical Genomics*, vol. 6, 2013.

[89] W. Zhang, T. Spector, P. Deloukas, J. Bell, and B. Engelhardt, "Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements." *Genome Biology*, vol. 16, 2015.

[90] A. Prochenka and et al., "A cautionary note on using binary calls for analysis of dna methylation." *Bioinformatics*, vol. 31, 2015.

[91] N. Plongthongkum, D. Diep, and K. Zhang, "Advances in the profiling of dna modifications: cytosine methylation and beyond." *Nature*, vol. 15, 2014.

[92] J. Wang and et al., "Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide cpg methylation profiling by reduced representation bisulfite sequencing." *BMC Genomics*, vol. 11, 2013.

[93] P. Boyle and et al., "Gel-free multiplexed reduced representation bisulfite sequencing for large-scale dna methylation profiling." *Genome Biology*, vol. 13, 2012.

[94] M. Schillebeeckx and et al., "Laser capture microdissection reduced representation bisulfite sequencing (lcm-rrbs) maps changes in dna methylation associated with gonadectomy-induced adrenocortical neoplasia in the mouse." *Nucleic Acids Research*, vol. 41, 2013.

[95] H. Guo and et al., "Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing." *Genome Research*, vol. 23, p. 2126, 2013.

[96] A. Adey and J. Shendure, "Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing." *Genome Research*, vol. 22, p. 1139, 2012.

[97] F. Miura and et al., "Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging." *Nucleic Acids Research*, vol. 40, 2012.

[98] M. J. Booth and et al., "Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine." *Nature Protocols*, vol. 8, p. 1841, 2013.

[99] A. Petterson and et al., "Rrhp: a tag-based approach for 5-hydroxymethylcytosine mapping at single-site resolution." *Genome Biology*, vol. 15, p. 456, 2014.

[100] M. Yu and et al., "Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome." *Cell*, vol. 149, p. 1368, 2012.

[101] J. Eid and et al., "Real-time dna sequencing from single polymerase molecules." *Science*, vol. 323, p. 133, 2009.

[102] B. A. Flusberg and et al., "Direct detection of dna methylation during single-molecule, real-time sequencing." *Nature Methods*, vol. 7, p. 461, 2010.

[103] Y. Yang and et al., "Quantitative and multiplexed dna methylation analysis using long-read single-molecule real-time bisulfite sequencing (smrt-bs)." *BMC Genomics*, vol. 16, p. 350, 2015.

[104] X. Chen and R. Doerge, "Towards better fdr procedures for discrete test statistics." *The Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture. Manhattan, KS*, p. 294, 2012.

[105] ——, "A weighted fdr procedure under discrete and heterogeneous null distributions." *ArXiv e-prints, http://arxiv.org/pdf/1502.00973v2.pdf.*, Last modified: Oct 19th, 2015.

[106] R. Heller and H. Gur, "False discovery rate controlling procedures for discrete tests." *ArXiv e-prints, http://arxiv.org/abs/1112.4627.*, 2012.

[107] Y. Benjamini and W. Liu, "A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence." *Journal of Statistical Planning and Inference*, vol. 82, p. 163, 1999.

[108] T. Wu and et al., "Fast and snp-tolerant detection of complex variants and splicing in short reads." *Bioinformatics*, vol. 26, pp. 873–881, 2010.

[109] S. Yuan and et al., "Read-mapping using personalized diploid reference genome for rna sequencing data reduced bias for detecting allele-specific expression." *BIBM 2012 Workshop on Data-Mining of Next Generation Sequencing*, 2013.

VITA

VITA

Longjie Cheng was born and raised in Wuhan, China. She majored in Statistics in college, and received her bachelor degree in 2010 from Wuhan University. She then joined Department of Statistics at Purdue University in 2010.