A MODIFIED INFORMATION CRITERION IN THE 1D FUSED LASSO FOR DNA COPY NUMBER VARIANT DETECTION USING NEXT GENERATION SEQUENCING DATA

By

Jaeeun Lee

Submitted to the Faculty of the Graduate School
of Augusta University in partial fulfillment
of the Requirements of the Degree of
Doctor of Philosophy

August
2017

ACKNOWLEDGEMENTS

I am heartily thankful to my advisor Dr. Jie Chen. I was fortunate to have had her valuable guidance and support in both research and life throughout the years I worked on my Ph.D. Her patience and knowledge have helped me learn how to carry out solid research. I could not have finished my dissertation without her.

I would like to express my gratitude to my committee members, listed below, for their unwavering support and sincere feedback. I am indebted to them for their time and effort in helping me succeed.

· Dr. Varghese George

· Dr. Santu Ghosh

· Dr. Hongyan Xu

· Dr. Xiaoling Wang

Dr. Stephan Looney has also been an important supporter during my time at Augusta University. I would like to thank him deeply for his help and encouragement of my academic progress over the past five years.

I would like to thank my classmates Jeannie and Taejin for their friendship throughout our parallel academic journeys.

Finally, I am grateful to my family for their help and support.

ABSTRACT

JAEEUN LEE
A Modified Information Criterion in the 1d Fused Lasso for DNA Copy Number Variant Detection using Next Generation Sequencing Data
(Under the direction of Dr. JIE CHEN)

DNA Copy Number Variations (CNVs) are associated with many human diseases. Recently, CNV studies have been carried out using Next Generation Sequencing (NGS) technology that produces millions of short reads. With NGS reads ratio data, we use the 1d fused lasso regression for CNV detection. Given the number of copy number changes, the corresponding genomic locations are estimated by fitting the 1d fused lasso. Estimation of the number of copy number changes depends on a tuning parameter in the 1d fused lasso. In this dissertation, we propose a new modified Bayesian information criterion, called JMIC, to estimate the optimal tuning parameter in the 1d fused lasso. In theoretical studies, we prove that the number of change points estimated by JMIC converges the true number of changes. Also, our simulation studies show that JMIC outperforms the other criteria considered. Finally, we apply our proposed method to the reads ratio data from the breast tumor cell HCC1954 and its matched cell line provided by Chiang et al. (2009).

KEY WORDS: Copy Number Variation (CNV), Next Generation Sequencing (NGS) reads ratio data, change point analysis, fused lasso, Bayesian information criterion, path algorithm

Table of Contents

List of Tables

List of Figures

1 Introduction

Copy number variations (CNVs) are associated with several complex diseases, such as autism spectrum disorder (Chung et al., 2014; Pinto et al., 2010), schizophrenia (Stone et al., 2008), obesity (Jacquemont et al., 2011), Alzheimer's (Rovelet-Lecrux et al., 2006), prostate (Yan et al., 2015) and breast cancers(Stephens et al., 2009), etc. Deletions and duplications of DNA segments greater than 50 base pairs are called copy number variations (MacDonald et al., 2014; Zarrei et al., 2015). Over the past decade, DNA CNVs have been studied using Next Generation Sequencing (NGS) that produces millions of short sequence reads. NGS allows us to achieve higher resolutions and more accurate CNV detection than earlier sequencing technologies such as array comparative genomic hybridization (aCGH) (Alkan et al., 2011; Chiang et al., 2009).

Once NGS reads are aligned to the genomic positions along chromosomes, copy number changes can be detected using multiple change point analysis that identifies the number of change points and their locations. However, in NGS CNV studies, it is computationally infeasible to search multiple copy number changes among thousands of genomic locations. Moreover, CNVs may occur on a relatively small number of genomic regions (Zhang et al., 2009). This property is called sparsity (i.e., a relatively small number of variables are meaningful) (Hastie et al., 2015). Since CNVs are sparse, l-1 norm penalized regression methods, also called lasso-based methods, have recently been applied to the CNV detection problem. Tibshirani and Taylor (2011) introduced the lasso model for a 1-dimensional data structure, called the 1d fused lasso. The l-1norm penalty term in the 1d fused lasso forces sparsity of the differences between neighboring coefficients. Therefore, it has special abilities such as: 1) encouraging neighboring genomic regions to have similar copy numbers, which is desirable in CNV detection because copy numbers have an inherent structure along chromosomes; 2) forcing some of the similar copy numbers to be exactly the same, which is suitable in change point analysis because exact locations of copy number changes can be identified. Given the number of copy number changes, their locations are then estimated by fitting the 1d fused lasso using the dual path algorithm proposed by Tibshirani and Taylor (2011).

However, the main challenge is to determine the number of change points. Although several scholars have applied lasso-based methods to the CNV detection problem, they have not yet intensely focused on how to determine the number of copy number changes. The estimated number of change points in the 1d fused lasso depends on which tuning parameter value is selected. Therefore in this dissertation, to estimate the optimal tuning parameter in the 1d fused lasso, we propose a modified Bayesian information criterion in terms of multiple change point problems. Then, we show theoretically that our proposed criterion consistently identifies the true number of change points.

The rest of the dissertation is organized as follows. Chapter 2 reviews DNA CNV detection methods using aCGH and NGS data in Section 2.1 and 2.2, respectively. As statistical CNV detection approaches, Section 2.3 presents change point detection methods. Section 2.4 covers lasso-based approaches to multiple change point problems. Section 2.5 presents optimization methods for fitting lasso-based models. Section 2.6 reviews tuning parameter selection methods in lasso-based models. In Chapter 3, we describe our approach to the CNV detection problem using NGS reads ratio data. Under the multiple change point model in Section 3.1, we introduce the 1d fused lasso in Section 3.2 to identify change points. Section 3.3 illustrates an optimization method, called the dual path algorithm, for fitting the 1d fused lasso. In Section 3.4, we propose a new modified information criterion, called JMIC, for selecting the tuning parameter in the 1d fused lasso. Chapter 4 derives theoretical properties of JMIC. Our simulation studies and real data applications are given in Chapter 5 and 6, respectively. Finally, we finish with a discussion in Chapter 7.

2 Literature review

2.1 CNV detection in aCGH

DNA Copy number variations (CNVs) have been studied using genomic data from array-based Comparative Genomic Hybridization (aCGH) technology. In a general aCGH process, genomic DNA segments first are extracted from test and control cells. The test and control DNA samples are differentially labeled with two fluorescent dyes. The labeled test and control samples are mixed and then competitively applied to a microarray slide. On the microarray slide, thousands of tiny spots are arrayed in a specific genomic order. Each spot represents a specific probe and consists of information about its DNA sequence and genomic location. Millions of copies of single-stranded DNA fragments are then built up on each spot. Because the mixed samples are single-stranded, when applied to the microarray chip, the test and control samples competitively attempt to bind to their corresponding single-strands and form two-stranded DNA. This process is called hybridization. If the tumor fragment contains DNA segments with duplication (or deletion), it is more (or less) likely to bind to its probe. Following the hybridization process, the array slide is placed in a scanner. Through scanning, the ratio of the fluorescence intensities between the two samples is calculated to detect CNVs. The array CGH intensity ratio is roughly proportional to the ratio of DNA copy numbers between the two samples (Feuk et al., 2006; Pinkel et al., 1998; Pollack et al., 1999).

Many CNV detection methods with aCGH data have been developed using likelihood ratio (Chen and Wang, 2009; Olshen et al., 2004), smoothing function (Eilers and De Menezes, 2005; Hsu et al., 2005; Huang et al., 2005), Bayesian (Baladandayuthapani et al., 2012; Chen et al., 2010), and Hidden Markov Model (Fridlyand et al., 2004; Guha et al., 2008; Sebat et al., 2004), etc. Lai et al.

(2005) and Willenbrock and Fridlyand (2005) reviewed existing CNV detection methods using array-based data.

However, although many researchers have used aCGH data in CNV studies, since the distances between genomic positions of probes are predefined on the array, the precision of boundaries of copy number changes is limited at relatively low resolution (>1000 base pairs of DNA, Shen and Zhang (2012)). More recently, next-generation sequencing (NGS) technology allows us to achieve much higher resolution and more accurate CNV detection. With NGS data, boundaries of CNVs can be precisely identified even at the single-base-pair level (Alkan et al., 2011; Chiang et al., 2009).

2.2 CNV detection in NGS

Over the past decade, next generation sequencing (NGS) data have been widely used in CNV detection studies. We describe NGS data processing in Section 2.2.1 and introduce CNV detection methods with NGS data in Section 2.2.2.

2.2.1 NGS technology

In NGS processing, a library is first constructed. During preparation of the library, a DNA sample is randomly fragmented, and adapters are attached to both ends of each DNA fragment. Then, the library fragments bind to a solid surface, either a bead or a glass slide, depending on the platform. Each fragment is amplified to produce millions of copies. This amplification step is required to allow us to detect the reaction signal in the following step. During the sequencing process, we sequence each of the millions of amplified copies simultaneously, which includes an imaging step to detect each nucleotide base one at a time until we generate the full sequence run along the length of the fragment. Finally, short-base pair sequenced fragments, called reads, are obtained. Short reads can be generated from sequencing either only one end or both ends of a fragment. Compared with single-end reads, paired-end sequencing improves read alignment of repetitive regions and increases the depth of coverage across a genome (Abyzov et al., 2011; Maher et al., 2009). In the last step, we align the reads to the reference genome to identify the true genomic locations of the reads. Then we are ready to analyze the reads data for CNV detection. For more details on NGS technologies, see Metzker (2010) and Mardis (2013).

2.2.2 CNV detection methods using NGS data

Early CNV studies with NGS data used paired-mapping methods (Korbel et al., 2007) that rely on only paired-end read data. To detect CNVs, paired-mapping methods are based on a distance between the two ends of aligned paired reads, also known as insert size. For example, if the observed insert size from a test genome is significantly longer (or shorter) than the expected insert size from the reference genome, insertion (or deletion) may occur in the genomic region.

CNV detection also has been carried out using read-count methods that can use paired end-reads as well as single-end reads. Read-count methods assume that the number of reads of a genomic region is approximately proportional to the copy number of that region. Then, a relatively small (or large) number of reads indicates a copy number loss (or gain) for any genomic region, compared to the normal cell. In contrast to paired-end mapping methods, read-count approaches not only estimate copy numbers but also less suffer from segmental duplications (Tuzun et al., 2005; Yoon et al., 2009). In addition, a paired-end mapping method cannot detect CNVs larger than its insert size because the insert size limits the maximum detectable size for CNVs (Yoon et al., 2009).

When reads are randomly sampled from a genome, read count data can be assumed to follow a Poisson distribution where the mean and variance are assumed to be equal. However, a Poisson distribution may be unsuitable for modeling read counts. This is because the variance of read counts is usually greater than their mean (i.e., overdispersion) due to systematic biases, such as GC (Guanine and Cytosine) content and mappability, etc (Cheung et al., 2011). For this reason, many studies have assumed other probabilistic distributions (negative binomial (Miller et al., 2011); nonhomogeneous poisson (Shen and Zhang, 2012); normal distributions (Yoon et al., 2009)) or used nonparametric approaches (Abyzov et al., 2011; Kim et al., 2010; Xi et al., 2011).

Miller et al. (2011) observed that a negative binomial model with the variance/mean ratio of 3 fits the read count data well in the Yoroban genome (Bentley et al, 2008). Shen and Zhang (2012) used two nonhomogeneous Poisson processes for test and control samples and detected CNVs based on the exact binomial generalized likelihood ratio statistic. Yoon et al. (2009) observed that read counts of 100-bp windows approximately follows a normal distribution at 30× coverage. Based on the standard normal distribution, they calculated a p-value for each read count data and tested a genomic region by comparing the maximum p-value with the corrected significance level. In nonparametric approaches, Abyzov et al. (2011) estimated a distribution of read counts using a two-dimensional Gauissian kernel. They identified copy number boundaries based on mean-shift vectors with a multiple-bandwidth partitioning strategy. Kim et al. (2010) assigned positive and negative weights to tumor and control reads, respectively and summed the weights by scanning along a chromosome. The large partial positive (or negative) cumulative sum indicates a local copy number gain (or loss). Xi et al. (2011) also used a nonparametric approach by constructing the joint likelihood of the indicator of being a tumor read and its genomic position.

Another read count approach is to use the ratio of read counts between tumor and its matched normal samples from the same patient, when both samples are available. Similar to fluorescence intensity ratios in aCGH data, read ratios represent DNA copy number ratios between the two samples. For any genomic region, a read ratio of 1 indicates no copy number change. A read ratio that deviates from 1 indicates that CNVs may occur on that region. By using the ratio of reads, GC content can be normalized because the paired tumor and normal samples may have similar GC content variations (Chiang et al., 2009; Klambauer et al., 2012). Chiang et al. (2009) derived that copy number ratio data approximately follow a normal distribution and proposed the test statistic for the log ratio of a difference in reads between neighboring regions. Xie and Tammi (2009) also used read ratio data for CNV detection and assumed a normal distribution. See Magi et al. (2012), Teo et al. (2012), and Zhao et al. (2013) for a general review of CNV detection methods with NGS data.

In this dissertation, we focus on NGS read count ratio data for CNV detection. Also, note that we do not go into data processing nor bias normalization. We focus on a segmentation step that estimates the number of copy number changes and their corresponding locations. The next section 2.3 presents a statistical framework to achieve this objective.

2.3 Change point methods for CNV detection

The purpose of CNV detection is to find genomic regions of a copy number gain or loss on chromosomes. After filtering out the genomic errors, a change point detection method identifies the boundaries between neighboring genomic regions where copy numbers change.

2.3.1 Change point detection methods

Change point analysis first appeared in the work of Page (1954) for quality control in a manufacturing process. Since then, it has widely become popular in various fields including finance (Chen and Gupta, 1997), climatology (Reeves et al., 2007), brain signal in neuroscience (Kirch et al., 2015), earthquake detection (Alawadhi and Alhulail, 2016), etc. The scope of change point analysis falls into two categories: online and offline. Online detection continuously observes data and tracks a change point, whereas offline analysis detects change points once all of the data are collected. Since all genomic data are observed before CNV detection analysis in this dissertation, we focus on the offline change point analysis. The (offline) change point analysis wishes to (1) decide if any change points occur and if so, (2) estimate the number of changes and their corresponding locations (Chen and Gupta, 2011).

Traditional change point detection methods are based on a likelihood ratio function to test whether a parameter changes at an unknown point. Hinkley (1970) first proposed the likelihood ratio test for a single mean change. He derived the asymptotic null distribution of the test statistic, assuming a normal distribution. The likelihood ratio approach to a single change problem has also been applied under other distributions, such as binomial (Hinkley and Hinkley, 1970), gamma (Hsu, 1979), and exponential (Haccou et al., 1987) distributions.

Another change point detection approach is based on an information model selection criterion. To detect a single change point, one can identify a possible change point by choosing between the model of no change point and the model of one change point at a possible location. Hence, a change point detection problem can be viewed as a model selection problem. A typical approach to model selection is to use an information criterion, such as AIC (Akaike, 1998) and SIC (Schwarz et al., 1978), based on the principle of minimum information criterion. Hsu (1979) performed the change point analysis by using SIC in stock market price data. Chen and Gupta (1999) tested a single mean and variance change point based on SIC in a sequence of independent normal random variables.

In real situations, however, data structures often change at more than one location, which is a multiple change point problem. When the number of observations in a sequence is large, the number of change points might also become large. Thus, it is computationally infeasible to examine all possible sets of the change point locations. To search multiple changes in an efficient way, Scott and Knott (1974) proposed the binary segmentation (BS) procedure that hierarchically identifies change point locations using the single change point analysis such as a likelihood ratio test or an information criterion approach. Given a sequence of observed data, if a single change point is detected, the sequence is divided into two subsequences. Then for each subsequence, another detected point splits the subsequence into two (sub) subsequences. The process recursively continues until no more changes are detected in any of the subsequences. Venkatraman (1992) derived the consistency of the BS procedure. Chen and Gupta (1997) detected multiple change points by combining the BS procedure with SIC (Schwarz et al., 1978).

A penalized regression approach has also been used in multiple change point analysis. Given the number of change points, a penalized regression model with the l-1 norm penalty estimates the locations of change points via an optimization procedure. When the number of change points is unknown, the number of change points is determined by selecting a tuning parameter. Levy-leduc and Harchaoui (2008) and Harchaoui and Lévy-Leduc (2012) first introduced the l-1 norm penalized regression approach for multiple change point estimation. They viewed a multiple change point problem as a variable selection problem. We focus on l-1 norm penalized regression approaches in this dissertation and will discuss in more detail in Section 2.4.

Change point problems have also been investigated using various statistical methods including cumulative sum type (Cho and Fryzlewicz, 2015), Bayesian (Du et al., 2016; Green, 1995), and nonparametric approaches (Loader et al., 1996; Pettitt, 1979), etc. For a general review of parametric change point analysis, see Chen and Gupta (2011).

2.3.2 CNV detection using change point methods

In aCGH data, Olshen et al. (2004) first applied the BS procedure to CNV detection, called the CBS (Circular Binary Segmentation). Many CNV detection analyses have been influenced by the CBS. Olshen et al. (2004) modified the likelihood ratio based test statistic given by Sen and Srivastava (1975) and computed critical values using approximate tail probabilities given by Siegmund (1986) or using Monte Carlo simulations. They also found that log intensity ratios fluctuate due to a biological reason, leading to false positive detection of copy number changes. To deal with this problem, a pruning procedure was used after the CBS procedure. Instead of using the likelihood ratio test approach in the BS procedure, Chen and Wang (2009) proposed the test statistic based on SIC, assuming a normal distribution. They derived approximate p-values of the test statistic and tested a single copy number change at a time in the BS procedure. Zhang and Siegmund (2007) modified SIC to determine the number of copy number changes and used a modified version of the CBS to localize the changes along a chromosome. Du et al. (2016) applied a Bayesian approach in a change point model framework to CNV detection. They investigated a prior distribution of change point locations and derived the asymptotic properties of the maximum marginal likelihood estimator.

In NGS read counts data, Miller et al. (2011) applied the CBS to CNV detection. Shen and Zhang (2012) proposed a change point model on inhomogeneous poisson processes for NGS read data. They adopted the CBS to search multiple change points and applied the modified SIC given by Zhang and Siegmund (2007) to choose the number of change points. Ji and Chen (2015) also analyzed NGS read counts based on a poisson change point model. They proposed a Bayesian approach and used a moving window algorithm given by Chen et al. (2010) to search multiple copy number changes.

Since high-throughput genomic data have been exploded in recent years, penalized regression approaches, especially l-1 norm penalized regression approaches, have also become popular in CNV detection studies with aCGH data (Eilers and De Menezes, 2005; Huang et al., 2005; Jong et al., 2003; Li and Zhu, 2007; Tibshirani and Wang, 2008) and NGS read data (Boeva et al., 2011; Duan et al., 2013). The next section briefly reviews l-1 penalized regression methods and their applications to CNV analysis. We refer to l-1 penalized regression methods as Lasso-based approaches in this dissertation.

2.4 Change point analysis using Lasso-based approaches

2.4.1 A brief introduction to Lasso-based approaches

The “Lasso" (Least Absolute Shrinkage and Selection Operator, Tibshirani (1996)) is a regression method that constraints the ordinary least-squares estimator by adding the l-1 penalty term. The l-1 norm penalty has a special property. It forces some coefficients to be exactly zero so that the Lasso performs variable selection with a sparse solution (i.e., it produces a relatively small number of variables that are meaningful). Because of sparsity of the l-1 penalty, Lasso-based approaches have received great attention in large data problems. Greenshtein et al. (2004) first used the lasso approach when the number of predictors is greater than the number of sample size. In such high-dimensional data, predictors tend to be highly correlated in groups. To deal with the correlated predictors, Zou and Hastie (2005) proposed the elastic net penalty that takes advantage of both the l-1 and l-2 penalties. For the situation where predictors are structurally grouped, Yuan and Lin (2006) introduced the group lasso that shrinks and selects groups together. Simon et al. (2013) extended the group lasso, called the sparse group lasso, to identify important variables within selected groups. Zou (2006) showed that the original lasso is inconsistent for variable selection in certain situations and proposed the adaptive lasso where coefficients are assigned to different weights according to their size. Park and Casella (2008) proposed the Bayesian lasso using a conditional Laplace prior.

In change point analysis, Harchaoui and Lévy-Leduc (2012) viewed a multiple change point problem as a variable selection problem and used the lasso model in a sequence of independent normal random variables. They demonstrated that the coefficient's estimates are consistent, but the estimates of change point locations are not consistent. Qian and Jia (2016) confirmed this inconsistency and provided a necessary and sufficient condition. Chan et al. (2014) focused changes in correlated structures using the group Lasso. Ciuperca (2014) considered different sub-models using the adaptive Lasso. In change point problems, observations are typically sequenced in a specific order. Tibshirani et al. (2005) designed the lasso model for sequence data and proposed the "fused lasso" that penalizes coefficients themselves as well as their neighboring differences. The penalty term for the differences between adjacent coefficients makes some similar coefficients identical, leading a piecewise constant model. Thus, the fused lasso is able to not only produce a sparse model for sequence data but also capture abrupt changes at the same time. For this reason, the fused lasso is an attractive method for change point analysis in high-dimensional data. Tibshirani and Taylor (2011) generalized the lasso model with a penalty matrix D. By the choice of D, the generalized lasso model becomes several shrinkage models such as the fused lasso, trend filtering, and wavelet smoothing, etc. It is worth mentioning that Tibshirani et al. (2005) included the two l-1 penalty terms for coefficients themselves and their neighboring differences, whereas Tibshirani and Taylor (2011) included the l-1 penalty only for the differences between neighboring coefficients and referred to their model as the “1d fused lasso". In this dissertation, we will use the 1d fused lasso to solve our multiple change point problem. In fact, the 1d fused lasso is known as a total variation (Rudin et al., 1992) in signal processing.

2.4.2 CNV detection using Lasso-based approaches

In CNV detection, as mentioned in the Introduction, neighboring genomic regions may contain similar copy numbers. Several studies have considered this structural dependency of copy numbers and applied the lasso-based regression models to CNVs detection. In aCGH data, Huang et al. (2005) constructed a regression model with the $l1$ penalty for differences in copy numbers between adjacent markers. To obtain a p-value and false discovery rate to decide whether a segment has a significant CNV, they proposed to use the stationary bootstrap. Tibshirani and Wang (2008) used the fused lasso of Tibshirani et al. (2005) to detect copy number alterations in aCGH data. They estimated the tuning parameters from smoothed observations using lowess and calculated the false-discovery rate assuming that the null distribution follows the standard normal. The optimization algorithm 'sqopt' of Gill et al. (1997) was used to fit the fused lasso. They compared the fused lasso with CGHseg (Picard et al., 2005), CBS (Olshen et al., 2004), and CLAC (Wang et al., 2005) and concluded that the fused lasso performs better than the other methods. Eilers and De Menezes (2005) and Li and Zhu (2007) also considered the structural dependency of copy numbers in aCGH data and used the fused quantile regression model. In NGS read counts data, Boeva et al. (2011) and Duan et al. (2013) applied the lasso-based model proposed by Levy-leduc and Harchaoui (2008) to the CNV detection problem.

In change point analysis along with Lasso-based regression, the locations of change points are identified by fitting the model. The next Section reviews optimization methods for fitting Lasso-based regression models.

2.5 Optimization methods for fitting Lasso-based models

Many optimization algorithms have been developed to solve Lasso-based regression problems. For the original lasso problem, the well-known optimization method is the Least Angle Regression (LARS) algorithm proposed by Efron et al. (2004). As the tuning parameter in the original lasso decreases, the LARS constructs the entire path of lasso coefficient estimates by adding a variable that is highly correlated with the residual. Also, since the optimal solution path is piecewise linear in the tuning parameter, it efficiently computes the solution. The LARS was extended by Rosset and Zhu (2007) for general regularization problems. Wu (2011) and Zhou and Wu (2014) also applied the LARS to generalized linear models based on an ordinary differential equation system.

Another optimization method for fitting lasso problems is pathwise coordinate descent proposed by Friedman et al. (2007). Rather than producing the entire solution path like the LARS, the pathwise coordinate descent algorithm computes solutions over pre-defined tuning parameter values. Given a tuning parameter value, the algorithm estimates each parameter while holding the other parameters fixed. This univariate optimization procedure continues over all parameters until convergence. Friedman et al. (2010) extended the pathwise coordinate descent algorithm to generalized linear models with elastic-net penalty. Mazumder et al. (2012) developed the path coordinate descent algorithm for nonconvex objective functions and investigated its convergence properties.

For fused lasso problems, Harchaoui and Lévy-Leduc (2012) used the LARS by transforming parameters. Although this reparameterized fused lasso can be solved by the LARS, Hastie et al. (2015) pointed out that this approach is inefficient because a design matrix in the transformed parameters will lead to high correlations among the parameters. The coordinate descent algorithm also has a limitation on the fused lasso. Tseng (2001) showed the global convergence properties of the coordinate descent algorithm. However, this result for convergence does not hold when the penalty function is not separable in each individual parameter. Unlike the lasso, the fused lasso has this issue. Therefore, a fused lasso solution by the pathwise coordinate descent algorithm may not be optimal (Friedman et al., 2007). To deal with this problem, Tibshirani and Taylor (2011) suggested a dual approach where the penalty function becomes separable. They proposed the dual path algorithm to solve the 1d fused lasso based on the Lagrange dual problem. The dual solution path is a piecewise linear function of the tuning parameter so that coordinates are adaptively split by moving the tuning parameter from ∞ to 0. In this dissertation, we use the dual path algorithm to fit the 1d fused lasso.

2.6 Tuning parameter selection in Lasso-based models

The path algorithms in Section 2.5 produce entire solutions at critical tuning parameter values. The total number of critical tuning parameters is $n - 1$, where n is the number of observations in a sequence. In the 1d fused lasso problem with the path algorithm, the first critical tuning parameter value corresponds to the model of no change. The second one corresponds to the model of a one change. The last one corresponds to the model at which all observations are change points. Thus, the detected number of change points depends on which tuning parameter value is selected among $n - 1$ critical values. When the tuning parameter is large (or small), the estimated number of change points is small (or large), leading a sparse (or dense) model. Thus, a tuning parameter selection problem becomes a model selection problem. In Subsection 2.6.1, we review model selection approaches to tuning parameter selection in penalized regression models and their applications to CNV detection.

2.6.1 Tuning parameter selection for penalized regression models

Traditional model selection methods are Akaike Information Criterion (AIC, Akaike (1998)), Bayes Information Criterion (SIC, Schwarz et al. (1978)), and Cross-validation (Stone, 1974). AIC evaluates a model in terms of Kullback-Leibler information between a fitted model and the true distribution. Cross-validation first splits data into a training set and a test set. It fits a model to the training set and predicts the test set using the fitted model. It is well known that the model selected by AIC is asymptotically equivalent to the model by cross-validation (Stone, 1977).

In penalized regression methods, the performance of penalized regression models heavily relies on the choice of tuning parameter. We can obtain the true model consistently when an appropriate tuning parameter is chosen (Fan and Li, 2001; Wang et al., 2009; Wang and Zhu, 2011). In the smoothly clipped absolute deviation (SCAD) penalized regression model, Wang et al. (2007) showed that AIC and generalized cross-validation (Craven and Wahba, 1978) tend to yield an overfitted model, whereas SIC approach consistently identifies the true model. Similar results are also identified by Zhang et al. (2010) in the generalized linear model with a nonconcave penalty function. Although SIC is consistent for model selection, the consistency of SIC may not be guaranteed in high dimensional settings because of overfitting (Chen and Chen, 2008; Fan and Tang, 2013; Siegmund, 2004). Chen and Chen (2008) modified SIC for high dimensional situations. They proved the consistency of the modified SIC under the normality assumption and applied it to the Lasso and SCAD penalized model. In ultra-high dimensional settings where the number of parameters exponentially increases as the sample size increases, Wang and Zhu (2011) extended the modified SIC by Chen and Chen (2008), and Fan and Tang (2013) proposed a modified generalized information criterion (Nishii et al., 1984) for a generalized penalized regression model.

In a change point model framework, Chen et al. (2006) modified SIC for a single change point problem, assuming that the true change point is located in the middle of a sequence. They showed that the limiting null distribution of their test statistic follows a chi-squares distribution and provided the best convergence rate for change point locations. Pan and Chen (2006) extended the criterion by Chen et al. (2006) to multiple change point problems and proved its consistency, assuming that true change points are uniformly distributed. Zhang and Siegmund (2007) proposed a modified SIC using Brownian motion with changing drift under a normal distribution and showed that the modified SIC is more robust than the original BIC.

2.6.2 Tuning parameter selection for CNV detection

For CNV detection using aCGH data, Huang et al. (2005) determined a tuning parameter value based on a specified threshold in the lasso problem. Eilers and De Menezes (2005) used cross-validation, while Li and Zhu (2007) used SIC for tuning parameter selection in a fused quantile regression. Tibshirani and Wang (2008) estimated the tuning parameters in the fused lasso from smoothed observations using a lowess method. In NGS read data, Boeva et al. (2011) used the penalized residual sum of squares to select the number of copy number changes. Duan et al. (2013) used SIC in the 1d fused lasso and showed that SIC gives more robust results than the criterion by Boeva et al. (2011).

Although several scholars have used lasso-based approaches to CNV detection, they have not yet intensely focused on tuning parameter selection. In this dissertation, we modify SIC to select the tuning parameter in the 1d fused lasso and apply our method to CNV detection.

3 Methodology

CNV detection can be viewed as a multiple change point detection problem. We introduce the multiple change point model and the 1d fused lasso in Section 3.1 and 3.2, respectively. Given the number of changes, we identify the locations of change points by fitting the 1d fused lasso using the dual path algorithm in Section 3.3. Then, we estimate the number of change points by selecting the tuning parameter in the 1d fused lasso. To estimate the optimal tuning parameter, we propose a modified Bayesian information criterion, called JMIC, in Section 3.4.

3.1 The multiple change point model

Suppose that we have read count data from the test and its matched normal samples from the same patient. Let y_i denote the $\log_2$ ratio of the number of reads between the two samples at the ith bin along a chromosome, $i = 1, ..., n$. The sequence of $y_1, ..., y_n$ is then split into non-overlapping $K+1$ segments by K change points such that

$$y_i = \beta_i + \varepsilon_i, \qquad i = 1, ..., n, \tag{3.1}$$

where $\beta_i = \mu_k$ for $i = t_{k-1}+1, ..., t_k$, $k = 1, ..., K+1$. The $t_1, ..., t_K$ are change point locations and ordered as $0 = t_0 < t_1 < t_2 < \cdots < t_K < t_{K+1} = n$. (3.1) is known as the multiple changes point model (Yao and Au, 1989). Our aim is then to estimate K and $t_1, ..., t_K$. The coefficient β_i represents the $\log_2$ ratio of the true copy numbers, and the ε_i is its random error. Chiang et al. (2009) showed that the $\log_2$ read ratio data approximately follow a normal distribution. In this dissertation, we assume that

$\varepsilon_1,...,\varepsilon_n$ are i.i.d. normal random variables with zero mean and finite variance σ^2. In this paper, the locations $t_1,...,t_K$ and the number K of change points are unknown and need to be estimated.

3.2 The 1d fused lasso for multiple change point problems

Let $\mathbf{y} \in \mathbb{R}^{n\times 1}$ be a response vector and $\beta \in \mathbb{R}^{n\times 1}$ be a vector of regression coefficients. Each observation y_i corresponds to the coefficient β_i, $i = 1,...,n$. Let $\mathbf{D} \in \mathbb{R}^{(n-1)\times n}$ be a penalty matrix. The 1d fused lasso problem (Tibshirani and Taylor, 2011) is then written as

$$\min_{\beta\in\mathbb{R}^n} \frac{1}{2}||\mathbf{y}-\beta||_2^2 + \lambda||\mathbf{D}\beta||_1, \tag{3.2}$$

where $\lambda \geq 0$ is a tuning parameter. The penalty matrix $\mathbf{D}$ is

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ & & & \cdots & & \\ 0 & 0 & 0 & ... & -1 & 1 \end{bmatrix}.$$

$||\cdot||_2^2$ and $||\cdot||_1$ denote the l-2 are l-1 norms, respectively. The $||\mathbf{y}-\beta||_2^2$ $(=\sum_{i=1}^n (y_i-\beta_i)^2)$ represents the squared error, and the $||\mathbf{D}\beta||_1$ $(=\sum_{i=1}^{n-1}|\beta_i-\beta_{i+1}|)$ represents the sum of the absolute differences between neighboring coefficients.

In CNV studies, copy numbers have an inherent structure along a chromosome. Copy numbers between neighboring genomic regions are then assumed to be similar. Thus, the penalty term $||\mathbf{D}\beta||_1$ is desirable in CNV detection because it encourages the differences between neighboring coefficients to be small. In addition, by using the l-1 norm, when the estimated values of neighboring coefficients are close to each other, the penalty term makes them to be exactly equal. Consequently, the solution,

$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_n)^T$, appears a piecewise constant shape. Then, the 1d fused lasso solution can capture the exact locations of change points. Thus, it is suitable in the multiple change point detection problem.

Furthermore, as mentioned in the Introduction, CNVs are sparse. In the true underlying model, only small number of regions may have CNVs among thousands of genomic locations. The sparsity in the 1d fused lasso depends on the tuning parameter selected. As λ increases, the $||\mathbf{D}\beta||_1$ encourages neighboring coefficients to be the same, leading a small number of change points (i.e., a sparse model); as λ decreases, neighboring coefficients tend to be unequal, leading a large number of change points (i.e., a dense model). We will describe how we select the optimal λ in Section 3.4. Now, given a λ value, we fit the 1d fused lasso to estimate the coefficients using the dual path algorithm in Section 3.3.

3.3 The dual path algorithm for solving the 1d fused lasso

In this section, we overview the dual path algorithm proposed by Tibshirani and Taylor (2011) for solving the 1d fused lasso problem. Although the 1d fused lasso (3.2) is convex, the penalty function, $||\mathbf{D}\beta||_1 = \sum_{i=1}^{n} |\beta_i - \beta_{i+1}|$, is not only non-differentiable but also non-seperable in each individual parameter. Thus, general optimization methods do not work for the 1d fused lasso. (Friedman et al., 2007). To deal with this issue, Tibshirani and Taylor (2011) proposed a path algorithm based on the Lagrange dual problem.

3.3.1 The dual problem of the 1d fused lasso

Kim et al. (2009) and Tibshirani and Taylor (2011) introduced a new variable $\mathbf{z} \in \mathbb{R}^{n-1}$ and a new equality constraint $\mathbf{z} = \mathbf{D}\beta$ to conveniently derive the dual problem of the 1d fused lasso. Then, (3.2) can be rewritten as

$$\min_{(\beta,\mathbf{z})\in\mathbb{R}^n\times\mathbb{R}^{n-1}} \frac{1}{2}||\mathbf{y}-\beta||_2^2+\lambda||\mathbf{z}||_1 \quad \text{subject to} \quad \mathbf{z}=\mathbf{D}\beta. \tag{3.3}$$

The Lagrangian of (3.3) follows that

$$\mathscr{L}(\beta,\mathbf{z},\mathbf{u}) = \frac{1}{2}||\mathbf{y}-\beta||_2^2+\lambda||\mathbf{z}||_1+\mathbf{u}^T(\mathbf{D}\beta-\mathbf{z}), \tag{3.4}$$

where $\mathbf{u}\in\mathbb{R}^{n-1}$ is the Lagrange multiplier vector, also known as a dual variable, associated with the equality constraint, $\mathbf{z}=\mathbf{D}\beta$. By minimizing $\mathscr{L}(\beta,\mathbf{z},\mathbf{u})$ in (3.4) over all β and $\mathbf{z}$, the dual function is obtained as

$$g(\mathbf{u}) = \min_{(\beta,\mathbf{z})\in\mathbb{R}^n\times\mathbb{R}^{n-1}} \mathscr{L}(\beta,\mathbf{z},\mathbf{u}) = \begin{cases} -\frac{1}{2}||\mathbf{y}-\mathbf{D}^T\mathbf{u}||_2^2 & \text{if } ||\mathbf{u}||_\infty \le \lambda \\ -\infty & \text{otherwise,} \end{cases} \tag{3.5}$$

where $||\mathbf{u}||_\infty = max_j|u_j|$. The dual function $g(\mathbf{u})$ yields lower bounds on the optimal value of (3.2) for any $\mathbf{u}\in\mathbb{R}^{n-1}$ satisfying the constraint. Then, the best lower bound of (3.2) is given by maximizing the dual function $g(\mathbf{u})$ that is concave. Maximizing the dual function is equivalent to minimizing $-g(\mathbf{u})$. Therefore, the optimization problem for (3.2) follows that

$$\min_{\mathbf{u}\in\mathbb{R}^{n-1}} \frac{1}{2}||\mathbf{y}-\mathbf{D}^T\mathbf{u}||_2^2 \quad \text{subject to} \quad ||\mathbf{u}||_\infty \le \lambda. \tag{3.6}$$

This is the dual problem of (3.2). Note that the dual problem is strictly convex, and thus it has a unique solution. Thus, the primal problem (3.2) can be solved by minimizing the objective function (3.6) with respect to the dual variable $\mathbf{u}\in\mathbb{R}^{n-1}$ under the constraint $||\mathbf{u}||_\infty \le \lambda$.

It is worth mentioning the relationship between the primal and dual problems. In optimization theory, if f^* is the optimal value of the primal problem and d^* is the optimal value of the dual problem (i.e., d^* is the best lower bound of f^*), then we always have $d^* \le f^*$. This property is called weak duality. Then, in optimization, we want to make the gap $f^* - d^*$ as tight as possible. When $f^* = d^*$, we say that strong duality holds. In general, strong duality does not hold. However, in the 1d fused lasso, since the primal problem (3.2) is strictly convex and unconstrained, strong duality holds (Boyd and Vandenberghe, 2004), suggesting that the values of the objective functions in the primal and dual problems are the same. Therefore, under the strong duality, the optimal solution of the 1d fused lasso problem can be efficiently obtained from solving the dual problem.

Let $(\hat{\beta}_\lambda, \hat{\mathbf{u}}_\lambda) \in \mathbb{R}^n \times \mathbb{R}^{n-1}$ be the pair of primal and dual solutions depending on λ. Under the strong duality, given a dual solution $\hat{\mathbf{u}}_\lambda$, any primal solution $\hat{\beta}_\lambda$ satisfies

$$\hat{\beta}_\lambda = \mathbf{y} - \mathbf{D}^T \hat{\mathbf{u}}_\lambda. \tag{3.7}$$

(3.7) indicates the linear relationship between $\hat{\beta}_\lambda$ and $\hat{\mathbf{u}}$. Also, $\hat{\beta}$ represents the residual of the dual solution.

The dual solution at each coordinate follows as

$$\hat{u}_{\lambda,i} \in \begin{cases} \{+\lambda\} & \text{if } (\mathbf{D}\hat{\beta}_\lambda)_i > 0 \quad i.e., \hat{\beta}_{\lambda,i+1} - \hat{\beta}_{\lambda,i} > 0 \\ (-\lambda, +\lambda) & \text{if } (\mathbf{D}\hat{\beta}_\lambda)_i = 0 \quad i.e., \hat{\beta}_{\lambda,i+1} - \hat{\beta}_{\lambda,i} = 0 \\ \{-\lambda\} & \text{if } (\mathbf{D}\hat{\beta}_\lambda)_i < 0 \quad i.e., \hat{\beta}_{\lambda,i+1} - \hat{\beta}_{\lambda,i} < 0. \end{cases} \tag{3.8}$$

(3.7) and (3.8) are obtained by taking the gradient and subgradient of the Lagrangian (3.4) with respect to β and $\mathbf{z}$, respectively. Based on the dual solution (3.8), we describe how the dual path algorithm solves the 1d fused lasso in the next subsection.

3.3.2 The dual path algorithm for the 1d fused lasso

Let B denote a boundary set that contains the coordinates, corresponding to the detected change points, and B^I denote an interior set for the rest. Suppose that the number of detected change points is K, $K = 0, ..., n-2$. Then, K coordinates that correspond to the change points are in the B. $n-K-1$ coordinates are in the B^I. Let $\mathbf{D}_B \in \mathbb{R}^{K\times n}$ and $\mathbf{D}_{B^I} \in \mathbb{R}^{(n-K-1)\times n}$ consist of the rows of $\mathbf{D}$ in the B and B^I, respectively. Let $\mathbf{s} \in \mathbb{R}^{K\times 1}$ consist of the signs of the coefficient estimates in the B.

Let λ_K denote the value of λ at the Kth iteration. Then, the dual solution in the B is estimated by (3.8) as

$$\hat{\mathbf{u}}_{\lambda_K,B} = \lambda_K \mathbf{s}. \tag{3.9}$$

In the interior set B^I, the dual problem (3.6) becomes

$$\min_{\mathbf{u}_{B^I}} \frac{1}{2}\|\mathbf{y} - \lambda \mathbf{D}_B{}^T\mathbf{s} - \mathbf{D}_{B^I}{}^T\mathbf{u}_{B^I}\|_2^2 \quad \text{subject to} \quad \|\mathbf{u}_{B^I}\|_\infty \le \lambda. \tag{3.10}$$

The dual solution in the B^I is then obtained as the least squares estimate of (3.10) at λ_K.

$$\begin{aligned} \hat{\mathbf{u}}_{\lambda_K,B^I} &= (\mathbf{D}_{B^I}\mathbf{D}_{B^I}{}^T)^{-1}\mathbf{D}_{B^I}(\mathbf{y} - \lambda_K \mathbf{D}_B{}^T\mathbf{s}) \\ &= (\mathbf{D}_{B^I}\mathbf{D}_{B^I}{}^T)^{-1}\mathbf{D}_{B^I}\mathbf{y} - \lambda_K(\mathbf{D}_{B^I}\mathbf{D}_{B^I}{}^T)^{-1}\mathbf{D}_{B^I}\mathbf{D}_B{}^T\mathbf{s}. \end{aligned} \tag{3.11}$$

For $\lambda < \lambda_K$, as λ decreases, $\hat{u}_{\lambda_K,i}$ will eventually hit the boundary of the constraint set $\|\mathbf{u}_{B^I}\|_\infty \le \lambda$ in (3.10). The corresponding $\hat{\beta}_{\lambda,i}$ and $\hat{\beta}_{\lambda,i+1}$ will split apart by (3.8), i.e., $\hat{\beta}_{\lambda,i} \neq \hat{\beta}_{\lambda,i+1}$. Then, the next critical λ_{K+1} is calculated as $\max_i |\hat{u}_{\lambda_K,i}|$. Moreover, once one of the dual interior solution hits the constraint boundary, it will stay on the boundary forever by the bound lemma 1 in Tibshirani and Taylor (2011). In other words, once neighboring coefficients split apart, they never merge again as λ decreases. Therefore, the dual path algorithm produces $n-1$ fused lasso solutions according to $n-1$ critical tuning parameter values without searching various λ values over a wide range.

In terms of change point problems, the dual path algorithm estimates the best locations of

---

Algorithm 1 Dual path algorithm for the 1d fused lasso

---

input: $\mathbf{y} \in \mathbb{R}^n$

initialize: $\lambda_0 = \infty$
$B = \emptyset$ (i.e., $\mathbf{D}_B =$ and $\mathbf{D}_{B^I} = (n-1) \times n$)
$\mathbf{s} = \emptyset$

iterate for $K = 0, ..., n-2$

1. Calculate $\hat{\mathbf{u}}_{\lambda_K, B^I}$ at λ_K as least-squares estimates, as in (3.11).
- Using the primal-dual relationship, $\hat{\beta}_{\lambda_K}$ at λ_K is calculated as
$\hat{\beta}_{\lambda_K} = \mathbf{y} - (\mathbf{D}_{B^I}{}^T \hat{\mathbf{u}}_{\lambda_K, B^I} + \mathbf{D}_B{}^T \hat{\mathbf{u}}_{\lambda_K, B})$, where $\hat{\mathbf{u}}_{\lambda_K, B} = \lambda_K \mathbf{s}$.

2. For $\lambda \leq \lambda_K$ and $i \in B^I$, update $\lambda_{K+1} = \max_i |\hat{u}_{\lambda_K, i}|$ and $i_{K+1} = \arg\max_i |\hat{u}_{\lambda_K, i}|$

3. Add the coordinate i_{K+1} to B and its sign to $\mathbf{s}_{K+1}$.

output: the critical tuning parameters $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{n-2}$.
the entire 1d fused lasso solutions $\hat{\beta}_{\lambda_K}$, where $K = 0, ..., n-2$.

---

change points according to the number of change points. The critical Kth tuning parameter corresponds to the model of K changes. At $\lambda_0 = \infty$, all fitted response values are the same as their mean, indicating the model of no change. At λ_1, the best location i_1 is estimated at one change point where $\hat{\beta}_{\lambda_1, i_1} \neq \hat{\beta}_{\lambda_1, i_1+1}$. Based on the boundary lemma, with fixing the detected change point at i_1 included in B, another change point in B^I is identified at i_2 where $\hat{\beta}_{\lambda_2, i_2} \neq \hat{\beta}_{\lambda_2, i_2+1}$, leading to the 2 change point model. Similarly at λ_K, with holding the detected change points at $i_1, ..., i_{K-1}$, the best location i_K is estimated where $\hat{\beta}_{\lambda_K, i_K} \neq \hat{\beta}_{\lambda_K, i_K+1}$, leading to the K change point model.

Therefore, once the dual path algorithm identifies the change point locations according to the number of change points, the next step is to estimate the true number of change points by selecting the optimal tuning parameter among $n-1$ candidate values. We select the optimal λ based on Bayesian information criterion in the next section.

3.4 A modified information criterion for tuning parameter selection

Let $\theta = (\theta_1, ..., \theta_K)$ be a parameter vector, $\mathbf{t}_K = (t_1, ..., t_K)$ be a location vector of K change points, where $0 = t_0 < t_1 < t_2 < \cdots < t_K < t_{K+1} = n$. Then, given λ_K in the 1d fused lasso, the original Bayesian information criterion (SIC, Swartz 1979) is calculated as

$$SIC(\lambda_K) = -2ln(\hat{\theta}, \hat{\mathbf{t}}_K, K) + dln(n), \tag{3.12}$$

where d is the number of parameters in each segment. In the multiple change points model (3.1), the log likelihood function has the form

$$ln(\theta, \mathbf{t}_K, K) = \sum_{k=1}^{K+1} \sum_{i=t_{k-1}+1}^{t_k} lnf(y_i, \theta_k), \tag{3.13}$$

where $\theta_k = (\mu_k, \sigma_k)$ are maximum likelihood estimators in each segment.

Although SIC has been widely used in model selection, since the multiple change point model has additional parameters, K and $\mathbf{t}$, SIC does not penalize enough the model complexity. For this reason, Pan and Chen (2006) modified SIC for the multiple change point problem. The model selection criterion by Pan and Chen (2006) is denoted as PMIC and defined as

$$PMIC(\lambda_K) = -2ln(\hat{\theta}, \hat{\mathbf{t}}_K, K) + d(K+1)ln(n) + C\sum_{k=1}^{K+1} \left(\frac{\hat{t}_k - \hat{t}_{k-1}}{n} - \frac{1}{K+1}\right)^2 ln(n). \tag{3.14}$$

Our proposed criterion in this dissertation was motivated by Pan and Chen (2006). The second term in PMIC is the penalty for the number of change points. For given K change points, there are $K+1$ segments, and each segment has d parameters. In this dissertation, $d = 2$ for mean and variance. The third term including C is the penalty for a distribution of change point locations. The $(t_k - t_{k-1})/n$ is the observed distance between change points scaled by n, while $1/(K+1)$ is the expected distance.

Thus, it indicates that PMIC favors a model where change points are evenly distributed with an equal distance of $1/(K+1)$. Pan and Chen (2006) states that if observed change points are closely located together or near the end of the sequence, the model is complex because it may include some unnecessary change points. Although the idea seems to be reasonable, PMIC may suffer when the true model is not uniformly distributed. In addition, for a constant C, Pan and Chen (2006) used $C = 1$ or 10 in their paper. However, PMICs with $C = 0, 1, 10, 100$ detected a similar number of change points in our simulation studies. Therefore, we removed the third penalty term including C from PMIC and considered another way to penalize multiple change point models.

Our proposed model selection criterion, called JMIC, is defined as

$$JMIC(\lambda_K) = -2ln(\hat{\theta}, \hat{\mathbf{t}}_K, K) + d(K+1)^{\gamma} n\rho_n, \tag{3.15}$$

where $1 < \gamma < 2$, and $\rho_n = n^{\alpha-1}$, $0 < \alpha < 1$. Through simulations, we found that SIC and PMIC tend to overestimate the number of change points. Thus, we put more penalty on the number of segments $(K+1)$ by introducing the exponent γ. In simulated data, when $0 < \gamma \leq 1$, there was no noticeable difference between JMIC and PMIC. Also, when $2 < \gamma$, JMIC tended to underestimate the number of changes. Thus, the range of γ was defined as being between 1 and 2. Furthermore, we included ρ_n to establish the consistency of JMIC. The ρ_n with $0 < \alpha < 1$ satisfies Assumption A6 in Chapter 4.

Let us now apply JMIC to tuning parameter selection in the 1d fused lasso. We have estimated $\hat{\mathbf{t}}_K$ by the dual path algorithm along with the 1d fused lasso at $n-2$ critical λ values. Then, based on the principle of minimum information criterion, we select the optimal $\hat{\lambda}_{\hat{K}}$, where $\hat{K} = arg\{\min_{0 \leq K \leq n-2} JMIC(\lambda_K)\}$. The entire steps for detecting change points using the 1d fused lasso with JMIC are given in Algorithm 2 below.

Zhang and Siegmund (2007) also proposed a modified Bayesian information criterion for multiple change point problems. We denote it as ZMIC.

$$ZMIC(\lambda_K) = -2ln(\hat{\boldsymbol{\theta}}, \hat{\mathbf{t}}, K) - \frac{1}{2}\sum_{k=1}^{K+1}\log(\hat{t}_k - \hat{t}_{k-1}) + (\frac{1}{2} - K)ln(n). \tag{3.16}$$

We will compare JMIC with SIC, PMIC, and ZMIC via simulations and real data analysis in Chapter 5 and 6, respectively.

Algorithm 2 Multiple change points detection using the 1d fused lasso with *JMIC*

---

input: $\mathbf{y} \in \mathbb{R}^n$, constants γ and α.

Step1 Do Algorithm 1: the dual path algorithm for the 1d fused lasso.

output: the critical tuning parameters $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{n-2}$.
the entire 1d fused lasso solutions $\hat{\beta}_{\lambda_K}$, where $K = 0, ..., n-2$.
the critical change point locations $\hat{\mathbf{t}} = (\hat{t}_1, ..., \hat{t}_K)$ for K change points, where $K = 0, .., n-2$.

Step2 for $K = 0, ..., n-2$
1. Calculate $JMIC(\lambda_K)$, as in (3.15).
2. Select the optimal $\hat{\lambda}_{\hat{K}}$, where $\hat{K} = arg\{\min_{0 \leq K \leq n-2} JMIC(\lambda_K)\}$.

output: the optimal number of change points $\hat{K}$ corresponding to $\hat{\lambda}_{\hat{K}}$.
the optimal change point locations $\hat{\mathbf{t}}_{\hat{K}} = (\hat{t}_1, ..., \hat{t}_{\hat{K}})$ corresponding to $\hat{\lambda}_{\hat{K}}$.
$\hat{\mathbf{y}}$, the mean values of observations in each segment.

---

4 Properties of JMIC

JMIC estimates the true number of change points via the selection of the optimal tuning parameter in the 1d fused lasso. In this chapter, we will prove that the number of change points estimated by JMIC converges to the true number of change points.

4.1 Consistency

Throughout this section, we denote the true value of a parameter with a superscript 0.

Let $I_k^0 = t_k^0 - t_{k-1}^0$ for $k = 1, ..., K^0 + 1$. Define

$$I_{min}^0 = \min_{1 \le k \le K^0+1} |I_k^0|,\ J_{min}^0 = \min_{1 \le k \le K^0} |\mu_{k+1}^0 - \mu_k^0|, \text{ and } J_{max}^0 = \max_{1 \le k \le K^0} |\mu_{k+1}^0 - \mu_k^0|.$$

I_{min}^0 denotes the minimum interval length, and J_{min}^0 and J_{max}^0 denote the minimum and maximum jump sizes, respectively.

We make the following assumptions. [1]

(A1) The $\varepsilon_1, ..., \varepsilon_n$ are iid random variables with $E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$ satisfying: there exists a positive constant B such that for all $\nu \in \mathbb{R}$, $\mathrm{E}(\exp(\nu\varepsilon_1)) \le \exp(B\nu^2)$.

(A2) The $\{\delta_n\}$ is a non-increasing sequence of positive constants such that $\delta_n \to 0$ and $n\delta_n(J_{min}^0)^2/\log(n) \to \infty$ as $n \to \infty$.

(A3) $K^0 = O(\log(n))$ and $I_{min}^0/(n\delta_n) \to \infty$ as $n \to \infty$.

[1] Throughout the dissertation, for two sequences a_n and b_n, $a_n = O(b_n)$ if there exist positive constants M and n_0 such that $|a_n| \le M|b_n|$ for all $n > n_0$; $a_n = \Omega(b_n)$ if there exist positive constants M and n_0 such that $|a_n| \ge M|b_n|$ for all $n > n_0$; $a_n = \Theta(b_n)$ if there exist positive constants $m < M$ and n_0 such that $m|b_n| \le |a_n| \le M|b_n|$ for all $n > n_0$.

(A4) $J^0_{max} = O(1)$, $J^0_{min} = \Omega(1)$, $I^0_{min} = \Theta(n)$.

(A5) The tuning parameter $\lambda = \lambda_n$ satisfies $n\lambda/(n\delta_n J^0_{min}) \to 0$ as $n \to \infty$.

(A6) The $\{\rho_n\}$ is a non-increasing sequence of positive constants such that $\rho_n \to 0$ and $\rho_n/\delta_n \to \infty$ as $n \to \infty$.

Lemma 1. Consider the 1d fused lasso problem in (3.2). Then

$$\begin{aligned} &\text{(i)} \quad \sum_{i=\hat{t}_l+1}^{n} (y_i - \hat{\beta}_i) = n\lambda \hat{s}_l \qquad \text{for} \quad l = 1, ..., \hat{K}; \\ &\text{(ii)} \quad |\sum_{i=j}^{n} (y_i - \hat{\beta}_i)| \le n\lambda \qquad \text{for} \quad j = 1, ..., n, \end{aligned}$$

using the convention $\hat{s}_l = +1$ if $\hat{\beta}_{\hat{t}_l+1} > \hat{\beta}_{\hat{t}_l}$ and $\hat{s}_l = -1$, otherwise.

Proof. Lemma 1 is adapted from Lemma 1 in Harchaoui and Lévy-Leduc (2012) and Lemma 1 in Qian and Su (2016). Let us first transform the 1d fused lasso in (3.2). Let $\mathbf{X} \in \mathbb{R}^{n \times n}$ be a lower triangular matrix with all nonzero elements equal to one. Let $\theta \in \mathbb{R}^{n \times 1}$ be a vector, having $\theta_1 = \beta_1$ and $\theta_i = \beta_i - \beta_{i-1}$ for $i = 2, ...n$. Then (3.2) is equivalent to the original Lasso problem (Tibshirani, 1996).

$$\min_{\theta \in \mathbb{R}^n} \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + ||\lambda\theta||_1, \tag{4.1}$$

By taking the subgradient of (4.1) with respect to θ, we have a necessary and sufficient condition for a solution $\hat{\theta}$ as follows:

$$(\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\theta}))_j = n\lambda \operatorname{sign}(\hat{\theta}_j), \quad \text{if} \quad \hat{\theta}_j \ne 0,$$

$$|(\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\theta}))_j| \le n\lambda, \quad \text{if} \quad \hat{\theta}_j = 0.$$

Therefore, since $(\mathbf{X}^T\mathbf{y})_j = \sum_{k=j}^n y_k$ and $(\mathbf{X}^T\mathbf{X}\hat{\theta})_j$=$(\mathbf{X}^T\hat{\beta})_j = \sum_{k=j}^n \hat{\beta}_k$, the proof is complete. □

Lemma 2. Let $\{\nu_n\}$ and $\{x_n\}$ be two positive sequences such that $\nu_n x_n^2/\log(n) \to \infty$. Then

$$P\Big(\max_{\substack{1\le r_n<s_n\le n\\ |r_n-s_n|\ge \nu_n}} \Big|(s_n-r_n)^{-1}\sum_{i=r_n+1}^{s_n}\varepsilon_i\Big| \ge x_n\Big) \to 0 \quad \text{as} \quad n\to\infty.$$

Proof. Lemma 2 closely follows Lemma 2 in Harchaoui and Lévy-Leduc (2012). Note that

$$P\Big(\max_{\substack{1\le r_n<s_n\le n\\ |r_n-s_n|\ge \nu_n}} \Big|(s_n-r_n)^{-1}\sum_{i=r_n+1}^{s_n}\varepsilon_i\Big| \ge x_n\Big) \le \sum_{\substack{1\le r_n<s_n\le n\\ |r_n-s_n|\ge \nu_n}} P\Big(\Big|(s_n-r_n)^{-1}\sum_{i=r_n+1}^{s_n}\varepsilon_i\Big| \ge x_n\Big).$$

For all $\eta > 0$,

$$P\Big(\sum_{i=r_n+1}^{s_n}\varepsilon_i > (s_n-r_n)x_n\Big) \le \exp(-\eta(s_n-r_n)x_n)E(\exp(\eta\varepsilon))^{s_n-r_n} \tag{4.2}$$

$$\le \exp(-\eta(s_n-r_n)x_n + B\eta^2(s_n-r_n)), \tag{4.3}$$

where B is a positive constant. To show (4.2), let Y be a random variable and $\varepsilon > 0$. Then, for any $t > 0$, $P(Y > \varepsilon) \le \exp(-t\varepsilon)E(\exp(tY))$ by Markov's inequality. Substituting $Y = \sum_{i=r_n+1}^{s_n}\varepsilon_i$, $\varepsilon = (s_n-r_n)x_n$, and $t = \eta$ yields (4.2). (4.3) follows from Assumption (A1).

Since the sharpest bound of (4.3) holds for $\eta = \frac{x_n}{2B}$, we obtain

$$P\Big((s_n-r_n)^{-1}\textstyle\sum_{i=r_n+1}^{s_n}\varepsilon_i > x_n\Big) \le \exp(-x_n^2(s_n-r_n)/(4B)).$$

Consider the bound for $|\varepsilon_i|$. Then,

$$P\Big(|(s_n-r_n)^{-1}\textstyle\sum_{i=r_n+1}^{s_n}\varepsilon_i| > x_n\Big) \le 2\exp(-x_n^2(s_n-r_n)/(4B)).$$

Hence,

$$P\Big(\max_{\substack{1\le r_n<s_n\le n\\ |r_n-s_n|\ge \nu_n}}\Big|(s_n-r_n)^{-1}\sum_{i=r_n+1}^{s_n}\varepsilon_i\Big|\ge x_n\Big)\le 2n\exp(-\nu_n x_n^2/(4B))$$

$$\to 0,$$

as $n\to\infty$ if $\nu_n x_n^2/\log(n)\to\infty$, which completes the proof. □

We introduce two techniques that will be useful in the proofs of the following theorems. In the two techniques, we will use the notion of random variables corresponding to a probability space.

Let $(\Omega,\mathscr{F},P)$ be a probability space (Laha and Rohatgi, 1979, p.3). Let E be a set. A real-valued function Z defined on Ω is said to be a random variable if

$$Z^{-1}(E)=\{\omega\in\Omega: Z(\omega)\in E\}\in\mathscr{F}\quad\text{for}\quad\text{all}\quad E\in\mathscr{B},$$

where $\mathscr{B}$ is the σ-field of Borel sets in $\mathbb{R}=(-\infty,\infty)$; that is, a random variable Z is a measurable transformation of $(\Omega,\mathscr{F},P)$ into $(\mathbb{R},\mathscr{B})$. We denote Z_1, Z_2, Z_3, and Z_4 as such defined random variables in the following two techniques.

Technique 1.

$$\begin{aligned}
&P\Big(\{\omega: Z_1(\omega)\ge Z_2(\omega)-Z_3(\omega)-Z_4(\omega)\}\Big)\\
&\quad\le P\Big(\{\omega: Z_1(\omega)\ge \frac{1}{3}Z_2(\omega)\}\Big)+P\Big(\{\omega: Z_3(\omega)\ge \frac{1}{3}Z_2(\omega)\}\Big)\\
&\qquad+P\Big(\{\omega: Z_4(\omega)\ge \frac{1}{3}Z_2(\omega)\}\Big).
\end{aligned}$$

Proof.

$$
\begin{aligned}
&\{\omega : Z_1(\omega) \geq Z_2(\omega) - Z_3(\omega) - Z_4(\omega)\} \\
&= \{\omega : Z_1(\omega) \geq Z_2(\omega) - Z_3(\omega) - Z_4(\omega)\} \\
&\quad \cap \Big(\{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : Z_1(\omega) < \frac{1}{3}Z_2(\omega)\}\Big) \\
&= \Big(\{\omega : Z_1(\omega) \geq Z_2(\omega) - Z_3(\omega) - Z_4(\omega)\} \cap \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\}\Big) \\
&\quad \cup \Big(\{\omega : Z_1(\omega) \geq Z_2(\omega) - Z_3(\omega) - Z_4(\omega)\} \cap \{\omega : Z_1(\omega) < \frac{1}{3}Z_2(\omega)\}\Big) \\
&\subseteq \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \\
&\qquad \cup \Big(\{\omega : Z_1(\omega) \geq Z_2(\omega) - Z_3(\omega) - Z_4(\omega)\} \cap \{\omega : Z_1(\omega) < \frac{1}{3}Z_2(\omega)\}\Big) \\
&= \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \\
&\qquad \cup \{\omega : Z_2(\omega) - Z_3(\omega) - Z_4(\omega) \leq Z_1(\omega) < \frac{1}{3}Z_2(\omega)\} \\
&\subseteq \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : Z_2(\omega) - Z_3(\omega) - Z_4(\omega) \leq \frac{1}{3}Z_2(\omega)\} \\
&= \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : \frac{2}{3}Z_2(\omega) \leq Z_3(\omega) + Z_4(\omega)\} \\
&= \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \\
&\quad \cup \Big(\{\omega : \frac{2}{3}Z_2(\omega) \leq Z_3(\omega) + Z_4(\omega)\} \cap (\{\omega : Z_3 \geq \frac{1}{3}Z_2\} \cup \{\omega : Z_3(\omega) < \frac{1}{3}Z_2(\omega)\})\Big) \\
&= \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \\
&\quad \cup \Big((\{\omega : \frac{2}{3}Z_2(\omega) \leq Z_3(\omega) + Z_4(\omega)\} \cap \{\omega : Z_3(\omega) \geq \frac{1}{3}Z_2(\omega)\}) \\
&\qquad \cup (\{\omega : \frac{2}{3}Z_2(\omega) \leq Z_3(\omega) + Z_4(\omega)\} \cap \{\omega : Z_3(\omega) < \frac{1}{3}Z_2(\omega)\})\Big) \\
&\subseteq \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : Z_3(\omega) \geq \frac{1}{3}Z_2(\omega)\} \\
&\quad \cup \Big(\{\omega : \frac{2}{3}Z_2(\omega) \leq Z_3(\omega) + Z_4(\omega)\} \cap \{\omega : Z_3(\omega) < \frac{1}{3}Z_2(\omega)\}\Big) \\
&= \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : Z_3(\omega) \geq \frac{1}{3}Z_2(\omega)\} \\
&\quad \cup \{\omega : \frac{2}{3}Z_2(\omega) - Z_4(\omega) \leq Z_3(\omega) < \frac{1}{3}Z_2(\omega)\} \\
&\subseteq \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : Z_3(\omega) \geq \frac{1}{3}Z_2(\omega)\}
\end{aligned}
$$

$$\cup \{\omega : \frac{2}{3}Z_2(\omega) - Z_4(\omega) \leq \frac{1}{3}Z_2(\omega)\}$$
$$= \{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : Z_3(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : Z_4(\omega) \geq \frac{1}{3}Z_2(\omega)\}.$$

Therefore,

$$\begin{aligned}
&P\Big(\{\omega : Z_1(\omega) \geq Z_2(\omega) - Z_3(\omega) - Z_4(\omega)\}\Big) \\
&\leq P\Big(\{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : Z_3(\omega) \geq \frac{1}{3}Z_2(\omega)\} \cup \{\omega : Z_4(\omega) \geq \frac{1}{3}Z_2(\omega)\}\Big) \\
&\leq P\Big(\{\omega : Z_1(\omega) \geq \frac{1}{3}Z_2(\omega)\}) + P\Big(\{\omega : Z_3(\omega) \geq \frac{1}{3}Z_2(\omega)\}\Big) \\
&\qquad + P\Big(\{\omega : Z_4(\omega) \geq \frac{1}{3}Z_2(\omega)\}\Big).
\end{aligned}$$

□

Technique 2.

$$\begin{aligned}
&P\Big(\{\omega : Z_1(\omega) + Z_2(\omega) \geq Z_3(\omega)/2\}\Big) \\
&\leq P\Big(\{\omega : Z_1(\omega) \geq Z_3(\omega)/2\}\Big) + P\Big(\{\omega : Z_2(\omega) \geq Z_3(\omega)/2\}\Big).
\end{aligned}$$

Proof.

$$\{\omega : Z_1(\omega) + Z_2(\omega) \geq Z_3(\omega)/2\}$$
$$= \{\omega : Z_1(\omega) \geq Z_3(\omega)/2\} \cap \{\omega : Z_2(\omega) \geq Z_3(\omega)/2\}$$
$$\subseteq \{\omega : Z_1(\omega) \geq Z_3(\omega)/2\} \cup \{\omega : Z_2(\omega) \geq Z_3(\omega)/2\}.$$

Therefore,

$$P\Big(\{\omega : Z_1(\omega)+Z_2(\omega) \geq Z_3(\omega)/2\}\Big)$$
$$\leq P\Big(\{\omega : Z_1(\omega) \geq Z_3(\omega)/2\} \cup \{\omega : Z_2(\omega) \geq Z_3(\omega)/2\}\Big)$$
$$\leq P\Big(\{\omega : Z_1(\omega) \geq Z_3(\omega)/2\}\Big) + P\Big(\{\omega : Z_2(\omega) \geq Z_3(\omega)/2\}\Big).$$

□

Theorem 1. Under Assumptions A1- A5, if $\hat{K} = K^0$, then

$$P(\max_{1\leq k\leq K^0} |\hat{t}_k - t_k^0| \leq n\delta_n) \to 1 \qquad \text{as } n \to \infty.$$

Proof. Our proof closely follows Proposition 3 in Harchaoui and Lévy-Leduc (2012) and Theorem 3.1 in Qian and Su (2016). Now we will prove that $P(\max_{1\leq k\leq K^0} |\hat{t}_k - t_k^0| > n\delta_n) \to 0$ as $n \to \infty$ as follows. Since $P(\max_{1\leq k\leq K^0} |\hat{t}_k - t_k^0| > n\delta_n) \leq \sum_{k=1}^{K^0} P(|\hat{t}_k - t_k^0| > n\delta_n)$, it suffices to prove that for all $k = 1,...,K^0$, $P(|\hat{t}_k - t_k^0| \geq n\delta_n) \to 0$. Let us denote $A_{n,k} = \{|\hat{t}_k - t_k^0| \geq n\delta_n\}$ and define the set

$$C_n = \{\max_{1\leq k\leq K^0} |\hat{t}_k - t_k^0| < I_{min}^0/2\}. \tag{4.4}$$

The proof of Theorem 1 is then completed by showing that (1-i) $P(A_{n,k} \bigcap C_n) \to 0$ and (1-ii) $P(A_{n,k} \bigcap C_n^c) \to 0$, where C_n^c denotes the complement of C_n.

We first prove (1-i). Denote $A_{n,k}^+ = \{t_k^0 - \hat{t}_k \geq n\delta_n\}$ and $A_{n,k}^- = \{\hat{t}_k - t_k^0 \leq n\delta_n\}$. Without loss of generality, we prove that $P(A_{n,k}^+ \bigcap C_n) \to 0$ as the other case follows analogously. Our goal is now to find an upper bound of $P(A_{n,k}^+ \bigcap C_n)$.

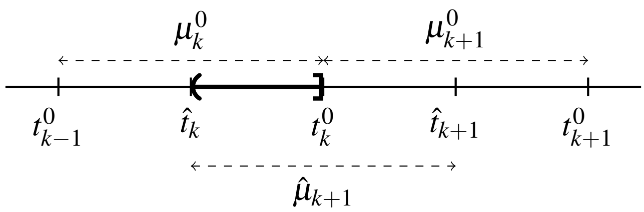
On the event $A_{n,k}^+ \bigcap C_n$, we have

$$t_{k-1}^0 < \hat{t}_k < t_k^0 < t_{k+1}^0 \quad \text{for all} \quad k \in \{1,...,K^0\}. \tag{4.5}$$

Applying (i) in Lemma 1 with $l = k$ and (ii) in Lemma 1 with $j = t_k^0 + 1$ gives, respectively, $\sum_{i=\hat{t}_k+1}^{n}(y_i - \hat{\beta}_i) = n\lambda\hat{s}_l$ and $|\sum_{i=t_k^0+1}^{n}(y_i - \hat{\beta}_i)| \leq n\lambda$. By the triangle inequality and the fact that $|\hat{s}_l| = 1$, we have

$$
\begin{aligned}
2n\lambda &\geq \Big|\sum_{i=\hat{t}_k+1}^{n}(y_i - \hat{\beta}_i) - \sum_{i=t_k^0+1}^{n}(y_i - \hat{\beta}_i)\Big| \\
&= \Big|\sum_{i=\hat{t}_k+1}^{t_k^0}(y_i - \hat{\beta}_i)\Big| \\
&= \Big|\sum_{i=\hat{t}_k+1}^{t_k^0}((\beta_i^0 + \varepsilon_i) - \hat{\beta}_i)\Big| \qquad (4.6) \\
&= \Big|\sum_{i=\hat{t}_k+1}^{t_k^0}(\mu_k^0 - \hat{\mu}_{k+1}) + \sum_{i=\hat{t}_k+1}^{t_k^0}\varepsilon_i\Big| \qquad (4.7) \\
&= \Big|(\hat{t}_k - t_k^0)(\hat{\mu}_{k+1} - \mu_k^0) + \sum_{i=\hat{t}_k+1}^{t_k^0}\varepsilon_i\Big| \\
&= \Big|(\hat{t}_k - t_k^0)(\mu_{k+1}^0 - \mu_k^0) + (\hat{t}_k - t_k^0)(\hat{\mu}_{k+1} - \mu_{k+1}^0) + \sum_{i=\hat{t}_k+1}^{t_k^0}\varepsilon_i\Big| \\
&\geq \Big|(\hat{t}_k - t_k^0)(\mu_{k+1}^0 - \mu_k^0)\Big| - \Big|(\hat{t}_k - t_k^0)(\hat{\mu}_{k+1} - \mu_{k+1}^0)\Big| - \Big|\sum_{i=\hat{t}_k+1}^{t_k^0}\varepsilon_i\Big| \\
&\equiv R_{n,k1} - R_{n,k2} - R_{n,k3}. \qquad (4.8)
\end{aligned}
$$

(4.6) is obtained from (3.1). (4.7) follows from the fact that $\hat{\beta}_i = \hat{\mu}_{k+1}$ and $\beta_i^0 = \mu_k^0$ for $i \in [\hat{t}_k + 1, t_k^0]$ by (4.5).



Define the event $R_{n,k} = \{2n\lambda \geq R_{n,k1} - R_{n,k2} - R_{n,k3}\}$ from (4.8). Since $P(R_{n,k}) = 1$, $P(A_{n,k}^{+} \cap C_n) = P(A_{n,k}^{+} \cap C_n \cap R_{n,k})$. Then,

$$
\begin{aligned}
P(A_{n,k}^{+} \cap C_n) &= P\Big(A_{n,k}^{+} \cap C_n \cap \{2n\lambda \geq R_{n,k1} - R_{n,k2} - R_{n,k3}\}\Big) \\
&\leq P\Big(A_{n,k}^{+} \cap C_n \cap \{2n\lambda \geq \frac{1}{3} R_{n,k1}\}\Big) \\
&\quad + P\Big(A_{n,k}^{+} \cap C_n \cap \{R_{n,k2} \geq \frac{1}{3} R_{n,k1}\}\Big) \\
&\quad + P\Big(A_{n,k}^{+} \cap C_n \cap \{R_{n,k3} \geq \frac{1}{3} R_{n,k1}\}\Big) \qquad (4.9) \\
&\equiv P(A_{n,k1}) + P(A_{n,k2}) + P(A_{n,k3}).
\end{aligned}
$$

(4.9) is obtained using Technique 1 by replacing Z_1, Z_2, Z_3, Z_4 with $2n\lambda$, $R_{n,k1}$, $R_{n,k2}$, $R_{n,k3}$, respectively. Let us first bound $P(A_{n,k1})$.

$$
\begin{aligned}
P(A_{n,k1}) &= P\Big(A_{n,k}^{+} \cap C_n \cap \{2n\lambda \geq \frac{1}{3} R_{n,k1}\}\Big) \\
&= P\Big(A_{n,k}^{+} \cap C_n \cap \{2n\lambda \geq |(\hat{t}_k - t_k^0)(\mu_{k+1}^0 - \mu_k^0)|/3\}\Big) \\
&\leq P\Big(2n\lambda \geq |(\hat{t}_k - t_k^0)(\mu_{k+1}^0 - \mu_k^0)|/3; \quad t_k^0 - \hat{t}_k \geq n\delta_n\Big) \\
&\leq P\Big(n\lambda/(n\delta_n J_{min}^0) \geq 1/6\Big).
\end{aligned}
$$

By Assumption (A5), $n\lambda/(n\delta_n J_{min}^0) < 1/6$ for n large enough, and thus, $P(A_{n,k1}) \to 0$ as $n \to \infty$. Next, we bound $P(A_{n,k2})$. Applying (ii) in Lemma 1 with $j = t_k^0 + 1$ and $j = (t_k^0 + t_{k+1}^0)/2 + 1$ and using the triangle inequality, we have

$$
\begin{aligned}
2n\lambda &\geq \Big| \sum_{i=t_k^0+1}^{(t_k^0+t_{k+1}^0)/2} (y_i - \hat{\beta}_i) \Big| \\
&= \Big| \sum_{i=t_k^0+1}^{(t_k^0+t_{k+1}^0)/2} (\mu_{k+1}^0 - \hat{\mu}_{k+1}) + \sum_{i=t_k^0+1}^{(t_k^0+t_{k+1}^0)/2} \varepsilon_i \Big|
\end{aligned}
$$

$$\geq (t_{k+1}^0 - t_k^0)|\mu_{k+1}^0 - \hat{\mu}_{k+1}|/2 - \Big|\sum_{i=t_k^0+1}^{(t_k^0+t_{k+1}^0)/2} \varepsilon_i\Big|, \tag{4.10}$$

where $\hat{\beta}_i = \hat{\mu}_{k+1}$ and $\beta_i^0 = \mu_{k+1}^0$ for $i \in [t_k^0, (t_k^0 + t_{k+1}^0)/2]$.

Then, we may upper bound $P(A_{n,k2})$ as follows:

$$\begin{aligned}
P(A_{n,k2}) &= P\Big(A_{n,k}^+ \cap C_n \cap \{R_{n,k2} \geq \frac{1}{3} R_{n,k1}\}\Big) \\
&= P\Big(A_{n,k}^+ \cap C_n \cap |(\hat{t}_k - t_k^0)(\hat{\mu}_{k+1} - \mu_{k+1}^0)| \geq |(\hat{t}_k - t_k^0)(\mu_{k+1}^0 - \mu_k^0)|/3\}\Big) \\
&\leq P\Big(4n\lambda\,(t_{k+1}^0 - t_k^0)^{-1} + 2\Big|(t_{k+1}^0 - t_k^0)^{-1} \textstyle\sum_{i=t_k^0+1}^{(t_k^0+t_{k+1}^0)/2} \varepsilon_i\Big| \geq |\mu_{k+1}^0 - \mu_k^0|/3\Big) \qquad (4.11) \\
&\leq P\Big(4n\lambda\,(t_{k+1}^0 - t_k^0)^{-1} \geq |\mu_{k+1}^0 - \mu_k^0|/6\Big) \\
&\quad + P\Big(2\Big|(t_{k+1}^0 - t_k^0)^{-1} \textstyle\sum_{i=t_k^0+1}^{(t_k^0+t_{k+1}^0)/2} \varepsilon_i\Big| \geq |\mu_{k+1}^0 - \mu_k^0|/6\Big). \qquad (4.12)
\end{aligned}$$

(4.11) follows from (4.10). (4.12) is obtained using Technique 2 by replacing Z_1, Z_2, Z_3 with $4n\lambda\,(t_{k+1}^0 - t_k^0)^{-1}$, $2\Big|(t_{k+1}^0 - t_k^0)^{-1} \sum_{i=t_k^0+1}^{(t_k^0+t_{k+1}^0)/2} \varepsilon_i\Big|$, $|\mu_{k+1}^0 - \mu_k^0|/3$, respectively.

Consider the first term on the right-hand side of (4.12). By Assumption A3, we obtain

$$P\Big(4n\lambda\,(t_{k+1}^0 - t_k^0)^{-1} \geq |\mu_{k+1}^0 - \mu_k^0|/6\Big) \leq P\Big(n\lambda/(n\delta_n J_{min}^0) \geq 1/24\Big). \tag{4.13}$$

Then, by Assumption (A5), (4.15) $\to 0$ as $n \to \infty$.

Next, consider the second term on the right-hand side of (4.12). By Assumption A3,

$$\begin{aligned}
&P\Big(2\Big|(t_{k+1}^0 - t_k^0)^{-1} \textstyle\sum_{i=t_k^0+1}^{(t_k^0+t_{k+1}^0)/2} \varepsilon_i\Big| \geq |\mu_{k+1}^0 - \mu_k^0|/6\Big) \\
&\leq P\Big(\max_{\substack{1 \leq t_k^0 < (t_k^0+t_{k+1}^0)/2 \leq n \\ |t_{k+1}^0 - t_k^0| \geq n\delta_n}} \Big|(t_{k+1}^0 - t_k^0)^{-1} \sum_{i=t_k^0+1}^{(t_{k+1}^0+t_k^0)/2} \varepsilon_i\Big| \geq J_{min}^0/12\Big), \qquad (4.14)
\end{aligned}$$

Then, by Lemma 2 with $x_n = J_{min}^0/12$ and $\nu_n = n\delta_n$, (4.14) $\to 0$ as $n \to \infty$. Therefore, $P(A_{n,k2}) \to 0$ as

$n \to \infty$.

Lastly, let us bound $P(A_{n,k3})$.

$$
\begin{aligned}
P(A_{n,k3}) &= P\Big(A_{n,k}^{+} \cap C_n \cap \{R_{n,k3} \geq \frac{1}{3} R_{n,k1}\}\Big) \\
&\leq P\Big(A_{n,k}^{+} \cap C_n \cap \Big\{\Big| \sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i \Big| \geq |(\hat{t}_k - t_k^0)(\mu_{k+1}^0 - \mu_k^0)|/3\Big\}\Big) \\
&\leq P\Big(\Big\{\Big|(t_k^0 - \hat{t}_k)^{-1} \sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i\Big| \geq |\mu_{k+1}^0 - \mu_k^0|/3\Big\} \quad ; t_k^0 - \hat{t}_k \geq \ n\delta_n\Big) \\
&\leq P\Big(\max_{\substack{1 \leq \hat{t}_k < t_k^0 \leq n \\ |t_k^0 - \hat{t}_k| \geq n\delta_n}} \Big|(t_k^0 - \hat{t}_k)^{-1} \sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i\Big| \geq J_{min}^0/3\Big).
\end{aligned}
$$

Then, by Lemma 2 with $x_n = J_{min}^0/3$ and $v_n = n\delta_n$, $P(A_{n,k3}) \to 0$ as $n \to \infty$. Consequently, $P(A_{n,k}^{+} \cap C_n) \to 0$ as $n \to \infty$.

Now, it remains to prove (1-ii). Without loss of generality, we will show that $P(A_{n,k}^{+} \cap C_n^c) \to 0$ as the other event, $A_{n,k}^{-} \cap C_n^c$, follows analogously. Define

$$
\begin{aligned}
D_n^{(l)} &\equiv \{\text{there exists } k \in \{1, ..., K^0\}, \hat{t}_k \leq t_{k-1}^0\} \cap C_n^c, \\
D_n^{(m)} &\equiv \{\text{for all } k \in \{1, ..., K^0\}, t_{k-1}^0 < \hat{t}_k < t_{k+1}^0\} \cap C_n^c, \\
D_n^{(r)} &\equiv \{\text{there exists } k \in \{1, ..., K^0\}, \hat{t}_k \geq t_{k+1}^0\} \cap C_n^c.
\end{aligned}
$$

Then, $P(A_{n,k}^{+} \cap C_n^c) = P(A_{n,k}^{+} \cap D_n^{(l)}) + P(A_{n,k}^{+} \cap D_n^{(m)}) + P(A_{n,k}^{+} \cap D_n^{(r)})$.

Let us first bound $P(A_{n,k}^{+} \cap D_n^{(m)})$.

$$
\begin{aligned}
P(A_{n,k}^{+} \cap D_n^{(m)}) &= P(A_{n,k}^{+} \cap \{\hat{t}_{k+1} - t_k^0 \geq I_{min}^0/2\} \cap D_n^{(m)}) \\
&\quad + P(A_{n,k}^{+} \cap \{\hat{t}_{k+1} - t_k^0 < I_{min}^0/2\} \cap D_n^{(m)}) \\
&\leq P(A_{n,k}^{+} \cap \{\hat{t}_{k+1} - t_k^0 \geq I_{min}^0/2\} \cap D_n^{(m)}) \\
&\quad + P(A_{n,k}^{+} \cap \{t_{k+1}^0 - \hat{t}_{k+1} \geq I_{min}^0/2\} \cap D_n^{(m)}) \qquad (4.15)
\end{aligned}
$$

$$\leq P\big(A_{n,k}^{+} \cap \{\hat{t}_{k+1} - t_k^0 \geq I_{min}^0/2\} \cap D_n^{(m)}\big) + \sum_{l=k+1}^{K^0-1} P\Big(\{t_l^0 - \hat{t}_l \geq I_{min}^0/2\} \cap \{\hat{t}_{l+1} - t_l^0 \geq I_{min}^0/2\} \cap D_n^{(m)}\Big) \tag{4.16}$$

For (4.15), $\hat{t}_{k+1} - t_k^0 \leq I_{min}^0/2$ implies that $t_{k+1}^0 - \hat{t}_{k+1} = (t_{k+1}^0 - t_k^0) - (\hat{t}_{k+1} - t_k^0) \geq I_{min}^0 - I_{min}^0/2 = I_{min}^0/2$. For (4.16), observe that $\{A_{n,k}^{+} \cap \{t_{k+1}^0 - \hat{t}_{k+1} \geq I_{min}^0/2\} \cap D_n^{(m)}\} \subset \cup_{l=k+1}^{K^0-1}\Big(\{t_l^0 - \hat{t}_l \geq I_{min}^0/2\} \cap \{\hat{t}_{l+1} - t_l^0 \geq I_{min}^0/2\} \cap D_n^{(m)}\Big)$.

We now bound the first term on the right-hand side of (4.16). Applying (i) in Lemma 1 with $l = k$ and (ii) in Lemma 1 with $j = t_k^0 + 1$, and using the triangle inequality and the fact that $|\hat{s}_l| = 1$, we have

$$|\hat{\mu}_{k+1} - \mu_k^0| \leq 2n\lambda\,(t_k^0 - \hat{t}_k)^{-1} + \Big|(t_k^0 - \hat{t}_k)^{-1} \sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i\Big|. \tag{4.17}$$

Also, applying (i) in Lemma 1 with $l = k+1$ and (ii) in Lemma 1 with $j = t_k^0 + 1$ yields

$$|\hat{\mu}_{k+1} - \mu_{k+1}^0| \leq 2n\lambda\,(\hat{t}_{k+1} - t_k^0)^{-1} + \Big|(\hat{t}_{k+1} - t_k^0)^{-1} \sum_{i=t_k^0+1}^{\hat{t}_{k+1}} \varepsilon_i\Big|. \tag{4.18}$$

Note that $|\mu_{k+1}^0 - \mu_k^0| = |\mu_{k+1}^0 - \hat{\mu}_{k+1} + \hat{\mu}_{k+1} - \mu_k^0| \leq |\mu_{k+1}^0 - \hat{\mu}_{k+1}| + |\hat{\mu}_{k+1} - \mu_k^0| = |\hat{\mu}_{k+1} - \mu_{k+1}^0| + |\hat{\mu}_{k+1} - \mu_k^0|$. Let us now define the event $E_{n,k} \equiv \{|\mu_{k+1}^0 - \mu_k^0| \leq |\hat{\mu}_{k+1} - \mu_{k+1}^0| + |\hat{\mu}_{k+1} - \mu_k^0|\}$. Then, substituting (4.17) and (4.18) into $E_{n,k}$ yields

$$\begin{aligned}
E_{n,k} &= \Big\{\Big|\mu_{k+1}^0 - \mu_k^0\Big| \leq 2n\lambda\,((t_k^0 - \hat{t}_k)^{-1} + (\hat{t}_{k+1} - t_k^0)^{-1}) \\
&\quad + \Big|(t_k^0 - \hat{t}_k)^{-1} \sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i\Big| + \Big|(\hat{t}_{k+1} - t_k^0)^{-1} \sum_{i=t_k^0+1}^{\hat{t}_{k+1}} \varepsilon_i\Big|\Big\} \\
&= \Big\{2n\lambda\,((t_k^0 - \hat{t}_k)^{-1} + (\hat{t}_{k+1} - t_k^0)^{-1}) \\
&\quad \geq |\mu_{k+1}^0 - \mu_k^0| - \Big|(t_k^0 - \hat{t}_k)^{-1} \sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i\Big| - \Big|(\hat{t}_{k+1} - t_k^0)^{-1} \sum_{i=t_k^0+1}^{\hat{t}_{k+1}} \varepsilon_i\Big|\Big\} \\
&\equiv \Big\{2n\lambda\,((t_k^0 - \hat{t}_k)^{-1} + (\hat{t}_{k+1} - t_k^0)^{-1})) \geq E_{n,k1} - E_{n,k2} - E_{n,k3}\Big\}.
\end{aligned} \tag{4.19}$$

Therefore, since $P(E_{n,k}) = 1$, the first term on the right hand side of (4.16) follows that

$$\begin{aligned}
&P\Big(A^+_{n,k} \cap \{\hat{t}_{k+1} - t^0_k \geq I^0_{min}/2\} \cap D^{(m)}_n\Big) \\
&\quad = P\Big(A^+_{n,k} \cap \{\hat{t}_{k+1} - t^0_k \geq I^0_{min}/2\} \cap D^{(m)}_n \cap E_{n,k}\Big) \\
&\quad \leq P\Big(A^+_{n,k} \cap \{\hat{t}_{k+1} - t^0_k \geq I^0_{min}/2\} \\
&\qquad\quad \cap \Big\{2n\lambda\,((t^0_k - \hat{t}_k)^{-1} + (\hat{t}_{k+1} - t^0_k)^{-1})) \geq \frac{1}{3}E_{n,k1}\Big\}\Big) \\
&\qquad + P\Big(A^+_{n,k} \cap \{\hat{t}_{k+1} - t^0_k \geq I^0_{min}/2\} \cap \{E_{n,k2} \geq \frac{1}{3}E_{n,k1}\}\Big) \\
&\qquad + P\Big(A^+_{n,k} \cap \{\hat{t}_{k+1} - t^0_k \geq I^0_{min}/2\} \cap \Big\{E_{n,k3} > \frac{1}{3}E_{n,k1}\Big\}\Big) \qquad (4.20) \\
&\quad \equiv P(A^{(m)}_{n,k1}) + P(A^{(m)}_{n,k2}) + P(A^{(m)}_{n,k3}).
\end{aligned}$$

(4.20) is obtained using Technique 1 by replacing Z_1, Z_2, Z_3, Z_4 with $2n\lambda\,((t^0_k - \hat{t}_k)^{-1} + (\hat{t}_{k+1} - t^0_k)^{-1}))$, $E_{n,k1}$, $E_{n,k2}$, $E_{n,k3}$, respectively.

Let us now bound $P(A^{(m)}_{n,k1})$.

$$\begin{aligned}
P(A^{(m)}_{n,k1}) &= P\Big(A^+_{n,k} \cap \{\hat{t}_{k+1} - t^0_k \geq I^0_{min}/2\} \\
&\qquad\quad \cap \Big\{2n\lambda\,((t^0_k - \hat{t}_k)^{-1} + (\hat{t}_{k+1} - t^0_k)^{-1})) \geq \frac{1}{3}E_{n,k1}\Big\}\Big) \\
&= P\Big(\Big\{2n\lambda\,((t^0_k - \hat{t}_k)^{-1} + (\hat{t}_{k+1} - t^0_k)^{-1})) \geq |\mu^0_{k+1} - \mu^0_k|/3\Big\} \\
&\qquad\quad ; t^0_k - \hat{t}_k \geq n\delta_n; \quad \hat{t}_{k+1} - t^0_k \geq I^0_{min}/2\Big) \\
&\leq P\Big(n\lambda\,(n\delta_n J^0_{min})^{-1} + 2n\lambda\,(I^0_{min} J^0_{min})^{-1} \geq 1/6\Big).
\end{aligned}$$

Then, by Assumptions A3 and A5, $n\lambda\,(n\delta_n J^0_{min})^{-1} = o(1)$ and $n\lambda\,(I^0_{min} J^0_{min})^{-1} = o(1)$. Thus, $P(A^{(m)}_{n,k1}) \to 0$ as $n \to \infty$.

Next we bound $P(A^{(m)}_{n,k2})$.

$$
\begin{aligned}
P(A^{(m)}_{n,k2}) &= P\Big(A^{+}_{n,k} \cap \{\hat{t}_{k+1} - t^0_k \geq I^0_{min}/2\} \cap \{E_{n,k2} \geq \frac{1}{3} E_{n,k1}\}\Big) \\
&\leq P\Big(\Big|(t^0_k - \hat{t}_k)^{-1} \sum_{i=\hat{t}_k+1}^{t^0_k} \varepsilon_i\Big| \geq |\mu^0_{k+1} - \mu^0_k|/3 \quad ; t^0_k - \hat{t}_k \geq n\delta_n\Big) \\
&\leq P\Big(\max_{\substack{1 \leq \hat{t}_k < t^0_k \leq n \\ t^0_k - \hat{t}_k \geq n\delta_n}} \Big|(t^0_k - \hat{t}_k)^{-1} \sum_{i=\hat{t}_k+1}^{t^0_k} \varepsilon_i\Big| \geq J^0_{min}/3\Big).
\end{aligned}
$$

Then, by Lemma 2 with $x_n = J^0_{min}/3$ and $\nu_n = n\delta_n$, $P(A^{(m)}_{n,k2}) \to 0$ as $n \to \infty$. Similarly, by Lemma 2 with $x_n = J^0_{min}/3$ and $\nu_n = I^0_{min}/2$, $P(A^{(m)}_{n,k3}) \to 0$ as $n \to \infty$. Therefore, we have proved that the first term on the right-hand side of (4.16) tends to zero as n tends to infinity. By a similar argument, the second term on the right-hand side of (4.16) also tends to zero. Hence, $P(A^{+}_{n,k} \cap D^{(m)}_n) \to 0$ as $n \to \infty$.

Now consider $P(A^{+}_{n,k} \cap D^{(l)}_n)$. Observe that $P(A^{+}_{n,k} \cap D^{(l)}_n) \leq P(D^{(l)}_n) \leq \sum_{k=1}^{K^0} P(\max\{1 \leq l \leq K^0, \hat{t}_l \leq t^0_{l-1}\} = k)$. The event $\{\max\{1 \leq l \leq K^0, \hat{t}_l \leq t^0_{l-1}\} = k\}$ implies that $\hat{t}_k \leq t^0_{k-1}$ and $\hat{t}_{m+1} > t^0_m$ for all $m \in \{k, ..., K^0\}$. Thus, $\{\max\{1 \leq l \leq K^0, \hat{t}_l \leq t^0_{l-1}\} = k\} \subset \cup_{m=k}^{K^0-1} \Big(\{t^0_m - \hat{t}_m \geq I^0_{min}/2\} \cap \{\hat{t}_{m+1} - t^0_m \geq I^0_{min}/2\}\Big)$. Then,

$$
\begin{aligned}
P(A^{+}_{n,k} \cap D^{(l)}_n) &\leq \sum_{k=1}^{K^0} \sum_{m=k}^{K^0-1} P\Big(\{t^0_m - \hat{t}_m \geq I^0_{min}/2\} \cap \{\hat{t}_{m+1} - t^0_m \geq I^0_{min}/2\}\Big) \\
&\leq \sum_{k=1}^{K^0-1} \sum_{m=k}^{K^0-1} P\Big(\{t^0_m - \hat{t}_m \geq I^0_{min}/2\} \cap \{\hat{t}_{m+1} - t^0_m \geq I^0_{min}/2\}\Big) \\
&\quad + P\Big(\{t^0_{K^0} - \hat{t}_{K^0} \geq I^0_{min}/2\} \cap \{\hat{t}_{K^0+1} - t^0_{K^0} \geq I^0_{min}/2\}\Big). \qquad (4.21)
\end{aligned}
$$

Let us first bound the term in the summation on the right-hand side of (4.21). Applying $k = m$ in (4.19) suggests that $P(E_{n,m}) = 1$. Then,

$$
P\Big(\{t^0_m - \hat{t}_m \geq I^0_{min}/2\} \cap \{\hat{t}_{m+1} - t^0_m \geq I^0_{min}/2\}\Big)
$$

$$
\begin{aligned}
&= P\Big(\{t_m^0 - \hat{t}_m \geq I_{min}^0/2\} \cap \{\hat{t}_{m+1} - t_m^0 \geq I_{min}^0/2\} \cap E_{n,m}\Big)\\
&= P\Big(\{t_m^0 - \hat{t}_m \geq I_{min}^0/2\} \cap \{\hat{t}_{m+1} - t_m^0 \geq I_{min}^0/2\}\\
&\qquad \cap \{2n\lambda((t_m^0 - \hat{t}_m)^{-1} + (\hat{t}_{m+1} - t_m^0)^{-1})) \geq E_{n,m1} - E_{n,m2} - E_{n,m3}\}\Big)\\
&\leq P\Big(\{t_m^0 - \hat{t}_m \geq I_{min}^0/2\} \cap \{\hat{t}_{m+1} - t_m^0 \geq I_{min}^0/2\}\\
&\qquad \cap \{2n\lambda((t_m^0 - \hat{t}_m)^{-1} + (\hat{t}_{m+1} - t_m^0)^{-1}) \geq |\mu_{m+1}^0 - \mu_m^0|/3\}\Big)\\
&\quad + P\Big(\{t_m^0 - \hat{t}_m \geq I_{min}^0/2\} \cap \{\hat{t}_{m+1} - t_m^0 \geq I_{min}^0/2\}\\
&\qquad \cap \{\Big|(t_m^0 - \hat{t}_m)^{-1} \sum_{i=\hat{t}_m+1}^{t_m^0} \varepsilon_i\Big| \geq |\mu_{m+1}^0 - \mu_m^0|/3\}\Big)\\
&\quad + P\Big(\{t_m^0 - \hat{t}_m \geq I_{min}^0/2\} \cap \{\hat{t}_{m+1} - t_m^0 \geq I_{min}^0/2\}\\
&\qquad \cap \{\Big|(\hat{t}_{m+1} - t_m^0)^{-1} \sum_{i=t_m^0+1}^{\hat{t}_{m+1}} \varepsilon_i\Big| \geq |\mu_{m+1}^0 - \mu_m^0|/3\}\Big) \qquad (4.22)\\
&\equiv P(A_{n,m1}^{(l)}) + P(A_{n,m2}^{(l)}) + P(A_{n,m3}^{(l)}),
\end{aligned}
$$

where $E_{n,m1} = |\mu_{m+1}^0 - \mu_m^0|$, $E_{n,m2} = \Big|(t_m^0 - \hat{t}_m)^{-1} \sum_{i=\hat{t}_m+1}^{t_m^0} \varepsilon_i\Big|$, and $E_{n,m3} = \Big|(\hat{t}_{m+1} - t_m^0)^{-1} \sum_{i=t_m^0+1}^{\hat{t}_{m+1}} \varepsilon_i\Big|$. (4.22) is obtained using Technique 1 by replacing Z_1, Z_2, Z_3, Z_4 with $2n\lambda((t_m^0 - \hat{t}_m)^{-1} + (\hat{t}_{m+1} - t_m^0)^{-1})$, $E_{n,m1}$, $E_{n,m2}$, $E_{n,m3}$, respectively.

For $P(A_{n,m1}^{(l)})$,

$$
\begin{aligned}
P(A_{n,m1}^{(l)}) &= P\Big(\{2n\lambda((t_m^0 - \hat{t}_m)^{-1} + (\hat{t}_{m+1} - t_m^0)^{-1}) \geq |\mu_{m+1}^0 - \mu_m^0|/3\}\\
&\qquad ; t_m^0 - \hat{t}_m \geq I_{min}^0/2 \quad ; \hat{t}_{m+1} - t_m^0 \geq I_{min}^0/2\Big)\\
&\leq P\Big(n\lambda(I_{min}^0 J_{min}^0)^{-1} \geq 1/24\Big)
\end{aligned}
$$

Then, by Assumptions A3 and A5, $n\lambda(I^0_{min}J^0_{min})^{-1} = o(1)$, and thus $P(A^{(l)}_{n,m1}) \to 0$ as $n \to \infty$.

Next,

$$
\begin{aligned}
P(A^{(l)}_{n,m2}) &\le P\Big(\Big|(t^0_m - \hat{t}_m)^{-1} \sum_{i=\hat{t}_m+1}^{t^0_m} \varepsilon_i\Big| \ge |\mu^0_{m+1} - \mu^0_m|/3 \quad ; t^0_m - \hat{t}_m \ge I^0_{min}/2\Big) \\
&\le P\Big(\max_{\substack{1\le \hat{t}_m < t^0_m \le n \\ t^0_m - \hat{t}_m \ge I^0_{min}/2}} \Big|(t^0_m - \hat{t}_m)^{-1} \sum_{i=\hat{t}_m+1}^{t^0_m} \varepsilon_i\Big| \ge J^0_{min}/3\Big).
\end{aligned}
$$

Then, by Lemma 2 with $x_n = J^0_{min}/3$ and $\nu_n = I^0_{min}/2$, $P(A^{(l)}_{n,m2}) \to 0$ as $n \to \infty$. Similarly, by Lemma 2 with $x_n = J^0_{min}/3$ and $\nu_n = I^0_{min}/2$, $P(A^{(l)}_{n,m3}) \to 0$. Hence, we have proved that the term in the summation on the right-hand side of (4.21) tends to zero as n tends to infinity.

Similarly, we now bound the last term on the right-hand side of (4.21). Applying $k = K^0$ yields that

$$
\begin{aligned}
&P\Big(\{t^0_{K^0} - \hat{t}_{K^0} \ge I^0_{min}/2\} \cap \{\hat{t}_{K^0+1} - t^0_{K^0} \ge I^0_{min}/2\}\Big) \\
&= P\Big(\{t^0_{K^0} - \hat{t}_{K^0} \ge I^0_{min}/2\} \cap \{\hat{t}_{K^0+1} - t^0_{K^0} \ge I^0_{min}/2\} \cap E_{n,K^0}\Big) \\
&= P\Big(\{t^0_{K^0} - \hat{t}_{K^0} \ge I^0_{min}/2\} \cap \{\hat{t}_{K^0+1} - t^0_{K^0} \ge I^0_{min}/2\} \\
&\qquad \cap \{2n\lambda((t^0_{K^0} - \hat{t}_{K^0})^{-1} + (\hat{t}_{K^0+1} - t^0_{K^0})^{-1}) \ge E_{n,K^0_1} - E_{n,K^0_2} - E_{n,K^0_3}\}\Big) \\
&\le P\Big(\{t^0_{K^0} - \hat{t}_{K^0} \ge I^0_{min}/2\} \cap \{\hat{t}_{K^0+1} - t^0_{K^0} \ge I^0_{min}/2\} \\
&\qquad \cap \{2n\lambda((t^0_{K^0} - \hat{t}_{K^0})^{-1} + (\hat{t}_{K^0+1} - t^0_{K^0})^{-1}) \ge |\mu^0_{K^0+1} - \mu^0_{K^0}|/3\}\Big) \\
&\quad + P\Big(\{t^0_{K^0} - \hat{t}_{K^0} \ge I^0_{min}/2\} \cap \{\hat{t}_{K^0+1} - t^0_{K^0} \ge I^0_{min}/2\} \\
&\qquad \cap \{\Big|(t^0_{K^0} - \hat{t}_{K^0})^{-1} \sum_{i=\hat{t}_{K^0}+1}^{t^0_{K^0}} \varepsilon_i\Big| \ge |\mu^0_{K^0+1} - \mu^0_{K^0}|/3\}\Big) \\
&\quad + P\Big(\{t^0_{K^0} - \hat{t}_{K^0} \ge I^0_{min}/2\} \cap \{\hat{t}_{K^0+1} - t^0_{K^0} \ge I^0_{min}/2\} \\
&\qquad \cap \{\Big|(\hat{t}_{K^0+1} - t^0_{K^0})^{-1} \sum_{i=t^0_{K^0}+1}^{\hat{t}_{K^0+1}} \varepsilon_i\Big| \ge |\mu^0_{K^0+1} - \mu^0_{K^0}|/3\}\Big). \qquad (4.23) \\
&\equiv P(A^{(l)}_{n,K^0_1}) + P(A^{(l)}_{n,K^0_2}) + P(A^{(l)}_{n,K^0_3}),
\end{aligned}
$$

where $E_{n,K_1^0} = |\mu^0_{K^0+1} - \mu^0_{K^0}|$, $E_{n,K_2^0} = \left|(t^0_{K^0} - \hat{t}_{K^0})^{-1}\sum_{i=\hat{t}_{K^0}+1}^{t^0_{K^0}} \varepsilon_i\right|$, and $E_{n,K_3^0} = \left|(\hat{t}_{K^0+1} - t^0_{K^0})^{-1}\sum_{i=t^0_{K^0}+1}^{\hat{t}_{K^0+1}} \varepsilon_i\right|$. (4.23) is obtained using Technique 1 by replacing Z_1, Z_2, Z_3, Z_4 with $2n\lambda((t^0_{K^0} - \hat{t}_{K^0})^{-1} + (\hat{t}_{K^0+1} - t^0_{K^0})^{-1})$, E_{n,K_1^0}, E_{n,K_2^0}, E_{n,K_3^0}, respectively. Then, as $n \to \infty$, $P(A^{(l)}_{n,K_1^0}) \to 0$ by Assumptions A3 and A5, $P(A^{(l)}_{n,K_2^0}) \to 0$ by Lemma 2 with $x_n = J^0_{min}/3$ and $v_n = I^0_{min}/2$, and $P(A^{(l)}_{n,K_3^0}) \to 0$ by Lemma 2 with $x_n = J^0_{min}/3$ and $v_n = I^0_{min}/2$. Consequently, $P(A^+_{n,k} \cap D^{(l)}_n)$ tends to zero as n tends to infinity. Also, in a similar way, we can prove that $P(A^+_{n,k} \cap D^{(r)}_n) \to 0$, which yields that $P(A^+_{n,k} \cap C^c_n) \to 0$. Therefore, we complete the proof. □

Theorem 2. Under Assumptions A1-A5, if $\hat{K} = K^0$, then

$$|\mu^0_k - \hat{\mu}_k| = O_p(n\lambda/I^0_k + n\delta_n/I^0_k + (I^0_k)^{-1/2}) \text{ for each } k = 1, \ldots, K^0 + 1.$$

Proof. Our proof closely follows Theorem 3.1(ii) in Qian and Su (2016). By the result in Theorem 1 and Assumption A3, $|\hat{t}_k - t^0_k| = O_p(n\delta_n) = o_p(I^0_{min})$ uniformly in $k = 1, \ldots, K^0$. It follows that either $(t^0_{k-1} + t^0_k)/2 < \hat{t}_k < t^0_k$ or $t^0_k \le \hat{t}_k < (t^0_k + t^0_{k+1})/2$ holds for each k. Fix $k \in \{1, \ldots, K^0\}$. Without loss of generality, we assume that $(t^0_{k-1} + t^0_k)/2 < \hat{t}_k < t^0_k$ and consider two subcases: (2-i) $(t^0_k + t^0_{k+1})/2 < \hat{t}_{k+1} < t^0_{k+1}$, and (2-ii) $t^0_{k+1} < \hat{t}_{k+1}$.

First, we will prove that Theorem 2 holds for all $k = 2, \ldots, K^0 + 1$ in subcase (2-i). Applying (i) in Lemma 1 with $l = k$ and $l = k+1$, and using the triangle inequality and the fact that $|\hat{s}_l| = 1$, we have

$$\begin{aligned}
2n\lambda &\ge \Big| \sum_{i=\hat{t}_k+1}^{n} (y_i - \hat{\beta}_i) - \sum_{i=\hat{t}_{k+1}+1}^{n} (y_i - \hat{\beta}_i) \Big| \\
&= \Big| \sum_{i=\hat{t}_k+1}^{t^0_k} ((\mu^0_k - \hat{\mu}_{k+1}) + \varepsilon_i) + \sum_{i=t^0_k+1}^{\hat{t}_{k+1}} ((\mu^0_{k+1} - \hat{\mu}_{k+1}) + \varepsilon_i) \Big| \\
&\ge \left| (\hat{t}_{k+1} - t^0_k)(\mu^0_{k+1} - \hat{\mu}_{k+1}) \right| - \Big| \sum_{i=t^0_k+1}^{\hat{t}_{k+1}} \varepsilon_i \Big| - \left| (t^0_k - \hat{t}_k)|\mu^0_k - \hat{\mu}_{k+1}| + \sum_{i=\hat{t}_k+1}^{t^0_k} \varepsilon_i \right| \\
&= (\hat{t}_{k+1} - t^0_k)|\mu^0_{k+1} - \hat{\mu}_{k+1}| - O_p((I^0_{k+1})^{1/2}) - O_p(n\delta_n) \qquad (4.24)
\end{aligned}$$

$$\geq (t_{k+1}^0 - t_k^0)|\mu_{k+1}^0 - \hat{\mu}_{k+1}|/2 - O_p((I_{k+1}^0)^{1/2}) - O_p(n\delta_n), \tag{4.25}$$

where $\hat{\beta}_i = \hat{\mu}_{k+1}$, $\beta_i^0 = \mu_k^0$ for $i \in [\hat{t}_k + 1, t_k^0]$, and $\hat{\beta}_i = \hat{\mu}_{k+1}, \beta_i^0 = \mu_{k+1}^0$ for $i \in [t_k^0 + 1, \hat{t}_{k+1}]$.

For (4.24), $\sum_{i=t_k^0+1}^{\hat{t}_{k+1}} \varepsilon_i = O_p\Big([(\hat{t}_{k+1} - t_k^0)\sigma^2]^{1/2}\Big)$ by Chebyshev's inequality. Then, since $\hat{t}_{k+1} < t_{k+1}^0$ in subcase (2-i), $\hat{t}_{k+1} - t_k^0 < t_{k+1}^0 - t_k^0 = I_{k+1}^0$. Thus, $\sum_{i=t_k^0+1}^{\hat{t}_{k+1}} \varepsilon_i = O_p((I_{k+1}^0)^{1/2})$. Similarly, by Chebyshev's inequality, $\sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i = O_p\Big([(t_k^0 - \hat{t}_k)\sigma^2]^{1/2}\Big)$. Then, since $t_k^0 - \hat{t}_k \leq n\delta_n$ by Theorem 1, $\sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i = O_p((n\delta_n)^{1/2})$. Also, since $|\mu_k^0 - \hat{\mu}_{k+1}|$ is a constant as n tends to infinity, $(t_k^0 - \hat{t}_k)|\mu_k^0 - \hat{\mu}_{k+1}| = O_p(n\delta_n)$.

For (4.25), in subcase (2-i), $(t_k^0 + t_{k+1}^0)/2 < \hat{t}_{k+1}$, implying that $(t_{k+1}^0 - t_k^0)/2 < \hat{t}_{k+1} - t_k^0$.

Next, consider subcase (2-ii). Applying (i) in Lemma 1 with $l = k$ and $l = k+1$, and using the triangle inequality and the fact that $|\hat{s}_l| = 1$, we obtain

$$\begin{aligned}
2n\lambda &\geq \Big| \sum_{i=\hat{t}_k+1}^{n} (y_i - \hat{\beta}_i) - \sum_{i=\hat{t}_{k+1}+1}^{n} (y_i - \hat{\beta}_i) \Big| \\
&\geq \Big|(t_{k+1}^0 - t_k^0)(\mu_{k+1}^0 - \hat{\mu}_{k+1}) + \sum_{i=t_k^0+1}^{t_{k+1}^0} \varepsilon_i\Big| - \Big|(t_k^0 - \hat{t}_k)(\mu_k^0 - \hat{\mu}_{k+1}) + \sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i\Big| \\
&\quad - \Big|(\hat{t}_{k+1} - t_{k+1}^0)(\mu_{k+2}^0 - \hat{\mu}_{k+1}) + \sum_{i=t_{k+1}^0+1}^{\hat{t}_{k+1}} \varepsilon_i\Big| \\
&\geq (t_{k+1}^0 - t_k^0)|\mu_{k+1}^0 - \hat{\mu}_{k+1}| - \Big|\sum_{i=t_k^0+1}^{t_{k+1}^0} \varepsilon_i\Big| - \Big|(t_k^0 - \hat{t}_k)(\mu_k^0 - \hat{\mu}_{k+1}) + \sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i\Big| \\
&\quad - \Big|(\hat{t}_{k+1} - t_{k+1}^0)(\mu_{k+2}^0 - \hat{\mu}_{k+1}) + \sum_{i=t_{k+1}^0+1}^{\hat{t}_{k+1}} \varepsilon_i\Big| \\
&= (t_{k+1}^0 - t_k^0)|\mu_{k+1}^0 - \hat{\mu}_{k+1}| - O_p((I_{k+1}^0)^{1/2}) - O_p(n\delta_n) - O_p(n\delta_n),
\end{aligned} \tag{4.26}$$

$\hat{\beta}_i = \hat{\mu}_{k+1}, \beta_i^0 = \mu_k^0$ for $i \in [\hat{t}_k + 1, t_k^0]$ and $\hat{\beta}_i = \hat{\mu}_{k+1}, \beta_i^0 = \mu_{k+1}^0$ for $i \in [t_k^0 + 1, t_{k+1}^0]$, and $\hat{\beta}_i = \hat{\mu}_{k+1}, \beta_i^0 = \mu_{k+2}^0$ for $i \in [t_{k+1}^0 + 1, \hat{t}_{k+1}]$.

For (4.26), by Chebyshev's inequality, $\sum_{i=t_k^0+1}^{t_{k+1}^0} \varepsilon_i = O_p((I_{k+1}^0)^{1/2})$. By Chebyshev's inequality and Theorem 1, $\sum_{i=\hat{t}_k+1}^{t_k^0} \varepsilon_i = O_p((n\delta_n)^{1/2})$ and $\sum_{i=t_{k+1}^0+1}^{\hat{t}_{k+1}} \varepsilon_i = O_p((n\delta_n)^{1/2})$. Also, since $|\mu_k^0 - \hat{\mu}_{k+1}|$ and

$|\mu^0_{k+2} - \hat{\mu}_{k+1}|$ are constants as n tends to infinity, $(t^0_k - \hat{t}_k)|\mu^0_k - \hat{\mu}_{k+1}| = O_p(n\delta_n)$ and $(\hat{t}_{k+1} - t^0_{k+1})|\mu^0_{k+2} - \hat{\mu}_{k+1}| = O_p(n\delta_n)$.

Therefore, by (4.25) and (4.26), Theorem 2 holds for all $k = 2, ..., K^0 + 1$.

Let us prove now Theorem 2 for $k = 1$. Applying (i) in Lemma 1 with $l = 1$ and (ii) in Lemma 1 with $j = 1$, and using the triangle inequality and the fact that $|\hat{s}_1| = 1$, we have

$$2n\lambda \geq \Big|\sum_{i=1}^{\hat{t}_1}(\beta^0_i - \hat{\beta}_i) + \sum_{i=1}^{\hat{t}_1}\varepsilon_i\Big|. \tag{4.27}$$

Consider the case when $t^0_1/2 < \hat{t}_1 < t^0_1$. Then, $\hat{\beta}_i = \hat{\mu}_1$ and $\beta^0_i = \mu^0_1$ for $i \in [1, \hat{t}_1]$. Thus, (4.27) follows that

$$\begin{aligned}
2n\lambda &\geq \Big|\sum_{i=1}^{\hat{t}_1}(\mu^0_1 - \hat{\mu}_1) + \Big(\sum_{i=1}^{t^0_1}\varepsilon_i - \sum_{i=\hat{t}_1+1}^{t^0_1}\varepsilon_i\Big)\Big| \\
&\geq \hat{t}_1|\mu^0_1 - \hat{\mu}_1| - \Big|\sum_{i=1}^{t^0_1}\varepsilon_i\Big| - \Big|\sum_{i=\hat{t}_1+1}^{t^0_1}\varepsilon_i\Big| \\
&= \hat{t}_1|\mu^0_1 - \hat{\mu}_1| - O_p((I^0_1)^{1/2}) - O_p(n\delta_n). \\
&\geq \frac{1}{2}t^0_1|\mu^0_1 - \hat{\mu}_1| - O_p((I^0_1)^{1/2}) - O_p(n\delta_n).
\end{aligned}$$

By Chebyshev's inequality, $\sum_{i=1}^{t^0_1}\varepsilon_i = O_p((I^0_1)^{1/2})$. Also, Chebyshev's inequality and Theorem 1, $\sum_{i=\hat{t}_1+1}^{t^0_1}\varepsilon_i = O_p(n\delta_n)$.

In a similar way, in the case when $\hat{t}_1 > t^0_1$, we obtain $|\mu^0_1 - \hat{\mu}_1| = O_p(n\lambda/I^0_1 + n\delta_n/I^0_1 + (I^0_1)^{-1/2})$. Therefore, we complete the proof. □

Theorem 3. Under A1-A5, if $K^0 \leq \hat{K} \leq K_{max}$, then

$$P(\mathscr{D}((\hat{\mathbf{t}}_{\hat{K}}, \mathbf{t}^0_{K^0}) \leq n\delta_n) \to 1 \text{ as } n \to \infty,$$

where $\mathscr{D}(A,B) \equiv \sup_{b\in B}\inf_{a\in A}|a-b|$ for any two sets A and B.

Proof. Our proof is adapted from Proposition 4 of Harchaoui and Lévy-Leduc (2012) and Theorem 3.2 in Qian and Su (2016). Given Theorem 1, the proof is completed by showing that $P(\{\mathscr{D}(\hat{\mathbf{t}}_{\hat{K}}, \mathbf{t}^0_{K^0}) > n\delta_n)\} \cap \{K^0 < \hat{K} \leq K_{max}\}) \to 0$ as $n \to \infty$. We start with the observation that

$$\begin{aligned}
&P(\{\mathscr{D}(\hat{\mathbf{t}}_{\hat{K}}, \mathbf{t}^0_{K^0}) > n\delta_n)\} \cap \{K^0 < \hat{K} \leq K_{max}\}) \\
&\leq \sum_{K=K^0+1}^{K_{max}} P(\mathscr{D}(\hat{\mathbf{t}}_K, \mathbf{t}^0_{K^0}) > n\delta_n) \\
&\leq \sum_{K=K^0+1}^{K_{max}} \sum_{k=1}^{K^0} P(\forall l, 1 \leq l \leq K, |\hat{t}_l - t^0_k| > n\delta_n) \\
&\equiv \sum_{K=K^0+1}^{K_{max}} \sum_{k=1}^{K^0} (P(R_{K,k1}) + P(R_{K,k3}) + P(R_{K,k3})),
\end{aligned} \tag{4.28}$$

where

$$\begin{aligned}
R_{K,k1} &= \{\text{for all} \quad l \in \{1, ..., K\}, \quad |\hat{t}_l - t^0_k| \geq n\delta_n \quad \text{and} \quad \hat{t}_l < t^0_k\} \\
R_{K,k2} &= \{\text{for all} \quad l \in \{1, ..., K\}, \quad |\hat{t}_l - t^0_k| \geq n\delta_n \quad \text{and} \quad \hat{t}_l > t^0_k\} \\
R_{K,k3} &= \{\text{there exists} \quad l \in \{1, ..., K-1\}, \\
&\qquad |\hat{t}_l - t^0_k| \geq n\delta_n, \quad |\hat{t}_{l+1} - t^0_k| \geq n\delta_n, \quad \text{and} \quad \hat{t}_l < t^0_k < \hat{t}_{l+1}\}.
\end{aligned}$$

Then, the procedure to find an upper bound of $P(\{\mathscr{D}(\hat{\mathbf{t}}_{\hat{K}}, \mathbf{t}^0_{K^0}) > n\delta_n)\} \cap \{K^0 < \hat{K} \leq K_{max}\})$ is similar to the one we used in Theorem 1. Let us first bound $P(R_{K,k1})$. Note that

$$\begin{aligned}
P(R_{K,k1}) &= P(R_{K,k1} \cap (\{\hat{t}_K > t^0_{k-1}\} \cup \{\hat{t}_K \leq t^0_{k-1}\})) \\
&= P(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\})) + P(R_{K,k1} \cap \{\hat{t}_K \leq t^0_{k-1}\})).
\end{aligned} \tag{4.29}$$

Without loss of generality, our goal is to prove that $P(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\})) \to 0$ as the other case follows analogously.

In the event $R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\}$, observe that

$$t^0_{k-1} < \hat{t}_K < t^0_k. \tag{4.30}$$

Applying (i) in Lemma 1 with $l = K$ and (ii) in Lemma 1 with $j = t^0_k + 1$, and using the triangle inequality and the fact that $|\hat{s}_K| = 1$, we obtain

$$\begin{aligned} 2n\lambda &\geq \Big| \sum_{i=\hat{t}_K+1}^{n} (y_i - \hat{\beta}_i) - \sum_{i=t^0_k+1}^{n} (y_i - \hat{\beta}_i) \Big| \\ &\geq (t^0_k - \hat{t}_K)|\mu^0_k - \mu^0_{k+1}| - (t^0_k - \hat{t}_K)|\mu^0_{k+1} - \hat{\mu}_{K+1}| - \Big| \sum_{i=\hat{t}_K+1}^{t^0_k} \varepsilon_i \Big| \\ &\equiv Q_{n,k1} - Q_{n,k2} - Q_{n,k3}, \end{aligned} \tag{4.31}$$

where $\hat{\beta}_i = \hat{\mu}_{K+1}$ and $\beta^0_i = \mu^0_k$ for $i \in [\hat{t}_K + 1, t^0_k]$.

Define the event $Q_{n,k} = \{2n\lambda \geq Q_{n,k1} - Q_{n,k2} - Q_{n,k3}\}$. Since $P(Q_{n,k}) = 1$, $P\Big(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\}\Big) = P\Big(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\} \cap Q_{n,k}\Big) = P\Big(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\} \cap \{2n\lambda \geq Q_{n,k1} - Q_{n,k2} - Q_{n,k3}\}\Big)$. Then, using Technique 1, we have

$$\begin{aligned} P\Big(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\}\Big) &\leq P\Big(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\} \cap \{2n\lambda \geq \frac{1}{3}Q_{n,k1}\}\Big) \\ &\quad + P\Big(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\} \cap \{Q_{n,k2} \geq \frac{1}{3}Q_{n,k1}\}\Big) \\ &\quad + P\Big(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\} \cap \{Q_{n,k3} \geq \frac{1}{3}Q_{n,k1}\}\Big) \\ &\equiv P(R^{(1)}_1) + P(R^{(2)}_1) + P(R^{(3)}_1). \end{aligned} \tag{4.32}$$

Let us first bound $P(R^{(1)}_1)$.

$$P(R^{(1)}_1) = P\Big(R_{K,k1} \cap \{\hat{t}_K > t^0_{k-1}\} \cap \{2n\lambda \geq (t^0_k - \hat{t}_K)|\mu^0_k - \mu^0_{k+1}|/3\}\Big)$$

$$\le P\Big(2n\lambda \ge (t_k^0 - \hat{t}_K)|\mu_k^0 - \mu_{k+1}^0|/3; \quad t_k^0 - \hat{t}_K \ge n\delta_n\Big)$$
$$\le P\Big(n\lambda(n\delta_n J_{min}^0)^{-1} \ge 1/6\Big). \tag{4.33}$$

Then, by Assumption A5, $n\lambda(n\delta_n J_{min}^0)^{-1} < 1/6$ for n large enough, and thus, $P(R_1^{(1)}) \to 0$ as $n \to \infty$.

Next we bound $P(R_1^{(2)})$. Applying (ii) in Lemma 1 with $j = t_k^0 + 1$ and $j = t_{k+1}^0 + 1$, and using the triangle inequality, we obtain

$$|\mu_{k+1}^0 - \hat{\mu}_{K+1}| \le 2n\lambda(t_{k+1}^0 - t_k^0)^{-1} + \Big|(t_{k+1}^0 - t_k^0)^{-1} \sum_{i=t_k^0+1}^{t_{k+1}^0} \varepsilon_i\Big|. \tag{4.34}$$

Therefore, we upper bound $P(R_1^{(2)})$ as follows:

$$P(R_1^{(2)}) = P\Big(R_{K,k1} \cap \{\hat{t}_K > t_{k-1}^0\} \cap \{(t_k^0 - \hat{t}_K)|\mu_{k+1}^0 - \hat{\mu}_{K+1}| \ge (t_k^0 - \hat{t}_K)|\mu_k^0 - \mu_{k+1}^0|/3\}\Big)$$
$$\le P\Big(|\mu_{k+1}^0 - \hat{\mu}_{K+1}| \ge |\mu_k^0 - \mu_{k+1}^0|/3\Big)$$
$$\le P\Big(2n\lambda(t_{k+1}^0 - t_k^0)^{-1} + \Big|(t_{k+1}^0 - t_k^0)^{-1} \sum_{i=t_k^0+1}^{t_{k+1}^0} \varepsilon_i\Big| \ge |\mu_k^0 - \mu_{k+1}^0|/3\Big) \tag{4.35}$$
$$\le P\Big(n\lambda(t_{k+1}^0 - t_k^0)^{-1} \ge |\mu_k^0 - \mu_{k+1}^0|/12\Big)$$
$$+ P\Big(\Big|(t_{k+1}^0 - t_k^0)^{-1} \sum_{i=t_k^0+1}^{t_{k+1}^0} \varepsilon_i\Big| \ge |\mu_k^0 - \mu_{k+1}^0|/6\Big) \tag{4.36}$$

(4.35) follows from (4.34). (4.36) is obtained using Technique 2 by replacing Z_1, Z_2, Z_3 with $2n\lambda(t_{k+1}^0 - t_k^0)^{-1}$, $\Big|(t_{k+1}^0 - t_k^0)^{-1} \sum_{i=t_k^0+1}^{t_{k+1}^0} \varepsilon_i\Big|$, $|\mu_k^0 - \mu_{k+1}^0|/3$, respectively.

Furthermore, as n goes to infinity, the first term on the right-hand side of (4.36) tends to zero by Assumptions A3 and A5, and also the second term tends to zero by Assumption A3 and Lemma 2 with $x_n = J_{min}^0/6$ and $v_n = n\delta_n$. Therefore, $P(R_1^{(2)}) \to 0$ as $n \to \infty$.

For $P(R_1^{(3)})$,

$$P(R_1^{(3)}) = P\Big(R_{K,k1} \cap \{\hat{t}_K > t_{k-1}^0\} \cap \{\Big|\sum_{i=\hat{t}_K+1}^{t_k^0} \varepsilon_i\Big| \geq (t_k^0 - \hat{t}_K)|\mu_k^0 - \mu_{k+1}^0|/3\}\Big)$$

$$\leq P\Big(\Big|(t_k^0 - \hat{t}_K)^{-1} \sum_{i=\hat{t}_K+1}^{t_k^0} \varepsilon_i\Big| \geq J_{min}^0/3 \quad ; t_k^0 - \hat{t}_K \geq n\delta_n\Big)$$

$$\leq P\Big(\max_{\substack{1 \leq t_k^0 < \hat{t}_K \leq n \\ t_k^0 - \hat{t}_K \geq n\delta_n}} \Big|(t_k^0 - \hat{t}_K)^{-1} \sum_{i=\hat{t}_K+1}^{t_k^0} \varepsilon_i\Big| \geq J_{min}^0/3\Big).$$

Then, by Lemma 2 with $x_n = J_{min}^0/3$ and $v_n = n\delta_n$, $P(R_1^{(3)}) \to 0$ as $n \to \infty$. Hence, we have proved that $P(R_{K,k1}) \to 0$ as $n \to \infty$. Similarly, we can also prove that $P(R_{K,k2}) \to 0$ as $n \to \infty$.

Lastly, let us bound $P(R_{K,k3})$. Note that $P(R_{K,k3})$ can be split into four terms as follows:

$$P(R_{K,k3}) = P(R_3^{(1)}) + P(R_3^{(2)}) + P(R_3^{(3)}) + P(R_3^{(4)})$$

$$\begin{aligned} \text{,where} \quad R_3^{(1)} &= R_{K,k3} \cap \{t_{k-1}^0 < \hat{t}_l < \hat{t}_{l+1} < t_{k+1}^0\} \\ R_3^{(2)} &= R_{K,k3} \cap \{t_{k-1}^0 < \hat{t}_l < t_{k+1}^0, \quad \hat{t}_{l+1} > t_{k+1}^0\} \\ R_3^{(3)} &= R_{K,k3} \cap \{\hat{t}_l \leq t_{k-1}^0, \quad t_{k-1}^0 < \hat{t}_{l+1} < t_{k+1}^0\}, \quad \text{and} \\ R_3^{(4)} &= R_{K,k3} \cap \{\hat{t}_l \leq t_{k-1}^0, \quad t_{k+1}^0 < \hat{t}_{l+1}\}. \end{aligned}$$

In the event $R_3^{(1)}$, there exists $l \in \{1, ..., K-1\}$, $\hat{t}_l < t_k^0 < \hat{t}_{l+1}$, and $t_{k-1}^0 < \hat{t}_l < \hat{t}_{l+1} < t_{k+1}^0$, which implies that

$$t_{k-1}^0 < \hat{t}_l < t_k^0 < \hat{t}_{l+1} < t_{k+1}^0. \tag{4.37}$$

Applying (i) in Lemma 1 with $l = l$ and (ii) in Lemma 1 with $j = t_k^0 + 1$, and using the triangle inequality and the fact that $|\hat{s}_l| = 1$, we have

$$|\mu_k^0 - \hat{\mu}_{l+1}| \leq 2n\lambda\,(t_k^0 - \hat{t}_l)^{-1} + \Big|(t_k^0 - \hat{t}_l)^{-1} \sum_{t=\hat{t}_l+1}^{t_k^0} \varepsilon_i\Big| \tag{4.38}$$

Also, applying (i) in Lemma 1 with $l = l+1$ and (ii) in Lemma 1 with $j = t_k^0 + 1$, and using the triangle inequality and the fact that $|\hat{s}_l| = 1$, we obtain

$$|\hat{\mu}_{l+1} - \mu_{k+1}^0| \leq 2n\lambda\,(\hat{t}_{l+1} - t_k^0)^{-1} + \Big|(\hat{t}_{l+1} - t_k^0)^{-1} \sum_{t=t_k^0+1}^{\hat{t}_{l+1}} \varepsilon_i\Big|. \tag{4.39}$$

Note that $|\mu_{k+1}^0 - \mu_k^0| = |\mu_{k+1}^0 - \hat{\mu}_{l+1} + \hat{\mu}_{l+1} - \mu_k^0| \leq |\hat{\mu}_{l+1} - \mu_k^0| + |\hat{\mu}_{l+1} - \mu_{k+1}^0|$ by the triangle inequality. Let us now define the event $W_{n,k} \equiv \{|\mu_{k+1}^0 - \mu_k^0| \leq |\hat{\mu}_{l+1} - \mu_k^0| + |\hat{\mu}_{l+1} - \mu_{k+1}^0|\}$.
Then by combining (4.38) and (4.39),

$$\begin{aligned}
W_{n,k} &= \Big\{|\mu_{k+1}^0 - \mu_k^0| \leq 2n\lambda\,(t_k^0 - \hat{t}_l)^{-1} + \Big|(t_k^0 - \hat{t})^{-1} \sum_{t=\hat{t}_l+1}^{t_k^0} \varepsilon_i\Big| \\
&\qquad + 2n\lambda\,(\hat{t}_{l+1} - t_k^0)^{-1} + \Big|(\hat{t}_{l+1} - t_k^0)^{-1} \sum_{t=t_k^0+1}^{\hat{t}_{l+1}} \varepsilon_i\Big|\Big\} \\
&= \Big\{2n\lambda\,(t_k^0 - \hat{t}_l)^{-1} + 2n\lambda\,(\hat{t}_{l+1} - t_k^0)^{-1} \\
&\qquad \geq |\mu_{k+1}^0 - \mu_k^0| - \Big|(t_k^0 - \hat{t}_l)^{-1} \sum_{t=\hat{t}_l+1}^{t_k^0} \varepsilon_i\Big| - \Big|(\hat{t}_{l+1} - t_k^0)^{-1} \sum_{t=t_k^0+1}^{\hat{t}_{l+1}} \varepsilon_i\Big|\Big\} \\
&\equiv \Big\{2n\lambda\,(t_k^0 - \hat{t}_l)^{-1} + 2n\lambda\,(\hat{t}_{l+1} - t_k^0)^{-1} \geq W_{n,k1} - W_{n,k2} - W_{n,k3}\Big\}.
\end{aligned} \tag{4.40}$$

Thus, by $P(W_{n,k}) = 1$ and Technique 1, we have

$$\begin{aligned}
P(R_3^{(1)}) &= P(R_3^{(1)} \cap W_{n,k}) \\
&\leq P\Big(R_3^{(1)} \cap \{2n\lambda\,(t_k^0 - \hat{t}_l)^{-1} + 2n\lambda\,(\hat{t}_{l+1} - t_k^0)^{-1} \geq \frac{1}{3} W_{n,k1}\}\Big)
\end{aligned}$$

$$P\Big(R_3^{(1)} \cap \{W_{n,k2} > \frac{1}{3} W_{n,k1}\}\Big)$$

$$P\Big(R_3^{(1)} \cap \{W_{n,k3} > \frac{1}{3} W_{n,k1}\}\Big). \tag{4.41}$$

Then, as n goes to infinity, the first term on the right-hand side of (4.41) tends to zero by Assumption A3. The second and third terms also tend to zero by Lemma 2 with $x_n = J_{min}^0/3$ and $\nu_n = n\delta_n$. Therefore, we have proved that $P(R_3^{(1)}) \to 0$. In a similar way, $P(R_3^{(2)}) \to 0$, $P(R_3^{(3)}) \to 0$, and $P(R_3^{(4)}) \to 0$, which implies that $P(R_{K,k3}) \to 0$ as $n \to \infty$. Consequently, we complete the proof. □

Theorem 4. Under Assumptions A1-A5, if $K \leq K^0$, for some $c > 0$,

$$P\Big(\inf_{0\leq K\leq K^0} \inf_{\mathbf{t}_K} \frac{n}{I_{min}^0 (J_{min}^0)^2} [\hat{\sigma}^2_{\hat{\mathbf{t}}_K} - \hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}] \geq c + o_p(1)\Big) \to 1 \text{ as } n \to \infty.$$

Proof. The following proof is adapted from Theorem 3.3 in Qian and Su (2016). To prove Theorem 4, it suffices to show (i) $\frac{n}{I_{min}^0 (J_{min}^0)^2} [\hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}} - \bar{\sigma}^2_n] = o_p(1)$, and (ii)$P\Big(\inf_{0\leq K\leq K^0} \inf_{\mathbf{t}_K} \frac{n}{I_{min}^0 (J_{min}^0)^2} [\hat{\sigma}^2_{\hat{\mathbf{t}}_K} - \bar{\sigma}^2_n] \geq c + o_p(1)\Big) \to 1$ as $n \to \infty$,

where

$$\bar{\sigma}^2_n \equiv \frac{1}{n} \sum_{k=1}^{K^0+1} \sum_{i=t_{k-1}^0+1}^{t_k^0} (y_i - \mu_k^0)^2 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2,$$

$$\hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}} \equiv \frac{1}{n} \sum_{k=1}^{K^0+1} \sum_{i=\hat{t}_{k-1}+1}^{\hat{t}_k} (y_i - \hat{\mu}_k(\hat{\mathbf{t}}_{K^0}))^2,$$

$$\hat{\sigma}^2_{\hat{\mathbf{t}}_K} \equiv \frac{1}{n} \sum_{k=1}^{K+1} \sum_{i=\hat{t}_{k-1}+1}^{\hat{t}_k} (y_i - \hat{\mu}_k(\hat{\mathbf{t}}_K))^2.$$

Let us first show (i). We proceed by making the following decomposition:

$$\hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}} - \bar{\sigma}^2_n = \frac{1}{n} \sum_{k=1}^{K^0+1} \sum_{i=\hat{t}_{k-1}+1}^{\hat{t}_k} (y_i - \hat{\mu}_k(\hat{\mathbf{t}}_{K^0}))^2 - \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2$$

$$
\begin{aligned}
&= \sum_{k=1}^{K^0+1} \frac{1}{n} \sum_{i=\hat{t}_{k-1}+1}^{\hat{t}_k} \left\{ (y_i - \hat{\mu}_k)^2 - \varepsilon_i^2 \right\} \\
&\equiv \sum_{k=1}^{K^0+1} Q_{n,1k}. \qquad (4.42)
\end{aligned}
$$

We now bound $Q_{n,1k}$ in the following four subcases: i1) $\hat{t}_{k-1} < t_{k-1}^0$ and $\hat{t}_k < t_k^0$; i2) $\hat{t}_{k-1} < t_{k-1}^0$ and $\hat{t}_k \geq t_k^0$; i3) $\hat{t}_{k-1} \geq t_{k-1}^0$ and $\hat{t}_k < t_k^0$; and i4) $\hat{t}_{k-1} \geq t_{k-1}^0$ and $\hat{t}_k \geq t_k^0$.

In (i1),

$$
\begin{aligned}
Q_{n,1k} &= \frac{1}{n} \sum_{i=t_{k-1}^0+1}^{t_k^0} \left((y_i - \hat{\mu}_k)^2 - \varepsilon_i^2 \right) + \frac{1}{n} \sum_{i=\hat{t}_{k-1}+1}^{t_{k-1}^0} \left((y_i - \hat{\mu}_k)^2 - \varepsilon_i^2 \right) \\
&\quad - \frac{1}{n} \sum_{i=\hat{t}_k+1}^{t_k^0} \left((y_i - \hat{\mu}_k)^2 - \varepsilon_i^2 \right) \\
&\equiv Q_{n,1k}(1) + Q_{n,1k}(2) - Q_{n,1k}(3).
\end{aligned}
$$

Let us first bound $Q_{n,1k(1)}$.

$$
\begin{aligned}
Q_{n,1k(1)} &= \frac{1}{n} \sum_{i=t_{k-1}^0+1}^{t_k^0} \left((y_i - \hat{\mu}_k)^2 - \varepsilon_i^2 \right) \\
&= \frac{1}{n} \sum_{i=t_{k-1}^0+1}^{t_k^0} \left((\mu_k^0 + \varepsilon_i - \hat{\mu}_k)^2 - \varepsilon_i^2 \right) \\
&= \frac{1}{n} \sum_{i=t_{k-1}^0+1}^{t_k^0} (\mu_k^0 + \varepsilon_i - \hat{\mu}_k + \varepsilon_i)(\mu_k^0 + \varepsilon_i - \hat{\mu}_k - \varepsilon_i) \\
&\leq \frac{1}{n} \sum_{i=t_{k-1}^0+1}^{t_k^0} (2\varepsilon_i + |\mu_k^0 - \hat{\mu}_k|)(|\mu_k^0 - \hat{\mu}_k|) \\
&= 2|\mu_k^0 - \hat{\mu}_k| \frac{1}{n} \sum_{i=t_{k-1}^0+1}^{t_k^0} \varepsilon_i + \frac{1}{n} \sum_{i=t_{k-1}^0+1}^{t_k^0} |\mu_k^0 - \hat{\mu}_k|^2 \\
&= 2|\mu_k^0 - \hat{\mu}_k| \frac{1}{n} O_p(n^{1/2}) + \frac{1}{n}(t_k^0 - t_{k-1}^0)|\mu_k^0 - \hat{\mu}_k|^2 \qquad (4.43) \\
&= |\mu_k^0 - \hat{\mu}_k| O_p(n^{-1/2}) + |\mu_k^0 - \hat{\mu}_k|^2 O_p(1), \quad \text{uniformly in} \quad k.
\end{aligned}
$$

For (4.43), by Chebyshev's inequality, $\sum_{i=t_{k-1}^0+1}^{t_k^0} \varepsilon_i = O_p\Big([(t_k^0 - t_{k-1}^0)\sigma^2]^{1/2}\Big)$. Then, since $t_k^0 - t_{k-1}^0 = O_p(n)$, $\sum_{i=t_{k-1}^0+1}^{t_k^0} \varepsilon_i = O_p(n^{1/2})$.

Similarly,

$$\begin{aligned}
Q_{n,1k(2)} &= \frac{1}{n}\sum_{i=\hat{t}_{k-1}+1}^{t_{k-1}^0}\Big((y_i - \hat{\mu}_k)^2 - \varepsilon_i^2\Big) \\
&\leq 2|\mu_k^0 - \hat{\mu}_k|\frac{1}{n}\sum_{i=\hat{t}_{k-1}+1}^{t_{k-1}^0}\varepsilon_i + \frac{1}{n}\sum_{i=\hat{t}_{k-1}+1}^{t_{k-1}^0}|\mu_k^0 - \hat{\mu}_k|^2 \\
&= |\mu_k^0 - \hat{\mu}_k|\frac{1}{n}O_p((n\delta_n)^{1/2}) + \frac{1}{n}O_p(n\delta_n)|\mu_k^0 - \hat{\mu}_k|^2, \quad \text{uniformly in} \quad k.
\end{aligned}$$

$$\begin{aligned}
Q_{n,1k(3)} &= \frac{1}{n}\sum_{i=\hat{t}_{k}+1}^{t_{k}^0}\Big((y_i - \hat{\mu}_k)^2 - \varepsilon_i^2\Big) \\
&\leq 2|\mu_k^0 - \hat{\mu}_k|\frac{1}{n}\sum_{i=\hat{t}_{k}+1}^{t_{k}^0}\varepsilon_i + \frac{1}{n}\sum_{i=\hat{t}_{k}+1}^{t_{k}^0}|\mu_k^0 - \hat{\mu}_k|^2 \\
&= |\mu_k^0 - \hat{\mu}_k|\frac{1}{n}O_p((n\delta_n)^{1/2}) + \frac{1}{n}O_p(n\delta_n)|\mu_k^0 - \hat{\mu}_k|^2, \quad \text{uniformly in} \quad k.
\end{aligned}$$

Consequently,

$$\begin{aligned}
Q_{n,1k} &\equiv Q_{n,1k}(1) + Q_{n,1k}(2) - Q_{n,1k}(3) \\
&= |\mu_k^0 - \hat{\mu}_k|O_p(n^{-1/2}) + |\mu_k^0 - \hat{\mu}_k|^2 O_p(1), \quad \text{uniformly in} \quad k.
\end{aligned}$$

Analogously, we can show that the result above also holds in (i2)-(i4). Furthermore, from Theorem 2, we know that $|\mu_k^0 - \hat{\mu}_k| = O_p(n\lambda/I_k^0 + n\delta_n/I_k^0 + (I_k^0)^{-1/2})$ for each $k = 1,...,K^0+1$. Also, by Assumption A3, A5, and $I_k^0 \geq I_{min}^0$, we obtain that $n\lambda/I_k^0 \leq n\lambda/I_{min}^0 \to 0$ and $n\delta/I_k^0 \leq n\delta_n/I_{min}^0 \to 0$ as $n \to \infty$. Therefore, $|\mu_k^0 - \hat{\mu}_k| = O_p((I_{min}^0)^{-1/2})$ as $n \to \infty$.

Hence,

$$Q_{n,1k} = O_p((I^0_{min})^{-1/2}))O_p(n^{-1/2}) + O_p((I^0_{min})^{-1/2})O_p((I^0_{min})^{-1/2})O_p(1)$$
$$= O_p((I^0_{min})^{-1/2}n^{-1/2} + (I^0_{min})^{-1}). \tag{4.44}$$

Using (4.44) and multiplying both sides of (4.42) by $\frac{n}{I^0_{min}(J^0_{min})^2}$ yields that

$$\frac{n}{I^0_{min}(J^0_{min})^2}[\hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}} - \bar{\sigma}^2_n] = \frac{n}{I^0_{min}(J^0_{min})^2}\sum_{k=1}^{K^0+1} Q_{n,1k}$$
$$= \frac{n(K^0+1)}{I^0_{min}(J^0_{min})^2} O_p((I^0_{min})^{-1/2}n^{-1/2} + (I^0_{min})^{-1})$$
$$= O_p\Big(\frac{\log n}{n}\frac{1}{(J^0_{min})^2}\Big) \tag{4.45}$$
$$= o_p(1) \tag{4.46}$$

We get (4.45) because $K^0 = O_p(\log n)$ and $I^0_{min} = \Theta(n)$ by Assumptions A3 and A4. For (4.46), as $n \to \infty$, $\log n/n \to 0$ by L'Hospital's rule, and J^0_{min} is a constant by Assumption A4. Therefore, we complete the proof of (i).

Next, let us prove (ii). We assume that $K^0 = 1$ and $\hat{t}_1 < t^0_1$ from now as the other cases follow analogously. When $K^0 = 1$, $K = 0$ and $\mathbf{t}_0 = \{\emptyset\}$. Thus, $\hat{\mu}_{\mathbf{t}_0}$ becomes the OLS estimate of y_i on index $1,\ldots,n$ using all n observations. That is, $\hat{\mu}_{\mathbf{t}_0} = \hat{\mu}_{OLS} \equiv \frac{1}{n}\sum_{i=1}^n y_i$. Then,

$$\hat{\sigma}^2_{\mathbf{t}_0} - \bar{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^n \Big((y_i - \hat{\mu}_{OLS})^2 - \varepsilon_i^2\Big)$$
$$= \frac{1}{n}\Big\{\Big(\sum_{i=1}^{t^0_1}(y_i - \hat{\mu}_{OLS})^2 - \varepsilon_i^2\Big) + \Big(\sum_{i=t^0_1+1}^{n}(y_i - \hat{\mu}_{OLS})^2 - \varepsilon_i^2\Big)\Big\}. \tag{4.47}$$

Also, using (3.1) with $K^0 = 1$ and the fact that $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i = O_p(n^{-1/2})$, we obtain

$$\begin{aligned}\hat{\mu}_{OLS} &= \frac{1}{n}\sum_{i=1}^{t_1^0}\mu_1^0 + \frac{1}{n}\sum_{i=t_1^0+1}^{n}\mu_2^0 + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \\ &= \frac{1}{n}\sum_{i=1}^{t_1^0}\mu_1^0 + \frac{1}{n}\sum_{i=t_1^0+1}^{n}\mu_2^0 + O_p(n^{-1/2}). \qquad (4.48)\end{aligned}$$

Then, through tedious calculations, using (4.48) and the fact that $\sum_{i=1}^{t_1^0}\varepsilon_i = Op((I_1^0)^{1/2})$ and $\sum_{i=t_1^0+1}^{n}\varepsilon_i = Op((I_2^0)^{1/2})$, (4.47) follows that

$$\begin{aligned}\hat{\sigma}_{\hat{\mathbf{t}}_0}^2 - \bar{\sigma}_n^2 &= \frac{I_1^0}{n}\Big(\frac{I_2^0}{n}J_{min}^0\Big)^2 + \frac{I_2^0}{n}\Big(\frac{I_1^0}{n}J_{min}^0\Big)^2 + O_p(n^{-1/2}) \\ &\geq c\frac{I_{min}^0(J_{min}^0)^2}{n} + O_p(n^{-1/2}) \qquad \text{for some} \quad c > 0. \qquad (4.49)\end{aligned}$$

(4.49) follows from $I_{min}^0 = \Theta(n)$ in Assumption (A3).

Then, multiplying both sides of (4.49) by $\frac{n}{I_{min}^0(J_{min}^0)^2}$ yields that

$$\begin{aligned}\frac{n}{I_{min}^0(J_{min}^0)^2}[\hat{\sigma}_{\hat{\mathbf{t}}_K}^2 - \bar{\sigma}_n^2] &\geq \frac{n}{I_{min}^0(J_{min}^0)^2}\Big\{c\frac{I_{min}^0(J_{min}^0)^2}{n} + O_p(n^{-1/2})\Big\} \\ &= c + \frac{n}{I_{min}^0(J_{min}^0)^2}O_p(n^{-1/2}) \\ &= c + o_p(1) \qquad \text{for some} \quad c > 0.\end{aligned}$$

The last equality is obtained by Assumption A3. Therefore, this completes the proof of (ii) for the case $K^0 = 1$. Also, using similar arguments the same conclusion can be drawn for the case when $K^0 \geq 2$. □

Theorem 5. Let $\hat{\lambda}_{\hat{K}} = \arg\min_{\lambda_K} JMIC(\lambda_K)$. Then, under Assumptions A1-A6,

$$P\Big(\inf_{\lambda_K\in\Omega_-\cup\Omega_+} JMIC(\lambda_K) > JMIC(\lambda^0)\Big) \to 1 \text{ as } n\to\infty.$$

Proof. $JMIC(\lambda_K)$ in (3.15) is given by

$$JMIC(\lambda_K) = -2ln(\hat{\theta}, \hat{\mathbf{t}}_K, K) + d(K+1)^{\gamma} n\rho_n.$$

In terms of residual sum of squares,

$$JMIC(\lambda_K) \propto ln(\hat{\sigma}^2_{\hat{\mathbf{t}}_K}) + d(K+1)^{\gamma}\rho_n, \tag{4.50}$$

under a normal distribution with $K+1$ differient means and a common variance σ^2.

Let $\Omega_- = \{\lambda_K \in \Omega : \hat{K} < K^0\}$, $\Omega_o = \{\lambda_K \in \Omega : \hat{K} = K^0\}$, and $\Omega_+ = \{\lambda_K \in \Omega : \hat{K} > K^0\}$. Ω_-, Ω_o, and Ω_+ denote the subsets of all λ values, producing the under-, correct-, and over-number of change points, respectively. Then, it suffices to show that

$$\text{(i)} \quad P\Big(\inf_{\lambda_K \in \Omega_-} JMIC(\lambda_K) > JMIC(\lambda^0)\Big) \to 1, \quad \text{and}$$
$$\text{(ii)} \quad P\Big(\inf_{\lambda_K \in \Omega_+} JMIC(\lambda_K) > JMIC(\lambda^0)\Big) \to 1,$$

as $n \to \infty$.

(i) $\lambda_K \in \Omega_-$ (i.e., $\hat{K} < K^0$)

$$\begin{aligned} & P\Big(\inf_{\lambda_K \in \Omega_-} JMIC(\lambda_K) > JMIC(\lambda^0)\Big) \\ & = P\Big((\ln(\hat{\sigma}^2_{\hat{\mathbf{t}}_{\hat{K}}}) + d(\hat{K}+1)^{\gamma}\rho_n) - (\ln(\hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}) + d(K^0+1)^{\gamma}\rho_n) > 0\Big) \\ & = P\Big(\ln(\hat{\sigma}^2_{\hat{\mathbf{t}}_{\hat{K}}} / \hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}) + \rho_n d((\hat{K}+1)^{\gamma} - (K^0+1)^{\gamma}) > 0\Big) \\ & \to 1, \end{aligned} \tag{4.51}$$

as $n \to \infty$. (4.51) follows from $\ln(\hat{\sigma}^2_{\hat{\mathbf{t}}_{\hat{K}}} / \hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}) > 0$ as $n \to \infty$ by Theorem 4, and from $\rho_n \to 0$ by Assumption A6.

(ii) $\lambda_K \in \Omega_+$ (i.e., $K^0 < \hat{K}$)

$$
\begin{aligned}
& P\Big(\inf_{\lambda_K \in \Omega_+} JMIC(\lambda_K) > JMIC(\lambda^0)\Big) \\
& = P\Big((\delta_n)^{-1}\Big\{(\ln(\hat{\sigma}^2_{\hat{\mathbf{t}}_{\hat{K}}}) + d(\hat{K}+1)^\gamma \rho_n) - (\ln(\hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}) + d(K^0+1)^\gamma \rho_n))\Big\} > 0\Big) \\
& = P\Big((\delta_n)^{-1} \ln(\hat{\sigma}^2_{\hat{\mathbf{t}}_{\hat{K}}} / \hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}) + (\delta_n)^{-1} \rho_n d((\hat{K}+1)^\gamma - (K^0+1)^\gamma) > 0\Big) \\
& = P\Big((\hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}})^{-1} (\delta_n)^{-1} (\hat{\sigma}^2_{\hat{\mathbf{t}}_{\hat{K}}} - \hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}) + (\delta_n)^{-1} \rho_n d((\hat{K}+1)^\gamma - (K^0+1)^\gamma) + o_p(1) > 0\Big) \qquad (4.52) \\
& = P\Big((\sigma_0^2)^{-1} (\delta_n)^{-1} (\hat{\sigma}^2_{\hat{\mathbf{t}}_{\hat{K}}} - \hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}) + (\delta_n)^{-1} \rho_n d((\hat{K}+1)^\gamma - (K^0+1)^\gamma) + o_p(1) > 0\Big) \qquad (4.53) \\
& \to 1, \qquad (4.54)
\end{aligned}
$$

as $n \to \infty$. (4.52) is obtained by the fact that $\ln(1+x) = x + O(x^2)$ for $|x| < 1$, where $x = (\hat{\sigma}^2_{\hat{\mathbf{t}}_{\hat{K}}} - \hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}) / \hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}$. For (4.53), observe that $\hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}} = \sigma_0^2 + o_p(1)$. (4.54) follows from $(\delta_n)^{-1}(\hat{\sigma}^2_{\hat{\mathbf{t}}_{\hat{K}}} - \hat{\sigma}^2_{\hat{\mathbf{t}}_{K^0}}) = O_p(1)$ by Lemma E.1 in Qian and Su (2016), and from $(\delta_n)^{-1}\rho_n \to \infty$ by Assumption A6. Therefore, we complete the proof.

□

5 Simulation studies

5.1 Simulation setup

We evaluate the following tuning parameter selection criteria: JMIC ($\alpha = 1/2$), JMIC ($\alpha = 1/3$), PMIC (C=1), PMIC (C=10), SIC, and ZMIC. For γ in JMIC, γ was set at $5/4$.

We generated a sequence of 300 observations ($i = 1, ..., 300$) from the multiple change points model (3.1). The 3, 4, or 8 change points were assumed in a sequence. To set the parameter values $\theta = ((\mu_1, \sigma_1), ..., (\mu_{K+1}, \sigma_{K+1}))$ in $K+1$ segments, the means and variances of the $\log_2$ ratio of reads data were estimated from the breast tumor cell line HCC1954 and its matched normal cell line BL1954 on chromosome 8 provided by Chiang et al. (2009). We observed that the mean values range from -1.04 to 3 or more in a segment and that the standard deviations range from 0.009 to 0.5814. Thus, μ was set between -1 and 2, and σ was set between 0.1 and 0.4. Then, we considered various situations, including when mean differences between neighboring segments are relatively small and large, and when variances are small, moderate, and large. For the change point locations, the following scenarios were considered: when change points are uniformly spread out (for example, for the 3 change point model, $(t_1, t_2, t_3) = (75, 150, 225)$); when some change points are located at the front and the others are at the center ($(t_1, t_2, t_3) = (50, 150, 185)$); and when some change points are located at the center and the others are at the end of the sequence ($(t_1, t_2, t_3) = (150, 240, 270)$).

Therefore, for each of the possible combinations for θ=((μ_1,σ_1),...,(μ_{K+1},σ_{K+1})), $\mathbf{t} = (t_1, ..., t_K)$ above, we simulated 1000 datasets of a sequence of 300 observations for the 3, 4, and 8 change point models. To each simulated dataset, Algorithm 2 in Section 3.4 was applied. In step 1, the dual path algorithm was performed using the R package *genlasso*. Then, step 2 was completed for each criterion to obtain its estimated number of change points $\hat{K}$.

5.2 Simulation results

To compare the performance of each criterion, we ran 1000 simulations in each scenario to calculate (1) the number of times the 1d fused lasso with each criterion correctly estimates the true change point locations. Furthermore, we analyzed (2) frequencies of the estimated number of change points to investigate model complexity. Note that the true change point was assumed to stay in the interval $t_i \pm 1$, $i = 1, \ldots, \hat{K}$. The results of (1) for the 3, 4, and 8 change point models are reported in Tables 7.1, 7.3, and 7.5, respectively. Tables 7.2, 7.4, and 7.6 show the results of (2) under the scenario with large means and small variances for the 3, 4, and 8 change point models, respectively.

5.2.1 The 3, 5, and 8 change point models

Tables 7.1, 7.3, and 7.5 show that, in all of the cases, the 1d fused lasso with JMIC correctly identified the true change point locations more times than it did with the other criteria. PMIC was competitive, but JMIC outperformed PMIC even when the true change points were evenly located. The 1d fused lasso with JMICs ($\alpha = 1/2$ and $1/3$) had similar performance. However, in the 8 change point model, the 1d fused lasso with JMIC ($\alpha = 1/3$) tended to detect the true change points more times than with JMIC ($\alpha = 1/2$). The 1d fused lasso with PMIC ($C = 10$) slightly performed better than with PMIC ($C = 1$). The 1d fused lasso with SIC or ZMIC had very low relative frequencies. It turned out that the 1d fused lasso with SIC and ZMIC detected too many false change points, which will be shown in Table 7.2,7.4, and 7.6. Furthermore, when the mean differences were small, variances were large, and change points were not evenly located, the 1d fused lasso failed to detect the true locations many times regardless of the criteria used.

Tables 7.2, 7.4, and 7.6 show that, in all of the cases, the estimated number of change points was greater than or equal to the true number of change points. This means that all of the criteria tend to overestimate the number of change points with different penalty amounts. For example, in the 3 change point model, when one change point was at the front and the others were at the center ((t_1,t_2,t_3) =(50,150,185)), JMIC ($\alpha = 1/2$) chose 4 changes 9 times, while PMIC ($C = 1$) selected 4 changes 52 times and 5 changes 9 times. This indicates that PMIC has a larger penalty than JMIC. SIC and ZMIC selected too many false change points, suggesting that their penalty terms do not penalize enough the model complexity.

6 Application to NGS data

6.1 Breast tumor data

We applied our algorithm to detect CNVs in the reads ratio data of the breast tumor cell line HCC1954 and its matched normal cell line BL1954 provided by Chiang et al. (2009). The reads data are freely available from the National Center for Biotechnology Information (NCBI) short read archive. The data used were normalized by using the R package QDNAseq (Scheinin et al., 2014) and binned into 100Kbp (100,000 base pairs). Due to space limitations, we have only presented the results for chromosomes 1, 8, 19, and X.

6.2 Application to CNV detection using NGS reads ratio data

Since SIC and ZMIC perform very poorly, we have presented the results only for JMIC and PMIC in this section. Figures 7.1, 7.2, 7.3, and 7.4 show the number of copy number changes and the locations estimated by the 1d fused lasso with JMIC and PMIC on chromosomes 1, 8, 19, X, respectively. In each segment, the log2 read ratio values were estimated as their mean, which is shown by the solid lines. Overall, the 1d fused lasso with JMIC splits the chromosomes into a smaller number of segments than with PMIC. The 1d fused lasso with PMIC tends to detect many unnecessary points. PMICs with C=1 and 10 select the same number of change points. The 1d fused lasso with JMIC($\alpha = 1/3$) selects more change points than with JMIC($\alpha = 1/2$).

For chromosome 1 in Figure 7.1, we can clearly see that the number of copy number changes estimated by JMIC is less than by PMIC. The 1d fused lasso with PMIC seems to detect many redundant points. The 1d fused lasso with JMIC ($\alpha = 1/3$) captures the narrow region of a possible copy number gain in the range of 120,000-130,000 Kbp, while it does not with JMIC ($\alpha = 1/2$). However, the 1d fused lasso with JMIC ($\alpha = 1/3$) tends to identify some redundant points in the 45,000 to 120,000 Kbp range. For chromosome 8 in Figure 7.2, in the 80,000-120,000 Kbp range where the data are largely spread out, the 1d fused lasso with JMIC ($\alpha = 1/2$) does not detect any change points, while it tends to detect many redundant change points with JMIC ($\alpha = 1/3$) and PMIC ($C = 1, 10$). For chromosome 19 in Figure 7.3, the 1d fused lasso with JMIC ($\alpha = 1/3$) identifies a possible copy number loss in the range of 0-8,000 Kbp, while it does not with JMIC ($\alpha = 1/2$). For chromosome X in Figure 7.4, the 1d fused lasso with JMIC ($\alpha = 1/3$) detects the narrow region of a possible copy gain in the 0-5,000 Kbp range, while it does not with JMIC ($\alpha = 1/3$). Also, the 1d fused lasso with JMIC ($\alpha = 1/3$) captures the region of a possible copy loss in the 60,000-80,000 Kbp range, while it does not detect any copy number changes with JMIC ($\alpha = 1/3$) in the region.

7 Discussion and future work

7.1 Discussion

We have proposed a new modified Bayesian information criterion, called JMIC, to estimate the optimal tuning parameter in the 1d fused lasso. The development of JMIC was motivated by the need to address multiple change point problems such as copy number variation detection. The 1d fused lasso encourages adjacent coefficients to be similar and even forces some of the similar adjacent coefficients to be exactly the same. As a result, the 1d fused lasso not only incorporates the idea of the structural dependency of copy numbers but also captures exact locations of copy number changes. Given the number of copy number changes, the corresponding locations are estimated by fitting the 1d fused lasso using the dual path algorithm. The number of copy number changes is estimated by selecting the tuning parameter using JMIC.

Here are some of the advantages of JMIC: (1) since the optimal number of change points is estimated by selecting λ that gives the smallest JMIC, the 1d fused lasso with JMIC can automatically estimate copy number changes. With other multiple change point search algorithms, such as binary segmentation, one can also use JMIC to automatically identify change points; (2) the 1d fused lasso with JMIC outperforms the other criteria considered in our numerical studies. Furthermore, when assuming the conditions in Chapter 4, through theoretical studies we have proved that JMIC consistently selects the true tuning parameter in the 1d fused lasso. That is, the number of change points estimated by JMIC converges to the true number of change points; and (3) we have established the reasonable ranges of JMIC's two constants, γ and α. Depending on the purpose, JMIC can be flexibly used without exhaustive experiments because the suggested ranges are narrow.

The simulation studies confirm the superiority of JMIC. The 1d fused lasso with JMIC detected the true change points more times than with the other criteria. However, although the 1d fused lasso with JMIC performed better than it did with the other criteria, it did not perform well with any of the criteria considered in the following circumstances: (1) when the true means in neighboring segments were similar; (2) variances were large; (3) change points were unevenly located. Especially, the 1d fused lasso performed poorly when the true change points were not uniformly spread out.

The results of the real data analysis suggest that the 1d fused lasso with JMIC gives a more reasonable number of changes than with the other criteria. However, although the 1d fused lasso with JMIC successfully detects the wide or narrow regions with possible CNVs, we did not further investigate whether the areas contain the genes that relate to breast cancer. In this dissertation, we focus on a segmentation method that estimates the number of copy number changes and their locations.

7.2 Future work

(1) In our simulation studies when the true mean consecutively jumped from one value to another in an increasing or decreasing direction, we found that the 1d fused lasso detected many false change points. The false change points occurred around the true change point locations, even when the mean differences were large. This phenomenon was identified in 2014 by Rojas and Wahlberg and was named the staircase effect. Rojas and Wahlberg (2014) showed that Karush-Kuhn-Tucker (KKT) optimality conditions are violated in the staircase situation. They also noted that the staircase effect problem is not related to the level of noise or the tuning parameter used. The staircase effect issue has been resolved in the context of its appearance in the image de-noising process, but it has been observed very recently in the change point analysis. As such there are only a few papers written about it (Rojas and Wahlberg, 2014, 2015), and thus there is clearly room for improvement of the 1d fused lasso in multiple change point problems.

(2) We assumed that the random errors are independent in our current work. However, since NGS reads data are sequenced according to their genomic locations, the error terms of the observations that are close to each other are likely to be correlated in the sequence. Therefore, since the assumption of independent error terms may be violated in NGS reads data, we will develop a new copy number change point detection model that incorporates the assumption of data being correlated.

(3) In addition to DNA copy number variations, epigenetic changes such as DNA methylation influence gene expression (Bell et al., 2011; Razin and Cedar, 1991; Wagner et al., 2014). As a result, different methylation patterns are associated with many human diseases including cancers (Hansen et al., 2011; Robertson, 2005). Over the past decade, with the development of data processing technology, many researchers have begun integrating and analyzing data from multiple biological systems (copy number and gene expression (Hassan et al., 2014; Lai et al., 2017; Menezes et al., 2009); methylation and gene expression (Li et al., 2016); copy number and methylation (Castellani et al., 2015); copy number, methylation, and gene expression data integration (Louhimo et al., 2012; Sun et al., 2011). Since many complex diseases are caused by interactions among biological variations at various developmental stages, integrated data analyses will provide a more profound insight into disease biology and help us identify or predict genomic risks for diseases.

Therefore, for future research, we are considering an integrated data analysis that identifies the relationship among DNA copy number variations, different methylation levels, and gene expression patterns and how the combination may affect the development of certain diseases. As a result, we may be able to provide biological evidence that the CNV regions detected by our proposed method contain the genes that relate to specific diseases. Also, we may be able to identify other genes that are influenced by epigenetic modifications.

$(\mu_1,\mu_2,\mu_3,\mu_4)$	$(\sigma_1,\sigma_2,\sigma_3,\sigma_4)$	(t_1,t_2,t_3)	JMIC $\alpha=1/2$	JMIC $\alpha=1/3$	PMIC $C=1$	PMIC $C=10$	SIC	ZMIC
(1,2,-1,0)	(0.1,0.1,0.1,0.1)	75,150,225	**1.000**	**1.000**	0.965	0.977	0.264	0.117
		50,150,185	**0.976**	0.975	0.936	0.958	0.233	0.076
		150,240,270	**0.823**	0.819	0.802	0.805	0.128	0.041
	(0.1,0.3,0.1,0.3)	75,150,225	**0.955**	0.953	0.922	0.939	0.164	0.000
		50,150,185	**0.577**	**0.577**	0.563	0.570	0.057	0.008
		150,240,270	**0.591**	0.580	0.550	0.556	0.052	0.011
	(0.2,0.4,0.2,0.4)	75,150,225	**0.870**	0.868	0.826	0.843	0.049	0.003
		50,150,185	**0.383**	**0.383**	0.371	0.374	0.005	0.000
		150,240,270	**0.293**	**0.293**	0.288	0.291	0.004	0.000
(0,1,0,1)	(0.1,0.1,0.1,0.1)	75,150,225	**0.999**	0.998	0.962	0.979	0.247	0.063
		50,150,185	**0.967**	0.963	0.919	0.929	0.170	0.031
		150,240,270	**0.825**	0.818	0.788	0.793	0.126	0.037
	(0.1,0.3,0.1,0.3)	75,150,225	**0.890**	0.888	0.855	0.872	0.111	0.013
		50,150,185	**0.451**	0.447	0.429	0.438	0.025	0.001
		150,240,270	**0.564**	0.549	0.516	0.518	0.052	0.004
	(0.2,0.4,0.2,0.4)	75,150,225	**0.724**	0.723	0.691	0.705	0.014	0.000
		50,150,185	**0.295**	**0.295**	0.290	0.293	0.002	0.000
		150,240,270	0.277	**0.288**	0.286	0.285	0.004	0.000

TABLE 7.1: The relative frequency (out of 1000 simulations) that the estimates $\hat{\mathbf{t}} = (\hat{t}_1,...,\hat{t}_{\hat{K}=3})$ of the change point locations equal the true change point locations. The true number of change points K is 3.

(μ_1,μ_2,μ_3)	$(\sigma_1,\sigma_2,\sigma_3)$	(t_1,t_2,t_3)		2	3	4	5+	Total
(1,2,-1,0)	(0.1,0.1,0.1,0.1)	75,150,225	JMIC($\alpha = 1/2$)	0	1000	0	0	1000
			JMIC($\alpha = 1/3$)	0	1000	0	0	
			PMIC(C=1)	0	965	29	6	
			PMIC(C=10)	0	977	20	3	
			SIC	0	264	90	646	
			ZMIC	0	117	33	850	
		50,150,185	JMIC($\alpha = 1/2$)	0	991	9	0	
			JMIC($\alpha = 1/3$)	0	979	20	1	
			PMIC(C=1)	0	939	52	9	
			PMIC(C=10)	0	962	37	1	
			SIC	0	234	58	708	
			ZMIC	0	76	21	903	
		150,240,270	JMIC($\alpha = 1/2$)	1	823	161	15	
			JMIC($\alpha = 1/3$)	0	819	164	17	
			PMIC(C=1)	0	802	174	24	
			PMIC(C=10)	0	805	171	24	
			SIC	0	128	113	759	
			ZMIC	0	41	31	928	

TABLE 7.2: The frequencies of the number of change points $\hat{K}$ estimated by the 1d fused lasso with each criterion over 1000 simulations. The true number of change points K is 3.

$(\mu_1,...,\mu_6)$	$(\sigma_1,...,\sigma_6)$	$(t_1,...,t_5)$	JMIC $\alpha=1/2$	JMIC $\alpha=1/3$	PMIC $C=1$	PMIC $C=10$	SIC	ZMIC
(1,-1,0,-1,2,1)	(0.1,0.1,...,0.1)	50,100,150,200,250	**1.000**	0.999	0.952	0.961	0.255	0.067
		40,70,140,175,230	**0.991**	0.986	0.941	0.953	0.238	0.054
		90,135,170,240,270	**0.958**	0.956	0.926	0.937	0.176	0.040
	(0.1,0.3,...,0.1,0.3)	50,100,150,200,250	**0.858**	0.856	0.801	0.810	0.130	0.019
		40,70,140,175,230	**0.693**	0.690	0.650	0.659	0.090	0.011
		90,135,170,240,270	**0.706**	0.700	0.656	0.665	0.082	0.020
	(0.2,0.4,...,0.2,0.4)	50,100,150,200,250	**0.661**	0.654	0.619	0.622	0.025	0.000
		40,70,140,175,230	**0.543**	**0.543**	0.510	0.522	0.013	0.000
		90,135,170,240,270	0.403	**0.407**	0.388	0.395	0.009	0.000
(0,1,0,1,0,1)	(0.1,0.1,...,0.1)	50,100,150,200,250	**1.000**	0.999	0.948	0.961	0.193	0.016
		40,70,140,175,230	**0.990**	0.988	0.942	0.950	0.180	0.022
		90,135,170,240,270	**0.957**	**0.957**	0.917	0.923	0.183	0.019
	(0.1,0.3,...,0.1,0.3)	50,100,150,200,250	**0.844**	0.841	0.807	0.819	0.064	0.001
		40,70,140,175,230	**0.702**	**0.702**	0.669	0.677	0.054	0.000
		90,135,170,240,270	**0.759**	0.757	0.718	0.722	0.048	0.004
	(0.2,0.4,...,0.2,0.4)	50,100,150,200,250	**0.653**	**0.653**	0.630	0.638	0.000	0.000
		40,70,140,175,230	**0.530**	**0.530**	0.501	0.504	0.000	0.000
		90,135,170,240,270	0.305	**0.408**	0.398	0.399	0.004	0.000

TABLE 7.3: The relative frequency (out of 1000 simulations) that the estimates $\hat{t} = (\hat{t}_1,...,\hat{t}_{\hat{K}=5})$ of the change point locations equal the true change point locations. The true number of change points K is 5.

$(\mu_1,\ldots,\mu_6)$	$(\sigma_1,\ldots,\sigma_6)$	$(t_1,\ldots,t_5)$		4	5	6	7+	Total
(1,-1,0,-1,2,1)	(0.1,...,0.1)	750,100,150,200,250	JMIC($\alpha=1/2$)	0	1000	0	0	1000
			JMIC($\alpha=1/3$)	0	999	1	0	
			PMIC(C=1)	0	952	44	4	
			PMIC(C=10)	0	961	37	2	
			SIC	0	255	78	667	
			ZMIC	0	67	18	915	
		40,70,140,175,230	JMIC($\alpha=1/2$)	0	995	5	0	
			JMIC($\alpha=1/3$)	0	990	9	1	
			PMIC(C=1)	0	944	46	10	
			PMIC(C=10)	0	957	35	8	
			SIC	0	239	82	679	
			ZMIC	0	54	23	923	
		90,135,170,240,270	JMIC($\alpha=1/2$)	0	958	40	2	
			JMIC($\alpha=1/3$)	0	956	42	2	
			PMIC(C=1)	0	926	65	9	
			PMIC(C=10)	0	937	55	8	
			SIC	0	176	98	726	
			ZMIC	0	40	19	941	

TABLE 7.4: The frequencies of the number of change points $\hat{K}$ estimated by the 1d fused lasso with each criterion over 1000 simulations. The true number of change points K is 5.

$(\mu_1,\ldots,\mu_9)$	$(\sigma_1,\ldots,\sigma_9)$	$(t_1,\ldots,t_8)$	JMIC $\alpha=1/2$	JMIC $\alpha=1/3$	PMIC $C=1$	PMIC $C=10$	SIC	ZMIC
(1,-1,0,-1,2,1,2,0,1)	(0.1,0.1,...,0.1, 0.1)	33,66,99,132,165,198,231,264	**1.000**	0.999	0.946	0.952	0.400	0.135
		25,50,75,135,160,185,210,235	**0.926**	**0.926**	0.892	0.893	0.243	0.077
		70,95,120,145,175,225,250,275	**0.910**	**0.910**	0.866	0.871	0.226	0.076
	(0.1,0.3,...,0.1,0.3)	33,66,99,132,165,198,231,264	0.947	**0.949**	0.894	0.898	0.211	0.028
		25,50,75,135,160,185,210,235	0.555	**0.569**	0.542	0.546	0.111	0.012
		70,95,120,145,175,225,250,275	0.388	**0.660**	0.642	0.644	0.104	0.020
	(0.2,0.4,...,0.2,0.4)	33,66,99,132,165,198,231,264	0.315	**0.695**	0.652	0.659	0.036	0.002
		25,50,75,135,160,185,210,235	0.039	**0.256**	0.247	0.247	0.008	0.001
		70,95,120,145,175,225,250,275	0.002	**0.316**	0.309	0.311	0.019	0.001
0,1,0,1,0,1,0,1,0	(0.1,0.1,...,0.1, 0.1)	33,66,99,132,165,198,231,264	**0.999**	0.996	0.939	0.948	0.354	0.061
		25,50,75,135,160,185,210,235	**0.923**	0.922	0.878	0.882	0.213	0.039
		70,95,120,145,175,225,250,275	**0.921**	**0.921**	0.886	0.887	0.214	0.035
	(0.1,0.3,...,0.1,0.3)	33,66,99,132,165,198,231,264	**0.945**	**0.945**	0.900	0.907	0.143	0.005
		25,50,75,135,160,185,210,235	0.547	**0.584**	0.564	0.568	0.054	0.006
		70,95,120,145,175,225,250,275	0.640	**0.696**	0.662	0.664	0.074	0.003
	(0.2,0.4,...,0.2,0.4)	33,66,99,132,165,198,231,264	0.002	**0.664**	0.627	0.635	0.003	0.000
		25,50,75,135,160,185,210,235	0.000	**0.251**	0.239	0.239	0.000	0.000
		70,95,120,145,175,225,250,275	0.000	**0.297**	0.282	0.284	0.000	0.000

TABLE 7.5: The relative frequency (out of 1000 simulations) that the estimates $\hat{t} = (\hat{t}_1,\ldots,\hat{t}_{\hat{K}=8})$ of the change point locations equal the true change point locations. The true number of change points K is 8.

$(\mu_1,\ldots,\mu_9)$	$(\sigma_1,\ldots,\sigma_9)$	$(t_1,\ldots,t_8)$		6	7	8	9	10+	Total
(1,-1,0,-1,2,1,2,0,1)	(0.1,...,0.1)	33,66,99,132,165,198,231,264	JMIC($\alpha = 1/2$)	0	0	1000	0	0	1000
			JMIC($\alpha = 1/3$)	0	0	999	1	0	
			PMIC(C=1)	0	0	946	51	3	
			PMIC(C=10)	0	0	952	45	3	
			SIC	0	0	400	108	492	
			ZMIC	0	0	135	39	826	
		25,50,75,135,160,185,210,235	JMIC($\alpha = 1/2$)	0	0	938	60	2	
			JMIC($\alpha = 1/3$)	0	0	938	60	2	
			PMIC(C=1)	0	0	903	83	14	
			PMIC(C=10)	0	0	904	83	13	
			SIC	0	0	245	126	629	
			ZMIC	0	0	77	34	889	
		70,95,120,145,175,225,250,275	JMIC($\alpha = 1/2$)	3	1	910	83	3	
			JMIC($\alpha = 1/3$)	0	0	910	87	3	
			PMIC(C=1)	0	0	866	117	17	
			PMIC(C=10)	0	0	871	115	14	
			SIC	0	0	226	173	601	
			ZMIC	0	0	76	58	866	

TABLE 7.6: The frequencies of the number of change points $\hat{K}$ estimated by the 1d fused lasso with each criterion over 1000 simulations. The true number of change points K is 8.

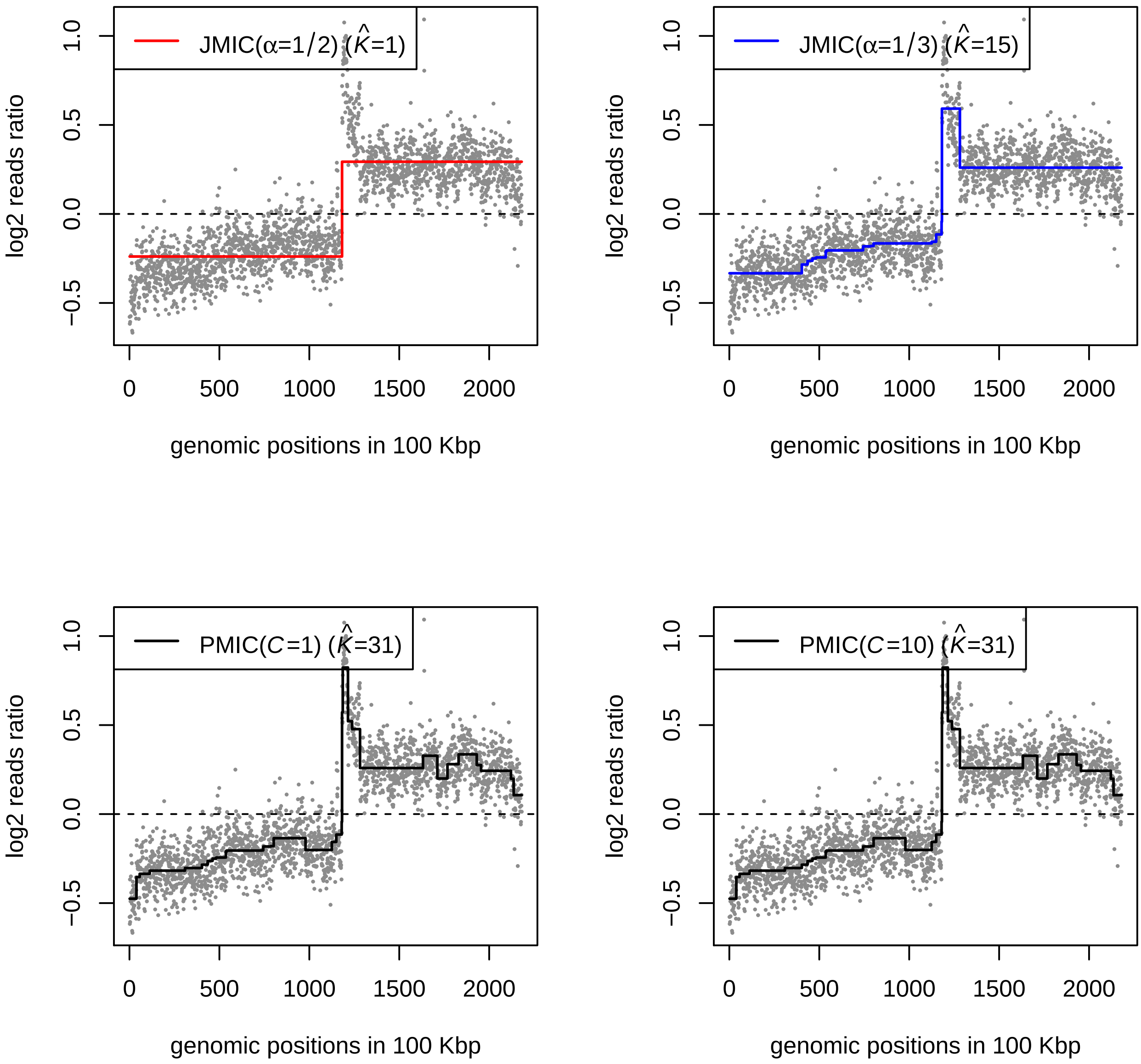


FIGURE 7.1: Chromosome 1 with the identified change points by the 1d fused lasso models based on JMIC($\alpha = 1/2,\ 1/3$) and PMIC($C = 1,\ 10$). $\hat{K}$ denotes the estimated number of change points. The solid line represents the average reads ratio in each segment. The dashed line is for $\hat{y} = 0$, indicating no copy number change.

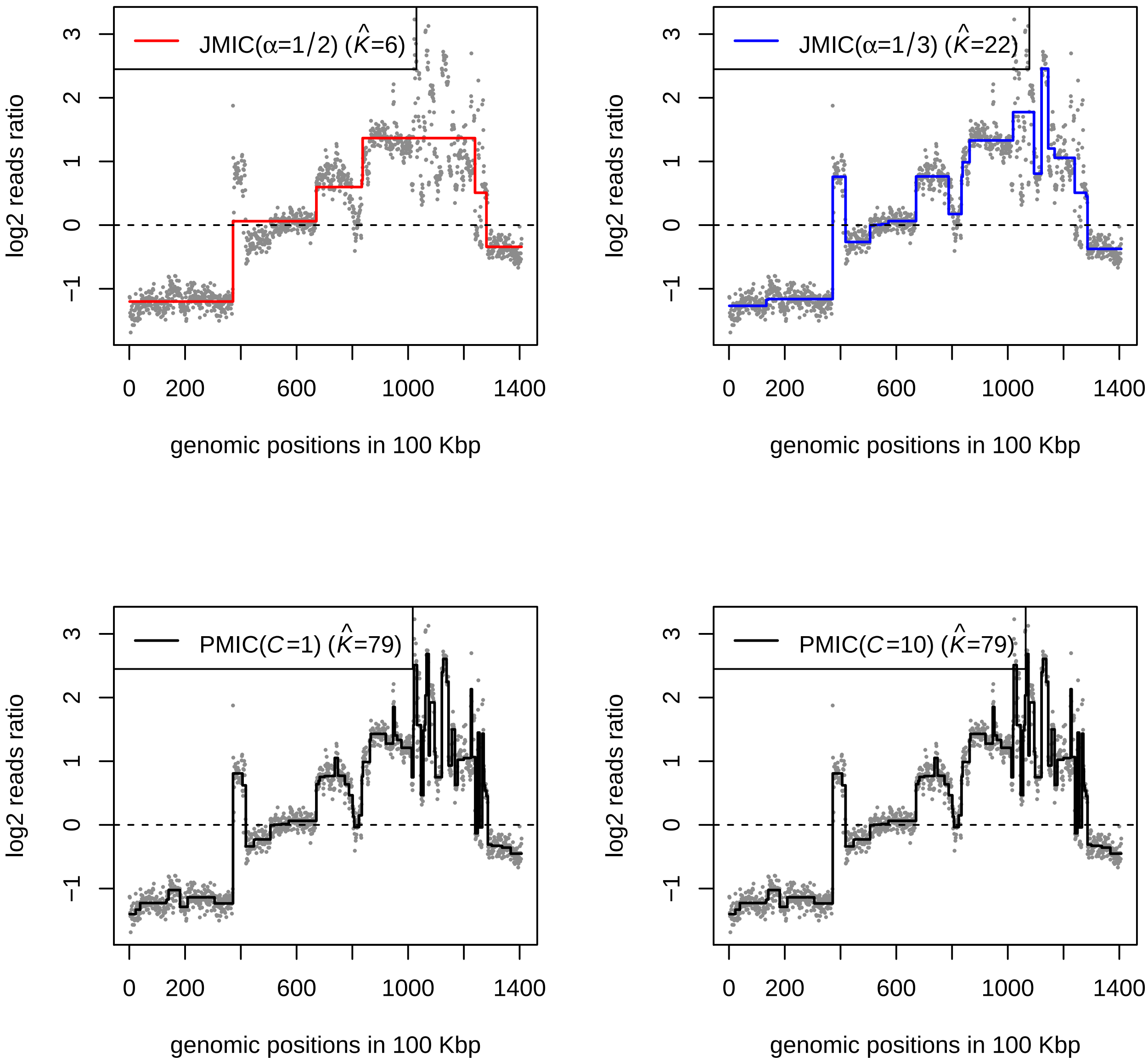


FIGURE 7.2: Chromosome 8 with the identified change points by the 1d fused lasso models based on JMIC($\alpha = 1/2,\ 1/3$) and PMIC($C = 1,\ 10$). $\hat{K}$ denotes the estimated number of change points. The solid line represents the average reads ratio in each segment. The dashed line is for $\hat{y} = 0$, indicating no copy number change.

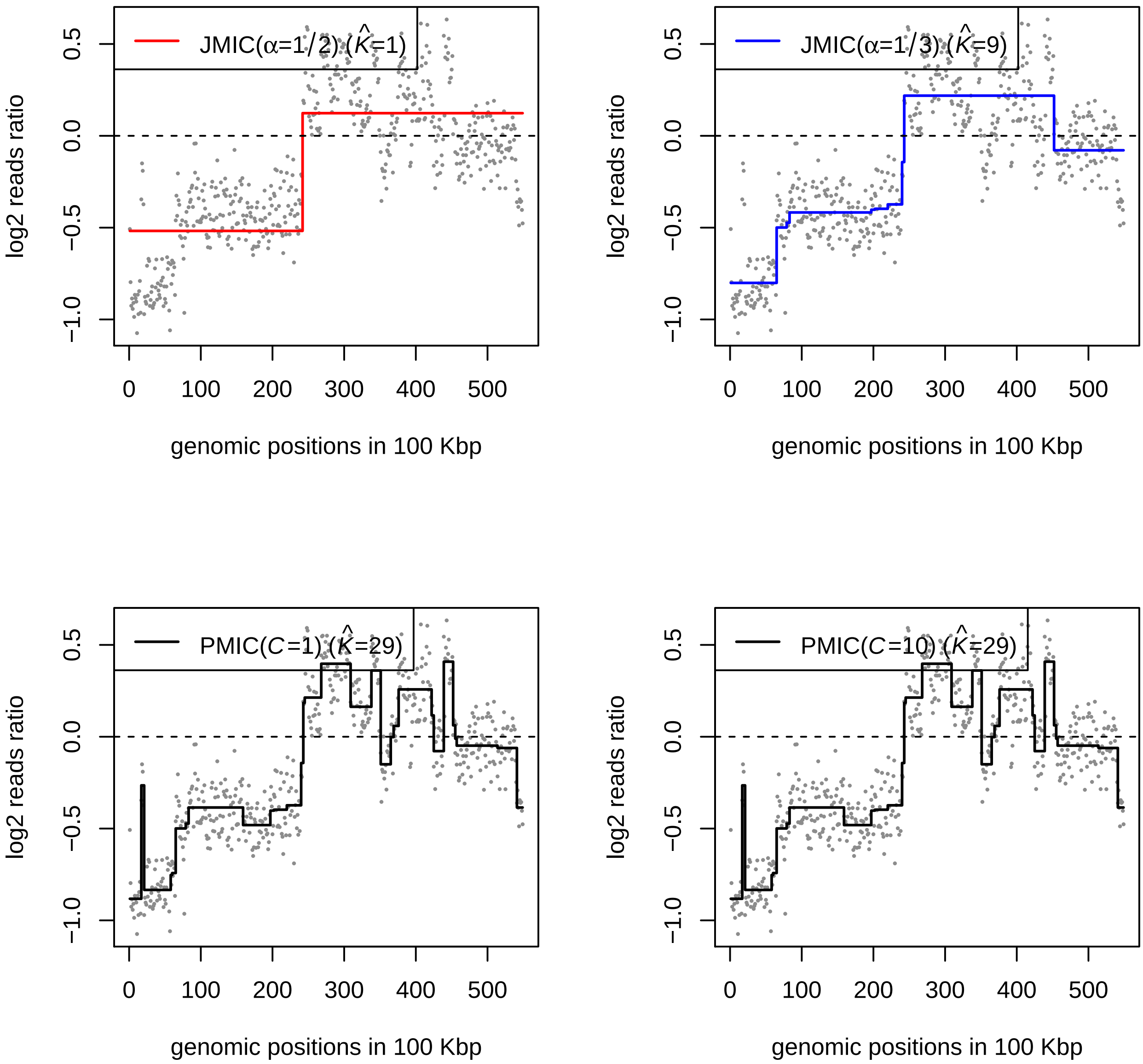


FIGURE 7.3: Chromosome 19 with the identified change points by the 1d fused lasso models based on JMIC($\alpha = 1/2,\ 1/3$) and PMIC($C = 1,\ 10$). $\hat{K}$ denotes the estimated number of change points. The solid line represents the average reads ratio in each segment. The dashed line is for $\hat{y} = 0$, indicating no copy number change.

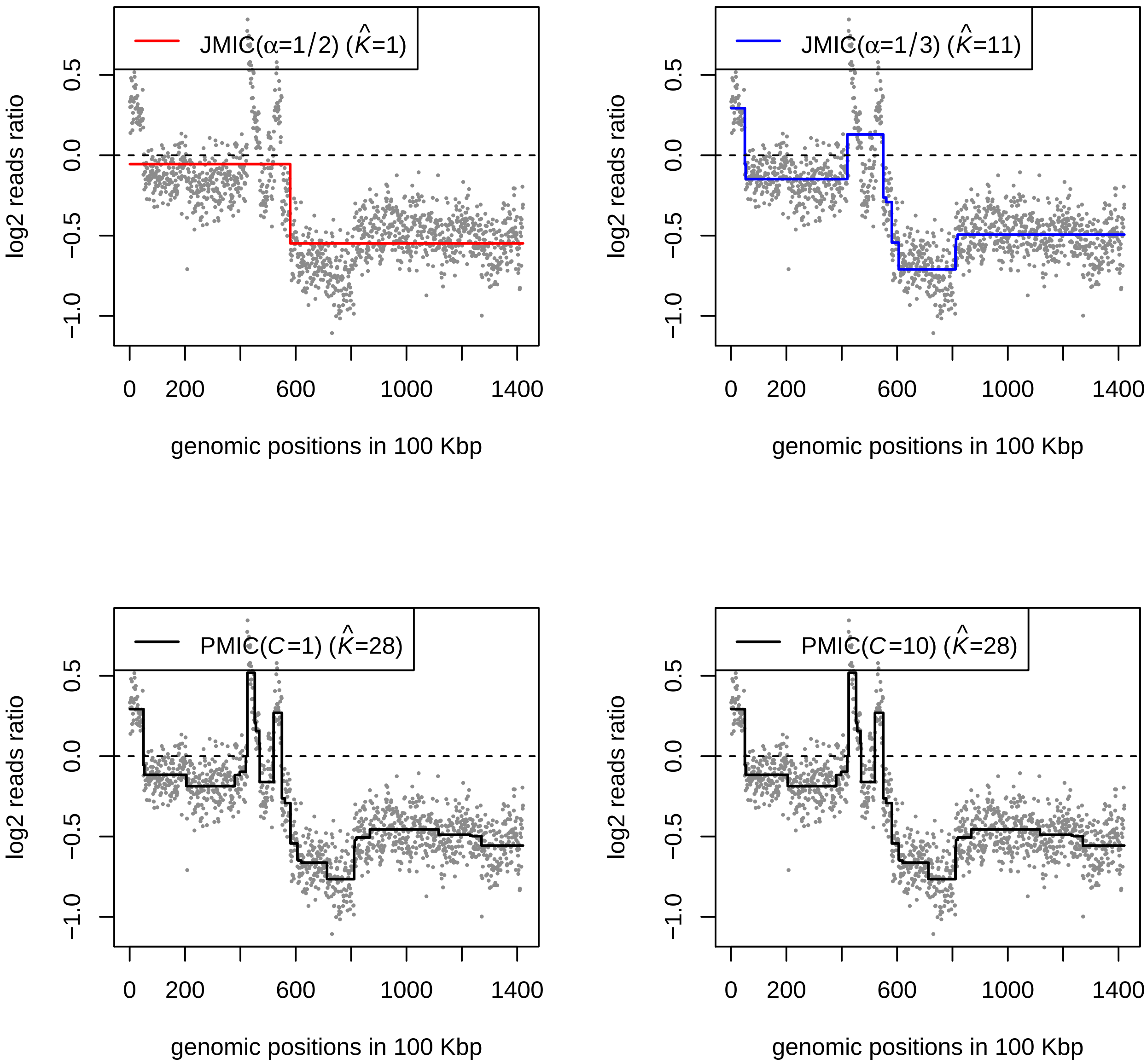


FIGURE 7.4: Chromosome X with the identified change points by the 1d fused lasso models based on JMIC($\alpha = 1/2,\ 1/3$) and PMIC($C = 1,\ 10$). $\hat{K}$ denotes the estimated number of change points. The solid line represents the average reads ratio in each segment. The dashed line is for $\hat{y} = 0$, indicating no copy number change.

References

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). Cnvnator: an approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome research*, 21(6):974–984.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

Alawadhi, F. and Alhulail, D. (2016). Bayesian change points analysis for earthquakes body wave magnitude. *Journal of Applied Statistics*, 43(9):1567–1582.

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376.

Baladandayuthapani, V., Ji, Y., Talluri, R., Nieto-Barajas, L. E., and Morris, J. S. (2012). Bayesian random segmentation models to identify shared copy number aberrations for array cgh data. *Journal of the American Statistical Association*.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., and Pritchard, J. K. (2011). Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome biology*, 12(1):R10.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J.-P., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using gc-content normalization. *Bioinformatics*, 27(2):268–269.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Castellani, C., Melka, M., Gui, J., O'Reilly, R., and Singh, S. (2015). Integration of dna sequence and dna methylation changes in monozygotic twin pairs discordant for schizophrenia. *Schizophrenia research*, 169(1):433–440.

Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2014). Group lasso for structural break time series. *Journal of the American Statistical Association*, 109(506):590–599.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, J. and Gupta, A. (1999). Change point analysis of a gaussian model. *Statistical Papers*, 40(3):323–333.

Chen, J., Gupta, A., and Pan, J. (2006). Information criterion and change point problem for regular models. *Sankhyā: The Indian Journal of Statistics*, pages 252–282.

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

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

Chen, J. and Wang, Y.-P. (2009). A statistical change point model approach for the detection of dna copy number variations in array cgh data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(4):529–541.

Chen, J., Yiğiter, A., Wang, Y.-P., and Deng, H.-W. (2010). A bayesian analysis for identifying dna copy number variations using a compound poisson process. *EURASIP Journal on Bioinformatics and Systems Biology*, 2010(1):1.

Cheung, M.-S., Down, T. A., Latorre, I., and Ahringer, J. (2011). Systematic bias in high-throughput sequencing data and its correction by beads. *Nucleic acids research*, page gkr425.

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L., Russ, C., Nusbaum, C., Meyerson, M., and Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507.

Chung, B. H.-Y., Tao, V. Q., and Tso, W. W.-Y. (2014). Copy number variation and autism: new insights and clinical implications. *Journal of the Formosan Medical Association*, 113(7):400–408.

Ciuperca, G. (2014). Model selection by lasso methods in a change-point model. *Statistical Papers*, 55(2):349–374.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

Du, C., Kao, C.-L. M., and Kou, S. (2016). Stepwise signal extraction via marginal likelihood. *Journal of the American Statistical Association*, 111(513):314–330.

Duan, J., Zhang, J.-G., Deng, H.-W., and Wang, Y.-P. (2013). Cnv-tv: A robust method to discover copy number variation from short sequencing reads. *BMC bioinformatics*, 14(1):1.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Eilers, P. H. and De Menezes, R. X. (2005). Quantile smoothing of array cgh data. *Bioinformatics*, 21(7):1146–1153.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552.

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004). Hidden markov models approach to the analysis of array cgh data. *Journal of multivariate analysis*, 90(1):132–153.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Gill, P. E., Murray, W., and Saunders, M. A. (1997). User's guide for sqopt 5.3: A fortran package for large-scale linear and quadratic programming. Technical report, Tech. Rep. NA 97-4, Department of Mathematics, University of California, San Diego.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Greenshtein, E., Ritov, Y., et al. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Guha, S., Li, Y., and Neuberg, D. (2008). Bayesian hidden markov modeling of array cgh data. *Journal of the American Statistical Association*, 103(482):485–497.

Haccou, P., Meelis, E., and Van De Geer, S. (1987). The likelihood ratio test for the change point problem for exponentially distributed random variables. *Stochastic processes and their applications*, 27:121–139.

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, 43(8):768–775.

Harchaoui, Z. and Lévy-Leduc, C. (2012). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*.

Hassan, N. Z. A., Mokhtar, N. M., Sin, T. K., Rose, I. M., Sagap, I., Harun, R., and Jamal, R. (2014). Integrated analysis of copy number variation and genome-wide expression profiling in colorectal cancer tissues. *PLoS One*, 9(4):e92553.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

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

Hinkley, D. V. and Hinkley, E. A. (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika*, 57(3):477–488.

Hsu, D. (1979). Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis. *Journal of the American Statistical Association*, 74(365):31–40.

Hsu, L., Self, S. G., Grove, D., Randolph, T., Wang, K., Delrow, J. J., Loo, L., and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2):211–226.

Huang, T., Wu, B., Lizardi, P., and Zhao, H. (2005). Detection of dna copy number alterations using penalized least squares regression. *Bioinformatics*, 21(20):3811–3817.

Jacquemont, S., Reymond, A., Zufferey, F., Harewood, L., Walters, R. G., Kutalik, Z., Martinet, D., Shen, Y., Valsesia, A., Beckmann, N. D., et al. (2011). Mirror extreme bmi phenotypes associated with gene dosage at the chromosome 16p11. 2 locus. *Nature*, 478(7367):97–102.

Ji, T. and Chen, J. (2015). Modeling the next generation sequencing read count data for dna copy number variant study. *Statistical applications in genetics and molecular biology*, 14(4):361–374.

Jong, K., Marchiori, E., Van Der Vaart, A., Ylstra, B., Weiss, M., and Meijer, G. (2003). Chromosomal breakpoint detection in human cancer. In *Workshops on Applications of Evolutionary Computation*, pages 54–65. Springer.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). \ell_1 trend filtering. *SIAM review*, 51(2):339–360.

Kim, T.-M., Luquette, L. J., Xi, R., and Park, P. J. (2010). rsw-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC bioinformatics*, 11(1):432.

Kirch, C., Muhsal, B., and Ombao, H. (2015). Detection of changes in multivariate time series with application to eeg data. *Journal of the American Statistical Association*, 110(511):1197–1216.

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S. (2012). cn. mops: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids research*, page gks003.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, 21(19):3763–3770.

Lai, Y.-P., Wang, L.-B., Wang, W.-A., Lai, L.-C., Tsai, M.-H., Lu, T.-P., and Chuang, E. Y. (2017). igc?an integrated analysis package of gene expression and copy number alteration. *BMC bioinformatics*, 18(1):35.

Levy-leduc, C. and Harchaoui, Z. (2008). Catching change-points with lasso. In *Advances in Neural Information Processing Systems*, pages 617–624.

Li, Y. and Zhu, J. (2007). Analysis of array cgh data for cancer studies using fused quantile regression. *Bioinformatics*, 23(18):2470–2476.

Li, Z., Heng, J., Yan, J., Guo, X., Tang, L., Chen, M., Peng, L., Wu, Y., Wang, S., Xiao, Z., et al. (2016). Integrated analysis of gene expression and methylation profiles of 48 candidate genes in breast cancer patients. *Breast cancer research and treatment*, 160(2):371–383.

Loader, C. R. et al. (1996). Change point estimation using nonparametric regression. *The Annals of Statistics*, 24(4):1667–1678.

Louhimo, R., Lepikhova, T., Monni, O., and Hautaniemi, S. (2012). Comparative analysis of algorithms for integration of copy number and expression data. *Nature methods*, 9(4):351–355.

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., and Scherer, S. W. (2014). The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic acids research*, 42(D1):D986–D992.

Magi, A., Tattini, L., Pippucci, T., Torricelli, F., and Benelli, M. (2012). Read count approach for dna copy number variants detection. *Bioinformatics*, 28(4):470–478.

Maher, C. A., Palanisamy, N., Brenner, J. C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T. R., Grasso, C., Yu, J., et al. (2009). Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings of the National Academy of Sciences*, 106(30):12353–12358.

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual review of analytical chemistry*, 6:287–303.

Mazumder, R., Friedman, J. H., and Hastie, T. (2012). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*.

Menezes, R. X., Boetzer, M., Sieswerda, M., van Ommen, G.-J. B., and Boer, J. M. (2009). Integrated analysis of dna copy number and gene expression microarray data using gene sets. *BMC bioinformatics*, 10(1):203.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31–46.

Miller, C. A., Hampton, O., Coarfa, C., and Milosavljevic, A. (2011). Readdepth: a parallel r package for detecting copy number alterations from short sequencing reads. *PloS one*, 6(1):e16327.

Nishii, R. et al. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765.

Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572.

Page, E. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Pan, J. and Chen, J. (2006). Application of modified information criterion to multiple change point problems. *Journal of multivariate analysis*, 97(10):2221–2241.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pettitt, A. (1979). A non-parametric approach to the change-point problem. *Applied statistics*, pages 126–135.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J. (2005). A statistical approach for array cgh data analysis. *BMC bioinformatics*, 6(1):1.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., et al. (1998). High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, 20(2):207–211.

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., Abrahams, B. S., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–372.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nature genetics*, 23(1):41–46.

Qian, J. and Jia, J. (2016). On stepwise pattern recovery of the fused lasso. *Computational Statistics & Data Analysis*, 94:221–237.

Qian, J. and Su, L. (2016). Shrinkage estimation of regression models with multiple structural changes. *Econometric Theory*, 32(6):1376–1433.

Razin, A. and Cedar, H. (1991). Dna methylation and gene expression. *Microbiological reviews*, 55(3):451–458.

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915.

Robertson, K. D. (2005). Dna methylation and human disease. *Nature reviews. Genetics*, 6(8):597.

Rojas, C. R. and Wahlberg, B. (2014). On change point detection using the fused lasso method. *arXiv preprint arXiv:1401.5408*.

Rojas, C. R. and Wahlberg, B. (2015). How to monitor and mitigate stair-casing in l1 trend filtering. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3946–3950. IEEE.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030.

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerrière, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., et al. (2006). App locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy. *Nature genetics*, 38(1):24–26.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268.

Scheinin, I., Sie, D., Bengtsson, H., Van De Wiel, M. A., Olshen, A. B., Van Thuijl, H. F., Van Essen, H. F., Eijk, P. P., Rustenburg, F., Meijer, G. A., et al. (2014). Dna copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome research*, 24(12):2022–2032.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528.

Sen, A. and Srivastava, M. S. (1975). On tests for detecting change in mean. *The Annals of statistics*, pages 98–108.

Shen, J. J. and Zhang, N. R. (2012). Change-point model on nonhomogeneous poisson processes with application in copy number profiling by next-generation dna sequencing. *The Annals of Applied Statistics*, pages 476–496.

Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *The Annals of Statistics*, pages 361–404.

Siegmund, D. (2004). Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika*, 91(4):785–800.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Stephens, P. J., McBride, D. J., Lin, M.-L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., et al. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005–1010.

Stone, J. L., O'Donovan, M. C., Gurling, H., Kirov, G. K., Blackwood, D. H., Corvin, A., Craddock, N. J., Gill, M., Hultman, C. M., Lichtenstein, P., et al. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210):237–241.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 44–47.

Sun, Z., Asmann, Y. W., Kalari, K. R., Bot, B., Eckel-Passow, J. E., Baker, T. R., Carr, J. M., Khrebtukova, I., Luo, S., Zhang, L., et al. (2011). Integrated analysis of gene expression, cpg island methylation, and gene copy number in breast cancer cells by deep sequencing. *PloS one*, 6(2):e17490.

Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., and Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21):2711–2718.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1):18–29.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. (2005). Fine-scale structural variation of the human genome. *Nature genetics*, 37(7):727–732.

Venkatraman, E. S. (1992). *Consistency results in multiple change-point problems*. PhD thesis, to the Department of Statistics.Stanford University.

Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014). The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome biology*, 15(2):R37.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and Tibshirani, R. (2005). A method for calling gains and losses in array cgh data. *Biostatistics*, 6(1):45–58.

Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, 102(7):1141–1151.

Willenbrock, H. and Fridlyand, J. (2005). A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21(22):4084–4091.

Wu, Y. (2011). An ordinary differential equation-based solution path algorithm. *Journal of nonparametric statistics*, 23(1):185–199.

Xi, R., Hadjipanayis, A. G., Luquette, L. J., Kim, T.-M., Lee, E., Zhang, J., Johnson, M. D., Muzny, D. M., Wheeler, D. A., Gibbs, R. A., et al. (2011). Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proceedings of the National Academy of Sciences*, 108(46):E1128–E1136.

Xie, C. and Tammi, M. T. (2009). Cnv-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC bioinformatics*, 10(1):80.

Yan, P. Y., Liu, S., Huo, Z., Martin, A., Nelson, J. B., Tseng, G. C., and Luo, J.-H. (2015). Genomic copy number variations in the genomes of leukocytes predict prostate cancer clinical outcomes. *PloS one*, 10(8):e0135982.

Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome research*, 19(9):1586–1592.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3):172–183.

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, 10:451.

Zhang, N. R. and Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323.

Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics*, 14(11):1.

Zhou, H. and Wu, Y. (2014). A generic path algorithm for regularized statistical estimation. *Journal of the American Statistical Association*, 109(506):686–699.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

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