

Hybrid Bootstrap for Mapping Quantitative Trait Loci and Change Point Problems

by
Ho Keun Sun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2008

Doctoral Committee:

Professor Robert W. Keener, Chair
Professor George Michailidis
Associate Professor Kerby Shedden
Assistant Professor Zhaohui Steve Qin

© Hokeun Sun 2008
All Rights Reserved

To my parents

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Robert Keener, for his constant support and guidance during my graduate study and thesis preparation. It has been a great pleasure to work under his supervision.

I would also like to thank Kerby Shedden, George Michailidis, and Steve Qin for serving on my thesis committee. They have provided thoughtful suggestions and helped me to improve the dissertation work.

I have been fortunate to work in the pleasant atmosphere of the Department of Statistics at the University of Michigan. I appreciate to all the members and staff of the department for their support.

Last but not least at all, I want to thank my parents for their unconditional support and patience. I dedicate this thesis to them.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	vi
LIST OF TABLES	viii
 CHAPTER	
I. Introduction	1
1.1 Hybrid Bootstrap Resampling	1
1.2 Application Examples	4
1.2.1 Mapping Quantitative Trait Loci	4
1.2.2 Change Point Problems	7
1.2.3 Poisson Example from Physics	10
 II. Mapping Quantitative Trait Loci	 12
2.1 Experimental Crosses	12
2.2 Interval Mapping Mixture Model	15
2.3 Estimation and Likelihood Ratio Test	18
2.4 Determination of Thresholds	20
2.4.1 LOD Scores and Likelihood Ratio Test	21
2.4.2 The Permutation Test	22
2.4.3 Nonparametric bootstrap	23
2.5 Hybrid Confidence Regions	23
2.6 Simulation Study	26
2.6.1 Hybrid Bootstrapping Procedures	26
2.6.2 Simulation Results	29
2.7 Data Analysis	32
2.7.1 Missing Genotypes	33
2.7.2 Results	36
2.8 Future Research for Multiple QTL model	40
 III. Change Point Problems	 53
3.1 Shewhart Control Chart	53
3.2 Estimation for Post-Change Mean in a Normal Shift	55
3.2.1 Model and Estimation	56
3.2.2 Normal Example	58
3.2.3 Bayesian Test Statistics	59
3.2.4 Simulation Study	61

3.3	More Topics and Future Research	62
3.3.1	Poisson Process Change Point Problems	63
3.3.2	Change Detection for Variation in Economic Time Series	67
IV.	Inconsistent Hybrid Bootstrap Confidence Region	72
4.1	Model and Estimation	72
4.2	Inconsistent Coverage Probability	75
V.	Summary and Conclusion	82
BIBLIOGRAPHY	85

LIST OF FIGURES

Figure

2.1	A backcross experiment, with four progeny. (Typical experiments contain more than 100 progeny.) Only one pair of homologous chromosome is shown.	43
2.2	A intercross experiment, with four progeny. (Typical experiments contain more than 100 progeny.) Only one pair of homologous chromosome is shown.	44
2.3	Histograms of the phenotype distributions in the parental strains, the F_1 generation, and the backcross generation.	45
2.4	The likelihood ratio test statistics (solid line) and the 95% hybrid quantile estimates (dotted line) for each (θ, J) are shown. The hybrid confidence regions for $(\theta, J) = (0, 1)$ are corresponding to the areas where the quantiles estimates are higher than the test statistics.	46
2.5	The likelihood ratio test statistics (solid line) and the 95% hybrid quantile estimates (dotted line) for each (θ, J) are shown. The hybrid confidence regions for $(\theta, J) = (0.4, 3)$ are corresponding to the areas where the quantiles estimates are higher than the test statistics.	47
2.6	The box plots of the distribution of the 95% quantile estimates of the log likelihood ratio test statistics, using the non-parametric bootstrap(NPB), the permutation(PER), the parametric bootstrap(PBT), and the hybrid bootstrap(HBT) with equally spaced marker distances of $5cM$ in the upper panel and $20cM$ in the lower panel, respectively.	48
2.7	The genetic marker distance map along 12 chromosomes. These chromosome have 18, 15, 21, 14, 12, 17, 15, 17, 13, 9, 13, and 11 markers, respectively.	49
2.8	Plot grid of genotype data along a total of 175 markers and 107 individuals. The genotypes HH , HL , and missing markers are displayed in the colors red, blue, and white, respectively.	49
2.9	The log likelihood ratio test statistics for (θ, J) on the whole genome are shown on the upper plot. The following plots are the log likelihood ratio test statistics for each chromosome.	50
2.10	The log likelihood ratio test statistics for (θ, J) on the whole genome with thresholds computed by the non-parametric bootstrap(NPB, dotted line), the permutation(PER, dash-dot line), and the parametric bootstrap(PBT, dashed line) are shown.	51

2.11	The log likelihood ratio test statistics for (θ, J) on the chromosome 3 and 12 with thresholds computed by the non-parametric bootstrap(NPB, dotted line), the permutation(PER, dash-dot line), and the parametric bootstrap(PBT, dashed line) are shown, respectively.	51
2.12	The constrained maximum likelihood estimates for $\eta_0(\theta, J)$ in red and $\eta_1(\theta, J)$ in blue are shown on the whole genome.	52
3.1	The confidence intervals for θ based on the likelihood ratio test statistics on the upper panel and the Bayesian test statistics on the lower panel.	70
3.2	The coverage probabilities of the confidence interval for θ based on the likelihood ratio test statistics (solid line) and the Bayesian test statistics (dashed line). . . .	71

LIST OF TABLES

Table

2.1	The coverage probabilities of 95% confidence intervals constructed by different methods; the non-parametric bootstrap(NPB), the permutation(PER), the parametric bootstrap(PBT), and the hybrid bootstrap(HBT) with the different marker distance.	31
2.2	The coverage probabilities of hybrid bootstrap regions under different sample sizes (n=200, 500, and 1000), nominal levels (95% and 90%), and QTL locations. . . .	32
2.3	The quantile estimates for the log likelihood ratio $\hat{q}(\theta, J)$ and the widths of the confidence intervals for the QTL location on the whole genome are shown. 95% confidence regions were computed by the non-parametric bootstrap(NPB), the permutation(PER), and the parametric bootstrap(PBT).	37
2.4	The quantile estimates for the log likelihood ratio $\hat{q}(\theta, J)$ and the widths of the confidence intervals for the QTL location on the chromosome 3 and 12 are shown, respectively. 95% confidence regions were computed by the non-parametric bootstrap(NPB), the permutation(PER), the parametric bootstrap(PBT), and the hybrid bootstrap(HBT).	39

CHAPTER I

Introduction

The hybrid bootstrap, introduced in Chuang and Lai [7], uses resampling ideas to extend the duality approach to interval estimation for a parameter of interest when there are nuisance parameters. It is called hybrid resampling because it “hybridizes” the exact method, which uses test inversion, with the bootstrap method that uses the observed data to determine the resampling distribution.

There are several interesting examples in which the data provide substantial information about the nuisance parameter, but limited information about the parameter of interest. In these cases, the confidence region constructed by the hybrid bootstrap may perform much better than the ordinary bootstrap region.

In this chapter we first give a brief explanation of the hybrid bootstrap approach. We then introduce three application examples where the hybrid bootstrap confidence region for a parameter of interest seems to be appealing.

1.1 Hybrid Bootstrap Resampling

A standard approach used to find a confidence set S for an unknown parameter $\theta \in \Omega$ based on data $X \sim P_\theta$ is to invert the family of likelihood ratio tests. Specifically,

if $l(\cdot)$ is the log likelihood function, if $\hat{\theta}$ is the maximum likelihood estimator, if

$$\Lambda(\theta_0) = l(\hat{\theta}) - l(\theta_0)$$

is the generalized log likelihood ratio test statistic used to test $\theta = \theta_0$ versus $\theta \neq \theta_0$,

and if $q(\theta)$ is the upper α -th quantile for the P_θ distribution of $\Lambda(\theta)$, then

$$S = S(X) = \{\theta : \Lambda(\theta) \leq q(\theta)\}$$

is a $1 - \alpha$ confidence region for θ .

Let θ and η denote the parameter of interest and the nuisance parameter, and let $\hat{\theta}$ and $\hat{\eta}$ be the maximum likelihood estimators for these parameters. If $\hat{\eta}_\theta$ maximizes the log likelihood $l(\theta, \eta)$ over η with θ fixed, then the log likelihood test statistic to test $\theta = \theta_0$ versus $\theta \neq \theta_0$ is now

$$\Lambda(\theta_0) = l(\hat{\theta}, \hat{\eta}) - l(\theta_0, \hat{\eta}_{\theta_0}).$$

Let $q(\theta, \eta)$ denote the upper α -th quantile for $\Lambda(\theta)$ under $P_{\theta, \eta}$. The region

$$(1.1) \quad \{\theta : \Lambda(\theta) < q(\theta, \eta), \forall \eta\}$$

has coverage probability at least $1 - \alpha$. But to find it, quantiles $q(\theta, \eta)$ are needed for all θ and η . The region may be too conservative. The ordinary bootstrap confidence region is

$$(1.2) \quad \left\{ \theta : \Lambda(\theta) < q(\hat{\theta}, \hat{\eta}) \right\}.$$

The only quantile necessary to compute this region is $q(\hat{\theta}, \hat{\eta})$. This quantile can be found, if necessary, by bootstrap simulation generating data X^* from $P_{\hat{\theta}, \hat{\eta}}$. The $P_{\theta, \eta}$ coverage of this interval will be approximately $1 - \alpha$ if $q(\hat{\theta}, \hat{\eta})$ accurately estimates $q(\theta, \eta)$. In regular models with large samples this will be the case for two reasons: the

maximum likelihood estimators $\hat{\theta}$ and $\hat{\eta}$ are consistent, and the null distributions for $\Lambda(\theta)$ are approximately independent of θ . In practice, the bootstrap region (1.2) often works well with moderate sample sizes, but with smaller samples its performance is suspect.

The hybrid bootstrap confidence region is

$$(1.3) \quad S = S(X) = \{\theta : \Lambda(\theta) < q(\theta, \hat{\eta}_\theta)\}.$$

To compute S , quantiles $q(\theta, \hat{\eta}_\theta)$ are necessary for all θ . These can be found by bootstrap simulation generating data X_θ^* from $P_{\theta, \hat{\eta}_\theta}$ for values of θ in a reasonably fine grid. Since multiple simulations are required, the computational burden to compute the hybrid region S is greater than that for the ordinary bootstrap, but with modern computing this is often feasible. Note that bootstrap simulations to find $q(\theta, \eta)$ for all θ and η to compute the first interval (1.1) would need to be done for a grid of values for θ and η , posing a greater burden than the simulations necessary for the hybrid region S in (1.3).

The $P_{\theta, \eta}$ coverage for the hybrid region S will be approximately $1 - \alpha$ if $q(\theta, \hat{\eta}_\theta)$ accurately estimates $q(\theta, \eta)$. As with the bootstrap region, this should be the case in large samples but may be suspect with small samples. But there are several interesting examples in which the data provide substantial information about the nuisance parameter η , but limited information about the parameter of interest θ . In these cases, $q(\theta, \hat{\eta}_\theta)$ may be a much better estimator of $q(\theta, \eta)$ than $q(\hat{\theta}, \hat{\eta})$ and the hybrid region may perform much better than the ordinary bootstrap region.

1.2 Application Examples

1.2.1 Mapping Quantitative Trait Loci

Quantitative trait loci (QTL) are of important scientific and economic value in medical research, and in plant and animal breeding. These are the genes responsible for variation in quantitative traits. The development of biochemical markers has led to a proliferation of studies aimed at identifying and characterizing QTL responsible for variation in quantitative traits such as blood pressure, tumor mass, and survival time after an infection.

Knowledge of the locations and actions of the QTLs helps us to understand the biochemical basis of these traits and of their evolution over time. In agricultural experiments, this knowledge may be used to design crosses leading to improved products. In biomedical experiments, the enhancement of understanding of the biochemical basis in the traits aids in identifying new drug targets.

Since the seminal paper of Lander and Botstein [19] there has been ongoing interest in experiments and statistical methods to find and locate quantitative trait loci. Lander and Botstein [19] introduced the concept of interval mapping based on likelihood functions. Interval mapping is currently the most popular approach for QTL mapping in experimental crosses. The method makes use of a genetic map of the typed markers, and assumes the presence of a single QTL. Each location in the genome is posited, one at a time, as the location of the putative QTL.

The LOD score has been proposed as a test statistic in order to detect QTL position. It is simply the log likelihood ratio test statistic scaled by the factor $1/4.61$. Interval mapping links the LOD scores of each typed maker loci, and estimates the QTL position as the location where the LOD curve achieves its maximum. Also, an interval estimate for QTL location can be determined using thresholds for LOD

scores. Lander and Botstein [19] performed extensive computer simulations to find an appropriate LOD threshold for various genome sizes and marker densities, and gave analytical calculations for the case of a very dense marker map. Dupuis and Siegmund [11] also have reported approximate thresholds of LOD scores and provided power calculations for identifying QTL. Churchill and Doerge [8] suggested permutation testing to obtain empirical distributions for a maximum LOD score for each possible location of QTL along chromosomes. Visscher *et al.*[34] used a bootstrap method to generate the sampling distribution of the maximum likelihood estimate (MLE) for QTL. These repeated sampling techniques provide thresholds for significant tests and critical values for interval estimates of the location of QTL. Recently, Chen and Chen [6] established the consistency of the maximum likelihood estimates and found the asymptotic distribution of the likelihood ratio test statistic for the mixture model of the interval mapping when the conditional distribution of phenotypes, given QTL genotypes, are assumed to be normal. Thresholds of the distribution can be approximated easily by using a Monte Carlo simulation.

Manichaikul *et al.*[23] investigated the performance of bootstrap confidence intervals and concluded that they provided very poor performance with respect to the coverage probability and interval width, compared with LOD support intervals and Bayes credible regions(Dupuis and Siegmund [11]). It was also pointed out that an unusual feature of the MLE for QTL is the reason for the failure of the bootstrap. Since the profile likelihood function of the location of QTL exhibits cusps at each genetic marker, the MLE is likely to occur at marker loci due to the change in the direction of the likelihood. After all, the distribution of the MLE depends on the position of the QTL relative to the markers, and this mainly contributes to the breakdown of the bootstrap confidence estimation based on the MLE.

Even though interval mapping models are currently among the most common methods for finding and locating QTL, different approaches has been suggested by other researchers, including Haley and Knott [13], Haley *et al.*[14], and Dupuis and Siegmund [11]. Most of these papers base their analysis on regression models. Dupuis and Siegmund [11] pointed out the similarity between estimating the location of a change-point and estimating the location of a trait locus from data on mapped markers. Consequently, they expected that a Bayesian credible region for a uniform prior distribution on the location of the QTL would provide satisfactory confidence regions. Deng *et al.*[10] proposed the finite logistic regression mixture models for binary trait locus. Cui *et al.*[9] used the generalized Poisson distribution model for count traits data in order to resolve over or under dispersion problems.

Various approaches for multiple QTLs have also been considered by Jansen [16], Broman and Speed [3], and Kao *et al.*[18]. Methods in which only one QTL is considered at a time can be biased for QTL identification and estimation if indeed multiple QTL are located in the same linkage group. The composite interval mapping model suggested by Jansen [16] incorporates multiple regression analysis into interval mapping by conditioning on markers outside an interval of interest. Kao *et al.*[18] proposed using multiple marker intervals simultaneously to map multiple QTLs of epistatic interactions throughout a linkage map. Broman and Speed [3] regarded the QTL mapping problem as one of model selection and provide a modified Bayes Information Criterion (BIC). Although multiple QTLs models seem to be more realistic, we expect that our approach based on interval mapping with a single QTL still provides improved performance and progress in the future for the QTL locating problem.

In the interval mapping model, the recombination rate between the genetic marker

and a QTL often represents the location of QTL. But, recombination events are unlikely to happen and the number of the progenies generated in the experiment is usually limited, so observations which give information about the location of a QTL are very few. This might be another reason that the ordinary bootstrap method using MLE is vulnerable since the quantile estimates of test statistic may vary heavily as the location of the QTL changes. In chapter II, we propose the hybrid bootstrap confidence region for QTL as an alternative of the ordinary bootstrap. The quantiles of test statistic in the hybrid region are estimated for each QTL location, a reasonable grid of QTL, so the region is less affected by unstable MLEs. As a result, the hybrid region performs better than the ordinary bootstrap region for mapping QTL problems.

We describe in details the typical experiments and statistical models used to locate a QTL in chapter II. It contains the results of some large simulation studies to demonstrate the performance of hybrid bootstrap in terms of coverage probabilities and widths. The analysis of a real data set of rice tiller number from Yan *et al.*[39] is then presented.

1.2.2 Change Point Problems

Change point problems arise in applications when observation distributions change at some point in time. For instance, potential observations X_1, X_2, \dots might be modeled as independent with $X_i \sim Q_\eta$ for $i = 1, \dots, \nu$, and $X_i \sim Q_\theta$ for $i > \nu$, where $\eta \neq \theta$. Here $\{Q_\eta\}$ and $\{Q_\theta\}$ lies in some specified parametric family of distributions, and the change point ν is viewed as an unknown parameter. η is often known or simply estimated from historical data, so primary interest in change point problems is estimation for either ν or θ .

There is much literature on change point problems. Interval estimation of the change point ν has been developed by Siegmund [31], Siegmund [32] and Worsley [36]. Smith [33] has employed Bayesian approach to estimate a change point. Testing problems of a sequence of change points in Gaussian model have been investigated by Hawkins [15]. Chen and Gupta [4], and Chen and Gupta [5] estimated variance change points for normal distributions, and multiple covariance change points for Gaussian random vectors, respectively. Inference problems for a Poisson process change point also have been analyzed by Akman and Raftery [1], Raftery and Akman [26], Loader [21], and West and Ogden [35]. The statistical methods and procedures discussed in the literatures above are all based on the specified number of observations where change(s) in distributions takes place in some unknown point(s). This is often called an off-line experiment. Estimation for a change point is usually of main interest in this experiment.

In contrast, on-line monitoring problems are involved in detecting the occurrence of the change as soon as possible. Particularly, in industrial and other applications, the distributional change may be associated with a problem for the underlying process, and data collection is done mainly to detect whether the change has occurred. In the on-line framework, the detection is based on a stopping rule, which usually has the form

$$\tau = \inf\{n : g(x_1, \dots, x_n) \geq \lambda\}.$$

Accordingly, the data are only sampled until the stopping time τ , and the threshold λ is chosen to keep $P(\tau \leq \nu)$ and $E(\tau - \nu)_+$ as small as possible. Representative examples include stopping times given by Shewhart, moving average, or cumulative sum control charts(Montgomery [25]).

One important issue when designing change detection algorithms is the use of

prior information about the changes. Specifically, the situation that the post change parameter θ is unknown is obviously the most interesting from a practical point of view, but is also challenging since there is limited observations about θ due to curtailed data by a stopping procedure. Compared with the off-line estimation problems of the change point, little research has been done on estimation of the post change parameter associated with data observed through a detection time τ . In recent developments, Wu [37] and Wu [38] derived the first-order bias of the post-change mean estimate and a corrected asymptotic normal pivot based on the estimate, assuming that the change point is large and the monitoring limit approaches infinity. He also showed that the estimate for the post change mean is robust even when the variance is also subject to change.

We propose another approach to estimate the post change parameter θ by using the hybrid bootstrap. Since detection stopping times are chosen to make $E(\tau - \nu)_+$ small, the data will provide only limited information about θ . In contrast, unless τ is small, there should be considerable information about η , so this provides yet another example in which the hybrid bootstrap approach seems natural. And since distribution theory in change point problems is generally a challenge, an approach based in part on simulation seems particularly appealing.

In chapter III, the hybrid confidence region for a post change mean is considered after a change is detected by a Shewhart control chart in a sequence of independent normal variables. The hybrid regions are constructed in two different methods: likelihood ratio and Bayesian statistics. Their performance are compared in the simulation study. Poisson process change point problems are then discussed.

1.2.3 Poisson Example from Physics

Researchers in high energy physics are at times interested in estimating a rate $\theta \geq 0$ from a Poisson measurement X with mean $\theta + \eta$. Here η represents a background rate, often considered as known from prior or “off-line” experiments. Also, in many cases $\theta = 0$ is a definite possibility, corresponding to the absence of the particle or phenomena the experiment is trying to detect. This problem is a bit nonstandard since EX is known to be at least η , and there has been some discussion in the physics literature about the proper way to set a confidence interval for θ . The “unified method” of Feldman and Cousins [12] amounts to inverting the family of likelihood ratio tests, and has seen wide interest in physics since its appearance. Bayesian credible regions for θ were developed in Roe and Woodroffe [27] and Roe and Woodroffe [28], when the background parameter is known. Zhang and Woodroffe [40] have shown that the Bayesian approach is robust when the background rate is regarded as a nuisance parameter.

In practice, an assumption that the background rate η is known may be too optimistic. More realistically, information about η may come from count data Y modeled as Poisson with mean $\gamma\eta$. Here the scale factor γ , represents the ratio of the observation times for Y and X . With large γ there is considerable information about the background η , exactly the setting in which the hybrid bootstrap approach seems most promising. Sen and Woodroffe [30] have investigated the performance of the hybrid bootstrap confidence interval in this example, and in their numerical work it seems to perform well.

Despite the positive accounts of the hybrid bootstrap interval’s performance, our investigation shows that the hybrid region S in (1.3) is not consistent—as $\gamma \rightarrow \infty$, $P_{\theta,\eta}(\theta \in S) \not\rightarrow 1 - \alpha$. This surprising problem has to do with discreteness. If $\Lambda_0(\theta)$

denotes the log likelihood ratio test statistic when the background η is known, then Taylor expansion gives

$$\Lambda(\theta) = \Lambda_0(\theta) + \gamma^{-1/2} Z_\gamma + o_p(\gamma^{-1/2})$$

as $\gamma \rightarrow \infty$ with Z_γ asymptotically normal. Since $\Lambda_0(\theta)$ is discrete, Z_γ typically remains relevant in testing at one of the atoms for $\Lambda_0(\theta)$, even if γ is large. Unfortunately, it does not do so in a fashion that preserves consistency for S . In chapter IV, we show that the coverage probability of hybrid confidence regions does not converge to the desired nominal value as information about the nuisance parameter increases.

CHAPTER II

Mapping Quantitative Trait Loci

In this section we consider the problem of estimating the location of quantitative trait loci (QTL) in genetics. QTL are the genes responsible for variation in quantitative traits such as blood pressure, tumor mass, and survival time after an infection. Background on experimental crosses and interval mapping mixture models are first described, and a likelihood based approach to estimation is introduced. Hybrid confidence regions are then proposed for mapping a QTL. In an extensive simulation study, these regions are compared with other approaches including permutation, nonparametric, and ordinary parametric bootstrap. The hybrid method is then employed to analyze a real data set of rice tiller number.

2.1 Experimental Crosses

Most experiments aimed at identifying quantitative trait loci (QTL) begin with two pure-breeding lines which differ in the trait of interest. We will call these the low (L) and high (H) parental lines. The lines are the result of intensive inbreeding, so that each is essentially homozygous at all loci (meaning that, at each locus, offspring receive the same allele from each of their two parents). Crossing these two parental lines gives the first filial (or F_1) generation. The F_1 individuals receive a copy of

each chromosome from each of two parental lines, and so, whenever the parental lines differ, they are heterozygous. All F_1 individuals will be genetically identical, just as the individuals in each of the parental lines were.

In a backcross (See Figure 2.1), the F_1 individuals are crossed to one of the two parental lines, for example, the low line. The backcross progeny, which may number from 100 to over 1000, receive one chromosome from the low parental line, and one from the F_1 . Thus, at each locus, they have genotype LL or HL . As a result of crossing over during meiosis¹, the chromosome received from the F_1 parent is a mosaic of the two parental chromosomes. At each locus, there is a half a chance of receiving the allele from the low parental line (L) and a half a chance of receiving the allele from the high parental line (H). The chromosome received will alternate between stretches of L 's and H 's.

Another common experiment is an intercross (See Figure 2.2). Here, the F_1 individuals are either selfed or crossed to each other. The individuals in the resulting F_2 generation each receive two chromosomes from the F_1 generation, each of which will be a combination of the two parental chromosomes. Thus, at each locus, the F_2 individuals will have genotypes LL , HL or HH .

Investigators produce a number of backcross progeny, generally around 100 individuals, and determine the phenotype trait value for each individual. This value could be quantitative, such as blood pressure or tumor mass, or binary data, like the presence or absence of some disease. Each individual is genotyped at a number of genetic markers, generally spread 10-20 centiMorgans(cM) apart, chosen to cover the genome uniformly². For each marker and each individual, it is observed whether F_1 parent transmitted the L or the H allele. A genetic map specifying the order

¹The process during which gametes or sex cells are formed.

²The cM is the unit of genetic distance, and is equivalent to 1% recombination

of the markers and the intermarker distances will be known or estimated based on data. The objective is to identify genome regions for which there is an association between the phenotype of a backcross individuals and whether it received the L or H allele from the F_1 parent.

Figure 2.3 contains histograms of the phenotype distributions for the paternal strains, the F_1 generation, and the backcross generation, for an imaginary backcross experiment. The paternal strains were chosen to have markedly different phenotype distributions; the L and H strains have average phenotypes of 80 and 20, respectively. While individuals within each strain are genetically identical, there is some variation in the phenotypes due to environmental differences and measurement error. Here, the phenotype distribution for F_1 generation is intermediate between the two parental strains, but shows approximately the same degree of variation, with a standard deviation of about 5.³

It is often assumed, though not always observed, that the degree of environmental variation will be independent of genotypes, as is seen in Figure 2.3—The standard deviation in the parental strains and in the F_1 generation are all about 5. The backcross generation, however, shows greater variation in phenotype because of genetic variation.

The aim of QTL mapping is to identify regions of the genome that are contributing to variation in the trait of interest. In agricultural experiments, this knowledge may be used to design crosses leading to improved products. In biomedical experiments, the goal is to enhance understanding of the biochemical basis of the trait and to identify new drug targets.

Our method will be applied initially to statistical models for the backcross exper-

³The standard deviation may be interpreted as the typical difference from the average. Individuals in the F_1 generation have an average phenotype of about 40, but they typically deviate from that by 5, having a phenotype between 35 and 45.

iment, because of its simplicity. At each locus in the genome, the backcross progeny have one of only two possible genotypes. The intercross is more commonly used in practice, but the analysis of the two types of experiments is similar. The strategies developed for analyzing backcross experiments will generally work for intercross experiments as well.

2.2 Interval Mapping Mixture Model

Lander and Botstein [19] introduced a new approach for mapping a QTL by considering flanked markers. Their method has been called “interval mapping”, and is currently the most popular method for identifying a QTL in experimental crosses. The method makes use of a genetic map of the typed markers, and assumes the presence of a single QTL. Each location in the genome is posited, one at a time, as the location of the putative QTL.

Let us consider a backcross population to the progeny of P_1 and F_1 so that the individuals in the backcross population have four different genotypes at marker 1 and marker 2: H_1H_2/H_1H_2 , H_1H_2/H_1L_2 , H_1H_2/L_1H_2 , and H_1H_2/L_1L_2 . Since the paternal genotype of this generation is fixed, we can code genotype pairs at markers by zero and ones, with a zero representing a maternal gene from one of the pure lines (L_j), and a one representing a gene from the other line (H_j).

Next, consider the case of $k+1$ consecutive markers along a strand of DNA. $M_j = 0$ for the j -th marker genotype of L_j and $M_j = 1$ for the j -th marker genotype of H_j . Then $M = (M_0, \dots, M_k)$ codes genotypes at $k + 1$ consecutive markers. Note that without recombination, M with either be $(0, \dots, 0)$ or $(1, \dots, 1)$. If recombination events along this strand of DNA are modeled as a Poisson process with an assumption

of no crossover interference, then

$$(2.1) \quad \gamma_j = P(M_{j-1} \neq M_j) = \frac{1}{2}(1 - e^{-2d_j/100}),$$

where d_j denotes the distance between the two markers, measured in cM .

Let Y denote the quantitative trait of phenotype measured, and assume that its conditional distribution given the QTL comes from some parametric family $\{Q_\eta\}$. Although there are various possibilities, mixture models, such as those in Chen and Chen [6], seem most natural for the conditional distribution of Y given M . The genotype of the putative QTL cannot be observed but can be inferred from the genotypes of the flanking markers.

Let us begin considering the two marker case, so $k = 1$. If the QTL lies between the two markers, then the distribution of Y given the genotype of the QTL will be Q_{η_0} if $Q = 0$, and Q_{η_1} if $Q = 1$, i.e., $Y|Q = 0 \sim Q_{\eta_0}$ and $Y|Q = 1 \sim Q_{\eta_1}$. If $d = d_1 = a + b$ with a the distance from the first marker to the QTL, and b the distance from the QTL to the second flanked marker, then, with the Poisson model for recombinations in (2.1), the conditional probability for the putative QTL to take genotype $Q = 1$ is

$$P(Q = 1|M = m) = \begin{cases} \frac{1}{4(1-\gamma_1)}(1 + e^{-2a/100})(1 + e^{-2b/100}), & m = (1, 1); \\ \frac{1}{4\gamma_1}(1 + e^{-2a/100})(1 - e^{-2b/100}), & m = (1, 0); \\ \frac{1}{4\gamma_1}(1 - e^{-2a/100})(1 + e^{-2b/100}), & m = (0, 1); \\ \frac{1}{4(1-\gamma_1)}(1 - e^{-2a/100})(1 - e^{-2b/100}), & m = (0, 0), \end{cases}$$

where $\gamma_1 = 0.5(1 - e^{-2d_1/100})$ with d_1 the distance between the first two markers, which is considered known from prior experiments. Generally, γ_1 lies substantially below 0.5. Since d_1 is known and $a + b = d_1$, these conditional probabilities can all

be considered as a known function of a . Letting

$$p_m(a) = P(Q = 1 | M = m),$$

the conditional distribution of Y given M is then mixture

$$Y | M = m \sim [1 - p_m(a)]Q_{\eta_0} + p_m(a)Q_{\eta_1}$$

Along with the marginal probabilities for M given above, this formula specifies a family of joint distribution for M and Y , parameterized by nuisance parameters η_0 and η_1 and the mixture probabilities $p_m(a)$.

Since d is typically quite small, recombinations are fairly rare and most observations of M will be either $(0, 0)$ or $(1, 1)$. Also, given $M = (0, 0)$, the conditional distribution of Y is approximately Q_{η_0} and given $M = (1, 1)$, the conditional distribution of Y is approximately Q_{η_1} , i.e., $p_{(0,0)}(a) \approx 0$ and $p_{(1,1)}(a) \approx 1$. By symmetry,

$$p_{(0,1)}(a) = P(Q = 1 | M = (0, 1)) = 1 - P(Q = 1 | M = (1, 0)) = 1 - p_{(1,0)}(a),$$

and the mixture probabilities $p_{(0,1)}(a)$ and $p_{(1,0)}(a)$ can be parameterized by a single value $\theta \stackrel{\text{def}}{=} p_{(0,1)}(a)$ and $1 - \theta \stackrel{\text{def}}{=} p_{(1,0)}(a)$, respectively. Then $\theta \approx a/d \in [0, 1]$ is a proportional distance between the left flanking marker and the QTL.

So, in this model the only observations that provide information about θ , our surrogate for QTL location, are those with $M = (0, 1)$ or $M = (1, 0)$, i.e., the observations in which there is a recombination between two markers. If recombinations are unlikely, we will have much less information about the location θ of the QTL than the nuisance parameters η_0 and η_1 . This makes a hybrid bootstrap approach to interval estimation of θ appealing, especially since the distribution theory necessary, which is often a challenge for mixture models, can be handled by bootstrap simulations.

2.3 Estimation and Likelihood Ratio Test

In our model, the effect of the putative QTL is represented by the difference $|\eta_0 - \eta_1|$ and its position is indicated by θ , if it exists. In real interval mapping, the whole genome, or the whole of several chromosomes, is searched for the detection of a QTL. This involves a collection of marker intervals on each of which a likelihood ratio test is conducted. Suppose a total of k intervals, or $k + 1$ consecutive markers along a strand of DNA are considered. We assume that there is at most one QTL in this strand. With multiple intervals, two parameters are needed to specify QTL location, the mixture probability θ defined before and the index J for the marker interval containing the QTL.⁴ Thus, the location (θ, J) means that a QTL is located $\theta \cdot d_J$ cM to the right of marker $(J - 1)$, where d_J is the known distance of the J -th marker interval.

Let $(Y_i, M_{i0}, \dots, M_{ik})$, $i = 1, \dots, n$ be the observed quantitative trait value and the marker genotypes of individual i from a random sample of sized n from a backcross population. Let us define two genotypes of the J -th marker interval of individual i as $\tilde{M}_i = (M_{iJ}, M_{i(J+1)})$. Since $Y|Q = 0 \sim Q_{\eta_0}$ and $Y|Q = 1 \sim Q_{\eta_1}$, the joint probability density function (pdf) of (Y_i, \tilde{M}_i) is

$$\prod_{i=1}^n q(m_i) f(y_i | m_i, \theta, J, \eta_0, \eta_1),$$

where $q(m_i)$ is the probability mass function of marker genotypes. Since the $q(m_i)$ do not involve any unknown parameters, they can be dropped from the likelihood function. Then the joint pdf of (Y_i, \tilde{M}_i) is proportional to the conditional distribution

⁴Some of authors do not separate these location parameters for QTL because they define the location of QTL as the distance from the very first genetic marker on a chromosome or a whole genome.

of Y_i given \tilde{M}_i , which can be written as

$$Y_i | \tilde{M}_i = (m_j, m_{j+1}) \sim \begin{cases} Q_{\eta_0}, & (m_j, m_{j+1}) = (0, 0) \\ (1 - \theta)Q_{\eta_0} + \theta Q_{\eta_1}, & (m_j, m_{j+1}) = (0, 1) \\ \theta Q_{\eta_0} + (1 - \theta)Q_{\eta_1}, & (m_j, m_{j+1}) = (1, 0) \\ Q_{\eta_1}, & (m_j, m_{j+1}) = (1, 1). \end{cases}$$

Notice that a QTL is assumed to be located exactly at a marker loci if $\theta = 0$ or 1 .

If we let $f_{\eta_0}(f_{\eta_1})$ denote the pdf of $Q_{\eta_0}(Q_{\eta_1})$, the log likelihood function of θ, η_0 and η_1 for the J -th marker interval is

$$\begin{aligned} l(\theta, J, \eta_0, \eta_1) = & \sum_{i \in \{i: \tilde{M}_i = (0,0)\}} \log f_{\eta_0}(Y_i) + \sum_{i \in \{i: \tilde{M}_i = (0,1)\}} \log \{(1 - \theta)f_{\eta_0}(Y_i) + \theta f_{\eta_1}(Y_i)\} \\ & + \sum_{i \in \{i: \tilde{M}_i = (1,0)\}} \log \{\theta f_{\eta_0}(Y_i) + (1 - \theta)f_{\eta_1}(Y_i)\} + \sum_{i \in \{i: \tilde{M}_i = (1,1)\}} \log f_{\eta_1}(Y_i). \end{aligned}$$

With a fixed J , the maximum likelihood estimates (MLE) of $\hat{\theta} = \hat{\theta}(J)$, $\hat{\eta}_0 = \hat{\eta}_0(J)$ and $\hat{\eta}_1 = \hat{\eta}_1(J)$ solve

$$(2.2) \quad \frac{\partial}{\partial \hat{\theta}} l(\hat{\theta}, J, \hat{\eta}_0, \hat{\eta}_1) = 0, \quad \frac{\partial}{\partial \hat{\eta}_0} l(\hat{\theta}, J, \hat{\eta}_0, \hat{\eta}_1) = 0$$

and

$$\frac{\partial}{\partial \hat{\eta}_1} l(\hat{\theta}, J, \hat{\eta}_0, \hat{\eta}_1) = 0.$$

Since the estimates above are rarely explicitly available, some computational algorithm such as the Newton-Raphson method will generally be employed to find these estimates. Then, the maximum likelihood estimator for J is determined as:

$$(2.3) \quad \hat{J} = \arg \max_{J \in \{1, \dots, k\}} l(\hat{\theta}(J), J, \hat{\eta}_0(J), \hat{\eta}_1(J)).$$

Note that the phenotype values of Y_i will be re-used for the MLE profiled with various values for J , but the estimates give different values for each J since the corresponding flanked marker genotypes are different. In the similar manner, the constrained MLE

for η_0 and η_1 can be calculated for a fixed (θ, J) . Let them be denoted by $\hat{\eta}_0(\theta, J)$ and $\hat{\eta}_1(\theta, J)$, respectively.

Finally, the log likelihood ratio test statistic for testing $H_0 : J = j_0, \theta = \theta_0$ is written as

$$(2.4) \quad \Lambda(\theta_0, j_0) = l(\hat{\theta}, \hat{J}, \hat{\eta}_0(\hat{\theta}, \hat{J}), \hat{\eta}_1(\hat{\theta}, \hat{J})) - l(\theta_0, j_0, \hat{\eta}_0(\theta_0, j_0), \hat{\eta}_1(\theta_0, j_0)).$$

If $q(\theta, J, \eta_0, \eta_1)$ denotes the upper α -th quantile for the distribution of $\Lambda(\theta, J)$ under true value of (θ, J) , then the region

$$(2.5) \quad \{(\theta, J) : \Lambda(\theta, J) < q(\theta, J, \eta_0, \eta_1)\},$$

has coverage $1 - \alpha$. Of course, (2.5) is not a confidence region since it depends on the unknown nuisance parameters, but natural confidence intervals arise estimating the quantile or quantile function.

2.4 Determination of Thresholds

A confidence region in (2.5) can be used to identify a chromosomal region in which to concentrate the search for the exact location of a QTL. Since the likelihood ratio test statistic is based on a mixture distribution, the normal asymptotic chi-square distribution theory may fail, and there has been a fair bit of effort estimating the quantiles in (2.5). These quantiles for the test statistic should depend on the size of the genome, the number and spacing of genetic markers, the amount and pattern of missing genotype information, and the true phenotype distribution. Various simulation studies have been conducted to examine distributions of the test statistic to determine threshold values.

Churchill and Doerge [8] suggested permutation testing to obtain empirical distributions for test statistics. Visscher *et al.*[34] used bootstrap resampling procedures

for a threshold value. Recently, Chen and Chen [6] establish the consistency of the maximum likelihood estimates and found the asymptotic distribution of the likelihood ratio test statistic for the mixture models of the interval mapping when the conditional distribution of phenotypes, given QTL genotypes, are normal. Here, we first introduce the LOD scores and its relation to the likelihood ratio because many approaches geneticists have designed and used are based on the LOD scores. Secondly, the permutation method and the non-parametric bootstrap are discussed to determine the thresholds of the LOD scores. In the next section the hybrid bootstrap approach for the likelihood ratio test statistics of interval mapping is described and compared with these methods.

2.4.1 LOD Scores and Likelihood Ratio Test

In the genetics community the LOD score statistic is more popular for inference than the log likelihood ratio test statistic $\Lambda(\theta, J)$ used here. The LOD statistic is essentially the log likelihood ratio test statistic testing whether a QTL exists at (θ, J) against a null hypothesis that there is no QTL, meaning that the individuals' phenotypes follow a single distribution, $Y_i \sim Q_\eta, i = 1, \dots, n$, where $\eta = \eta_0 = \eta_1$. Specifically,

$$(2.6) \quad \text{LOD}(\theta, J) = \frac{l(\theta, J, \hat{\eta}_0(\theta, J), \hat{\eta}_1(\theta, J)) - l_0(\hat{\eta})}{\log 10},$$

where $l_0(\eta) = \sum_{i=1}^n \log f_\eta(y_i)$, the log likelihood when there is no QTL, and $\hat{\eta}$ maximizes l_0 .

The LOD score measures the strength of the evidence for the presence of a QTL at the location (θ, J) , compared to there being no segregating QTL in the backcross. It would aim to test if a QTL exists at a specific location rather than to estimate

the location of the QTL. But, both LOD score and the likelihood ratio test statistic have functional relation each other,

$$(2.7) \quad \Lambda(\theta, J) = \log 10 \left\{ \text{LOD}(\hat{\theta}, \hat{J}) - \text{LOD}(\theta, J) \right\}.$$

Let us define $\text{LOD}(\hat{\theta}, \hat{J}) - \text{LOD}(\theta, J)$ as the re-centered LOD score with its minimum of 0 indicates the most likely location of a QTL. Then, it can be easily shown that the re-centered LOD score multiplied by $\log 10$ is equivalent to the likelihood ratio test statistic in (2.4).

2.4.2 The Permutation Test

One of the most common methods to find the thresholds for LOD scores is using of the permutation distribution for likelihood ratio. This approach has the advantage that it makes no assumptions on the distribution of the phenotype. However, it requires substantial computation for each study since the thresholds depend on the observed data.

Suppose we permute (Y_1, \dots, Y_n) with marker genotypes fixed. Repeating this gives a simulation approximation to the no QTL permutation distribution of the maximum LOD score. If q is the α -th quantile for this distribution, the LOD confidence region contains all locations with an LOD score above q ,

$$\{(\theta, J) : \text{LOD}(\theta, J) > q\}.$$

By (2.7), this region is the same as

$$\{(\theta, J) : \Lambda(\theta, J) < [\text{LOD}(\hat{\theta}, \hat{J}) - q] \log 10\},$$

so $[\text{LOD}(\hat{\theta}, \hat{J}) - q] \log 10$ should estimate the upper α -th quantile for the distribution

$\Lambda(\theta, J)$. This estimate is reported later in the simulation study.

2.4.3 Nonparametric bootstrap

In bootstrap simulation, $(Y_i^*, M_{i0}^*, \dots, M_{ik}^*)$, $i = 1, \dots, n$, are i.i.d from the empirical distribution of $(Y_i, M_{i0}, \dots, M_{ik})$, $i = 1, \dots, n$. The log likelihood ratio test statistic is then computed from each bootstrap sample. It is

$$\Lambda^*(\theta, J) = l(\hat{\theta}^*, \hat{J}^*, \hat{\eta}_0(\hat{\theta}^*, \hat{J}^*), \hat{\eta}_1(\hat{\theta}^*, \hat{J}^*)) - l(\theta, J, \hat{\eta}_0(\theta, J), \hat{\eta}_1(\theta, J)),$$

where $\hat{\theta}^*$ and \hat{J}^* are the maximum likelihood estimates of θ and J based on the resampled data, respectively.

If \hat{q} is the upper α -th quantile for $\Lambda^*(\hat{\theta}, \hat{J})$, the confidence region will be

$$\{(\theta, J) : \Lambda(\theta, J) < \hat{q}\},$$

so the bootstrap quantile \hat{q} should also estimate the upper α -th quantile for the distribution $\Lambda(\theta, J)$. Notice that unlike permutation testing, the observed combinations of the phenotypes and markers remain together in the bootstrap method.

2.5 Hybrid Confidence Regions

The hybrid bootstrap finds the estimate for the quantile in (2.5) by simulation from a reasonable parametric distribution. The conditional distribution of phenotypes given a QTL genotype are often assumed to follow some parametric distribution. Many quantitative traits observed, such as body mass index and insulin concentration, are regarded as normally distributed. Another type of data arise when the phenotype of interest is measured in counts. The number of roots generated in a plant, and the number of doubled haploid rice tiller [39] are examples of

phenotypes measured in counts. In this case a Poisson model for quantitative traits counts seems appropriate. Finally, some traits are observed as binary data, such as the presence or absence of some disease, and the genes for these traits are known as binary trait loci. A logistic regression model is the most common method for identifying a QTL for these binary data.

In situations where we can reasonably assume that quantitative traits come from some parametric family, the parametric bootstrap should be more powerful than a nonparametric bootstrap. If we define $\hat{\eta}_0(\theta, J)$ and $\hat{\eta}_1(\theta, J)$ as the constrained maximum likelihood estimates for each (θ, J) , and the genotypes on the J -th marker interval for i -th individual is $\tilde{M}_i = (m_{iJ}, m_{i(J+1)})$, then the phenotype values $Y_i^*, i = 1, \dots, n$, are randomly generated from the following resampling distribution:

$$Y_i^* | \tilde{M}_i = (m_j, m_{j+1}) \sim \begin{cases} Q_{\hat{\eta}_0(\theta, J)}, & (m_j, m_{j+1}) = (0, 0); \\ (1 - \theta)Q_{\hat{\eta}_0(\theta, J)} + \theta Q_{\hat{\eta}_1(\theta, J)}, & (m_j, m_{j+1}) = (0, 1); \\ \theta Q_{\hat{\eta}_0(\theta, J)} + (1 - \theta)Q_{\hat{\eta}_1(\theta, J)}, & (m_j, m_{j+1}) = (1, 0); \\ Q_{\hat{\eta}_1(\theta, J)}, & (m_j, m_{j+1}) = (1, 1). \end{cases}$$

If the likelihood ratios $\Lambda^*(\theta, J)$ are computed based on $(Y_i^*, M_{i0}, \dots, M_{ik})$ for fixed (θ, J) , and the upper α -th quantile of the distribution $\Lambda^*(\theta, J)$ is $\hat{q}(\theta, J, \hat{\eta}_0(\theta, J), \hat{\eta}_1(\theta, J))$, then the $(1 - \alpha)$ hybrid confidence regions are defined as

$$(2.8) \quad S_H = \{(\theta, J) : \Lambda(\theta, J) < \hat{q}(\theta, J, \hat{\eta}_0(\theta, J), \hat{\eta}_1(\theta, J))\}.$$

The hybrid bootstrap is also a generalization of the parametric bootstrap, which uses the unconstrained maximum likelihood estimates $(\hat{\theta}, \hat{J})$ instead. So, in the parametric bootstrap simulation the pseudo samples are generated from $Q_{\hat{\eta}_0(\hat{\theta}, \hat{J})}$ and $Q_{\hat{\eta}_1(\hat{\theta}, \hat{J})}$, regardless of the values of (θ, J) . The quantile estimate of parametric bootstrap is then a single value for all (θ, J) , so the computation time is drastically

reduced, compared with the hybrid bootstrap. If the maximum likelihood estimates of $(\hat{\theta}, \hat{J})$ are very near the true values of (θ, J) , the ordinary (parametric) bootstrap gives almost the same quantile estimate as the hybrid bootstrap does, but with much less computation time.

Interval mapping, however, is based on the recombination fraction between two genes, and information about the QTL location parameters (θ, J) is only available if the recombination events occur. In a real experimental cross the sample size is limited, and recombinations observed are extremely rare. In particular, dense genetic markers are usually spaced between 1 to $5cM$, so the recombination rate observed on these marker intervals is approximately 1% – 4.7%. The hybrid approach should be appealing in the interval mapping model with the rare recombination, because it considers the LRT for all θ and J . Although it requires more computation than the ordinary bootstrap, the burden can be reduced by aiming to find only the end points of the confidence interval.

As an illustration, Figure 2.4 and 2.5 display the log LRT and 95% hybrid quantile estimates for a grid of (θ, J) . The results are based on 6 equally spaced genetic markers with an intermarker separation of $20cM$ and sample of 500 obtained from the backcross design. Two models were considered each with a single QTL but at different positions: the first locates the QTL exactly at the first marker loci (2.4), and the other locates the QTL $8cM$ away from the third marker loci to the right (2.5), so that the QTL true location parameters are $(\theta_0, J_0) = (0, 1)$ and $(\theta_0, J_0) = (0.4, 3)$, respectively. Also, the phenotype values are assumed to follow the Poisson distribution with $\eta_0 = 5$ and $\eta_1 = 4$ so the ratio of two phenotype parameters is just 1.25.

In Figure 2.4 and 2.5 the solid line indicates the log LRT $\Lambda(\theta, J)$ in (2.8) over (θ, J) of the whole chromosome. The dotted line is the 95% quantiles of the log

LRT $\Lambda^*(\theta, J)$ based on resampled data, estimating $q(\theta, J, \hat{\eta}_0(\theta, J), \hat{\eta}_1(\theta, J))$ in (2.8). The corresponding region of (θ, J) in (2.8) is then the hybrid bootstrap confidence interval. It is given by the range of θ in J -th marker interval, where the quantiles (the dotted line) are higher than the test statistics (the solid line). It seems that the hybrid confidence regions for QTL give a reasonable range and both include the true QTL location parameter, even with the small shift in the phenotype values. In the next simulation study, we investigate the performance of hybrid confidence regions in terms of coverage probability and distribution of quantile estimates with different marker distance, samples size, and location of QTL.

2.6 Simulation Study

Computer simulation studies are crucial for understanding the relative performance of different methods for locating a QTL. The simulation study reported here includes the comparison of the distribution of quantile estimates and coverage probabilities of the confidence regions, constructed by the permutation, nonparametric bootstrap, ordinary bootstrap and hybrid bootstrap. Also, more extensive simulations for coverage probabilities of hybrid confidence regions under the different sample size and location of a QTL were conducted. First, we describe the procedure of the hybrid bootstrap method used in the simulation study. The extensive simulation results are then reported.

2.6.1 Hybrid Bootstrapping Procedures

Suppose that we have a total of k marker intervals and n progeny samples from a backcross population. Each sample includes $k + 1$ marker genotypes and the pheno-

type values.⁵ The pre-specified values are d_j for all $j = 1, \dots, k+1$, i.e., the distance between $(j-1)$ -st and j -th markers. Also, the true QTL location (θ_0, J_0) , the QTL phenotype parameters η_0 and η_1 , and the family of phenotype distributions are given in the simulation study. The procedure begins with generating a marker interval matrix.

(a) First, we generate $n \times (k+1)$ marker interval matrix with a specified recombination rate. Let M_{ij} , $i = 1, \dots, n$, $j = 0, \dots, k$ be elements of the matrix. Take a random sample of X_{i0} from the uniform distribution with 0 and 1. If $X_{i0} \leq 0.5$ then put 0 into M_{i0} ; otherwise, put 1 into M_{i0} . Let $\gamma_j = (1 - e^{-2d_j})/2$, where d_j is the specified value. Then take X_{ij} from the uniform distribution with 0 and 1 again. If $M_{i(j-1)} = 0$ and $X_{ij} \leq 1 - \gamma_j$, or $M_{i(j-1)} = 1$ and $X_{ij} \leq \gamma_j$, then put 0 into M_{ij} ; otherwise put 1 into M_{ij} . We proceed this until the marker interval matrix is complete. Notice that the M_{ij} for each chromosome forms a Markov chain, with transition probabilities $P(M_{ij} = 1 | M_{i(j-1)} = 0) = P(M_{ij} = 0 | M_{i(j-1)} = 1) = \gamma_j$ and with $P(M_{ij} = 1) = P(M_{ij} = 0) = 0.5$.

(b) Next, phenotype values are simulated based on the QTL location (θ_0, J_0) . With the specified Q_{η_0} and Q_{η_1} , the phenotype $Y = (Y_1, \dots, Y_n)$ can be generated from

$$Y_i | (M_{i(J_0-1)} M_{iJ_0} = m) \sim \begin{cases} Q_{\eta_0}, & m = (0, 0) \\ (1 - \theta_0)Q_{\eta_0} + \theta_0 Q_{\eta_1}, & m = (0, 1) \\ \theta_0 Q_{\eta_0} + (1 - \theta_0)Q_{\eta_1}, & m = (1, 0) \\ Q_{\eta_1}, & m = (1, 1) \end{cases}$$

⁵For the simplicity it is often assumed that the markers are equally spaced on a chromosome.

Then, complete $k + 1$ marker genotypes and a phenotype value are obtained for each individual.

- (c) Set a grid of θ , e.g., from 0 to 1 with increments of 0.02. The maximum likelihood estimates for θ , J , η_0 , and η_1 are numerically computed by (2.2) and (2.3). Two dimensional Newton-Raphson algorithm might be necessary for some parametric distributions of the QTL. Then, the log likelihood ratio test statistics for each (θ, J) are computed by (2.4). Here, the constrained maximum likelihood estimates for $\eta_0(\theta, J)$ and $\eta_1(\theta, J)$ should be reserved for all (θ, J) .
- (d) For each (θ, J) , pseudo samples of phenotype values $Y^* = (Y_1^*, \dots, Y_n^*)$ are now generated based on $\eta_0(\theta, J)$ and $\eta_1(\theta, J)$.

$$Y_i^* | (M_{i(J-1)} M_{iJ} = m) \sim \begin{cases} Q_{\hat{\eta}_0(\theta, J)}, & m = (0, 0) \\ (1 - \theta)Q_{\hat{\eta}_0(\theta, J)} + \theta Q_{\hat{\eta}_1(\theta, J)}, & m = (0, 1) \\ \theta Q_{\hat{\eta}_0(\theta, J)} + (1 - \theta)Q_{\hat{\eta}_1(\theta, J)}, & m = (1, 0) \\ Q_{\hat{\eta}_1(\theta, J)}, & m = (1, 1) \end{cases}$$

Let $\Lambda_l^*(\theta, J), l = 1, \dots, N$ be a log likelihood ratio test statistic based on the l -th resampled phenotype values of $Y_l^* = (Y_1^*, \dots, Y_n^*)_l$. Then, we can obtain a total of N log likelihood ratio test statistics for each (θ, J) . If $\hat{q}(\theta, J)$ is the upper α -th quantile of these N test statistics, it is the hybrid bootstrap quantile estimate.

- (e) Finally, the $1 - \alpha$ confidence regions for (θ, J) are obtained from

$$S_H = \{(\theta, J) : \Lambda(\theta, J) < \hat{q}(\theta, J)\}.$$

2.6.2 Simulation Results

In this study we assume a single QTL is present on a chromosome, the genetic markers are equally spaced, and the genotype data are complete. We first simulated 9 equally spaced genetic markers with the marker distance of $5cM$ for a dense set and $20cM$ for a sparse set. The corresponding recombination rates are $\gamma = 0.0475$ and $\gamma = 0.1648$, respectively. The process is assumed to exhibit no crossover interference so that generated markers form a Markov chain, with transition probabilities $P(M_{ij} = 1|M_{i(j-1)} = 0) = P(M_{ij} = 0|M_{i(j-1)} = 1) = \gamma$, $j = 1, \dots, 8$ and with $P(M_{i0} = 1) = P(M_{i0} = 0) = 0.5$. We then locate a single QTL in the different position of two models; $2cM$ in a dense marker set, and $8cM$ in a sparse marker set away from the third marker loci to the right. The true location parameter is then $(\theta_0, J_0) = (0.4, 3)$ in both models. Finally, the phenotype values were generated from the Poisson distribution with $\eta_0 = 6$ and $\eta_1 = 4$. We also fixed the sample sizes as 100 and 200 for the simulation, and we believe they are reasonable sizes for real experimental crosses.

The first simulation is based on 1000 replicates from the model just described. We find four different quantile estimates for $q(\theta, J, \eta_0, \eta_1)$ in (2.5) for each replicate. They are the nonparametric bootstrap(NPB), the permutation(PER), the parametric bootstrap(PBT), and the hybrid bootstrap(HBT) estimates. These quantile estimates are constant in (θ, J) except the hybrid estimate. For the purpose of the comparison, we use the hybrid quantile estimate under true (θ_0, J_0) , i.e., $\hat{q}(\theta_0, J_0)$ ⁶. We now investigate the distribution of these quantile estimates, using box-plots.

Figure 2.6 shows four box plots of the 95% quantile estimates for the log likelihood ratio test statistics based on four different methods. The box plots in the left show

⁶If the quantile estimate $\hat{q}(\theta, J)$ varies heavily in (θ, J) , the estimate $\hat{q}(\theta_0, J_0)$ is not representative.

the distributions of the quantile estimates for dense marker sets ($5cM$), and the box plots in the right are those for sparse marker sets ($20cM$). The true 95% quantiles of $\Lambda(\theta_0, J_0)$ based on 1000 replicates are estimated by 2.7006 for dense sets, and 2.4505 for sparse sets, respectively. The size of quantiles is related to the width of confidence interval, because the confidence regions consist of all values of (θ, J) such that $\Lambda(\theta, J)$ is less than the quantile. The smaller the quantile is, the narrower the confidence interval is.

The box plots in Figure 2.6 show the variation in quantile estimates in our simulation. Permutation estimates have the greatest variability, including some negative values, corresponding to cases in which the QTL is not detected. On occasion, this QTL is even larger than all LOD scores (2.6) across the entire genome. This behavior arises since the method focuses on detecting, rather than locating a QTL. Quantile estimates based on nonparametric bootstrapping are less variable than permutation estimates, but more variable than estimates based on parametric or hybrid bootstrapping. This is natural since the latter methods are based on correct parametric assumptions.

The parametric and hybrid bootstrap quantiles do not seem to be much different and both distributions are quite centered around the true quantiles for both dense and sparse marker sets, although it seems that both tend to under-estimate the true quantile for dense marker sets. However, narrow confidence intervals does not always mean better performance. One prefers intervals to be as small as possible, while maintaining the appropriate level of coverage. Coverage probability is another important measure to compare the performance of confidence intervals.

Table 2.1 shows the coverage probabilities of nominal 95% confidence intervals constructed by the four different methods. We notice that confidence regions by both

distance	NPB	PER	PBT	HBT
$5cM$	0.8660	0.9610	0.9200	0.9370
$20cM$	0.9300	0.8800	0.9280	0.9550

Table 2.1: The coverage probabilities of 95% confidence intervals constructed by different methods; the non-parametric bootstrap(NPB), the permutation(PER), the parametric bootstrap(PBT), and the hybrid bootstrap(HBT) with the different marker distance.

nonparametric and permutation methods have quite different coverage probabilities in dense and sparse marker sets. We suspect this is due to variation in the quantile estimates in the previous simulation. Large quantiles cover more regions, but small quantiles are likely to miss the true location and lead to poor coverage. In contrast, the parametric and hybrid bootstrap quantile estimates are less variable than the other estimates, and their coverage probabilities are near nominal coverage for both the marker distance of $5cM$ and $20cM$. But, the hybrid regions have closer nominal coverage than the parametric regions does. In particular, the hybrid regions have better coverage than the ordinary bootstrap regions in the dense marker sets, where the recombination events are rare.

The last simulation shows coverage probabilities of hybrid regions in various situations. In Table 2.2 the coverage probabilities of hybrid regions with different sample sizes ($n=200$, 500 , and 1000), nominal levels (95% and 90%), and QTL location (the end of the chromosome, and the middle of the chromosome, $(\theta, J) = (0, 1)$, and $(\theta, J) = (0.4, 3)$, respectively) are shown. It seems that the hybrid regions provide the desired nominal coverage regardless of the sample size, and QTL location.

In the first two simulation, hybrid regions were compared with other confidence regions. It was seen that quantile estimate of hybrid approach are less variable than that of any other methods, and coverage probabilities of hybrid regions are the closest to the nominal coverage. The last simulation demonstrates that hybrid regions are reliable and robust. Real QTL experiments do not have equally spaced markers

and exhibit complex patterns of missing genotype data. Although the simulation studies reported are criticized as not being sufficiently realistic, they are among the most complete and realistic such studies, and the results are of considerable value for the assessment of the performance of the QTL mapping methods included.

n	$1 - \alpha$	$(\theta, J) = (0, 1)$	$(\theta, J) = (0.4, 3)$
1000	0.90	0.882	0.897
	0.95	0.944	0.943
500	0.90	0.899	0.898
	0.95	0.953	0.945
200	0.90	0.906	0.893
	0.95	0.956	0.943

Table 2.2: The coverage probabilities of hybrid bootstrap regions under different sample sizes (n=200, 500, and 1000), nominal levels (95% and 90%), and QTL locations.

2.7 Data Analysis

Data for rice tiller number is originally is given and analyzed by Yan *et al.*[39]. In their experiment, two inbred lines, semidwarf IR64 and tall Azucena, were crossed to generate an F_1 progeny population. By doubling haploid chromosomes of the gametes derived from the heterozygous F_1 , a double-haploid(DH) population of 123 lines was founded, which is genetically equivalent to a backcross population. A genetic linkage map was constructed using 175 genetic markers, with a total length of 2005cM, representing a good coverage of 12 rice chromosomes.

The 123 DH lines were planted in a completely randomized design with two replications. Each replicate was divided into different plots, each containing eight plants per line. Tiller numbers were measured for five central plants in each plot 40 days after transplanting. The tiller numbers were averaged from the two replicates. Given that tiller number can be only an integer, the averaged tiller number was rounded to the nearest integer for QTL analysis.

Our analysis is based on the data including marker distance map, genotype information, and rice tiller numbers. The genetic marker distance map along 12 chromosomes is present in Figure 2.7. The distances vary from $0.8cM$ to $43.8cM$. Figure 2.8 is a plot grid of genotype data. The genotypes HH and HL are displayed in the colors red and blue, respectively. The white color means the missing genotype at that loci. A total of 107 individuals only are analyzed, since the phenotypes of 16 remaining individuals are missing and we left out these cases. 90% of the markers on the whole chromosomes are genotyped, but still many markers are missing.

Inference for QTL location in interval mapping models is based on flanked marker genotypes, so we cannot directly apply hybrid bootstrap approach to these data with missing marker genotypes. So, we first need to derive modified equations for the likelihood when some marker information is missing. Then, estimation for QTL location should be based on this modified likelihood. In the next section an approach for missing genotypes are described and the data analysis results are followed.

2.7.1 Missing Genotypes

When marker genotypes are missing, it is generally known that information from other markers in a linkage group can be used to recover some missing information. In their original paper on linkage map reconstruction, Lander and Green [20] outlined a Markov chain method to recover missing information. Jiang and Zeng [17] derived a general algorithm to systematically deal with dominant and missing markers in F_2 and other populations derived from two inbred lines. Particularly, they formulated the algorithm in a way that can efficiently calculate QTL genotype distribution given observed marker phenotypes. Martinez and Curnow [24] proposed using nearby markers to recover information for the individuals with missing markers in QTL

regression mapping model. We apply this approach to our interval mapping model to recover marker information.

Let us revisit the interval mapping model with unknown parameters. We observe $(Y_i, M_{i0}, \dots, M_{ik})$, $i = 1, \dots, n$, where Y_i is a phenotype value and M_{ij} is j -th marker genotype for i -th individual. Assume that $Y|Q = 1 \sim \text{Poisson}(\eta_1)$ and $Y|Q = 0 \sim \text{Poisson}(\eta_0)$. Let d_j be a distance of j -th marker interval, i.e., the distance between $M_{i(j-1)}$ and M_{ij} , $j = 1, \dots, k$ for all i . The proportional location of a QTL is represented by $\theta = a/d_j$, if the QTL lies within j -th marker interval and the distance from $M_{i(j-1)}$ to the QTL is a . Since we estimate a putative QTL location, a grid of either $\theta \in [0, 1]$ or $a \in [0, d_j]$ should be considered.

We can find the nearest known flanking marker genotypes at any given θ , say u for one from the left and v from the right. They may or may not be the marker loci $(j-1)$ and j when θ is in that interval since some individuals have missing markers. Here are two possible cases we have to consider when θ lies between $(j-1)$ -st and j -th marker.

1. Both nearest flanking marker genotypes are known, i.e., the marker genotypes at u and v are known.
2. The very first or last marker on a chromosome is missing so one of u and v is still unknown.

In the first case, the condition probability of $Q = 1$ given flanking marker genotypes is

$$P(Q = 1 | (M_{iu}, M_{iv}) = m) \approx \begin{cases} 0, & m = (0, 0); \\ \theta_{uv}, & m = (0, 1); \\ 1 - \theta_{uv}, & m = (1, 0); \\ 1, & m = (1, 1), \end{cases}$$

where

$$\theta_{uv} = \frac{d_{u+1} + \cdots + d_{j-1} + \theta d_j}{d_{u+1} + \cdots + d_j + \cdots + d_v}.$$

Notice that θ_{uv} is equal to θ in case $u = j - 1$ and $v = j$.

Let

$$p_m(\theta_{uv}) = P(Q = 1 | (M_{iu}, M_{iv}) = m),$$

and the conditional distribution of Y given (M_{iu}, M_{iv}) is then mixture

$$Y | (M_{iu}, M_{iv}) = m \sim [1 - p_m(\theta_{uv})]Q_{\eta_0} + p_m(\theta_{uv})Q_{\eta_1},$$

where Q_η is a Poisson distribution with mean η .

In the second case, if the first marker of i -th individual is missing, i.e., M_{i0} is unknown, the conditional distribution of Y given M_{iv} is

$$(2.9) \quad Y | M_{iv} = m \sim \begin{cases} (1 - \delta_v)Q_{\eta_0} + \delta_v Q_{\eta_1}, & m = 0 \\ \delta_v Q_{\eta_0} + (1 - \delta_v)Q_{\eta_1}, & m = 1, \end{cases}$$

where δ_v is a recombination rate between the known marker M_{iv} and putative QTL location.

$$\delta_v = \frac{1}{2} [1 - \exp(-2((1 - \theta)d_j + \cdots + d_{j+1} + \cdots + d_v)/100)]$$

Similarly, if the last marker is missing, i.e., M_{ik} is unknown, the conditional distribution of Y given M_{iu} is same as the distribution (2.9), but δ_v should be replaced by δ_u , which is

$$\delta_u = \frac{1}{2} [1 - \exp(-2(d_{u+1} + \cdots + d_{j-1} + \cdots + \theta d_j)/100)].$$

The total likelihood function of $(\theta, J, \eta_0, \eta_1)$ is then given by

$$\begin{aligned}
l(\theta, J, \eta_0, \eta_1) = & \sum_i (Y_i \log \eta_0 - \eta_0) I_{(i \in \{i: (M_{iu}, M_{iv}) = (0,0)\})} \\
& + \sum_i (Y_i \log \eta_1 - \eta_1) I_{(i \in \{i: (M_{iu}, M_{iv}) = (1,1)\})} \\
& + \sum_i \log[(1 - \theta_{uv})\eta_0^{y_i} e^{-\eta_0} + \theta_{uv}\eta_1^{y_i} e^{-\eta_1}] I_{(i \in \{i: (M_{iu}, M_{iv}) = (0,1)\})} \\
& + \sum_i \log[\theta_{uv}\eta_0^{y_i} e^{-\eta_0} + (1 - \theta_{uv})\eta_1^{y_i} e^{-\eta_1}] I_{(i \in \{i: (M_{iu}, M_{iv}) = (1,0)\})} \\
& + \sum_i \log[(1 - \delta_v)\eta_0^{y_i} e^{-\eta_0} + \delta_v\eta_1^{y_i} e^{-\eta_1}] I_{(i \in \{i: (M_{i0}, M_{iv}) = (c,0)\})} \\
& + \sum_i \log[\delta_v\eta_0^{y_i} e^{-\eta_0} + (1 - \delta_v)\eta_1^{y_i} e^{-\eta_1}] I_{(i \in \{i: (M_{i0}, M_{iv}) = (c,1)\})} \\
& + \sum_i \log[(1 - \delta_u)\eta_0^{y_i} e^{-\eta_0} + \delta_u\eta_1^{y_i} e^{-\eta_1}] I_{(i \in \{i: (M_{iu}, M_{ik}) = (0,c)\})} \\
& + \sum_i \log[\delta_u\eta_0^{y_i} e^{-\eta_0} + (1 - \delta_u)\eta_1^{y_i} e^{-\eta_1}] I_{(i \in \{i: (M_{iu}, M_{iv}) = (1,c)\})} \\
& - \sum_{i=1}^n \log Y_i!,
\end{aligned}$$

where c means a missing genotype, and the marker genotypes (M_{iu}, M_{iv}) are determined for each $J \in \{1, \dots, k\}$.

2.7.2 Results

Figure 2.9 shows the log likelihood ratio test statistics for each (θ, J) on the whole genome, and on 12 separate chromosome, respectively. The genomic positions corresponding to the lowest point of the curves are the maximum likelihood estimates for the QTL. It turns out that $(\hat{\theta}, \hat{J}) = (0.58, 17)$ on chromosome 3 has the strongest evidence for a QTL on the genome-wide scan. For each chromosome, we find permutation threshold of the likelihood ratio, and it turns out that chromosome 3 and 12 have a QTL because the test statistics on the other chromosomes are all lower than

Table 2.3: The quantile estimates for the log likelihood ratio $\hat{q}(\theta, J)$ and the widths of the confidence intervals for the QTL location on the whole genome are shown. 95% confidence regions were computed by the non-parametric bootstrap(NPB), the permutation(PER), and the parametric bootstrap(PBT).

method	$\hat{q}(\theta, J)$	Chromosome	width	total width
PER	2.2250	3	5.01	5.01
NPB	3.3150	3	6.06	6.34
		12	0.28	
PBT	3.5001	3	6.2	7.12
		12	0.92	

the thresholds. Notice that the test statistics are all less than 2, and mostly around 1 except chromosome 3 and 12.

Confidence regions for the QTL location could be constructed with an appropriate threshold of the test statistics. The regions where the test statistics are less than the threshold are taken as confidence regions for a QTL. For comparison, we apply 4 different methods previously described to compute thresholds. They are the non-parametric bootstrap, the permutation test, the parametric bootstrap, and the hybrid bootstrap.

Figure 2.10 shows the results of the genome-wide scan. The log likelihood ratio test statistics for (θ, J) on the whole genome are solid lines, and three thresholds computed by the non-parametric bootstrap(NPB)—dotted line, the permutation(PER)—dash-dot line, and the parametric bootstrap(PBT)—dashed line are present. Table 2.3 shows the corresponding quantiles and confidence widths for each method. Since the quantile estimate of the parametric bootstrap is the largest among three methods, it produces the widest confidence regions. The permutation method has the narrowest confidence interval and the non-parametric bootstrap is followed.

Both the parametric and non-parametric bootstrap locate a QTL either between marker 13 and marker 20 on chromosome 3 (marker interval RG179–RG910) or between marker 9 and marker 11 on chromosome 12 (marker interval CDO345–

RG181). But, the permutation confidence regions locate a QTL only on chromosome 3 between marker 14 and marker 19 (marker interval CDO337–CDO87). Please refer to Yan *et al.*[39] for the molecular linkage map information of the data, including marker names and distances.

The hybrid bootstrap method was applied to the data, and it turns out that the hybrid quantile estimates $\hat{q}(\theta, J)$ are greater than the log likelihood ratio $\Lambda(\theta, J)$ for all (θ, J) . This means that the entire genome forms the hybrid confidence regions for a QTL. So, hybrid bootstrap fails to properly estimate for the location of a QTL on the genome-wide scan. The result is interesting because the other methods give some thresholds of likelihood ratio test. We investigate the reason that the hybrid bootstrap fails on the genom-wide scan later in the section.

However, scans on individual chromosomes give different results. Fig 2.11 shows the log likelihood ratio test statistics for (θ, J) on the chromosomes 3 and 12 with thresholds computed by the non-parametric bootstrap(NPB)— dotted line, the permutation(PER) — dash-dot line, and the parametric bootstrap(PBT)— dashed line, respectively. A QTL was not detected on the other chromosomes, i.e., the thresholds are higher than the likelihood ratio for all (θ, J) . Table 2.4 shows the corresponding quantile estimates and confidence widths for each method.

The non-parametric confidence regions have the narrowest interval on chromosome 3, but the widest interval on chromosome 12. In contrast, the permutation region is the widest on chromosome 3, but the narrowest on chromosome 12. This may be due to large variation for the quantile estimates noted in the simulation study. The parametric bootstrap regions locate a QTL on chromosome 3, but fail to detect a QTL on chromosome 12 since the quantile estimate is greater than the likelihood ratio for all (θ, J) . The hybrid regions are similar to the parametric bootstrap, but

Table 2.4: The quantile estimates for the log likelihood ratio $\hat{q}(\theta, J)$ and the widths of the confidence intervals for the QTL location on the chromosome 3 and 12 are shown, respectively. 95% confidence regions were computed by the non-parametric bootstrap(NPB), the permutation(PER), the parametric bootstrap(PBT), and the hybrid bootstrap(HBT).

Chromosome	method	$\hat{q}(\theta, J)$	width
3	NPB	2.1902	4.94
	PER	3.4506	6.16
	PBT	2.9625	5.88
	HBT	(3.406, 2.648)	5.82
12	NPB	1.9968	6.92
	PER	0.4964	1.41
	PBT	2.4884	—
	HBT	—	—

are a little narrower on chromosome 3. Since the quantile estimates of the hybrid bootstrap are not constant, the endpoints of the quantiles are reported in the Table 2.4.

The hybrid quantile estimates are based on the resampling distribution of $Q_{\eta_0(\theta, J)}$ and $Q_{\eta_1(\theta, J)}$ for each (θ, J) . So, if $\eta_0(\theta, J)$ and $\eta_1(\theta, J)$ do not vary in (θ, J) , the hybrid regions are not much different from the parametric bootstrap regions. Figure 2.12 shows the constrained maximum likelihood estimates $\hat{\eta}_0(\theta, J)$ in red and $\hat{\eta}_1(\theta, J)$ in blue on the whole genome for each (θ, J) . The two curves are quite symmetric around a horizontal line at 11. The largest difference between the two estimates occurs near marker 17 on chromosome 3, corresponding to the location of the maximum likelihood estimate for (θ, J) . The averages of the estimates $\hat{\eta}_0(\theta, J)$ and $\hat{\eta}_1(\theta, J)$ over (θ, J) are 10.9752 and 11.1607, respectively, and the average difference of $\hat{\eta}_0(\hat{\theta}, \hat{J})$ and $\hat{\eta}_1(\hat{\theta}, \hat{J})$ is 2.3052.

All quantile estimates except the hybrid bootstrap are based on the maximum likelihood estimate $(\hat{\theta}, \hat{J})$ for their computation. In these methods the estimate of the QTL effect, $\hat{\eta}_0(\hat{\theta}, \hat{J}) - \hat{\eta}_1(\hat{\theta}, \hat{J})$ is employed for resampling distribution. However, the hybrid regions consider the estimate $\hat{\eta}_0(\theta, J) - \hat{\eta}_1(\theta, J)$ for all (θ, J) . It turns out that they are almost 0 for some (θ, J) , and very tiny for most of (θ, J) . So the

phenotype distributions for QTL genotype Q and q are not much different from each other in the hybrid method. This is the reason that the hybrid regions fails to locate a QTL on the whole genome of the data.

2.8 Future Research for Multiple QTL model

Interval mapping assumes the presence of a single QTL for a chromosome. One may use interval mapping to identify multiple QTLs, especially when they are on separate chromosomes. Recent efforts in developing methods to identify QTLs have focused on multiple QTL methods. When several QTLs are modelled, one can control for much of the genetic variation in a cross, and thus individual QTLs can be more clearly seen. In contrast, when one models a single QTL at a time, the genetic variation due to other segregating QTLs is incorporated into the “environmental” variation. When two QTLs are linked, the single QTL method of the interval mapping often view them as a single QTL. Searches which allow multiple QTLs do a better job of separating the two loci, and identifying them as distinct. The presence of the interaction between the multiple QTLs, which is also called epistasis in genetics, can only be detected and estimated using models which include multiple QTLs. Incorporating epistatic effects into multiple QTL models will be very difficult, however. If one were to include all possible pairwise interactions, the number of parameters in the model would quickly explode. For this reason most of work actually neglect the possibility of epistasis.

Lander and Botstein [19] briefly mentioned a method for distinguishing linked loci. If, when performing interval mapping, the LOD curve for a linkage group shows two peaks, or a single very broad peak, Lander and Botstein recommended to fix the

position of one QTL at the location of the maximum LOD, and then search for a second QTL on that linkage group. They fix the location of the first QTL, and vary the location of the second QTL along the linkage group. At each location for the second QTL, we calculate a LOD score, comparing the maximum likelihood under the hypothesis of two QTLs at these locations, to that with a single QTL, located where the first QTL was placed. Each individual's contribution to the likelihood has the form of a mixture of four distributions, the four components corresponding to the four possible QTL genotypes, such as Q/Q , Q/q , q/Q , and q/q on the first QTL and the second QTL, respectively.

However, this method has been criticized, pointed out the phenomenon of “ghost QTLs.” When two or more QTLs are linked in coupling, meaning that their effects have the same sign, interval mapping often gives a maximum LOD score at a location in between the two QTLs, even if there does not exist a QTL near that location.

Broman and Speed [3] viewed the problem of the multiple QTLs mapping as one of model selection. Their method assumes that the genetic markers are sufficiently dense, and dispense with interval mapping, considering only the marker loci as putative locations for QTLs. Additionally, with the assumption of additive QTLs and no epistasis, their method focuses on identifying the number and locations of the QTLs with a developed selection method. Let y_i denote the phenotype of individual i , and let $x_{ij} = 1$ or $x_{ij} = 0$ according to whether individual i has genotype MM or Mm respectively, at marker j . Then the linear model

$$y_i = \mu + \sum_{j=1}^P \beta_j x_{ij} + \epsilon_i,$$

is considered, with the ϵ_i independent and identically distributed $N(0, \sigma^2)$. The selection method seeks to identify the subset of markers for which $\beta_j \neq 0$.

They assumes that QTLs are located exactly at marker loci but we can relax this

assumption and bring the interval mapping method back to consider the multiple QTLs problem.

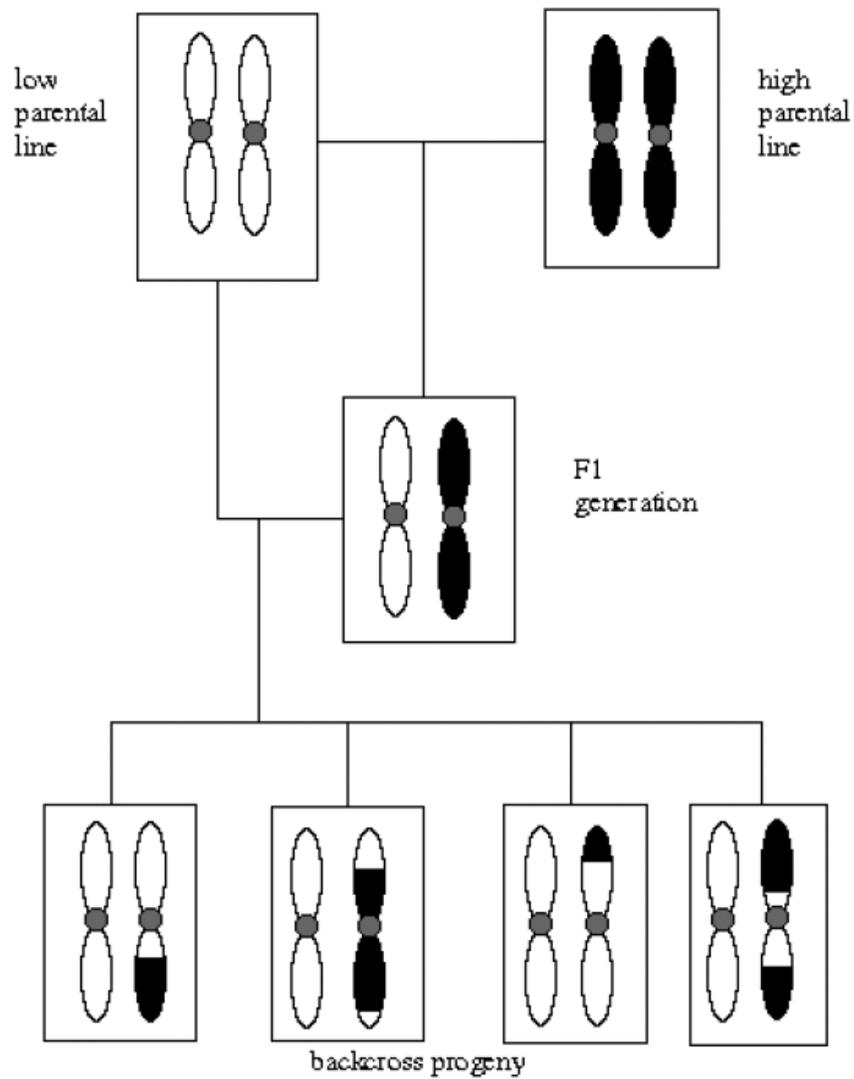


Figure 2.1: A backcross experiment, with four progeny. (Typical experiments contain more than 100 progeny.) Only one pair of homologous chromosome is shown.

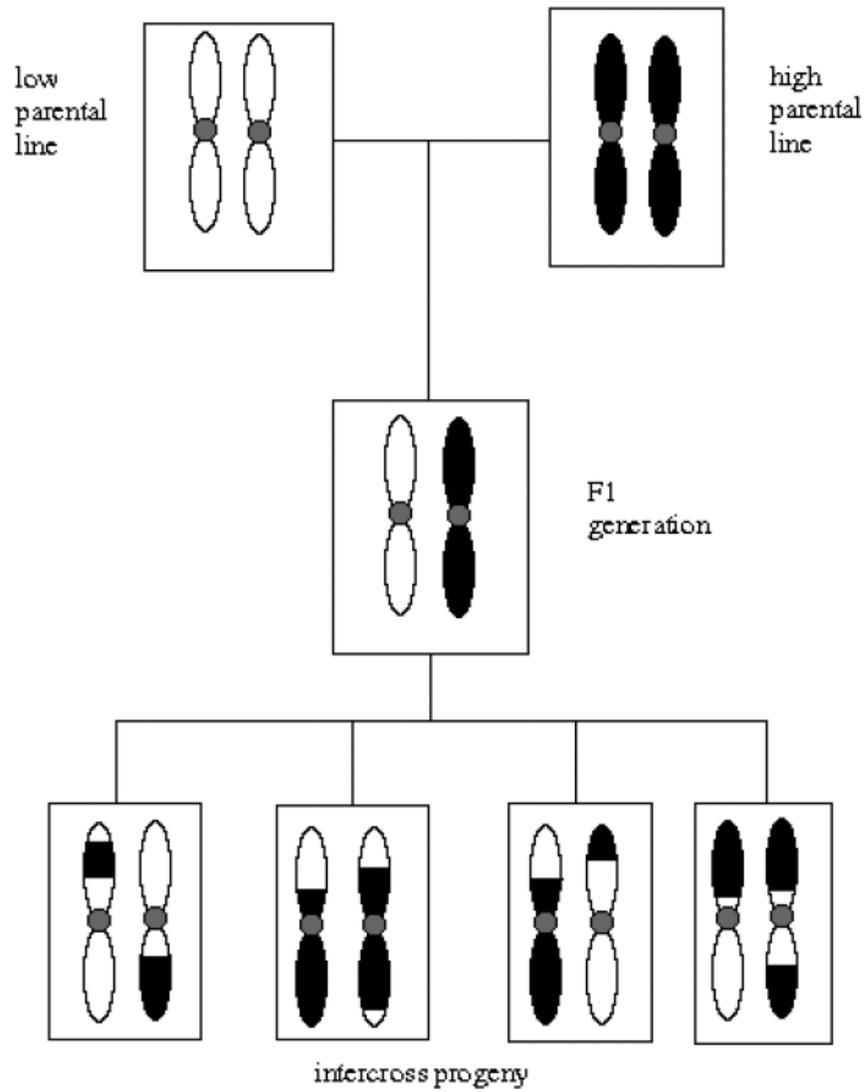


Figure 2.2: A intercross experiment, with four progeny. (Typical experiments contain more than 100 progeny.) Only one pair of homologous chromosome is shown.

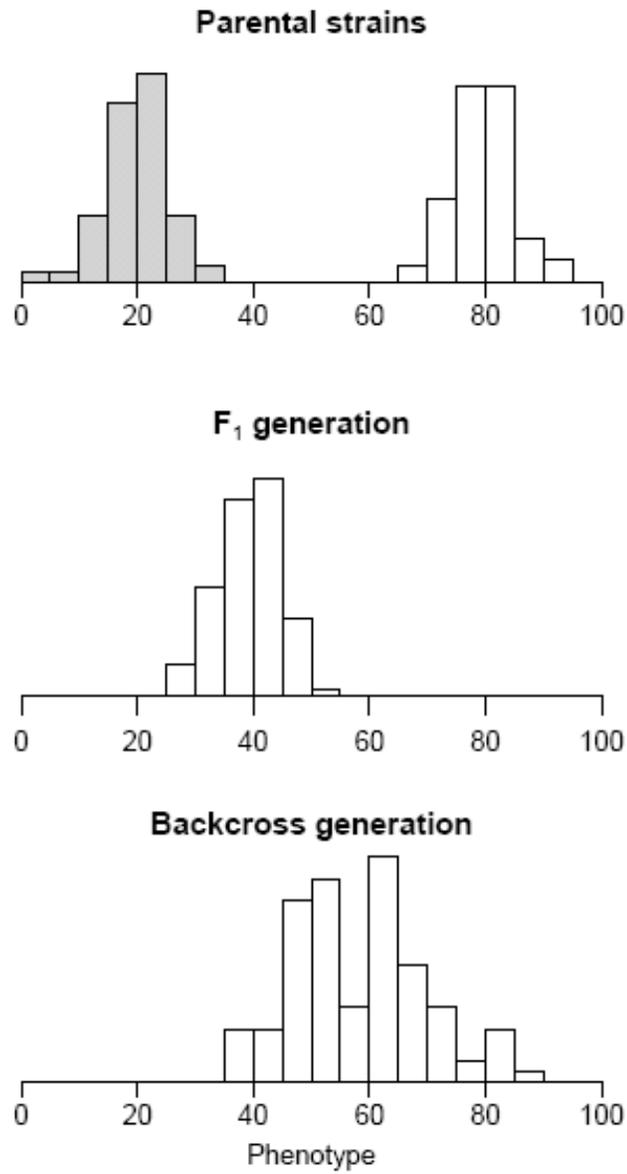


Figure 2.3: Histograms of the phenotype distributions in the parental strains, the F_1 generation, and the backcross generation.

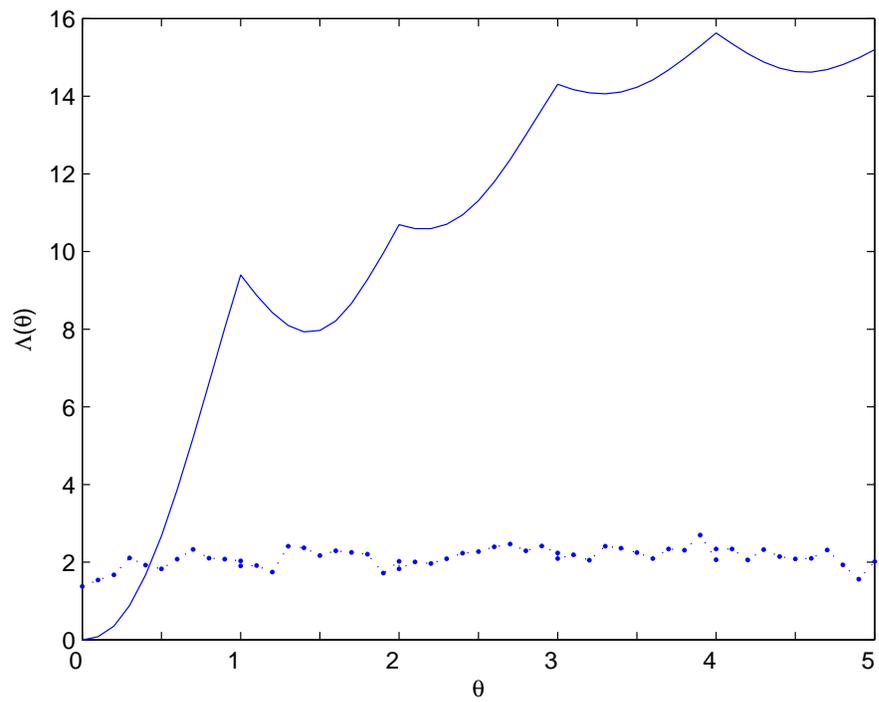


Figure 2.4: The likelihood ratio test statistics (solid line) and the 95% hybrid quantile estimates (dotted line) for each (θ, J) are shown. The hybrid confidence regions for $(\theta, J) = (0, 1)$ are corresponding to the areas where the quantiles estimates are higher than the test statistics.

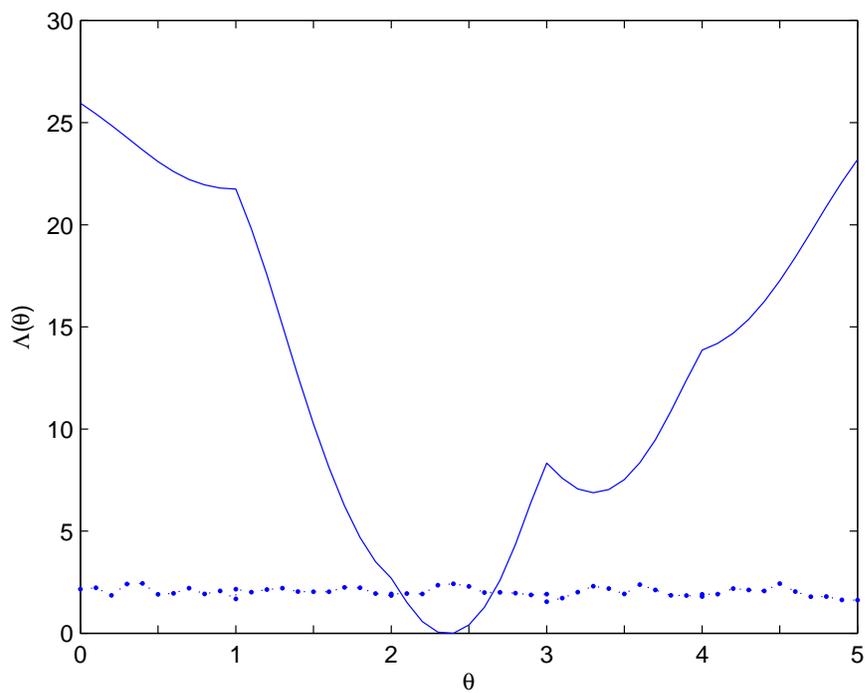


Figure 2.5: The likelihood ratio test statistics (solid line) and the 95% hybrid quantile estimates (dotted line) for each (θ, J) are shown. The hybrid confidence regions for $(\theta, J) = (0.4, 3)$ are corresponding to the areas where the quantiles estimates are higher than the test statistics.

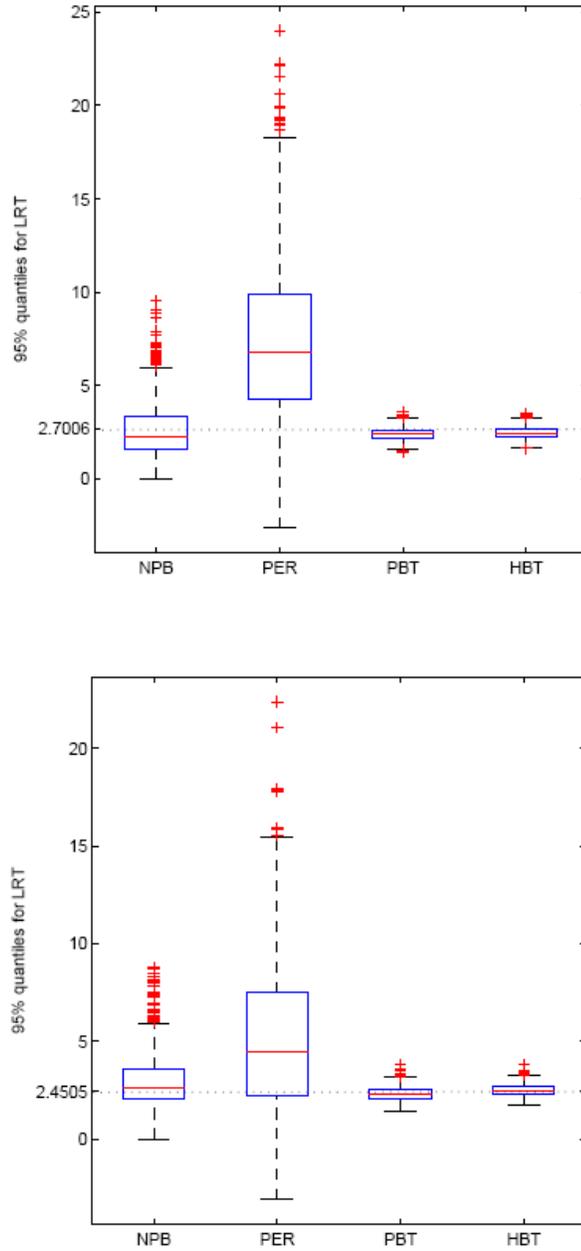


Figure 2.6: The box plots of the distribution of the 95% quantile estimates of the log likelihood ratio test statistics, using the non-parametric bootstrap(NPB), the permutation(PER), the parametric bootstrap(PBT), and the hybrid bootstrap(HBT) with equally spaced marker distances of $5cM$ in the upper panel and $20cM$ in the lower panel, respectively.

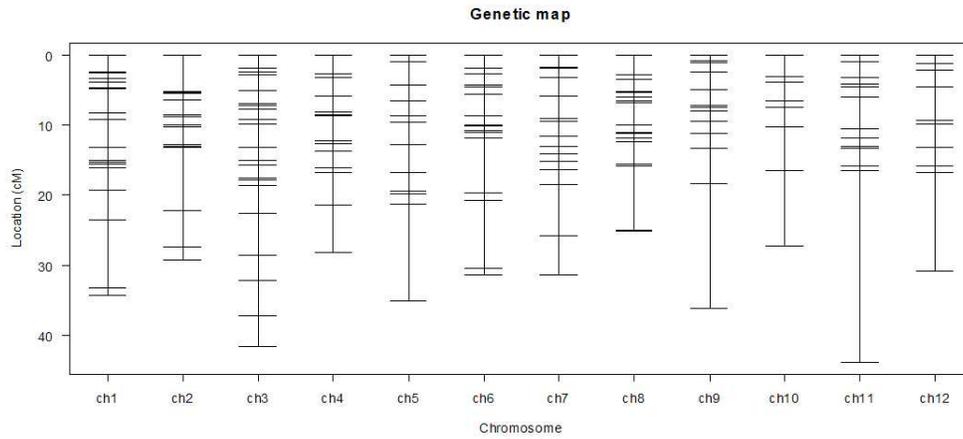


Figure 2.7: The genetic marker distance map along 12 chromosomes. These chromosome have 18, 15, 21, 14, 12, 17, 15, 17, 13, 9, 13, and 11 markers, respectively.

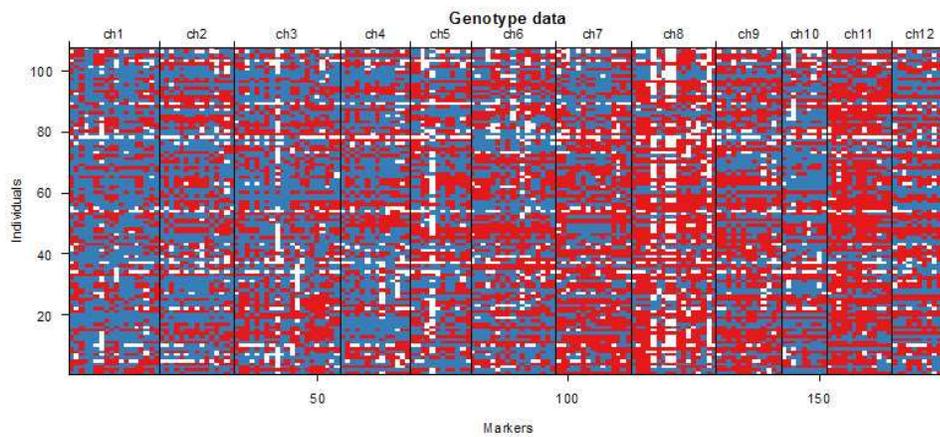


Figure 2.8: Plot grid of genotype data along a total of 175 markers and 107 individuals. The genotypes HH , HL , and missing markers are displayed in the colors red, blue, and white, respectively.

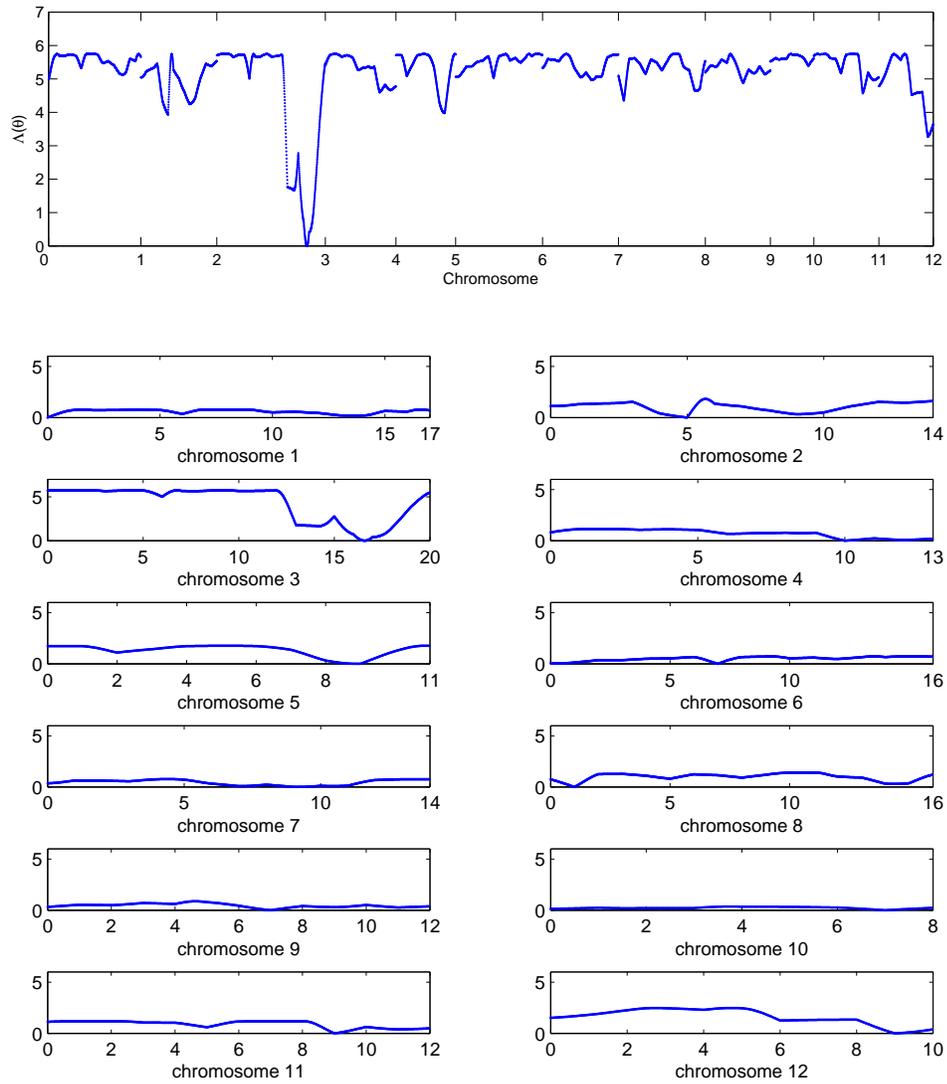


Figure 2.9: The log likelihood ratio test statistics for (θ, J) on the whole genome are shown on the upper plot. The following plots are the log likelihood ratio test statistics for each chromosome.

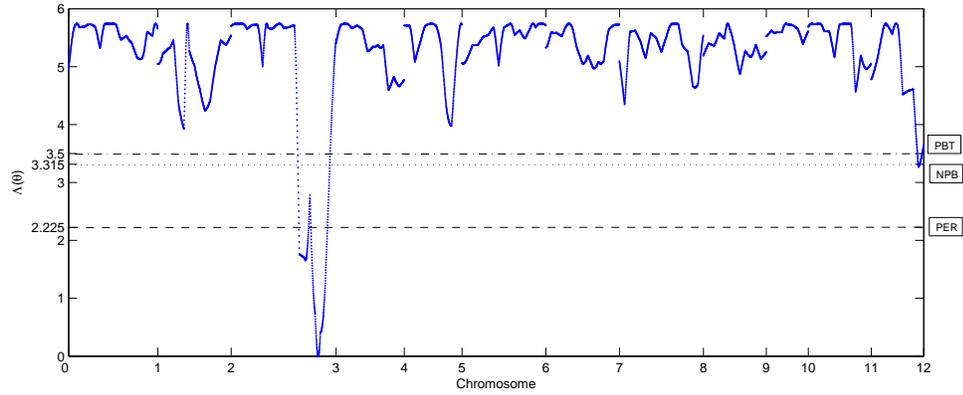


Figure 2.10: The log likelihood ratio test statistics for (θ, J) on the whole genome with thresholds computed by the non-parametric bootstrap(NPB, dotted line), the permutation(PER, dash-dot line), and the parametric bootstrap(PBT, dashed line) are shown.

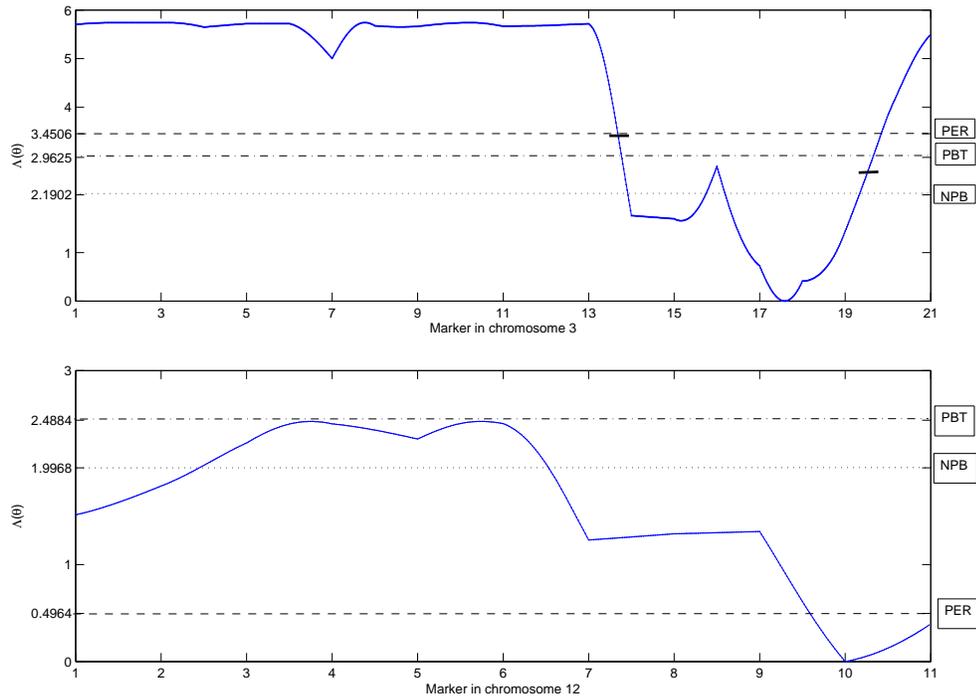


Figure 2.11: The log likelihood ratio test statistics for (θ, J) on the chromosome 3 and 12 with thresholds computed by the non-parametric bootstrap(NPB, dotted line), the permutation(PER, dash-dot line), and the parametric bootstrap(PBT, dashed line) are shown, respectively.

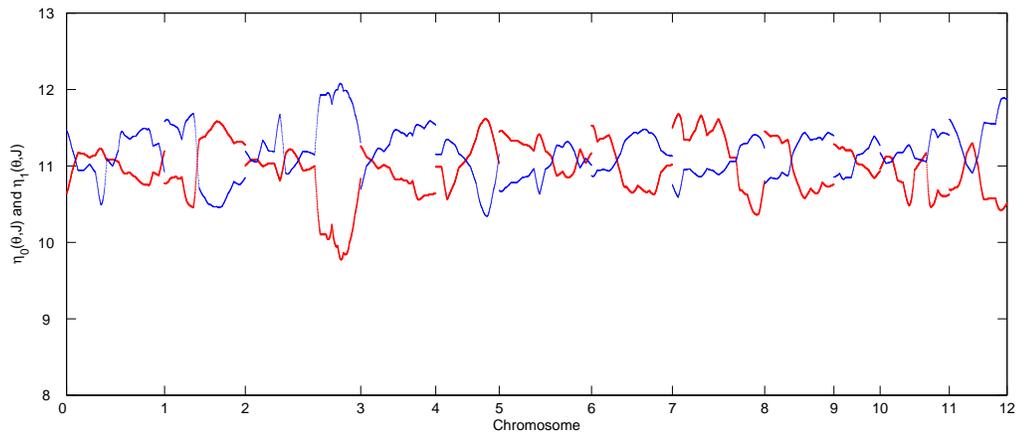


Figure 2.12: The constrained maximum likelihood estimates for $\eta_0(\theta, J)$ in red and $\eta_1(\theta, J)$ in blue are shown on the whole genome.

CHAPTER III

Change Point Problems

In this chapter we first give brief explanation about Shewhart control chart. We then consider the estimation problem for a post change parameter after a change is detected by a Shewhart stopping procedure. In particular, we model independent normal variables with unit variance but shift in mean at some unknown point. Both likelihood ratio and Bayesian statistics are used to find hybrid confidence regions for the post change mean. In the simulation study, their coverage probabilities are compared at each mean difference, including the distribution without changes. Change point Poisson process models are also described.

3.1 Shewhart Control Chart

On-line quality control procedures are used when decisions are to be reached sequentially, as measurements are taken. Situation where the process leaves a controlled condition and enters an out of control state are called disorders. For reasons of safety of the technological process, or quality of production it is necessary to detect disorder quickly with as few false alarms as possible.

These problems are often investigated using a statistical approach. From the statistical point of view, measurement sample is a realization of a random process.

Because of random behavior, large fluctuations can occur in the measurements even when the process is in control, and many result in false alarms if they go beyond certain boundaries. If the measurements when the process is in control have a specific probability distribution, this distribution is assumed to change at some point, once the process is out of control. In a parametric approach, the change leads to different values for distributional parameters.

There are many change detection algorithms also known as control charts in industrial applications. Shewhart charts, the CUSUM charts, and the moving average control charts are the most well-known algorithms. For more a general setting and background, we refer interested readers to Basseville and Nikiforov [2], Montgomery [25], and Lorden [22].

Since Walter A. Shewhart originated the concept of the control chart in the early 1920s, it has become popular in statistical process control. Shewhart-type control charts consist of a graph with time on the horizontal axis and a characteristic of interest (individual measurements or statistics such as mean or range) on the vertical axis. Control limits drawn on the graph provide easy checks on the stability of the process, with values beyond these limits signalling the presence of special causes. We consider the usual control charts with a lower control limit (LCL) and an upper control limit (UCL). If the measurement value x is lower than LCL or higher than UCL, then the process is called out of control.

If the underlying distribution of the observed process is assumed to be normal, then the traditional Shewhart individuals control chart has limits defined by

$$\text{UCL} = \mu + \Phi^{-1}(1 - \alpha/2)\sigma$$

and

$$\text{LCL} = \mu - \Phi^{-1}(\alpha/2)\sigma,$$

where Φ^\leftarrow is the inverse of the standard normal cumulative distribution function Φ , and μ is the mean and σ is the standard deviation of the normal distribution. Level α is the false alarm rate. Typically, μ and σ are unknown. However, we shall assume that they can be estimated from a sample x_1, \dots, x_n of independently and identically distributed random variables. Classical estimators of μ and σ are the sample mean $\bar{x}_n = n^{-1} \sum_{i=1}^n x_i$ and the sample standard deviation $\hat{\sigma} = [(n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2]^{-1/2}$. Thus, typically the alarm is set the first time at which,

$$|x_n - \bar{x}_n| \geq c\hat{\sigma}$$

where c is a tuning parameter and set to control the false alarm rate. Shewhart charts are sensitive to large process shifts but the probability of detecting small shifts fast is rather small.

3.2 Estimation for Post-Change Mean in a Normal Shift

To begin our exploration of hybrid bootstrapping in change point problems, we start with considering the estimation problems of post change parameter θ for the general model. Next, we will see how this approach can be applied to an example in which the data are normal with unit variance, with mean η before the change and mean θ after the change, using the Shewhart detection time

$$(3.1) \quad \tau = \inf\{n \geq 2 : |x_n - \bar{x}_n| \geq c\sqrt{(n-1)/n}\}.$$

A Bayesian approach is then introduced and compared with one based on likelihood estimation in a simulation study.

3.2.1 Model and Estimation

Let x_1, x_2, \dots be the independent random variables with x_1, x_2, \dots, x_ν having distribution F and $x_{\nu+1}, x_{\nu+2}, \dots$ having distribution $G \neq F$. The change point ν , where the distribution shifts from F to G , is regarded as an unknown parameter. In the experiment we observe sequentially x_1, x_2, \dots, x_τ and τ is a stopping time by some procedure, which depends only on the x observed values. Suppose that the stopping time is likely to be occurred after the change point. We also assume that F and G are one-parameter distributions with density f_η for $i \leq \nu$ and g_θ for $i > \nu$, respectively.

Then the likelihood function is

$$L(\theta, \nu, \eta) = \begin{cases} \prod_{i=1}^{\nu} f_\eta(x_i) \prod_{i=\nu+1}^{\tau} g_\theta(x_i) & , \tau > \nu \\ \prod_{i=1}^{\tau} f_\eta(x_i) & , \tau \leq \nu, \end{cases}$$

and the log likelihood function is

$$l(\theta, \nu, \eta) = \sum_{i=1}^{\nu \wedge \tau} \log f_\eta(x_i) + I_{(\tau > \nu)} \sum_{i=\nu+1}^{\tau} \log g_\theta(x_i),$$

where $I_{(\cdot)}$ is an indicator function. The maximum likelihood estimates of η and θ can be obtained for a fixed ν ,

$$\hat{\eta}(\nu) \text{ solves } \sum_{i=1}^{\nu \wedge \tau} \frac{\partial}{\partial \eta} \log f_\eta(x_i) = 0$$

and

$$\hat{\theta}(\nu) \text{ solves } \sum_{i=\nu+1}^{\tau} \frac{\partial}{\partial \theta} \log g_\theta(x_i) = 0 \quad \text{for } \tau > \nu.$$

If observing the data is stopped at the exact change point or before the change actually occurs, i.e., $\tau \leq \nu$, the post change parameter θ can not be estimated. This is the case of the false alarm. Since the change point ν is not known in practice, we cannot distinguish the alarms due to ‘out of control’ process from the false alarm.

In this problem we consider only the cases that the stopping procedure detects a change point, i.e., $\tau < \nu$.

Let us denote $\hat{\nu}$ as a maximum likelihood estimate for the change point. It maximizes the profile likelihood function $l(\hat{\theta}(\nu), \nu, \hat{\eta}(\nu))$ over ν . So,

$$\hat{\nu} = \arg \sup_{1 < \nu < \tau} l(\hat{\theta}(\nu), \nu, \hat{\eta}(\nu)),$$

Similarly, the constrained maximum likelihood estimate of ν is

$$\hat{\nu}_\theta = \arg \sup_{1 < \nu < \tau} l(\theta, \nu, \hat{\eta}(\nu)),$$

for each fixed θ . Notice that $\hat{\nu}_\theta = \hat{\nu}$ if $\theta = \hat{\theta}(\hat{\nu})$. Since the change point ν is discrete here, the estimate $\hat{\nu}_\theta$ is an integer, but may vary for different values of θ .

The log likelihood ratio test statistic of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is then

$$\Lambda(\theta_0) = l(\hat{\theta}(\hat{\nu}), \hat{\nu}, \hat{\eta}(\hat{\nu})) - l(\theta_0, \hat{\nu}_{\theta_0}, \hat{\eta}(\hat{\nu}_{\theta_0})),$$

and the confidence set for θ is

$$S = \{\theta : \Lambda(\theta) < q(\theta)\},$$

where $q(\theta)$ should be the smallest value of q for which

$$P_{\theta, \nu, \eta}(\Lambda(\theta) < q) \geq 1 - \alpha.$$

The quantile $q(\theta)$ is estimated, using the hybrid bootstrap. We can start with simulating the distribution of $\Lambda(\theta)$. Let $\Lambda^*(\theta)$ be a log likelihood ratio test based on the samples generated from $P_{\theta, \hat{\nu}_\theta, \hat{\eta}(\hat{\nu}_\theta)}$.

If $\hat{q}(\theta)$ be the smallest value of q for which

$$P_{\theta, \hat{\nu}_\theta, \hat{\eta}(\hat{\nu}_\theta)}(\Lambda^*(\theta) < q) \geq 1 - \alpha,$$

the hybrid confidence regions are

$$(3.2) \quad S_H = \{\theta : \Lambda(\theta) < \hat{q}(\theta)\}.$$

3.2.2 Normal Example

If the data are normal with unit variance, with mean η before the change and mean θ after the change, i.e., $x_1, x_2, \dots, x_\nu \sim N(\eta, 1)$ and $x_{\nu+1}, \dots, x_\tau \sim N(\theta, 1)$, then the log likelihood function is

$$l(\theta, \nu, \eta) = -\frac{1}{2} \sum_{i=1}^{\nu \wedge \tau} (x_i - \eta)^2 - \frac{1}{2} \sum_{i=\nu+1}^{\tau} (x_i - \theta)^2 I_{(\tau > \nu)} - \tau \log \sqrt{2\pi}.$$

The maximum likelihood estimators of η and θ for a fixed ν are

$$\begin{aligned} \hat{\eta}(\nu) &= \frac{s_\nu}{\nu}, \quad \text{where } s_\nu = \sum_{i=1}^{\nu} x_i \\ \hat{\theta}(\nu) &= \frac{s_\tau - s_\nu}{\tau - \nu}, \end{aligned}$$

and the maximum likelihood estimate for ν is

$$\begin{aligned} \hat{\nu} &= \arg \sup_{1 < \nu < \tau} l(\hat{\theta}(\nu), \nu, \hat{\eta}(\nu)) \\ &= \arg \sup_{1 < \nu < \tau} \left\{ \frac{s_\nu^2}{\nu} + \frac{(s_\tau - s_\nu)^2}{\tau - \nu} \right\}. \end{aligned}$$

The constrained maximum likelihood estimator of ν for each θ is then

$$\begin{aligned} \hat{\nu}_\theta &= \arg \sup_{1 < \nu < \tau} l(\theta, \nu, \hat{\eta}(\nu)) \\ &= \arg \sup_{1 < \nu < \tau} \left\{ -\frac{(\tau - \nu)}{2} \theta^2 + (s_\tau - s_\nu) \theta + \frac{s_\nu^2}{2\nu} \right\}. \end{aligned}$$

Finally, the log likelihood ratio test for each θ is given by

$$\begin{aligned} \Lambda(\theta) &= l(\hat{\theta}(\hat{\nu}), \hat{\nu}, \hat{\eta}(\hat{\nu})) - l(\theta, \hat{\nu}_\theta, \hat{\eta}(\hat{\nu}_\theta)) \\ (3.3) \quad &= \frac{1}{2} \left\{ (\tau - \hat{\nu}_\theta) \theta^2 - 2(s_\tau - s_{\hat{\nu}_\theta}) \theta + \frac{(s_\tau - s_{\hat{\nu}_\theta})^2}{(\tau - \hat{\nu}_\theta)} + \frac{s_{\hat{\nu}_\theta}^2}{\hat{\nu}_\theta} - \frac{s_{\hat{\nu}}^2}{\hat{\nu}} \right\}. \end{aligned}$$

Now, we can find the quantile estimate for the distribution of $\Lambda(\theta)$, using the hybrid bootstrap. For each θ , we first generate $X^*(\theta) = \{x_1^*, x_2^*, \dots, x_\tau^*\}$ from the normal distribution with a mean of $\hat{\eta}(\hat{\nu}_\theta)$ for $i \leq \hat{\nu}_\theta$ and a mean of θ for $i > \hat{\nu}_\theta$, i.e.,

$$\begin{aligned} x_1^*, x_2^*, \dots, x_{\hat{\nu}_\theta}^* &\sim N(\hat{\eta}(\hat{\nu}_\theta), 1) \\ x_{\hat{\nu}_\theta+1}^*, \dots, x_\tau^* &\sim N(\theta, 1), \end{aligned}$$

where $\tau = \tau(x_1^*, x_2^*, \dots)$ is not a fixed stopping time but determined based on the resampled data each time. If $\Lambda^*(\theta)$ is a log likelihood ratio computed from $X^*(\theta)$, and $\hat{q}(\theta)$ is the upper α -th quantile of the distribution of $\Lambda^*(\theta)$, the hybrid confidence region S_H in (3.2) can be easily obtained.

If the difference between θ and η is large, stopping is likely to occur immediately after the change point ν . In this case we have a few observations associated with θ , and this makes estimation for θ to be difficult so hybrid bootstrap seems to be appealing. However, if the difference between θ and η is very small or almost 0, stopping time could occur much later than the change point does. In this case estimation for the change point ν is difficult because the process seems to be from a single distribution. We investigate the estimation results in the simulation study when θ approaches η .

3.2.3 Bayesian Test Statistics

A Bayesian approach to change point problems was suggested by Smith [33]. In his work, inference is based on the posterior probabilities of the possible change points. He considers the cases where the underlying distributions are normal and binomial. We use the Bayesian test statistics in the previous normal example, and compare this approach with one based on the likelihood ratio test statistics in the simulation study.

Let us start with a general Bayesian framework in change point problems. Assuming that the distributions have densities $f(x|\eta)$ and $g(x|\theta)$, the joint distribution

of x_1, \dots, x_τ conditional on η, θ and the change having taken place at ν is given by

$$p(x_1, \dots, x_\tau | \eta, \theta, \nu) = \begin{cases} \prod_{i=1}^{\nu} f(x_i | \eta) \prod_{i=\nu+1}^{\tau} g(x_i | \theta), & \tau > \nu \\ \prod_{i=1}^{\tau} f(x_i | \eta), & \tau \leq \nu. \end{cases}$$

We further assume a prior distribution to be specified over the set of possible change points, given by a mass function $p_0(\nu)$ such that $\sum_{\nu} p_0(\cdot) = 1$. Independently of the assignment of $p_0(\nu)$, we assign a prior density $p_0(\eta, \theta)$ over Θ , the range of possible values of (η, θ) . We then obtain the posterior density of ν

$$p(\nu | x_1, \dots, x_\tau) \propto \int_{\Theta} p(x_1, \dots, x_\tau | \eta, \theta, \nu) p_0(\eta, \theta) p_0(\nu) d\eta d\theta,$$

and the marginal posterior density for θ is given by

$$p(\theta | x_1, \dots, x_\tau) \propto \sum_{\nu} \int_{\eta} p(x_1, \dots, x_\tau | \eta, \theta, \nu) p_0(\eta, \theta) p_0(\nu) d\eta.$$

The test statistic of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is

$$\Lambda(\theta_0) = \log \sum_{\nu} p(\nu | x_1, \dots, x_\tau) - \log p(\theta_0 | x_1, \dots, x_\tau).$$

In the normal example with uniform priors for ν, η , and θ , the Bayesian test statistic is given by

$$\begin{aligned} \Lambda(\theta) &= \log \left\{ \sum_{\nu=1}^{\tau-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{i=1}^{\nu} \phi(x_i - \eta) \prod_{i=\nu+1}^{\tau} \phi(x_i - \theta) d\eta d\theta \right\} \\ &\quad - \log \left\{ \sum_{\nu=1}^{\tau-1} \int_{-\infty}^{\infty} \prod_{i=1}^{\nu} \phi(x_i - \eta) \prod_{i=\nu+1}^{\tau} \phi(x_i - \theta) d\eta \right\} \\ (3.4) \quad &= \log \left\{ \sum_{\nu=1}^{\tau-1} \frac{1}{\sqrt{\nu(\tau - \nu)}} \exp \left(\frac{s_{\nu}^2}{2\nu} + \frac{(s_{\tau} - s_{\nu})^2}{2(\tau - \nu)} \right) \right\} \\ &\quad - \log \left\{ \sum_{\nu=1}^{\tau-1} \frac{1}{\sqrt{\nu}} \exp \left(\frac{s_{\nu}^2}{2\nu} - \frac{(\tau - \nu)}{2} \theta^2 + (s_{\tau} - s_{\nu}) \theta \right) \right\} + \log \sqrt{2\pi}, \end{aligned}$$

where $\phi(x)$ is the standard normal density. This is similar to (3.3) except that the likelihood test statistic takes the ratio of the maximized likelihood functions over ν ,

while the Bayesian test statistic takes the ratio of the summed posterior functions over ν . The numerical comparison will be shown in the next simulation example.

3.2.4 Simulation Study

In the simulation study we find the confidence interval for the post-change mean θ , using the hybrid bootstrapping, based on the likelihood ratio test statistic (3.3) and the Bayesian test statistic (3.4). Besides the comparison of two intervals, the coverage probabilities of the hybrid confidence regions are computed for each test statistic and compared with a sequence of θ .

The original data were randomly generated from $x_i \sim N(0, 1)$ for $i \leq 10$ and $x_i \sim N(3, 1)$ for $i > 10$, with the stopping time

$$(3.5) \quad \tau = \inf \{n > 1 : |x_n - \bar{x}_n| > 3\sqrt{(n-1)/n}\} \wedge 100.$$

The largest sample size allowed is 100. For instance, in the first replication the sampling actually stopped at $\tau = 12$. Based on this sample $X = (x_1, \dots, x_{12})$, the maximum likelihood estimates of $\hat{\eta}(\hat{\nu})$, $\hat{\nu}$ and $\hat{\theta}(\hat{\nu})$ were computed and then for each θ , 500 independent samples of (x_1^*, x_2^*, \dots) were generated from the normal distribution with a mean of $\hat{\eta}(\hat{\nu}_\theta)$ for $i \leq \hat{\nu}_\theta$ and a mean of θ for $i > \hat{\nu}_\theta$. Finally, we estimated quantiles of the distributions for both test statistics of (3.3) and (3.4).

In the Figure 3.1 the solid lines denote the test statistic of the original data over a grid of θ and the dotted lines indicate the 95% quantiles of the test statistics based on the resampled values. The plot on the upper panel is for the log likelihood ratio test statistics and that on the lower panel is for the Bayesian test statistics. Then the confidence regions are given by the areas of θ at which $\{\Lambda(\theta) < q(\theta)\}$. It appears that both confidence intervals for θ are very similar.

However, we do notice one anomaly in the Figure 3.1. The quantiles of the likelihood ratio test statistics tend to increase as θ tends to $\eta = 0$, i.e., when there has been a small shift in the mean. This seems to occur because the change point ν is hard to estimate in this case, and maximization over the tenable values for ν increases the test statistic more than it would increase if there were more precise information about ν . Although this rise in quantiles is not a major problem in this case, the quantiles based on the Bayesian statistics seems more stable for θ near η . This is because the Bayesian test statistics tend to penalize the flexibility mentioned in a fairly natural fashion.

In the next simulation we compare the coverage probabilities of the hybrid bootstrap intervals for θ based on the likelihood ratio test statistics and the Bayesian test statistics. The data is randomly generated from $x_i \sim N(0, 1)$ for $i \leq 10$ and $x_i \sim N(\theta, 1)$ for $i > 10$, $\theta \in [0, 4]$ and the stopping time was followed by (3.5). For each θ , 1000 original data sets of $X_i = (x_1, x_2, \dots, x_\tau), i = 1, \dots, 1000$ were generated and 100 resamplings for each set were carried out to find the 95% quantiles of test statistics. In the Figure 3.2 the coverage probabilities of the confidence intervals over a grid of θ were shown based on the likelihood test statistics (solid line) and the Bayesian test statistics (dashed line). It seems that the coverage probabilities based on the Bayesian test statistics are much more stable than those for the likelihood test statistics as θ tends to η .

3.3 More Topics and Future Research

In this section we introduce more interesting topics about change point problems. First, Poisson process with rate change model is described. The basic framework of

Poisson model is similar to the normal shift model, but the change point in Poisson process is considered as a continuous time point. Secondly, time series model with change in variation is introduced. The model plays an important role in calculating value at risk of a financial position in risk management and in asset allocation. We are still working on these problems for the future research.

3.3.1 Poisson Process Change Point Problems

The task of detecting a change point in the number of daily defects in an industrial process or in the number of annual cases of a particular genetic disease may be considered in the context of a Poisson process. Raftery and Akman [26] consider Bayesian inference for a Poisson process with a single change point at an unknown time. Akman and Raftery [1] study asymptotic estimation for a Poisson process change point. West and Ogden [35] suggests a simplified grid search to find maximum likelihood estimates for the change point. All work above are based on off-line experiments without optional stopping to estimate the change point in a Poisson process. In this section, we consider on-line monitoring for a Poisson process with a possible abrupt change, seeking estimates of the rate of occurrence after the change.

Suppose that the data x_1, x_2, \dots , over unit time periods are sequentially observed until a stopping time τ . Here, the observation x_i represents the number of events that occurred in the i -th time period. A natural model for x_i is the Poisson distribution. The question of interest is whether there has been an abrupt change in the rate parameter defining the Poisson distribution over the τ periods. Let ν represent such a continuous time change point, η represent the Poisson rate parameter before the change, and θ represent the rate parameter after the change. To detect a change we continue using the Shewhart control chart, a natural monitoring procedure for the

Poisson Process. So, the stopping time for the Poisson process is

$$\tau = \inf\{n \geq 2 : |x_n - \bar{x}_n| \geq c\sqrt{\bar{x}}\}.$$

The basic scheme of this problem is similar to the normal mean shift model in the previous section, but differs in the type of the change point, that is, the change point is regarded as a continuous time point here. So, we have to consider three different distributions for individual observations.

Denote by $\lfloor x \rfloor$ the greatest integer function of x such that $x \geq \lfloor x \rfloor$, and $\langle x \rangle$ the fractional part of x , so $x = \lfloor x \rfloor + \langle x \rangle$. If a change occurs at ν , then the change occurs in the $(\lfloor \nu \rfloor + 1)$ -st interval. The observation for this period can be thought of as a sum of two independent Poisson random variables which are not observed directly.

Then,

$$x_i \sim \begin{cases} \text{Poisson}(\eta), & i = 1, \dots, \lfloor \nu \rfloor \\ \text{Poisson}(\langle \nu \rangle \eta + (1 - \langle \nu \rangle) \theta), & i = \lfloor \nu \rfloor + 1 \\ \text{Poisson}(\theta), & i = \lfloor \nu \rfloor + 2, \dots, \tau \end{cases}$$

If the process generating these observations is Poisson, these observations will be mutually independent. We also assume that the stopping is likely to occur after the change point and the change does not occur in either extreme periods, so $1 \leq \nu < \tau - 1$. The log likelihood function for η, ν , and θ is then given by

$$\begin{aligned} l(\theta, \nu, \eta) &= -\eta \lfloor \nu \rfloor - \theta(\tau - \lfloor \nu \rfloor - 1) + s_{\lfloor \nu \rfloor} \log \eta + (s_\tau - s_{\lfloor \nu \rfloor + 1}) \log \theta \\ &\quad - \{\langle \nu \rangle \eta + (1 - \langle \nu \rangle) \theta\} + x_{\lfloor \nu \rfloor + 1} \log \{\langle \nu \rangle \eta + (1 - \langle \nu \rangle) \theta\}, \end{aligned}$$

where $s_j = \sum_{i=1}^j x_i$. If we assume that ν is known and lies in $\nu \in [j, j + 1), j = 1, \dots, \tau - 2$, the maximum likelihood estimates for η and θ can be solved by differentiating $l(\theta, \nu, \eta)$ with respect to η and θ , respectively. They can be found by solving

the following two quadratic equations:

$$\hat{\eta}(\nu) = \frac{1}{\nu} \left[s_j + \frac{(\nu - j)\hat{\eta}(\nu)x_{j+1}}{(\nu - j)\hat{\eta}(\nu) + (1 - \nu + j)\hat{\theta}(\nu)} \right]$$

and

$$\hat{\theta}(\nu) = \frac{1}{\tau - \nu} \left[(s_\tau - s_{j+1}) + \frac{(1 - \nu + j)\hat{\theta}(\nu)x_{j+1}}{(\nu - j)\hat{\eta}(\nu) + (1 - \nu + j)\hat{\theta}(\nu)} \right],$$

but their explicit forms may not be obtained. The maximum likelihood estimate for ν is then

$$\hat{\nu} = \arg \sup_{\substack{j \leq \nu < j+1 \\ j \in \{1, \dots, \tau-2\}}} l(\hat{\theta}(\nu), \nu, \hat{\eta}(\nu)).$$

Since the profile likelihood function of ν is not differentiable with respect to ν due to the discontinuities of $\langle \nu \rangle$, the estimate $\hat{\nu}$ should be searched from $l(\hat{\theta}(\nu), \nu, \hat{\eta}(\nu))$ over a set of points of ν on the interval $[1, \tau - 1)$.

After we obtain $\hat{\nu}$ and plug it into the likelihood function, the log likelihood ratio test for fixed θ is given by

$$\Lambda(\theta) = l(\hat{\theta}(\hat{\nu}), \hat{\nu}, \hat{\eta}(\hat{\nu})) - \sup_{\substack{j \leq \nu < j+1 \\ j \in \{1, \dots, \tau-2\}}} l(\theta, \nu, \hat{\eta}(\nu)).$$

To simulate the distribution of $\Lambda(\theta)$, a double grid search is required varying both ν and θ , and this will increase explosively the burden of computation. Since the profile likelihood function of ν , $l(\hat{\theta}(\nu), \nu, \hat{\eta}(\nu))$ is piecewise smooth over each of the time interval, the problem can be considered one interval at a time. West and Ogden [35] suggest a search involving only one calculation for each of the $\tau - 2$ intervals.

If it is known if $\hat{\nu} \in (j^*, j^* + 1)$, then the maximum likelihood estimates solve

$$\frac{\partial}{\partial \hat{\eta}} l(\hat{\theta}, \hat{\nu}, \hat{\eta}) = 0, \quad \frac{\partial}{\partial \hat{\nu}} l(\hat{\theta}, \hat{\nu}, \hat{\eta}) = 0, \quad \text{and} \quad \frac{\partial}{\partial \hat{\theta}} l(\hat{\theta}, \hat{\nu}, \hat{\eta}) = 0,$$

for $\hat{\eta} = \hat{\eta}(\hat{\nu})$, $\hat{\nu}$, and $\hat{\theta} = \hat{\theta}(\hat{\nu})$, respectively. They are given by

$$\hat{\eta} = \frac{s_{j^*}}{j^*}, \quad \hat{\nu} = j^* + \frac{x_{j^*+1} - \hat{\theta}}{\hat{\eta} - \hat{\theta}}, \quad \text{and} \quad \hat{\theta} = \frac{s_\tau - s_{j^*+1}}{\tau - j^* - 1}.$$

However, there is no guarantee that $\hat{\nu}$ will in fact fall in the specified interval, so in that case, the endpoints of the interval should be examined. By approaching the problem in this manner, computation time to search $\hat{\nu}$ is significantly reduced. Since each estimate above is a function of $j = j^*$, $j = 1, \dots, \tau - 2$, and the maximized likelihood function $l(\hat{\theta}, \hat{\nu}, \hat{\eta})$ over all parameters also becomes a function of j , the problem concerns now a discrete change time point. Let us denote $\tilde{l}(j) = l(\hat{\theta}, \hat{\nu}, \hat{\eta})$, and \hat{j} maximizes $\tilde{l}(j)$ over $j \in \{1, \dots, \tau - 2\}$.

Similarly, suppose $\hat{\nu}_\theta$ solves

$$\frac{\partial}{\partial \hat{\eta}} l(\theta, \hat{\nu}, \hat{\eta}) = 0 \quad \text{and} \quad \frac{\partial}{\partial \hat{\nu}} l(\theta, \hat{\nu}, \hat{\eta}) = 0$$

for $\hat{\nu}$, assuming $\hat{\nu} \in [j, j + 1)$. Then, $\tilde{l}_\theta(j) = l(\theta, \hat{\nu}_\theta, \hat{\eta})$ is now a function of both θ and j , and \hat{j}_θ maximizes $\tilde{l}_\theta(j)$ over j for fixed θ . The log-likelihood ratio test for θ is then

$$\Lambda(\theta) = \tilde{l}(\hat{j}) - \tilde{l}_\theta(\hat{j}_\theta).$$

Finally, the hybrid bootstrapping is applied to find the confidence interval for θ .

For fixed θ we can generate

$$x_i \sim \begin{cases} \text{Poisson}(\hat{\eta}), & i = 1, \dots, \hat{j}_\theta \\ \text{Poisson}(\lambda(\theta)), & i = \hat{j}_\theta + 1 \\ \text{Poisson}(\theta), & i = \hat{j}_\theta + 2, \dots, \tau, \end{cases}$$

where $\lambda(\theta) = (\hat{\nu}_\theta - \hat{j}_\theta)\hat{\eta} + (1 - \hat{\nu}_\theta + \hat{j}_\theta)\theta$, and $\hat{\eta} = \frac{s_{\hat{j}}}{\hat{j}}$. The stopping times τ should depend on the generated samples each time. The confidence region for θ is now easily constructed as the same way in the normal example.

3.3.2 Change Detection for Variation in Economic Time Series

Detecting deviations from a supposed process is a problem which appears in many fields of sciences. Particularly, in finance a broker wants to detect trends in the course of a stock. Several authors show that control charts for independent variables, such as the Shewhart, the EWMA (exponentially weighted moving average), and the CUSUM scheme cannot be directly applied to time series data. It is necessary to take the structure of the time series into account. In recent years several control charts for time series has been introduced. One of them is a residual chart and it makes use of a transformation of the data. The aim is to derive statistics which are again independent variables since then it is possible to apply well-known methods to this quantities. Modified control scheme is another alternative for time series data and it is based on similar statistics as the classical procedures for independent samples.

One of the main processes is a GARCH (generalized autoregressive conditional heteroscedasticity) process in economic time series. The main property of this process is that their conditional variance is not constant. For that reason they are able to describe some behavior frequently observed in economics, for example, periods of large fluctuations alternating with relatively quiet phase. The control charts for GARCH processes were introduced by Schipper and Schmid [29]. They dealt with detecting changes in the variance. This problem is of great interest in practice since the variance measures the risk of the asset. For this reason it is a fundamental quantity for a portfolio manager. As the returns of an asset react very sensibly to new information, it is not surprising that changes in the variance of stock market returns can frequently be observed.

They distinguished between the target process, $\{Y_t\}$ and the observed process,

$\{X_t\}$. Suppose that sequentially data x_1, x_2, \dots are taken from a quantity of interest. The data are realizations of the observed process $\{X_t\}$. Each observation is examined to determine whether it can reasonably be explained by the distribution law of $\{Y_t\}$ or not. While the distribution of $\{X_t\}$ is unknown, the distribution of $\{Y_t\}$ is assumed to be known. For instance it can be obtained by fitting a suitable process to historical data for the characteristic. In most cases the target process is assumed to be a GARCH process.

A stochastic process $\{Y_t\}$ is called a GARCH(p, q) process if

$$Y_t = \sigma_t \epsilon_t,$$

with $\sigma_t > 0$ and

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^q \alpha_i Y_{t-i}^2 + \sum_{j=1}^p \beta_j \sigma_{t-j}^2$$

for $t \in \mathbb{Z}$. It is assumed that $\alpha_0 > 0, \alpha_i \geq 0, \beta_j \geq 0$ for all i, j . Moreover, the random variables $\{\epsilon_t\}$ are supposed to be independent with $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = 1$.

First, let $\{Y_t\}$ be an arbitrary process with mean μ_0 and variance γ_0 . The observed process $\{X_t\}$ is generated from the target process by a scale transform, more precisely

$$X_t = \begin{cases} Y_t & \text{for } 1 \leq t < \nu \\ \mu_0 + \Delta(Y_t - \mu_0) & \text{for } t \geq \nu \end{cases}$$

with $\Delta \geq 1$ and $\nu \in \mathbb{N}$. Thus, a change in the scale appears at position ν if $\Delta > 1$.

In this case the process $\{X_t\}$ is out-of-control. If $\Delta = 1$, it is called in-control. For the observed process, $E(X_t) = \mu_0, \text{Var}(X_t) = \gamma_0$ for $t < \nu$, but $\text{Var}(X_t) = \Delta^2 \gamma_0$ for $t \geq \nu$. When the target process $\{Y_t\}$ is a GARCH process, $\mu_0 = 0$ and $\gamma_0 = \sigma_t^2$.

The conditional variance of the underlying asset return σ_t^2 is also known as a volatility. It is an important factor in option trading and has many other financial applications. The volatility modeling provides a simple approach to calculating value

at risk of a financial position in risk management. It also plays an important role in asset allocation under the mean-variance framework. Furthermore, modeling the volatility of a time series can improve the efficiency in parameter estimation and the accuracy in interval forecast. However, the volatility is not directly observable from the return data.

Change detection for the scale Δ should be meaningful because it is associated with σ_t^2 . Since the control scheme stops the process as soon as possible once it detects deviation from the target process, the limited information about Δ is obtained. This should be another interesting example of the estimation of the post change parameter based on the hybrid bootstrapping. Since the underlying process is time series model, the likelihood approach is quite challenging.

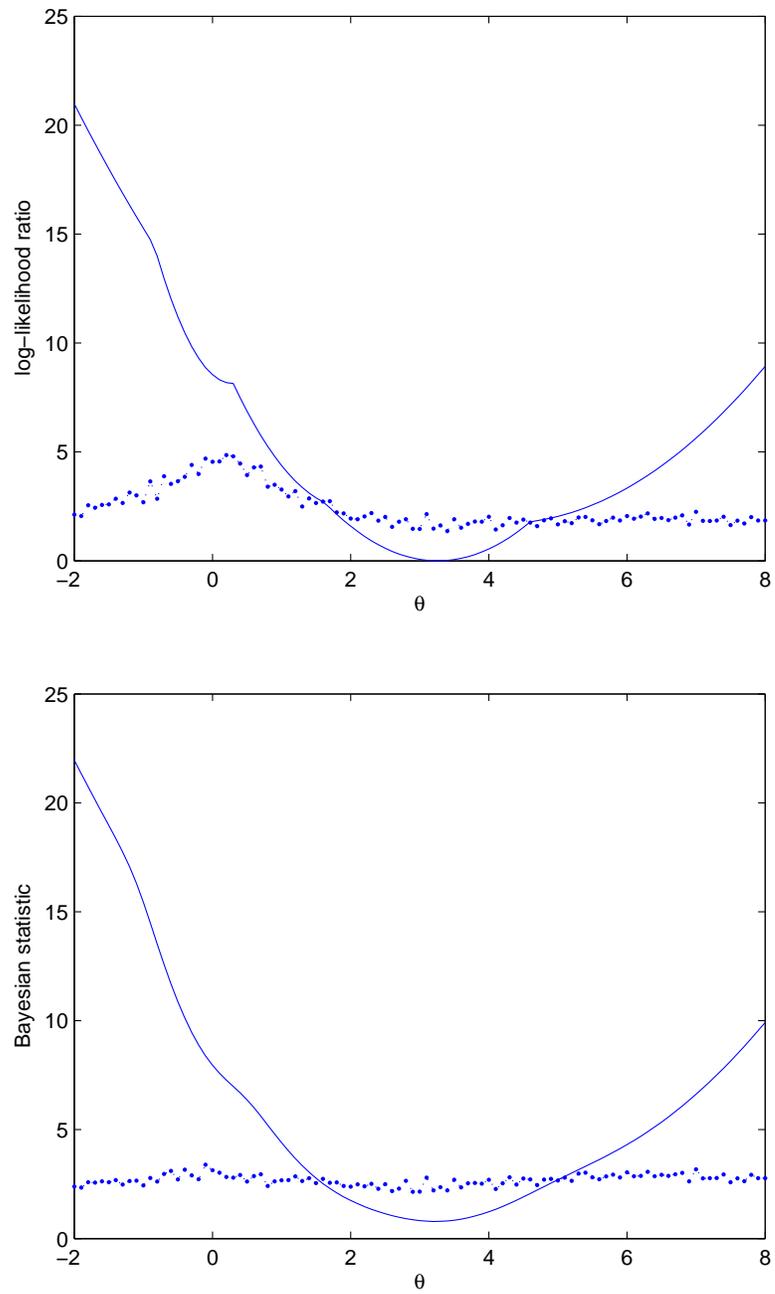


Figure 3.1: The confidence intervals for θ based on the likelihood ratio test statistics on the upper panel and the Bayesian test statistics on the lower panel.

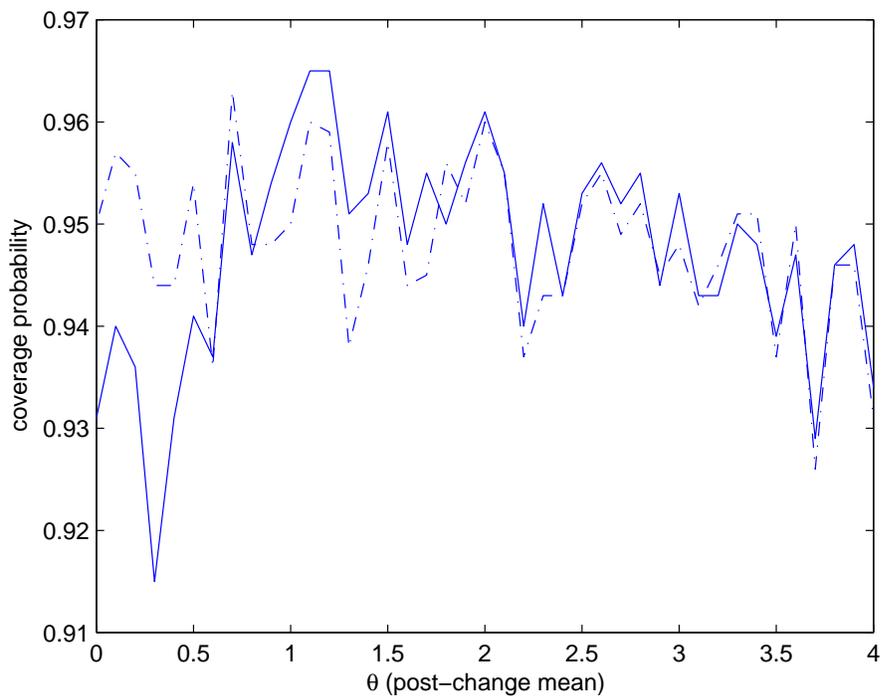


Figure 3.2: The coverage probabilities of the confidence interval for θ based on the likelihood ratio test statistics (solid line) and the Bayesian test statistics (dashed line).

CHAPTER IV

Inconsistent Hybrid Bootstrap Confidence Region

In this section we investigate whether the coverage probability of hybrid regions converges to the desired nominal value as the information of nuisance parameters increases. The problem is based on a signal plus noise model for Poisson data, of interest in high energy physics. We show that the coverage probability is not consistent in this example.

4.1 Model and Estimation

Suppose that $Z = (X, Y)$ where X and Y are independent, X has the Poisson distribution with mean $\theta + \eta$, and Y has the Poisson distribution with mean $\gamma\eta$. Here $\theta \geq 0$ and $\eta \geq 0$ are unknown but only θ is of interest. The scale factor γ is assumed to be known, and large values of γ are considered.

For notation, let

$$l_0(\theta, \eta) = X \log(\theta + \eta) - \theta - \eta - \log(X!),$$

the log likelihood from X , and

$$l_\gamma(\eta) = Y \log(\gamma\eta) - \gamma\eta - \log(Y!),$$

the log likelihood from Y . Then the total log likelihood is

$$l(\theta, \eta) = l_0(\theta, \eta) + l_\gamma(\eta).$$

The score functions for θ and η are

$$\frac{\partial l(\theta, \eta)}{\partial \theta} = \frac{X}{\theta + \eta} - 1, \quad \text{and} \quad \frac{\partial l(\theta, \eta)}{\partial \eta} = \frac{Y}{\eta} + \frac{X}{\theta + \eta} - (\gamma + 1).$$

Taking $\tilde{\eta} = Y/\gamma$, the maximum likelihood estimator for η based on Y , the maximum likelihood estimator for θ is

$$\hat{\theta} = (X - \tilde{\eta})_+,$$

and the maximum likelihood estimator for η is

$$\hat{\eta} = \begin{cases} \tilde{\eta}, & X \geq \tilde{\eta}; \\ \frac{X+Y}{1+\gamma}, & X \leq \tilde{\eta}. \end{cases}$$

Finally, let $\hat{\theta}_\eta = (X - \eta)_+$, the maximum likelihood estimator for θ when η is known, and let $\hat{\eta}_\theta$ denote the maximum likelihood estimator for η when θ is known, given explicitly as

$$(4.1) \quad \hat{\eta}_\theta = \frac{X + Y - \theta(1 + \gamma) + \sqrt{[X + Y - \theta(1 + \gamma)]^2 + 4\theta Y(1 + \gamma)}}{2(1 + \gamma)}$$

Thus, the log likelihood ratio test statistic for θ

$$(4.2) \quad \Lambda(\theta) = -X - Y + \theta + \hat{\eta}_\theta(1 + \gamma) + X \log \left(\frac{\hat{\eta} + \hat{\theta}}{\hat{\eta}_\theta + \theta} \right) + Y \log \left(\frac{\hat{\eta}}{\hat{\eta}_\theta} \right)$$

after some simple algebra.

The hybrid bootstrap confidence interval for θ can be computed over a grid of θ values. The method requires computing (4.2) for a single $\eta = \hat{\eta}_\theta$ over a grid of θ . In some cases this can be done by numerical integration. But it can always be done by simulation by following these steps:

1. Generate independent $(X_1^*, Y_1^*), (X_2^*, Y_2^*), \dots, (X_K^*, Y_K^*)$ (psuedo samples) from the distribution of $P_{\theta, \hat{\eta}_\theta}$ with $X_i^* \sim \text{Poisson}(\theta + \hat{\eta}_\theta)$ and $Y_i^* \sim \text{Poisson}(\gamma \hat{\eta}_\theta)$, $i = 1, 2, \dots, K$.
2. Compute $\Lambda_i(\theta)$ in (4.2) based on each generated (X_i^*, Y_i^*) .
3. Let $\hat{q}(\theta)$ be the smallest value of q for which

$$\frac{\#\{i \leq K : \Lambda_i(\theta) \leq q\}}{K} \geq 1 - \alpha,$$

so $\hat{q}(\theta)$ is a Monte Carlo Estimate for the upper α -th quantile of the test statistic under $P_{\theta, \hat{\eta}_\theta}$.

4. The hybrid confidence region is given by

$$(4.3) \quad \{\theta \geq 0 : \Lambda(\theta) \leq \hat{q}(\theta)\}.$$

Sen and Woodroffe [30] have shown good performance of the hybrid bootstrap confidence interval in (4.3) in their simulation work. If the method works reasonably, we should have

$$P_{\theta, \eta}[\Lambda(\theta) \leq \hat{q}(\theta)] \approx 1 - \alpha.$$

Chuang and Lai [7] argued both theoretically and by example that this should be true. However, the asymptotic coverage probability may not be $1 - \alpha$ if we let $\gamma \rightarrow \infty$. The next section will show the inconsistent coverage probability of this confidence region.

4.2 Inconsistent Coverage Probability

Equation (4.1) can be rewritten in terms of $\tilde{\eta} = Y/\gamma$, as

$$\hat{\eta}_\theta = \frac{1}{2}(\tilde{\eta} - \theta) + \frac{X - \tilde{\eta}}{2(1 + \gamma)} + \frac{1}{2} \sqrt{(\tilde{\eta} + \theta)^2 + \frac{2(\tilde{\eta} - \theta)X - 2\tilde{\eta}(\tilde{\eta} + \theta)}{1 + \gamma} + \frac{(X - \tilde{\eta})^2}{(1 + \gamma)^2}}.$$

From the second equation here and the equation above for $\hat{\eta}$, it is easy to see that

$$(4.4) \quad \hat{\eta}_\theta = \tilde{\eta} + O_p(1/\gamma) \quad \text{and} \quad \hat{\eta} = \tilde{\eta} + O_p(1/\gamma)$$

as $\gamma \rightarrow \infty$. By Taylor expansion about $\tilde{\eta}$, with η_* an intermediate value between $\hat{\eta}$ and $\tilde{\eta}$,

$$(4.5) \quad l_\gamma(\hat{\eta}) = l_\gamma(\tilde{\eta}) - \frac{\gamma\tilde{\eta}(\hat{\eta} - \tilde{\eta})^2}{2\eta_*^2} = l_\gamma(\tilde{\eta}) + O_p(1/\gamma).$$

Similarly,

$$(4.6) \quad l_\gamma(\hat{\eta}_\theta) = l_\gamma(\tilde{\eta}) + O_p(1/\gamma).$$

Since $\hat{\eta}_\theta \xrightarrow{p} \eta$ as $\gamma \rightarrow \infty$, using (4.5) and (4.6)

$$\begin{aligned} \Lambda(\theta) &\xrightarrow{p} \Lambda_0(\theta, \eta) \\ &\stackrel{\text{def}}{=} l_0(\hat{\theta}_\eta, \eta) - l_0(\theta, \eta) \\ &= X \log \left(\frac{(X - \eta)_+ + \eta}{\theta + \eta} \right) + \theta - (X - \eta)_+, \end{aligned}$$

the log likelihood ratio test statistic when the background η is known. Let $q_0(\theta, \eta)$ denote the upper α -th quantile for $\Lambda_0(\theta, \eta)$. Since $\Lambda_0(\theta, \eta)$ is a function of X , it is a discrete variable, and to be precise and avoid ambiguity we will take¹

$$q_0(\theta, \eta) = \sup \{x : P_{\theta, \eta}(\Lambda_0(\theta, \eta) \leq x) \leq 1 - \alpha\}.$$

¹The quantile $q(\theta, \eta)$ for $\Lambda(\theta)$ should also be defined similarly, although this feels less important since probabilities for the atoms of the distribution of $\Lambda(\theta)$ are all small.

Then $q_0(\theta, \eta)$ will be an atom for the $P_{\theta, \eta}$ distribution of $\Lambda_0(\theta, \eta)$, and since cumulative distribution functions are right continuous,

$$P_{\theta, \eta}(\Lambda_0(\theta, \eta) < q_0(\theta, \eta)) \leq 1 - \alpha$$

and

$$P_{\theta, \eta}(\Lambda_0(\theta, \eta) \leq q_0(\theta, \eta)) > 1 - \alpha.$$

Define

$$\Delta_+(\theta, \eta) = P_{\theta, \eta}(\Lambda_0(\theta, \eta) \leq q_0(\theta, \eta)) - (1 - \alpha) > 0,$$

$$\Delta_-(\theta, \eta) = 1 - \alpha - P_{\theta, \eta}(\Lambda_0(\theta, \eta) < q_0(\theta, \eta)) \geq 0,$$

and

$$\Delta(\theta, \eta) = \Delta_+(\theta, \eta) + \Delta_-(\theta, \eta) = P_{\theta, \eta}(\Lambda_0(\theta, \eta) = q_0(\theta, \eta)).$$

If $\Delta_-(\theta, \eta) \neq 0$, it is fairly easy to argue that $q(\theta, \eta)$ will converge to $q_0(\theta, \eta)$ as $\gamma \rightarrow \infty$. But a more careful analysis is necessary to approximate probabilities. For this, define $Z_\gamma = \sqrt{\gamma}(\tilde{\eta} - \eta)/\sqrt{\eta}$, so that

$$Z_\gamma \Rightarrow N(0, 1)$$

as $\gamma \rightarrow \infty$, by normal approximation for the Poisson distribution.

The main results below require two regularity assumptions:

$$A1: \Delta_-(\theta, \eta) > 0.$$

$$A2: \text{There is a unique constant } x^* \text{ so that } \Lambda_0(\theta, \eta) = q_0(\theta) \text{ if and only if } X = x^*.$$

A1 ensures that q_0 is continuous at (θ, η) , so that small changes for the values of these parameters will not have a large effect on this quantile. As a function of X , $\Lambda_0(\theta, \eta)$ is unimodal, achieving its minimum when $X = \theta + \eta$. So it is possible for two integral values for X to give the same value for $\Lambda_0(\theta, \eta)$, but the assertion in *A2* should be typical.

Lemma 4.2.1. *Define*

$$\sigma^2(\theta, \eta) = \eta \left[\frac{(\eta - x^*)_+}{\eta} + \frac{(x^* - \theta - \eta)}{\theta + \eta} \right]^2.$$

If assumptions A1 and A2 hold and if $\sigma(\theta, \eta) > 0$, then for any $c \in \mathbb{R}$, as $\gamma \rightarrow \infty$,

$$\begin{aligned} P_{\theta, \eta}[\Lambda(\theta) \leq q_0(\theta, \eta) + c/\sqrt{\gamma} + o(1/\sqrt{\gamma})] \\ \rightarrow 1 - \alpha - \Delta_-(\theta, \eta) + \Delta(\theta, \eta)\Phi(c/\sigma(\theta, \eta)) \end{aligned}$$

and

$$q(\theta, \eta) = q_0(\theta, \eta) + \frac{\sigma(\theta, \eta)}{\sqrt{\gamma}} \Phi^{-1} \left(\frac{\Delta_-(\theta, \eta)}{\Delta(\theta, \eta)} \right) + o(1/\sqrt{\gamma}),$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function Φ .

Also, the results here hold uniformly for η sufficiently close a value satisfying A1 and A2.

Proof. By (4.4)

$$l_0(\hat{\theta}, \hat{\eta}) - l_0(\theta, \hat{\eta}_\theta) = l_0[(X - \tilde{\eta})_+, \tilde{\eta}] - l_0(\theta, \tilde{\eta}) + O_p(1/\gamma).$$

Also, note that

$$(X - \tilde{\eta})_+ = (X - \eta)_+ - (\tilde{\eta} - \eta)I\{X > \eta\} + (\tilde{\eta} - \eta)_-I\{X = \eta\} + O_p(1/\gamma),$$

which follows because the equation holds exactly unless $\eta < X < \tilde{\eta}$ or $\tilde{\eta} < X < \eta$, and probabilities for these events tend to zero. By (4.5), (4.6) and the delta method (Taylor expansion),

$$\begin{aligned} (4.7) \quad \Lambda(\theta) &= l_0(\hat{\theta}, \hat{\eta}) - l_0(\theta, \hat{\eta}_\theta) + O_p(1/\gamma) \\ &= \Lambda_0(\theta, \eta) - \frac{\sqrt{\eta}}{\sqrt{\gamma}} \left[\frac{(\eta - X)_+}{\eta} + \frac{X - \theta - \eta}{\theta + \eta} \right] Z_\gamma + O_p(1/\gamma). \end{aligned}$$

The first assertion in the lemma follows from this representation using the next lemma, and the second assertion of the lemma follows easily from the first. Uniformity in η is evident from the method of proof. \square

Lemma 4.2.2. *Suppose $\tilde{Z}_\gamma \Rightarrow \tilde{Z}$ as $\gamma \rightarrow \infty$ and W is a discrete random variable with W and \tilde{Z}_γ asymptotically independent. If $P(g(W) = q) > 0$ and $g(W) = q$ if and only if $W = w^*$, then*

$$\begin{aligned} P[g(W) + h(W)\tilde{Z}_\gamma/\sqrt{\gamma} + o_p(1/\sqrt{\gamma}) \leq q + c/\sqrt{\gamma}] \\ \rightarrow P(g(W) < q) + P(g(W) = q)P(h(w^*)\tilde{Z} \leq c) \end{aligned}$$

as $\gamma \rightarrow \infty$, whenever c is a continuity point for the distribution of $h(w^*)\tilde{Z}$.

Proof. The o_p term can be incorporated into \tilde{Z}_γ , so we can assume it is zero. If $g(w) < q$,

$$\begin{aligned} P[g(W) + h(W)\tilde{Z}_\gamma/\sqrt{\gamma} > q + c/\sqrt{\gamma}, W = w] \\ \leq P[h(w)\tilde{Z}_\gamma - c > (q - g(w))\sqrt{\gamma}] \rightarrow 0, \end{aligned}$$

and so

$$P[g(W) + h(W)\tilde{Z}_\gamma/\sqrt{\gamma} \leq q + c/\sqrt{\gamma}, W = w] \rightarrow P(W = w).$$

Then

$$\begin{aligned} P[g(W) + h(W)\tilde{Z}_\gamma/\sqrt{\gamma} \leq q + c/\sqrt{\gamma}, g(W) < q] \\ = \sum_{w: g(w) < q} P[g(W) + h(W)\tilde{Z}_\gamma/\sqrt{\gamma} \leq q + c/\sqrt{\gamma}, W = w] \\ \rightarrow P(g(W) < q), \end{aligned}$$

by dominated convergence. Similarly,

$$P[g(W) + h(W)\tilde{Z}_\gamma/\sqrt{\gamma} \leq q + c/\sqrt{\gamma}, g(W) > q] \rightarrow 0.$$

Finally, by the asymptotic independence,

$$\begin{aligned} \limsup_{\gamma \rightarrow \infty} P[g(W) + h(W)\tilde{Z}_\gamma/\sqrt{\gamma} \leq q + c/\sqrt{\gamma}, g(W) = q] \\ = \limsup_{\gamma \rightarrow \infty} P[h(w^*)\tilde{Z}_\gamma \leq c, W = w^*] \\ \leq P(h(w^*)\tilde{Z} \leq c)P(W = w^*), \end{aligned}$$

since the set (\tilde{Z}_γ, W) lies in is closed; and for $\epsilon > 0$

$$\begin{aligned} & \liminf_{\gamma \rightarrow \infty} P[g(W) + h(W)\tilde{Z}_\gamma/\sqrt{\gamma} \leq q + c/\sqrt{\gamma}, g(W) = q] \\ & \geq \liminf_{\gamma \rightarrow \infty} P[h(w^*)\tilde{Z}_\gamma < \epsilon + c, |W - w^*| < \epsilon] \\ & \geq P(h(w^*)\tilde{Z} < \epsilon + c)P(|W - w^*| < \epsilon), \end{aligned}$$

since the set (\tilde{Z}_γ, W) lies in is open. Letting $\epsilon \downarrow 0$,

$$\begin{aligned} & P[g(W) + h(W)\tilde{Z}_\gamma/\sqrt{\gamma} \leq q + c/\sqrt{\gamma}, g(W) = q] \\ & \rightarrow P(h(w^*)\tilde{Z} \leq c)P(W = w^*), \end{aligned}$$

and the lemma follows. \square

To study coverage probabilities for hybrid bootstrap intervals, we will need the following lemma describing how $q_0(\theta, \eta)$ varies with η .

Lemma 4.2.3. *If A2 holds and $\Delta_-(\theta, \eta) > 0$, then*

$$D(\theta, \eta) \stackrel{\text{def}}{=} \frac{\partial q_0(\theta, \eta)}{\partial \eta} = \begin{cases} \frac{\theta + \eta - x^*}{\theta + \eta}, & x^* \geq \eta; \\ \frac{\theta x^*}{\eta(\theta + \eta)}, & x^* \leq \eta. \end{cases}$$

Proof. Since $\Delta_-(\theta, \eta) > 0$, the equation

$$q_0(\theta, \bar{\eta}) = x^* \log \left[\frac{(x^* - \bar{\eta})_+ + \bar{\eta}}{\theta + \bar{\eta}} \right] + \theta - (x^* - \bar{\eta})_+,$$

which holds automatically when $\bar{\eta} = \eta$, will also hold for $\bar{\eta}$ sufficiently close to η .

If $x^* \neq \eta$, the desired result then follows easily by ordinary calculus. If $x^* = \eta$, by

Taylor expansion as $\epsilon \rightarrow 0$,

$$\begin{aligned} q_0(\theta, \eta + \epsilon) &= x^* \log \left(\frac{\eta + \epsilon + \epsilon_+}{\theta + \eta + \epsilon} \right) + \theta - \epsilon_+ \\ &= q_0(\theta, \eta) + \frac{\epsilon\theta}{\theta + \eta} + o(\epsilon), \end{aligned}$$

and the desired result still holds. \square

The following theorem is our main result about coverage probabilities for hybrid bootstrap confidence regions in this Poisson example.

Theorem 4.2.4. *Assume A1 and A2 and define*

$$\tilde{\sigma}^2(\theta, \eta) = \eta \left[\frac{(\eta - x^*)_+}{\eta} + \frac{(x^* - \theta - \eta)}{\theta + \eta} + D(\theta, \eta) \right]^2.$$

If $\tilde{\sigma}(\theta, \eta) > 0$, then

$$\begin{aligned} P_{\theta, \eta}(\theta \in S) &= P_{\theta, \eta}[\Lambda(\theta) \leq q(\theta, \hat{\eta}_\theta)] \\ &\rightarrow 1 - \alpha - \Delta_-(\theta, \eta) + \Delta(\theta, \eta) \Phi \left(\frac{\sigma(\theta, \eta) \Phi^{-1}(\Delta_-(\theta, \eta)/\Delta(\theta, \eta))}{\tilde{\sigma}(\theta, \eta)} \right). \end{aligned}$$

The limit will only be $1 - \alpha$ if $\tilde{\sigma}(\theta, \eta) = \sigma(\theta, \eta)$ or $\Delta_-(\theta) = \Delta(\theta, \eta)/2$.

Proof. By Lemma 4.2.3, the equation (4.4), and the approximation in Lemma 4.2.1 for $q(\theta, \eta)$,

$$\begin{aligned} q(\theta, \hat{\eta}_\theta) &= q_0(\theta, \hat{\eta}_\theta) + \frac{\sigma(\theta, \eta)}{\sqrt{\gamma}} \Phi^{-1} \left(\frac{\Delta_-(\theta, \eta)}{\Delta(\theta, \eta)} \right) + o_p(1/\sqrt{\gamma}) \\ &= q_0(\theta, \eta) + \frac{1}{\sqrt{\gamma}} \left[D(\theta, \eta) \sqrt{\eta} Z_\gamma + \sigma(\theta, \eta) \Phi^{-1} \left(\frac{\Delta_-(\theta, \eta)}{\Delta(\theta, \eta)} \right) \right] \\ &\quad + o_p(1/\sqrt{\gamma}). \end{aligned}$$

Using this and the representation for $\Lambda(\theta)$ in (4.7), the coverage probability in the lemma can be written as

$$\begin{aligned} P_{\theta, \eta} \left[\Lambda_0(\theta, \eta) - \frac{\sqrt{\eta}}{\sqrt{\gamma}} \left(\frac{(\eta - X)_+}{\eta} + \frac{X - \theta - \eta}{\theta + \eta} + D(\theta, \eta) \right) Z_\gamma \right. \\ \left. \leq q_0(\theta, \eta) + \frac{\sigma(\theta, \eta)}{\sqrt{\gamma}} \Phi^{-1} \left(\frac{\Delta_-(\theta, \eta)}{\Delta(\theta, \eta)} \right) + o_p(1/\sqrt{\gamma}) \right] \end{aligned}$$

The stated result now follows by Lemma 4.2.2. \square

In this result, the reason consistency fails is related to the fact that quantiles $q_0(\theta, \eta)$ for $\Lambda_0(\theta, \eta)$ vary with η . In essence, the hybrid-bootstrap region estimates

$q(\theta, \eta)$ by $q(\theta, \hat{\eta}_\theta)$. If the estimation error could be $o_p(1/\sqrt{\gamma})$, then coverage probabilities would converge to $1 - \alpha$. But that is a bit better than what is possible. It is well known that ordinary bootstrapping can fail if the functionals of interest are not smooth. This seems similar—the quantiles of interest vary a bit too rapidly to be estimated with the necessary precision. When $X = x^*$, the hybrid regions uses $\tilde{\eta}$ in a natural way to decide whether to include a value θ . Regions that use $\tilde{\eta}$ naturally cannot have coverage probabilities converging properly. However, regions that use $\tilde{\eta}$ essentially for randomization (perhaps through a variable U_γ defined as the fractional part of $Y/\sqrt{\gamma}$) can have coverage probabilities converging to $1 - \alpha$.

CHAPTER V

Summary and Conclusion

This thesis surveys results involving hybrid bootstrap method and related statistical applications. The method is based on the model where the data have substantial information about the nuisance parameter, but limited information about the parameter of interest.

The first part was concerned with a genetics application. Particularly, we considered the mapping problem of quantitative trait loci in an experimental population. In this example the parameter of interest is the QTL location and the effects of QTL are regarded as nuisance parameters. We constructed the hybrid confidence region for the QTL location with phenotype measurements and marker genotypes from a backcross experiment.

The simulation studies have demonstrated that the hybrid quantile estimates are less variable than other quantile estimates, and the hybrid regions have almost exact $1 - \alpha$ coverage for each study, regardless of sample size and marker distance.

Even though the performance of hybrid regions is excellent in the simulation studies, compared with other methods such as permutation, non-parametric bootstrap, and parametric bootstrap, the method failed to detect a QTL in the data set of rice tiller numbers.

We believe that one of the main reasons for this failure is a lack of information necessary to resolve a small shift in nuisance parameters. The QTL effect of rice tiller numbers seems to be too weak to locate a QTL. The hybrid method considers all possible QTL locations of (θ, J) , and average of $|\hat{\eta}_0(\theta, J) - \hat{\eta}_1(\theta, J)|$ over (θ, J) is almost 0.

However, other methods, which are based only on unconstrained maximum likelihood estimates $(\hat{\theta}, \hat{J})$, identify a QTL effect and locate a QTL. But, since the difference $|\hat{\eta}_0(\theta, J) - \hat{\eta}_1(\theta, J)|$ is maximized when $(\theta, J) = (\hat{\theta}, \hat{J})$, the estimate of the magnitude of the QTL effect may well have a positive bias and could be spurious. So, their detection could be too optimistic with this sample size. This possibility deserves further research. Perhaps future simulation, studying detection error fraction when there are a small shift of or no QTL.

The second part covered change point problems in industrial application. In particular, we gave interval estimates for a post change parameter θ in sequentially observed data which is truncated by stopping times. In the experiment the change point was considered as a nuisance parameter, and we also assumed that the stopping time occurs after the distributional change does. A simulation study shows reasonable performance for the hybrid regions in a simple normal example. Since real industrial processes are more complicated patterns, extensions to more realistic models would be of interest.

The limiting coverage probability of hybrid confidence region was investigated in the last part. It is based on a signal plus noise Poisson model in high energy physics problems. Here we theoretically showed that the coverage probability is inconsistent as information about the nuisance parameter increases. Modification to reserve consistency essentially seems to be impossible due to the natural estimation error. This

problem is associated with discreteness of the likelihood ratio test statistic Λ_0 when the nuisance parameters are known. The magnitude of the limiting discrepancy is at most the size of an atom for the distribution, so in practice this may not be a big concern if Λ_0 takes on many values each with small probability.

Finally, future topics we could study were introduced at the end of each chapter. They include application to an intercross population in the QTL mapping problem, the model with multiple QTLs, Poisson change point problems, and financial time series with change in variation. Although we cannot directly apply our methods to these problems, they are all based on the basic models we have investigated through this thesis. In the future we hope we can resolve these problems.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] V. E. Akman and A. E. Raftery. Asymptotic inference for a change point poisson process. *Annals of Statistics*, 14:1583–1590, 1986.
- [2] Michele Basseville and Igor v. Nikiforov. *Detection of abrupt changes theory and application*. Prentice Hall, New Jersey, 1993.
- [3] Karl W. Broman and Terence P. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of Royal Statistical Society, B*, 64:641–656, 2002.
- [4] Jie Chen and A. K. Gupta. Testing and locating varicance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92:739–747, 1997.
- [5] Jie Chen and A. K. Gupta. Statistical inference of covariance change points in gaussian model. *Statistics*, 38:17–28, 2004.
- [6] Zehua Chen and Hanfeng Chen. On some statistical aspects of the interval mapping for qtl detection. *Statistica Sinica*, 15:909–925, 2005.
- [7] Chin-Shan Chuang and Tze Leung Lai. Hybrid resampling methods for confidence intervals. *Statistica Sinica*, 10:1–50, 2000.
- [8] G. A. Churchill and R.W. Doerge. Empirical threshold values for quantitative trait mapping. *the Genetics Society of America*, 138:963–971, 1994.
- [9] Yuehua Cui, Dong-Yun Kim, and Jun Zhu. On the generalized poisson regression mixture model for mapping quantitative trait loci with count data. *Genetics*, 174:2159–2172, 2006.
- [10] Weiping Deng, Hanfeng Chen, and Zhaohai Li. A logistic regression mixture model for interval mapping of genetic trait loci affecting binary phenotypes. *Genetics*, 172:1349–1358, 2006.
- [11] Josee Dupuis and David Siegmund. Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics*, 151:373–386, 1999.
- [12] Gary J. Feldman and Robert D. Cousins. A unified approach to the classical statistical analysis of small signals. *Physical Review D*, 57:3873–3889, 1998.
- [13] C. S. Haley and S. A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69:315–324, 1992.
- [14] C. S. Haley, S. A. Knott, and J. M. Elsen. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, 136:1195–1207, 1994.
- [15] Douglas M. Hawkins. Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, 72:180–186, 1977.
- [16] R. C. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135:205–211, 1993.

- [17] Changjian Jiang and Zhao-Bang Zeng. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica*, 101:47–58, 1997.
- [18] C. H. Kao, Z. B. Zeng, and R. D. Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 152:1203–1216, 1999.
- [19] Eric S. Lander and David Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121:185–199, 1989.
- [20] Eric S. Lander and Philip Green. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci.*, 84:2363–2367, 1987.
- [21] Clive R. Loader. A log-linear model for a poisson process change point. *Annals of Statistics*, 20:1391–1411, 1992.
- [22] G. Lorden. Procedure for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42:1897–1908, 1971.
- [23] A. Manichaikul, J. Dupuis, S. Sen, and K. W. Broman. Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. *Genetics*, 174:481–489, 2006.
- [24] Octavio Martinez and Robert N. Curnow. Missing markers when estimating quantitative trait loci using regression mapping. *Genetica*, 101:47–58, 1994.
- [25] Douglas C. Montgomery. *Introduction to Statistical Quality Control*. Wiley, New York, 1996.
- [26] A.E. Raftery and V.E. Akman. Bayesian analysis of a poisson process with a change-point. *Biometrika*, 73:85–89, 1986.
- [27] Byron P. Roe and Michael B. Woodroffe. Improved probability method for estimating signal in the presence of background. *Physical Review D*, 60:053009, 1999.
- [28] Byron P. Roe and Michael B. Woodroffe. Setting confidence belts. *Physical Review D*, 63:013009, 2000.
- [29] S. Schipper and W. Schmid. Control charts for garch processes. *Nonlinear Analysis*, 47:2049–2060, 2001.
- [30] Bodhisattva Sen and Michael B. Woodroffe. On the unified method with nuisance parameters. working paper, University of Michigan, Department of Statistics, 2006.
- [31] David Siegmund. Boundary crossing probabilities and statistical applications. *Annals of Statistics*, 14:361–404, 1986.
- [32] David Siegmund. Confidence sets in change-point problems. *International Statistical Review*, 56:31–48, 1988.
- [33] A.F.M. Smith. A bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62:407–416, 1975.
- [34] Peter M. Visscher, Robin Thompson, and Chris S. Haley. Confidence intervals in qtl mapping by bootstrapping. *the Genetics Society of America*, 143:1013–1020, 1996.
- [35] R.Webster West and Ogden R.Todd. Continuous-time estimation of a change-point in a poisson process. *Journal of Statistical Computation and Simulation*, 56:293–302, 1997.
- [36] K.J. Worsley. Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, 73:91–104, 1986.
- [37] Yanhong Wu. Inference for change-point and post-change mean with possible change in variance. *Sequential Analysis*, 24:279–302, 2005.

- [38] Yanhong Wu. Inference for post-change mean by a cusum procedure. *Journal of Statistical Planning and Inference*, 136:3625–3646, 2006.
- [39] J. Q. Yan, J. Zhu, C. X. He, M. Benmoussa, and P. Wu. Quantitative trait locus analysis for the development behavior of tiller number in rice. *Theor. Appl. Genet.*, 97:267–274, 1998.
- [40] Tonglin Zhang and Michael B. Woodroffe. Credible and confidence sets for restricted parameter. *Journal of Statistical Planning and Inference*, 115:479–490, 2003.