Stony Brook University

Penalization for Gaussian mixture model and its application

A Dissertation Presented

by

Ziqi Meng

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

May 2017

Stony Brook University

The Graduate School

Ziqi Meng

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Song Wu – Dissertation Advisor
Associate Professor, Department of Applied Mathematics and Statistics

Wei Zhu - Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Jie Yang
Assistant Professor, Department of Applied Mathematics and Statistics

Christine DeLorenzo
Associate Professor, Department of Psychiatry, Stony Brook University

This dissertation is accepted by the Graduate School

Charles Taber

Dean of the Graduate School

Abstract of the Dissertation

Penalization for Gaussian mixture model and its application

by

Ziqi Meng

Doctor of Philosophy

in

Department of Applied Mathematics and Statistics

Stony Brook University

2017

Interval and linkage mapping are currently the most popular approaches for identifying quantitative trait loci (QTL). If phenotypic traits of interest are continuous, they are often assumed to follow a Gaussian mixture model. In this way, standard ML approach and LR test can be used to find the estimates of parameters and the position of a QTL. However, the assumption of homogeneity of variance across different genotypic groups can be violated in real data, and heterogeneous variances may lead to the unbounded likelihood function. Under this circumstance, the ML approach cannot be applied appropriately as the global ML estimate always fails to exist.

In order to solve this problem, we derived a suitable penalty function to regularize the likelihood function. It allows heterogeneous variances in the Gaussian mixture model. We applied this penalized method to both interval mapping and Linkage disequilibrium mapping and tested the presence of single QTL on a genome. We performed extensive simulation studies to compare the penalized method with standard ML approach on the power of detecting the existence of QTL

and the accuracy of estimated parameters for the Gaussian mixture model under different scenarios. We find that the penalized method is preferred to the ML when the true model has heterogeneous variance and the sample size is small.

We also applied the penalized and standard ML methods to a real data set with 96 markers genotyped for 502 F2 mice. The results suggest that both penalized and ML method were able to detect one genome-wide significant QTL. However, the penalized method offered additional information on the possible existence of QTLs at chromosome level.

Table of Contents

List of Figures

List of Tables

Acknowledgment

First of all, I would like to thank my PhD advisor, Professor Song Wu, who took me in his group at the second year of my PhD study. Prof. Wu was very enlightening and patient during this time and with his guide, I acquired deeper understanding of my research topic and improve my research ability. In addition, I did appreciate Prof. Wu's respect for students' own will and value on the development of the all-around ability of a PhD, so that I was able to have such an enriched experience during my these years. I was very grateful and honored to be one of his team members.

Secondly, I would like to thank my other committee members, Prof. Wei Zhu, Prof. Jie Yang and Prof. Christine DeLorenzo who kindly helped me through my PhD study and read through my dissertation with constructive comments and helps. Especially Professor Yang, who was my supervisor when I was working in the BCC. She helped me to apply all my statistical knowledge to solve real problems and taught me how to be a qualified consultant. These professional skills help me find the satisfactory job before my graduation and I am sure that in future, it will influence my career positively as well.

I would also like to express my thanks to my parents Xianguo Meng and Yiling Liu, who were very supportive, financially and spiritually with my PhD study, especially during the hardest time when I was looking for a job and preparing my thesis. Thanks for listening to my complaints and comforting me all the time and I would not be able to finish all this without their support.

In the end, I want to thank all my friends in Stony Brook who accompany me during these years and bring me a lot of good memories that will remain in my life.

Chapter 1 Introduction

1.1 Quantitative Traits Loci

Quantitative traits are phenotypes that demonstrate continuous variation within or among populations. Examples of the quantitative trait are crop yield, blood pressure, resistance to certain diseases, life span of mice or milk production in animals. Variation in such quantitative traits is associated not only with the environment they were exposed to, but also with the genetic loci they carry. Knowledge on genetic basis of this variation for quantitative traits is critical for addressing many important questions, from increasing the rate of selective improvement of important species in agriculture, to developing new and more personalized interventions to human beings and animals. Quantitative trait loci (QTLs) are stretches of DNA containing or linked to the genes that underlie a quantitative trait. Traits, such as grain yield, reproductive behavior and cancer, which are important in economics, biology and clinics are controlled by a series of QTL.

One of the main goals of QTL studies is to find out whether the variation in a phenotypic trait is controlled by any genetic loci, and if so, what are the locations of these loci. It is also interesting to know the phenotypic variation is attributed to a few loci each with relatively large effects, or to many loci, each with small effects. It is also possible that a few loci explain a substantial proportion of the difference in many quantitative traits while the remaining is due to many loci of minor effects [1] [2] [3]. For example, in domesticated rice[4], studies of flowering time have identified six QTL, which explains 84% of the variation in this trait [5]. This information is useful for applications such as breeding for better rice lineages.

1.2 Genetic markers

In general, QTL is unobservable. What can be observed are genetic markers that consist of genes or DNA sequences with known locations on chromosomes. Traditional genetic markers include single nucleotide polymorphisms (SNPs), simple sequence repeats (SSRs, or microsatellites), restriction fragment length polymorphisms (RFLPs), and transposable element positions [6-9]. Usually genetic markers themselves do not have direct causal effects on the phenotypes of interest. However, since they occupy positions that near or linked to the unknown QTL, they are indirectly associated with the phenotypes, which forms the basis of most genetic mapping methods.

One example is the hyper data from Hara (2001) [10]. Mice were given water containing 1% NaCl for two weeks. The phenotype is blood pressure (actually the average of 20 blood pressure measurements from 5 days). The blood pressure is plotted against the genotype at two genetic markers D4Mit214 and D12Mit20, as shown in Figure 1-1. For D4Mit214, the homozygous individuals exhibit a larger average phenotype than the heterozygotes, indicating that this marker is linked to a QTL that may affect the blood pressure. For D12Mit20, on the other hand, the two genotype groups show similar phenotypes, and therefore D12Mit20 is not likely associated with a QTL.

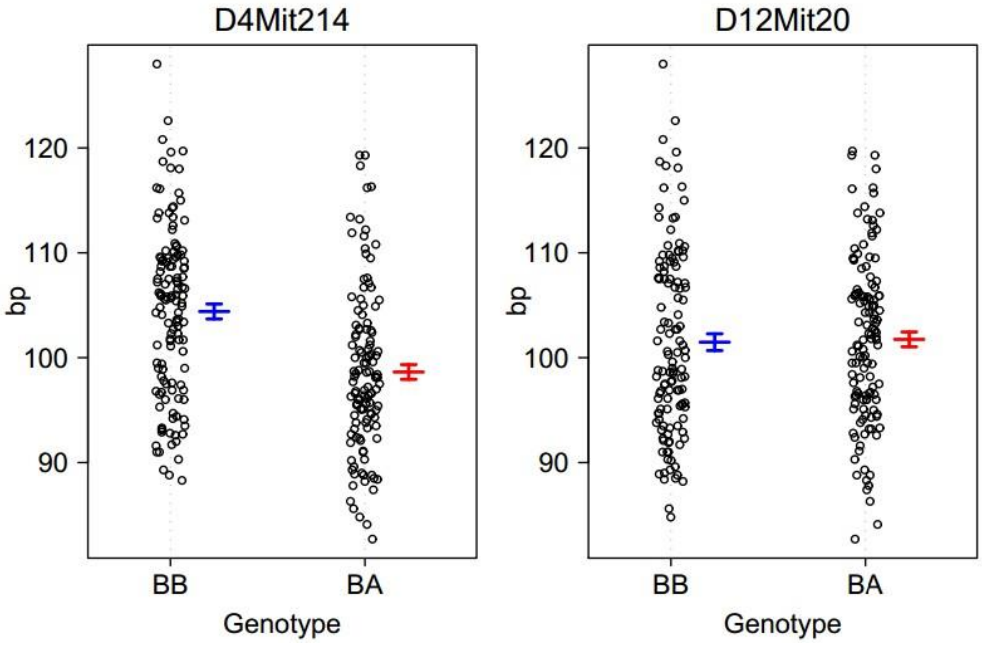


Figure 1-1: Blood pressure against the genotypes at two selected markers. [11]

1.3 Overview of the QTL mapping

The method of obtaining knowledge of the number, locations and effects of these genetic loci is broadly referred as quantitative trait loci (QTL) mapping. It studies the association between two types of information — phenotypic data (trait measurements) and genotypic data (usually molecular markers) — to explain the genetic basis of variation in complex traits. The basic question of this process is how we can efficiently and effectively determine the association between a quantitative trait and its corresponding QTLs, and subsequently find their locations and the genetic effects through QTL-linked genetic markers.

Depending on the biological nature of the organism and traits studied, either controlled populations, such as mice from experimental crosses, or populations arising naturally, such as human, can be used to map the QTL of interest. There are important distinctions for QTL mapping in controlled or natural populations. In natural populations, gene components are mainly separated

through random mating. For example, if we want to understand hypertension, we may be able to study genetic associations with hypertension in a large cohort such as the Nurses Health Study [12]. In addition, due to the different environments each subject was exposed to, the phenotypic characterization in natural populations typically have large noise. QTL mapping in experimental crosses provides an excellent alternative. In experimental crosses, individual gene components, including QTLs, are separated in a controlled manner, which allows us to magnify the genetic effects of a QTL. We can also perform phenotypes that may be impractical in humans (such as examining the liver in mice on a high-fat diet). We are able to control the environment of the subjects as well. For example, we may feed all mice the same diet, and keep the mouse rooms at the same temperature, and measure the phenotype of the mice at the same day in identical experimental conditions.

Several types of mapping populations generated from different experimental crosses can be constructed to map the QTL of interest. Among those, backcross and F2 intercross are probably two of the most widely used techniques and have been applied in many areas, such as maize and mice studies[13-15].

1.4 Basic QTL mapping methods

Mapping QTL on the genome is of great scientific importance and economical values for plant and animal breeding as well as for medical research. The discoveries from the detection and localization of QTLs may be used for genetic modification of genes that are important in breeding programs, for development of efficient vaccines etc. The last 25 years or so has witnessed a rapid advancement of statistical methods for mapping QTL in experimental organisms.

Statistical methods for QTL mapping in experimental organisms started with the naive single marker locus analysis [16]. It uses t-statistics or ANOVA to test the equality of the trait means in different marker genotype classes. For each genetic marker, the progenies are splitted into groups based on their genotypes. The mean square calculated from the difference among the different genotypes reflects the degree to which the maker is associated with the QTL affecting the trait of interest, and the mean square from the difference within the genotype groups are the residual variance. The ratio of these two is the F test statistic which will be compared with the threshold value obtained from the theoretical F distribution to determine if the marker is linked to a QTL for a particular trait. The marker genotype analysis was later extended to regression models with multiple markers. The measurement of trait are regressed on the genotypes of the multiple markers[17] [18].

As mentioned above, markers are the genes that linked to QTL which affects the trait of interest. However, in these early methods, marker loci are actually treated as QTL, which is not sensible. More advanced QTL mapping methods have then been developed after taking consideration of this problem. Weller (1986) [19] considered mixture models with a single marker locus. Weller (1987) [20] and Lander and Botstein (1989) [21] considered mixture models with two marker loci flanking a putative QTL. With the availability of maps of molecular markers covering the whole genome, Lander and Botstein (1989)[21] proposed a single interval mapping approach. Later on, a method combining single interval mapping approach and linear regression was then introduced by Jansen (1993) [22], Jansen and Stam (1994) [23] and Zeng (1993, 1994) [24, 25]. After this, Kao et al. (1999) [26] and Kao and Zeng (2002) [27] developed a more complicated approach — the multiple interval mapping method which used mixture models that

consider the effects of multiple QTL simultaneously and was more powerful in detecting multiple QTLs.

Basically, these statistical methods can be classified into “single marker method”, “Flanking marker methods” and “multiple marker methods” based on the number of markers used in each method [28, 29]. Meanwhile, these methods can also be grouped as “least square methods”, “regression methods”, “maximum likelihood methods”, and “mixed linear model approach methods”, etc, as the statistical techniques employed by them. In summary, these methods vary from simple to complicated, from detecting QTL-marker association to locating QTLs position and estimation their effects, with their own advantages and limitations. In this study, we will mainly applied our penalized models in interval mapping and linkage disequilibrium mapping.

Chapter 2 Finite mixture of Gaussian regressions model

2.1 Mixture model

The past few decades have witnessed tremendous statistical methodological development in QTL mapping, such as analysis of variance, interval mapping, multiple interval mapping [30-34]. Usually, one important assumption of these statistical methods is that the phenotypic values of a trait followed a known parametric distribution, such as a normal distribution. By estimating the parameters and comparing the phenotypic distributions under each genotype of QTL, the existence of a QTL and its genetic effects can be inferred.

Mixture model is one of the most important methodologies in QTL mapping. The phenotypic values corresponding to specific QTL genotypes can be modeled by known parametric distributions, e.g. normal for continuous and binomial for binary traits. If we can calculate the conditional probability of the QTL genotypes given a marker's genotypes, we will be able to develop a mixture model. Statistical approaches for parameter estimation with the mixture model are typically derived using maximum likelihood (ML) method because of many good properties of a ML estimator, such as asymptotical unbiasedness and asymptotical efficiency.

Mixture models have been used for more than one hundred year and its applications have emerged in many areas. The first attempt to analyze mixture models are often attributed to Pearson (1894) [35]. Since then, mixture models have been applied in a wide range of applications. For example, independently and identically distributed (i.i.d.) mixture models can well fit problems in signal and image processing. An example of applying mixture models in biological (plant morphology measures) and physiological (EEG signal) data modeling is presented by Roberts et al. (1998) [36]. In the field of geophysical data processing, the work by Kormylo and Mendel

(1982) [37] has introduced a Bernoulli-Gaussian description for sparse spike trains, i.e. a particular case of a two class Gaussian mixture model. McLachlan and Basford (1987) [38] highlighted the important role of mixture models in the field of cluster analysis and Biernacki et al. (1997) [39] proposed a model selection criterion applied to multivariate real data sets. Markovian mixture models are also commonly used, as in Ridolfi (1997) [40] or Idier (2001) [41] , where an application to medical image segmentation is considered.

Compared to the non-mixture models, more intensive computing is usually required by mixture models. Thanks to the more and more powerful computing technologies, the application of mixture models has tremendously increased in the last decades.

2.2 Gaussian mixture regressions model

Finite mixture models have been used for model-based clustering, in which a convex combination of a finite number of different distributions have been used to represent cluster features. For example, for a mixture of K univariate normal densities, it is defined as:

$$h(y;\theta) = \sum_{k=1}^{K} \pi_k f_k(y;u_k,\sigma_k) \tag{0-1}$$

where

$$f_k(y;u_k,\sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{(y-u_k)^2}{2\sigma_k^2}\right) \tag{2-2}$$

The parameters $\theta = (\pi, u, \sigma) = (\pi_1 \ldots . \pi_K, u_1 \ldots . u_K, \sigma_1 \ldots \sigma_K)$ are the mixture parameters, belonging to the parameter space.

$$\Theta = \{\theta = (\pi, u, \sigma) \mid 0 \le \pi_k \le 1, \sum_{k=1}^{K} \pi_k = 1, u_k \in R, \sigma_k > 0, k = 1,....,K\}$$

The data $y_1, y_2 \ldots\ldots y_n$ are assumed to be i.i.d samples. From a clustering point of view, we can say that each observed quantity $y_n, n = 1,2 \ldots N$ has been sampled from one of the K Gaussian distributions, according to the proportions $\pi_1, \pi_2 \ldots\ldots \pi_K$, i.e. each y_n belongs to one of K classes.

For the problem of QTL mapping, if we assume *n* independent subjects, each with a measurement of trait (y) and p marker information as showed in the following table:

Subject	Phenotype	Marker				
		m1	m2	m3	..	mp
1	$y_1 = 23.0$	0	1	2		2
2	$y_2 = 11.2$	2	2	1		2
3	$y_3 = 15.7$	2	1	0		0
…		…	…			
n	$y_n = 27.4$	1	0	0		1

The likelihood of the measurements of trait (y) and the marker (M) of the underlying QTL is constructed through a mixture model, expressed as:

$$\mathrm{L(Y,M)} = \prod_{i=1}^{n} \sum_{j=0}^{2} w_{j|i_k} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - \mu_j)^2}{2\sigma^2}))$$

(2-3)

Where $w_{j|i_k}$ is the conditional probability of the *i*th subject to carry QTL genotype j, given that its marker type. μ_j is the mean for the *j*th QTL genotypic subgroup. We also assume a common variance σ^2 for all the subgroups.

To obtain the maximum likelihood estimates (MLEs) of the unknown parameters, it can be achieved by solving the likelihood equations by differentiating the log-likelihood with respect to each parameter and setting the derivatives to zero, *i.e*,

$$\frac{\partial}{\partial \beta} \log L(Y, M) = \sum_{n=1}^{n} \sum_{j=0}^{2} \frac{w_{j|i_k} f_j(y_i)}{\sum_{j=0}^{2} w_{j|i_k} f_j(y_i)} = 0 \tag{2-4}$$

It is usually very difficult to solve these likelihood equations in an explicit form. Hence, we implement an EM algorithm, which has been shown to be very efficient for the parameter estimation for problems with a mixture of densities.

2.3 EM algorithm for Gaussian mixture model

Maximization of the log-likelihood of a mixture density is often done using the traditional EM algorithm proposed by Dedmpster et al. [42].

To maximize the likelihood function in (2-3), we first augment the observed data (y) with $Z_i's$. Z = $(Z_1, Z_2 \ldots . Z_n)$ are the latent variables (Z_i = 0,1,2) that determine the component from which the observation comes, so that the complete likelihood for (Y, Z) is then:

$$\mathrm{L(Y, Z)} = \prod_{i=1}^{n} \sum_{j=0}^{2} I(Z_i = j) w_{j|i_k} \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x_i - u_j)^2}{2\sigma^2}\right),$$

and the complete log-likelihood:

$$\log L(Y, Z) = \sum_{i=1}^{n} \sum_{j=0}^{2} I(Z_i = j) \cdot \left[\log w_{j|i_k} - \frac{1}{2} \log 2\pi - \log \sigma - \frac{(y_i - \mu_j)^2}{2\sigma^2}\right] \tag{2-5}$$

The detailed EM algorithm is given as follows:

E-step: it calculates the expected value of the log likelihood function, with respect to the conditional distribution of latent variable Z, given the observed data y under the current estimate of the parameters $\boldsymbol{\theta}^{(t)}$:

$$E_{Z|Y}[\log L(Y,Z) = \sum_{i=1}^{n}\sum_{j=0}^{2} T_{j,i_k}^{(t)} \cdot [\log w_{j|i_k} - \frac{1}{2}\log 2\pi - \log\sigma - \frac{(y_i-\mu_j)^2}{2\sigma^2}]$$

(2-6)

Where $T_{j|i_k}$ is the conditional distribution of the latent variable Z given the observed data and current estimate of $\theta^{(t)}$:

$$T_{j,i_k} = P\left(Z_i = j \middle| Y_i = y_i; \theta^{(t)}\right) = \frac{w_{j|i_k} f_j(y_i)}{\sum_{l=0}^{2} w_{l|i_k} f_l(y_i)} \quad (2\text{-}7)$$

M-step: since the conditional distribution of the latent variable $w_{j|i_k}$ is known given the markers information, the only thing we need to maximize is (μ, σ) in (2-6).

Taking the derivative of (2-6) with respect of μ and σ, setting them to zeros and solving the equations yield the updated estimates:

$$\mu_j^{(t+1)} = \frac{\sum_i^n T_{j,i_k} y_i}{\sum_i^n T_{j,i_k}}$$

(2-8)

And

$$\sigma_j^{(t+1)} = \sqrt{\frac{\sum_{i=1}^n T_{j,i_k}(y_i - u_j)^2}{\sum_{i=1}^n T_{j,i_k}}}$$

(2-9)

The algorithm iterates between the 2 steps until:

$$E_{Z|Y}[\log L\left(\theta^{(t)}; Y, Z\right) - E_{Z|Y}[\log L\left(\theta^{(t-1)}; Y, Z\right) \leq \epsilon$$

where ϵ is some a pre-set threshold.

The general convergence of the EM-type algorithms was discussed by Alfred [43].

2.4 Main goal of our study

In the previous framework about Gaussian mixture model in QTL mapping, we assume that each subgroup has a common variance because in this way the maximum likelihood estimate exists as the global maximization of the likelihood function [44]. However, this is a rather strict constraint and may be violated in real data. In the case of heteroscedastic variance, the likelihood function is always unbounded [45] and global ML estimates fail to exist, and therefore the EM algorithm would diverge toward a degenerated solution.

Intuitively, the degeneracy happens because in the sum of Gaussian densities, the variance parameter appears in the denominator. When a sample point y_k happens to equal to one of the means u_i, and at the same time, the corresponding σ_i goes to 0, the likelihood value goes to infinity. Indeed, points such ($\sigma_i^2 = 0, \mu_i = y_k$) yield singularities.

From a theoretical point of view, as stated by McLachlan and Peel (2000) [44], the non-existence of a global maximizer of the likelihood function does not exclude the ML approach, since its essential aim is to find a sequence of (local) maximize that is consistent [46]. Studies such as Peters and Walker (1978) [47], Kiefer (1978) [48], Redner (1981) [49] and Redner and Walker (1984) [50] focused on local ML estimation and mathematically investigated the existence of a consistent sequence of local maximizers. Unfortunately, in practice, it is hard to conceive a local maximization technique that could avoid global maxima. Actually, all the existent optimization techniques, including the very popular EM algorithm, are likely to converge to degenerated global solutions, depending on the initialization point. This is a severe drawback for the Gaussian mixture model with heterogeneous variances.

Hathaway (1985) [51] proposed a constrained formulation of the ML approach, which is based on the conditions

$$\forall\, k, k' \in \{1, \ldots . K\} \quad \sigma_k / \sigma'_k \geq c > 0$$

Where c is a constant to be chosen a priori. Moreover, Hathaway proves that this estimator is strongly consistent over the constrained parameter space. The numerical constrained maximization of the likelihood function is performed by a constrained EM algorithm [52], which, for sake of numerical robustness, implements an additional condition

$$\forall\, k \in \{1, \ldots . K\} \quad \pi_k \geq \varepsilon > 0$$

Where ε is another constant to be chosen a priori.

2.5 The Penalized Method

To avoid the degeneracy in the QTL mapping due to heteroscedastic variances, we propose a solution by adding a penalized term to the likelihood function that penalizes small variances. The penalized likelihood function therefore stay finite whenever σ_k goes to zero so that heterogeneous variances in the Gaussian mixture model are allowed and we will be able to obtain a penalized maximum estimates.

We penalize the mixture likelihood with a term p(σ) so that the penalized likelihood function can be expressed as:

$$\mathrm{L}(\mathrm{y}, \mathrm{M}; \theta) = L(y, M; \theta) p(\sigma)$$

The penalty term $p(\sigma)$ is adjusted to make the penalized likelihood bounded.

We choose $p(\sigma) = e^{-\lambda \sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k}}$ to be the penalty term. The reason for choosing such penalty term for two reasons: (1) it solves the degeneracy problem and guarantees the existence of maximum likelihood estimators, and (2) it also weights over each subgroup and automatically balances the weight in the sparse groups. The properties of the penalized likelihood and penalized maximum likelihood estimators will be showed by the following proposition.

***Proposition 1*:** Suppose the penalty term is $p(\sigma) = e^{-\lambda \sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k}}$, the penalized likelihood is bounded over Θ and penalized maximum likelihood estimators exist.

Proof:

Since $\exp(\frac{-(x-\mu_k)^2}{2\sigma_k^2}) \le 1$, we can have:

$$
\begin{aligned}
h_{penalized}(y;\mu,\sigma) &= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right) \exp\left(-\lambda \sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k}\right) \\
&\le \left(\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}}\right)^n \exp\left(-\lambda \sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k}\right) \\
&\le \left(\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_j^2}}\right)^n \exp\left(-\lambda \sum_{k=1}^{K} \pi_k \frac{1}{\sigma_k}\right) \qquad \sigma_j = \min_k \sigma_k \\
&= (2\pi)^{-\frac{n}{2}} \sigma_j^{-n} e^{-\lambda \pi_j \frac{1}{\sigma_j}} e^{-\lambda \sum_{k \ne j}^{K} \pi_k \frac{1}{\sigma_k}}
\end{aligned}
$$

$$
\lim_{\sigma_j \to 0^+} \sigma_j^{-n} e^{-\lambda \pi_j \frac{1}{\sigma_j}} = \lim_{\sigma_j \to 0^+} \frac{\left(\frac{1}{\sigma_j}\right)^n}{e^{\lambda \pi_j \frac{1}{\sigma_j}}} = 0
$$

And $(2\pi)^{-\frac{n}{2}} e^{-\lambda \sum_{k \neq j}^{K} \pi_k \frac{1}{\sigma_k}}$ is bounded over Θ.

This shows that the penalized likelihood goes to 0 as one of the σ tends to 0 and it can handle the problem of degeneracy caused by the heterogeneous variances from the different genotypic groups.

Chapter 3 Interval mapping

Interval mapping, as first introduced by Lander and Botstein in 1989 [21], is one of the most influential statistical models used to determine the positions of QTLs. It is an extension of one marker analysis. The term 'interval mapping' is used because it uses two flanking markers to construct an interval for searching a putative QTL within the interval. The idea of viewing QTL genotypes as missing data leads to the use of a mixture model for maximum likelihood analysis as mentioned in the previous chapter. To better understand the interval mapping, we need to introduce a few other concepts: the fraction of recombination rate and map function.

3.1 Fraction of recombination

Suppose the parental genotype is AB|ab for two loci on the genome, and the two homologous chromosomes lie side by side as showed in Figure 3-1 (A). Each of the paired chromosomes is then duplicated to generate 2 sister chromatids connected to each other at a region called the centromere. The homologous chromosomes form pairs and the four chromatids in Figure 3-1 (B) is known as a tetrad. In Figure 3-1 (C) the non-sister chromatids adhere to each other where crossing over can occur and the regions it occur we call them chiasmata. Chiasmata do not happen entirely at random since they are more likely to be further away from the centromere. The chiasmata are then divided into four gametes, each of which corresponds to one chromatid from a tetrad to make up the haploid complement (Figure 3-1 (D)). If no crossover occurs, then the chromosome just completely replicates the parental chromosome which means the gametes must be AB or ab. However, when there is one exchange between loci A and B, new combinations occurs as Ab and aB, which are referred to be the recombinant types. Genes on the same chromosome do not sort independently and particularly, genes on the same chromosome tend to

stay together. The extent to which loci remain together depends on their physical closeness, and this property of co-inheritance is called as linkage. If loci A and B are close, i.e. tight linkage, the probability of crossing over is usually small and therefore, gametes would carry more AB and ab and fewer Ab and aB. The recombination fraction (r) is defined as the proportion of recombinant gametes, and 1-r corresponds to the proportion of parental type AB and a. For example, suppose we observe the following gamete frequencies:

Gamete	AB	Ab	aB	ab
Recombination probability	0.49	0.01	0.01	0.49

It is quite straightforward to calculate the recombination rate as r=.01+.01=.02=2%.

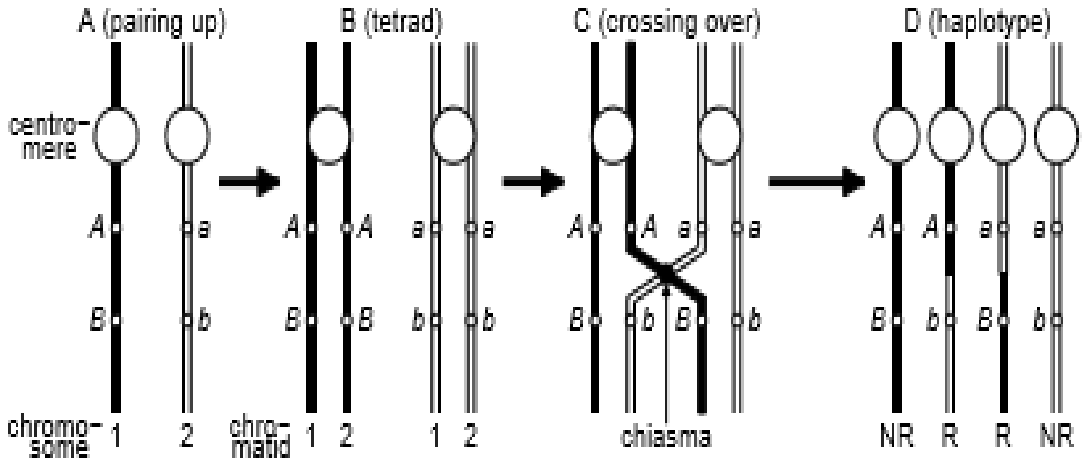


Figure 3-1: Diagram for crossing-over between linked loci A and B*[53]*.In general, for two linked loci, their recombination fraction should be smaller than 0.5, and for two unrelated (or independent) lock, their recombination fraction is 0.5. Hence recombination fraction is a measure of genetic linkage and is usually used in the creation of a genetic linkage map.

3.2 Map distance and map function

One limitation of recombination fraction is that it is not additive. Suppose we have 3 markers A, B and C and let AB be the event that odd number of crossovers occur between A and B and the recombination fraction $r_{ab} = P(AB)$. Assuming crossovers between A and B are independent of crossovers between B and C, we can get:

$$\begin{aligned} \mathrm{P(AC)} &= P(AB \cap BC^C) + P(AB^c \cap BC) \\ &= P(AB)P(BC^C) + P(AB^c)P(BC) \end{aligned}$$

So that:

$$\begin{aligned} r_{ac} &= r_{ab}(1 - r_{bc}) + (1 - r_{ab})r_{bc} \\ &= r_{ab} + r_{bc} - 2r_{ab}r_{bc} \end{aligned}$$

Which shows the non-additiveness of recombination fraction.

To find some additive distance measure, the genetic distance (d) between two loci is defined, which is the expected number of crossovers occurring between them on a single meiosis. Each chiasma (i.e. crossover) involves 2 of the 4 the potential gametes. Hence, the expected number of chiasma between the loci on the tetrad is 2d. Also, Genetic distance is always additive, since expectations are additive. We usually measure genetic distance in Morgans or centi-Morgans: 100cM = 1 Morgan. The unit Morgan is defined so that crossovers occur at an average rate 1 per Morgan (M) or 0.01 per centi-Morgan (cM).

If there is no crossover or there is an even number of crossovers between two loci, the two haplotype generated from meiosis will be exactly the same as parental one; while on the other hand, if there is an odd number of crossovers, two recombinant haplotypes will occur. Based on these, a

theoretical model has been derived to present the recombination fraction between two loci using their map distance.

Map function is a mathematical transformation that connects the recombination fraction (r) between two loci to the genetic distance between them. There are several types of map functions, including the Mather's formula (1938), the Morgan function, the Haldane map function and the Kosambi map function.

3.2.1 Mather's function

In Mather [54] derivation, the recombination fraction (r) between two loci is half the probability of the crossover occurring in all four strands of tetrad between the loci.

As shown below:

$$\mathrm{r} = \frac{1}{2}Prob(X > 0) = \frac{1}{2}Prob(1 - Prob(X = 0))$$

The $Prob(X = 0)$ is the probability that there is no crossover occurring between the two loci. According to the definition, the map distance (d) is the expected number of crossovers occurring on a single chromatid during meiosis. Then:

$$\mathrm{d} = \frac{1}{2}E(X)$$

Because each crossover involves two chromatids.

3.2.2 Haldane Map Function

The Haldane map function [55] is one of the simplest map function. It assumes that crossovers occur randomly and independently of each other. Under this assumption, we can view the occurrence of crossovers between two loci on a chromosome as a Poisson process, which

means at any point between the loci, it happens with the equal probability. The probability of the number of crossovers can be expressed using the Poisson distribution in terms of the genetic distance (d) as follows:

Crossover	0	1	2	3	…	k	…
Probability	e^{-d}	$\frac{d}{1!}e^{-d}$	$\frac{d^2}{2!}e^{-d}$	$\frac{d^3}{3!}e^{-d}$		$\frac{k^2}{k!}e^{-d}$	

The recombination rate is the sum of the probabilities of all the odd numbers of crossovers so that we will have the following formula:

$$r = e^{-d}\left(\frac{d}{1!} + \frac{d^3}{3!} + \frac{d^5}{5!} + \cdots\right)$$

$$= \frac{1}{2}(1 - e^{-2d})$$

In this thesis, we will mainly use the Haldane map function due to its simplicity.

3.2.3 The Morgan map Function

The Morgan is another simple map function. It assumes single crossover occurring between adjacent loci, and the probability of a crossover is proportional to the map length of the interval (Morgan 1928). Under these assumptions, we can get:

$$\mathrm{r} = \frac{1}{2}[1 - Prob(X = 0)] = \frac{1}{2}[1 - (1 - 2d)] = d$$

As stated before, when $\mathrm{d} > \frac{1}{2}$ this function cannot be considered true since r cannot exceed $\frac{1}{2}$.

3.2.4 Kosambi Map Function

In the previous map function, crossovers are assumed to be random and independent of each other. In practice, however, this is not always the case. We call this non-randomness interference. In general, occurrence of a crossover tends reduces the probability of other crossover in its nearby region.

Kosambi [56] mapping function assumes a constant and specific level of interference with the following formula:

$$\mathrm{d} = \frac{1}{4}\ln\left(\frac{1+2r}{1-2r}\right)$$

Figure 3-2 shows the relationship between the genetic distance (in centi-Morgan) and recombination fraction using Kosambi map function. Kosambi is preferred for most cases when there is evidence for interference. However, in reality the level of interference is usually unknown, and seems to vary across the genome (Sherman & Stack, 95)

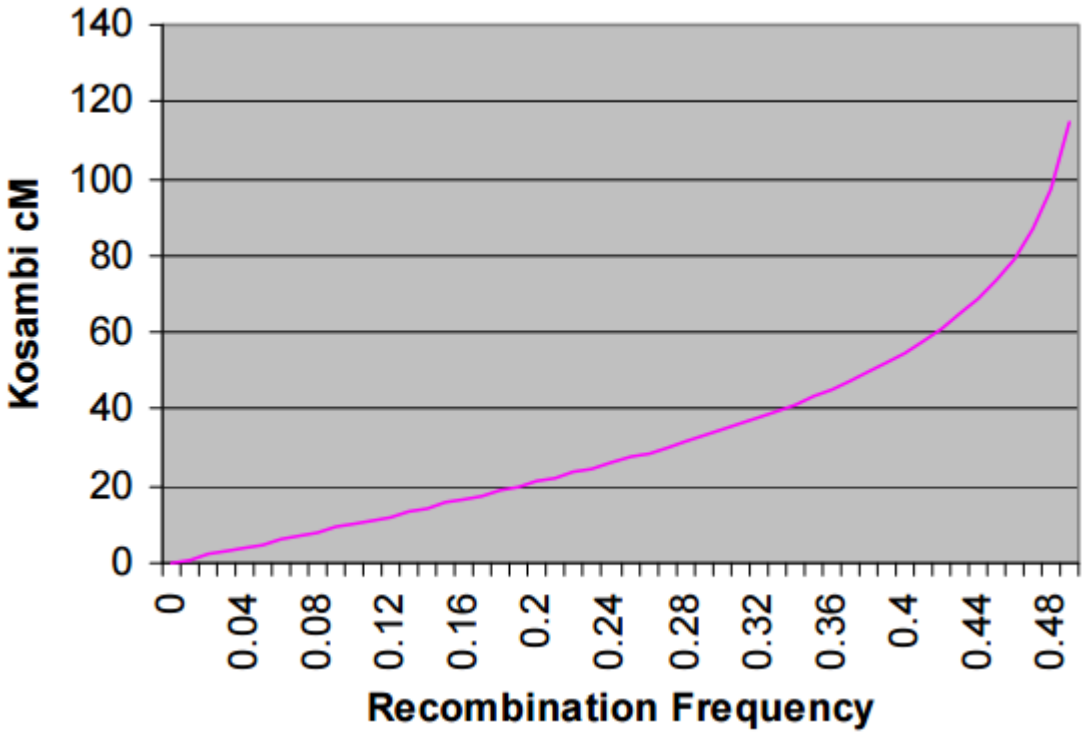


Figure 3-2: The relationship between the genetic distance and recombination fraction.

3.3 Interval Mapping

The basic idea of interval mapping is straightforward. We first consider an interval between two observable markers M1 and M2, each having two possible alleles M1, m1, M2, m2. Suppose the genetic distance and recombination frequency between the two markers have been previously estimated, and a map function (either Haldane or Kosambi) is used to convert between them.

For an F2 population, the two markers form 9 genotypic groups. The trait measurements are then from a mixture distribution with 3 components, corresponding to the three genotype of an unobserved QTL, as showed in the following table.

Table 3-1: The genotypes formed by the 2 markers

	L	R	N	QQ (0) $f(u_0, \sigma_0^2)$	Q|q + q|Q (1) $f(u_1, \sigma_1^2)$	Qq (2) $f(u_2, \sigma_2^2)$
1	M1M1	M2M2	n_1	$w_{0\|1}$	$w_{1\|1}$	$w_{2\|1}$
2	M1M1	M2m2	n_2	$w_{0\|2}$	$w_{1\|2}$	$w_{2\|2}$
3	M1M1	m2m2	n_3		..	
..			..	..	..	..
9	m1m1	M2m2	n_9	$w_{0\|9}$	$w_{1\|9}$	$w_{2\|9}$

The genotypes formed by the 2 markers and the conditional probability of the genotypes of QTL given the markers genotype. The three QTL genotypes (QQ, Qq and qq) are denoted as 0, 1 and 2, respectively.

For joint probabilities involving more than two loci (e.g. three), all recombination rates among these loci need to be considered. For a single QTL flanked by two markers M1 and M2, the gamete frequencies then depend on three parameters: the recombination frequency r between the two markers, the recombination fraction r_1 between marker M1 and the QTL, and the recombination fraction r_2 between the QTL and marker M2.

From the definition of recombination fraction, the probabilities of haplotypes can be expressed as:

$$\mathrm{P(M1M2)} = P(m1m2) = \frac{(1-r)}{2}$$

$$\mathrm{P(M1m2)} = P(m1M2) = \frac{r}{2}$$

$$\mathrm{P(M1M2Q)} = P(m1m2q) = \frac{1}{2}(1-r_1)(1-r_2)$$

$$\mathrm{P(M1M2q)} = P(m1m2Q) = \frac{1}{2}r_1 r_2$$

$$\mathrm{P(M1m2Q)} = P(m1M2q) = \frac{1}{2}(1 - r_1)r_2$$

$$\mathrm{P(M1m2q)} = P(m1M2Q) = \frac{1}{2}r_1(1 - r_2)$$

Based on these, the joint marker-QTL genotype frequencies in Table 3-2 can be derived and the conditional probability of the genotype of QTL given a certain type of paired markers can also be computed.

As introduced in the previous chapter, the likelihood of the measurements (y) and the marker (M) at the underlying QTL is constructed through a Gaussian mixture model, expressed as:

$$\mathrm{L(Y, M)} = \prod_{i=1}^{n}\sum_{j=0}^{2} w_{j|i_k}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y_i - \mu_j)^2}{2\sigma^2})$$

Where the $w_{j|i_k}$ is the conditional probability of the *i*th subject to carry QTL genotype *j*, given that his marker type, and u_j is the mean for the *j*th QTL genotypic subgroup.

Table 3-2: Joint marker-QTL genotype frequencies in a F2 population

L	R	Frequency	QQ	Q|q + q|Q	qq
M1M1	M2M2	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}(1-r_1)^2(1-r_2)^2$	$\frac{1}{2}r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{4}r_1^2r_2^2$
M1M1	M2m2	$\frac{1}{2}r(1-r)$	$\frac{1}{2}r_2(1-r_1)^2(1-r_2)$	$\frac{1}{2}r_1(1-r_1)(1-2r_2+2r_2^2)$	$\frac{1}{2}r_1^2r_2(1-r_2)$
M1M1	m2m2	$\frac{1}{4}r^2$	$\frac{1}{4}(1-r_1)^2r_2^2$	$\frac{1}{2}r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{4}r_1^2(1-r_2)^2$
M1m1	M2M2	$\frac{1}{2}r(1-r)$	$\frac{1}{2}r_1(1-r_1)(1-r_2)^2$	$\frac{1}{2}r_2(1-r_2)(1-2r_1+2r_1^2)$	$\frac{1}{2}r_1(1-r_1)r_2^2$
M1m1	M2m2	$\frac{1}{2}(1-2r+2r^2)$	$r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{2}(1-2r_1+2r_1^2)(1-2r_2+2r_2^2)$	$r_1r_2(1-r_1)(1-r_2)$
M1m1	m2m2	$\frac{1}{2}r(1-r)$	$\frac{1}{2}r_1(1-r_1)r_2^2$	$\frac{1}{2}r_2(1-r_2)(1-2r_1+2r_1^2)$	$\frac{1}{2}r_1(1-r_1)(1-r_2)^2$
m1m1	M2M2	$\frac{1}{4}r^2$	$\frac{1}{4}r_1^2(1-r_2)^2$	$\frac{1}{2}r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{4}(1-r_1)^2r_2^2$
m1m1	M2m2	$\frac{1}{2}r(1-r)$	$\frac{1}{2}r_1^2r_2(1-r_2)$	$\frac{1}{2}r_1(1-r_1)(1-2r_2+2r_2^2)$	$\frac{1}{2}r_2(1-r_1)^2(1-r_2)$
m1m1	M2m2	$\frac{1}{4}(1-r)^2$	$\frac{1}{4}r_1^2r_2^2$	$\frac{1}{2}r_1r_2(1-r_1)(1-r_2)$	$\frac{1}{4}(1-r_1)^2(1-r_2)^2$

For the interval mapping, we calculate a LOD score at an equal step size, for example, each 2cM in the interval and finally get the profile of LOD score for the whole genome. The LOD score is defined as:

$$\text{LOD} = log_{10}\frac{\text{likelihood of qtl exists at this loci}}{\text{likelihood of no qtl at this loci}}$$

Large values of LOD score indicate high probability of the corresponding loci being a QTL. When some peak of the profile exceeds a pre-set threshold value, we say that a QTL have been found at that location.

3.4 Penalized EM algorithm in interval mapping

As stated previously, to allow difference variances for different QTL groups, a penalized likelihood needs to be employed.

$$\mathrm{L(Y,M)} = \prod_{i=1}^{n}\sum_{j=0}^{2} w_{j|i_k}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i-\mu_j\right)^2}{2\sigma^2}\right)\ e^{-\lambda\sum_{j=1}^{2} w_{j|i_k}\frac{1}{\sigma_j}}$$

In order to obtain the maximum likelihood estimates, we need to implement a penalized version of the EM algorithm.

Since given marker information of each subject, the conditional probability for each QTL genotype is fixed and known, in the maximization step, we only need to update the value of (μ, σ) in each iterative process.

In the E-step, under the current estimate of the parameters $\boldsymbol{\theta}^{(t)}$, the expected value of the log likelihood function, with respect to the conditional distribution of latent variable Z and the observed data y, is :

$$E_{Z|Y}[\log L(Y,Z)] = \sum_{i=1}^{n}\sum_{j=0}^{2} T_{j,i_k}^{(t)} \cdot [\log w_{j|i_k} - \frac{1}{2}\log 2\pi - \log\sigma - \frac{(y_i-\mu_j)^2}{2\sigma_j^2}]$$

(3.2.1)

In this step, we will add the penalty term $\lambda\sum_{j=0}^{2}\sum_{i_k=1}^{9}\frac{n_{i_k}}{n} w_{j|i_k}\frac{1}{\sigma_j}$ to the corresponding likelihood as:

$$E_{Z|Y}\left[\log L(Y,Z)\right]=\sum_{i=1}^{n}\sum_{j=0}^{2}T_{j,i_k}^{(t)}(\log w_{j|i_k}-\frac{1}{2}\log 2\pi-\log\sigma_j-\frac{(y_i-\mu_j)^2}{2\sigma_j^2})-\lambda\sum_{i_k=1}^{9}\frac{n_{i_k}}{n}\sum_{j=0}^{2}w_{j|i_k}\frac{1}{\sigma_j}$$

(3.2.2)

n_{i_k} is the number of observations that in the $i_k th$ marker group, $i_k = 1,2,3 \ldots..9$.

From previous EM algorithm chapter, we know that $T_{j|i_k}$ is defined as the conditional distribution of the latent variable Z given the observed data and current estimate of $\theta^{(t)}$

$$T_{j,i_k}=P(Z_i=j\,|\,Y_i=y_i;\theta^{(t)})=\frac{w_{j|i_k}\,f_j(y_i)}{\sum_{l=0}^{2}w_{l|i_k}\,f_l(y_i)}$$

(3.2.3)

In the M-step, from Table 3-2, we will be able to calculate the conditional probability $w_{j|i_k}$ and these values are fixed and are not updated in the EM algorithm:

$$w_{j|i_k=l}=\frac{P(Marker=l\ and\ QTL=j\ )}{P(Markder=l)}$$

(3.2.4)

For the maximization of (μ,σ):

For u_j, since it is not involved in the penalty term, the updated value of $u_j^{(t+1)}$remains the same as that in the standard EM algorithm.

$$\mu_j^{(t+1)}=\frac{\sum_{i}^{n}T_{j,i_k}y_i}{\sum_{i}^{n}T_{j,i_k}}$$

(3.2.5)

For σ_j, the derivative of (3.2.2) with respect of σ_j yields:

$$\sigma_j^2\left(\sum_{i=1}^{n} T_{j,i_k}\right) - \lambda\sigma_j\left(\sum_{i_k=1}^{9}\frac{n_{i_k}}{n}w_{j|i_k}\right) - \sum_{i=1}^{n} T_{j,i_k}\left(y_i - u_j\right)^2 = 0 \tag{3.2.6}$$

Solving the above equation gives the updated σ_j:

$$\sigma_j^{(t+1)} = \frac{\lambda c + \sqrt{(\lambda c)^2 + 4T_{j,i_k}\cdot\sum_{i=1}^{n} T_{j,i_k}\left(y_i - u_j\right)^2}}{2\sum_{i=1}^{n} T_{j,i_k}} \tag{3.2.7}$$

Where $c = \sum_{i_k=1}^{9}\frac{n_{i_k}}{n}w_{j|i_k}$

These 2 steps are repeated iteratively until 3.2.2 converges.

Below is the proof to show that for any initial point θ^0, the penalized likelihood is non-decreasing at each step, i.e

$$L_{pen}\left(\theta^{i+1}\right) > L_{pen}\left(\theta^{i}\right), \quad i = 0,1,\ldots.$$

Proof:

$$\log\left(L(Y|\theta)\right) = \log\left(L(Y,Z|\theta)\right) - logL(Z|Y,\theta))$$

We take the expectation over values of Z by multiplying both sides by $L(Z|Y,\theta^t))$ and summing over Z.

$$\log\left(L(Y|\theta)\right) = \sum_{Z} L(Z|Y,\theta^t)\ \log\left(L(Y,Z|\theta)\right) - logL(Z|Y,\theta))$$

$$= E_{Z|Y,\theta^{(t)}}\log\left(L(Y,Z|\theta)\right) - logL(Z|Y,\theta))$$

$$= Q\left(\theta\middle|,\theta^{(t)}\right) + H(\theta|,\theta^{(t)})$$

Where Q is $E_{Z|Y,\theta^{(t)}}\log\left(L(Y,Z|\theta)\right)$, and H is $-logL(Z|Y,\theta)$.

If we add a penalty term on both sides:

$$\log(\, L_{pen}(Y|\theta)) = Q_{pen}(\left(\theta\right|, \theta^{(t)}) + H(\theta|, \theta^{(t)})$$

For any $\theta = \theta^{(t)}$,

$$log\ (L_{pen}(Y|\theta^{(t)})) = Q_{pen}\left(\theta^{t}\right|, \theta^{(t)}) + H(\theta^{t}|, \theta^{(t)})$$

And

$$\log(L_{pen}(Y|\theta)) - \log L_{pen}(Y|\theta^{(t)})$$

$$= Q_{pen}(\left(\theta\right|, \theta^{(t)}) - Q_{pen}(\left(\theta^{(t)}\right|, \theta^{(t)}) + H\left(\theta\right|, \theta^{(t)}) - H\left(\theta^{(t)}\right|, \theta^{(t)})$$

Jensen's inequality tells us that $H\left(\theta\right|, \theta^{(t)}) > H\left(\theta^{(t)}\right|, \theta^{(t)})$,

So

$$log\ (L_{pen}(Y|\theta)) - log\ L_{pen}(Y|\theta^{(t)}) \geq Q_{pen}\left(\theta\right|, \theta^{(t)}) - Q_{pen}\left(\theta^{(t)}\right|, \theta^{(t)})$$

For $\theta = \theta^{(t+1)}$ that maximizes $Q_{pen}\left(\theta\right|, \theta^{(t)})$, i.e. $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q_{pen}\left(\theta\right|, \theta^{(t)})$,

$$\log(L_{pen}(Y|\theta^{(t+1)})) - \log L_{pen}(Y|\theta^{(t)}) \geq Q_{pen}\left(\theta^{(t+1)}\right|, \theta^{(t)}) - Q_{pen}\left(\theta^{(t)}\right|, \theta^{(t)})$$

$$\geq 0$$

3.5 Hypothesis test

To test the existence of a QTL at each scanned position, the hypotheses are formulated as:

H0: There is no QTL in the whole interval so that:

$$u_j = \mu\ and\ \sigma_j = \ \sigma\ for\ j = 0,1,2$$

H1: Single QTL exists in this interval so that:

$$At\ least\ one\ of\ the\ equalities\ above\ not\ hold$$

We define the LOD score to be:

$$\text{LOD} = log_{10} \frac{\text{likelihood of qtl exists at this loci}}{\text{likelihood of no qtl at this loci}}$$

Larger LOD score indicate higher probability of QTL existence so that the test statistic is max(LOD) in the whole interval.

3.6 Permutation test

In hypothesis testing, a decision will be made by comparing the value of a test statistic, which usually has different values under the null hypothesis and the alternatives. The sampling distribution of the test statistic under the null hypothesis needs to be identified. For some test statistics, we are able to find a parametric distribution of the test statistics, and the p-value for the test, which is the probability that the test statistic would be at least as extreme as observed value under the null hypothesis, is calculated to compare with the threshold value to make decision on the test. However, a theoretical distribution cannot always be obtained. In this case, we can use a permutation test to solve this problem.

Permutation testing can be dated back to Fisher (1935) [57], and its essential idea is to compute the null distribution of the test statistics from the data. Suppose we have n observation, each having a measurement of trait, $y_1, y_2, \ldots . y_n$ and the genotypes of p markers, as showed in the following table.

Table 3-3: Samples with the observed information of p markers and the phenotype

Subject	Phenotype	Marker				
		m1	m2	m3	..	mp
1	$y_1 = 23.0$	0	1	2		2
2	$y_2 = 11.2$	2	2	1		2
3	$y_3 = 15.7$	2	1	0		0

…		…	…			
n	$y_n = 27.4$	1	0	0		1

If the null hypothesis is true, the pairings between the outcome y and the marker information should not be unique. That is, the pairings found in the observed data is just one of the possible, equally likely many other pairings. A realization of the null hypothesis can be obtained by randomly shuffling or permuting the observed data, e.g., by randomly matching different outcome values with genotype data of all subjects.

For example, after shuffling the values of y while keep the order of the markers, we might get the following realization:

Table 3-4: Samples with the permuted information of p markers and the phenotype

Subject	Phenotype	Marker				
		m1	m2	m3	..	mp
1	$y_2 = 11.2$	0	1	2		2
2	$y_6 = 9.9$	2	2	1		2
3	$y_n = 27.4$	2	1	0		0
…		…	…			
n	$y_4 = 19.2$	1	0	0		1

The random shuffling/permuting can be repeated for a large number of times, e.g. 1000 times, and each permutation can then generate a value of test statistic under null hypothesis. The values of the test statistics for all permutations can be used to derive an empirical distribution of the test statistics, and p-value can be obtained for the observed test statistic.

The number of permutations to be conducted is a trade-off between precision and computation time. Usually the more permutations the better, since probability estimates are subject to error due to sampling the population of possible permutations. But as the number of permutation increased, the requirement for the computation power and time is growing linearly. The permutation test can be very time-consuming, especially when we study large datasets. In exploratory analyses, 500 to 1000 permutations are sufficient to ensure the stability of the probability estimates. If the computed p-value is close to the preselected significance level, more permutation runs are needed.

Permutation provides an efficient and effective approach and the key advantage is that one does not have to worry about the assumption of the distribution of the test statistic, whether it is unknown or poorly assumed, as we do in traditional testing procedures. The disadvantage is, as stated in the last paragraph, the amount of computation time required to actually perform a large number of permutations sometimes can be impractical. Good news is this disadvantage has been weakened with development of computer and paralleling computing.

To apply the permutation procedures in QTL mapping, we use 1000 permutations to get the empirical distribution of the test statistic: max (LOD). Figure 3-3 shows an example of scanning profiles for typical QTL mapping. The data is a realization as describe in Table 3-4, where there are 500 samples and each has a measurement of trait (y) and 20 markers. The y-axes are the LOD score. The x-axes are the scanning positions and the 20 markers. The black dots are the LOD scores computed from the penalized method at each scanning position while the red ones are from the standard ML method. We can see that the max (LOD) is somewhere between the 12^{nd} and 13^{th} marker. The observed value is 11.38 and 2.25from the penalized and the standard ML method, respectively.

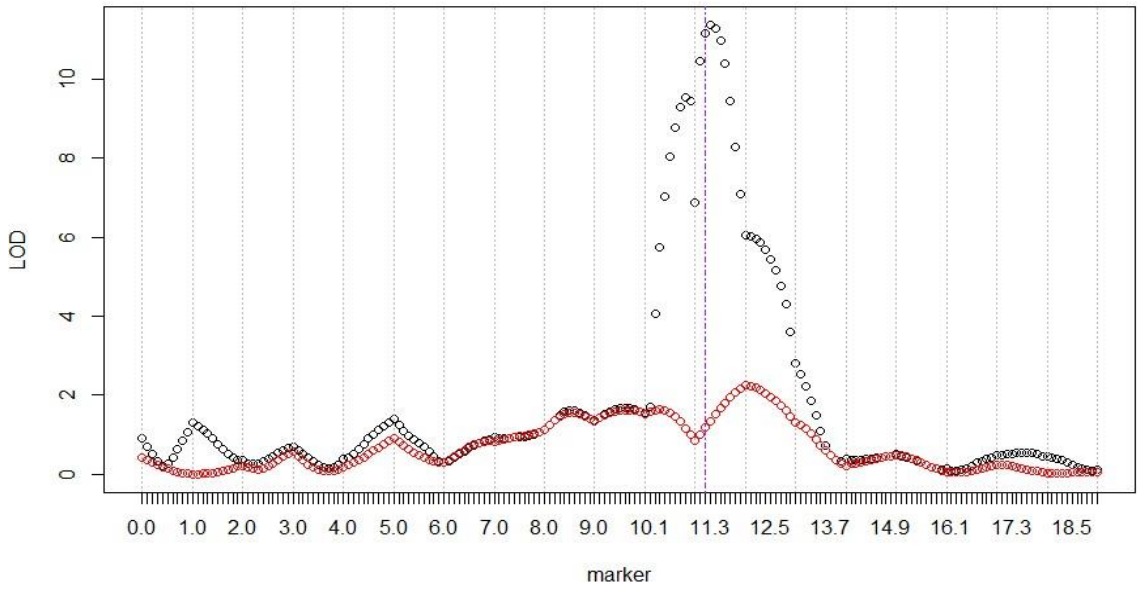


Figure 3-3: Scanning profile of QTL controlling the trait of interest. The y-axes are the LOD score. The x-axes are the scanning positions and the 20 markers. The black dots are the LOD scores computed from the penalized method while the red ones are from the standard ML method.

Then we permute the data and scan the interval again to calculate the LOD score at each scanning position and record the largest LOD value. We repeat this for 1000 times to form our reference distribution of the test statistic: max (LOD).

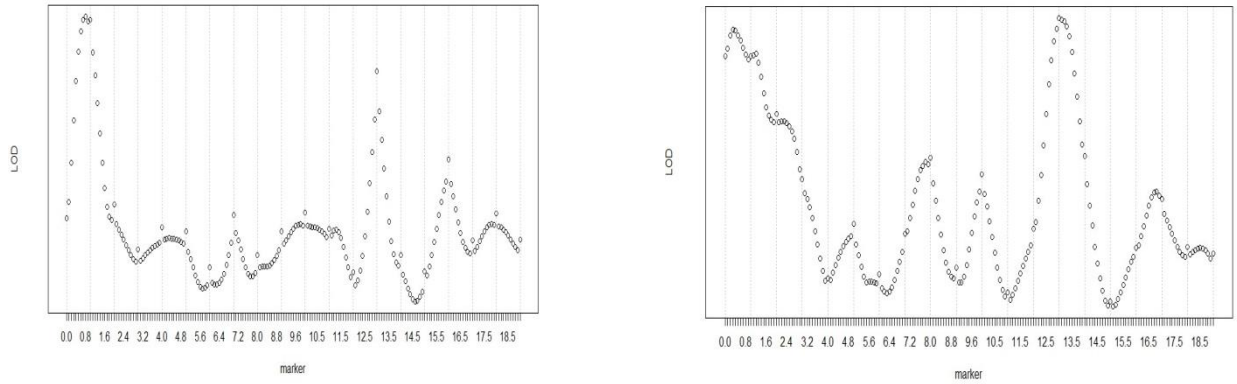


Figure 3-4: Examples of scanning profile of QTLs using permutated data. The y-axes are the LOD score. The x-axes are the scanning positions and the 20 markers.

After obtaining the 1000 values from the permutations, a histogram can be generated, as showed in the following figures for both un-penalized and penalized methods.

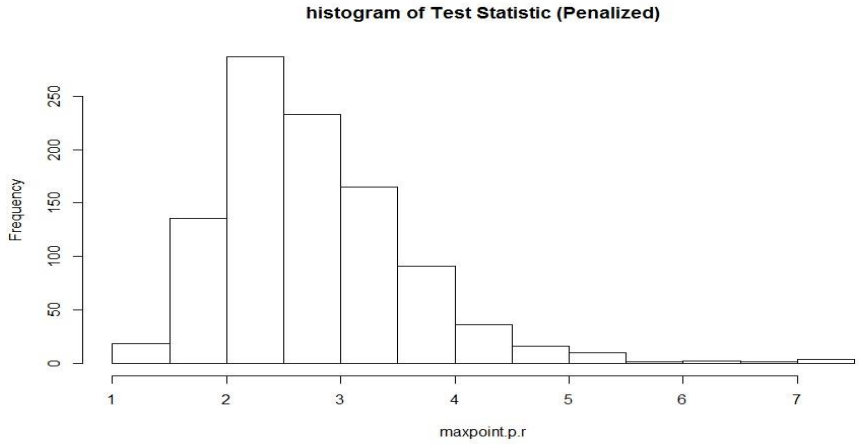


Figure 3-5: Values of 1000 max (LOD) using the penalized method.

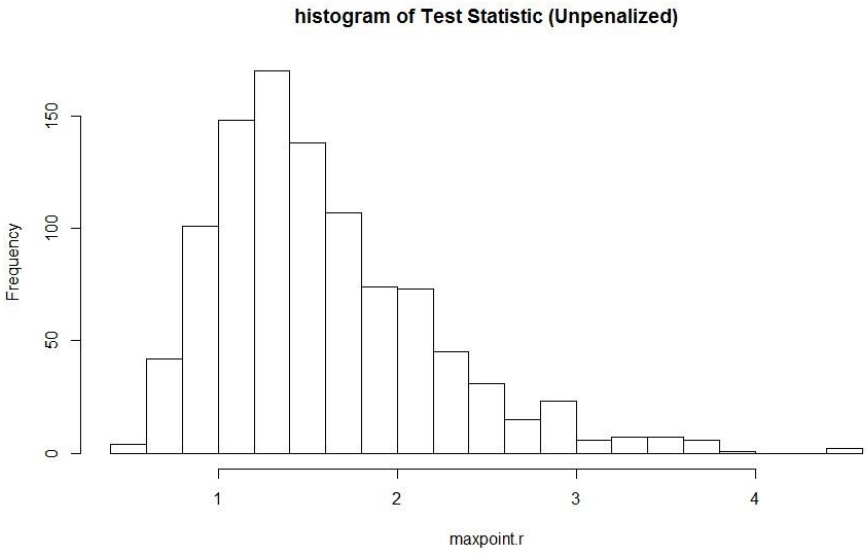


Figure 3-6: Values of 1000 max (LOD) using the standard ML method.

The p-value for the penalized method is then calculated as,

$$p - value = \frac{\#(\mathrm{TS} > 11.38)}{1000} = 0$$

Where the 11.38 is the max(LOD) calculated using the penalized method from the observed dataset in the profile showed in Figure 3-3 (Red).

For the standard unpenalized ML method, the p-value is:

$$p - value = \frac{\#(\mathrm{TS} > 2.25)}{1000} = 0.123$$

Similarly, 2.25 is the max (LOD) calculated using the standard ML method from the observed dataset in the profile showed in Figure 3-3 (Black)

We compare the calculated p-value with the pre-selected significance level, for example, 0.05 to makes decisions about rejecting the null hypothesis or not.

3.7 Simulation

We performed Monte Carlo simulation to examine the statistical property of the penalized method in interval mapping. The simulation contains two parts, one for generating the marker/QTL information and one for generating the trait value based on the simulated QTL genotypes. The QTL information has been removed for analyses to mimic real data in which QTL is unknown. The detailed setting is described as follows:

(1) For F2, suppose QTL has three genotypes (QQ, Qq, qq) and is flanked by two markers with genotypes (M1M1, M1m1, m1m1) and (M2M2, M2m2, m2m2), respectively. For each individual, the first marker on the chromosome is MM, Mm, mm with the probability (0.25,0.5,0.25). For the next position, the chance of obtaining certain type of genotype is determined by its recombination frequency with the previous marker as the conditional probability expressed in the following table:

Marker 1 genotype	Marker 2 genotype

	M2M2	M2m2	m2m2
M1M1	$(1-r)^2$	$2r(1-r)$	r^2
M1m1	$r(1-r)$	$(1-r)^2+r^2$	$r(1-r)$
m1m1	r^2	$2r(1-r)$	$(1-r)^2$

For example, if the distance between each marker is 20cM, then the next position, if it is a marker, by using the Haldane map function, the recombination frequency should be 0.165 and the probability of current position falling into the 3 genotypes will be determined by the table above. After determining the genotype for current position, we can record the genotype value and proceed to the next one. The process will continue until all markers and QTLs have been reached.

(2) After the QTL genotypes are simulated in step (1), we will be able to generate the measurement of the trait, which are from 3 different normal distributions with different mean and variance. The trait value for each individual was obtained according to the observation's genotype of QTL.

In our simulation, the number of markers is set to be 20 and the distance between each marker is 20cM. The location of QTL is assumed to be 4cM left to the 12nd marker. The penalized method is compared to the standard ML approach in which the variances are assumed to be homogenous. We set the sample size to be from 100 to 500. The means of the trait measurement for 3 different QTL families are (1,2,3) and the variances are adjusted to see how the difference in variances may affect power of detecting a significant QTL. The σ is set to be (3-δ, 3, 3+δ), where δ varies from 0 to 1. We scanned the simulated chromosome with a step size of 2cM from the leftmost marker to

the rightmost. At each scanning point, we compute the LOD value. The maximum LOD in the whole profile will be recorded. A detected QTL is defined by having a LOD score value that is greater than a predefined threshold. The QTL will then be counted for calculating power of QTL detection. The predefined threshold is calculated from permutation test (permuted for 1000 times) with type I error $\alpha = 0.05$. The power and parameter estimates are computed from 100 simulation replicates. The simulations are conducted to compare the consistency and efficiency between the penalized and standard ML methods under different sample size and variation of the variance in 3 QTL subgroups.

Table 3-5: Parameter estimation of interval mapping when δ = 0

			100		200
Parameter	True	Pen	ML	Pen	ML
u0	1	1.26(0.092)	1.15(0.106)	1.08(0.068)	0.98(0.065)
u1	2	1.96(0.061)	1.95(0.067)	2.02(0.04)	2.04(0.044)
u2	3	2.74(0.105)	2.88(0.108)	2.88(0.058)	2.95(0.063)
sigma0	2.2	2.86(0.064)	2.84(0.022)	3.01(0.035)	2.95(0.015)
sigma1	3	2.96(0.037)	2.84(0.022)	3.02(0.027)	2.95(0.015)
sigma2	3.8	2.85(0.06)	2.84(0.022)	2.99(0.037)	2.95(0.015)
pos	224	212.54(9.143)	228.48(7.928)	215.54(6.3)	222.86(5.685)
			300		400
Parameter	True	Pen	ML	Pen	ML
u0	1	1.11(0.051)	1.06(0.052)	1.1(0.045)	1.05(0.044)
u1	2	2.01(0.031)	2.01(0.034)	2(0.025)	2(0.027)
u2	3	2.93(0.048)	2.98(0.045)	2.97(0.04)	3.01(0.039)

sigma0	2.2	3.03(0.031)	2.95(0.014)	3.04(0.027)	2.97(0.011)
sigma1	3	2.97(0.022)	2.95(0.014)	2.98(0.019)	2.97(0.011)
sigma2	3.8	3(0.033)	2.95(0.014)	3.04(0.028)	2.97(0.011)
pos	224	223.96(4.375)	226.06(3.661)	225.26(2.904)	224.34(2.799)
			500		
Parameter	True	Pen	ML		
u0	1	1.06(0.037)	1.03(0.037)		
u1	2	2(0.023)	2.01(0.023)		
u2	3	2.98(0.037)	3(0.036)		
sigma0	2.2	3.03(0.022)	2.97(0.01)		
sigma1	3	2.98(0.016)	2.97(0.01)		
sigma2	3.8	3.04(0.025)	2.97(0.01)		
pos	224	224.04(2.628)	224.18(2.014)		

The penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 100, 200, 300, 400,500 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations

Table 3-6: Parameter estimation of interval mapping when $\delta = 0.2$

		100		200	
Parameter	True	Pen	ML	Pen	ML
u0	1	1.25(0.09)	1.2(0.099)	1.09(0.066)	1(0.062)
u1	2	1.96(0.06)	1.93(0.067)	2.02(0.04)	2.02(0.044)
u2	3	2.77(0.103)	2.88(0.117)	2.86(0.064)	2.97(0.067)
sigma0	2.8	2.71(0.062)	2.85(0.022)	2.83(0.033)	2.95(0.015)

sigma1	3	2.97(0.037)	2.85(0.022)	3.04(0.028)	2.95(0.015)
sigma2	3.2	2.98(0.064)	2.85(0.022)	3.16(0.039)	2.95(0.015)
pos	224	215.32(8.93)	223(8.079)	213.02(6.539)	224.16(5.654)
			300		400
Parameter	True	Pen	ML	Pen	ML
u0	1	1.06(0.044)	1.08(0.05)	1.08(0.038)	1.06(0.042)
u1	2	2.02(0.029)	1.99(0.034)	2.01(0.026)	1.99(0.027)
u2	3	2.97(0.043)	3(0.048)	2.98(0.042)	3.03(0.041)
sigma0	2.8	2.85(0.031)	2.95(0.014)	2.84(0.027)	2.97(0.012)
sigma1	3	2.97(0.022)	2.95(0.014)	2.99(0.019)	2.97(0.012)
sigma2	3.2	3.19(0.034)	2.95(0.014)	3.24(0.03)	2.97(0.012)
pos	224	224.78(3.863)	224.84(3.846)	219.62(3.179)	224.26(2.801)
			500		
Parameter	True	Pen	ML		
u0	1	1.05(0.034)	1.04(0.033)		
u1	2	2(0.022)	1.99(0.024)		
u2	3	2.98(0.036)	3.02(0.038)		
sigma0	2.8	2.83(0.021)	2.97(0.01)		
sigma1	3	2.99(0.016)	2.97(0.01)		
sigma2	3.2	3.23(0.026)	2.97(0.01)		
pos	224	220.22(2.732)	224.74(2.34)		

The Penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 100, 200, 300, 400,500 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations

Table 3-7: Parameter estimation of interval mapping when $\delta = 0.5$

		100		200	
Parameter	True	Pen	ML	Pen	ML
u0	1	1.22(0.084)	1.22(0.095)	1.07(0.056)	1.04(0.059)
u1	2	1.97(0.059)	1.9(0.067)	2.02(0.037)	1.98(0.043)
u2	3	2.79(0.106)	2.94(0.119)	2.9(0.066)	3(0.072)
sigma0	2.5	2.54(0.061)	2.87(0.022)	2.55(0.026)	2.97(0.016)
sigma1	3	2.99(0.038)	2.87(0.022)	3.05(0.027)	2.97(0.016)
sigma2	3.5	3.15(0.071)	2.87(0.022)	3.44(0.044)	2.97(0.016)
pos	224	217.74(8.508)	220.18(7.986)	219.52(4.63)	224.84(5.761)
			300		400
Parameter	True	Pen	ML	Pen	ML
u0	1	1.05(0.036)	1.1(0.047)	1.06(0.033)	1.07(0.039)
u1	2	2.02(0.026)	1.97(0.034)	2.01(0.024)	1.97(0.027)
u2	3	2.97(0.045)	3.03(0.052)	2.99(0.041)	3.05(0.046)
sigma0	2.5	2.55(0.027)	2.97(0.015)	2.54(0.022)	2.99(0.012)
sigma1	3	2.98(0.021)	2.97(0.015)	3(0.018)	2.99(0.012)
sigma2	3.5	3.51(0.037)	2.97(0.015)	3.55(0.032)	2.99(0.012)
pos	224	224.72(2.408)	223.42(4.214)	225.88(1.251)	224.04(2.813)
			500		
Parameter	True	Pen	ML		
u0	1	1.04(0.029)	1.05(0.029)		
u1	2	2(0.02)	1.97(0.024)		

u2	3	2.99(0.035)	3.04(0.042)
sigma0	2.5	2.53(0.019)	2.99(0.01)
sigma1	3	3(0.016)	2.99(0.01)
sigma2	3.5	3.54(0.027)	2.99(0.01)
pos	224	224.46(0.873)	224.74(2.348)

The Penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 100, 200, 300, 400,500 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations

Table 3-8: Parameter estimation of interval mapping when $\delta = 0.8$

			100		200
Parameter	True	Pen	ML	Pen	ML
u0	1	1.16(0.068)	1.26(0.088)	1.06(0.047)	1.12(0.06)
u1	2	1.96(0.055)	1.83(0.066)	2.01(0.034)	1.93(0.045)
u2	3	2.88(0.109)	3.05(0.124)	2.95(0.064)	3.04(0.079)
sigma0	2.2	2.28(0.054)	2.89(0.025)	2.27(0.023)	2.99(0.017)
sigma1	3	2.98(0.036)	2.89(0.025)	3.04(0.027)	2.99(0.017)
sigma2	3.8	3.49(0.079)	2.89(0.025)	3.75(0.044)	2.99(0.017)
out	224	222.44(7.006)	213.66(8.53)	225.04(2.71)	228.4(5.651)
			300		400
Parameter	True	Pen	ML	Pen	ML
u0	1	1.04(0.031)	1.11(0.045)	1.05(0.027)	1.09(0.034)
u1	2	2.01(0.027)	1.93(0.033)	2.01(0.023)	1.95(0.028)
u2	3	2.98(0.046)	3.07(0.057)	3.01(0.041)	3.07(0.054)

sigma0	2.2	2.26(0.022)	3(0.015)	2.25(0.017)	3.02(0.013)
sigma1	3	3(0.021)	3(0.015)	3.01(0.016)	3.02(0.013)
sigma2	3.8	3.81(0.036)	3(0.015)	3.85(0.033)	3.02(0.013)
pos	224	224.96(0.883)	221.16(4.204)	225(0.501)	224.04(2.838)

			500
Parameter	True	Pen	ML
u0	1	1.04(0.025)	1.06(0.026)
u1	2	2(0.02)	1.96(0.024)
u2	3	2.99(0.038)	3.06(0.045)
sigma0	2.2	2.24(0.016)	3.02(0.011)
sigma1	3	3(0.015)	3.02(0.011)
sigma2	3.8	3.84(0.028)	3.02(0.011)
pos	224	225.22(0.424)	224.82(2.344)

The Penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 100, 200, 300, 400,500 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations

Table 3-9: Parameter estimation of interval mapping when $\delta = 1$

		100		200	
Parameter	True	Pen	ML	Pen	ML
u0	1	1.22(0.084)	1.22(0.095)	1.07(0.056)	1.04(0.059)
u1	2	1.97(0.059)	1.9(0.067)	2.02(0.037)	1.98(0.043)
u2	3	2.79(0.106)	2.94(0.119)	2.9(0.066)	3(0.072)
sigma0	2	2.54(0.061)	2.87(0.022)	2.55(0.026)	2.97(0.016)

sigma1	3	2.99(0.038)	2.87(0.022)	3.05(0.027)	2.97(0.016)
sigma2	4	3.15(0.071)	2.87(0.022)	3.44(0.044)	2.97(0.016)
pos	224	217.74(8.508)	220.18(7.986)	219.52(4.63)	224.84(5.761)
			300		400
Parameter	True	Pen	ML	Pen	ML
u0	1	1.05(0.036)	1.1(0.047)	1.06(0.033)	1.07(0.039)
u1	2	2.02(0.026)	1.97(0.034)	2.01(0.024)	1.97(0.027)
u2	3	2.97(0.045)	3.03(0.052)	2.99(0.041)	3.05(0.046)
sigma0	2	2.55(0.027)	2.97(0.015)	2.54(0.022)	2.99(0.012)
sigma1	3	2.98(0.021)	2.97(0.015)	3(0.018)	2.99(0.012)
sigma2	4	3.51(0.037)	2.97(0.015)	3.55(0.032)	2.99(0.012)
pos	224	224.72(2.408)	223.42(4.214)	225.88(1.251)	224.04(2.813)
			500		
Parameter	True	Pen	ML		
u0	1	1.04(0.029)	1.05(0.029)		
u1	2	2(0.02)	1.97(0.024)		
u2	3	2.99(0.035)	3.04(0.042)		
sigma0	2	2.53(0.019)	2.99(0.01)		
sigma1	3	3(0.016)	2.99(0.01)		
sigma2	4	3.54(0.027)	2.99(0.01)		
pos	224	224.46(0.873)	224.74(2.348)		

The Penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 100, 200, 300, 400,500 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations

Below are the summaries of the results:

(1) When the variances of the 3 genotype groups are homogeneous and the sample size is small, say, N =100 and N = 200 in table 3-5, the standard ML method works better than the penalized method in estimating the parameters. As the sample size grows, it will improve the performance of both methods and the penalized method becomes as good as the standard ML method.

(2) When the variances of the three genotypic groups start to vary, the penalized estimators are consistent while the ML estimators become biased as the variation is larger.

(3) When the variance gets bigger, larger sample size will lead to better performance of both methods in estimating the parameters and reducing the standard errors. For example, when $\delta = 0.8$ and sample size 100 in Table 3-8, the standard errors of the position of the QTL for the Penalized and standard ML are 7.006 and 8.53, respectively. While when the sample size is 500, these values are reduced to 0.424 and 2.344. This is pretty satisfying for the penalized method while 2.344 is still a relative large value for the standard ML method.

(4) As the delta becomes larger and larger, the penalized method has an increasing power of detecting a significant QTL while the ML method shows an opposite tendency after a specific value as showed in Figure 3-7. For instance, when N = 300, the power for standard ML grows as δ grows until it reaches 0.2. However, it decreases after 0.2 from 0.75 to 0.73 then goes to 0.72 when δ exceeds 0.8. At the same time, the penalized method can reach a power as high as 1 when $\delta = 1$ and the sample size larger than 300.

Table 3-10: Power of two methods in interval mapping

delta	0		0.2		0.5		0.8		1	
	Pen	ML	Pen	ML	Pen	ML	Pen	ML	Pen	ML
N = 100	0.2	0.25	0.19	0.22	0.28	0.24	0.41	0.23	0.54	0.23
N = 200	0.37	0.52	0.43	0.53	0.57	0.49	0.85	0.47	0.96	0.47
N = 300	0.59	0.72	0.65	0.75	0.87	0.73	0.98	0.72	1	0.72
N = 400	0.81	0.88	0.87	0.88	0.96	0.86	0.99	0.83	1	0.82
N = 500	0.9	0.93	0.91	0.93	0.99	0.93	1	0.92	1	0.9

Interval mapping under when $\delta = 0, 0.2, 0.5, 0.8, 1$, the Penalized ML and standard ML's power of detecting a QTL using an F2 population of sample size N = 100, 200, 300, 400, 500 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. The results are calculated from 100 simulations.

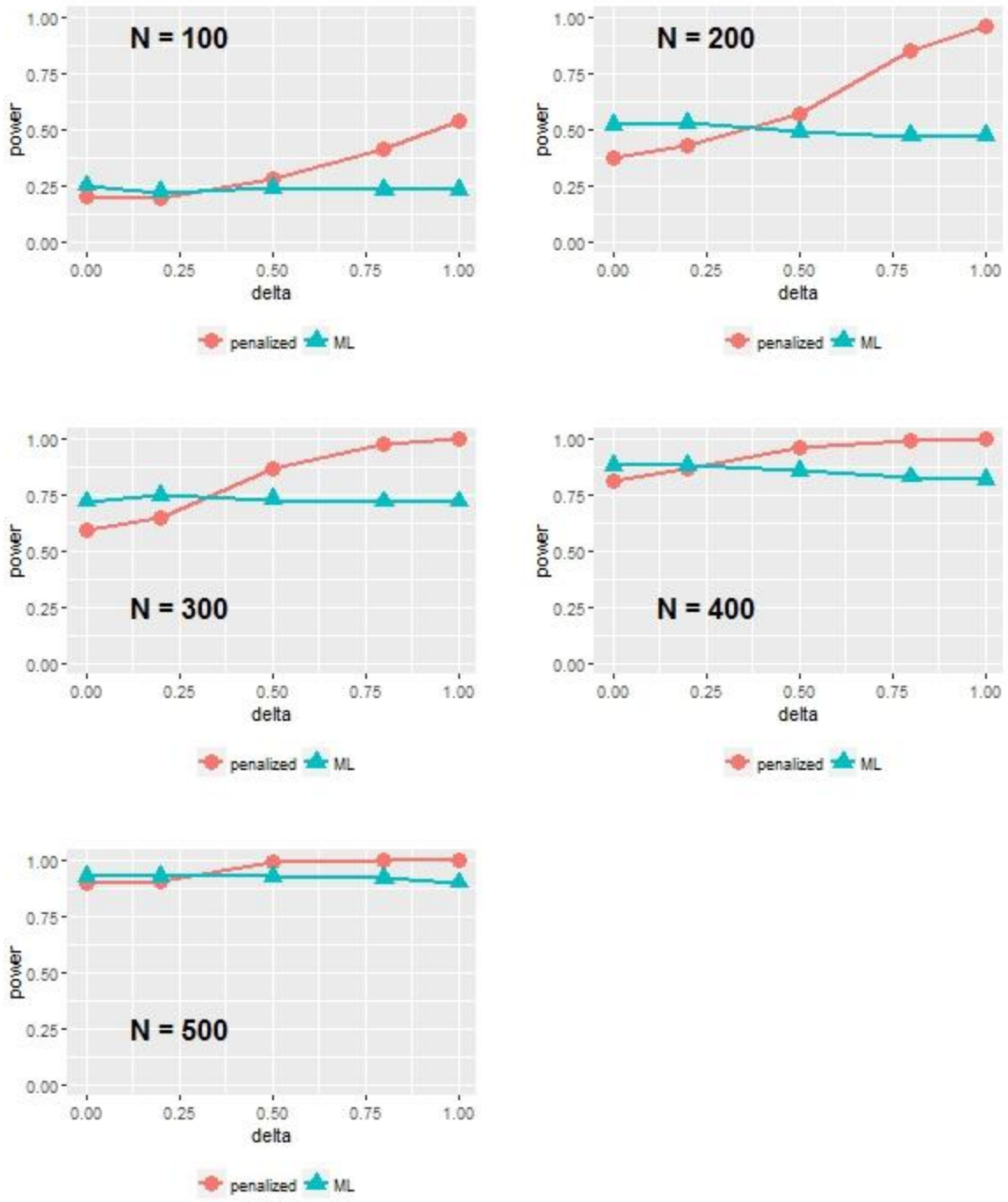


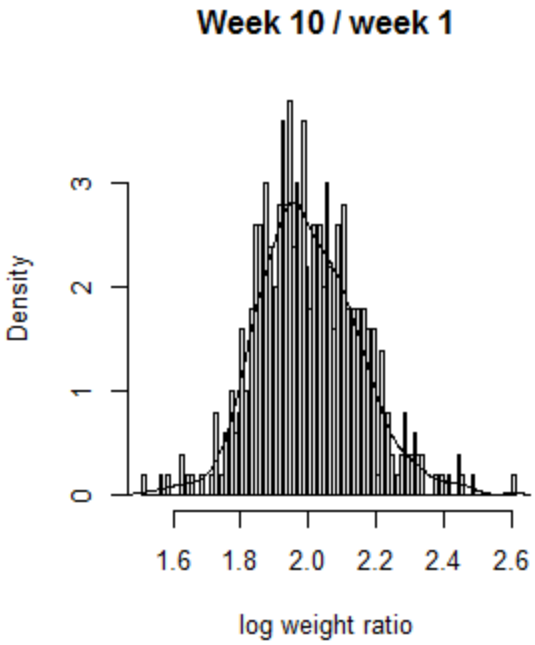
Figure 3-7: Power of Penalized and Standard ML of detecting QTL in interval mapping. Phenotypic data is simulated from Gaussian mixture distributions with 3 components. The results are calculated from 100 simulations under scenarios δ = 0, 0.2, 0.5, 0.8, 1 with sample size N = 100, 200, 300, 400,500.

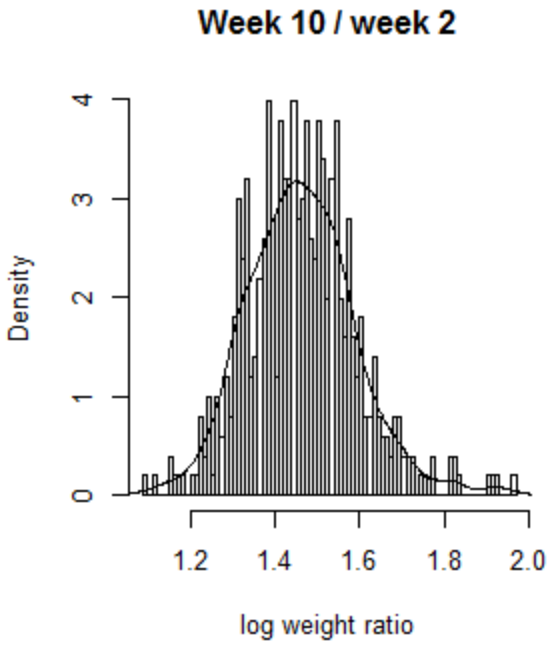
3.8 Real data analysis for interval mapping

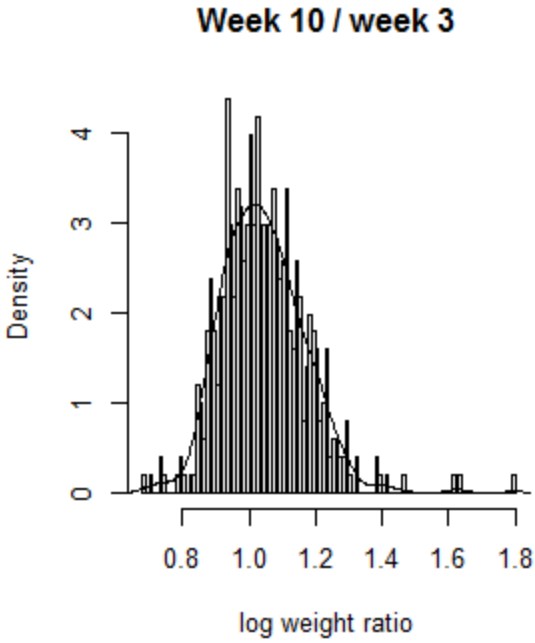
In this section, we will apply our proposed penalized method to a real data. We will compare it with the standard ML method that assumes the homogeneous variance across the genotypic groups of QTL.

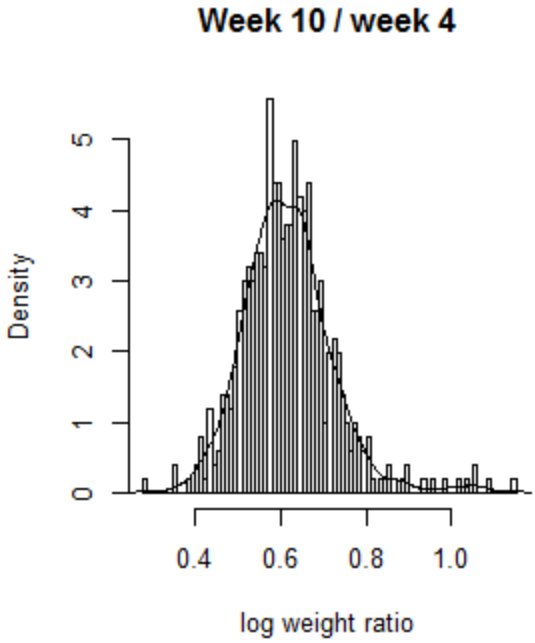
Vaughn et Al [58] constructed a QTL map with 96 markers for 502 F2 mice (259 males and 243 females) which are derived from two inbreed strains. The F2 progeny's body mass was measured weekly for 10 weeks after they were born.

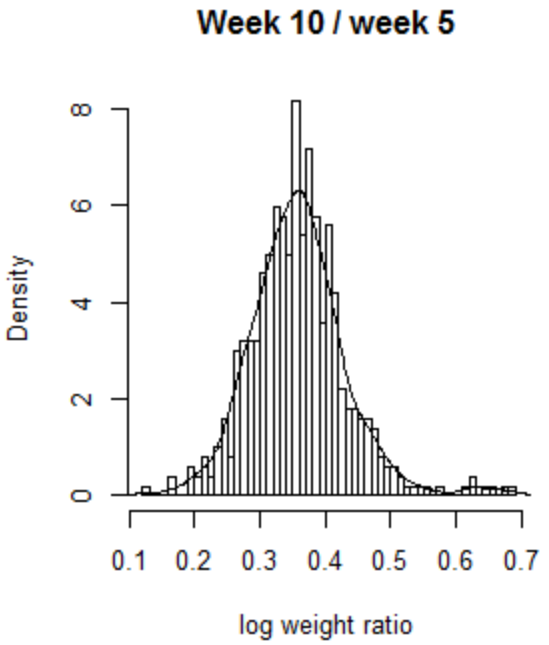
We calculated the log body mass ratio of week 10 over week 1 to week 9 and plotted the histogram as presented in Figure 3-8. These ratios indicate how fast the mice grow. We choose week 10 over week 3 as our traits of interest.

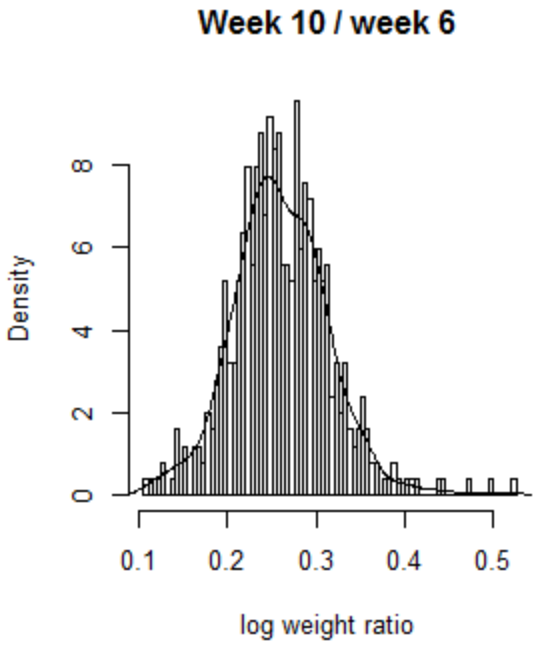

Week 10 / week 1
Density
0
1
2
3
1.6
1.8
2.0
2.2
2.4
2.6
log weight ratio

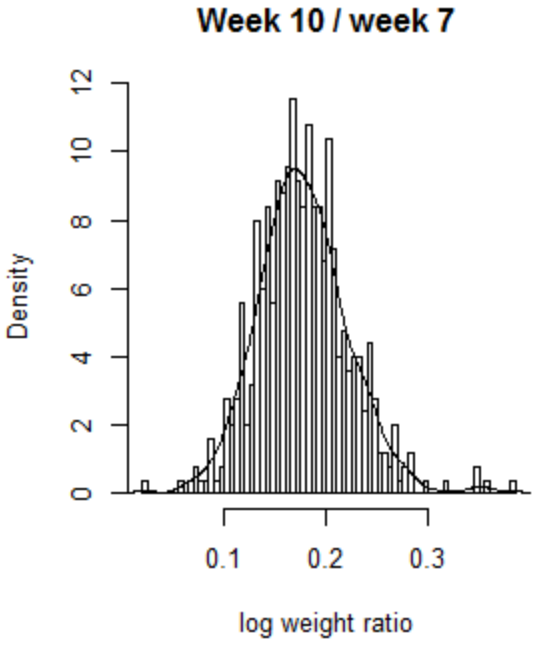

Week 10 / week 2
Density
0
1
2
3
4
1.2
1.4
1.6
1.8
2.0
log weight ratio

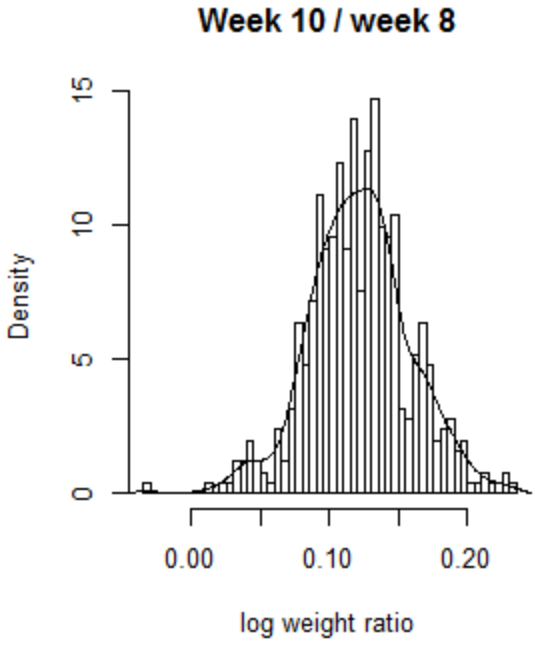

Week 10 / week 3
Density
0
1
2
3
4
0.8
1.0
1.2
1.4
1.6
1.8
log weight ratio


Week 10 / week 4
Density
0
1
2
3
4
5
0.4
0.6
0.8
1.0
log weight ratio


Week 10 / week 5
Density
0
2
4
6
8
0.1
0.2
0.3
0.4
0.5
0.6
0.7
log weight ratio


Week 10 / week 6
Density
0
2
4
6
8
0.1
0.2
0.3
0.4
0.5
log weight ratio


Week 10 / week 7
Density
0
2
4
6
8
10
12
0.1
0.2
0.3
log weight ratio


Week 10 / week 8
Density
0
5
10
15
0.00
0.10
0.20
log weight ratio

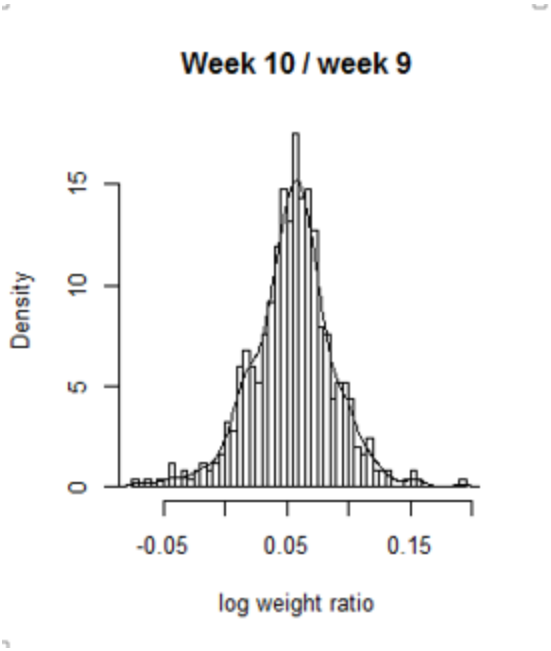
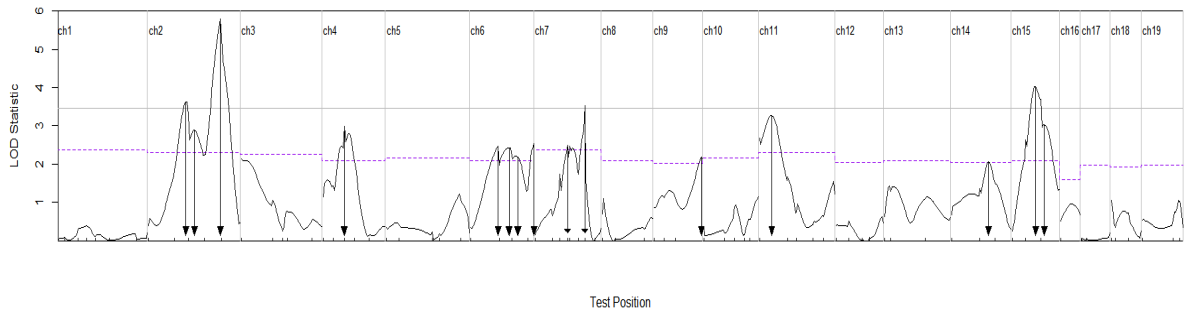
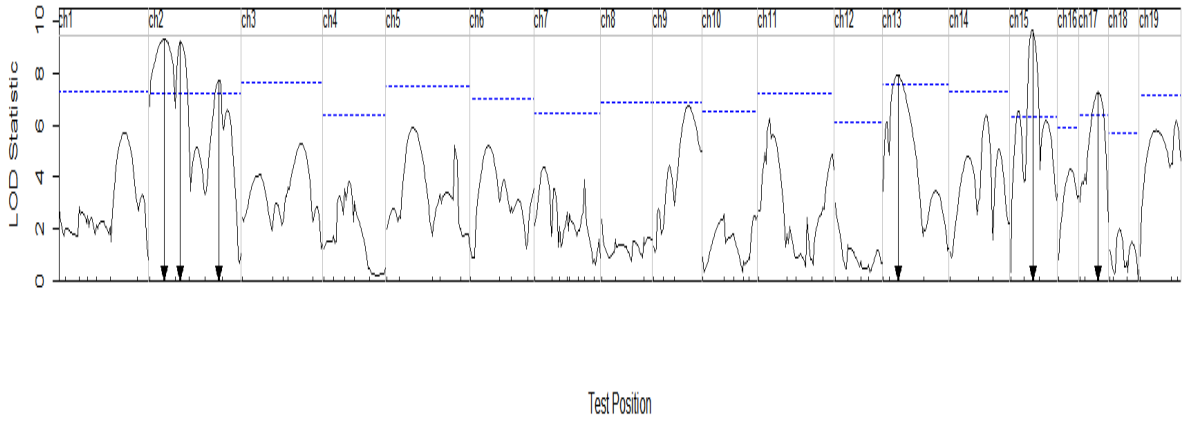


Figure 3-8: Body mass ratio for the 502 mice, between week 10 and week 1 to week 9.

Our goal is to identify QTL that affect the growth rate of body mass from week 3 to week 10. These 96 markers spread on 19 chromosomes. The profile of the test statistics, LOD score across the whole mice genome is shown in the Figure 3-9. The empirical distribution of the test statistics were obtained from 1000 permutations and the significance level was set at 5%.



(a)



(b)

Figure 3-9: QTL scanning profiles by standard ML (a) and penalized ML (b). The y-axes are the LOD test scores. The dash lines are the 0.05 significant level chromosome-wide while the solid line is the significant level genome-wide based on 1000 permutations. The x-axis ticks is the marker positions.

Results:

Both the penalized likelihood and ML detected one genome-wide significant QTL on the chromosome 15. Although the overall profiles of the penalized and un-penalized ML look similar, they did detect different chromosome-wide significant QTL. Additional chromosome-wide significant QTL locations are identified in chromosomes 13 and 17 by the penalized approach while the standard ML shows no sign of an existence of QTL in these 2 chromosomes, suggesting that the penalized method may provide insights that the standard ML method cannot provide.

Table 3-11: Genome-wide significant QTL detected in an F2 mouse population

Chromosome	Position*	Left Marker	Right Marker
15	12.2	D15Mit5	D15Mit2

*position in cM from the leftmax marker on the chromosome.
Significant QTL for body mass ratio between week 10 and week 3 in an F2 mouse population detected from the genome-wide interval mapping scan by the penalized and ML methods, at the 0.05 significance level from the permutation.

Table 3-12: Chromosome-wide significant QTL detected in an F2 mouse population

Chromosome	Position*	Left Marker	Right Marker
2	24	D2Mit1	D2Mit370
2	6.8	D2Mit370	D2Mit380
2	21.5	D2Mit17	D2Mit22
13	13.5	D13Mit115	D13Mit9
17	20.5	D17Mit16	D17Mit39

*position in cM from the leftmax marker on the chromosome.
Significant QTL for body mass ratio between week 10 and week 3 in an F2 mouse population detected from the chromosome-wide interval mapping scan by the penalized methods, at the 0.05 significance level from the permutation.

Chapter 4 Linkage disequilibrium mapping

In this chapter, we will describe how to apply our penalized approach to the linkage disequilibrium mapping framework. The linkage disequilibrium mapping is mainly applied for data collected from natural populations, such as humans, where controlled mating is not possible.

4.1 Linkage disequilibrium

Linkage disequilibrium occurs when genotypes at two loci are not independent of another. In another word, two loci are in linkage disequilibrium when the haplotype frequencies of the two loci are different from random associations of composing alleles. Consider two loci (A and B), each having two alleles (A1, A2, B1, and B2). Therefore four possible haplotypes may present in the population:

Table 4-1: The Frequency of the 4 haplotypes formed by two loci.

Haplotype	Frequency
A1B1	p_{11}
A1B2	p_{12}
A2B1	p_{21}
A2B2	p_{22}

Then, the allele frequency can be represented by the haplotype counts as:

Table 4-2: The Frequency of the 4 types of Allele in the two loci

Allele	Frequency
A1	$p = p_{11} + p_{12}$
A2	$1 - p = p_{21} + p_{22}$
B1	$q = p_{11} + p_{21}$
B2	$1 - q = n_{12} + n_{22}$

If alleles at the two loci are randomly associated with each other, then the frequencies of the four haplotype are equal to the product of the frequencies of alleles it contains. In this case, there is no linkage disequilibrium and gamete frequencies can be expressed as:

$$p_{11} = pq$$

$$p_{12} = p\,(1 - q)$$

$$p_{21} = (1 - p)q$$

$$p_{22} = (1 - p)(1 - q)$$

However, if alleles at the two loci are not randomly associated, then there will a deviation (D) in the expected frequencies:

$$p_{11} = pq + D$$

$$p_{12} = p\,(1 - q) - D$$

$$p_{21} = (1 - p)q - D$$

$$p_{22} = (1 - p)(1 - q) + D$$

This parameter D is called the coefficient of linkage disequilibrium and it quantifies the deviations from random association of alleles. This coefficient was first proposed by Lewontin and Kojima (1960) [59] and it is defined for a specific pair of alleles, A and B. If D = 0, it means linkage

equilibrium (LE), implying a statistical independence. The coefficient of linkage disequilibrium is a descriptive statistics. Their magnitude does not indicate whether or not there is a statistically significant association between alleles in haplotypes. Standard statistical tests, including the chi-squared and Fisher's exact test, are commonly used to test for significance:

$$\chi^2 = \frac{\sum(obs - exp)^2}{exp}$$

Linkage disequilibrium is influenced by many factors, including selection, the rate of recombination, the rate of mutation, genetic drift, the system of mating, population structure, and genetic linkage.

Several statistics have been proposed to measure the amount of linkage disequilibrium, and they have different advantages. Although the measure D has the most intuitive concepts of linkage disequilibrium, its numerical value is not very useful for measuring the strength of linkage disequilibrium or comparing different levels of linkage disequilibrium, because D is affected by allele frequencies. Researchers suggested the value should be normalized based on the theoretical maximum and minimum relative to the value of D. so the standardized value of D is proposed by Lewontin (1964) [60]. It is estimated as:

- When $D \geq 0$:

$$D' = \frac{D}{D_{max}}$$

D_{max} is the smaller of p(1-q) and (1-p)q.

- When $D < 0$:

$$D' = \frac{D}{D_{min}}$$

D_{min} is the larger of –pq and –(1-p)(1-q).

When D'=1, it is known as complete linkage disequilibrium, which means that two markers have not been separated by recombination in the population and occurs only when some haplotypes have frequency equals to zero. D' = 0 represents linkage equilibrium.

The other measure R^2 is defined as:

$$R^2 = \frac{D^2}{pq(1-p)(1-q)}$$

Where the nominator is the square of the linkage disequilibrium parameter D and the denominator is the product of the four allele frequencies. The range of R^2 is between 0 and 1. R^2 = 1 means complete linkage disequilibrium while R^2 = 0 means complete linkage equilibrium. In general, R^2 is used to measure the statistical association between marker pairs and is related to the power of the LD mapping. It is preferred when the focus is on the predictability of one polymorphism given the other.

4.2 Linkage disequilibrium mapping

Linkage disequilibrium mapping is performed by scanning the entire genome for significant associations between markers and a particular phenotype. Suppose a QTL has the

alleles Q and q. The allele frequencies of Q and q are expressed as q and 1 - q, respectively. Three genotypes can be formed (j=0 for QQ, 1 for Qq and 2 for qq) for this QTL.

Assume this QTL is genetically associated with a genetic marker with three genotypes MM, Mm and mm. Let p and 1-p be the allele frequencies of M and m, respectively, and D be the coefficient of linkage disequilibrium between the marker and QTL.

The marker and QTL together form four haplotypes, MQ, Mq, mQ and mq, with respective frequencies expressed as $p_{11} = \mathrm{pq} + \mathrm{D}$, $p_{10} = \mathrm{p}(1-\mathrm{q}) - \mathrm{D}$, $p_{01} = (1-\mathrm{p})\mathrm{q} - \mathrm{D}$, $p_{00} = (1-\mathrm{p})(1-\mathrm{q}) + \mathrm{D}$. These four haplotypes randomly unite to generate 16 combinations, some of which are collapsed to form nine distinguishable genotypes with frequencies presented in Table 4-3.

Table 4-3: Genotypic and diplotypic frequencies for the maker and QTL

		QTL = 2	QTL = 1	QTL = 0
		QQ	Q|q + q|Q	qq
Marker = 2	MM	p_{11}^2	$2p_{11}p_{10}$	p_{10}^2
Marker = 1	Mm	$2p_{11}p_{01}$	$2p_{11}p_{00} + 2p_{10}p_{01}$	$2p_{10}p_{00}$
Marker = 0	Mm	p_{01}^2	$2p_{01}p_{00}$	p_{00}^2

From this table, we will be able to calculate the conditional probability of the QTL genotypes given a specific genotype of marker so that a mixture model can be constructed as followed:

$$\mathrm{L}(\mathrm{Y},\mathrm{M}) = \prod_{i=1}^{n}\sum_{j=0}^{2} w_{j|i_k}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\mu_j)^2}{2\sigma_j^2}\right))$$

Similar to the interval mapping, the $w_{j|i_k}$ is the conditional probability of the QTL given the genotype of *i*th subject's marker and (μ_j, σ_j^2) are the mean and variance of *j*th QTL genotypic subgroup.

4.3 Penalized EM algorithm in linkage disequilibrium mapping

From previous EM algorithm chapter, we defined $T_{j|i_k}$ to be the conditional distribution of the latent variable Z given the observed data and current estimate of $\theta^{(t)}$:

$$T_{j,i_k} = P(Z_i = j \mid Y_i = y_i; \theta^{(t)}) = \frac{w_{j|i_k} f_j(y_i)}{\sum_{l=0}^{2} w_{l|i_k} f_l(y_i)}$$

So that the expected value of the log likelihood function, with respect to the conditional distribution of latent variable Z, given the observed data y under the current estimate of the parameters $\boldsymbol{\theta}^{(t)}$ is :

$$E_{Z|Y}[\log L(Y,Z)] = \sum_{i=1}^{n} \sum_{j=0}^{2} T_{j,i_k}^{(t)} \cdot [\log w_{j|i_k} - \frac{1}{2}\log 2\pi - \log \sigma - \frac{(y_i-\mu_j)^2}{2\sigma_j^2}]$$

The penalty term we add in linkage disequilibrium mapping is $\lambda \sum_{j=0}^{2} \frac{1}{\sigma_j}$. We do not consider the weight in order to have an explicit form of estimate of $w_{j|i_k}$. So the the penalized expected value of the log likelihood function is:

$$Q_{pen} = \sum_{i=1}^{n} \sum_{j=0}^{2} T_{j,i_k}^{(t)} \cdot [\log w_{j|i_k} - \frac{1}{2}\log 2\pi - \log \sigma - \frac{(y_i-\mu_j)^2}{2\sigma_j^2}] - \lambda \sum_{j=0}^{2} \frac{1}{\sigma_j} \quad (4\text{-}1)$$

The major difference between the linkage disequilibrium and interval mapping with respect to the EM algorithm is that the conditional probability is not fixed in the linkage disequilibrium mapping and should be updated in each M-step.

In the M-step, since $w_{j|i_k}$ is only involved the first term, we need to maximize

$$K(Y) = \sum_{i=1}^{n} \sum_{j=0}^{2} T_{j,i_k}^{(t)} \cdot \log w_{j|i_k}$$

$$= \sum_{j=0}^{2} (\sum_{i_k=2} T_{j,i_k}^{(t)} \cdot logw_{j|i_k} + \sum_{i_k=1} T_{j,i_k}^{(t)} \cdot logw_{j|i_k} + \sum_{i_k=0} T_{j,i_k}^{(t)} \cdot logw_{j|i_k}) \quad (4\text{-}2)$$

Let ${N_{j|l}}^{(t)}$ be the number of observations that have the Maker = l and QTL = j. We can get:

$${N_{j|l}}^{(t)} = \sum_{i_k = l} T_{j,i_k}^{(t)}$$

Then the likelihood function can be expressed as the product of the nine cells:

$$K(Y) = \sum_{j=0}^{2} ({N_{j|2}}^{(t)} logw_{j|2} + ({N_{j|1}}^{(t)} logw_{j|1} + {N_{j|0}}^{(t)} logw_{j|0})$$

$$= {N_{2|2}}^{(t)} \log p_{11}^2 + {N_{2|1}}^{(t)} \log(2p_{11}p_{10}) + {N_{2|0}}^{(t)} \log(p_{10}^2)$$

$$+ {N_{1|2}}^{(t)} \log(2p_{11}p_{01}) + {N_{1|1}}^{(t)} \log(2p_{11}p_{00} + 2p_{10}p_{01})$$

$$+ {N_{1|0}}^{(t)} \log(2p_{10}p_{00}) + {N_{0|2}}^{(t)} \log(p_{01}^2) + {N_{0|1}}^{(t)} \log(2p_{01}p_{00})$$

$$+ {N_{0|0}}^{(t)} \log(p_{00}^2)$$

$$\propto \left(2{N_{2|2}}^{(t)} + {N_{1|2}}^{(t)} + {N_{2|1}}^{(t)}\right) \log(p_{11})$$

$$+ (2{N_{0|2}}^{(t)} + {N_{1|2}}^{(t)} + {N_{0|1}}^{(t)}) \log(p_{10})$$

$$+ (2{N_{2|0}}^{(t)} + {N_{1|0}}^{(t)} + {N_{2|1}}^{(t)}) \log(p_{01})$$

$$+ (2{N_{0|0}}^{(t)} + {N_{1|0}}^{(t)} + {N_{0|1}}^{(t)})\log(p_{00})$$

$$+ {N_{1|1}}^{(t)}\log(p_{11}p_{00} + p_{10}p_{01})$$

(4-3)

To simply the notation, let

$${N_1}^{(t)} = 2{N_{2|2}}^{(t)} + {N_{1|2}}^{(t)} + {N_{2|1}}^{(t)}$$

$${N_2}^{(t)} = 2{N_{0|2}}^{(t)} + {N_{1|2}}^{(t)} + {N_{0|1}}^{(t)}$$

$${N_3}^{(t)} = 2{N_{2|0}}^{(t)} + {N_{1|0}}^{(t)} + {N_{2|1}}^{(t)}$$

$${N_4}^{(t)} = 2{N_{0|0}}^{(t)} + {N_{1|0}}^{(t)} + {N_{0|1}}^{(t)}$$

$${N_5}^{(t)} = {N_{1|1}}^{(t)}$$

We can apply another EM algorithm here to solve for haplotype frequencies. Let R be a latent variable follow a binomial distribution Bin($ {N_5}^{(t)}, \frac{p_{11}p_{00}}{p_{11}p_{00}+p_{10}p_{01}}$). So that:

$$\varphi = \mathrm{E(R)} = \frac{{N_5}^{(t)}\, p_{11}p_{00}}{p_{11}p_{00} + p_{10}p_{01}}$$

and:

$$\log K(Y,R) = ({N_1}^{(t)} + \mathrm{R})\log p_{11} + ({N_2}^{(t)} + {N_5}^{(t)} + \mathrm{R})\log p_{10} + \left({N_3}^{(t)} + {N_5}^{(t)} - \mathrm{R}\right)\log p_{01}$$
$$+ {N_4}^{(t)} + \mathrm{R})\log p_{10}$$

(4-4)

In the E-step:

$$E_{Z|Y}\mathrm{K(Y,R)} = (N_1 + \varphi)\log p_{11} + (N_2 + N_5 + \varphi)\log p_{10} + (N_3 + N_5 - \varphi)\log p_{01}$$
$$+ (N_4 + \varphi)\log p_{10}$$

(4-5)

After maximizing the above function subject to $p_{11} + p_{10} + p_{01} + p_{00} = 1$, we can get:

$$\widehat{p_{11}}^{(t+1)} = \frac{N_1{}^{(t)} + \varphi}{N_1{}^{(t)} + N_2{}^{(t)} + N_3{}^{(t)} + N_4{}^{(t)} + 2N_5{}^{(t)}}$$

$$\widehat{p_{10}}^{(t+1)} = \frac{N_2{}^{(t)} + N_5{}^{(t)} + \varphi}{N_1{}^{(t)} + N_2{}^{(t)} + N_3{}^{(t)} + N_4{}^{(t)} + 2N_5{}^{(t)}}$$

$$\widehat{p_{01}}^{(t+1)} = \frac{N_3{}^{(t)} + N_5{}^{(t)} - \varphi}{N_1{}^{(t)} + N_2{}^{(t)} + N_3{}^{(t)} + N_4{}^{(t)} + 2N_5{}^{(t)}}$$

$$\widehat{p_{00}}^{(t+1)} = \frac{N_4{}^{(t)} + \varphi}{N_1{}^{(t)} + N_2{}^{(t)} + N_3{}^{(t)} + N_4{}^{(t)} + 2N_5{}^{(t)}}$$

From Table 3-3, we will be able to calculate the updated conditional probability $w_{j|i_k}{}^{(t+1)}$

$$w_{j|i_k=l}{}^{(t+1)} = \frac{P(Marker = l \text{ and } QTL = j)}{P(Markder = l)}$$

We take the derivative of Equation (3.3) with respect of μ and we will get the updated estimate:

$$\mu_j^{(t+1)} = \frac{\sum_i^n T_{j,i_k} y_i}{\sum_i^n T_{j,i_k}}$$

For σ_j, if we take the derivative of (3.3) with respect of σ_j, we will get:

$$\sigma_j^2\left(\sum_{i=1}^n T_{j,i_k}\right) - \lambda\sigma_j\left(\sum_{i_k=1}^9 \frac{n_{i_k}}{n} w_{j|i_k}\right) - \sum_{i=1}^n T_{j,i_k}\left(y_i - u_j\right)^2 = 0 \quad (4\text{-}6)$$

And

$$\sigma_j^{(t+1)} = \frac{\lambda + \sqrt{\lambda^2 + 4T_{j,i_k} \cdot \sum_{i=1}^n T_{j,i_k}\left(y_i - u_j\right)^2}}{2\sum_{i=1}^n T_{j,i_k}} \quad (4\text{-}7)$$

We can see that when $\lambda = 0$, the updated estimate $\sigma_j^{(t+1)} = \sqrt{\frac{\sum_{i=1}^{n} T_{j,i_k}(y_i - u_j)^2}{\sum_{i=1}^{n} T_{j,i_k}}}$, which is just the one from standard EM algorithm.

The algorithm iterates between the 2 steps until it converges.

4.4 Hypothesis test

The hypothesis testing tests the marker is linked to a QTL that affects the trait of interest. The hypothesis is formulated as:

$$H_0: u_j = \mu \; and \; \sigma_j = \sigma \; for \; j = 0,1,2$$

$$H_1 = at \; least \; one \; of \; the \; equalities \; above \; not \; hold.$$

The null hypothesis states that there is no QTL affecting the interested quantitative trait and H_1 proposes that such QTL exists. The test statistics for this hypothesis testing is the likelihood ratio (LR) test statistic. Similar to Interval mapping, larger LR values indicate higher probability of existence of a QTL and we also use the permutation test to get an empirical distribution for determining the critical threshold.

4.5 Simulation

We performed Monte Carlo simulations to examine the statistical property of the penalized method in the linkage disequilibrium mapping framework. The simulation design here is similar to those used in the interval mapping, first generating the marker/QTL information and then generating the trait value based on the simulated QTL. The penalized method is again compared to the standard ML method.

(1) Suppose a QTL has three genotypes (QQ, Qq, qq) and is linked with a marker with genotypes MM, Mm, mm. The allele frequencies for M and m are set as p and 1-p respectively. We first generate the markers, which are randomly selected from one of the three genotypes MM, Mm, mm with the probability $(p^2, 2p(1-p), (1-p)^2)$.

(2) The allele frequencies for Q and q are set as q and 1-q. The coefficient of linkage disequilibrium between the marker and QTL is set to be D. Then we calculate the four haplotypes, MQ, Mq, mQ and mq, with respective frequencies expressed as $p_{11} = \mathrm{pq} + \mathrm{D}$, $p_{10} = \mathrm{p}(1-\mathrm{q}) - \mathrm{D}$, $p_{01} = (1-\mathrm{p})\mathrm{q} - \mathrm{D}$, $p_{00} = (1-\mathrm{p})(1-\mathrm{q}) + \mathrm{D}$, which will give us the joint distribution of the Marker and QTL.

(3) Then we calculate the conditional probability for the 3 genotypes QQ, Qq and qq given the genotypes of the marker and generate the QTL based on the simulated markers.

(4) Lastly, we simulated the measurement of the traits which are from 3 different normal distributions with different mean and variance according to the observation's genotype of QTL.

In our simulation, we set p = 0.6, q = 0.4 and the linkage disequilibrium coefficient D = 0.05. The sample size is set to be 500, 1000, 1500 or 2000. The means of the trait measurement for 3 different QTL families are (1, 2, 3) and the variances are adjusted to see how the difference in variances may affect power. The σ is (3-δ, 3, 3+δ) and we vary δ from 0 to 2. A detected QTL is defined by having a LR value that is greater than a predefined threshold. The significant QTL will then be counted for calculating power of QTL detection. The predefined threshold is calculated from the permutation test (permuted for 1000 times) with type I error $\alpha = 0.05$. The power and parameter estimates are calculated from 100 simulation replicates.

Table 4-4: Parameter estimation of LD mapping under the scenario $\delta = 0$

		N = 500		N=1000	
Names	TRUE	Penalized	ML	Penalized	ML
mu0	1	1.12(0.14)	1.16(0.119)	1.06(0.093)	1.15(0.081)
mu1	2	1.97(0.097)	1.97(0.078)	1.88(0.07)	1.82(0.051)
mu2	3	2.21(0.174)	2.22(0.134)	2.4(0.086)	2.49(0.098)
sigma0	3	2.62(0.056)	2.89(0.026)	2.82(0.047)	2.96(0.015)
sigma1	3	2.86(0.039)	2.89(0.026)	2.95(0.027)	2.96(0.015)
sigma2	3	2.73(0.061)	2.89(0.026)	2.94(0.045)	2.96(0.015)
p	0.4	0.4(0.002)	0.4(0.002)	0.4(0.001)	0.4(0.001)
q	0.6	0.5(0.007)	0.51(0.007)	0.5(0.003)	0.51(0.003)
D	0.05	0.07(0.006)	0.04(0.005)	0.07(0.005)	0.04(0.004)

		N = 1500		N=2000	
mu0	1	1.01(0.077)	1.08(0.074)	1.05(0.068)	1.09(0.061)
mu1	2	1.88(0.047)	1.88(0.036)	1.85(0.043)	1.81(0.034)
mu2	3	2.39(0.074)	2.36(0.067)	2.48(0.061)	2.51(0.067)
sigma0	3	2.93(0.04)	2.99(0.011)	2.95(0.029)	2.99(0.011)
sigma1	3	2.99(0.023)	2.99(0.011)	2.98(0.018)	2.99(0.011)
sigma2	3	2.91(0.041)	2.99(0.011)	2.99(0.029)	2.99(0.011)
p	0.4	0.4(0.001)	0.4(0.001)	0.4(0.001)	0.4(0.001)
q	0.6	0.5(0.002)	0.5(0.001)	0.5(0.001)	0.5(0.001)
D	0.05	0.05(0.004)	0.03(0.003)	0.05(0.003)	0.04(0.003)

The Penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 500, 1000, 1500, 2000 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations.

Table 4-5: Parameter estimation of LD mapping under the scenario $\delta = 0.5$

		N = 500		N=1000	
Names	TRUE	Penalized	ML	Penalized	ML
mu0	1	1.34(0.129)	1.5(0.077)	1.19(0.079)	1.5(0.07)
mu1	2	1.75(0.084)	1.63(0.071)	1.79(0.069)	1.59(0.065)
mu2	3	2.52(0.142)	3.11(0.212)	2.58(0.105)	3.64(0.229)
sigma0	2.5	2.45(0.061)	2.8(0.024)	2.56(0.051)	2.82(0.019)
sigma1	3	2.71(0.043)	2.8(0.024)	2.83(0.032)	2.82(0.019)
sigma2	3.5	3.06(0.064)	2.8(0.024)	3.15(0.054)	2.82(0.019)
p	0.4	0.4(0.002)	0.4(0.002)	0.4(0.001)	0.4(0.001)
q	0.6	0.5(0.008)	0.54(0.01)	0.51(0.007)	0.55(0.011)
D	0.05	0.08(0.005)	0.04(0.004)	0.07(0.004)	0.03(0.003)
		N = 1500		N=2000	
mu0	1	1(0.073)	1.35(0.041)	1.03(0.055)	1.36(0.036)
mu1	2	1.9(0.057)	1.56(0.038)	1.68(0.04)	1.47(0.035)
mu2	3	2.52(0.099)	3.32(0.158)	2.88(0.065)	3.82(0.159)
sigma0	2.5	2.48(0.04)	2.83(0.017)	2.52(0.031)	2.8(0.016)
sigma1	3	2.81(0.031)	2.83(0.017)	2.8(0.021)	2.8(0.016)
sigma2	3.5	3.23(0.045)	2.83(0.017)	3.31(0.03)	2.8(0.016)

p	0.4	0.4(0.001)	0.4(0.001)	0.4(0.001)	0.4(0.001)
q	0.6	0.51(0.004)	0.54(0.007)	0.5(0.002)	0.56(0.008)
D	0.05	0.06(0.003)	0.03(0.003)	0.06(0.002)	0.03(0.002)

The Penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 500, 1000, 1500, 2000 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations.

Table 4-6: Parameter estimation of LD mapping under the scenario $\delta = 1$

		N = 500		N=1000	
Names	TRUE	Penalized	ML	Penalized	ML
mu0	1	1.26(0.096)	1.66(0.113)	1.16(0.059)	1.53(0.073)
mu1	2	1.85(0.091)	1.68(0.061)	1.8(0.067)	1.75(0.085)
mu2	3	2.63(0.17)	4.4(0.298)	2.73(0.113)	5.13(0.309)
sigma0	2	2.18(0.066)	2.73(0.024)	2.09(0.051)	2.71(0.021)
sigma1	3	2.64(0.046)	2.73(0.024)	2.75(0.04)	2.71(0.021)
sigma2	4	3.45(0.077)	2.73(0.024)	3.6(0.065)	2.71(0.021)
p	0.4	0.4(0.002)	0.4(0.002)	0.4(0.001)	0.4(0.001)
q	0.6	0.53(0.01)	0.61(0.014)	0.54(0.008)	0.64(0.013)
D	0.05	0.07(0.004)	0.03(0.003)	0.06(0.003)	0.03(0.003)
		N = 1500		N=2000	
mu0	1	1.05(0.053)	1.44(0.029)	0.98(0.048)	1.43(0.024)
mu1	2	1.76(0.051)	1.73(0.063)	1.69(0.034)	1.51(0.038)
mu2	3	2.77(0.09)	4.78(0.279)	2.91(0.051)	5.37(0.224)

sigma0	2	2.03(0.047)	2.73(0.021)	1.95(0.029)	2.68(0.018)
sigma1	3	2.76(0.039)	2.73(0.021)	2.74(0.023)	2.68(0.018)
sigma2	4	3.57(0.053)	2.73(0.021)	3.75(0.031)	2.68(0.018)
p	0.4	0.4(0.001)	0.4(0.001)	0.4(0.001)	0.4(0.001)
q	0.6	0.52(0.005)	0.63(0.012)	0.51(0.004)	0.65(0.011)
D	0.05	0.05(0.002)	0.02(0.002)	0.05(0.001)	0.03(0.002)

The Penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 500, 1000, 1500, 2000 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations.

Table 4-7: Parameter estimation of LD mapping under the scenario $\delta = 1.5$

		N = 500		N=1000	
names	TRUE	Penalized	ML	Penalized	ML
mu0	1	1.19(0.072)	1.6(0.099)	1.07(0.047)	1.55(0.076)
mu1	2	1.9(0.072)	1.67(0.061)	1.86(0.052)	1.62(0.063)
mu2	3	2.46(0.18)	4.94(0.35)	2.85(0.14)	5.59(0.313)
sigma0	1.5	1.7(0.067)	2.69(0.029)	1.57(0.054)	2.66(0.025)
sigma1	3	2.69(0.06)	2.69(0.029)	2.71(0.053)	2.66(0.025)
sigma2	4.5	3.91(0.095)	2.69(0.029)	4.1(0.073)	2.66(0.025)
p	0.4	0.4(0.002)	0.4(0.002)	0.4(0.001)	0.4(0.001)
q	0.6	0.55(0.01)	0.63(0.014)	0.54(0.01)	0.65(0.013)
D	0.05	0.06(0.003)	0.03(0.003)	0.05(0.002)	0.02(0.002)
		N = 1500		N=2000	
mu0	1	1.06(0.045)	1.47(0.024)	0.98(0.016)	1.45(0.021)

mu1	2	1.8(0.044)	1.63(0.055)	1.76(0.026)	1.53(0.028)
mu2	3	2.78(0.081)	5.28(0.303)	2.93(0.042)	5.99(0.275)
sigma0	1.5	1.45(0.037)	2.68(0.026)	1.39(0.018)	2.64(0.022)
sigma1	3	2.75(0.05)	2.68(0.026)	2.7(0.02)	2.64(0.022)
sigma2	4.5	4.07(0.062)	2.68(0.026)	4.27(0.024)	2.64(0.022)
p	0.4	0.4(0.001)	0.4(0.001)	0.4(0.001)	0.4(0.001)
q	0.6	0.53(0.008)	0.64(0.012)	0.53(0.004)	0.67(0.011)
D	0.05	0.05(0.001)	0.02(0.002)	0.05(0.001)	0.02(0.002)

The Penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 500, 1000, 1500, 2000 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations.

Table 4-8: Parameter estimation of LD mapping under the scenario $\delta = 2$

		N = 500		N=1000	
Names	TRUE	Penalized	ML	Penalized	ML
mu0	1	1.11(0.055)	1.6(0.103)	1.03(0.026)	1.56(0.08)
mu1	2	1.94(0.059)	1.66(0.078)	1.95(0.041)	1.54(0.056)
mu2	3	2.71(0.138)	5.41(0.403)	2.83(0.071)	5.94(0.337)
sigma0	1	1.24(0.076)	2.69(0.034)	1.11(0.063)	2.67(0.03)
sigma1	3	2.79(0.081)	2.69(0.034)	2.82(0.056)	2.67(0.03)
sigma2	5	4.4(0.1)	2.69(0.034)	4.78(0.061)	2.67(0.03)
p	0.4	0.4(0.002)	0.4(0.002)	0.4(0.001)	0.4(0.001)
q	0.6	0.57(0.01)	0.64(0.014)	0.57(0.008)	0.65(0.013)
D	0.05	0.05(0.002)	0.02(0.002)	0.05(0.001)	0.02(0.002)

		N = 1500		N=2000	
mu0	1	1.06(0.041)	1.47(0.022)	1.03(0.02)	1.46(0.02)
mu1	2	1.93(0.032)	1.57(0.053)	1.92(0.024)	1.5(0.028)
mu2	3	2.89(0.055)	5.5(0.334)	2.99(0.045)	6.13(0.317)
sigma0	1	1.02(0.045)	2.7(0.03)	0.99(0.02)	2.66(0.027)
sigma1	3	2.9(0.05)	2.7(0.03)	2.87(0.031)	2.66(0.027)
sigma2	5	4.74(0.065)	2.7(0.03)	4.89(0.028)	2.66(0.027)
p	0.4	0.4(0.001)	0.4(0.001)	0.4(0.001)	0.4(0.001)
q	0.6	0.58(0.007)	0.63(0.012)	0.58(0.005)	0.66(0.012)
D	0.05	0.05(0.001)	0.02(0.002)	0.05(0.001)	0.02(0.002)

The Penalized ML and standard ML estimates of QTL parameters from an F2 population of sample size N = 500, 1000, 1500, 2000 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. Numbers in the parentheses are the mean square errors (MSE) of the estimates. The results are calculated from 100 simulations.

Below are the summaries of the results:

(1) When the variances of the 3 genotype groups are homogeneous, both the penalized and standard ML methods work well as showed in Table 4-4. As the sample size grows, it will improve the performance of both methods.

(2) When the variances of the three genotypic groups start to vary, the penalized estimators perform consistently well while the ML estimators become biased as the variation is large and sample size is small, say, $\delta = 1.5$ and N = 500 in Table 4-7.

(3) When the variance becomes larger, larger sample size improves the performance of the penalized method while it does not help much on the estimation bias in standard ML method.

(4) As the delta becomes larger and larger, i.e the difference between the variances of the 3 genotype groups are large, the penalized method has an increasing power while the ML method shows an opposite tendency after a specific value. Especially when sample size is large, for example, when N=2000 in Figure 4-1, the power for standard ML grows when δ grows until it reaches 1, the power starts to decrease fast after that.

(5) In scenario with larger sample size and smaller delta, the penalized method tends to outperform the standard ML method. For example, when N = 500, the penalized method will have a higher power than the standard ML method at δ = 1.5. However, when N = 2000, this occurs at δ = 0.7 (Figure 4-1)

In summary, the penalized method consistently outperforms the ML method when δ is larger than 1.5 in both estimation (Table 4-7 and Table 4-8) and detection of significant QTL (Figure 4-1). The simulation results clearly demonstrate that the penalized method is preferred to the standard ML method when heterogeneous variances exist.

Table 4-9: Power of two methods in LD mapping

delta	δ = 0		δ = 0.5		δ = 1		δ =1.5		δ = 2	
N	Pen	ML	Pen	ML	Pen	ML	Pen	ML	Pen	ML
500	0.09	0.15	0.06	0.17	0.18	0.28	0.33	0.29	0.67	0.32
1000	0.22	0.26	0.29	0.37	0.52	0.53	0.72	0.52	0.93	0.51
1500	0.28	0.35	0.36	0.39	0.54	0.53	0.83	0.54	0.98	0.54
2000	0.37	0.53	0.55	0.62	0.84	0.71	0.98	0.69	1	0.64

LD mapping under when δ = 0,0.5,1,1.5,2, the Penalized ML and standard ML's power of detecting a QTL using an F2 population of sample size N = 500, 1000, 1500, 2000 for the phenotypic data simulated from Gaussian mixture distributions with 3 components. The results are calculated from 100 simulations.

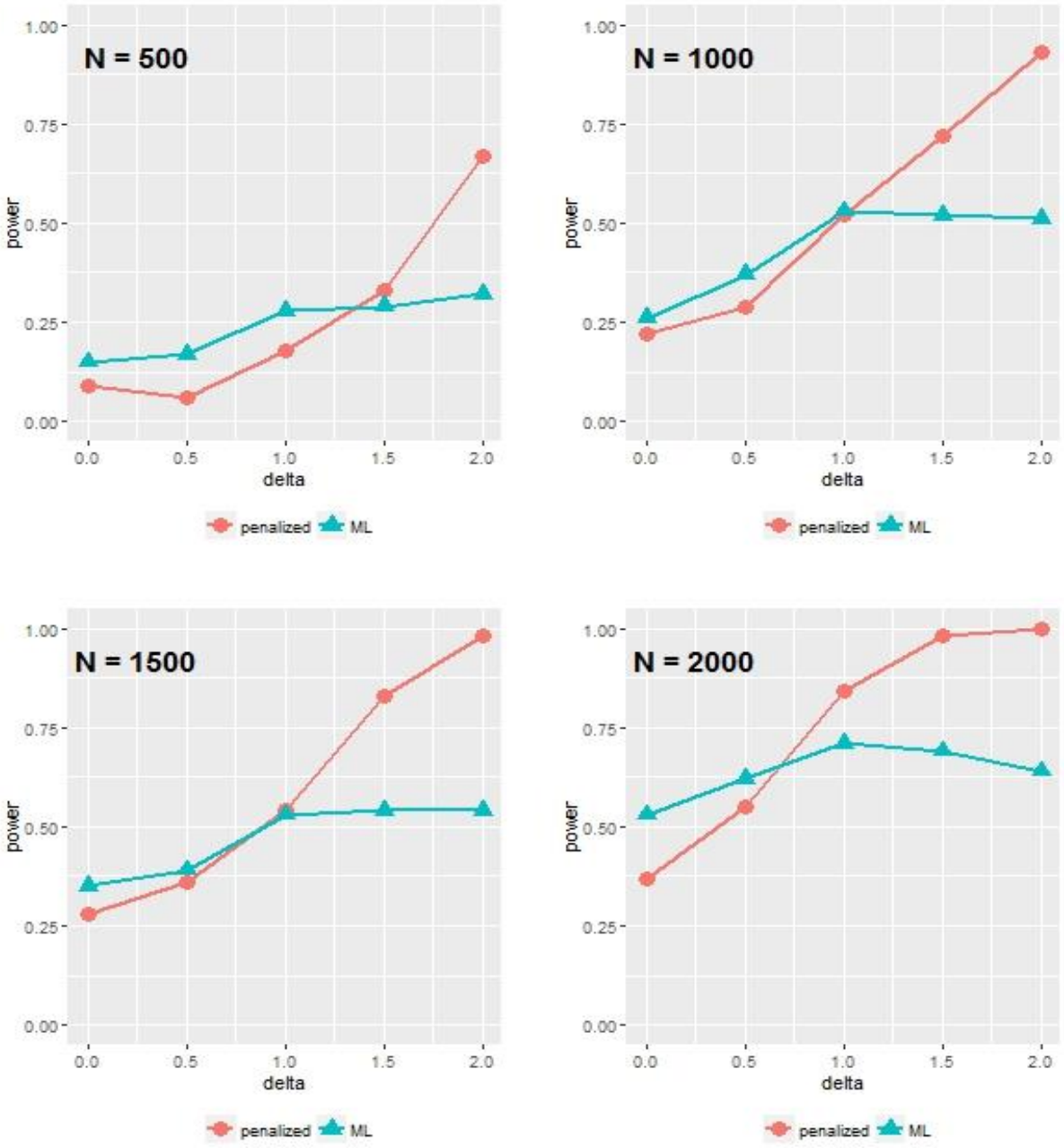


Figure 4-1: Power of Penalized and Standard ML of detecting QTL in LD mapping. Phenotypic data is simulated from Gaussian mixture distributions with 3 components. The results are calculated from 100 simulations under scenarios $\delta = 0,0.5,1,1.5,2$ with sample size N = 500, 1000, 1500, 2000.

Chapter 5 Discussion and Future Work

In the current work, we mainly adopted the penalized model for the F2 and natural populations. This method may also be extended to a more complex sampling schema, for example that involves cluster sampling, which may induce correlation between individual observations within sampled clusters. For example, the case-parent triad design is commonly used in genetic association studies, and in this design, N families with two parents and one child are recruited and families are assumed to independent of each other. For a locus with allele A and a, there are 27 possible genotype combinations among father, mother and child. When the genotypes are not linked with sex chromosome (i.e. genotypes of father and mother are exchangeable), these 27 combinations can be reduced to ten groups. Let p be the frequency of allele A and q be the frequency of allele a. the joint genotype distribution of parents-child triads are given in table 5-1.

Table 5-1: Joint genotype distribution of parents-child triads

	Genotype		
	Parents	Child	Joint probability
1	AA-AA	AA	p^4
2	AA-Aa	AA	$2p^3q$
3	AA-Aa	Aa	$2p^3q$
4	AA-aa	Aa	$2p^2q^2$
5	Aa-Aa	AA	p^2q^2
6	Aa-Aa	Aa	$2p^2q^2$

7	Aa-Aa	Aa	p^2q^2
8	Aa-aa	Aa	$2pq^3$
9	Aa-aa	aa	$2pq^3$
10	aa-aa	aa	q^4

One future direction is to explore how the penalized method introduced in this thesis can be used for genetic mapping in these types of designs.

Reference

1. Remington, D.L. and M.D. Purugganan, *Candidate genes, quantitative trait loci, and functional trait evolution in plants.* International Journal of Plant Sciences, 2003. **164**(S3): p. S7-S20.
2. Radonić, A., et al., *Guideline to reference gene selection for quantitative real-time PCR.* Biochemical and biophysical research communications, 2004. **313**(4): p. 856-862.
3. Roff, D.A., *A centennial celebration for quantitative genetics.* Evolution, 2007. **61**(5): p. 1017-1032.
4. Xiong, L., et al., *Identification of genetic factors controlling domestication-related traits of rice using an F2 population of a cross between Oryza sativa and O. rufipogon.* Theoretical and Applied Genetics, 1999. **98**(2): p. 243-251.
5. Lin, S., T. Sasaki, and M. Yano, *Mapping quantitative trait loci controlling seed dormancy and heading date in rice, Oryza sativa L., using backcross inbred lines.* Theoretical and Applied Genetics, 1998. **96**(8): p. 997-1003.
6. Casa, A.M., et al., *The MITE family Heartbreaker (Hbr): molecular markers in maize.* Proceedings of the National Academy of Sciences, 2000. **97**(18): p. 10083-10089.
7. Vignal, A., et al., *A review on SNP and other types of molecular markers and their use in animal genetics.* Genetics Selection Evolution, 2002. **34**(3): p. 1.
8. Henry, R.J., *Plant Conservation Genetics.* 2006.
9. Gupta, P. and S. Rustgi, *Molecular markers from the transcribed/expressed region of the genome in higher plants.* Functional & integrative genomics, 2004. **4**(3): p. 139-162.
10. Hara, J., et al., *Genetic ablation of orexin neurons in mice results in narcolepsy, hypophagia, and obesity.* Neuron, 2001. **30**(2): p. 345-354.

11. Broman, K.W. and S. Sen, *A Guide to QTL Mapping with R/qtl*. Vol. 46. 2009: Springer.

12. London, S.J., et al., *Prospective study of relative weight, height, and risk of breast cancer.* Jama, 1989. **262**(20): p. 2853-2858.

13. Coelho, C.M., et al., *Identification of quantitative trait loci that affect endoreduplication in maize endosperm.* Theoretical and Applied Genetics, 2007. **115**(8): p. 1147-1162.

14. Lynch, M. and B. Walsh, *Genetics and analysis of quantitative traits*. Vol. 1. 1998: Sinauer Sunderland, MA.

15. Wu, R., et al., *Functional mapping of quantitative trait loci that interact with the hg mutation to regulate growth trajectories in mice.* Genetics, 2005. **171**(1): p. 239-249.

16. Soller, M., T. Brody, and A. Genizi, *On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines.* Theoretical and applied genetics, 1976. **47**(1): p. 35-39.

17. Cowen, N., *Multiple linear regression analysis of RFLP data sets used in mapping QTLs.* Development and application of molecular markers to problems in plant genetics, 1989: p. 113-116.

18. Moreno-Gonzalez, J., *Genetic models to estimate additive and non-additive effects of marker-associated QTL using multiple regression techniques.* Theoretical and Applied Genetics, 1992. **85**(4): p. 435-444.

19. Weller, J., *Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers.* Biometrics, 1986: p. 627-640.

20. Weller, J., *Mapping and analysis of quantitative trait loci in Lycopersicon (tomato) with the aid of genetic markers using approximate maximum likelihood methods.* Heredity, 1987. **59**(3): p. 413-421.

21. Lander, E.S. and D. Botstein, *Mapping mendelian factors underlying quantitative traits using RFLP linkage maps.* Genetics, 1989. **121**(1): p. 185-199.

22. Jansen, R.C., *Interval mapping of multiple quantitative trait loci.* Genetics, 1993. **135**(1): p. 205-211.

23. Jansen, R.C. and P. Stam, *High resolution of quantitative traits into multiple loci via interval mapping.* Genetics, 1994. **136**(4): p. 1447-1455.

24. Zeng, Z.-B., *Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci.* Proceedings of the National Academy of Sciences, 1993. **90**(23): p. 10972-10976.

25. Zeng, Z.-B., *Precision mapping of quantitative trait loci.* Genetics, 1994. **136**(4): p. 1457-1468.

26. Kao, C.-H., Z.-B. Zeng, and R.D. Teasdale, *Multiple interval mapping for quantitative trait loci.* Genetics, 1999. **152**(3): p. 1203-1216.

27. Kao, C.-H. and Z.-B. Zeng, *Modeling epistasis of quantitative trait loci using Cockerham's model.* Genetics, 2002. **160**(3): p. 1243-1261.

28. Liu, B. and S. Knapp, *Computational tools for study of complex traits.* Molecular Dissection of Complex Traits, 1997: p. 43.

29. Hoeschele, I., et al., *Advances in statistical methods to map quantitative trait loci in outbred populations.* Genetics, 1997. **147**(3): p. 1445-1457.

30. Crosses, E., *Review of statistical methods for QTL mapping in experimental crosses.* Lab animal, 2001. **30**(7).

31. Broman, K.W. and S. Sen, *A Guide to QTL Mapping with R/qtl*. 2009: Springer.

32. Kruglyak, L. and E.S. Lander, *A nonparametric approach for mapping quantitative trait loci.* Genetics, 1995. **139**(3): p. 1421-1428.

33. Wu, S., et al., *Genetic mapping of complex traits by minimizing integrated square errors.* BMC genetics, 2012. **13**(1): p. 20.

34. Xu, S. and W.R. Atchley, *A random model approach to interval mapping of quantitative trait loci.* Genetics, 1995. **141**(3): p. 1189-1197.

35. Pearson, K., *Contributions to the mathematical theory of evolution.* Philosophical Transactions of the Royal Society of London. A, 1894. **185**: p. 71-110.

36. Roberts, S.J., et al., *Bayesian approaches to Gaussian mixture modeling.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(11): p. 1133-1142.

37. Kormylo, J. and J. Mendel, *Maximum likelihood detection and estimation of Bernoulli-Gaussian processes.* IEEE transactions on information theory, 1982. **28**(3): p. 482-488.

38. McLachlan, G.J. and K.E. Basford, *Mixture models. Inference and applications to clustering.* Statistics: Textbooks and Monographs, New York: Dekker, 1988, 1988. **1**.

39. Biernacki, C., G. Celeux, and G. Govaert, *Assessing a mixture model for clustering with the integrated completed likelihood.* IEEE transactions on pattern analysis and machine intelligence, 2000. **22**(7): p. 719-725.

40. Ridolfi, A., *Maximum likelihood estimation of hidden Markov model parameters, with application to medical image segmentation*. 1997.

41. Idier, J., Y. Goussard, and A. Ridolfi, *Unsupervised Image Segmentation Using a Telegraph Parameterization of Pickard Random Field*, in *Spatial Statistics: Methodological Aspects and Applications*. 2001, Springer. p. 115-140.

42. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the royal statistical society. Series B (methodological), 1977: p. 1-38.

43. Hero, A.O. and J.A. Fessler, *Convergence in norm for alternating expectation-maximization (EM) type algorithms.* Statistica Sinica, 1995. **5**(1): p. 41-54.

44. McLachlan, G., *Peel., D.(2000). Finite mixture models*. New York: Wiley.

45. Kiefer, J. and J. Wolfowitz, *Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters.* The Annals of Mathematical Statistics, 1956: p. 887-906.

46. Lehmann, E., *Theory of Point Estimation, John Willey & Sons*. 1983, Inc.

47. Peters, J., B Charles and H.F. Walker, *An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions.* SIAM Journal on Applied Mathematics, 1978. **35**(2): p. 362-378.

48. Kiefer, N.M., *Discrete parameter variation: Efficient estimation of a switching regression model.* Econometrica: Journal of the Econometric Society, 1978: p. 427-434.

49. Redner, R., *Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions.* The Annals of Statistics, 1981: p. 225-228.

50. Redner, R.A. and H.F. Walker, *Mixture densities, maximum likelihood and the EM algorithm.* SIAM review, 1984. **26**(2): p. 195-239.

51. Hathaway, R.J., *A constrained formulation of maximum-likelihood estimation for normal mixture distributions.* The Annals of Statistics, 1985: p. 795-800.

52. Hathaway, R.J., *A constrained EM algorithm for univariate normal mixtures.* Journal of Statistical Computation and Simulation, 1986. **23**(3): p. 211-230.

53. Wu, R., C. Ma, and G. Casella, *Statistical genetics of quantitative traits: linkage, maps and QTL*. 2007: Springer Science & Business Media.

54. Mather, K., *Crossing-over.* Biological Reviews, 1938. **13**(3): p. 252-292.

55. Haldane, J., *The combination of linkage values and the calculation of distances between the loci of linked factors.* J Genet, 1919. **8**(29): p. 299-309.

56. Kosambi, D.D., *The estimation of map distances from recombination values.* Annals of eugenics, 1943. **12**(1): p. 172-175.

57. Fisher, S.R.A., et al., *The design of experiments.* 1960.

58. Vaughn, T.T., et al., *Mapping quantitative trait loci for murine growth: a closer look at genetic architecture.* Genetical research, 1999. **74**(03): p. 313-322.

59. Lewontin, R. and K.-i. Kojima, *The evolutionary dynamics of complex polymorphisms.* Evolution, 1960: p. 458-472.

60. Lewontin, R., *The interaction of selection and linkage. I. General considerations; heterotic models.* Genetics, 1964. **49**(1): p. 49-67.