UNIVERSITY OF CALIFORNIA
RIVERSIDE

Bayesian Statistics and Its Application to Quantitative Trait Loci Mapping

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Xiaohong Che

December 2011

Dissertation Committee:

Dr. Shizhong Xu , Chairperson
Dr. Jun Li
Dr. James Flegal

Copyright by
Xiaohong Che
2011

The Dissertation of Xiaohong Che is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to take this opportunity to express my sincere gratitude to my advisor, Prof. Shizhong Xu, for his professional advice, guidance and help. He has provided me with numerous valuable ideas, insights, encouragements and comments. He has also been very understanding and supportive. I am very fortunate to have him as my advisor.

I also would like to express my appreciation to Dr. Jun Li for serving on my oral exam committee and dissertation committee, Dr. James Flegal for serving on my dissertation exam committee, Dr. Changxuan Mao, Dr. Daniel Jeske, Dr. Laosheng Wu, and Dr. Mohsen El Hafsi for serving on my oral exam committee. A sincere thank goes to Statistics Department of University of California, Riverside, for its support for me during the past five years. And many thanks go to all the faculty and staff members and my dear classmates in the statistics department. I have a wonderful time in the department as a student there. A special thank goes to all my colleagues in Prof. Xu's lab for their help during my past three years there.

The text of this dissertation, in part or in full, are reprints of the materials as they appear in "Bayesian Data Analysis for Agricultural Experiments"(*Canadian Journal of Plant Science*, 90: 575-603) and "Significance Test and Genome Selection in Bayesian Shrinkage Analysis"(*International Journal of Plant Genomics*, Vol 2010, Article ID 893206, 11 pages, doi:10.1155/2010/893206). The co-author, Dr. Shizhong Xu, listed in the publication directed and supervised the researches which form the basis of this dissertation.

Most of all, I would like to thank God and my family for their encouragements and support during the past so many years!

To my family for all the support.

ABSTRACT OF THE DISSERTATION

Bayesian Statistics and Its Application to Quantitative Trait Loci Mapping

by

Xiaohong Che

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, December 2011
Dr. Shizhong Xu , Chairperson

Quantitative trait loci (QTL) mapping is one of the applications of statistics in genetics. This dissertation focuses two problems on QTL mapping which include a new permutation method used to find the thresholds for the shrinkage Bayesian estimation of quantitative trait loci parameters and three algorithms of handling the missing genotype problems in multiple QTL mapping under the generalized linear mixed model framework. In addition, this dissertation includes a review on Bayesian statistics and some data analyses using Markov chain Monte Carlo (MCMC).

Chapter 2 is a review of the Bayesian statistics and some data analyses using MCMC. It includes almost all the aspects of Bayesian statistics such as Bayes' theorem, prior and posterior distributions, Bayesian inference, and Markov chain Monte Carlo (MCMC) algorithms.

In Chapter 3, a new way to conduct the permutation test under the Shrinkage Bayesian method is developed. Permutation test is the most frequently used method for statistical test for QTL mapping. And it was applied on the QTL mapping based on the Bayesian approach. While using the traditional permutation test to get the thresholds for QTL mapping from the MCMC algorithms in the Bayesian models is

quite time-consuming, a new way to permute the samples from the MCMC algorithms is performed in Chapter 3. Empirical power analysis is done to test the method through the simulations.

Generalized linear mixed model has been applied to analyze the discrete traits. Research on handling the missing genotype problems in multiple QTL mapping under the generalized linear mixed model framework is presented in Chapter 4. Three algorithms were proposed: (1) expectation algorithm, (2) overdispersion model algorithm and (3) mixture model algorithm.

Contents

List of Figures **xi**

List of Tables **xiv**

1 Introduction **1**
1.1 An introduction to Quantitative Trait Loci Mapping 1
1.2 Brief review of Quantitative Trait Loci Mapping Techniques 2
1.3 Bayesian Methods and Quantitative Trait Loci Mapping 4
1.3.1 Bayesian Statistics and Data Analysis 4
1.3.2 Bayesian Shrinkage Method and Permutation Test for Quantitative Trait Loci Mapping 5
1.3.3 Quantitative Trait Loci Mapping based on Generalized Linear Mixed Models 7

2 Bayesian Data Analysis for Agricultural Experiments **9**
2.1 Introduction 9
2.2 Theory 11
2.2.1 Bayes' Theory 11
2.2.2 Prior distribution 14
2.2.2.1 Informative and uninformative priors 14
2.2.2.2 Proper and improper priors 15
2.2.2.3 Conjugate prior 15
2.2.2.4 Jeffreys' prior 16
2.2.3 Posterior distribution 17
2.2.4 Bayesian inference 22
2.2.4.1 Point estimation 23
2.2.4.2 Hypothesis testing 24
2.2.4.3 Credibility set 25
2.3 MCMC Algorithm 26
2.3.1 Gibbs sampler 28
2.3.2 Metropolis algorithm 29
2.3.3 Metropolis-Hastings algorithm 31
2.3.4 Assessment of Markov chain convergence 33
2.3.4.1 Burn-in and thinning 33
2.3.4.2 Visual analysis of trace plots 34
2.3.4.3 Statistical diagnosis for convergence 36

2.3.4.4 Autocorrelation . . . 37
2.3.4.5 Effective sample size (ESS) . . . 39
2.3.5 Post MCMC analysis . . . 39
2.3.5.1 Posterior sample and marginal posterior distribution . . . 40
2.3.5.2 Summary statistics . . . 41
2.4 Software packages . . . 43
2.4.1 WinBUGS . . . 43
2.4.2 PROC MCMC . . . 45
2.5 Data analysis . . . 46
2.5.1 The damage data . . . 46
2.5.1.1 Model . . . 47
2.5.1.2 Prior and posterior . . . 48
2.5.1.3 SAS code . . . 51
2.5.1.4 Result . . . 55
2.5.2 The seeds data . . . 60
2.5.2.1 Model . . . 62
2.5.2.2 Prior and posterior . . . 63
2.5.2.3 SAS code . . . 64
2.5.2.4 Result . . . 66
2.5.3 The fertility data . . . 71
2.5.3.1 Model . . . 72
2.5.3.2 Prior and posterior . . . 73
2.5.3.3 SAS code . . . 74
2.5.3.4 Result . . . 77
2.6 Discussion . . . 84

3 Significance Test and Genome Selection in Bayesian Shrinkage Analysis 86
3.1 Introduction . . . 86
3.2 Methods . . . 89
3.2.1 Model . . . 89
3.2.2 Permutation between Markov chains . . . 91
3.2.3 Permutation within Markov chain . . . 92
3.2.4 Genome selection . . . 93
3.3 Results and Discussion . . . 94
3.3.1 Simulation study . . . 94
3.3.2 Permutation outside Markov chain . . . 99
3.3.3 Permutation inside Markov chain . . . 99
3.3.4 Power analysis . . . 103
3.3.5 False positive rate . . . 105
3.3.6 Cross validation for genome selection . . . 107
3.3.7 Real data analysis . . . 108
3.3.7.1 Arabidopsis data . . . 108
3.3.7.2 Barley data . . . 112
3.3.7.3 Wheat data . . . 115
3.4 Discussion . . . 118

4 Generalized Linear Mixed Models for Mapping Quantitative Trait Loci **121**
4.1 Introduction . . . 121
4.2 Methods . . . 124
4.2.1 Generalized linear mixed model . . . 124
4.2.2 Missing genotypes . . . 127
4.2.2.1 Expectation model . . . 128
4.2.2.2 Overdispersion model . . . 129
4.2.2.3 Mixture model . . . 129
4.3 Application . . . 130
4.3.1 Simulation study . . . 130
4.3.1.1 Binomial data . . . 130
4.3.1.2 Binary data . . . 134
4.3.2 Mapping wheat fertility QTL . . . 136
4.3.2.1 Binomial trait . . . 136
4.3.2.2 Binary trait . . . 140
4.3.2.3 Truncated binomial trait . . . 143
4.4 Discussion . . . 145

Bibliography **148**

List of Figures

2.1 Trace plots for Markov chain convergence diagnosis for parameter theta (θ). (a) The Markov chain quickly reaches the stationary distribution. (b) The Markov chain mixes poorly (high autocorrelation). (c) The Markov chain does not converge at all. 36
2.2 Auto correlation plots against lag: (a) Low autocorrelation, (b) High autocorrelation. 38
2.3 The posterior TAD panels (trace, autocorrelation and density) for parameter beta (β) of the damage data. 57
2.4 The posterior TAD panels for parameter rho ($\rho = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$) of the damage data. 58
2.5 The posterior TAD panels for parameter gamma1 (γ_1) of the seeds data. 68
2.6 The posterior TAD panels for parameter gamma2 (γ_2) of the seeds data. 69
2.7 The posterior TAD panels for parameter gamma3 (γ_3) of the seeds data. 70
2.8 The posterior TAD panels for parameter beta (β) of the fertility data. . 78
2.9 Posterior mean of QTL effect (panel a) and LOD score (panel b) plotted against the genome location of the wheat fertility trait (the fertility data) from the MCMC implemented Bayesian analysis (multiple QTL model). The five chromosomes are separated by the dotted reference lines. . . . 79
2.10 The estimated QTL effect (panel a) and LOD score (panel b) plotted against the genome location of the wheat fertility trait (the fertility data) from the maximum likelihood analysis implemented in the generalized linear model (single QTL model). The five chromosomes are separated by the dotted reference lines. 81
2.11 The trace plot (panel a) and the posterior density (panel b) of the QTL detected in chromosome 2 for the binary fertility trait of wheat (the fertility data). 83

3.1 The true and estimated QTL effects for the entire genome of the simulated data. (a) The true positions and effects of the simulated QTL, (b) The estimated positions and effects of QTL using the Bayesian shrinkage method. 97
3.2 Posterior distribution of QTL number 1 of the simulation experiment. The true effect of the simulated QTL is 4.47. There is a high probability mass at value zero, even though this is the largest QTL out of the 20 QTL simulated. 98

3.3 Empirical threshold values generated from permutation analysis and the estimated QTL effects (simulated data). Empirical threshold values generated from permutation analysis at $\alpha = 0.05$ (2.5%-97.5%) and $\alpha = 0.10$ (5%-95%) along with the estimated QTL effects (simulated data). Percentiles for the 2.5%-97.5% interval are plotted against the genome location as dashed lines (wider interval). Percentiles of the 5%-95% interval are plotted against the genome location as solid lines (narrower interval). (a) Shows the result of "permutation outside the Markov chain", (b) Result of "permutation within the Markov chain" with phenotype reshuffling in every iteration ($h = 1$). 100
3.4 Empirical threshold values generated from "permutation within Markov chain" and the estimated QTL effects (simulated data). Empirical threshold values generated from "permutation within Markov chain" analysis at $\alpha = 0.05$ (2.5%-97.5%) and $\alpha = 0.10$ (5%-95%) along with the estimated QTL effects (simulated data). Percentiles for the 2.5%-97.5% interval are plotted against the genome location as dashed lines (wider interval). Percentiles of the 5%-95% interval are plotted against the genome location as solid lines (narrower interval). (a) Phenotype reshuffling in every 5 iterations ($h = 5$), (b) Phenotype reshuffling in every 10 iterations ($h = 10$), (c) Phenotype reshuffling in every 100 iterations ($h = 100$). 102
3.5 Empirical statistical power for the simulated QTL. Empirical statistical powers for the simulated QTL obtained from 100 replicated experiments. (a) Statistical powers at Type I error of $\alpha = 0.05$; (b) Statistical power at Type I error of $\alpha = 0.10$. 104
3.6 False positive rate profiles for the simulated markers obtained from 100 replicated experiments. (a) False positive rate at $\alpha = 0.05$; (b) False positive rate at $\alpha = 0.10$. 106
3.7 Prediction error (PE) plotted against the Type I error for the simulated data. The squared prediction error (PE) plotted against the Type I error obtained from the five-fold cross validation test for the simulated data. . 108
3.8 Result of the Arabidopsis data analysis. (a) The upper panel shows the estimated QTL effects for the entire genome and the empirical thresholds drawn from permutation within the Markov chain analysis at $\alpha = 0.05$ (2.5%-97.5%, wider interval) and $\alpha = 0.10$ (5%-95%, narrower interval). (b) The lower panel shows the plot of the squared prediction error (PE) against the Type I error obtained from the five-fold cross validation test. 111
3.9 Result of the barley data analysis. (a) The upper panel shows the estimated QTL effects for the entire genome and the empirical thresholds drawn from permutation within the Markov chain analysis at $\alpha = 0.05$ (2.5%-97.5%, wider interval) and $\alpha = 0.10$ (5%-95%, narrower interval). (b) The lower panel shows the plot of the squared prediction error (PE) against the Type I error obtained from the cross validation test. 114
3.10 Result of the wheat data analysis. (a) The upper panel shows the estimated QTL effects for the entire genome and the empirical thresholds drawn from permutation within the Markov chain analysis at $\alpha = 0.05$ (2.5%-97.5%, wider interval) and $\alpha = 0.10$ (5%-95%, narrower interval). (b) The lower panel shows the plot of the squared prediction error (PE) against the Type I error obtained from the cross validation test. 117

4.1 True QTL effects (top panel) and their estimated values for the simulated binomial trait using the expectation model (panel in the middle) and overdispersion model (bottom panel). The estimated QTL effects are the averages of 1000 replicated samples. The positions of 20 simulated QTL are indicated by the inward ticks on the horizontal axis. 133
4.2 The estimated QTL effects for the simulated binary trait using the mixture model (top panel), expectation model (panel in the middle) and overdispersion model (bottom panel). The estimated QTL effects are the averages of 1000 replicated samples. The positions of 20 simulated QTL are indicated by the inward ticks on the horizontal axis. 135
4.3 Binomial trait analysis of the wheat experiment using the expectation model (blue) and the overdispersion model (red). The top panel shows the estimated QTL effects and the bottom panel shows the LOD scores. Chromosomes are separated by the dotted vertical lines. Positions of true markers are indicated by the inward ticks on the horizontal axis. 138
4.4 Frequency distribution of the number of seeded spikelets of the F_2 wheat population. Among the 243 plants, 39 of them have no seeds (zero category). 141
4.5 Binary trait (seed presence/absence) analysis using the expectation model (blue), overdispersion model (red) and mixture model (black). The top panel shows the estimated QTL effects and the bottom panel shows the LOD scores. Chromosomes are separated by the dotted vertical lines. Positions of true markers are indicated by the inward ticks on the horizontal axis. 142
4.6 Zero-truncated binomial trait (excluding plants with no seeds) analysis using the expectation model (blue) and overdispersion model (red). The top panel shows the estimated QTL effects and the bottom panel shows the LOD scores. Chromosomes are separated by the dotted vertical lines. Positions of true markers are indicated by the inward ticks on the horizontal axis. 144

List of Tables

2.1 The average degree of damage caused by insects for four randomly selected wheat varieties. The original dataset was published by Milliken and Johnson (2009, p. 314). This dataset is called the "damage data" here in this study. 47
2.2 Summary statistics of the posterior sample for the damage data. 59
2.3 Diagnostic test statistics for the Markov chain convergence of the damage data. 60
2.4 Seed germination of two different varieties of *Orobanche cernua aegyptiaca* plants (O.a75 and O.a73) in two different host plants or root extracts (bean and cucumber). The data set was published by Crowder (1978) and called the "seeds data" in this study. The column headed "germinated" is the numbers of germinated seeds. The column headed "seed" is the number of seeds planted. The column headed "rate" is the proportion of the germinated seeds (number of germinated seeds / total number of seeds planted). 61
2.5 Summary statistics of the posterior sample for the seeds data. 67
2.6 Diagnostic test statistics for the Markov chain convergence of the seeds data. 67

3.1 QTL parameters used in the simulation experiment. 96

4.1 Comparison of the mean squared errors (MSE) for the three models in the replicated simulation study. 132
4.2 QTL detected for the binomial trait of wheat fertility using the overdispersion model. 139

Chapter 1

Introduction

The Quantitative traits loci (QTL) techniques were developed in the late 1980s. QTL mapping is the statistical study of the association between the alleles that occur in a locus and the phenotypes (physical forms or traits) that they produce. In this chapter, we first give an introduction to QTL mapping and a brief review of the QTL mapping techniques. Then, an outline and introduction to my research and analysis on QTL mapping is presented.

1.1 An introduction to Quantitative Trait Loci Mapping

In order to understand what Quantitative Trait Loci mapping is, first, we need to understand what quantitative traits are and what quantitative trait loci are. Quantitative traits refer to phenotypes (characteristics) that vary in degree and can be attributed to polygenic effects, i.e., product of two or more genes, and their environment. For example, the weight or the height of a person is the quantitative trait. The eye color of a person is not quantitative trait. Quantitative trait loci (QTLs) are stretches of DNA containing or linked to the genes that underlie a quantitative trait. QTL mapping

studies the alleles that occur in a locus and the phenotypes (physical forms or traits) that they produce. QTL identify a particular region of the genome as containing a gene that is associated with the trait being assayed or measured. The basic idea of applying statistical methods to QTL mapping has been clear since Sax (1923). And since late 1980s, there has been a resurgence of interest in the development of statistical models and algorithms for genetic mapping, aiming to improve the precision of QTL mapping and equip the models to suit different genetic designs (F_2/backcross or full-sib family), marker types (dominant, or codominant) or marker spaces (sparse or dense).

1.2 Brief review of Quantitative Trait Loci Mapping Techniques

As stated earlier, the quantitative traits loci techniques were developed in the late 1980s. The basic idea has been clear since Sax (1923). And there had been some traditional methods of using genetic markers to study quantitative trait loci (Thoday 1960; Jayakar 1970; Soller and Brody 1976; Tanksley, Medina-Filho and Rick 1982; Edwards, Stuber and Wendell 1987). These methods involve comparing the trait means of different markers. The difference of the trait means provides an estimate for the phenotypic effect. To test whether the inferred phenotypic effect is significantly different from 0, one applies some simple statistical methods such as t-test and simple or multiple regressions.

But the approach did not become possible in principle until the advent of restriction fragment length polymorphisms (RFLPs) as genetic markers. With the facility of this technology, Lander and Botstein (1989) proposed a much-improved method for QTL mapping, named interval mapping (IM). And there are various modified versions of this

approach (Jansen 1993; Zeng 1994).The basic idea of these methods is to divide the entire genome into a finite number of points 1 or 2 cM apart. These points are subject to statistical test and evaluation and thus called putative QTL. Their genotypes are not observable but can be inferred from the genotypes of flanking markers. Two flanking markers define an interval that may contain several putative QTL which explains why the method is called interval mapping. Interval mapping is a one-dimensional search algorithm since it tests one putative position at a time. So, if one wants to test a series putative position in the entire genome, multiple tests will be required. Lander and Botstein (1989) used the logarithm of the odds (to the base 10) (LOD) as the test statistics. From their method, one can get a smoothed plot of the test-statistic value against the genome position which forms a continuous profile. A significant peak in the profile indicates a QTL located in the neighborhood of the peak. Since it is a one-dimensional search algorithm, the interval mapping can only handle the models with a single QTL. It only considers the effects of the putative QTL at the current position in the model, and all the other QTL effects are ignored. Those QTL effects not included in the model are thrown into the residual error. In contrast to interval mapping, multiple interval mapping (MIM) (Jansen 1993; Zeng 1994) treats QTL effects ignored by the interval mapping as the background effects, which are absorbed by the selective markers (called cofactors) outside the tested interval. If the cofactors are chosen properly, MIM can substantially improve the efficiency of QTL mapping. When multiple QTL are chosen, one may has to rewrite the model to include all the significant intervals in a single multiple-QTL model and reestimate the QTL effects with QTL positions fixed at their estimated values (Yano et. al. 1997; Hunt et. al. 1999; Bunyamin et. al. 2002). This two-step approach may not be optimal and is being replaced by a one-step multiple-QTL mapping where the values and the locations of the QTL are estimated

simultaneously (Kao et. al 1999).

Multiple interval mapping has been the most popular QTL mapping procedure. But it has some disadvantages such as the difficulties in the implementation of the multiple interval mapping. Recently, people applied the Bayesian methods to QTL mapping. Section 1.3.1 and section 1.3.2 will give an introduction to the Bayesian shrinkage method used in QTL mapping and a permutation test based on Bayesian shrinkage method respectively

1.3 Bayesian Methods and Quantitative Trait Loci Mapping

1.3.1 Bayesian Statistics and Data Analysis

Bayes' theorem originated from Thomas Bayes (1763). It has the general form introduced by Pierre-Simon Laplace as follow:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{\sum P(A)P(B|A)} \tag{1.1}$$

Starting from this theorem, some formulated a branch of statistics called Bayesian statistics. In Bayesian statistics, ones have the Bayesian inference to do the point estimation, hypothesis testing, credibility set calculation, analogous to the statistical method. In Bayesian inference, each parameter θ is regarded as a random variable, and assigned a prior distribution which has the probability density $\pi(\theta)$. Combining the information from the prior distribution and the likelihood $p(y|\theta)$ for the data $y = \{y_1, ..., y_n\}$, one

can compute the posterior distribution $p(\theta|y)$ through the following expression:

$$p(\theta|y) = \frac{\pi(\theta)p(y|\theta)}{p(y)} = \frac{\pi(\theta)p(y|\theta)}{\int \pi(\theta)p(y|\theta)\mathrm{d}x} \tag{1.2}$$

And the inference about the parameter θ can be drawn from the posterior distribution. For example, the posterior mean can be calculated by:

$$\hat{\theta} = E(\theta|y) = \int \theta p(\theta|y)\mathrm{d}\theta \tag{1.3}$$

The posterior variance can be calculated by:

$$var(\theta|y) = \int (\theta - \hat{\theta})^2 p(\theta|y)\mathrm{d}\theta \tag{1.4}$$

The above examples show that Bayesian inference involves integration. In order to numerically approximate high dimensional integrals in some complicated Bayesian inference cases, some developed Markov chain Monte Carlo (MCMC) algorithms. To date, two softwares, WinBUGS and PROC MCMC in SAS can perform MCMC algorithms for Bayesian analysis. We also demonstrated the implementation of the MCMC algorithm using professional software package-the MCMC procedure in SAS. Three data sets from agricultural experiments were analyzed to demonstrate the MCMC algorithm in Chapter 2. Also, in Chapter 2, there is a review about Bayesian statistics.

1.3.2 Bayesian Shrinkage Method and Permutation Test for Quantitative Trait Loci Mapping

Bayesian shrinkage method was satisfactorily applied to QTL mapping by Xu (2003). In this method, each marker is treated as a putative QTL and thus included in the model

as a variable. All the markers are analyzed simultaneously. The data are analyzed by using a Bayesian framework under a random regression coefficient model. In the model, each gene effect is assigned a normal prior with mean zero and a unique variance. The effect-specific prior variance is further assigned a vague prior so that the variance can be estimated from the data. This method can resolve two problems in the QTL mapping. First, it can handle a single model with a large number of markers. Second, it can deal with some close-to-zero gene effect in the model. This method will be briefly described as follows based on the backcross design:

The linear model for the phenotypic value y_i of the ith individual is:

$$y_i = b_0 + \sum_{j=1}^{p} x_{ij} b_j + e_i \tag{1.5}$$

where y_i is the phenotypic value for the individual i, p is the total number of markers in the entire genome, x_{ij} is a dummy variable indicating the genotype of the jth marker for individual i, b_j is the QTL effect associated with marker j, and e_i is the residual error with a $N(0, \sigma_0^2)$ distribution.

The Bayesian framework chooses the following prior distribution, $p(b_0) \propto 1$, $p(\sigma_0^2) \propto 1/\sigma_0^2$, $p(b_j) = N(0, \sigma_j^2)$, and $p(\sigma_j^2) \propto 1/\sigma_j^2$ for $j = 1, ..., p$. The joint prior of the unobservable $p(b, \nu)$ takes the product of the priors of individual parameters. The likelihood is $p(y|b, \nu) = \prod_{i=1}^{n} p(y_i|b, \sigma_0^2) \propto (\sigma_0^2)^{-n/2} exp\{-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{p} x_{ij} b_j^2)\}$. The joint posterior distribution has a form of $p(b, \nu|y) \propto p(y|b, \nu) p(b, \nu)$.

The Bayesian analysis was implemented via the Markov chain Monte Carlo (MCMC). The sampling is performed by sampling the unobservables from the above joint posterior distribution in certain sequences.

One of the challenges in the QTL mapping is the problem of setting up the sta-

tistical tests for the QTL effects. More specifically, we need to set a standard for the QTL effects, any estimated QTL effect beyond the standard considered to be significant. Permutation test (Churchill and Doerge 1994) is the most commonly used method for statistical test for QTL effect. It is very efficient in interval mapping under the maximum likelihood framework. Kopp et al. (2003) applied the permutation test to determine empirical thresholds for Bayesian shrinkage mapping. The problem with such a test for the MCMC implemented Bayesian mapping is the heavy computational burden. Each MCMC run may take one or a few hours to complete for a reasonable size mapping population. Performing thousands of permutation analyses is not realistic for the Bayesian method. To solve this problem, I conduct a within-chains permutation test under the shrinkage Bayesian framework. This procedure is less time-consuming, and according to the empirical power and false discovery rate analyses, the method turns out to be efficient. The permutation test based on the above Bayesian shrinkage method for QTL mapping is presented in Chapter 3.

1.3.3 Quantitative Trait Loci Mapping based on Generalized Linear Mixed Models

The Bayesian shrinkage QTL analysis is based on the normal error model which assumes the normal distribution for the traits. But in the reality, there are some traits which are not normally distributed, especially some discrete traits, like disease. For the QTL mapping of these non-normally distributed traits, the earliest analysis can be found in Kruglyak & Lander (1995) and Hackett & Weller (1995) (Henceforth abbreviated KL and HW). Basically two methods were proposed: nonparametric approach (KL) for continuous and categorical data and the logistic regression approach for ordinal data (HW). As for the logistic regression approach for ordinal data (HW), a similar analysis

can be found in a more recent paper (Rao and Xu 1998). Rebaï (1997) compared these two methods with a simulation approach.

Besides the nonparametric approaches (Kruglyak and Lander 1995), previously, most of these types of problems were analyzed based on the generalized linear models (Nelder and Wedderburn 1972). Lately, with the development of the generalized linear mixed models which is an extension of the generalized linear models, the non-normally distributed traits can be analyzed with more advanced models, i.e. generalized linear mixed models.

Generalized linear mixed models have been well developed till now, and have been applied to QTL mapping (Yi and Banerjee 2009). Three algorithms for the missing genotype problems in multiple QTL mapping under the generalized linear mixed model framework were proposed, which are (1) expectation algorithm, (2) overdispersion model algorithm and (3) mixture model algorithm. And the three methods were compared through simulations. The above work is presented in Chapter 4.

Chapter 2

Bayesian Data Analysis for Agricultural Experiments

2.1 Introduction

Bayesian statistics is a branch of statistics that originated from the Bayes' theorem developed by Thomas Bayes (1763) . The idea of Bayes' theorem first appeared in the publication -"A letter to John Canton" in 1763 where the author proved a special case of what is now called the Bayes' theorem. Stimulated by this idea, Pierre-Simon Laplace introduced a general version of the Bayes' theorem and used this theorem to solve problems in celestial mechanics, medical statistics, reliability and jurisprudence.

Nowadays, more and more attention has been given to the Bayesian statistics. It has been applied to many fields, including genetics (Beaumont et al. 2002; Beerli 2006; Efron and Tibshirani 2002; Hoeschele and VanRaden 1993; Holsinger and Wallace 2004; Murphy and Mutalik 1969; Sorensen and Gianola 2002; Xu 2003; Yi and Shriner 2008; Yi and Xu 2008) and medicine (Carlin et al. 1993; Desouza 1991; Halperin et al. 1990; Heitjan 1997; Loke et al. 2006; Palmer and Muller 1998; Racine et al. 1986; Spiegel-

halter et al. 2004; Turner et al. 2001; Wakefield and Racine-Poon 1995) . Numerous publications appeared in various different forms, such as tutorials, reviews and books (Besag et al. 1995; Casella and George 1992; Chib and Greenberg 1995; Kass et al. 1998) . Some of the review papers emphasized the theory of Bayesian statistics, which seemed to be too theoretical to applied scientists, and others gave brief summaries on the applications of Bayesian statistics to some special areas.

In this review, we introduced the basic concept of Bayesian statistics, the Markov chain Monte Carlo (MCMC) algorithm, the convergence diagnosis and the post MCMC summary statistics. A complete Bayesian analysis requires three steps: (1) statistical modeling, (2) scientific computing and (3) model checking. This review focuses on the first two steps, leaving the third step to the references, e.g., Dey et al. (2000) and Gelman et al. (2005). We also introduced two professional software packages commonly used for Bayesian analysis, which are WinBUGS and PROC MCMC. Three examples were used to demonstrate the applications of the Bayesian statistics to agricultural experiments. The first example was the "damage data" analysis, modeled using the simple ANOVA under the random model framework. This data analysis can be done easily with the maximum likelihood method. The Bayesian analysis gives an empirical posterior sample for each parameter of interest and the empirical posterior sample for any function of the parameters. The second example presents the Bayesian analysis for the "seeds data" under the generalized linear model framework using the logit link function. The difficult level of this data analysis is intermediate because maximum likelihood method can still be used to generate very similar result. The third data set was the "fertility data" collected from a QTL mapping experiment for wheat. The difficult level is high due to the large data set and the large model. We fit all the 75 pseudo markers in a single generalized linear mixed model. The maximum likelihood method may not be able to

fit that many effects and variance components to a single model. This will demonstrate the advantage of the Bayesian analysis over the maximum likelihood method.

The purpose of this review article is not to criticize the maximum likelihood (ML) method in favor of the Bayesian. Different methods have different pros and cons and users have their freedom to choose their favorite methods. The main advantage of Bayesian analysis is the ability to handle complicated models by taking advantage of the high power computers. The downside of the method is the intensive computational cost. The ML method provides hypothesis tests for the parameters of interest. Users are provided with a clear recommendation whether a parameter is real or not. In Bayesian analysis, however, users are provided with a posterior sample for each parameter, not a clear cut recommendation about the parameter. It is up to the users to decide whether a parameter is "significant" or not. This article targets the group of people (non-statisticians) who intend to use the Bayesian method to analyze their data but do not understand exactly what the Bayesian method is and how the method is implemented. The review is by no mean exhaustive but provides a simple guidance from which more advanced topics may be accessible. The three sample data analyses may be particularly helpful to give the users a quick start for the Bayesian tour.

2.2 Theory

2.2.1 Bayes' Theory

Bayes' theorem originated from Thomas Bayes (1763) who proved a special case of the theorem. Pierre-Simon Laplace, one of the main developers of Bayesian statistics,

introduced the following general form of the Bayes' theorem,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{\sum P(A)P(B|A)} \tag{2.1}$$

where $P(A|B)$ is the conditional probability of A given B, $P(B|A)$ is the conditional probability of B given A, $P(A)$ is the prior probability of A (before B is observed) and $P(B) = \sum P(A)P(B|A)$ is the marginal probability of B(acting as a normalizing constant). The summation is taken with respect to all possible events of A. This version of the Bayes' theorem applies to discrete variables.

Bayesian inference was formulated based on the above Bayes' theorem for the conditional probability for continuous variables. Unlike the frequentist (or classical) method in which the parameters are treated as fixed but unknown constants, the Bayesian method offers an alternative approach, that is to treat all parameters as random variables. Suppose that we are interested in estimating θ from data $y = \{y_1, ..., y_n\}$ using the following probability density $p(y|\theta)$. In Bayesian inference, we treat θ as a random variable because it cannot be determined exactly. We use a probability statement to describe the uncertainty of θ. This probability density is called the prior distribution. For example, we may say that the prior distribution of θ is normally distributed with mean 0 and variance 1, if it is believed that this distribution best describes the uncertainty associated with the parameter. Again, the prior distribution is not a true distribution, but a distribution that represents our lack of knowledge about the parameter.

In general, Bayesian inference follows three essential steps. First, a probability distribution for is assigned, denoted by $\pi(\theta)$, which is known as the prior distribution. This prior distribution expresses the prior belief of the investigator about the parameter before the data are examined. Second, a probability density for the data given the

parameter is chosen, denoted by $p(y|\theta)$, which is also called the model. Third, the prior belief about θ is updated by combining information from the prior distribution and the data through the calculation of the posterior distribution, denoted by $p(\theta|y)$. More specifically, in the third step, the prior and the model are combined through the Bayes' theorem in the following expression,

$$p(\theta|y) = \frac{\pi(\theta)p(y|\theta)}{p(y)} = \frac{\pi(\theta)p(y|\theta)}{\int \pi(\theta)p(y|\theta)\mathrm{d}\theta} \tag{2.2}$$

where

$$p(y) = \int \pi(\theta)p(y|\theta)\mathrm{d}\theta \tag{2.3}$$

is the normalizing constant of the posterior distribution $p(\theta|y)$. It is also the marginal distribution of y.Note that the summation of the normalizing factor in the discrete situation has been replaced by the integration for the continuous case. Both equations (2.1) and (2.2)are called the Bayes' theorem, but the former represents Bayes' theorem in the discrete case and the latter represents the continuous case. Since the marginal probability is not a function of the parameter, it can be ignored in the Bayesian inference. The likelihood function of θ,denoted by $L(\theta)$, is proportional to $p(y|\theta)$, i.e. $L(\theta) \propto p(y|\theta)$. The two differ by a constant factor, which is irrelevant to the parameter. We can rewrite equation (2.2) as

$$p(\theta|y) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)\mathrm{d}\theta} \tag{2.4}$$

The marginal distribution $p(y)$ in the continuous case is an integral. As long as the integral is finite, it will not have any impact on the posterior distribution $p(\theta|y)$. The reason is that the integral is a function of the data y but not a function of the parameter. Therefore, this constant does not affect the inference of the parameter. The posterior

distribution can be rewritten in the form of proportionality, as shown below.

$$p(\theta|y) \propto L(\theta)\pi(\theta) \tag{2.5}$$

2.2.2 Prior distribution

A prior distribution must be defined for each parameter before we can perform Bayesian inference. Based on different categorizing rules, the prior distribution can be categorized into different types. In this section, we introduce the following concepts of prior distributions.

2.2.2.1 Informative and uninformative priors

Informative priors provide specific, definite information about a variable. An informative prior usually has an impact on the posterior distributions. These priors must be handled with care in practice because when they are in use, people are combining the past experience and the data obtained from the current experiment to make statistical inference. An example of the informative priors is the normal distribution with a mean and a relatively small variance.

Uninformative priors are also known as vague, diffuse or flat priors. An uninformative prior has minimal impact on the posterior distribution of the parameter. The uninformative prior is used solely for allowing the investigator to conduct Bayesian inference because a prior is required for each parameter. It is one of many Bayesian statisticians' favorite priors due to its objectiveness. More references about the theories and applications of prior distributions can be found in Berger and Bernardo (1989) and Tibshirani (1989). The most commonly used uninformative priors include the uniform prior for θ, i.e. $\pi(\theta) \propto 1$, or a normal distribution with mean 0 and an extremely large

variance, say 10^{15}.

2.2.2.2 Proper and improper priors

The definitions of proper and improper priors are relatively straightforward. If the sum (discrete case) or integral (continuous case) of a prior density is finite, the prior is proper; otherwise, it is improper prior. There are many examples for proper and improper priors. A simple one occurs in the Bayesian linear regression analysis. One can use a normal distribution as the prior for a regression coefficient. The integral of the normal distribution is finite as long as the variance of the normal is finite. So, it is a proper prior. A typical example of improper prior is Beta(0,0), the beta distribution for $\alpha = 0$ and $\beta = 0$. Another example of improper is the logarithmic prior on a positive real number. A proper prior always leads to a proper posterior. However, an improper prior does not always lead to a proper posterior. Some improper prior may generate improper posterior and others may generate proper posteriors. A typical example is the unbounded uniform prior for the regression coefficient of a linear model. With this prior, the posterior distribution of the regression coefficient is normal. Although the unbounded uniform prior is improper, the posterior is proper. The property of the posterior distribution for the residual variance of the linear model, however, depends on the prior distribution. We often use the improper prior $\pi(\sigma^2) = 1/\sigma^2$ to describe the uncertainty of σ^2. This improper prior leads to an improper posterior for σ^2. Therefore, improper prior should be used with caution.

2.2.2.3 Conjugate prior

In Bayesian probability theory, a class of prior distributions is said to be conjugate to a class of likelihood functions $p(y|\theta)$ if the resulting posterior distributions $p(\theta|y)$

belong to the same family as $\pi(\theta)$. Such a prior is called a conjugate prior. The concept of "conjugate prior" was introduced by Raiffa and Schlaifer (1961). More discussions about conjugate priors can be found in Consonni and Veronese (1992) and Diaconis and Ylvisaker (1979).

The most common conjugate family is the Gaussian family. If the likelihood function is Gaussian, choosing a Gaussian prior over the mean will ensure that the resulting posterior is also Gaussian. Let $y = \{y_1, ..., y_n\}$ be n independent observations from a $y_j \sim N(\theta, \phi)$ distribution, where θ (the mean) is the parameter and ϕ (the scale or variance) is assumed to be known. Sometimes it is more convenient to denote the normal distribution by $p(y_j|\theta) = N(y_j|\theta, \phi)$. With this notation of the distribution, a normal prior for parameter θ can be written as $\pi(\theta) = N(\theta|\theta_0, \phi_0)$. The posterior distribution of is proportional to the product of the prior and the likelihood

$$p(\theta|y) \propto \pi(\theta)p(y|\theta) \propto exp[-\frac{1}{2}\theta^2(1/\phi_0 + n/\phi) + \theta(\theta_0/\phi_0 + \sum_{j=1}^{n} y_j/\phi)] \tag{2.6}$$

If we compare this posterior with the kernel of a normal distribution, we can see that it is normal $p(\theta|y) = N(\theta|\theta_1, \varphi_1)$ with mean $\theta_1 = \varphi_1(\theta_0/\varphi_0 + \sum y_j/\varphi)$ and variance $\phi_1 = (1/\phi_0 + n/\phi)^{-1}$. Therefore, the Gaussian family is conjugate. Details of the derivation and the kernel of the normal distribution will be described in a later section.

2.2.2.4 Jeffreys' prior

The Jeffreys' prior was introduced by Jeffreys (1939) in the following form,

$$\pi(\theta) \propto |I(\theta|y)|^{1/2} \tag{2.7}$$

where $I(\theta|y)$ is the Fisher information matrix defined as

$$I(\theta|y) = -E[\frac{\partial^2 \ln p(y|\theta)}{\partial \theta^2}] = E[\frac{\partial \ln p(y|\theta)}{\partial \theta}]^2 \tag{2.8}$$

Here,$p(y|\theta)$ is the likelihood function as defined earlier. The Jeffreys' prior is locally uniform and hence uninformative. Another property of the Jeffreys' prior is that it is invariant with respect to one-to-one transformation (Jeffreys 1946) . More specifically, in a Bayesian context, if we transform the unknown parameter θ to $\psi = \psi(\theta)$, then

$$\frac{\partial\{\ln p(y|\psi)\}}{\partial \psi} = \frac{\partial\{\ln p(y|\theta)\}}{\partial \theta}\frac{d\theta}{d\psi} \tag{2.9}$$

Squaring and taking expectations over values of y (please note that $d\theta/d\psi$ does not depend on y), then we have

$$I(\psi|y) = I(\theta|y)(d\theta/d\psi)^2 \tag{2.10}$$

Thus, if a prior density $p(\theta) \propto |I(\theta|y)|^{1/2}$ is used, then by the usual chain rule of differentiation, we have $p(\psi) \propto |I(\psi|y)|^{1/2}$. The transformation invariance property can save much effort in searching for new prior distribution for a transformed parameter, which is one of the reasons why Jeffreys suggested this prior.

2.2.3 Posterior distribution

In the section of Bayes' theorem, we introduced the concept of posterior distribution of a parameter. Like the prior distribution, the posterior distribution of a parameter is also not a real distribution. It represents our updated degree of belief about the parameter after information from experiments is combined with our subjective prior

knowledge. The posterior distribution, if exists, will be narrower than the prior distribution. In other words, we will be more certain about the true value of the parameter after we observe the data than we were before.

In general, there are three ways to find the posterior distribution. (1) The most important way is to take advantage of the literature. Most of the problems in agricultural experiments may be analyzed using similar designs of experiments and similar models found in other areas. In many situations, the posterior distributions are given by the investigators in those areas. (2) In the situations where no comparable problems can be found in the literature, we may choose a prior in the conjugate family. In this case, the posterior distribution is automatically given because it belongs to the same family as the prior distribution. (3) Finally, we may try to derive the posterior distribution by ourselves. The general rule for deriving a posterior distribution is to compare the kernel of the posterior distribution with the kernels of many existing distributions. The distribution in the list of existing distributions that has the same kernel as the posterior distribution is the posterior distribution of the parameter.

We now provide two examples to show how to derive the posterior distribution of a parameter. The first example is the normal distribution. The problem is slightly different from that given in the section of conjugate prior. Let $y = \{y_1, ..., yn\}$ be the data where $p(y_j) = N(y_j|\theta, \phi)$ for $j = 1, ..., n$. The mean θ is the parameter and ϕ is a known scalar. Let $\pi(\theta) = a$ be the prior distribution, where a is a constant (not a function of the parameter). This prior is improper. We want to derive the posterior distribution $p(\theta|y)$.First, we need to find the kernel of this posterior distribution. The

full expression of the posterior density is

$$p(\theta|y) = \pi(\theta)p(y|\theta) = \frac{a}{(2\pi\phi)^{n/2}}\exp[-\frac{1}{2\phi}\sum_{j=1}^{n}(y_j-\theta)^2] \tag{2.11}$$

where

$$\sum_{j=1}^{n}(y_j-\theta)^2 = \sum_{j=1}^{n}y_j^2 - 2\theta\sum_{j=1}^{n}y_j + n\theta^2 \tag{2.12}$$

Substituting this sum of squares into equation (2.11), we get

$$\begin{aligned} p(\theta|y) &= \frac{a}{(2\pi\phi)^{n/2}}\exp[-\frac{1}{2\phi}(\sum_{j=1}^{n}y_j^2 - 2\theta\sum_{j=1}^{n}y_j + n\theta^2)] \\ &= \frac{a}{(2\pi\phi)^{n/2}}\exp[-\frac{1}{2\phi}\sum_{j=1}^{n}y_j^2 + \theta\frac{1}{\phi}\sum_{j=1}^{n}y_j - \theta^2\frac{n}{2\phi}] \\ &= \frac{a}{(2\pi\phi)^{n/2}}\exp[-\frac{1}{2\phi}\sum_{j=1}^{n}y_j^2]\exp[\theta\frac{1}{\phi}\sum_{j=1}^{n}y_j - \theta^2\frac{n}{2\phi}] \\ &= C\exp[\theta\frac{1}{\phi}\sum_{j=1}^{n}y_j - \theta^2\frac{n}{2\phi}] \end{aligned} \tag{2.13}$$

where

$$C = \frac{a}{(2\pi\phi)^{n/2}}\exp[-\frac{1}{2\phi}\sum_{j=1}^{n}y_j^2] \tag{2.14}$$

is again a constant with respect to variable θ. In other words, C is not a function of θ, although it is a function of the data and other known quantities. Ignoring this constant, we get the kernel of the posterior,

$$K[p(\theta|y)] = \exp[\theta\frac{1}{\phi/n}\sum_{j=1}^{n}y_j/n - \theta^2\frac{1}{2\phi/n}] \tag{2.15}$$

We used the special notation K[density] to represent the kernel for the density specified within the brackets. We now compare this kernel with the kernels of existing distri-

butions to find a match. Let $\xi \sim N(\mu, \sigma^2)$ be a variable with the specified normal distribution. The density is

$$\begin{aligned} N(\xi|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}(\xi-\mu)^2] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}\mu^2]\exp[\frac{\mu}{\sigma^2}\xi - \frac{1}{2\sigma^2}\xi^2] \\ &= C\exp[\frac{\mu}{\sigma^2}\xi - \frac{1}{2\sigma^2}\xi^2] \end{aligned} \tag{2.16}$$

where

$$C = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}\mu^2] \tag{2.17}$$

is a constant with respect to variable ξ.Therefore, the kernel of this normal is

$$K[N(\xi|\mu,\sigma^2)] = \exp[\frac{\mu}{\sigma^2}\xi - \frac{1}{2\sigma^2}\xi^2] \tag{2.18}$$

We now compare equation (2.15) with equation (2.18),

$$\begin{cases} K[p(\theta|y)] = \exp\left[\dfrac{\sum_{j=1}^{n} y_j/n}{\phi/n}\theta - \dfrac{1}{2\phi/n}\theta^2\right] \\ K[N(\xi|\mu,\sigma^2)] = \exp[\frac{\mu}{\sigma^2}\xi - \frac{1}{2\sigma^2}\xi^2] \end{cases} \tag{2.19}$$

and realize that the two kernels have the same form. Therefore, the posterior distribution of θ is normal with mean and variance given below,

$$p(\theta|y) = N\left(\theta\middle|\frac{1}{n}\sum_{j=1}^{n} y_j, \frac{\phi}{n}\right) \tag{2.20}$$

Although the unbounded uniform prior for θ is improper, the posterior is normal with finite variance and thus it is proper.

The second example is the Beta prior for the parameter of a binomial distribution. Let $y = m/n$ be a binomial data with m events out of n trails. Let θ be the parameter of this binomial distribution. The density is

$$p(y|\theta) = \frac{n!}{m!n!}\theta^m(1-\theta)^{n-m} \tag{2.21}$$

Let

$$\pi(\theta) = \frac{1}{\Gamma(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{2.22}$$

be the Beta prior distribution for the parameter with shape parameter α and scale parameter β. The posterior of θ is

$$p(\theta|y) = \pi(\theta)p(y|\theta) = \frac{1}{\Gamma(\alpha,\beta)}\frac{n!}{m!n!}\theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1} \tag{2.23}$$

The posterior can be rewritten as

$$p(\theta|y) = C\theta^{(m+\alpha)-1}(1-\theta)^{(n-m+\beta)-1} \tag{2.24}$$

where

$$C = \frac{1}{\Gamma(\alpha,\beta)}\frac{n!}{m!n!} \tag{2.25}$$

is a constant with respect to θ. Therefore, the kernel of the posterior distribution is

$$K[p(\theta|y)] = \theta^{(m+\alpha)-1}(1-\theta)^{(n-m+\beta)-1} \tag{2.26}$$

Comparing this kernel with the kernel of the Beta distribution,

$$K[\pi(\theta)] = \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{2.27}$$

we realize that the two have the same form. Therefore, the posterior is Beta with a shape parameter $m + \alpha$ and a scale parameter $m + \beta$.The Beta prior is conjugate with respect to the binomial likelihood. In many problems, explicit forms of the posterior distributions do not exist. Therefore, special algorithms are required to implement the Bayesian analysis. These special algorithms will be discussed later.

2.2.4 Bayesian inference

Once the posterior distribution of the parameter of interest is derived, all information about this parameter can be found from this distribution. It is important to note that the posterior distribution $p(\theta|y)$ is also called the marginal posterior distribution. The reason we want to emphasize the term of "marginal" is that if there are more than one parameter involved in the problem, say $\theta = \{\theta_1, \theta_2\}$, the posterior distributions for the two components are

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2)\mathrm{d}\theta_2 \tag{2.28}$$

and

$$p(\theta_2|y) \int p(\theta_2|\theta_1, y)p(\theta_1)(d)\theta_1 \tag{2.29}$$

Each parameter component should be inferred from its own marginal posterior distribution. Theoretically, once we find the marginal posterior distribution of the parameter, we are done with the job because we can say that $p(\theta|y)$ is is the "Bayesian inference" of the parameter because all information about θ is contained in $p(\theta|y)$. However, just

providing the posterior distribution without summarizing the distribution seems to be incomplete. Therefore, we will discuss the following specific statistics drawn from the posterior distribution.

2.2.4.1 Point estimation

The classical statistical inference is based on the maximization of the likelihood function whereas the Bayesian inference is based on integration of the probability distribution. For example, the posterior mean of parameter θ is obtained by integrating the posterior distribution as

$$\hat{\theta} = E(\theta|y) = \int \theta p(\theta|y) \mathrm{d}\theta$$

Therefore, the posterior mean is considered as a Bayesian estimate. Similarly, the posterior mode can also be considered as a Bayesian estimate of parameter θ. The posterior mode is defined as the value of θ that maximize $p(\theta|y)$.In addition, the posterior median (the median of the posterior distribution) is also a candidate Bayesian estimate. All these candidate Bayesian estimates are called the point estimates.

A point estimate only gives a single value. For example, the posterior mean or median only represents the central location of the posterior distribution. The shape of the distribution, however, tells how spread of the distribution away from the central location. We normally use the square root of the posterior variance as a measurement of the shape. The posterior variance is defined as

$$\mathrm{var}(\theta|y) = \int (\theta - \hat{\theta})^2 p(\theta|y) \mathrm{d}\theta$$

2.2.4.2 Hypothesis testing

Hypothesis testing in Bayesian analysis is slightly different from that in the maximum likelihood analysis. The maximum likelihood analysis is often called the classical or the frequentist method. We first briefly describe hypothesis testing in the frequentist analysis. Let us denote the parameter space by Θ, which is partitioned into two disjoint subspaces, Θ_0 and Θ_1, so that $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$. The null hypothesis and the alternative hypothesis are defined as $\mathrm{H}_1 : \theta \in \Theta_0$ and $\mathrm{H}_1 : \theta \in \Theta_1$, respectively. A decision regarding which hypothesis should be accepted is made based on the p-value approach. The p-value is defined as the probability that a test statistic is at least as extreme as the one that is actually observed assuming that the null hypothesis is true. Let $T(y)$ be the observed test statistic (a single value calculated from the data of the current experiment) and $p[\xi(y)|\theta \in \Theta_0]$ be the probability density of the test statistic under the null model. The p-value is expressed as

$$p - \text{value} = \int_{T(y)}^{\infty} p[\xi(y)|\theta \in \Theta_0]\mathrm{d}\xi \tag{2.30}$$

If the p-value is small, say p-value<0.05, it means that $T(y)$ is unlikely to be drawn from $p[\xi(y)|\theta \in \Theta_0]$ distribution. Therefore, $\mathrm{H}_0 : \theta \in \Theta_0$ should be rejected. Rejection of $\mathrm{H}_0 : \theta \in \Theta_0$ means acceptance of $\mathrm{H}_1 : \theta \in \Theta_1$. The small probability for the p-value to compare is determined by the investigator. It is quite arbitrary, but people often choose 0.05 or 0.01.If $p>0.05$, we accept $\mathrm{H}_0 : \theta \in \Theta_0$. If $0.01<p<0.05$ we reject $\mathrm{H}_0 : \theta \in \Theta_0$ and say that the test is significant. If $p<0.01$,we reject $\mathrm{H}_0 : \theta \in \Theta_0$ and say that the test is very significant. The null distribution $p[\xi(y)|\theta \in \Theta_0]$ is purely hypothetical. It is a distribution of the test statistic under the null model assuming that we could repeat the experiment infinite number of times.

In Bayesian analysis, hypothesis testing is more intuitive than that in the frequentist analysis. We do not rely on hypothetical repeated experiments in the future under the null mode; rather, we only focus on the posterior distribution of the parameter given the data of the current experiment. Let $p_0 = \Pr(\theta \in \Theta_0|y)$ and $p_1 = \Pr(\theta \in \Theta_1|y)$ be the posterior probabilities under the two hypotheses. Define $\pi_0 = \Pr(\theta \in \Theta_0)$ and $\pi_1 = \Pr(\theta \in \Theta_1)$ as the prior probabilities. Let p_1/p_0 be the posterior odds ratio and π_1/π_0 be the prior odds ratio. For example, if $p_1/p_0 = 20$,we can say that H_1 is 20 times as likely to be true as H_0. The posterior odds ratio divided by the prior odds ratio gives the Bayes' factor

$$\text{BF} = \frac{p_1/p_0}{\pi_1/\pi_0} = \frac{p_1\pi_0}{p_0\pi_1} \tag{2.31}$$

The Bayes' factor can be interpreted as the "odds for H_1 to H_0".

2.2.4.3 Credibility set

In the frequentist analysis, we normally define a confidence set for parameters. In Bayesian analysis, we do not call it a confidence set; instead, we call it a credibility set. The two (confidence and credibility) sets are similar but defined under different methods of analysis. In the Bayesian analysis, a $100(1-\alpha)\%$ credibility set for θ is a subset C of Θ which satisfies

$$1-\alpha \leq P(C|y) = \begin{cases} \int_C P(\theta|y)\mathrm{d}\theta & \text{(continuous case)}, \\ \sum_{\theta\in C} p(\theta|y) & \text{(discrete case)}. \end{cases} \tag{2.32}$$

When choosing a credibility set for θ, we try to minimize the size of the set. We want to choose a set that includes the points with the largest posterior density. So, analogous to the smallest confidence set in frequentist analysis, we have the highest posterior density

(HPD) region in Bayesian statistics. For each parameter, we can also define the equal tail credible interval. Details of these credible intervals will be described latter in the section of post MCMC analysis.

2.3 MCMC Algorithm

Markov chain Monte Carlo (MCMC) refers to an algorithm to implement the Bayesian analysis. The term Monte Carlo in the context of statistics means computer simulation. Markov chain represents a special distribution in "time series" where the state of a variable in the current time point depends only on the state of the variable in the previous time point. Why do we have to use MCMC to implement the Bayesian analysis? The reason is simple, that is to numerically approximate high dimensional integral. The largest obstacle in Bayesian analysis is the integration. We are interested in the posterior distribution of parameters conditional on data of the current experiment. But the posterior distribution rarely has a closed form.

In Bayesian analysis, our goal is usually to obtain the expectation of a function of the unknown parameters from the posterior distribution. This can be expressed as

$$E[g(\theta)|y] = \int g(\theta)p(\theta|y)\mathrm{d}\theta \tag{2.33}$$

But in most cases, this integration is difficult to derive. The MCMC algorithm is a simulation method to sample parameters from the posterior distribution. In the resulting sampling consequence, each observation depends only on the previous one like a Markov chain in the time series. As in Monte Carlo integration, Monte Carlo is used to

approximate an expectation by using the Markov chain samples, as shown below,

$$E[g(\theta)|y] = \int g(\theta)p(\theta|y)\mathrm{d}\theta \cong \frac{1}{M}\sum_{t=1}^{M} g(\theta^{(t)}) \tag{2.34}$$

where $\theta^{(t)}$ is the sampled parameter vector at iteration t and M is the posterior sample size.

The MCMC algorithm may be accomplished in one of three different ways: The Gibbs sampler, the Metropolis algorithm and the Metropolis-Hastings algorithm. The earliest MCMC algorithm is the Metropolis algorithm introduced by Metropolis and Ulam (1949) and further detailed by Metropolis *et al.* (1953) . Hastings (1970) made a generalization of the Metropolis algorithm and developed the so called Metropolis-Hastings algorithm. Geman and Geman (1984) analyzed an image data set by using what is now called the Gibbs sampler, which is a special case of the Metropolis-Hastings algorithm. Gibbs sampler is named after physicist J.W. Gibbs, in reference to an analogy between the sampling algorithm and statistical physics. All these algorithms can draw a sequence of samples from the joint distribution of two or more variables. More references about the MCMC algorithm can be found in the literature given below. The properties of Markov chains were discussed by these authors, Feller (1968) , Breiman (1968) and Meyn and Tweedie (1993) . Conditions that govern the Markov chain convergence and rates of convergence can be found in Amit (1991), Applegate, Kannan and Polson (1990), Chan (1993), Geman and Geman (1984) , Liu, Wong and Kong (1991a; 1991b), Rosenthal (1991a;1991b), Tierney (1994), and Schervish and Carlin (1992). Papers that provide both theoretical and applied treatments of the MCMC algorithm are found in Tanner (1993), Gilks, Richardson and Spiegelhalter (1996), Chen, Shao and Ibrahim (2000), Liu (2001), Gelman et al (2004), Robert and Casella (2004), and Congdon (2001;

2003; 2005).

2.3.1 Gibbs sampler

Gibbs sampler is the simplest MCMC algorithm, in which one parameter is sampled at a time from its fully conditional posterior distribution. Let m be the number of parameters and $\theta = \{\theta_1, ..., \theta_m\}$ be the vector of parameters. The fully conditional posterior distribution of the kth parameter is denoted by $p(\theta_k|\theta_{-k}, y)$, where θ_k is the kth parameter and θ_{-k} is an $(m-1)\times 1$ vector for the remaining parameters. In order to perform the Gibbs sampler, $p(\theta_k|\theta_{-k}, y)$ must be a distribution with a closed form, i.e., this distribution must be known and a random number generator is available for this distribution. In most situations of the MCMC analysis, this fully conditional distribution has a simple form, e.g., normal distribution, Bernoulli distribution and so on. For example, if $p(\theta_k|\theta_{-k}, y)$ is normal with mean μ and variance σ^2, i.e., $p(\theta_k|\theta_{-k}, y) = N(\theta_k|\mu, \sigma^2)$, where μ and σ^2 are functions of θ_{-k} and y, then a normal random number generator is required so that the value of θ_k can be directly generated from the simulator. Most software packages do not have an existing generator for such a normal distribution with mean μ and variance σ^2, but they often have a generator for the standardized normal distribution. In such a case, we first generate a standardized random variable $z \sim N(0, 1)$ and then take a linear transformation to get

$$\theta_k = z\sigma + \mu \tag{2.35}$$

The original definition of Gibbs sampler is that θ_k must be a single component of vector θ. This has been generalized to the so called "block Gibbs sampler" in which θ_k contains more than one parameters. The block Gibbs sampler is more efficient than the single

element Gibbs sampler in terms of speed of convergence to the stationary distribution. However, it depends on a closed form of the multivariate version of the fully conditional posterior distribution $p(\theta_k|\theta_{-k}, y)$.

Gibbs sampler is a special case of the Metropolis-Hastings algorithm. The algorithm was named by Geman and Geman (1984) after the American physicist Josiah W. Gibbs. Gelfand et al. (1990) first used Gibbs sampler to solve problems in Bayesian statistics. Casella and George (1992) gave a tutorial on Gibbs sampler. The Gibbs sampler can be summarized as follows,

1. Let $t = 0$ and initialize all parameters by $\theta^{(t)} = \{\theta_1^{(t)}, ..., \theta_m^{(t)}\}$.

2. Generate each parameter in turn as follows:

   - draw $\theta_1^{(t+1)}$ from $p(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, ..., \theta_m^{(t)}, y)$
   - draw $\theta_2^{(t+1)}$ from $p(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_m^{(t)}, y)$
   - .....
   - draw $\theta_m^{(t+1)}$ from $p(\theta_m|\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_{m-1}^{(t+1)}, y)$

3. Increment $t = t + 1$ and repeat step 2 until a desired length of the Markov chain has been reached.

In the fully conditional posterior distribution, the values of the parameters appearing after the conditional sign "|" must be the most current values in order to have the maximum efficiency of the sampling.

2.3.2 Metropolis algorithm

Named after its inventor, the American physicist and computer scientist N. C. Metropolis, the Metropolis algorithm can be used to generate random samples from any

complicated target distribution of any form. This makes the algorithm more general than the Gibbs sampler in which the target distribution must be known. Like the Gibbs sampler, the Metropolis algorithm can also be performed either in a univariate fashion or a multivariate (block) fashion. Let us take the univariate Metropolis algorithm as an example. Suppose that we want to sample θ_k from $p(\theta_k|\theta_{-k}^{(t)}, y)$ where $p(\theta_k|\theta_{-k}^{(t)}, y)$ does not have a closed form, i.e., we do not know what type of distribution it is so that a Gibbs sampler cannot be used. To implement the Metropolis sampler, we first sample a random number from a symmetric distribution, e.g., a uniform distribution or a normal distribution. This symmetric distribution is called the proposal distribution. Let $\xi \sim \text{Uniform}(-\Delta, \Delta)$ be the proposal distribution where Δ is a positive number in the neighborhood of zero (Δ is also called the tuning parameter). Let $\theta_{(t)}$ be the current value of parameter θ_k. We then let $\theta_k^* = \theta_k^{(t)} + \xi$ be the proposed value for the next move.This proposed value may be accepted or rejected based on the Metropolis rule to be described shortly. If the proposed value is accepted, we let $\theta_k^{(t+1)} = \theta_k^* = \theta_k^{(t)} + \xi$, the parameter is updated. If the proposed value is rejected,$\theta_k^{(t+1)} = \theta_k^{(t)}$, the previous value is carried over to the next cycle. Unlike the Gibbs sampler where the acceptance rate is 100%, the Metropolis algorithm does not guarantee that the proposed value is always accepted. The Metropolis rule says that the proposed value θ_k^* is accepted with probability

$$\alpha = \min\left\{\frac{p(\theta_k^*|\theta_{-k}^{(t)}, y)}{p(\theta_k^{(t)}|\theta_{-k}^{(t)}, y)}, 1\right\} \tag{2.36}$$

In other words, the probability of acceptance is a function of the posterior ratio of the proposed value to the previous value. If the posterior ratio is greater than one, $\alpha = 1$ and the proposed value is always accepted. If the posterior ratio is less than one, the proposed value is accepted but only with a probability α; there is still a $1 - \alpha$ chance

that the old value is carried over to the next cycle. The Metropolis algorithm is also called the random walk algorithm.

The claim that the Metropolis algorithm can be applied to any arbitrary distribution is perhaps over exaggerated. Some adjustment needs to be done to make the algorithm sufficiently general for any distribution. For example, if the parameter is a variance component, which can only take positive number, and the previous value is already close to the boundary, we cannot simulate the proposed value from a symmetric distribution. Note that $\theta_k > 0$ is the domain of the parameter. Let $0 < \theta_k^{(t)} < \Delta$ be the current value and Δ be the tuning parameter of the proposal distribution. The proposed value must be sampled from

$$\theta_k^* = \theta_k^{(t)} + \text{Uniform}(-\Delta^*, \Delta) \tag{2.37}$$

where

$$\Delta^* = \begin{cases} \Delta & \text{for } \theta_k^{(t)} \geq \Delta, \\ \theta_k^{(t)} & \text{for } \theta_k^{(t)} < \Delta. \end{cases} \tag{2.38}$$

may be smaller than Δ to guarantee that the proposed value is within the legal domain. This proposal distribution is not symmetric and thus violates the basic requirement of the Metropolis algorithm. If we use this asymmetric proposal distribution as the sampler, the value of θ_k will be trapped at 0, even if the true value of θ_k may be much larger than zero. This problem can be avoided by a proper adjustment of the acceptance probability, an improvement made by Hastings (Hastings 1970).

2.3.3 Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm is a generalization of the Metropolis algorithm. It was proposed by Hastings (1970) to handle asymmetric proposal distribution.

We use the asymmetric uniform distribution as an example to demonstrate this algorithm. Let

$$q(\theta_k^*|\theta_k^{(t)}) = \text{Uniform}(\theta_k^*|\theta_k^{(t)} - \Delta^*, \theta_k^* + \Delta) \tag{2.39}$$

be the proposal distribution, where $\Delta^* = \theta_k^{(t)} < \Delta$. We need to examine a reverse probability of the above proposal distribution. This reverse distribution is the probability that, given the new value θ_k^*, the parameter eventually takes the previous value $\theta_k^{(t)}$. This reverse proposal distribution depends on the θ_k^*. If $\theta_k^* \geq \Delta$, the reverse proposal distribution is

$$q(\theta_k^{(t)}|\theta_k^*) = \text{Uniform}(\theta_k^{(t)}|\theta_k^* - \Delta, \theta_k^* + \Delta) \tag{2.40}$$

Otherwise,

$$q(\theta_k^{(t)}|\theta_k^*) = \text{Uniform}(\theta_k^{(t)}|\theta_k^* - \Delta^{**}, \theta_k^* + \Delta) \tag{2.41}$$

where

$$\Delta^{**} = \begin{cases} \Delta & \text{for } \theta_k^* > \Delta, \\ \theta_k^* & \text{for } \theta_k^* < \Delta. \end{cases} \tag{2.42}$$

The Metropolis-Hastings acceptance probability is defined as

$$\alpha = \min\left\{ \frac{p(\theta_k^*|\theta_{-k}^{(t)}, y)}{p(\theta_k^{(t)}|\theta_{-k}^{(t)}, y)} \frac{q(\theta_k^{(t)}|\theta_k^*)}{q(\theta_k^*|\theta_k^{(t)})}, 1 \right\} \tag{2.43}$$

The proposal ratio is

$$\frac{q(\theta_k^{(t)}|\theta_k^*)}{q(\theta_k^*|\theta_k^{(t)})} = \frac{\Delta + \Delta^*}{\Delta + \Delta^{**}} \tag{2.44}$$

When neither $\theta_k^{(t)}$ nor θ_k^* is near the boundary, we have $\Delta^* = \Delta^{**} = \Delta$. The proposal ratio is unity and the Metropolis-Hastings algorithm becomes the simple Metropolis algorithm. Therefore, the Metropolis algorithm is a special case of the Metropolis-

Hastings algorithm when the proposal ratio equals one.

We mentioned earlier in the review that the Gibbs sampler is a special case of the general MH algorithm. The Gibbs sampler draws θ_k directly from the fully conditional posterior distribution. Therefore, $q(\theta_k^*|\theta_k^{(t)}) = p(\theta_k^*|\theta_{-k}^{(t)}, y)$ and $q(\theta_k^{(t)}|\theta_k^*) = p(\theta_k^{(t)}|\theta_{-k}^{(t)}, y)$. The probability of accepting the new draw is

$$\alpha = \min\left\{ \frac{p(\theta_k^*|\theta_{-k}^{(t)}, y)}{p(\theta_k^{(t)}|\theta_{-k}^{(t)}, y)} \frac{q(\theta_k^{(t)}|\theta_k^*)}{q(\theta_k^*|\theta_k^{(t)})}, 1 \right\} = \min\{1, 1\} = 1 \tag{2.45}$$

The new draw is always accepted. Therefore, the Gibbs sampler is a special case of the MH algorithm. All the three algorithms introduced so far are called the MCMC algorithm.

Choosing the proposal distribution and the tuning parameter can be tedious. However, we can let the computer to figure out an optimal tuning parameter. Before the MCMC process starts, we set up a target acceptance probability, say 0.60, and let the computer to find the tuning parameter so that the acceptance rate is close to the target acceptance rate. This tuning process can take very long time.

2.3.4 Assessment of Markov chain convergence

2.3.4.1 Burn-in and thinning

The product of MCMC is a posterior sample for all parameters of interest. This sample is supposed to be generated from the posterior distribution of the parameters, $p(\theta|y)$. However, the MCMC algorithm draws parameters from $p(\theta_k|\theta_{-k}, y)$ in turn. The current observation drawn depends on the previous draw, explaining why the chain is called the Markov chain. The entire MCMC process is stochastic, meaning that the

parameters drawn do not converge to some fixed values; instead, they converge to a distribution. This distribution is the posterior distribution $p(\theta|y)$, also called the stationary distribution. The Markov chain takes some time to reach the stationary distribution. Before the stationary distribution is reached, the observations drawn cannot be used. The time from the start of the chain to the point where the stationary distribution is just reached is called the burn-in period. Observations from the burn-in period should be discarded. Once the stationary distribution is reached, we can collect the observations and store them in the computer as random draws from the posterior distribution. However, observations from consecutive draws may be highly correlated (autocorrelation). Therefore, we have to delete several observations and keep one observation repeatedly along the Markov chain. This process is called thinning. The rate of thinning depends on the degree of autocorrelation. An alternative way of collecting the posterior sample is to collect only one observation after the burn-in for each chain and restart the chains using a different set of initial values. For a posterior sample of M observations, we need M independent Markov chains. This alternative approach is time consuming but avoids the concern of autocorrelation.

2.3.4.2 Visual analysis of trace plots

One important convergence diagnostic checking method is to use the visual analysis. This can be done through drawing the trace plots of the parameters. A trace plot of a parameter is a plot of the sampled value of the parameter against the time (number of iterations). The trace plot of a parameter provides a lot of information about the chain, such as whether the chain has converged to the stationary distribution, whether we need a longer burn-in period or whether the chain mixes well or not. Figure 2.1 gives examples of the trace plots for several different situations. Figure 2.1a shows a

typical trace plot for a variable. The initial value of the parameter was -10. It takes less than 100 iterations for the chain to reach the stationary distribution where the mean of the stationary distribution was 1.0. Afterwards, the value of the parameter tends to stabilize around 1.0. The chain mixes very well after the stationary distribution is reached. Figure 2.1b shows an example where the Markov chain does not mix well, indicating a potential problem of the model. Figure 2.1c shows a situation where the chain does not converge to any distribution at all.

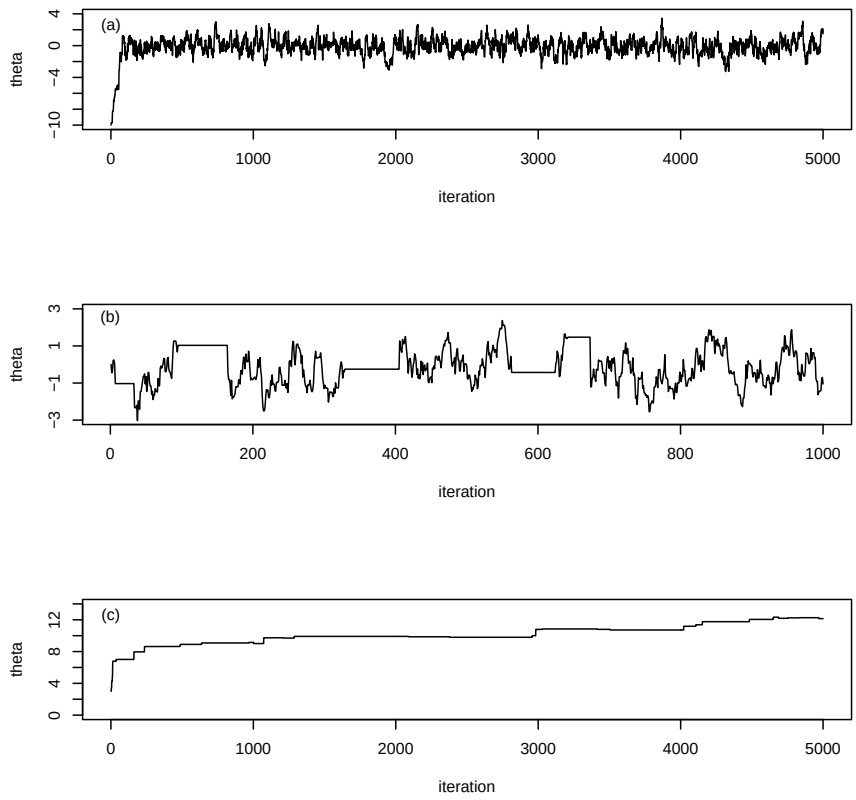


Figure 2.1: Trace plots for Markov chain convergence diagnosis for parameter theta (θ). (a) The Markov chain quickly reaches the stationary distribution. (b) The Markov chain mixes poorly (high autocorrelation). (c) The Markov chain does not converge at all.

2.3.4.3 Statistical diagnosis for convergence

Gelman-Rubin R test The Gelman and Rubin R test (Gelman and Rubin 1992; Brooks and Gelman 1997) requires multiple chains running simultaneously. This diagnostic test compares the between chain variance of the parameter of interest with the within chain variance. A significantly larger between chain variance than the within chain variance indicates that the chains have not converged to the stationary distribution. The idea is similar to the analysis of variance where we partition the total variance of a variable into between and within group variances. Technical details of the R test

statistics can be found in the original studies (Gelman and Rubin 1992; Brooks and Gelman 1997)

Geweke z-test Intuitively, for a Markov chain, if an early part of the chain is very different from a later part of the chain, then the Markov chain has not converged at the point where the early part is examined. Based on this idea, Geweke (1992) proposed a z-test for checking the convergence of a single Markov chain. The basic idea was to collect two subsamples from the Markov chain, one from an early stage of the chain with sample size and one from a later stage of the chain with sample size M_2, where $M_1 + M_2 < M$ and M is the total length of the Markov chain. A z-test can be performed for the sampled parameter of interest. If the means of the two samples are significantly different, the chain may not have converged at the point where the early stage of the sample is collected.

2.3.4.4 Autocorrelation

The posterior sample obtained through MCMC sampling is different from a sample of observations collected from a real agricultural experiment in that the observations from consecutive draws of the MCMC may be highly correlated. This type of correlation is called autocorrelation or series correlation. A high autocorrelation indicates poor mixing. Autocorrelation may be monitored by the plot of correlation between consecutive observations against the lag, as demonstrated in Figure 2.2 where panel (a) indicates low autocorrelation and panel (b) indicates high autocorrelation. The autocorrelation of lag h for parameter θ in a Markov chain is defined as

$$\rho_h(\theta) = \frac{\text{cov}_h(\theta)}{\text{cov}_0(\theta)}, \text{for} \quad 0 < h < M \tag{2.46}$$

where M is the posterior sample size and

$$\text{cov}_h(\theta) = \frac{1}{M-h}\sum_{i=1}^{M-h}(\theta^{(i+h)} - \bar{\theta})(\theta^{(i)} - \bar{\theta}), \text{for} \quad 0 \leq h < M \tag{2.47}$$

is called the autocovariance. The denominator in equation (2.46) is the autocovariance of lag 0, which is expressed as

$$\text{var}(\theta) = \text{cov}_0(\theta) = \frac{1}{M}\sum_{i=1}^{M}(\theta^{(i)} - \bar{\theta})(\theta^{(i)} - \bar{\theta}) \tag{2.48}$$

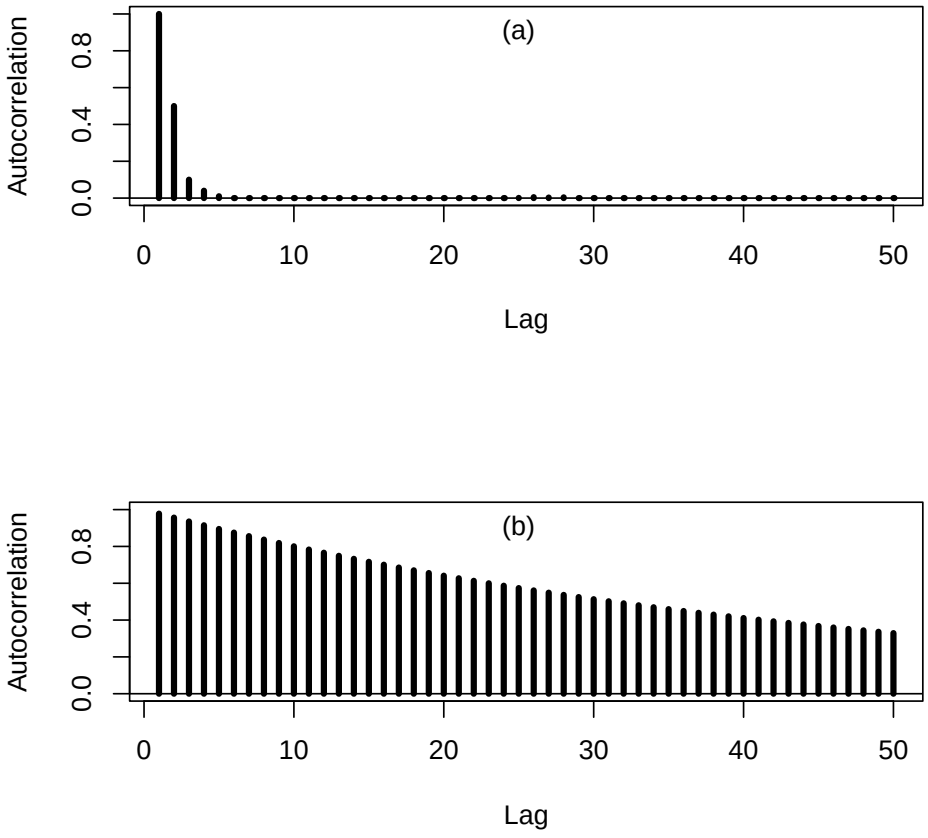


Figure 2.2: Auto correlation plots against lag: (a) Low autocorrelation, (b) High autocorrelation.

2.3.4.5 Effective sample size (ESS)

The effective sample size is defined as (Kass et al. 1998)

$$\text{ESS} = \frac{M}{1 + 2\sum_{h=1}^{\infty} \rho_h(\theta)} \tag{2.49}$$

where M is the actual posterior sample size and $\rho_h(\theta)$ is the autocorrelation of lag h for parameter θ. In practice $\sum_{h=1}^{\infty} \rho_h(\theta)$ is usually replaced by $\sum_{h=1}^{H} \rho_h(\theta)$ where H is called the correlation time, a positive number so that $\rho_h(\theta) \approx 0$ as $h > H$. Because the autocorrelation is always positive, the effective sample size is always less than the actual posterior sample size. A much smaller effective sample size than the actual size indicates poor mixing of the Markov chain. The concept of effective sample size is much the same as the effective population size in population genetics.

2.3.5 Post MCMC analysis

The product of the MCMC implemented Bayesian analysis is a posterior sample for all parameters. The posterior sample of a parameter contains all information about the parameter. The most informative information is the posterior distribution itself. However, the investigator needs to summarize the posterior distribution to convince the readers what he/she wants to tell the readers. Therefore, further analysis of the posterior sample is necessary to complete the Bayesian analysis. This further analysis is called the post MCMC analysis, which is often called post Bayesian analysis in the literature.

2.3.5.1 Posterior sample and marginal posterior distribution

Recall that the main purpose of Bayesian analysis is to infer the marginal posterior distribution of the parameters of interest. If there are multiple parameters of interest, the marginal posterior distribution for each parameter should be inferred. Let $\theta = \{\theta_1, ..., \theta_m\}$ be the vector of multiple parameters. Our main purpose is to infer $p(\theta_k|y)$ for $k = 1, ..., m$. However, we never sample a parameter from this marginal distribution; instead, we always sample θ_k from $p(\theta_k|\theta_{-k}, y)$. What is the relationship of the posterior sample drawn from $p(\theta_k|\theta_{-k}, y)$ and the marginal posterior distribution $p(\theta_k|y)$?

The burn-in period and the thinning of the Markov chain serve as a way to convert the sampled observations from $p(\theta_k|\theta_{-k}, y)$ into $p(\theta|y) = p(\theta_1, ..., \theta_m|y)$, the joint posterior distribution of the parameters. The complete MCMC chain values are not i.i.d. samples from the posterior distribution. After the burn-in samples are removed and the chains are thinned to remove the autocorrelation, the remaining samples may be regarded as i.i.d. samples from the marginal posterior of the parameters. Instead of inferring the posterior distribution explicitly, we now have a sample drawn from that distribution. With this sample in hands, we can look at each parameter of interest and ignore all other parameters. The distribution of this parameter (when other parameters are ignored) is equivalent to the empirical marginal posterior distribution. This is analogous to the situation of sampling (x, y) from the joint distribution $p(x, y)$. When examining the sampled values of x and ignoring the values of y, we are actually looking at the marginal distribution of x.

2.3.5.2 Summary statistics

Given the posterior sample drawn from the marginal distribution, we are ready to perform the post MCMC analysis. The following summary statistics are normally reported in a Bayesian analysis and are discussed in turn as follows. Again, the summary statistics are computed on chains after removal of burn-in and thinning.

Posterior mean The most important summary statistic for a parameter is the posterior mean, which is simply the arithmetic average of the sampled values of the parameter in the posterior sample,

$$E(\theta|y) = \hat{\theta} = \frac{1}{M}\sum_{t=1}^{M}(\theta^{(t)}) \tag{2.50}$$

where $\theta_{(t)}$ is the tth observation in the MCMC sample and is the posterior sample size. Posterior mode and median are also relevant posterior statistics. The posterior mode of a parameter is defined as the most frequent value of the sampled parameters in the posterior sample. The posterior median is the value of the parameter that divides the posterior sample into two equal parts. Depending on the shape of the distribution, the posterior mode and posterior median of the parameter may be different from the posterior mean. In this case, they should also be reported. These summary statistics are called the point estimates.

Posterior variance The posterior variance of a parameter represents how the sampled values spread (deviate from) around the posterior mean. It is calculated as the sample variance of the parameter in the posterior sample,

$$\text{var}(\theta|y) = \frac{1}{M-1}\sum_{t=1}^{M}(\theta^{(t)} - \hat{\theta})^2 \tag{2.51}$$

The square root of the posterior variance is the posterior standard deviation. We often call it the posterior "standard error". However, it should not be confused with the definition of standard error in a real sample of a variable, where the standard error is defined as the square root of the variance of the sampled mean. The posterior variance should be used with caution when the autocorrelation is high. If the Markov chain is thinned well, the posterior sample should have very low autocorrelation. The posterior variance is useful only if the autocorrelation is low. In practice, the adjusted posterior variance should be used, in which the effective sample size is taken into account for the adjustment.

The effective sample size introduced earlier is useful to adjust the posterior variance for Bayesian significance test. It is well known that the empirical posterior mean of parameter obtained from a posterior sample is not affected by the autocorrelation, but the posterior variance is strongly affected by the auto correlation. This leads to a serious downwards bias for the estimated posterior variance. Let $\text{var}(\theta|y)$ be the posterior variance obtained from the posterior sample with sample size M. We know that $\text{var}(\theta|y)$ is biased downwardly due to autocorrelation. The adjusted posterior variance may be defined as

$$\text{var}^*(\theta|y) = \frac{M}{\text{ESS}}\text{var}(\theta|y) \tag{2.52}$$

This adjusted posterior variance should be used when performing any significance test for parameter θ.

Credibility interval There are two types of credibility intervals to consider in Bayesian analysis. One is the so called $\alpha \times 100\%$ equal tail credibility interval and the other is the $(1-\alpha) \times 100\%$ highest posterior density (HPD) interval. The $\alpha \times 100\%$ equal tail credibility interval is defined as the interval bracketed by the $\alpha/2$ quantile and the

$1 - \alpha/2$ quantile of the posterior sample, where $0 < \alpha < 1$. Typical choice of the alpha value is $\alpha = 0.05$. The credibility interval is defined as an interval in the posterior distribution where the posterior density within the interval is higher than the posterior density outside the interval. If such an interval contains $1 - \alpha$ of the posterior sample, it is called the $(1 - \alpha) \times 100\%$ HPD credibility interval. In reality, we do not have the posterior density; instead, we only have M observations drawn from the posterior distribution. The credibility interval can be interpreted as the shortest interval that contains $(1 - \alpha) \times M$ observations of the posterior sample.

2.4 Software packages

There are many statistical software packages that can perform Bayesian analysis using the MCMC algorithm. Many of them are specialized for particular problems in some special areas. Here, we introduce two software packages that are sufficiently general to handle any problems of Bayesian analysis. One program is called WinBUGS, a free software package downloadable from the internet. This program uses its own computer language for program coding, called the BUGS language. The other program is a SAS procedure called the MCMC procedure. Since SAS is a commercialized statistical software package, the program is not free. However, the SAS Institute provides excellent service for technical support to all SAS users.

2.4.1 WinBUGS

WinBUGS is a stand-alone program for Bayesian analysis using the Gibbs sampler. It is based on the BUGS (Bayesian inference Using Gibbs Sampling) project that began in 1989 by the MRC (Medical Research Council) Biostatistics Unit. The BUGS project

was initially conducted under the "Classic" BUGS program, which uses a text-based model description and a command-line interface. This old version of the BUGS is still available for all the major computer platforms, e.g., the UNIX and the DOS. WinBUGS is the Windows version of BUGS that was initially released in 1997. The current version of the WinBUGS v1.4.3 was developed jointly by the original BUGS group and the Imperial College School of Medicine at St. Mary's, London.

The WinBUGS14.exe file can be downloaded from the following website,

http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

Once the program is installed in the computer, the user needs to use a patch to upgrade to WinBUGS version 1.4.3 and a free key to allow for unrestricted use of the program. The patch and the link to the free key are given at the top of the WinBUGS webpage. WinBUGS can handle models with high degree of complexity. Users need to provide the model (also called the log likelihood) to describe the relationship of the data and the parameters. A proper prior distribution should also be provided by the user for each parameter. WinBUGS can figure out the fully conditional posterior distribution for each parameter automatically so that Gibbs sampler can be used to draw the samples. WinBUGS does not take improper priors.

There are at least three ways that a user can run WinBUGS: (1) Using the WinBUGS window interface, (2) Using the script and (3) Running WinBUGS under R. Directly running WinBUGS under the WinBUGS window can be very tedious and is only performed for small and simple problems. It is also used for beginners to learn WinBUGS. Using the script to run WinBUGS is recommended for running large and complicated problems. Under the script running mode, users can store all the commands in a text file and execute the script file. The most efficient way of running WinBUGS is to run the program under R. Users need to download an R package called R2WinBUGS

(http://cran.r-project.org/web/packages/R2WinBUGS/index.html). Under R, the data input and output are handled by R. WinBUGS directly takes the variables from R for analysis. One of the nice properties of running WinBUGS under R is that users can fully take advantage of the R package to draw high resolution graphs.

2.4.2 PROC MCMC

Data analysis in SAS is divided into two steps, (1) the data step and (2) the procedure step. The data step allows the user to input the data. The procedure step is to perform the statistical analysis using a built-in subroutine within the SAS system. SAS calls each subroutine a procedure. There are numerous procedures within the SAS system, each performing some specific task. Users simply select the appropriate procedure to analyze the data with minimum requirement for coding. The SAS has virtually all procedures that users need for data analysis. The SAS syntax to call a procedure is "PROC name-of-the-procedure;"

There is a particular procedure named the MCMC procedure. The syntax to call this procedure is "PROC MCMC". The MCMC procedure is particularly designed for the MCMC implemented Bayesian analysis. It is sufficiently general to handle problems with high level of complexity. It is even more general than WinBUGS because PROC MCMC, by default, uses the general random walk Metropolis algorithm to sample all variables. The very reason that PROC MCMC does not use the Gibbs sampler to draw variable is the emphasis of generality. The Gibbs sampler is problem (distribution) specific while the random walk Metropolis algorithm is not. PROC MCMC can handle improper priors. Users have an option to choose any prior distributions, as long as the log of the prior densities is programmable. There is a tradeoff between generality and efficiency. PROC MCMC spends most of the time trying to tune the parameter

of the proposal distribution to optimize the acceptance rate (at about 60%. Therefore, PROC MCMC usually takes longer time for the MCMC analysis than WinBUGS. The MCMC procedure is still a trial procedure. Significant improvement is expected in future releases. For skilled users, PROC MCMC does provide an option to draw variables using the Gibbs sampler. This is called the User Defined Sampler (UDS) option. Detailed information about the MCMC procedure can be found in the SAS help and documentation.

2.5 Data analysis

We analyzed three datasets to demonstrate the Bayesian method and the software application. Although both WinBUGS and PROC MCMC were introduced, only PROC MCMC was used for the Bayesian analysis. For each data set, we introduced the data, selected a model for the data, chose a prior distribution for each parameter and provided the fully conditional posterior distribution for the parameter. We also provided the SAS code and reported the result for each data analysis.

2.5.1 The damage data

The first data set was an example to evaluate the variation of the degree of damage caused by insects for different varieties of wheat (Milliken and and Johnson 2009) . The data set is given in Table 2.1 and called the "damage" data. The experimenter randomly selected four varieties of wheat from a large number of varieties and conducted an experiment to evaluate damage caused by insects on the wheat plants just prior to heading. The design structure was a completely randomized design with four replications or plots per variety (the plot is the experimental unit). Because of environmental conditions,

some of the plots were destroyed (flooded out by excess rain). Therefore, the design was also an unbalanced design. The experimenter randomly selected 20 plants from each available plot and rated the amount of insect damage done to each plant using a scale from 0 to 10 where 0 indicates no damage and 10 indicates severe damage. Thus, the response measured on each plot is the mean of the ratings from the 20 plants within each plot.

Table 2.1: The average degree of damage caused by insects for four randomly selected wheat varieties. The original dataset was published by Milliken and Johnson (2009, p. 314). This dataset is called the "damage data" here in this study.

Plot	Wheat	Damage
1	A	3.90
2	A	4.05
3	A	4.25
4	B	3.60
5	B	4.20
6	B	4.05
7	B	3.85
8	C	4.15
9	C	4.60
10	C	4.15
11	C	4.40
12	D	3.35
13	D	3.80

2.5.1.1 Model

Let y be an $n \times 1$ vector of the response variable (damage) where $n = 13$, β be a scalar for the population mean, X be an $n \times 1$ vector of unity, $\gamma = \{\gamma_A, \gamma_B, \gamma_C, \gamma_D\}$ be an $m \times 1$ vector for the effects of wheat variety where $m = 4$ is the number of varieties, Z be an $n \times m$ design matrix (dummy variables) and ε be an $n \times 1$ vector of residual errors with an assumed $N(0, I\sigma_E^2)$ distribution. The linear model for the damage trait is

$$y = X\beta + Z\gamma + \varepsilon \tag{2.53}$$

Detail of this linear model is shown as follows,

$$\begin{bmatrix} 3.90 \\ 4.05 \\ 4.25 \\ 3.60 \\ 4.20 \\ 4.05 \\ 3.85 \\ 4.15 \\ 4.60 \\ 4.15 \\ 4.40 \\ 3.35 \\ 3.80 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \end{bmatrix} \tag{2.54}$$

where vector $\gamma = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$ is a numerical representation of vector $\gamma = \{\gamma_A, \gamma_B, \gamma_C, \gamma_D\}$.

2.5.1.2 Prior and posterior

The investigator randomly selected four out of many existing wheat varieties. Therefore, the effects of varieties are random effects. Frequentists call such a model the random model. Since the population mean β is a fixed effect, this model is represented in the form of a mixed model. We now choose the following prior distributions,

$$\pi(\beta) = \text{Normal}(\beta|0, 10^{15})$$

$$\pi(\gamma) = \text{Normal}(\gamma|0, I\sigma_A^2)$$

$$\pi(\sigma_A^2) = \text{Inv} - \chi^2(\sigma_A^2|\tau, \omega) = \text{Inv} - \chi^2(\sigma_A^2|0, 0) = 1/\sigma_A^2$$

$$\pi(\sigma_E^2) = \text{Inv} - \chi^2(\sigma_E^2|\tau, \omega) = \text{Inv} - \chi^2(\sigma_E^2|0, 0) = 1/\sigma_E^2 \tag{2.55}$$

We have used a new notation system to represent the density. For example, $\pi(\beta) = \text{Normal}(\beta|0, 10^{15})$ is equivalent to $\beta \sim N(0, 10^{15})$ and $\pi(\sigma_A^2) = \text{Inv}-\chi^2(\sigma_A^2|\tau, \omega)$ is equivalent to $\sigma_A^2 \sim \text{Inv} - \chi^2(\tau, \omega)$, which is called the scaled inverse chi-square distribution with degree of belief τ and scale ω. Under the mixed model framework, frequentists call γ random effects (not parameters), but Bayesians do not specifically distinguish parameters from random effects; they call everything, except the hyper parameters, random variable. The purpose of the analysis was to examine the relative importance of the variance of wheat varieties to the total variance of the variable damage. This relative importance is also called the intra class correlation and denoted by

$$\rho = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} \tag{2.56}$$

The γ vector is an unknown vector whose prior distribution is multivariate normal. The variance in the prior distribution is given a scaled inverse chi-square distribution with zero degree of belief and zero scale (also called the Jeffreys' prior). This prior is improper, but the MCMC procedure can handle such an improper prior. This mixed model is also called the Bayesian hierarchical model because of the multiple levels of

priors. Let us define

$$V = \text{Var}(y) = ZZ^T\sigma_A^2 + I\sigma_E^2 \tag{2.57}$$

We can derive the fully conditional posterior distribution for each unknown. The priors we chose are conjugate and hence they generated the following fully conditional posterior distributions,

$$\begin{aligned}
&p(\beta|...) = \text{Normal}[\beta|(X^TV^{-1}X)^{-1}(X^TV^{-1}y), (X^TV^{-1}X)^{-1}] \\
&p(\gamma|...) = \text{Norma}[\gamma|\sigma_A^2Z^TV^{-1}(y - X\beta), \sigma_A^2(I - Z^TV^{-1}Z\sigma_A^2)] \\
&p(\sigma_A^2|...) = \text{Inv} - \chi^2(\sigma_A^2|\tau + m, \omega + \gamma^T\gamma) = \text{Inv} - \chi^2(\sigma_A^2|4, \gamma^T\gamma) \\
&p(\sigma_E^2|...)\text{Inv} - \chi^2(\sigma^2|\tau + n, \omega + SS) = \text{Inv} - \chi^2(\sigma_E^2|13, SS)
\end{aligned} \tag{2.58}$$

where

$$SS = (y - X\beta - Z\gamma)^T(y - X\beta - Z\gamma) \tag{2.59}$$

is the sum of squares of the residual errors. Only two random number generators are required in this problem, the normal variable generator and the scaled inverse chi-square variable generator. To generate a scaled inverse chi-square variable, we need a chi-square variable generator, which is available in most software systems. Let χ^2_{df} be a random variable sampled from a chi-square distribution with degrees of freedom df. The scaled inverse chi-square variable with $p(\sigma_A^2|...) = \text{Inv} - \chi^2(\sigma_A^2|\tau + m, \omega + \gamma^T\gamma)$ distribution is simply obtained by

$$\sigma_A^2 = \frac{\omega + \gamma^T\gamma}{\chi^2_{\tau+m}} \tag{2.60}$$

Since every unknown has a closed form distribution, the Gibbs sampler algorithm can

be used for the MCMC experiment.

2.5.1.3 SAS code

```
%let dir=C:\Bayes\damage;

libname xx "&dir";
filename aa "&dir\post-sample.csv";

data damage;
   input plot y x z1-z4;
datalines;
 1  3.90  1 1 0 0 0

 2  4.05  1 1 0 0 0

 3  4.25  1 1 0 0 0

 4  3.60  1 0 1 0 0

 5  4.20  1 0 1 0 0

 6  4.05  1 0 1 0 0

 7  3.85  1 0 1 0 0

 8  4.15  1 0 0 1 0

 9  4.60  1 0 0 1 0

10  4.15  1 0 0 1 0

11  4.40  1 0 0 1 0

12  3.35  1 0 0 0 1

13  3.80  1 0 0 0 1
;
run;

ods graphics on;

proc mcmc data=damage outpost=xx.postsample nmc=100000
          thin=100 seed=246810 nbi=5000 ntu=3000
          monitor=(beta gamma1-gamma4 sigma2_a sigma2_e rho);
     ods select PostSummaries ESS Geweke PostIntervals TADpanel;
     array gamma[4];
     array z[4] z1-z4;
```

```
     parms beta 0;
     parms gamma: 0;
     parms sigma2_a 1;
     parms sigma2_e 1;
     begincnst;
        tau=1e-10;
        omega=1e-10;
     endcnst;
     prior beta ~ normal(mean = 0, var = 1e15);
     prior gamma: ~ normal(mean = 0, var = sigma2_a);
     prior sigma2_a ~ sichisq(tau,omega);
     prior sigma2_e ~ sichisq(tau,omega);
/*   prior sigma2_a ~ general(-log(sigma2_a));*/

/*   priorsigma2_e~ general(-log(sigma2_e));*/
     mu = x*beta;
     beginprior;
        rho=sigma2_a/(sigma2_a+sigma2_e);
     endprior;
     do k=1 to 4;
        mu = mu + z[k]*gamma[k];
     end;
     model y ~ normal(mean = mu, var = sigma2_e);
run;
ods graphics off;

proc export data=xx.postsample outfile=aa dbms=csv replace;
run;
```

Here is a brief explanation of the SAS code. The statements before "proc mcmc" are typical SAS statements for creating the SAS dataset for analysis. Readers are supposed to be familiar with the SAS language, and thus no explanation is given. The MCMC procedure starts with the statement "proc mcmc". Here are the explanations of the options of the "proc mcmc" statement.

`data=damage: Tells proc mcmc to use data with a name damage`

`outpost=xx.postsample: Tells proc mcmc to write the posterior sample to a SAS dataset named xx.postsample. The two level SAS dataset name means that the posterior sample will be stored in the folder with libname xx as a permanent SAS dataset. The dataset contains all variables defined in the parms statements plus the log likelihood, the log prior density and the log posterior density.`

`nmc=100000:This option defines the total length of the Markov chain`

after the burn-in deletion.

thin=100: This option defines the thinning rate. In this case, the posterior sample will keep one draw in every 100 iterations. In this example, the posterior sample size (named xx.postsample) will contain 100000/100=1000 observations.

seed=246810: This option allows users to set the seed for random number generators. Choosing the same seed will allow the users to duplicate the results. If no seed is given, proc mcmc assumes a default seed of zero, which will generate a different sequence of random numbers every time the program is executed. The difference between different runs is called the Monte Carlo error.

nbi=5000: Defines the number of iterations in the burn-in period. In this case, proc mcmc starts to collect posterior sample after 5000 iterations. The burn-in period does not affect the posterior sample size stored in the outpost dataset. For example, the current setting requires proc mcmc to run a total of 100000+5000=105000 iterations, although the posterior sample only contains 1000 observations.

ntu=3000: Define the number of iterations for tuning the parameter of the proposal distribution to reach a desirable acceptance rate. Users can ignore this option.

monitor=(beta gamma1-gamma4 sigma2_a sigma2_e rho): Variables included in the braces are subject to the post MCMC analysis. Note that variable rho is not a true variable but a function of sigma2_a and signa2_e. This new variable must be defined within the proc mcmc procedure in order to monitor this variable.

The following paragraph explains the statements within the MCMC procedure.

ods select PostSummaries ESS Geweke PostIntervals TADpanel; This statement tells proc mcmc to select the following items to be handled by the SAS output delivery system (ODS) for output: (1) The post MCMC summaries for the variables contained in the monitor() option, effective sample sizes, (2) the Geweke z-test diagnostic statistics for convergence, (3) the credibility intervals and (4) the trace-autocorrelation-density (TAD) panels. Each monitored variable has a TAD panel that contains three figures drawn in the same page (the trace plot, the autocorrelation plot against lag and the marginal posterior density).

array gamma[4]; Define an array named gamma. Later on, you can refer gamma1-gamma4 for the four variables defined by this array statement.

array z[4] z1-z4; Define an array named z which refers to z1-z4.

```
parms beta 0; Define a parameter named beta and assign a value 0 as
      the initial value.

parms gamma: 0; Define an array named gamma (four variables) as
parameters and assign each of the four elements an initial value of
zero. It also tells proc mcmc to draw the four variables
simultaneously as a block. If you want to draw each element
separately, you need four parms statements, one for each element.
Note that adding ":" after gamma means gamma1-gamma4.

parms sigma2_a 1; Define signma2_a as a parameter and assign an
     initial value 1 to the parameter.

parms sigma2_e 1; Define signma2_e as a parameter and assign an
     initial value 1 to the parameter.

begincnst; Start a "begin constant and end constant" block.
     tau=1e-10; Assign variable tau a constant (a very small number).
     omega=1e-10; Assign variable omega a constant (a very small number)

endcnst;  End the "begin constant and end constant" block. Within
     this block, each variable is assigned a constant. Throughout the
     entire MCMC sampling process, tau and omega are two constant values.

prior beta ~ normal(mean = 0, var = 1e15); Assign parameter beta a
     normal prior with mean zero and a very large variance.

prior gamma:~ normal(mean = 0, var = sigma2_a); Assign each of the
     gamma variables a normal prior with mean zero and a common variance
     sigma2_a.

prior sigma2_a ~ sichisq(tau,omega); Assign sigma2_a a scaled
     inverse chi-square prior distribution with tau and omega as
     the degree of belief and scale, respectively.

prior sigma2_e ~sichisq(tau,omega); Defined the same prior as
     sigma2_a.

/*prior sigma2_a ~ general(-log(sigma2_a));*/

/* prior sigma2_e ~ general(-log(sigma2_e));*/

The above two statements are commented out. These statements assign
the variance components a Jeffreys' prior (improper). Since the
Jeffreys' prior is a user defined prior distribution, you must use
the general function and put the log density of your prior density
inside the general function. Note that log(1/sigma2_a) =-log(sigma2_a).
```

```
beginprior; Define a "begin prior and end prior" block.

rho=sigma2_a/(sigma2_a+sigma2_e); Create a new variable rho.

endprior; End the "begin prior and end prior" block. The reason that
we placed this statement inside the beginprior and endprior block is
to save computing time. This assign statement will only be executed
twice per iteration when placed inside this block. Otherwise, it
will be executed n = 13 times per iteration. mu = x*beta; Define the
fixed effect.

do k=1 to 4;

mu = mu + z[k]*gamma[k];

end;

Define mu as the sum of the fixed and the random effects.

model y ~ normal(mean = mu, var = sigma2_e); Define the model, i.e.,
the likelihood function or the density of the data given the
parameters.
```

The last line of the code calls another SAS procedure named PROC EXPORT.

```
proc export data=xx.postsample outfile=aa dbms=csv replace;
```

The EXPORT procedure simply writes the posterior sample stored in the SAS dataset xx.postsample into a physical excel file with a name defined in the filename aa statement. The filename aa refers to a physical file "post-sample.csv" in the "C:\Bayes \damage" folder.

2.5.1.4 Result

Figure 2.3 shows the posterior TAD (trace-autocorrelation-density) panels for the population mean (beta or β). The Markov chain converges very well with very low autocorrelation and almost a perfect normal posterior distribution. Figure 2.4 gives the TAD panels for the intra class correlation (rho or ρ). Because of the small sample size ($n = 13$). the intra class correlation is hard to estimate accurately. The trace plot

shows that the Markov chain mixes poorly with very high autocorrelation. The posterior distribution seems to be bimodal (two peaks). The posterior mean estimate is 0.3779 with a large posterior standard deviation 0.3098.

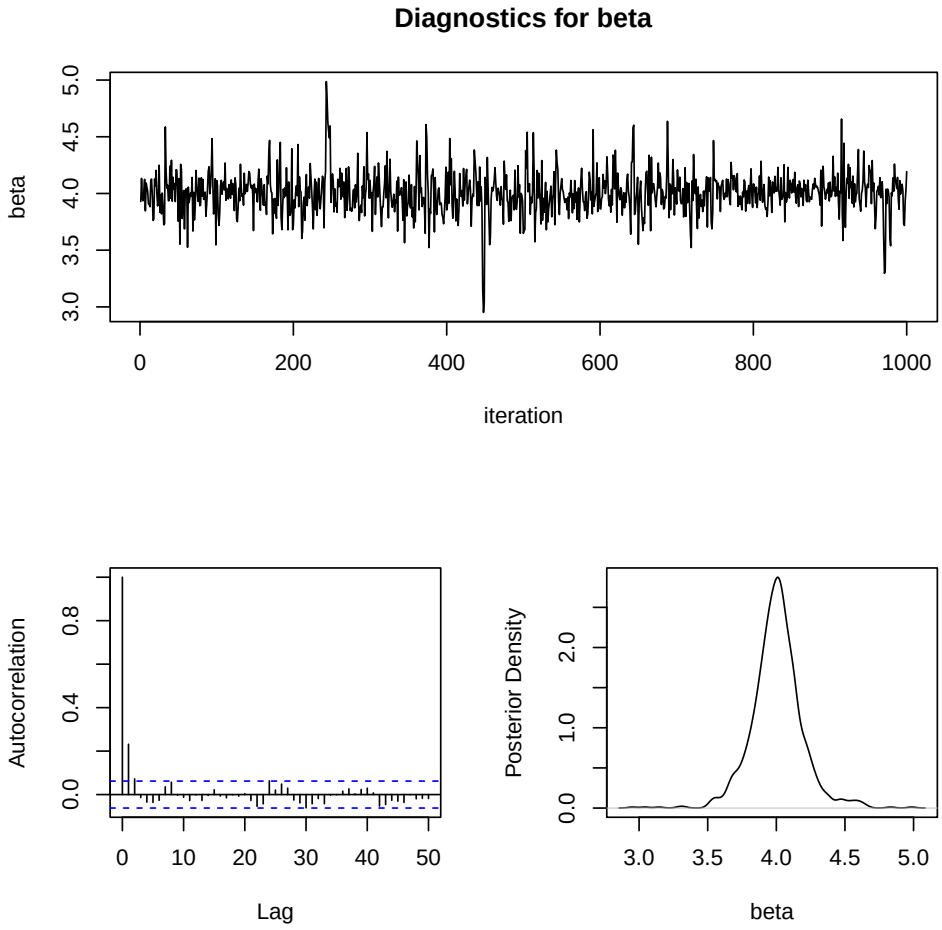


Figure 2.3: The posterior TAD panels (trace, autocorrelation and density) for parameter beta (β) of the damage data.

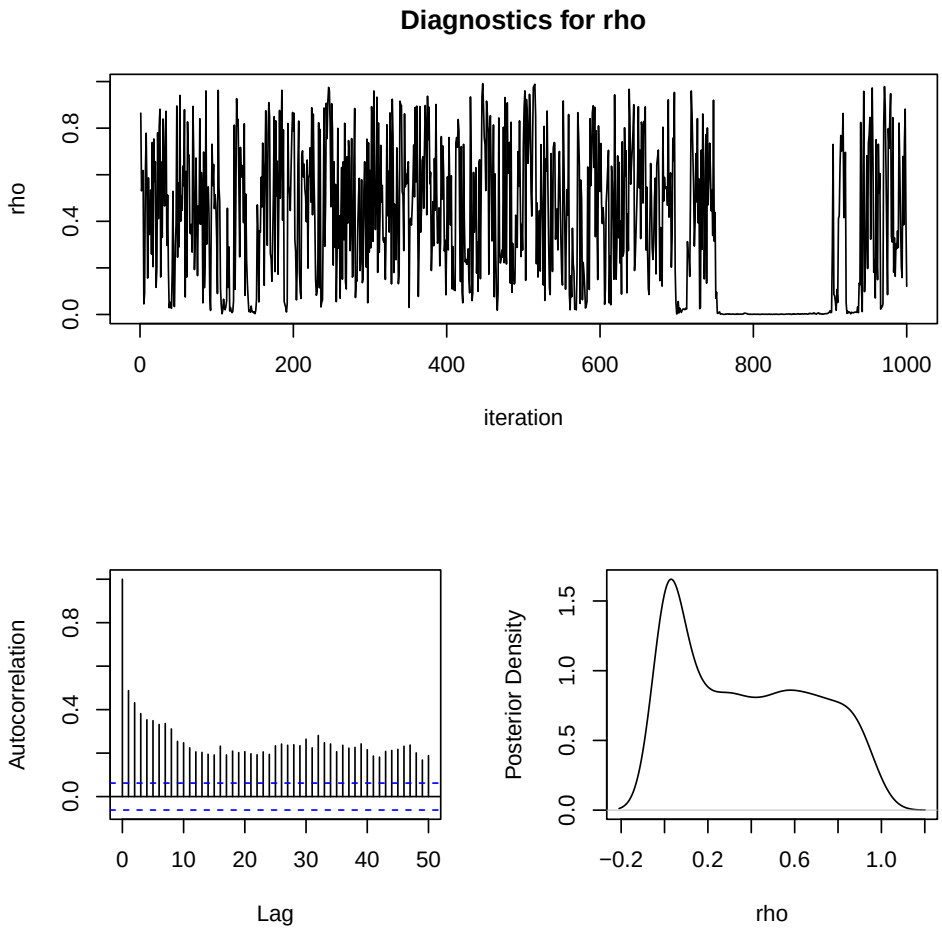


Figure 2.4: The posterior TAD panels for parameter rho ($\rho = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$) of the damage data.

Table 2.2 gives the summary statistics for all the variables, including parameters and the random effects. The posterior means for the between variety variance and within variety variance are $\hat{\sigma}_A^2 = E(\sigma_A^2|y) = 0.1197$ and $\hat{\sigma}_E^2 = E(\sigma_E^2|y) = 0.0913$, respectively. The posterior mean of the intra class correlation (relative importance of between variety variance) is $\hat{\rho} = E(\rho|y) = 0.3779$. This Bayesian estimate is not the same as

$$\frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_E^2} = \frac{0.1197}{0.1197 + 0.0913} = 0.5673 \tag{2.61}$$

This ratio would be the maximum likelihood estimate of the ratio if both variance components were the maximum likelihood estimates.

Table 2.2: Summary statistics of the posterior sample for the damage data.

		Standard	Equal-Tail-Interval		HPD(95%)	
Parameter	Mean	deviation	2.5%	97.5%	HPD(left)	HPD(right)
β	4.0044	0.1868	3.6492	4.4128	3.6479	4.3874
γ_1	0.0361	0.1905	-0.3605	0.4336	-0.3455	0.4466
γ_2	-0.0384	0.1822	-0.4713	0.3043	-0.4379	0.3227
γ_3	0.1900	0.2233	-0.1431	0.6834	-0.1590	0.6576
γ_4	-0.1965	0.2366	-0.7632	0.0956	-0.7197	0.0996
σ_A^2	0.1197	0.3052	0.0001	0.7357	0.0000	0.4497
σ_E^2	0.0913	0.0594	0.0289	0.2460	0.0219	0.1976
ρ	0.3779	0.3098	0.0008	0.9323	0.0002	0.8930

Table 2.3 shows the Geweke z test for convergence and other diagnostic statistics. Parameters that have converged well include β, γ_1, γ_2, γ_4 and σ_A^2. SS denotes the posterior sample size. ESS denotes the effective sample size. The most important parameter ρ behaves very badly, which is consistent with the trace plot. Because the sample size is so small, it is hard to obtain good estimate for this ratio parameter.

Table 2.3: Diagnostic test statistics for the Markov chain convergence of the damage data.

Parameter	Geweke-z	p-value	SS	ESS	Correlation time	Efficiency
β	-0.4524	0.6510	1000	636.8	1.5704	0.6368
γ_1	1.3413	0.1798	1000	537.0	1.8621	0.5370
γ_2	0.1325	0.8946	1000	601.6	1.6621	0.6016
γ_3	2.4789	0.0132	1000	67.4	14.8397	0.0674
γ_4	-1.9281	0.0538	1000	47.7	20.9737	0.0477
σ_A^2	-0.3054	0.7601	1000	279.4	3.5790	0.2794
σ_E^2	-4.2433	$<.0001$	1000	349.1	2.8649	0.3491
ρ	3.6545	0.0003	1000	21.6	46.1933	0.0216

2.5.2 The seeds data

The purpose of this experiment was to investigate the effect of host plants (bean and cucumber) on the seed germination of two *Orobanche cernua aegyptiaca* plant varieties (O.a75 and O.a73). The *Orobanche cernua* is a kind of parasitic plant that lives on other plants. In other words, roots of the parasitic plant penetrate into other plants to extract nutrition from the host plants. This is a 2×2 factorial design of experiment. The response variable is the germination rate and the depend variables are the plant variety and the root extract. The data are given in Table 2.4 as a binomial data set with the trial being the number of seeds and the event being the number of germinated seeds. This data set is called the "seed" data in this study. The original data were obtained from Crowder (1978)

Table 2.4: Seed germination of two different varieties of *Orobanche cernua aegyptiaca* plants (O.a75 and O.a73) in two different host plants or root extracts (bean and cucumber). The data set was published by Crowder (1978) and called the "seeds data" in this study. The column headed "germinated" is the numbers of germinated seeds. The column headed "seed" is the number of seeds planted. The column headed "rate" is the proportion of the germinated seeds (number of germinated seeds / total number of seeds planted).

Plate	Breed	Host	Germinated	Seed	Rate
1	O.a75	Bean	10	39	0.2564
2	O.a75	Bean	23	62	0.3710
3	O.a75	Bean	23	81	0.2840
4	O.a75	Bean	26	51	0.5098
5	O.a75	Bean	17	39	0.4359
6	O.a75	Cucumber	5	6	0.8333
7	O.a75	Cucumber	53	74	0.7162
8	O.a75	Cucumber	55	72	0.7639
9	O.a75	Cucumber	32	51	0.6275
10	O.a75	Cucumber	46	79	0.5823
11	O.a75	Cucumber	10	13	0.7692
12	O.a73	Bean	8	16	0.5000
13	O.a73	Bean	10	30	0.3333
14	O.a73	Bean	8	28	0.2857
15	O.a73	Bean	23	45	0.5111
16	O.a73	Bean	0	4	0.0000
17	O.a73	Cucumber	3	12	0.2500
18	O.a73	Cucumber	22	41	0.5366
19	O.a73	Cucumber	15	30	0.5000
20	O.a73	Cucumber	32	51	0.6275
21	O.a73	Cucumber	3	7	0.4286

2.5.2.1 Model

The response variable is binomial with two data points per observation. The numerator is the number of germinated seeds (event, denoted by m_j) and the denominator is the total number of seeds planted (trial, denoted by n_j). The observed germination rate is $r_j = m_j/n_j$ for $j = 1, ..., n$, where $n = 21$ is the sample size (number of plates in the seed data). Since the data point is not normally distributed, we used the generalized linear model to fit the data. Let $E(r_j) = \mu_j$ be the expectation of the binomial data point, which is connected to two independent variables, breed of the O.a plant (O.a 75 and O.a 73) and host plant (bean and cucumber). Let

$$\eta_j = X_j\beta + Z_{j1}\gamma_1 + Z_{j2}\gamma_2 + Z_{j3}\gamma_3 \tag{2.62}$$

be a linear combination of the model effects, where $X_j = 1$ for all $j = 1, ..., n$, β is the intercept, γ_1 is the effect of breed, $Z_{j1} = \{0, 1\}$ is a binary indicator variable assigned a value 0 for O.a 75 and 1 for O.a 73, γ_2 is the effect of host, $Z_{j2} = \{0, 1\}$ is a binary indicator variable assigned a value 0 for bean and 1 for cucumber, γ_3 is the interaction effect of the breed and host and $Z_{j3} = Z_{j1} \times Z_{j2} = \{0, 1\}$ is an binary indicator variable with value 1 for the (O.a 73,Cucumber) combination and 0 otherwise. We are interested in estimating the parameter vector $\theta = \{\beta, \gamma_1, \gamma_2, \gamma_3\}$. A significant γ_1 indicates that the two breeds of O.a plants have different germination rate. A significant γ_2 indicates that the two host plants cause different germination rates for the O.a plants. A significant γ_3 indicates a strong interaction effect between breed and host. In matrix notation, the linear model is

$$\eta = X\beta + Z\gamma \tag{2.63}$$

The relationship between μ_j and η_j is through the the logit link,

$$\eta_j = \text{logit}(\mu_j) = \log\left(\frac{\mu_j}{1-\mu_j}\right) \tag{2.64}$$

More intuitively, the inverse of the logit link is

$$\mu_j = \text{logistic}(\eta_j) = \frac{\exp(\eta_j)}{1+\exp(\eta_j)} = \frac{\exp(X_j\beta + Z_j\gamma)}{1+\exp(X_j\beta + Z_j\gamma)} \tag{2.65}$$

The likelihood function (model) of the jth data point is

$$p(m_j|n_j,\mu_j) = \text{Binomial}(m_j|n_j,\mu_j) = \frac{(n_j)!}{(m_j)!(n_j-m_j)!}\mu_j^{m_j}(1-\mu_j)^{n_j-m_j} \tag{2.66}$$

2.5.2.2 Prior and posterior

Each parameter is assigned a flat normal prior, i.e.,

$$\begin{aligned}
\pi(\beta) &= \text{Normal}(\beta|0,10^{15}) \\
\pi(\gamma_1) &= \text{Normal}(\gamma_1|0,10^{15}) \\
\pi(\gamma_2) &= \text{Normal}(\gamma_2|0,10^{15}) \\
\pi(\gamma_3) &= \text{Normal}(\gamma_3|0,10^{15})
\end{aligned} \tag{2.67}$$

None of the parameters has a closed form of the fully conditional posterior distribution. Therefore, Gibbs sampler cannot be used for the MCMC sampling; instead, the general random walk Metropolis algorithm must be used.

Although our parameter vector is $\theta = \{\beta, \gamma_1, \gamma_2, \gamma_3\}$, the investigator's real interest

was to conduct various comparisons for the germination rates. Let the average germination rates for the four combinations be

$$\begin{aligned}
\psi_{00} &= \text{logistic}(\beta) && (O.a75, \text{Bean}) \\
\psi_{01} &= \text{logistic}(\beta + \gamma_2) && (O.a75, \text{Cucumber}) \\
\psi_{10} &= \text{logistic}(\beta + \gamma_1) && (O.a73, \text{Bean}) \\
\psi_{11} &= \text{logistic}(\beta + \gamma_1 + \gamma_2 + \gamma_3) && (O.a73, \text{Cucumber})
\end{aligned} \tag{2.68}$$

The following comparisons are of interest,

$$\begin{aligned}
\psi_{Breed} &= (\psi_{10} + \psi_{11}) - (\psi_{00} + \psi_{01}) \\
\psi_{Host} &= (\psi_{01} + \psi_{11}) - (\psi_{00} + \psi_{10}) \\
\psi_{Breed \times Host} &= (\psi_{01} + \psi_{10}) - (\psi_{00} + \psi_{11})
\end{aligned} \tag{2.69}$$

Therefore, we can report the summary statistics of these additional parameters.

2.5.2.3 SAS code

```
%let dir=C:\Bayes\seeds;

libname xx "&dir";
filename aa "&dir\post-sample.csv";

data seeds;
    input  plate x z1 z2 z3 m n;
    r=m/n;
datalines;

1   1   0   0   0   10  39

2   1   0   0   0   23  62
```

```
3   1   0   0   0   23  81

4   1   0   0   0   26  51

5   1   0   0   0   17  39

6   1   0   1   0   5   6

7   1   0   1   0  53   74

8   1   0   1   0  55   72

9   1   0   1   0  32   51

10  1   0   1   0  46   79

11  1   0   1   0  10   13

12  1   1   0   0  8    16

13  1   1   0   0  10   30

14  1   1   0   0   8   28

15  1   1   0   0   23  45

16  1   1   0   0   0   4

17  1   1   1   1   3   12

18  1   1   1   1   22  41

19  1   1   1   1   15  30

20  1   1   1   1   32  51

21  1   1   1   1   3    7;

run
;

ods graphics on; proc mcmc data=seeds outpost=xx.postsample
seed=123456
          nmc=100000 thin=100 nbi=5000 ntu=3000
          monitor=(beta gamma1-gamma3
                   psi_00 psi_01 psi_10 psi_11
                   psi_breed psi_host psi_bxh)
          stats(percent=(2.5 5 50 95 97.5))=all simreport=5
          diagnostics=(all geweke(f1=0.3 f2=0.3));
```

```
     ods select PostSummaries ESS Geweke PostIntervals TADpanel;
     array gamma[3];
     array z[3];
     parms beta 0 gamma: 0;
     prior beta ~ normal(mean=0,var=1e15);
     prior gamma: ~ normal(mean=0, var=1e15);
     beginprior;
        psi_00=logistic(beta);
        psi_01=logistic(beta+gamma2);
        psi_10=logistic(beta+gamma1);
        psi_11=logistic(beta+gamma1+gamma2+gamma3);
        psi_breed=(psi_10+psi_11)-(psi_00+psi_01);
        psi_host=(psi_01+psi_11)-(psi_00+psi_10);
        psi_bxh=(psi_01+psi_10)-(psi_00+psi_11);
     endprior;
     eta=x*beta;
     do k=1 to 3;
        eta=eta+z[k]*gamma[k];
     end;
     mu = logistic(eta);
     model m ~ binomial(n = n, p = mu);
run;
ods graphics off;

proc export data=xx.postsample outfile=aa dbms=csv replace;

run;
```

2.5.2.4 Result

The summary statistics for all the parameters, including functions of the parameters, are listed in Table 2.5. The three most important parameters are $\gamma = \{\gamma_1, \gamma_2, \gamma_3\}$, which are translated into three comparisons of the average germination rates $\psi = \{\psi_{breed}, \psi_{host}, \psi_{b\times h}\}$. The equal-tail credibility intervals and the HPD intervals all show that the two breeds of O.a plants have no difference in germination rate, i.e., γ_1 and ψ_{breed} are not different from zero. The two host plants (bean and cucumber) have significant effects on the germination rate, i.e., γ_2 and ψ_{host} are significantly different from zero. The interaction effect is also significant, i.e., γ_3 and $\psi_{breed\times host}$ are different from

zero.

Table 2.5: Summary statistics of the posterior sample for the seeds data.

			Standard	Equal-Tail-Interval		HPD(95%)	
Parameter	Median	Mean	Deviation	2.5%	97.5%	HPD(left)	HPD(right)
β	-0.5596	-0.5603	0.1222	-0.8059	-0.3284	-0.7991	-0.3243
γ_1	0.1399	0.1372	0.219	-0.2903	0.5508	-0.2707	0.5674
γ_2	1.3175	1.3221	0.1738	0.9877	1.6703	0.981	1.6548
γ_3	-0.772	-0.7747	0.3095	-1.3689	-0.1827	-1.3622	-0.181
ψ_{00}	0.3637	0.3639	0.0282	0.3088	0.4186	0.3069	0.4162
ψ_{01}	0.6813	0.6812	0.0264	0.6267	0.7322	0.6313	0.7349
ψ_{10}	0.3962	0.3965	0.0433	0.319	0.4858	0.3176	0.4819
ψ_{11}	0.531	0.5308	0.0426	0.4494	0.6131	0.4559	0.6182
ψ_{Breed}	-0.1193	-0.1178	0.071	-0.2584	0.0164	-0.2505	0.0219
ψ_{Host}	0.4528	0.4515	0.0725	0.3102	0.5928	0.3185	0.5955
$\psi_{B\times H}$	0.1811	0.183	0.0732	0.0448	0.3242	0.0485	0.3269

Table 2.6 shows the convergence test for all the unknowns. The Markov chains behave very well for all the unknowns. The p-values for the Geweke z-test are larger than 0.05 for all unknowns. The effective sample sizes are very close to the actual posterior sample sizes. Figure 2.5-2.7 gives the posterior TAD panels for γ_1,γ_2 and γ_3 respectively. The autocorrelations are all very small. Overall, this data set is sufficient to allow more precise estimates of the parameters.

Table 2.6: Diagnostic test statistics for the Markov chain convergence of the seeds data.

Parameter	Geweke-z	p-value	SS	ESS	Correlation time	Efficiency
β	1.8167	0.0693	1000	1000.0	1.0000	1.0000
γ_1	-0.7885	0.4304	1000	974.5	1.0261	0.9745
γ_2	-1.6777	0.0934	1000	964.6	1.0367	0.9646
γ_3	0.4744	0.6352	1000	908.5	1.1007	0.9085
ψ_{00}	1.8031	0.0714	1000	1000.0	1.0000	1.0000
ψ_{01}	-0.4548	0.6493	1000	867.6	1.1526	0.8676
ψ_{10}	0.1273	0.8987	1000	902.5	1.1081	0.9025
ψ_{11}	-0.5571	0.5775	1000	1034.3	0.9668	1.0343
ψ_{Breed}	-0.7166	0.4736	1000	975.8	1.0248	0.9758
ψ_{Host}	-1.2368	0.2162	1000	1000.0	1.0000	1.0000
$\psi_{B\times H}$	-0.4182	0.6758	1000	907.1	1.1024	0.9071

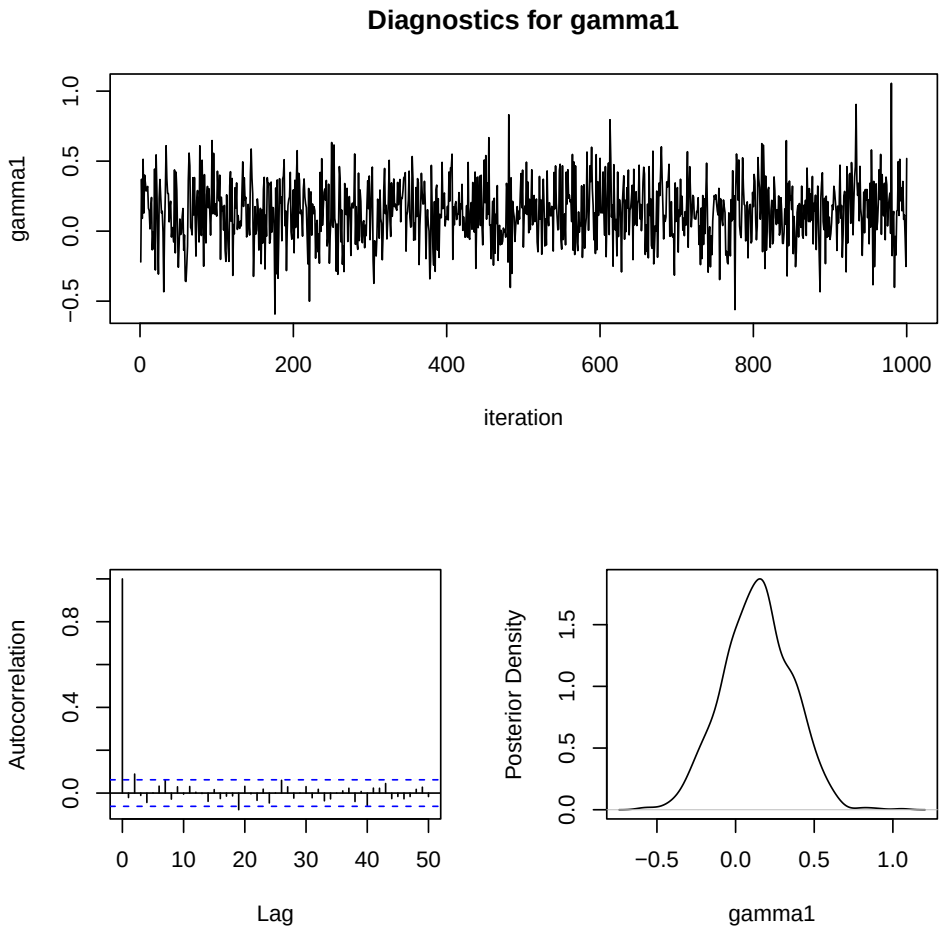


Figure 2.5: The posterior TAD panels for parameter gamma1 (γ_1) of the seeds data.

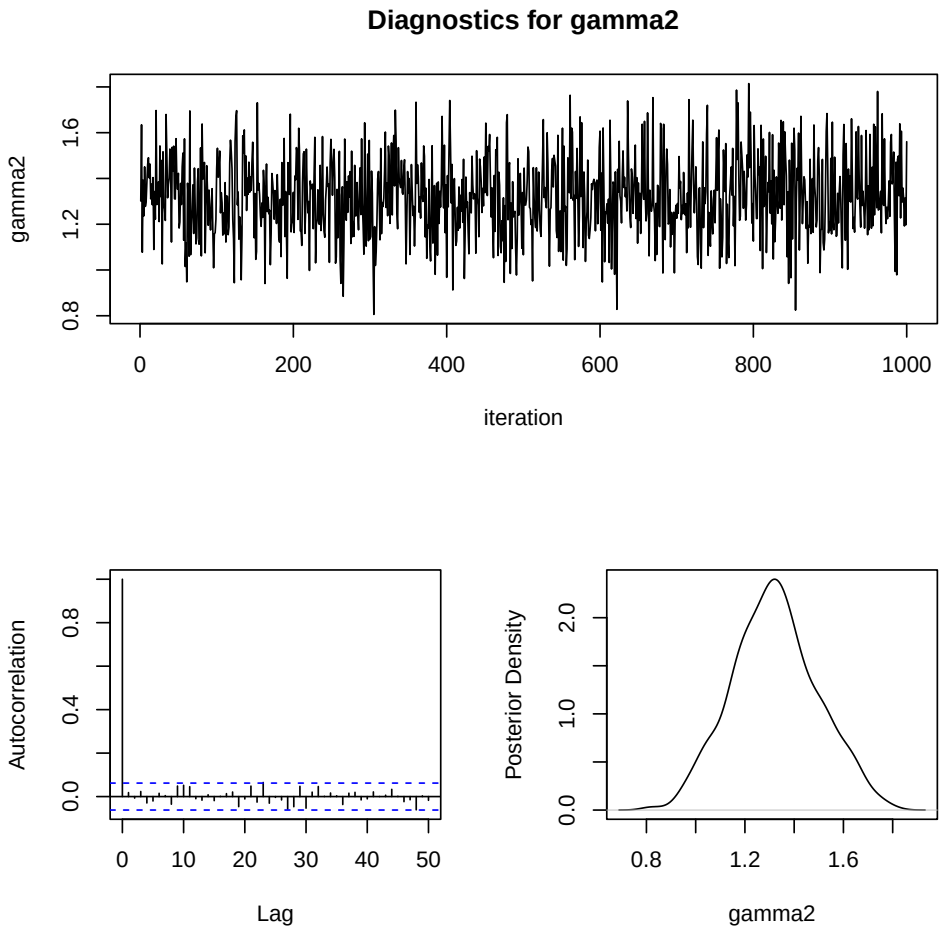


Figure 2.6: The posterior TAD panels for parameter gamma2 (γ_2) of the seeds data.

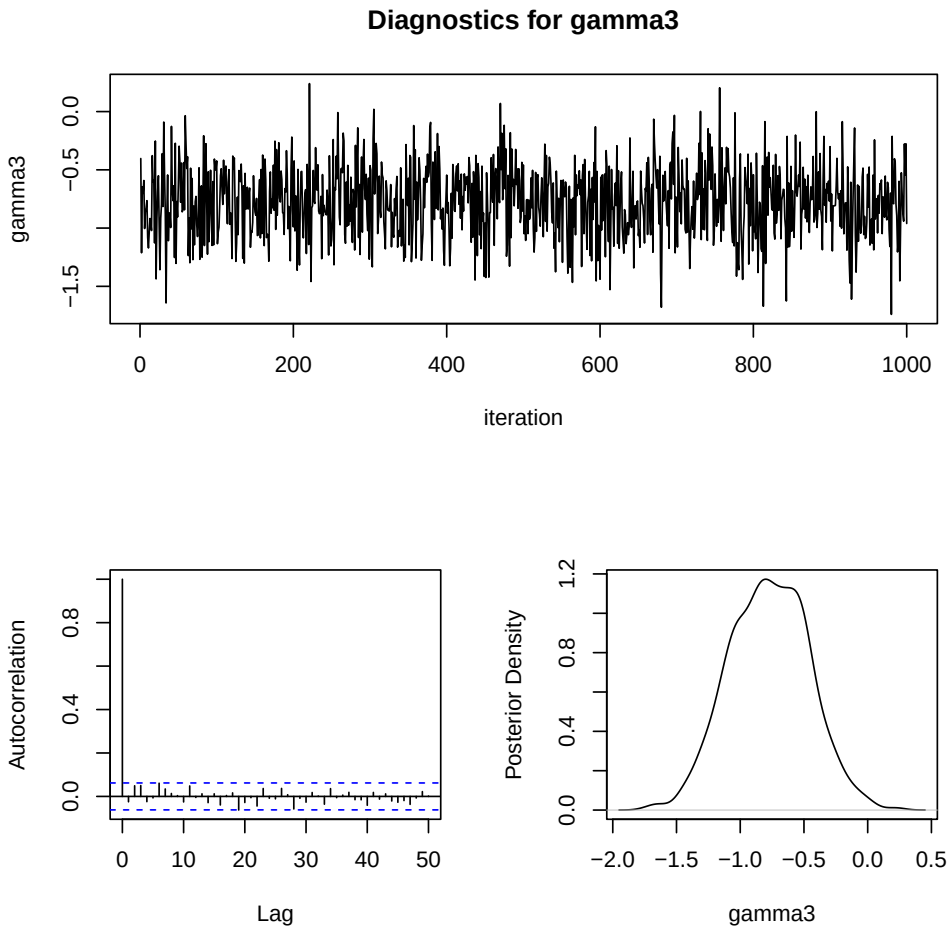


Figure 2.7: The posterior TAD panels for parameter gamma3 (γ_3) of the seeds data.

2.5.3 The fertility data

This example demonstrates the application of the generalized linear model to quantitative trait locus (QTL) mapping for binary traits in wheat. The experiment was conducted by Dou et al. (2009) who made the data available to us for this analysis. A female sterile line XND126 and an elite cultivar Gaocheng8901 with normal fertility were crossed for genetic analysis of female sterility measured as a binary trait. The parents, their F_1 and F_2 progeny were planted at the Huaian experimental station in China for the 2006-2007 growing season under the normal autumn sowing condition. The mapping population was an F_2 family consisting of 243 individual plants. The binary trait was the presence of seed setting of the female plants. The trait is called the wheat fertility and thus the dataset is named the "fertility data". About 84% of the F_2 progeny had seeded spikelets (phenotype 1) and the remaining 16% plants did not have seeded spikelets (phenotype 0). This is a typical binary trait regarding the presence of seeds. A total of 28 SSR markers were used in this experiment. These markers covered 5 chromosomes of the wheat genome with an average genome marker density of 15.5 cM per marker interval. The five chromosomes are only part of the wheat genome.

These chromosomes were scanned for QTL of the binary trait using the MCMC implemented Bayesian method. The purpose of QTL mapping was to identify chromosome regions that are associated with the binary fertility trait. The dependent variable is the binary trait phenotype while the independent variables are numerically coded genotype indicator variables for the part of genome under investigation. We emphasize the advantage of the Bayesian analysis over the frequentist method for detecting multiple QTL simultaneously within a single model. To conduct the multiple locus analysis, we placed one pseudo marker in every 5 centiMorgan (cM) of the genome. This generated 75

pseudo markers for the five chromosomes. Therefore, we have a total of 75 independent variables. For each independent variable, the numerically coded value was the difference between the conditional probabilities of the two homozygote genotypes. Let A_1A_1, A_1A_2 and A_2A_2 be the three genotypes for the kth pseudo marker of the genome. The numerically coded value for the locus is

$$Z_{jk} = p(G_{jk} = A_1A_1|\text{marker}) - p(G_{jk} = A_2A_2|\text{marker}) \tag{2.70}$$

for $k = 1, ..., 75$. The map of the 75 pseudo markers, the phenotypic values (binary phenotypes) of the 234 plants and the 75 numerically coded independent variables can be downloaded from our website (www.statgen.ucr.edu) in the Bayesian Review software section.

2.5.3.1 Model

The observed binary trait is $r_j = \{0, 1\}$ for $j = 1, ..., n$, where $n = 243$ is the sample size. Let $E(r_j) = \mu_j$ be the expectation of the binomial data point, which is connected to the $m = 75$ independent variables through

$$\eta_j = X_j\beta + \sum_{k=1}^{m} Z_{jk}\gamma_k \tag{2.71}$$

where $X_j = 1$ for all $j = 1, ..., n$, β is the intercept, γ_k is the effect of the kth pseudo marker and Z_{jk} is the conditional expectation defined earlier. We are interested in estimating the parameter vector $\theta = \{\beta, \gamma_1, ..., \gamma_{75}\}$.The relationship between μ_j and η_j is through the probit link,

$$\eta_j = \text{probit}(\mu_j) = \Phi^{-1}(\mu_j) \tag{2.72}$$

More intuitively, the inverse of the probit link is

$$\mu_j = \Phi(\eta_j) = \Phi\left(X_j\beta + \sum_{k=1}^{m} Z_{ij}\gamma_k\right) \tag{2.73}$$

The likelihood function (model) of the jth data point is

$$p(r_j|\mu_j) = \text{Bernoulli}(r_j|\mu_j) = \text{Binary}(r_j|\mu_j) = \mu_j^{r_j}(1-\mu_j)^{1-r_j} \tag{2.74}$$

2.5.3.2 Prior and posterior

The intercept is assigned a flat normal prior, i.e.,

$$\pi(\beta) = \text{Normal}(\beta|0, 10^{15}) \tag{2.75}$$

Each of the QTL (pseudo marker) effect is assigned a normal prior,

$$\pi(\gamma_k) = \text{Normal}(\gamma_k|0, \sigma_k^2) \tag{2.76}$$

which is QTL specific, i.e., each QTL has its own prior variance. The variance in the prior is assigned a higher level prior (hierarchical prior),

$$\pi(\sigma_k^2) = \text{Inv} - \chi^2(\gamma_k|\tau, \omega) = \text{Inv} - \chi^2(\gamma_k|10^{-10}, 10^{-10}) \tag{2.77}$$

This hierarchical model is also called the Bayesian shrinkage analysis . There is no closed form of the fully conditional posterior distribution for β and γ_k. However, given γ_k, the fully conditional posterior distribution of σ_k^2 remains scaled inverse chi-square,

i.e.,

$$p(\sigma_k^2|...) = \text{Inv} - \chi^2(\sigma_k^2|\tau+1, \omega+\gamma_k^2) = \text{Inv} - \chi^2(\gamma_k|10^{-10}+1, 10^{-10}+\gamma_k^2) \quad (2.78)$$

2.5.3.3 SAS code

The Bayesian shrinkage analysis for the generalized linear model is new and has never been used to map QTL for the binary fertility trait of wheat. We are not so sure about the efficiency of the method. Therefore, we analyzed the fertility data using both the MCMC procedure under the multiple QTL model and the GENMOD procedure under the single QTL model. The GENMOD procedure is the fixed model version of the generalized linear model producing the maximum likelihood estimates of the parameters.

PROC MCMC

```
%let dir=C:\Bayes\fertility;

libname xx "&dir";
filename aa "&dir\fertility.csv" lrecl=200000;
filename bb "&dir\post-sample.csv";

data fertility;
   infile aa dlm=',' firstobs=2;
   input plant fert_rat fert_bin z1-z75;
   r=fert_bin;
run;

%macro fertility;
ods graphics on; proc mcmc data=fertility outpost=xx.postsample
seed=12345
          nmc=50000 thin=50 nbi=5000 ntu=3000
          monitor=(beta gamma1-gamma5 sigmasqr1-sigmasqr5)
          stats(percent=(2.5 5 50 95 97.5))=all simreport=5
          diagnostics=(all geweke(f1=0.3 f2=0.3));
     ods select PostSummaries ESS Geweke PostIntervals TADpanel;
     array gamma[75];
     array sigmasqr[75];
     array z[75];
```

```
     parms beta 0;
     prior beta ~ normal(mean=0,var=1e15);
     %do k=1 %to 75;
         parms gamma&k 0;
         parms sigmasqr&k 1;
         prior gamma&k ~ normal(mean=0, var=sigmasqr&k);
         prior sigmasqr&k ~ sichisq(1e-10,1e-10);
     %end;
     eta=beta;
     do k=1 to 75;
        eta=eta+z[k]*gamma[k];
     end;
     mu = probnorm(eta);
     model r ~ binary(mu);
run;
ods graphics off;
%mend;

%fertility

proc export data=xx.postsample outfile=bb dbms=csv replace;

run;
```

Here is a brief explanation of the code. Since the data set is relatively large, it was stored in an excel file as an external file. In the input data (fertility data set), there are two fertility phenotypes, one is the ratio of the number of seeded spikelets to the total number of spikelets named fert_rat and the other is the binary seed presence and absence trait named fert_bin. We only analyzed the binary trait renamed r = fert_bin. Since the data set is large along with a large model, the MCMC procedure took much longer time to finish. We put an option in the proc mcmc statement called simreport=5. This option tells proc mcmc to report 5 times about the progress of program running to allow the programmer to monitor the remaining running time. For example, this data set took about 22 hours to finish and thus the program delivered a message in about every 4 hours on the SAS log to report the progress.

Another important difference between this program and the previous ones is the

use of SAS macro for handling large model. We used the array statement to define the gamma vector $\gamma = \{\gamma_1, ..., \gamma_{75}\}$ and the sigma^2(sigmasqr) vector $\sigma^2 = \{\sigma_1^2, ..., \sigma_{75}^2\}$. The problem with the MCMC procedure is that it does not take the following do loop,

```
do k=1 to 75;
    parms gamma[k] 0;
    parms sigmasqr[k] 1;
    prior gamma[k] ~ normal(mean=0, var=sigmasqr[k]);
    prior sigmasqr[k] ~ sichisq(1e-10,1e-10);
end;
```

The parms and prior statements only take the following forms of the do loop,

```
%do k=1 %to 75;
    parms gamma&k 0;
    parms sigmasqr&k 1;
    prior gamma&k ~ normal(mean=0, var=sigmasqr&k);
    prior sigmasqr&k ~ sichisq(1e-10,1e-10);
%end;
```

which must be executed as a macro.

Proc GenMod

```
%let dir=C:\Bayes\fertility;
libname xx "&dir";

filename aa "&dir\fertility.csv" lrecl=200000;

filename bb "&dir\genmod_out.csv" lrecl=200000;

data one;
     infile aa dlm=',' firstobs=2;
     input plant fert_rat fert_bin z1-z75;
     event=fert_bin;
     trial=1;
run;

proc genmod data=one;
     model event/trial =z1  / dist = bin
           link = probit lrci type1;
     ods output ParameterEstimates=parms Modelfit=fitness Type1=Type1;
run;

%macro genmod;
%do i=2 %to 75;
```

```
proc genmod data=one;
     model event/trial = z&i / dist = bin
           link = probit lrci type1;
     ods output ParameterEstimates=parms1 Modelfit=fitness Type1=type1;
run;
proc append base=parms data=parms1;
%end;
%mend genmod;

%genmod

data xx.parms;
   set parms;
   if (parameter ^= "Intercept" and parameter ^="Scale");
run;

proc export data=xx.parms outfile=bb dbms=csv replace; run;
```

2.5.3.4 Result

The burn-in period was 5000, after which one observation was collected in every 50 iterations until the posterior sample size reached 1000. Therefore, the total length of the Markov chain was $5000 + 50 \times 1000 = 55000$. The MCMC procedure took about 22 CPU hours to complete the MCMC sampling. About half of the computing time was spent on tuning the parameter of the proposal distribution, trying to reach the target acceptance probability.

We did not monitor all the parameters except the intercept and the first five pseudo markers. The posterior TAD panels for the intercept (beta or β) are presented in Figure 2.8, clearly showing that the chain has converged. Figure 2.9 shows the posterior means of the 75 pseudo markers (QTL) effects plotted against the genome location (panel a) and LOD (log of odds) profile (panel b). The LOD score for a particular pseudo marker was calculated as

$$\text{LOD}(\gamma_k) = \frac{\hat{\gamma}_k^2}{4.61 \times \text{var}(\gamma_k|\text{data})} \tag{2.79}$$

where $\hat{\gamma}_k$ is the posterior mean and var(γ_k|data) is the posterior variance of γ_k. The QTL effect profile indicates a major QTL on chromosome 2 and a few minor QTL on chromosome 5. However, the LOD score profile shows that only one QTL exits, that is the major one on chromosome 2. The other minor QTL all have large posterior standard errors and thus have very small LOD scores. Note that the LOD score cannot be used as significance test in Bayesian analysis. It only indicates a major QTL.

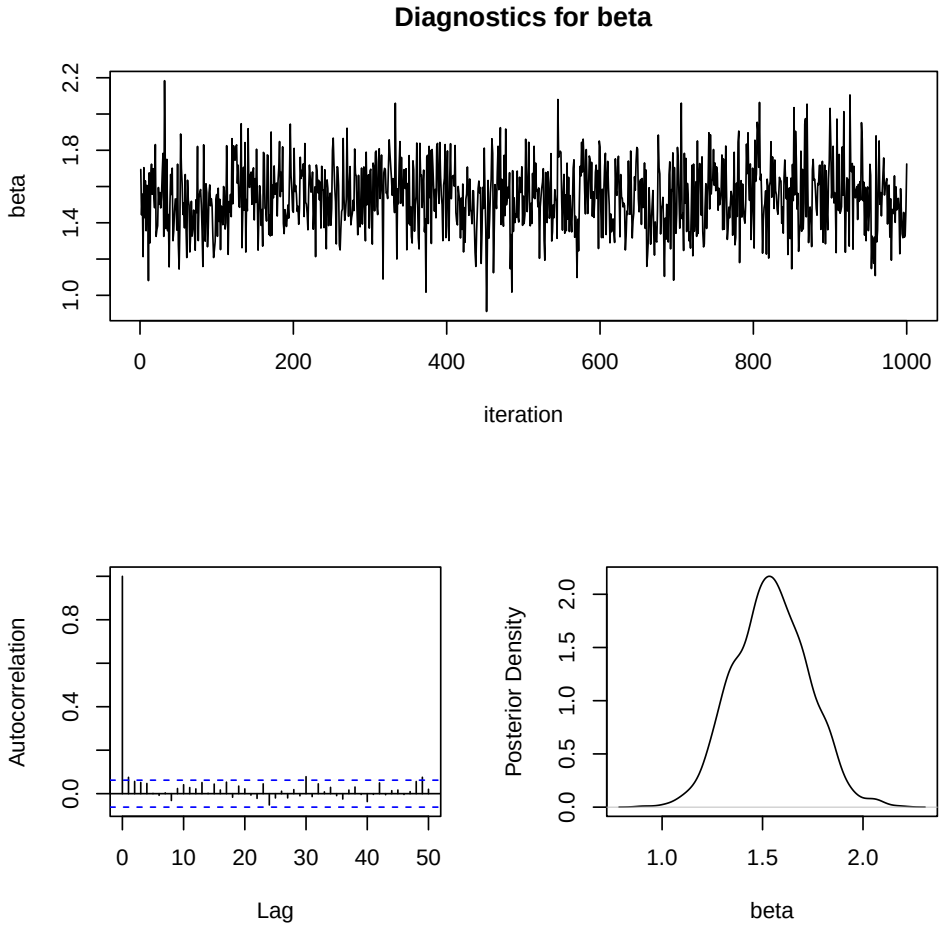


Figure 2.8: The posterior TAD panels for parameter beta (β) of the fertility data.

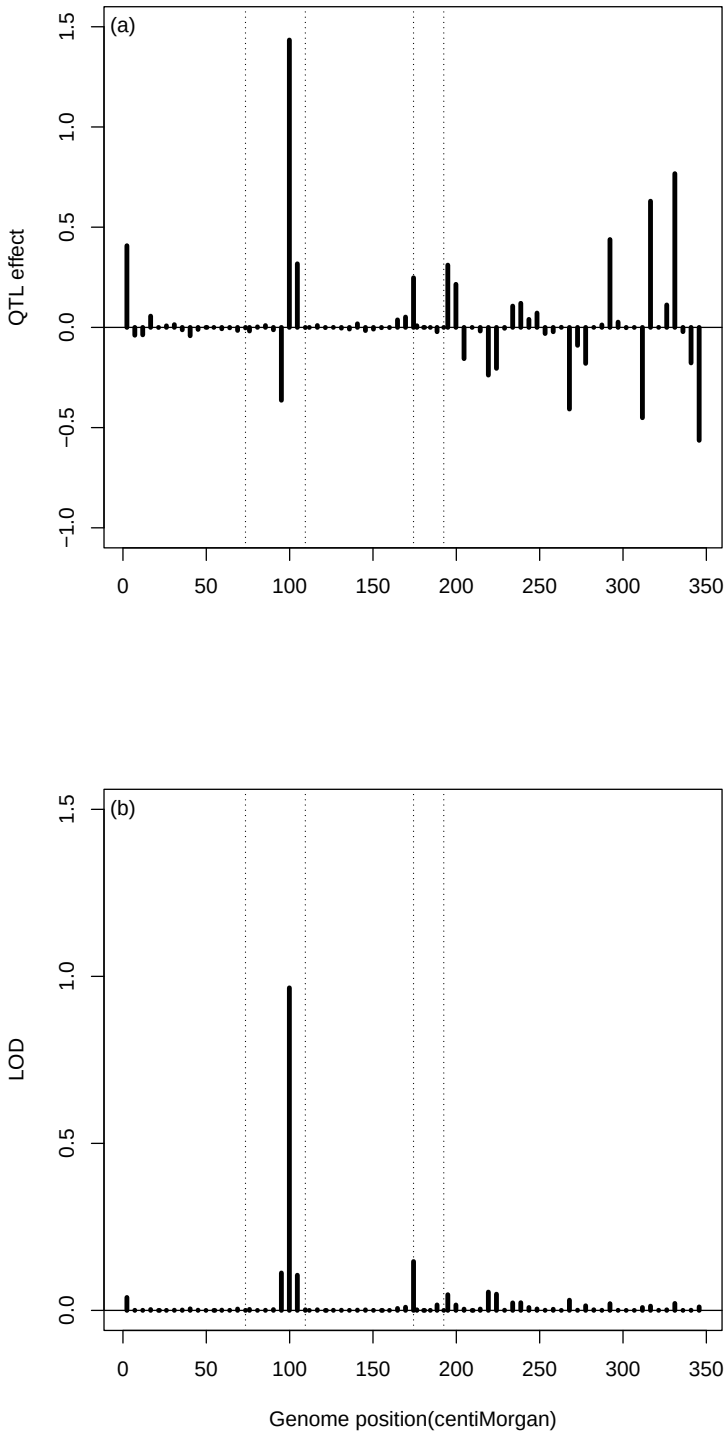


Figure 2.9: Posterior mean of QTL effect (panel a) and LOD score (panel b) plotted against the genome location of the wheat fertility trait (the fertility data) from the MCMC implemented Bayesian analysis (multiple QTL model). The five chromosomes are separated by the dotted reference lines.

To further validate the major QTL on chromosome 2, we re-analyzed the data using a single QTL model to scan the entire genome for the 75 pseudo markers. The maximum likelihood method was implemented using the GENMOD procedure of SAS (see the code given in the previous section). The corresponding QTL effect and LOD profiles are given in Figure 2.10, showing that a large peak occurs on chromosome 2 also. The peak position is off by one pseudo marker (5 cM) away from that of the Bayesian analysis.

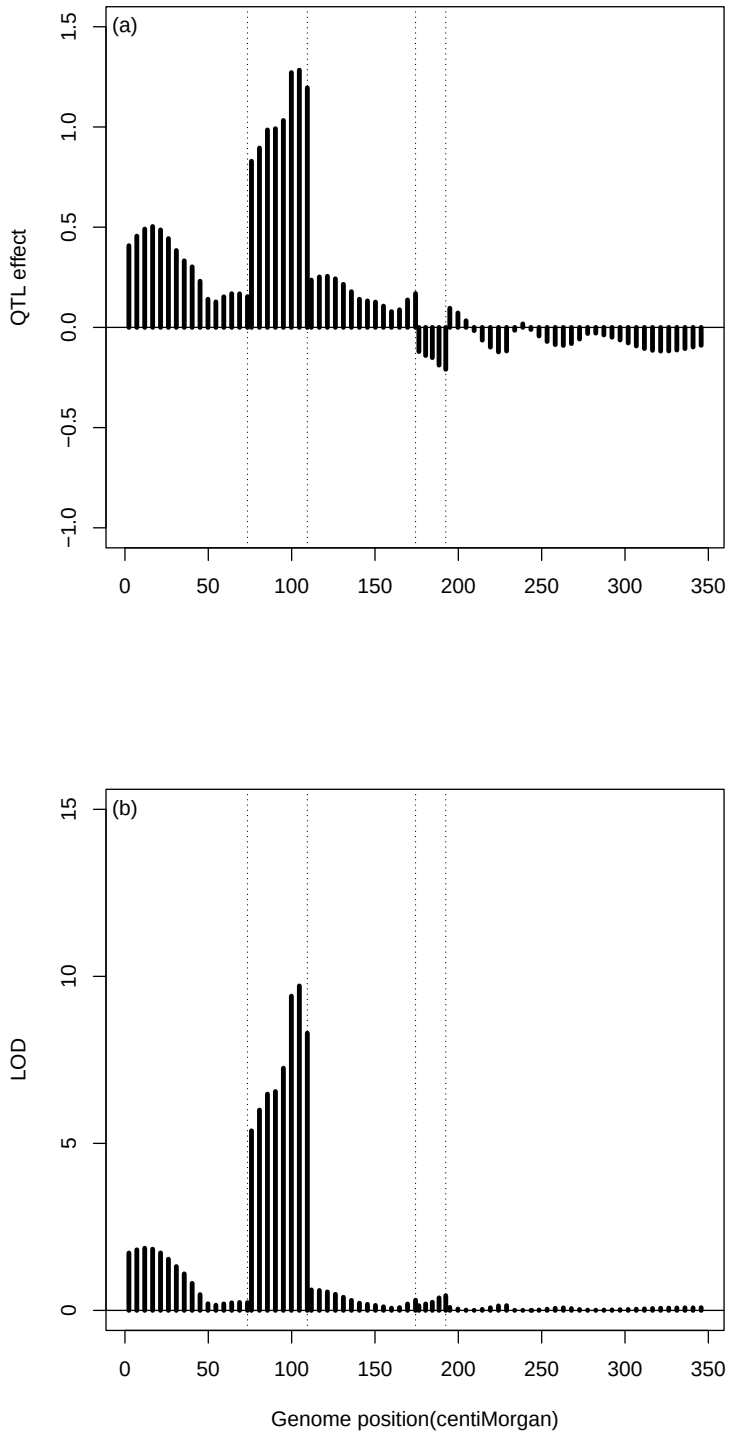
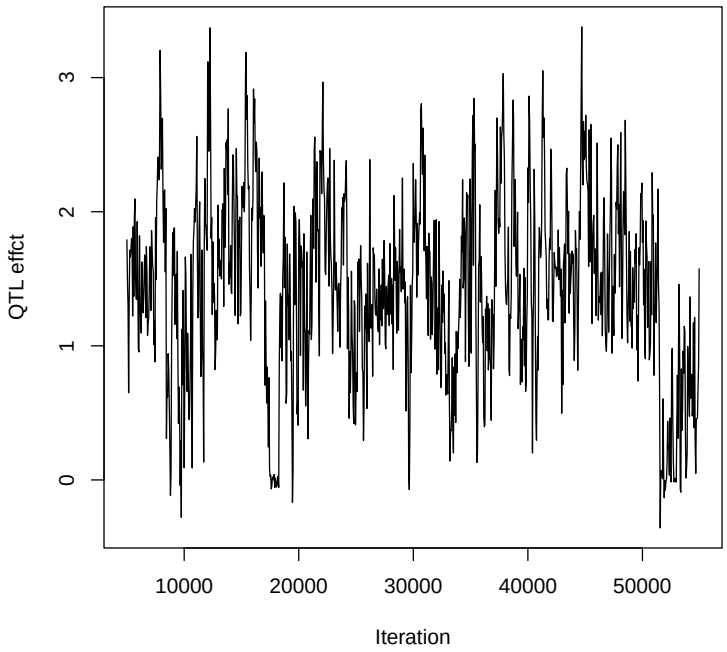


Figure 2.10: The estimated QTL effect (panel a) and LOD score (panel b) plotted against the genome location of the wheat fertility trait (the fertility data) from the maximum likelihood analysis implemented in the generalized linear model (single QTL model). The five chromosomes are separated by the dotted reference lines.

We now focus on the major QTL on chromosome 2 and present some details of this QTL. Figure 2.11 gives the trace plot and the posterior density for the QTL. The Markov chain reached the stationary distribution, but did not mix well (panel a). The posterior distribution fits a normal distribution very well (panel b). The QTL is represented by pseudo marker number 22 (γ_{22}) with the posterior mean $\pm$ standard deviation $\hat{\gamma}_{22} = 1.4338 \pm 0.6796$. The posterior median is 1.4579, almost overlapping with the posterior mean. The 5% equal tail (credible) interval is (0.00881,2.7167) and the 10% equal tail interval is (0.0901,2.5371). The 95% HPD interval and the 90% HPD are the same as the equal tail intervals due to the normality of the posterior distribution. Neither interval covers zero, meaning that the QTL is real with very high credibility.



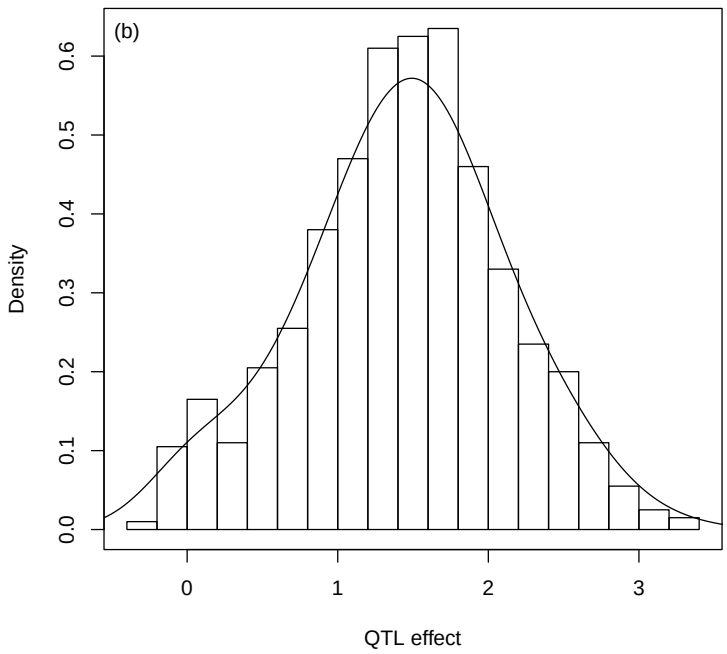


Figure 2.11: The trace plot (panel a) and the posterior density (panel b) of the QTL detected in chromosome 2 for the binary fertility trait of wheat (the fertility data).

2.6 Discussion

The original Bayesian method was more complicated than the classical maximum likelihood method because multiple integrals are often involved. In most situations, explicit form of the multiple integrals does not exist, and thus limits the application of Bayesian analysis. Although Bayesian inference was proposed earlier than the maximum likelihood inference, it becomes popular only more recently due to the advent of high computing power and the advanced MCMC algorithm for numerical integrations. With the MCMC implemented Bayesian method, we often adopt hierarchical models to fit a complicated data and partition the complicated models to many small parts, each is described by a simple model. We then draw one or a few parameters with other parameters fixed at values already drawn previously. Because the fully conditional posterior distribution is often very simple, the MCMC process is much easier to understand than the maximum likelihood method. The MCMC algorithm has revolutionarized the field of Bayesian inference. Thanks to the MCMC algorithm, we, the non-statisticians, can also perform Bayesian analysis. Conducting an MCMC sampling process is no more complicated than doing a field experiment. The MCMC turns a statistician into an experimentalist who plants the "seeds" of parameters and let them "grow" into full "parameters". The only difference is that the MCMC experimentalist does the experiment in computers rather than in the actual field.

The MCMC experiment is more of an art than a science, requiring a lot of experience before the first successful MCMC experiment happens. Without closely monitoring the convergence process of the MCMC sampler, the result can be misleading. This is why the WinBUGS software always prompts a warning message - "MCMC sampling can be dangerous!". However, we should not be discouraged by the warning message. People

should try to use different priors and use longer chains to draw the posterior samples. It does not cost anything other than a little longer CPU time to run more and longer MCMC experiments. Even if the results are bad and wrong, it is not "life threatening". The time when caution should be taken is when we are ready to report the result. It is always a good idea to compare the result with that of the ML analysis if possible. If the two results are not comparable or completely different, extra caution should be taken or more analyses with different models and different priors should be conducted before the final report. Analyzing simulated data is another way to verify the model and the priors. In the fertility data analysis, we used a new model under new priors, and produced result that has never been reported before. Therefore, we analyzed the same data using PROC GENMOD under the single QTL model. The two results do share some similarity. Such a comparison increased our confidence on the Bayesian analysis.

Chapter 3

Significance Test and Genome Selection in Bayesian Shrinkage Analysis

3.1 Introduction

Interval mapping (Lander and Botstein 1989) and multiple interval mapping (Kao et. al. 1999) are the most commonly used methods for QTL mapping. These methods are developed in the maximum likelihood framework, which has limitation in terms of handling large saturated models. Bayesian mapping (Satagopan et. al. 1996, Sillanpää and Arjas 1998, Yi and Xu 2000, Yi and Xu 2000, Yi and Xu 2001) deals with large models more efficiently through the reversible jump Markov chain Monte Carlo (RJMCMC) (Sillanpää 1998) , the shrinkage analysis (Wang et. al. 2005, Xu 2003) or the stochastic search variable selection (SSVS) (Yi et. al. 2005) . Shrinkage mapping and SSVS are more efficient in terms of whole genome evaluation because they are statistically easy

to understand and also provide better chance to evaluate the entire genome. These two methods are related to the Lasso method for regression analysis (Tibshirani 1996). Rather than deleting non-significant QTL explicitly from the model, these methods use a special algorithm to shrink estimated QTL effects to zero or close to zero. A QTL with zero estimated effect is treated the same as being excluded from the model. No statistical test is required because genome regions bearing no QTL often show no bumps (QTL effects) in the QTL effect profile (plot of QTL effects against genome location). The visual inspection on the QTL effect profile is not optimal because small QTL may come and go during the MCMC sampling process. It is desirable to provide some kind of statistical confidence on these small QTL.

Permutation test (Churchill and Doerge 1994) itself is not a method of QTL mapping; rather, it is a method to find the critical value used to declare the significance of QTL for any method of QTL mapping. It is very efficient in interval mapping under the maximum likelihood framework. A new resampling method was developed by Zou et al. (Zou et. al. 2004) for significance test under the composite interval mapping or other multiple effect based QTL mapping schemes. The new resampling method is computationally less intensive and may perform better than the permutation test. However, it has not been as popular as we would have thought. The reason for this is perhaps due to the fact that the theory behind the method is not straightforward to most QTL mapping experimentalists. The permutation test, although time consuming, does not requires any theory and easy to understand. People tend to trust a simple method they understand, rather than a comprehensive method they do not, even if the simple method is suboptimal. Therefore, the permutation test remains the most popular method for finding the critical value of a test statistic for QTL detection. Kopp et al. (Kopp et. al. 2003) applied the permutation test to determine empirical thresholds for Bayesian

shrinkage mapping. The problem with such a test for the MCMC implemented Bayesian mapping is the heavy computational burden. Each MCMC run may take one or a few hours to complete for a reasonable sample size of the mapping population. Performing thousands of permutation analyses is not realistic for the Bayesian method. Therefore, improvement of the permutation test applied to Bayesian analysis is required. This is the first objective of this study.

Broman and Speed (Broman and Speed 2002) treated multiple QTL mapping as a model (variable) selection problem and developed a new method called BIC_δ. More recently, Manichaikul et al. (Manichaikul et. al. 2009) extended the Broman and Speed (Broman and Speed 2002) model selection by allowing epistatic (non-allelic interaction) effects to be included in the model. They called the extended model selection method the penalized LOD score method (pLOD). Two versions of the penalized LOD score method were investigated, one is called heavy penalized LOD score (pLOD_H) and the other called the light penalized LOD score (pLOD_L). With this new notation, the original BIC_δ of Broman and Speed (Broman and Speed 2002) was renamed as pLOD_a, penalized LOD score for additive effects only. The authors compared these methods along with two other BIC-based methods and the Bayesian model selection method of Yi, Xu and Allison (Yi, Xu and Allison 2003) using both simulated data and real data. They concluded that the pLOD methods including epistatic effects and the Bayesian model selection method outperformed other methods in most cases they evaluated.

The model selection methods are alternative method of QTL analysis. They cannot replace the Bayesian shrinkage analysis because the two have quite different purposes. Model selection aims to detecting QTL while Bayesian shrinkage focuses on genome evaluation. We realized that if the Bayesian shrinkage analysis is accompanied with a significance test, it can serve both QTL detection and genome selection. The original

Bayesian shrinkage analysis (Wang et. al. 2005, Xu 2003) has no significance test associated with the method because the entire genome was evaluated simultaneously in a single model. More recently, researchers, especially animal and plant breeders, became interested in genome selection (Weuwissen et. al. 2001, Weuwissen et. al. 2009) using the Bayesian method. Application of genome selection to laboratory mice (Legarra et. al. 2008) and human (Lee et. al. 2008) were also reported. Genome selection does not require statistical tests because QTL of the entire genome, regardless the sizes, are included to predict the genomic effect of individuals. However, there is no report so far to investigate whether inclusion of small QTL will benefit genome selection. Cross validation can be used to determine how large a QTL should be included in genome selection. This is the second aim of this study.

3.2 Methods

3.2.1 Model

For the paper to be self contained, we briefly introduce the Bayesian shrinkage model here. Let y_j be the phenotypic value of a quantitative trait measured for individual j for $j = 1, ..., n$, where n is the sample size. Suppose that the individual is genotyped for m markers, which are more or less evenly distributed across the genome. Let X_{jk} be the genotype indicator variable for individual j at marker k, for $k = 1, ..., m$. The linear model describing the relationship between the phenotype and the genotypes of markers is

$$y_j = b_0 + \sum_{k=1}^{m} X_{jk} b_k + e_j \tag{3.1}$$

where b_0 is the intercept, b_k is the QTL effect for marker k and e_j is the residual error with an assumed $N(0, \sigma^2)$ distribution. The reason that the Bayesian shrinkage method can handle large m is the prior distribution assigned to each QTL effect,

$$p(b_k) = N(b_k|0, \sigma_k^2) \tag{3.2}$$

where σ_k^2 is a QTL specific prior variance. This prior alone is not sufficient to generate the desired shrinkage estimate of QTL effect. A hierarchical model with a higher level of prior assignment is necessary, in which the prior variance σ_k^2 is further assigned a scaled inverse chi-square distribution,

$$p(\sigma_k^2) = \text{Inv} - \chi^2(\sigma_k^2|\tau, \omega) \tag{3.3}$$

In the original shrinkage analysis, Xu (Xu 2003) set $\tau = \omega = 0$, leading to $p(\sigma_k^2) = 1/\sigma_k^2$. Ter Braak et al (ter Braak et. al. 2005) claimed that this prior is improper and leads to an improper posterior distribution. They revised the prior so that the posterior distribution becomes proper. Their revised prior is

$$p(\sigma_k^2) = \text{Inv} - \chi^2(\sigma_k^2| - 2\delta, 0) \propto \frac{1}{(\sigma_k^2)^{1-\delta}} \tag{3.4}$$

where $0 < \delta \leq 0.5$. If $\delta = 0$, this revised prior would be equivalent to Xu's (Xu 2003) vague prior. However, Xu's vague prior is just excluded from the revised prior. In this study, we used the proper prior of Ter Braak et al (ter Braak 2005) , as a precaution to avoid any potential problems caused by the improper posterior distribution of σ_k^2.

3.2.2 Permutation between Markov chains

In the MCMC implemented Bayesian shrinkage analysis, Xu (Xu 2003) plotted the estimated QTL effects against the genome location. We could have plotted a test statistic, say a t-test or an F-test, against the genome location. Unfortunately, the test statistic requires the posterior standard deviation of each sampled QTL. The empirical posterior standard deviation highly depends on the thinning rate of the Markov chain and thus is always underestimated due to possible autocorrelation. Therefore, we prefer to use the QTL effect profile rather than a test statistic profile. To determine the threshold values for the QTL effects under the null model, we employed a permutation test just like frequentists do in interval mapping (Churchill and Doerge 1994) . Let $y = \{y_j\}$ be the vector of the phenotypic values ordered according to the individuals' natural identification numbers, i.e., the original data set where the individuals' phenotypes match their marker genotypes in the files. Let $y^* = \{y_j^*\}$ be a randomly rearranged vector of phenotypes, called a permutation, in which the phenotypes do not match the marker genotypes. Performing a Bayesian shrinkage analysis on the permuted data by running a Markov chain with a desired length, we obtain a posterior sample for all the parameters. For the parameters of interest, say the QTL effects, we record their values and save them in a file as one observation from one permutation analysis. The permutation analysis is repeated independently for a thousand times, we then obtain a thousand observations for each of the interested parameters (QTL effects). This sample contains observations from the empirical distribution of the null model (no QTL effects). The $\frac{1}{2}\alpha \times 100\%$ and $(1 - \frac{1}{2}\alpha) \times 100\%$ percentiles of a parameter in the thousand permuted samples are the empirical critical values used to declare statistical significance for a QTL in the analysis of the original data set (phenotypes match the genotypes). This

permutation strategy was first applied by Kopp et al (Kopp et al. 2003) . This so called "permutation outside the Markov chain" approach is the traditional application of the permutation test (Churchill and Doerge 1994) to the Bayesian analysis. The problem with this strategy is the extensive CPU time. Each MCMC run may take an hour or so and a complete permutation experiment consisting of 1000 permutation analyses may take a month computing time. Therefore, we will invent a more efficient permutation method to replace this traditional method of permutation.

3.2.3 Permutation within Markov chain

As the name of the method implies, this permutation strategy permutes the phenotypes in every h-th iteration within a Markov chain, where $1 \leq h \leq L$ and L is the length of the Markov chain. If $h = L$, this approach is equivalent to the permutation-between-chains approach. If $h = 1$, we permute the phenotype in every iteration. The approach is implemented as follows. For each iteration, after all parameters are sampled, the phenotypes are reshuffled before the next round of sampling starts. The total length of the chain is not necessarily longer than a regular Markov chain for the un-shuffled data. Therefore, a complete data analysis requires only two chains, one for the original data and one for the reshuffled data. The reshuffled chain provides the $\frac{1}{2}\alpha \times 100\%$ and $(1 - \frac{1}{2}\alpha) \times 100\%$ percentiles used as critical values of the QTL effects.

The within-chain permutation is a strategy to generate the posterior distributions of the regression coefficients under the null model. If the genotypes do not match the phenotypes, the Bayesian estimates (posterior means) of the regression coefficients are expected to be zero across all loci. The posterior variances are determined by the residual variance and the variance of the genotypic indicator variables, which are preserved in the permuted sample, regardless how frequent the phenotypes are reshuffled. There is

not much theory behind this permutation test. We chose this test for the very reason of simplicity. As long as we can control the type I error for the entire genome and produce reasonable powers for all the large QTL, the permutation test should be admissible.

3.2.4 Genome selection

Genome selection aims to evaluate the genetic effect for the entire genome using dense markers for each individual. When all individuals in a population are evaluated, the genomic effects of different individuals can be compared and the "best" individuals are selected for breeding. How to combine the QTL mapping result with genome selection is an important but not yet answered question. We adopted a five-fold cross validation test (Tibshirani 1996) to answer this question. In the cross validation analysis, we partition the sample into five equal parts (subsamples). Each time, we use four parts ($4n/5$ individuals) to estimate the QTL effects and perform within-chain random shuffling to determine the empirical percentiles for QTL detection. Only significant QTL at the level are used to predict the total genomic effect for an individual in the remaining part ($n/5$ individuals). Note that that the training sample ($4n/5$ individuals) is used for parameter estimation and significance test and the testing sample ($n/5$ individuals) is used for prediction. The squared prediction error (PE) for the s-part is defined as

$$\Delta_s(\alpha) = \frac{5}{n}\sum_{j'=1}^{n/5}(y_{j'} - b_0 - \sum_{k=1}^{m} X_{kj'}b_k)^2 \tag{3.5}$$

where $y_{j'}$ is the phenotypic value of an individual in the test sample and j' indexes all individuals in the test sample. The intercept and the regression coefficients are estimated from the training sample. Note that $\hat{b}_k$ equals the shrinkage estimate if it passes the

thresholds and $\hat{b}_k = 0$ otherwise. The overall PE for the cross validation test is

$$\mathrm{PE}(\alpha) = \frac{1}{5}\sum_{s=1}^{5}\Delta_s(\alpha) \tag{3.6}$$

We vary α from 0 to 1 incremented by 0.5. The α value that minimizes the PE is the optimal one used as the criterion of QTL inclusion for genome selection.

3.3 Results and Discussion

3.3.1 Simulation study

The design of the simulation experiment conducted by Wang et al (Wang et. al. 2005) was adopted here expect that the population simulated was an F_2 rather than a BC population. The sample size was fixed at 500, which is a typical sample size used in most QTL mapping experiments. The genome size was 2400 cM long covered by 241 evenly distributed markers (10 cM per marker interval). A total of 20 QTL were placed on the genome and the positions and effects of the 20 QTL are presented in Table 3.1. The QTL size varied from 0.3% phenotypic variation to 13% phenotypic variation. These proportions of QTL explaining the total phenotypic variance were calculated based on the following method. The genotype indicator variable for individual j at locus k is defined as $X_{jk} = \{1, 0, -1\}$ for the three genotypes (A_1A_1, A_1A_2, A_2A_2), respectively. Dominance effects were not simulated and also not included in the model for this simulation experiment because they do not help answer questions addressed in this study These parameter values were used to generate a quantitative trait with a population mean $b_0 = 10.0$ and a residual error variance $\sigma^2 = 10.0$. The total genetic

variance for the trait is

$$V_G = \sum_{k=1}^{20}\sum_{k'=1}^{20} b_k b_{k'} \text{cov}(z_k, z_{k'}) = \frac{1}{2}\sum_{k=1}^{20}\sum_{k'=1}^{20} b_k b_{k'}(1 - 2r_{kk'}) \tag{3.7}$$

where $r_{kk'}$ is the recombination frequency between QTL k and k', $\text{cov}(z_k, z_{k'}) = \text{var}(z)(1-2r_{kk'})$ is the covariance between Z_k and $Z_{k'}$, and $\text{var}(z) = 1/2$ is the variance of Z (assuming no segregation distortion). The total genetic variance for the quantitative trait is $V_G = V_Q + V_L = 66.384$,which is the sum of the genetic variances due to QTL(V_Q) and covariance between linked QTL(V_L), where

$$V_Q = \frac{1}{2}\sum_{k=1}^{20} b_k^2 = 46.7806 \tag{3.8}$$

and

$$V_L = \sum_{k>k'}^{20} b_k b_{k'}(1 - 2r_{kk'}) = 19.6034 \tag{3.9}$$

The residual error variance for the trait is $\sigma^2 = V_E = 10.0$. Therefore, the total phenotypic variance is $V_P = V_G + V_E = 76.384$. The proportion of the genetic variance contributed by each QTL is $0.5b_k^2/V_G$ for the kth QTL (given in the column headed with Prop-G in Table 3.1). The corresponding proportion of the phenotypic variance contributed by the kth QTL is $0.5b_k^2/V_P$ and given in the column headed with Prop-P in Table 3.1. The true QTL effects are depicted in Figure 3.1.

All 241 markers were included in the model, leading to the dimensionality of the model of $n \times (m + 1) = 500 \times (241 + 1)$. The burn in period was 1000. The chain was thinned by keeping one observation out of 10 iterations until the posterior sample size reached 5000. The total number of iterations was $1000 + 5000 \times 10 = 51000$. The true values of the QTL effects and the locations of the simulated QTL are depicted in Table

3.1. In the table, Prop-G means the proportion of genetic variance contributed by the QTL and Prop-P means the proportion of phenotypic variance contributed by the QTL.

Table 3.1: QTL parameters used in the simulation experiment.

QTL	Position	Marker	Effect	Prop-G	Prop-P
1	50	11	4.47	0.1505	0.1308
2	125	26	3.16	0.0752	0.0654
3	205	42	-2.24	0.0378	0.0328
4	235	48	-1.58	0.0188	0.0163
5	355	72	2.24	0.0378	0.0328
6	360	73	3.16	0.0752	0.0654
7	610	123	1.10	0.0091	0.0079
8	630	127	-1.10	0.0091	0.0079
9	800	161	0.77	0.0045	0.0039
10	900	181	1.73	0.0225	0.0196
11	905	182	3.81	0.1093	0.0950
12	920	185	2.25	0.0381	0.0331
13	1100	221	-1.30	0.0127	0.0111
14	1210	243	-1.00	0.0075	0.0065
15	1305	262	-2.24	0.0378	0.0328
16	1335	268	1.58	0.0188	0.0163
17	1345	270	1.00	0.0075	0.0065
18	1365	274	-1.73	0.0225	0.0196
19	1800	361	0.71	0.0038	0.0033
20	2300	461	0.89	0.0060	0.0052

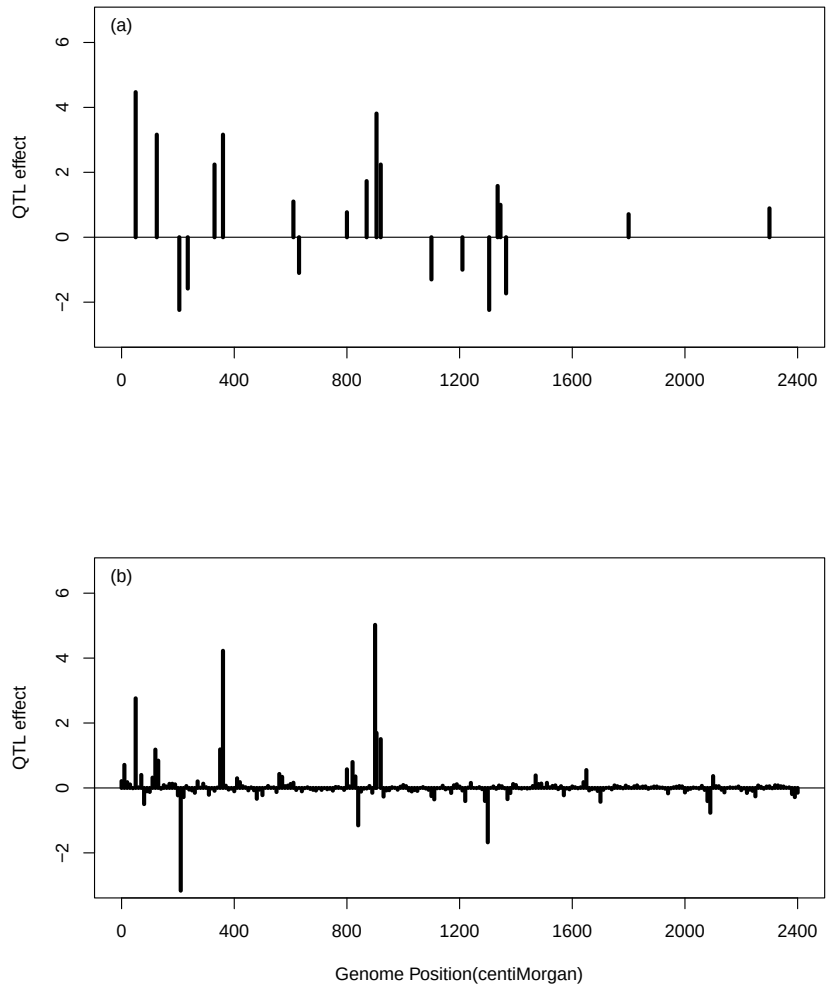


Figure 3.1: The true and estimated QTL effects for the entire genome of the simulated data. (a) The true positions and effects of the simulated QTL, (b) The estimated positions and effects of QTL using the Bayesian shrinkage method.

The true values and estimated values of QTL are depicted in Figure 3.1. Clearly, the Bayesian shrinkage method provides very good estimates to the true effects. Regions without QTL show no sign of major QTL. For the small QTL, say QTL numbers 19 and 20, the estimated effects are also small with values no larger than the bumps in the no QTL regions (noises).

We calculated the equal tail credible interval at $\alpha = 0.05$, i.e., the 2.5%-97.5% percentile range, for each marker. Only one (the largest) QTL was detected because the

interval excluded 0 (data not shown). The equal-tail credible intervals of all other QTL covered zero, and thus, they are "not significant" in terms of statistical testing. Using the equal tail credible interval at $\alpha = 0.10$, two more QTL were detected in addition to the largest QTL (data not shown). Certainly, the equal tail credible interval is not a good criterion for significance test. The posterior distributions for most estimated QTL effects have a special distribution with a spike at zero, which is the cause for the failure of equal tail credible interval as the criterion for significance test. These intervals cannot be used for significance test under the Bayesian shrinkage mapping. The reason is that almost all QTL have an equal-tail interval covering the null value, e.g., zero. Even the largest QTL in our simulation had a high probability mass at zero (see Figure 3.2). This spike-shaped or zero inflated posterior distribution for QTL effect is typical in Bayesian shrinkage mapping. If we had used the equal tail interval at $\alpha = 0.05$ as the significance test criterion, only one QTL (the largest one), out of the 20 simulated QTL, would have reached the statistical significance level. The permutation test, however, detected many major QTL, as demonstrated next in the permutation test sections.

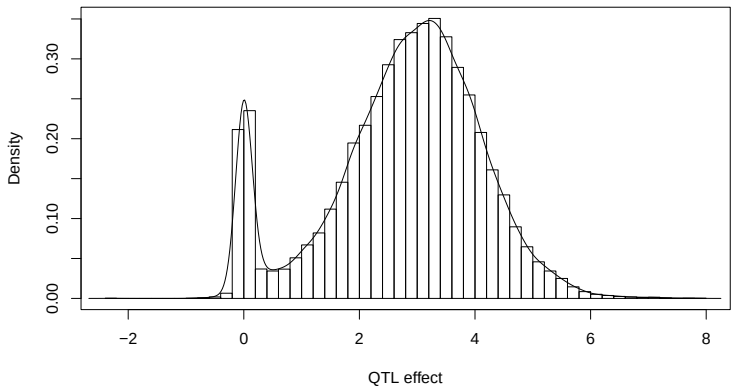


Figure 3.2: Posterior distribution of QTL number 1 of the simulation experiment. The true effect of the simulated QTL is 4.47. There is a high probability mass at value zero, even though this is the largest QTL out of the 20 QTL simulated.

3.3.2 Permutation outside Markov chain

We generated a total of 5000 permuted samples. Each permuted sample was subject to the same MCMC analysis as the original data (51000 iterations). The total computing time of the entire permutation experiment was approximately 20 days in a Dell PC (2.5 GHz and 3.25 Go of RAM). For each marker, the 2.5%-97.5% and 5%-95% intervals (corresponding to $\alpha = 0.05$ and $\alpha = 0.10$) were calculated. The profiles of these percentiles along with the estimated QTL effects are given in Figure 3.3a. Using the 2.5%-97.5% interval, we can detect 15 QTL out of the 20 simulated QTL. A few more QTL with small effects were detected when 5%-95% interval was used. The results here are more reasonable than that when the equal tail credible interval was used. The conclusion is that permutation test applies well to the Bayesian shrinkage mapping.

3.3.3 Permutation inside Markov chain

This permutation strategy only requires running one more chain in addition to the MCMC run of the original data. The phenotypes are reshuffled in every h-th iteration within the Markov chain. We first evaluated the following the performance of $h = 1$, i.e., reshuffling the phenotype in every iteration. The 2.5%, 5%, 95% and 97.5% percentiles plotted against the genome location are shown in Figure 3.3b to compare with the result of permutation outside the chains. These intervals (the within-chain permutation) appear to be wider that the intervals of the between-chain permutation analysis. Therefore, the tests for the within-chain permutation are more conservative than the between-chain permutation. Using the within-chain permutation, 13 QTL were detected for $\alpha = 0.05$ and 19 QTL were detected for $\alpha = 0.10$, not too much different from the result of the between-chain permutation. A more conservative test is better

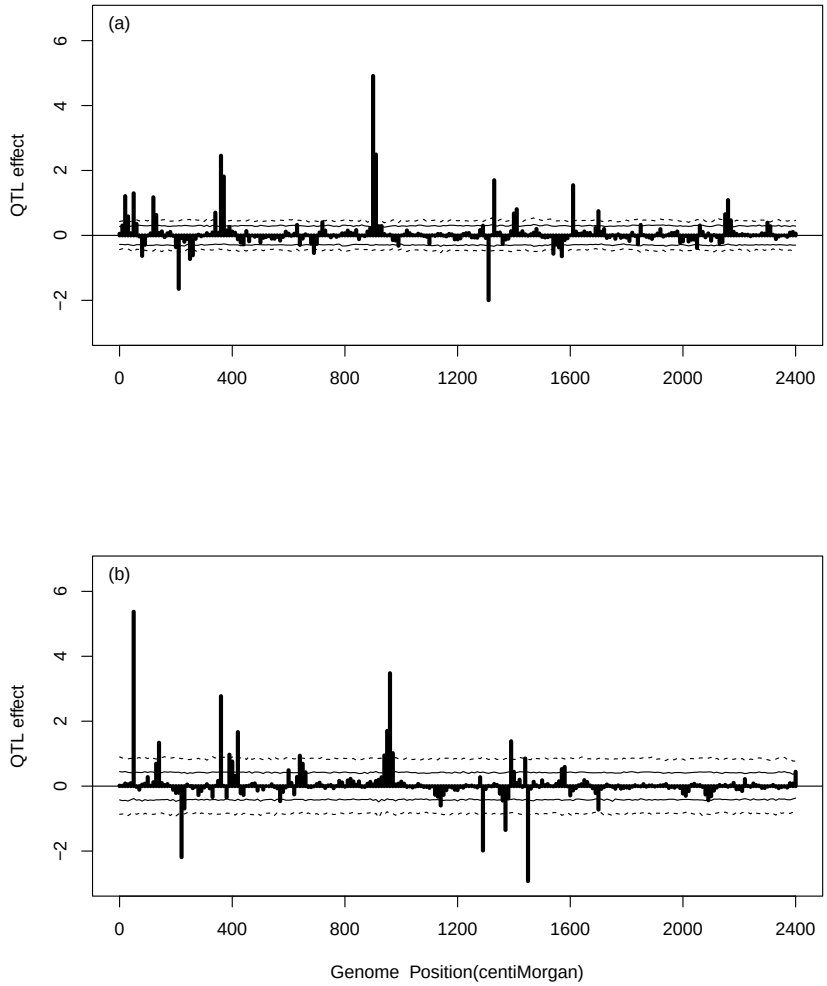


Figure 3.3: Empirical threshold values generated from permutation analysis and the estimated QTL effects (simulated data). Empirical threshold values generated from permutation analysis at $\alpha = 0.05$ (2.5%-97.5%) and $\alpha = 0.10$ (5%-95%) along with the estimated QTL effects (simulated data). Percentiles for the 2.5%-97.5% interval are plotted against the genome location as dashed lines (wider interval). Percentiles of the 5%-95% interval are plotted against the genome location as solid lines (narrower interval). (a) Shows the result of "permutation outside the Markov chain", (b) Result of "permutation within the Markov chain" with phenotype reshuffling in every iteration ($h = 1$).

than a more liberal test, as long as the statistical power is not compromised (examined later in the power study section).

We now evaluate situations where h is greater than one. This time we chose three different levels, $h = 5, 10 \quad \text{and} \quad 100$. The 2.5%, 5%, 95% and 97.5% percentiles plotted against the genome location are shown in Figure 3.4. These intervals appear to be similar to $h = 1$ except that the higher $h's$ tend to generate rougher percentile profiles. Therefore, $h = 1$ is more preferable than other values of h. Hereafter, we choose $h = 1$ for all subsequent analyses.

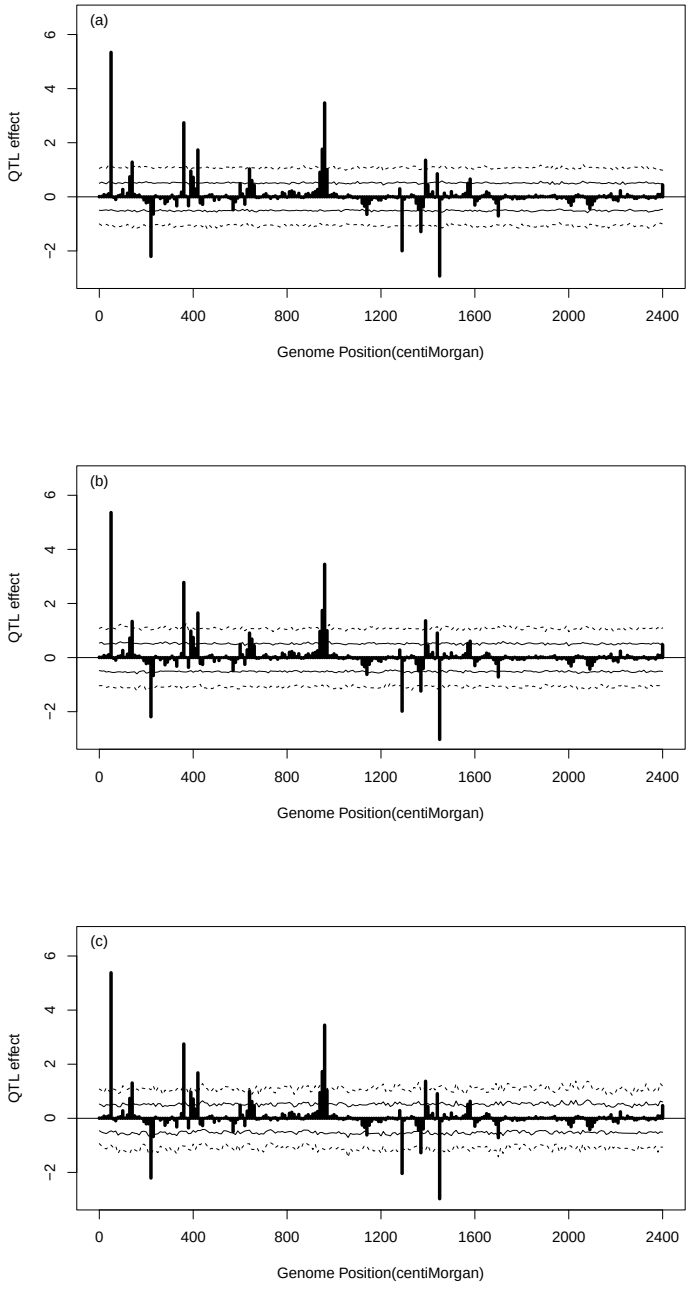
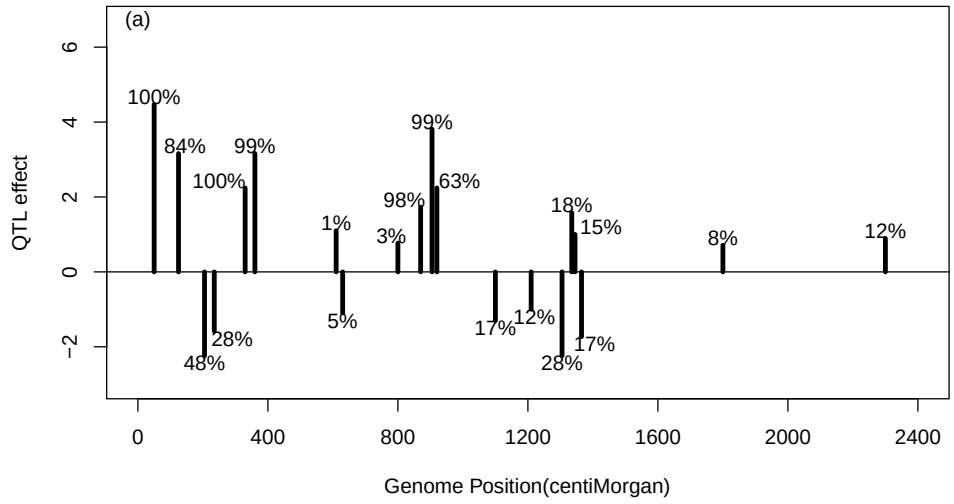


Figure 3.4: Empirical threshold values generated from "permutation within Markov chain" and the estimated QTL effects (simulated data). Empirical threshold values generated from "permutation within Markov chain" analysis at $\alpha = 0.05$ (2.5%-97.5%) and $\alpha = 0.10$ (5%-95%) along with the estimated QTL effects (simulated data). Percentiles for the 2.5%-97.5% interval are plotted against the genome location as dashed lines (wider interval). Percentiles of the 5%-95% interval are plotted against the genome location as solid lines (narrower interval). (a) Phenotype reshuffling in every 5 iterations ($h = 5$), (b) Phenotype reshuffling in every 10 iterations ($h = 10$), (c) Phenotype reshuffling in every 100 iterations ($h = 100$).

3.3.4 Power analysis

Using the same parameters given in Table 3.1, we simulated 100 more independent samples to investigate the statistical power of the Bayesian shrinkage method. Two MCMC runs were conducted for each sample. One run was the MCMC sampler on the original data to estimate QTL effects and the other run was the MCMC sampler on the within-chain reshuffled data to generate the critical values for QTL detection. The statistical power for each QTL was calculated based on the proportions of samples in which the QTL fell outside the empirical intervals. We observed that if a true QTL failed to be detected at the locus where it was placed, the effect was often picked up by a marker nearby (10 cM away). Therefore, a true QTL was claimed to be detected if one or more of the triplets (three loci) covering the true QTL (20 cM range) was detected. The statistical powers for the 20 QTL are depicted in Figure 3.5. The powers seem to be reasonable, seven out of the 20 simulated QTL have a power reached 80% at $\alpha = 0.10$. Therefore, the conserved within-chain permutation significance test does not sacrifice much statistical power.



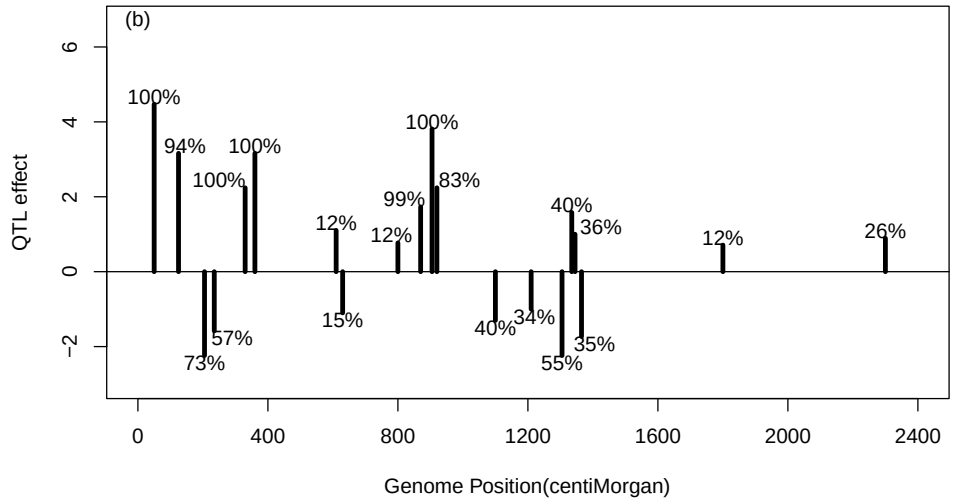


Figure 3.5: Empirical statistical power for the simulated QTL. Empirical statistical powers for the simulated QTL obtained from 100 replicated experiments. (a) Statistical powers at Type I error of $\alpha = 0.05$; (b) Statistical power at Type I error of $\alpha = 0.10$.

3.3.5 False positive rate

For the 241 marker effects included in the model, $20 \times 3 = 60$ loci were reserved for the true QTL (20 true QTL plus 40 flanking markers), leaving $241 - 60 = 181$ model effects as false QTL. If a false QTL was detected based in a particular sample, it was counted as one false positive. For each false QTL, we counted the total number of false positives among the 100 replicated experiments. The proportion of false positive (false positive rate or type I error) was recorded for each false QTL simulated. The false positive rate (FPR) profiles are depicted in Figure 3.6. The upper panel of this figure shows the observed false positive rate when $\alpha = 0.05$. Only two markers had false positive rate larger than the controlled value of 0.05. All other markers had false positive rate less than 0.05. The average false positive rate of all markers was about 0.02. The observed false positive rate is indeed less than 0.05, confirming our previous conclusion that the within-chain permutation approach is conservative. The lower panel of Figure 3.6 shows the observed false positive rate at $\alpha = 0.10$. Only four markers has false positive rate larger than 0.10. The average false positive rate for all these markers was about 0.05, again confirming the conservativeness of the within chain permutation approach.

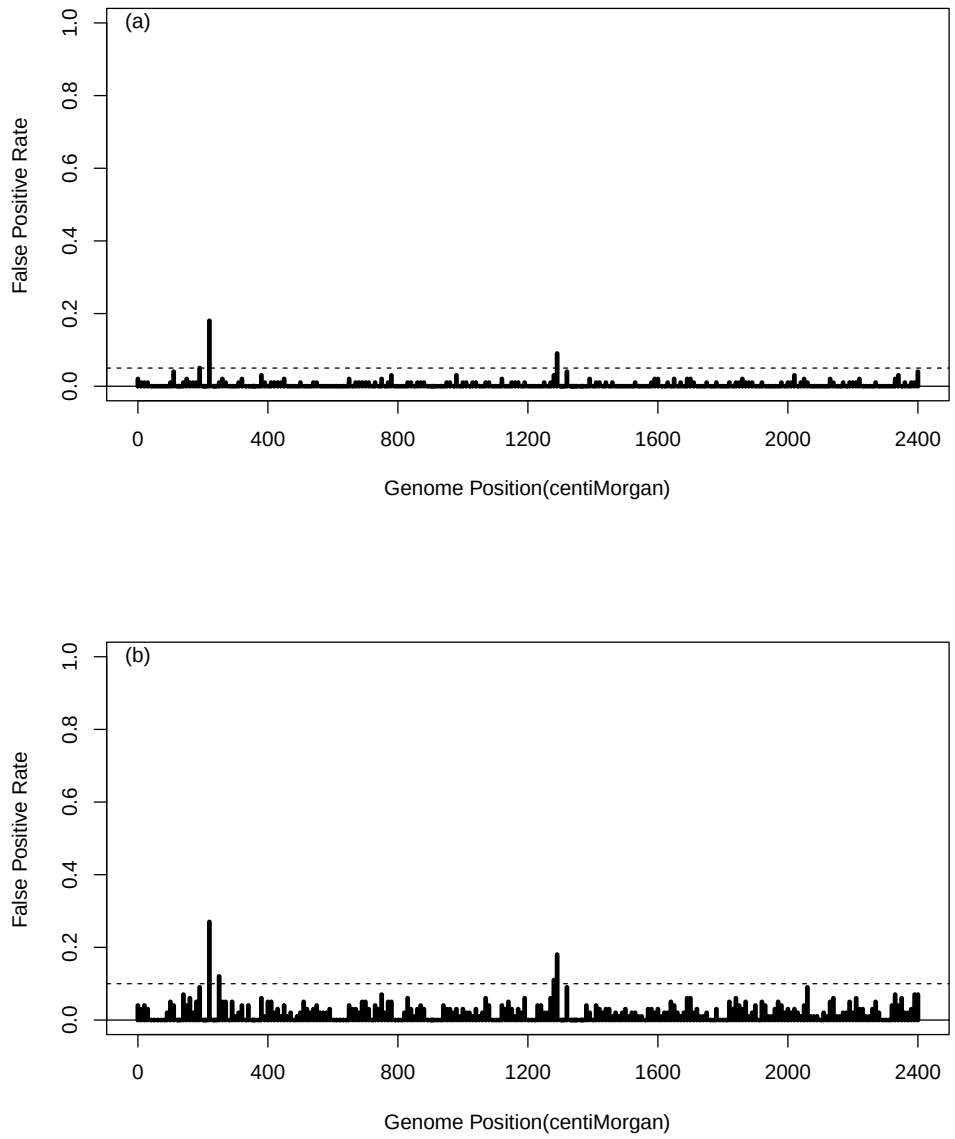


Figure 3.6: False positive rate profiles for the simulated markers obtained from 100 replicated experiments. (a) False positive rate at $\alpha = 0.05$; (b) False positive rate at $\alpha = 0.10$.

3.3.6 Cross validation for genome selection

Using the original data simulated in the beginning of the experiment (not a sample from the power study), we performed the five fold cross validation study to determine how large a QTL should be included in the model to predict the total genetic value of an individual. The PE (squared prediction error) values are plotted against the α value in Figure 3.7. The minimum PE value occurs when $\alpha = 0.2$. The decrease of the PE from $\alpha = 0.0$ to $\alpha = 0.2$ is very sharp, but after $\alpha = 0.2$, the PE value tends to be stabilized or slightly increased. The conclusion is that in genome selection, we should choose the α around 0.2. Of course, this optimal value may vary from sample to sample. We recommend such a cross validation test for each data analysis to determine how many QTL should be included. From the PE profile, including all QTL ($\alpha = 1$) into the prediction model (regardless the sizes of the QTL) does not lead to any significant loss in the precision of genome selection compared to the optimal number of QTL determined by the cross validation test. Therefore, a robust choice is simply to include all QTL in the model for genome selection.

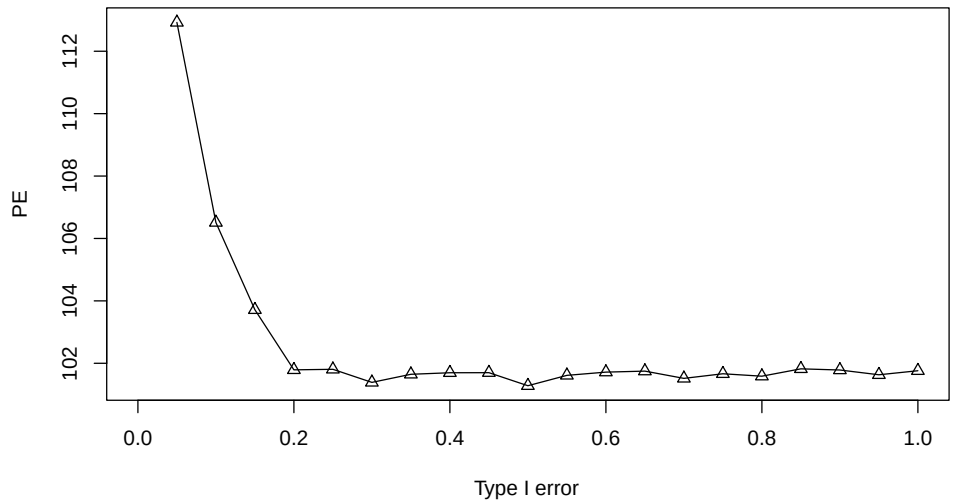


Figure 3.7: Prediction error (PE) plotted against the Type I error for the simulated data. The squared prediction error (PE) plotted against the Type I error obtained from the five-fold cross validation test for the simulated data.

3.3.7 Real data analysis

We now use three sample data to demonstrate the application of the permutation test associated Bayesian shrinkage analysis. These data were collected from QTL mapping experiments in model plants and agricultural crops. The original data are downloadable from the internet and also attached to this manuscript as supplemental material.

3.3.7.1 Arabidopsis data

The first data set is the recombinant inbred line data of Arabidopsis data (Loudet et. al. 2002) , where the two parents initiating the line cross were Bay-0 and Shahdara with Bay-0 as the female parent. The recombinant inbred lines were actually F_7 progeny of single seed descendants of the F_2 plants. The residual heterozygosity was low (Loudet et. al. 2003). Flowering time was recorded for each line in two environments: long day (16 hour photoperiod) and short day (8 hour photoperiod). We used the short day

flowering time as the quantitative trait for QTL mapping. The two parents had very little difference in short day flowering time. The sample size (number of recombinant lines) was 420. A couple of lines did not have the phenotypic records and their phenotypic values were replaced by the population mean for convenience of data analysis. A total of 38 microsatellite markers were used for QTL mapping. These markers are more or less evenly distributed along five chromosomes with an average 10.8 centiMorgan (cM) per marker interval. The marker names and positions are given in the original article (Loudet et. al. 2003) .

We inserted a pseudo marker in every 2 cM of the genome. Including the inserted pseudo markers, the total number of loci subject to analysis was 200 (38 true markers plus 162 pseudo markers). All the 200 putative loci were evaluated simultaneously in a single model. Therefore, the model for the short day flowering time trait is

$$y = b_0 + \sum_{k=1}^{200} X_k b_k + \varepsilon \tag{3.10}$$

where X_k is a 420×1 vector coded as 1 for one genotype and 0 for the other genotype for locus k. If locus k is a pseudo marker, $X_k = \Pr(\text{genotype} = 1)$, which is the conditional probabilities of marker k being of genotype 1. Finally, b_k is the QTL effect of locus k. For the original data analysis, the burn-in period was 1000. The thinning rate was 10. The posterior sample size was 10000, and thus the total number of iterations was $1000 + 10000 \times 10 = 101000$. The posterior sample size of the within-chain permutation analysis was 80000, i.e., $1000 + 80000 \times 10 = 801000$ iterations in total. The estimated QTL effects and the permutation generated 2.5%-97.5% and 5%-95% intervals are plotted in Figure 3.8a. A total of 4 QTL were detected on three chromosomes at $\alpha = 0.05$. Chromosomes 1 and 4, each has one QTL and chromosome 5 has two QTL. When $\alpha = 0.10$ was used,

one more QTL on chromosome 1 was detected.

The five-fold cross validation shows that the optimal strategy of genome selection for this data set was to include all QTL in the model, regardless the significance of the estimated QTL effects (see Figure 3.8b). The general pattern of the PE profile remains the same as that of the simulated data. Below $\alpha = 0.2$ the decrease of PE was dramatic but after $\alpha = 0.2$ the PE values approached a stable value.

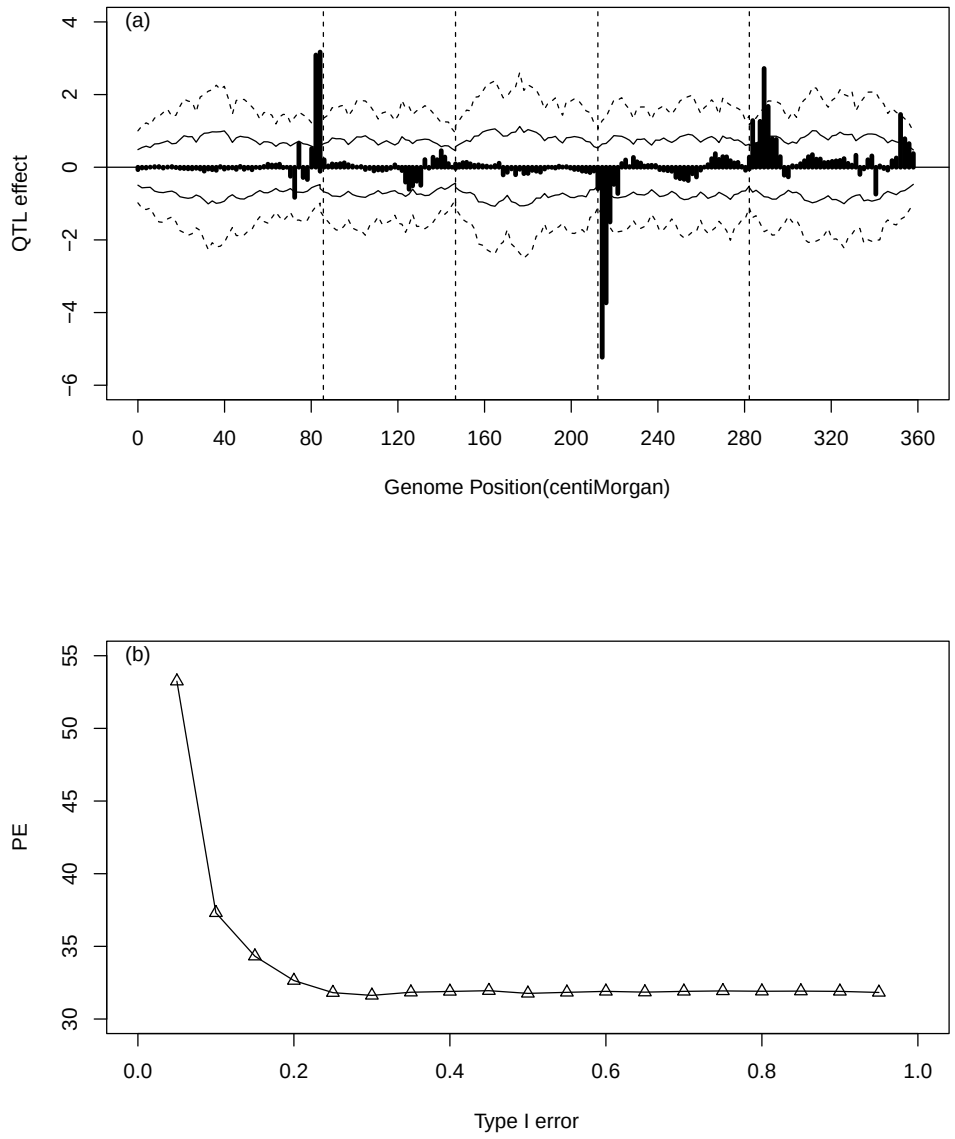


Figure 3.8: Result of the Arabidopsis data analysis. (a) The upper panel shows the estimated QTL effects for the entire genome and the empirical thresholds drawn from permutation within the Markov chain analysis at $\alpha = 0.05$ (2.5%-97.5%, wider interval) and $\alpha = 0.10$ (5%-95%, narrower interval). (b) The lower panel shows the plot of the squared prediction error (PE) against the Type I error obtained from the five-fold cross validation test.

3.3.7.2 Barley data

The second data are the double haploid (DH) data obtained from Luo et al (Luo et. al. 2007) . This data set consists of 150 double haploids (DH) derived from the cross of two spring barley varieties, Steptoe and Morex, designated as the cross. The phenotype was the spot blotch (a fungus *Cochliobolus sativus*) resistance measured as the lesion size on the leaves of barley seedlings. The total number of markers was 495 distributed along seven chromosomes of the barley genome. Because of the small sample size, we could not analyze all the 495 markers simultaneously (high collinearity). Therefore, we placed one pseudo marker in every 5 cM and overall obtained 225 pseudo markers for the entire genome. The genotypes of the pseudo markers were inferred from the multipoint method (Jiang and Zeng 1997)). All the 225 putative loci were evaluated simultaneously in a single model. Therefore, the model for the disease resistance trait is

$$y = b_0 + \sum_{k=1}^{225} X_k b_k + \varepsilon \tag{3.11}$$

where X_k is a 150×1 vector coded as 1 for one genotype and 0 for the other genotype for locus k. If locus k is a pseudo marker, $X_k = \Pr(\text{genotype} = 1)$, which is the conditional probabilities of marker k being of genotype 1. Finally, b_k is the QTL effect of locus k.

The parameters of the MCMC experiment (e.g., burn-in period, thinning rate and so on) were the same as the Arabidopsis data analysis. The estimated QTL effects and the permutation generated 2.5%-97.5% and 5%-95% intervals are plotted in Figure 3.9a. A total of two QTL were detected on chromosome 7 at $\alpha = 0.05$. These two are major because their estimated values are way over the critical value. When the critical values at $\alpha = 0.10$ were used, five more QTL were declared as significant.

The cross validation shows that the optimal strategy of genome selection for this

data set was to include all QTL that are significant at $\alpha = 0.15$ (see Figure 3.9b). Below $\alpha = 0.15$ the decrease of PE was dramatic but after $\alpha = 0.15$ the PE values increased slightly until it reached a plateau at $\alpha = 0.3$ (see Figure 3.9b). This example demonstrated the usefulness of using cross validation to select QTL for inclusion for prediction of genomic effect.

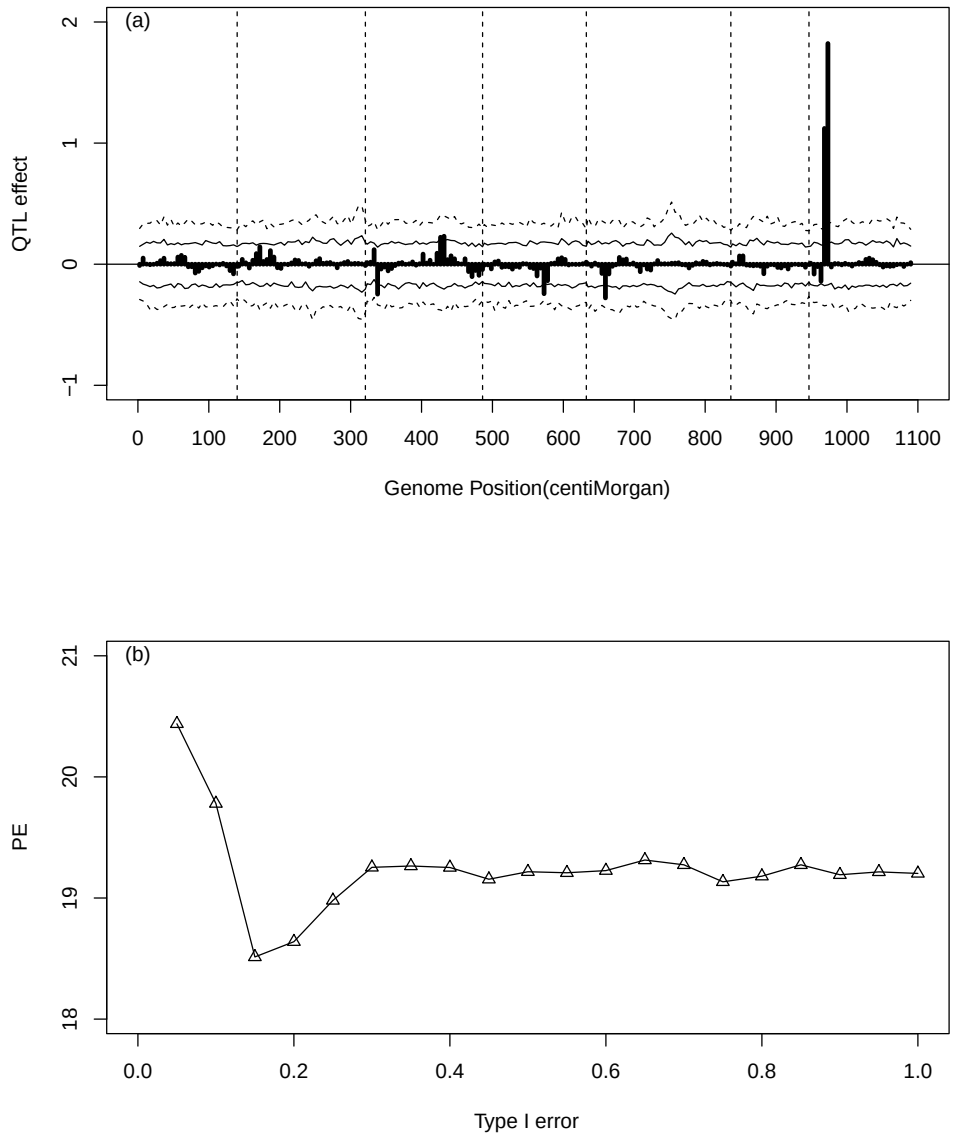


Figure 3.9: Result of the barley data analysis. (a) The upper panel shows the estimated QTL effects for the entire genome and the empirical thresholds drawn from permutation within the Markov chain analysis at $\alpha = 0.05$ (2.5%-97.5%, wider interval) and $\alpha = 0.10$ (5%-95%, narrower interval). (b) The lower panel shows the plot of the squared prediction error (PE) against the Type I error obtained from the cross validation test.

3.3.7.3 Wheat data

This example demonstrates the application of the Bayesian shrinkage analysis to QTL mapping for the number of seeded spikelets (a female fertility trait) in wheat. The experiment was conducted by Dou et al. (Dou et. al. 2009) who made the data available to us for this analysis. A female sterile line XND126 and an elite cultivar Gaocheng 8901 with normal fertility were crossed for genetic analysis of female sterility measured as a quantitative trait. The parents, their F_1 and F_2 progeny were planted at the Huaian experimental station in China for the 2006-2007 growing season under the normal autumn sowing condition. The mapping population was an F_2 family consisting of 243 individual plants. A total of 28 SSR markers were used in this experiment. These markers covered 5 chromosomes of the wheat genome with an average genome marker density of 15.5 cM per marker interval. The five chromosomes are only part of the wheat genome. These chromosomes were scanned for QTL of the fertility trait using the MCMC implemented Bayesian method. The dependent variable was the fertility phenotype while the independent variables were numerically coded genotype indicator variables for the part of genome under investigation. We placed one pseudo marker in every 5 centiMorgan (cM) of the genome. This generated 75 pseudo markers for the five chromosomes. Therefore, we have a total of 75 independent variables. For each independent variable, the numerically coded value was the difference between the conditional probabilities of the two homozygote genotypes. Let A_1A_1, A_1A_2 and A_2A_2 be the three genotypes for the kth pseudo marker of the genome. The numerically coded value for the locus is

$$X_{jk} = p(G_{jk} = A_1A_1|\text{marker}) - p(G_{jk} = A_2A_2|\text{marker}) \tag{3.12}$$

for $k = 1, ..., 75$. The map of the 75 pseudo markers, the phenotypic values of the 243 plants and the 75 numerically coded independent variables can be found from the supplemental material of this study.

The parameters of the MCMC experiment (e.g., burn-in period, thinning rate and so on) were the same as the previous two data analyses. The estimated QTL effects and the permutation generated 2.5%-97.5% and 5%-95% intervals are plotted in Figure 3.10a. A total of two QTL were detected on chromosome 2 at $\alpha = 0.05$. When we lowered the critical value to $\alpha = 0.10$, one more QTL was detected on chromosome 5.

The cross validation shows that the optimal strategy of genome selection for this data set was to include all QTL that are significant at $\alpha = 0.1$ (see Figure 3.10b). Below $\alpha = 0.1$, the decrease of PE was dramatic but after $\alpha = 0.1$ the PE values increased slightly until it reached a plateau at $\alpha = 0.3$.

In general, the optimal alpha value is somewhere between 0.1 to 0.2, but it varied from one experiment to another. The last two data analyses did indicate that including small QTL can be detrimental to genome selection. Cross validation is an experimental specific approach and is useful to decide how large a QTL should be included in the model for genome selection.

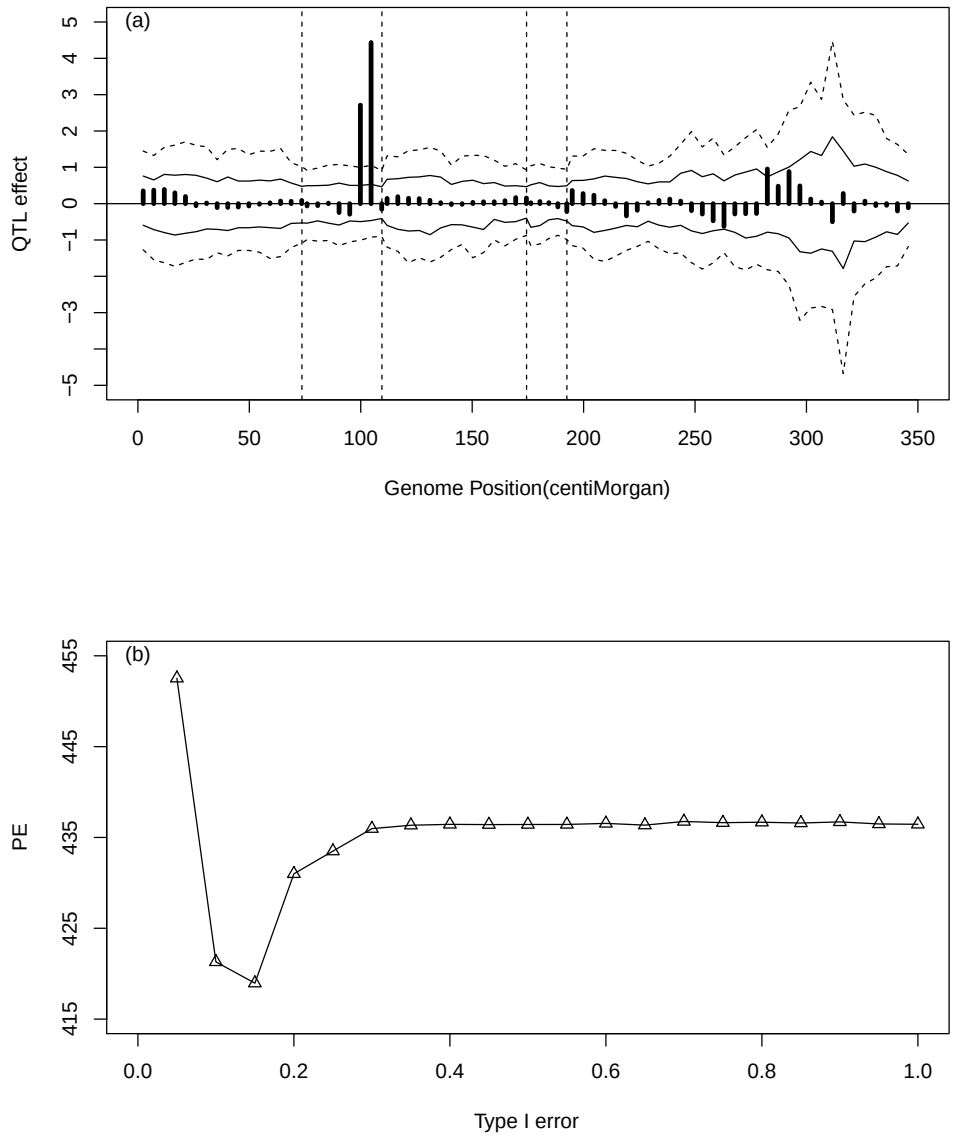


Figure 3.10: Result of the wheat data analysis. (a) The upper panel shows the estimated QTL effects for the entire genome and the empirical thresholds drawn from permutation within the Markov chain analysis at $\alpha = 0.05$ (2.5%-97.5%, wider interval) and $\alpha = 0.10$ (5%-95%, narrower interval). (b) The lower panel shows the plot of the squared prediction error (PE) against the Type I error obtained from the cross validation test.

3.4 Discussion

Bayesian shrinkage analysis can be used for both QTL mapping and genome selection. The two applications are quite different. QTL mapping aims to detect QTL with large effects while genome selection tries to predict the total genetic values of individuals using markers of the entire genome. In QTL mapping, significance test is important, but Bayesian inference usually does not mix with significance test. This is because Bayesian inference focuses on the probability statement of a parameter given the information drawn from the current data and it does not intend to extend the statement beyond the data. Significance test, however, assumes a null distribution and tries to compare the statistics against the null distribution. The null distribution is purely hypothetical and, therefore, significance test gives conclusion that applies to hypothetical future experiments. The permutation test adopted in the Bayesian analysis is a convenient way to connect significance test with Bayesian analysis. Permutation analysis is a way to draw the null distribution. If a statistics, e.g., estimated QTL effect, is far away from the null distribution, we are confident that this QTL is true. This type of significance test provides different conclusion from the Bayesian credible statement. In Bayesian analysis, people often report the α-equal-tail interval or α-highest posterior density (HPD) interval. These intervals cannot be used for significance test under the Bayesian shrinkage mapping. The reason is that almost all QTL have an equal-tail interval covering the null value, e.g., zero. Even the largest QTL in our simulation had a high probability mass at zero (see Figure 3.2). This zero inflated posterior distribution for QTL effect is typical in Bayesian shrinkage mapping. If we had used the equal tail interval at $\alpha = 0.05$ as the significance test criterion, only one QTL (the largest one), out of the 20 simulated QTL, would have reached the statistical significance level. The permutation test, however,

detected many major QTL.

In the simulation experiment, we observed that the percentile profiles for the $\frac{1}{2}\alpha \times 100\% - (1 - \frac{1}{2}\alpha) \times 100\%$ interval were pretty much constant across the entire genome (see Figures 3.3 and 3.4). This is due to the uniform information content across the genome. We simulated 241 markers covering the entire genome evenly with 10 cM per marker interval. These markers were co-dominant with no missing genotypes. In contrast, the three real data analyses showed that the percentile lines varied dramatically across the genome. The intervals were narrow at marker positions and wide when the positions are away from the markers. The lengths of marker intervals also varied across the genome, making the information content much uneven across the genome. The location specific empirical threshold values in real data analysis mean that different locations of the genome should use different criteria for QTL detection. Two QTL with the same estimated effect but located in different regions of the genome, one may be declared as significant but the other may not be significant due to the variation in information content. This actually justifies the use of estimated QTL effects, not some kind of test statistics, for significance test.

In classical QTL mapping experiments, investigators always use some kinds of test statistics (e.g., t-test, F-test, likelihood ratio test or LOD score) to decide whether a QTL is significant or not. A permutation test also draws critical values for the test statistic under consideration, not the critical values for the QTL effects. This merely reflects the tradition or convention of people who do statistical analysis and does not mean a test statistic is the only quantity that can be used in QTL mapping. The reason for using test statistics is that one can compare the observed test statistic (calculated values) with the critical values of some distribution, e.g. normal distribution, F-distribution, t-distribution and chi-square distribution. The critical values of these

standard distributions can be found from statistical tables or calculated from statistical analysis software. With the permutation test, we never need the critical values of the standard distributions. Therefore, there is no need to stick with the test statistics. Directly comparing the estimated QTL effects with the critical values is more intuitive.

Significance test can help us decide which QTL should be claimed as significance. The significant QTL will be the targets for further study, e.g., cloning or marker assisted selection. What do we do with those QTL whose effects do not reach the significance level? These QTL may not be significant individually, but collectively they may contribute to a large proportion of the phenotypic variance. This implies that they are perhaps useful to predict the total genetic effects of individuals (Meuwissen et. al. 2001) , a technology called genome selection. Our cross validation experiments showed that QTL should be used to predict the total genetic effects once they reached a certain critical value. Including many small QTL can be harmful to genome selection. Common sense tells us that estimated effects of small QTL are most likely caused by noises rather than by true signals and inclusion of the many small QTL to predict the genetic effects may be even worse than inclusion of only the significant QTL.

Chapter 4

Generalized Linear Mixed Models for Mapping Quantitative Trait Loci

4.1 Introduction

Linear mixed model methodology is a powerful technology to analyze models containing both the fixed and random effects. The model was first proposed to estimate genetic parameters for unbalanced data (Henderson 1950). This technique has been used to map genes controlling the variation of quantitative traits (Xu and Yi 2000; Boer et al. 2007). The mixed model methodology cannot be directly applied to traits with discrete distributions. Wedderburn (1974) proposed a linear predictor and a link function to handle discrete traits. The linear predictor is simply a linear model combining information from the independent variables. The link function is to describe the relationship between the linear predictor and the expectation of the response variable.

This approach eventually leads to a special area of statistics called the generalized linear model (McCullagh and Nelder 1989). The generalized linear model takes advantage of all theory and methods developed in the usual linear model methodology (Searle 1997). It has been applied to QTL mapping for some special traits, e.g., binary traits (Xu and Atchley 1996, Deng et al. 2006, Yi and Xu 1999a,b, Yi and Xu 2000), ordinal traits (Hackett and Weller 1995; Rao and Xu 1998) and Poisson traits (Cui et al. 2006, Cui and Yang 2009). Depending on the special characteristics of the traits, distribution specific generalized linear models have been developed for these traits. These methods are not sufficiently general to extend to all traits that can be modeled by the generalized linear model. For example, the EM algorithm developed by Xu et al. (2003, 2005) only applies to binary and ordinal traits. They treated both the marker genotypes and the latent variable as missing values. Although parameter estimation under the EM algorithm is simple, the information matrix of the estimated parameters is difficult to calculate. A more comprehensive analysis of the generalized linear model applied to QTL mapping is the seminal paper by Lange and Whitaker (2001). They adopted the generalized estimating equations (GEE) approach to analyze multiple traits with arbitrary combination of continuous and discrete trait components. The method replaces the unobserved QTL genotypes by the conditional expectations of the genotype indicator variable given flanking marker information. The uncertainties of the genotype indicator variables are ignored. In addition, detailed formulas for the partial derivatives of the expectation of the data with respect to the parameters are not given.

Xu and Hu (2010) recently developed a generalized linear model approach to interval mapping for traits with arbitrary distribution. The purpose of that study was to investigate the efficiencies of different methods for handling missing genotypes. Three algorithms have been proposed: (1) expectation algorithm, (2) heterogeneous variance

algorithm and (3) mixture model algorithm. The expectation algorithm is the simplest one in which the missing genotypes of QTL are replaced by the conditional expectation of the genotype indicator variable. The heterogeneous variance algorithm takes into account the heterogeneous variances of different genotypes due to heterogeneous information contents. The mixture model fully utilizes the conditional distribution of the missing genotypes. Theoretically, the mixture model approach should be optimal, followed by the heterogeneous variance model and the expectation algorithm. In practice, the heterogeneous model is more efficient because it is robust to departure from the assumed normal distribution of the residuals. On the contrary, the mixture model is very sensitive to the departure of an assumed distribution and the choice of the initial values of the parameters. The three algorithms of handling missing genotypes have not been applied to multiple QTL mapping and this study aims to explore their application.

When the number of QTL included in a model reaches a certain level, the model is oversaturated. In this case, some kind of penalty is required to shrink the superfluous QTL down to zero. When the linear predictor contains both fixed and random effects, the model is then called the generalized linear mixed model (Breslow and Clayton 1993; McCulloch and Neuhaus 2005). Special algorithms have been developed to estimate variance components and predict the random effects, e.g., the pseudo likelihood algorithm (Wolfiner and O'Connell 1993). However, existing generalized linear mixed models have not fully considered the missing genotype problem. The hierarchical generalized linear model for multiple QTL mapping developed by Yi and Banerjee (2009) also ignored the missing genotype problem and thus the method only applies to marker analysis.

When there are no missing values, commercial software packages are available to estimate parameters of a generalized linear mixed model under a wide range of distribution of the traits, e.g., GLIMMIX procedure in SAS (SAS Institute 2008). These

programs may handle missing values using the imputation algorithm, the missing patterns handled by these commercial programs are usually different from that of QTL mapping. In QTL mapping, genotypes are missing for every individual at a putative QTL position unless the QTL overlaps with a fully informative marker. In the statistics literature, generalized linear model with missing covariates is often handled with the EM algorithm (Horton and Laird 1998). However, other methods are also available, as summarized by Ibrahim et al. (2005), who reviewed four general approaches: maximum likelihood method implemented via the EM algorithm by the method of weights (Horton and Laird 1998), multiple imputation (Rubin 1987), fully Bayesian (Ibrahim et al. 2002) and weighted estimation equation (Robins and Rotnitzky 2001). Ibrahim et al (2005) concluded that the most accurate method is the fully Bayesian method, although the method is associated with a high cost in terms of computing time. The second best method is the EM algorithm via the method of weights. Applications of these methods to multiple QTL mapping have not been attempted. In this study, we proposed three algorithms to handle the missing genotype problems in multiple QTL mapping under the generalized linear mixed model framework.

4.2 Methods

4.2.1 Generalized linear mixed model

We use a binomial trait as an example to demonstrate the new methodology, although the method can be applied to other discrete traits. Let y_j be the number of events and t_j be the number of trials for individual j from a population of n individuals. Let $E(y_j/t_j) = \mu_j$ be the expectation of the binomial trait. Define $\eta_j = \Phi^{-1}(\mu_j)$ as a linear predictor with the probit link function. The logit link function may also be

applied, $\eta_j = \text{logit}(\mu_j) = \ln[\mu_j/(1-\mu_j)]$, but we prefer the probit link function because the normal latent variable seems to be more appropriate for modeling an underlying quantitative trait. The linear predictor is a function of marker genotypes, as described below,

$$\eta_j = \beta + \sum_{k=1}^{p} Z_{jk}\gamma_k \tag{4.1}$$

where β is the intercept, γ_k is the marker effect for locus k and Z_{jk} is an independent variable determined by the genotype of marker k of individual j and p is the total number of markers included in the model. Details about Z_{jk} will be described later.

The log likelihood function for parameters $\{\beta, \gamma\}$ is

$$L(\beta, \gamma) = \sum_{j=1}^{n} [y_j \ln(\mu_j) + (t_j - y_j)\ln(1-\mu_j)] \tag{4.2}$$

The prior distribution for β is assumed to be uniform, but $p(\gamma_k) = N(\gamma_k|0, \sigma_k^2)$ is chosen as the prior distribution for each γ_k. When p is large, a typical case for genome prediction, we need a hierarchical prior for each variance component,

$$p(\sigma_k^2) = \text{Inv}\chi^2(\sigma_k^2|\tau, \omega) \tag{4.3}$$

when (τ, ω) are the hyper-parameters. When $(\tau, \omega) = (-2, 0)$, this prior is equivalent to the uniform prior, leading to the usual generalized linear mixed model (GLMM, McCulloch and Neuhaus 2005). Let $G = \{\sigma_k^2\}$ be the array of variance components. In the GLMM framework,$\theta = \{\beta, G\}$ are treated as parameters and γ are considered as missing values. We now describe an EM algorithm to estimate the parameters θ, to infer the QTL effects γ and to test the significance of γ.

The EM algorithm used here is a sequential approach that updates one element of

the parameters (including missing values) at a time conditional on values of all other parameters. This approach is also called the descent coordinate algorithm. When all parameters are updated in turn, one cycle of the iterations is completed and the iteration process continues until a certain criterion of convergence is reached. The EM iteration process starts with initial values of all unknowns provided by the investigator, denoted by $\{\beta^{(0)}, \gamma^{(0)}, G^{(0)}\}$. Updating β and G represents the maximization steps and updating γ represents the expectation steps.

1. Update β using the following equation,

$$\beta^{(t+1)} = \beta^{(t)} - \left[L''(\beta^{(t)}, \gamma^{(t)})\right]^{-1} L'(\beta^{(t)}, \gamma^{(t)}) \tag{4.4}$$

   where $L'(\beta^{(t)}, \gamma^{(t)})$ and $L''(\beta^{(t)}, \gamma^{(t)})$ are, respectively, the first and second partial derivatives of the log likelihood function with respect to β evaluated at $\beta^{(t)}$ and $\gamma^{(t)}$. This equation is the first step of the Newton-Raphson iteration. Once β is updated, and β involved in the future steps will be replaced by the most current value of β.

2. Update each γ_k using the posterior mean, which is inferred by combining both the prior information and the likelihood function. Let

$$\tilde{\gamma_k} = \gamma_k^{(t)} - \left[L''(\beta^{(t)}, \gamma^{(t)})\right]^{-1} L'(\beta^{(t)}, \gamma^{(t)}) \tag{4.5}$$

   the first step Newton-Raphson update of γ_k, and

$$s_k^2 = -\left[L''(\beta^{(t)}, \gamma^{(t)})\right]^{-1} \tag{4.6}$$

be an approximate variance of this update, where the first and second partial derivatives are now taken with respect to γ_k. The QTL effect is now inferred from two sources of information, one from the likelihood and the other from the prior. Assume that the information from the likelihood is approximately normal, i.e., $\gamma_k \sim N_1(\tilde{\gamma_k}, s_k^2)$. Recall that the prior distribution of γ_k is $N_2(0, \sigma_k^2)$. Combining the likelihood and the prior, we get posterior distribution of γ_k, which is normal with mean

$$\mathrm{E}(\gamma_k) = \left(\frac{1}{\sigma_k^2} + \frac{1}{s_k^2}\right)^{-1} \frac{1}{s_k^2}\tilde{\gamma_k} \tag{4.7}$$

and variance

$$\mathrm{var}(\gamma_k) = \left(\frac{1}{\sigma_k^2} + \frac{1}{s_k^2}\right)^{-1} \tag{4.8}$$

The updated γ_k is the posterior mean, i.e., $\gamma_k^{(t+1)} = \mathrm{E}(\gamma_k)$.

3. Update σ_k^2 using the equation given by Xu (2010)

$$\sigma_k^2 = \frac{\mathrm{E}(\gamma_k^2) + \omega}{\tau + 2 + 1} = \frac{[\mathrm{E}(\gamma_k)]^2 + \mathrm{var}(\gamma_k) + \omega}{\tau + 2 + 1} \tag{4.9}$$

This equation is different from the posterior mode estimation of σ_k^2 in which $\mathrm{E}(\gamma_k^2)$ is simply replaced by $[\mathrm{E}(\gamma_k)]^2$ with $\mathrm{var}(\gamma_k)$ ignored. This explains why the algorithm is call the EM algorithm.

4.2.2 Missing genotypes

In QTL mapping, the genotype indicator variable (Z_{jk}) is missing if the QTL position does not overlap with a fully informative marker. However, partial information is available due to linkage disequilibrium. We propose three approaches to handling the missing genotypes. The differences of the three approaches are reflected by the

differences of the log likelihood functions, which are described in the following sections.

4.2.2.1 Expectation model

The linkage disequilibrium allows us to infer the conditional distribution of Z_{jk} given information from linked markers. Let A_1A_1, A_1A_2 and A_2A_2 be the three genotypes of a QTL for an individual in an F_2 population. The Z variable is determined by the genotype of locus k,

$$Z_{jk} = \begin{cases} +1 & \text{for } A_1A_1 \\ 0 & \text{for } A_1A_2 \\ -1 & \text{for } A_2A_2 \end{cases} \tag{4.10}$$

In the context of generalized linear mixed model, $\gamma_k = a_k$, where a_k is called the additive effect of locus k. When Z_{jk} is missing, the expectation and variance of it are inferred from the genotypes of flanking markers. Let $p_j(+1)$, $p_j(0)$ and $p_j(-1)$ be the conditional probabilities of the three genotypes inferred from flanking markers. The expectation and variance of Z_{jk} are

$$\mathrm{E}(Z_{jk}) = U_{jk} = p_j(+1) - p_j(-1) \tag{4.11}$$

and

$$\mathrm{var}(Z_{jk}) = \Sigma_{jk} = [p_j(+1) - p_j(-1)] - [p_j(+1) - p_j(-1)]^2 \tag{4.12}$$

With the expectation approach, we simply replace Z_{jk} by U_{jk}. Therefore, the linear predictor is defined as

$$\eta_j = \beta + \sum_{k=1}^{p} U_{jk}\gamma_k \tag{4.13}$$

Everything else remains the same as the situation with complete genotypic information.

4.2.2.2 Overdispersion model

The expectation method only takes advantage of the first moment of the distribution. The second moment information has been ignored, which will generate a situation called overdispersion. For locus k, the overdispersion is defined as

$$o_{jk} = \Sigma_{jk}\gamma_k^2 + 1 \tag{4.14}$$

Incorporating this overdispersion, we redefine the linear predictor as

$$\eta_{jk} = \frac{1}{\sqrt{o_{jk}}}(\beta + U_{jk}\gamma_k + \xi_{jk}) \tag{4.15}$$

where

$$\xi_{jk} = \sum_{k' \neq k}^{p} U_{jk'}\gamma_{k'} \tag{4.16}$$

is an offset of the linear predictor contributed by other loci. We now have a locus specific log likelihood function,

$$L(\gamma_k) = \sum_{j=1}^{n} [y_j \ln(\mu_{jk}) + (t_j - y_j)\ln(1 - \mu_{jk})] \tag{4.17}$$

where $\mu_{jk} = \Phi(\eta_{jk})$.Note that if a QTL overlaps with a fully informative marker, $\mathrm{E}(Z_{jk}) = U_{jk} = Z_{jk}$ and $\mathrm{var}(Z_{jk}) = \Sigma_{jk} = 0$, leading to $o_{jk} = 1$.

4.2.2.3 Mixture model

The optimal approach is to take advantage of the conditional distribution of the missing genotypes. When the genotype of a QTL is missing, the model becomes a

mixture model. Let us define

$$\mu_j(-1) = \Phi(\beta - \gamma_k + \xi_{jk})$$

$$\mu_j(0) = \Phi(\beta + \xi_{jk})$$

$$\mu_j(+1) = \Phi(\beta + \gamma_k + \xi_{jk}) \tag{4.18}$$

as the expectations of the trait for the three genotypes. The log likelihood of the mixture model for locus k is

$$L(\gamma_k) = \sum_{j=1}^{n} \ln\left\{ \sum_{i=1}^{3} p_j(i-2)[\mu_j(i-2)]^{y_j}[1-\mu_j(i-1)]^{t_j-y_j} \right\} \tag{4.19}$$

Note that $p_j(i-2)$ is the conditional probability, not p_j multiplied by $(i-2)$.

4.3 Application

4.3.1 Simulation study

4.3.1.1 Binomial data

We simulated a single large chromosome of 2400 cM long evenly covered by 241 co-dominance markers (10 cM per marker interval). The simulated population was an F_2 family derived from the cross of two inbred lines with sample size $n = 500$.The genotype indicator variable for individual j at locus k was defined as $Z_{jk} = \{-1, 0, 1\}$ for the three genotypes (A_1A_2, A_1A_2, A_2A_2), respectively. Dominance effects were not simulated and also not included in the model for this simulation experiment. A total of 20 QTL were simulated with the sizes and locations of the QTL depicted in Figure 1a (the top panel). Positions of the simulated QTL were not evenly placed, as indicated by the inward ticks

on the horizontal axis. Most QTL were placed in the left partof the genome. Some QTL were far apart from each other while others were clustered in some narrow regions. About half of the simulated QTL overlapped with true markers (known genotypes) and the remaining QTL were located between markers (having missing genotypes). We first generated a linear predictor, η_j, for each individual using the genotypes of the 20 simulated QTL and the true effects of these QTL. The linear predictor was then transformed into the probability of a binomial variable using $\mu_j = \Phi(\eta_j)$. We then simulated a zero-truncated Poisson variable with mean 4 as the number of trials for individual j, denoted by t_j (the number of trial must be greater than zero). We then simulated the number of events y_j from the corresponding binomial distribution defined by μ_j and t_j, i.e., $y_j \sim \text{Binomial}(\mu_j, t_j)$, The simulation experiment was replicated 1000 times.

In the binomial data analysis, we added a pseudo marker in every marker interval so that the total number of loci analyzed was $p = 241 + 240 = 481$ with 241 true markers and 240 pseudo markers. Genotypic probabilities of the pseudo markers were inferred from information of flanking markers. The hyper-parameter values $(\tau, \omega) = (-1, 0)$ were chosen for the analysis. We also tested $(\tau, \omega) = (0, 0)$ and the results were similar to the ones with $(\tau, \omega) = (-1, 0)$ and thus we only presented the results for $(\tau, \omega) = (-1, 0)$. The mixture model failed to converge and thus was not used in the binomial data analysis.

The average estimated QTL effects from the 1000 replicated simulations are depicted in Figure 4.1b (the panel in the middle) for the expectation model and Figure 4.1c (the bottom panel) for the overdispersion model. The true QTL effects are shown in the top panel of Figure 4.1 for comparison. The differences between the two models were barely noticed from the visual plots. The two models share the following common features:

(1) they both underestimated the QTL effects (bias towards zero due to shrinkage) and (2) a QTL with large effect was usually estimated as two or a few smaller QTL in the neighborhood of the true QTL. The simulation experiments allow us to evaluate the bias and estimation error of each QTL and eventually the mean squared error (MSE) for the QTL. Let $\bar{\gamma}_k$ be the average estimate of γ_k for the 1000 replicates and s_k^2 be the variance of the estimated γ_k across the replications, the MSE for γ_k is defined as

$$\text{MSE}_k = (\bar{\gamma}_k - \gamma_k)^2 + s_k^2 \tag{4.20}$$

The sum of MSE's for all QTL is

$$\text{MSE} = \sum_{k=1}^{p} (\bar{\gamma}_k - \gamma_k)^2 + \sum_{k=1}^{p} s_k^2 = \text{Bias} + \text{Error} \tag{4.21}$$

The MSE's for the two models (expectation and overdispersion) are shown in Table 4.1. The overdispersion model has a much smaller overall bias but the small bias was compromised by a slightly larger error. The overall MSE is 6.3081 for the overdispersion model and 7.1049 for the expectation model, indicating a noticeable improvement of the overdispersion model over the simple model.

Table 4.1: Comparison of the mean squared errors (MSE) for the three models in the replicated simulation study.

Data type	Model	Bias	Error	MSE
Binomial	Mixture	–	–	–
	Expectation	4.6925	2.4124	7.1049
	Overdispersion	3.5768	2.7313	6.3081
Binary	Mixture	4.6126	2.9622	7.5749
	Expected	4.7447	3.0791	7.8238
	Overdispersion	4.6055	2.9338	7.5393

The simulation experiment was replicated 1000 times. The mixture model did not work for the binomial data analysis.

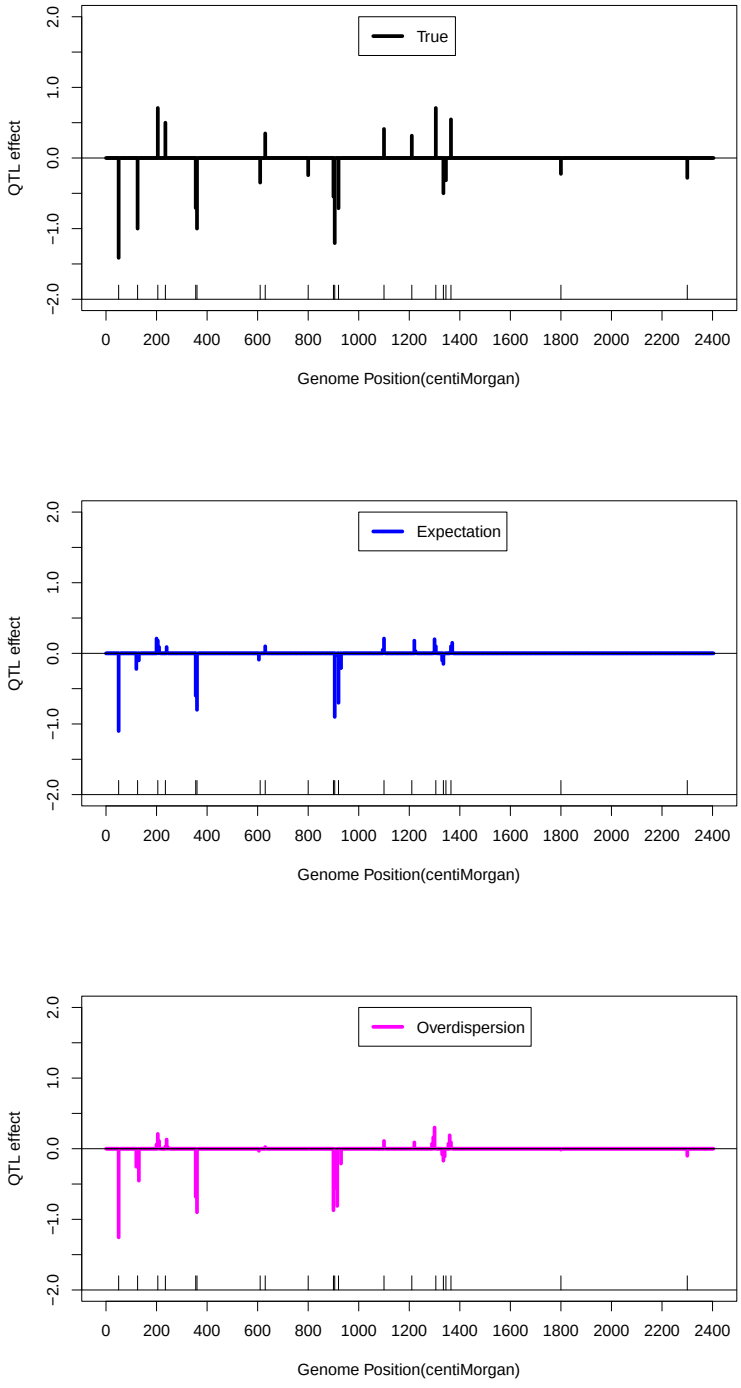


Figure 4.1: True QTL effects (top panel) and their estimated values for the simulated binomial trait using the expectation model (panel in the middle) and overdispersion model (bottom panel). The estimated QTL effects are the averages of 1000 replicated samples. The positions of 20 simulated QTL are indicated by the inward ticks on the horizontal axis.

4.3.1.2 Binary data

The experimental design was exactly the same as in the binomial experiment. The only difference in the simulation is that the trial was a fixed number of one for every individual in the binary data simulation experiment. For the binary data, the mixture model converged nicely and, therefore, each dataset was analyzed using three models (mixture, expectation and overdispersion). The average estimated QTL effects across the 1000 replicated simulations are depicted in Figure 4.2, where the mixture model is on the top panel, the expectation model on the panel in the middle and the overdispersion model on the bottom panel. The true QTL effects can be found in Figure 4.1 (the top panel). Again, the differences among the three models are barely noticeable. For the binary data, the models were only able to detect QTL with large effects and the estimated QTL effects were all biased towards zero (shrinkage). The biases, errors and MSE's of the binary data analysis are given in Table 4.1 also. The mixture and overdispersion models are much the same to each other but both have much smaller biases than the expectation model. The two models also have slightly smaller errors than the expectation model. As a result, both models have small MSE's relative to the expectation model.

The conclusion of the simulation studies were: (1) the mixture model did not work for binomial data analysis, (2) the overdispersion and mixture models worked equally well for binary traits, (3) both overdispersion and mixture models outperformed the expectation model in terms of generating smaller MSEs.

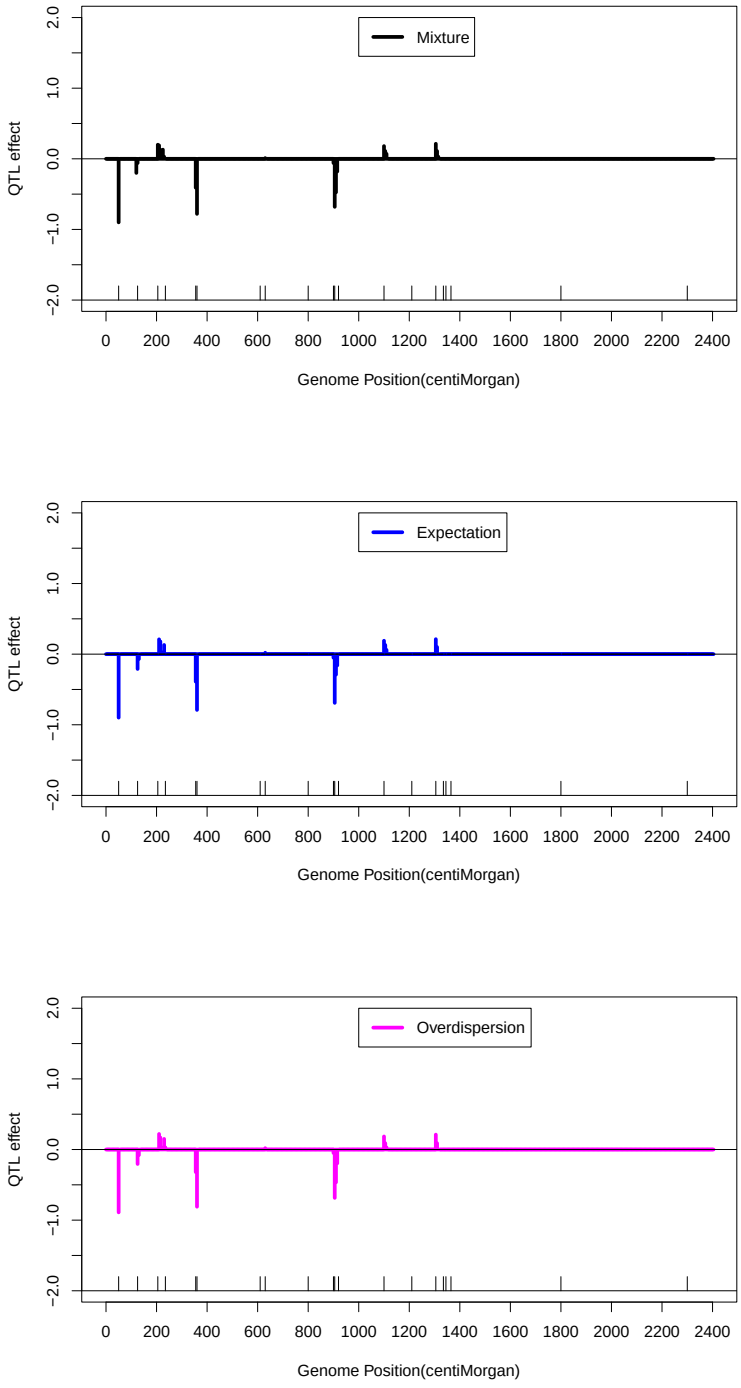


Figure 4.2: The estimated QTL effects for the simulated binary trait using the mixture model (top panel), expectation model (panel in the middle) and overdispersion model (bottom panel). The estimated QTL effects are the averages of 1000 replicated samples. The positions of 20 simulated QTL are indicated by the inward ticks on the horizontal axis.

4.3.2 Mapping wheat fertility QTL

The experiment was conducted by Dou et al. (2009). The mapping population contained 243 F_2 individuals derived from the cross of two inbred lines. The trait of interest is the female fertility measured as a binomial trait. The event is the number of seeded spikelets per plant (average 19.13 seeded spikelets) and the trial is the total number of spikelets per plant (average 25.15 spikelets). A total of 28 markers were genotyped in this experiment. These markers covered five chromosomes of the wheat genome with an average marker interval of 15.5 cM. The five chromosomes are only part of the wheat genome.

4.3.2.1 Binomial trait

Since the marker map is sparse, we inserted one pseudo marker in every two cM, generating a total of 197 loci (28 true markers and 169 pseudo markers). The pseudo markers have missing genotypes and the probability distributions of these pseudo markers were inferred from linked markers using the multipoint methods (Jiang and Zeng 1997). The sample size was $n = 243$ and the size of the model was $p = 197$. Since the sample size is much larger than the size of the model, we chose a weak shrinkage prior for the QTL variance represented by $(\tau, \omega) = (-2, 0)$, equivalent to the uniform prior for σ_k^2. Two models (expectation and overdispersion) were used for the binomial data analysis. Unfortunately, the mixture model approach did not work. The mixture model did work for binary traits, as demonstrated later. Therefore, we only presented the results for the expectation and overdispersion models for the binomial trait analysis.

For the real data analysis, we need to calculate the LOD score for each marker. The

LOD score test statistic was calculated using

$$\text{LOD}_k = \frac{\hat{\gamma}_k^2}{2\ln(10)\text{var}(\gamma_k)} \tag{4.22}$$

The estimated QTL effects for the two models are depicted in Figure 4.3a (the top panel). The LOD score profiles for the two models are depicted in Figure 4.3b (the bottom panel). The two models show some similarity and differences. Using the LOD = 3 as the threshold, the expectation model detected 17 QTL while the overdispersion model detected 15 QTL. Among these detected QTL, eight of them were detected by both models. The effects along with the estimation errors and the LOD scores obtained from the overdispersion model are listed in Table 4.2. Most detected QTL were located on chromosome II, IV and V. The QTL with the largest effect and LOD score occurred on the second chromosome at position 28.71 cM (cumulative position of 104.60 cM). Unlike the simulation study where the true effects of QTL were known, for the wheat data, the true QTL were not known. Therefore, we were not able to compare the biases and the mean square errors of the estimated QTL effects. We chose an alternative method to evaluate the two models, that is the leave-one-out cross validation. The cross validation only evaluates the predictabilities of the models. Due to linkage disequilibrium, a wrong model may have a high predictability. For the purpose of molecular breeding and marker assistant selection, higher predictability is more preferable. For the purpose of gene cloning, the biases of QTL estimates are of major concern. We used the Pearson correlation coefficient between the observed and predicted trait values as a measurement of the predictability. The Pearson correlation coefficients for the expectation model and overdispersion model were 0.5166 and 0.5290, respectively. The overdispersion model shows a slight advantage over the expectation model. We concluded that incorporation

of overdispersion does show the expected benefit (increase in predictability) in QTL mapping over the simple model.

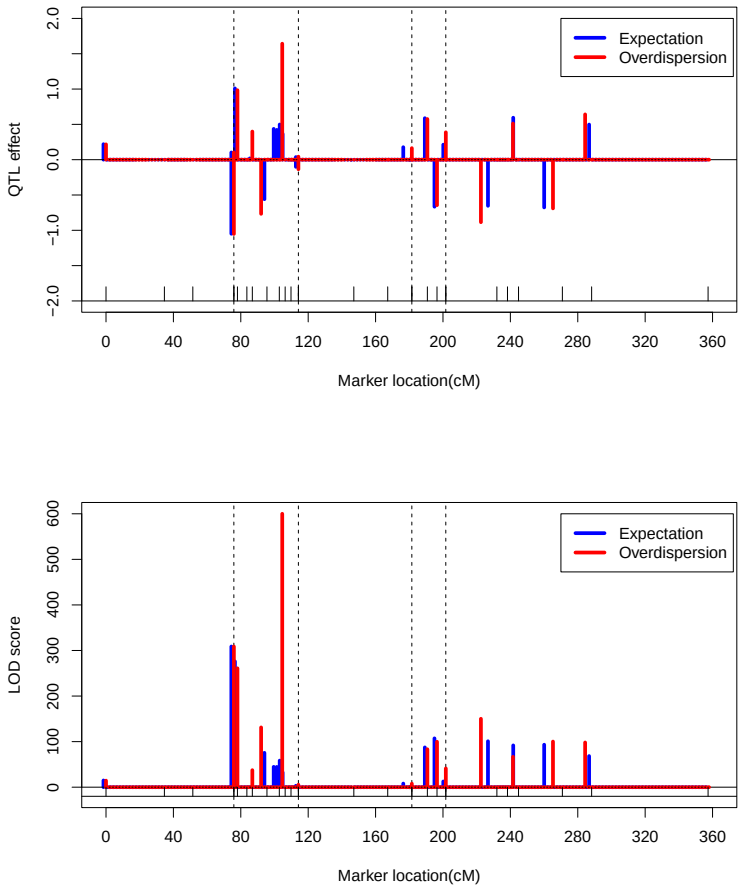


Figure 4.3: Binomial trait analysis of the wheat experiment using the expectation model (blue) and the overdispersion model (red). The top panel shows the estimated QTL effects and the bottom panel shows the LOD scores. Chromosomes are separated by the dotted vertical lines. Positions of true markers are indicated by the inward ticks on the horizontal axis.

Table 4.2: QTL detected for the binomial trait of wheat fertility using the overdispersion model.

QTL	Chromosome	Position	Marker [1]	Estimate [2]	StdErr [3]	LOD
1	1	0.00	1	0.2171	0.0266	14.40
2	2	0.00	1	-1.0517	0.0278	308.75
3	2	2.12	1	0.9841	0.0283	260.97
4	2	10.96	1	0.3985	0.0303	37.36
5	2	16.16	0	-0.7670	0.0311	131.24
6	2	28.70	0	1.6423	0.0306	621.89
7	2	38.29	1	-0.1356	0.0272	5.37
8	3	67.32	1	0.1635	0.0273	7.78
9	4	9.20	1	0.5755	0.0293	83.30
10	4	14.92	1	-0.6445	0.0300	99.78
11	5	0.00	1	0.3878	0.0281	41.17
12	5	20.83	0	-0.8852	0.0336	150.14
13	5	39.87	0	0.5121	0.0292	66.67
14	5	63.60	0	-0.6898	0.0321	100.17
15	5	82.68	0	0.6414	0.0301	98.17

[1] This column indicates whether the QTL overlaps with a true marker (1) or a pseudo marker (0).

[2] This column gives the estimated QTL effect.

[3] This column shows the standard error of the estimated QTL effect.

4.3.2.2 Binary trait

Among the 243 plants, 39 of them did not have seeds at all. The frequency distribution of the number of seeded spikelets is shown in Figure 4.4. It appears that the zero category was inflated. The binomial data analysis did not differentiate QTL responsible for seed presence and absence. We now defined a binary trait as seed presence/absence and used the three models (expectation, overdispersion and mixture) to analyze the binary trait. The estimated QTL effect profiles are shown in Figure 4.5a (the top panel) and the LOD score profiles are depicted in Figure 4.5b (the bottom panel). The three models appeared to generate much the same result. Using the LOD 3 criterion, we only detected a single QTL at position 28.71 cM of chromosome II (cumulative position 104.60 cM). This QTL was the same one detected for the binomial trait (the largest QTL for the binomial trait). Our conclusion was that, except this particular QTL, the multiple QTL detected for the binomial trait reported early were all responsible for the variation of the number of seeded spikelets, not the seed presence/absence trait. Leave-one-out cross validation analysis did not show much differences for the three models. The Pearson correlation coefficients between the observed and predicted trait values were 0.4715, 0.4729 and 0.4721, respectively, for the three models (expectation, overdispersion and mixture). As expected, the predictability for the binary trait is lower than that of the binomial trait.

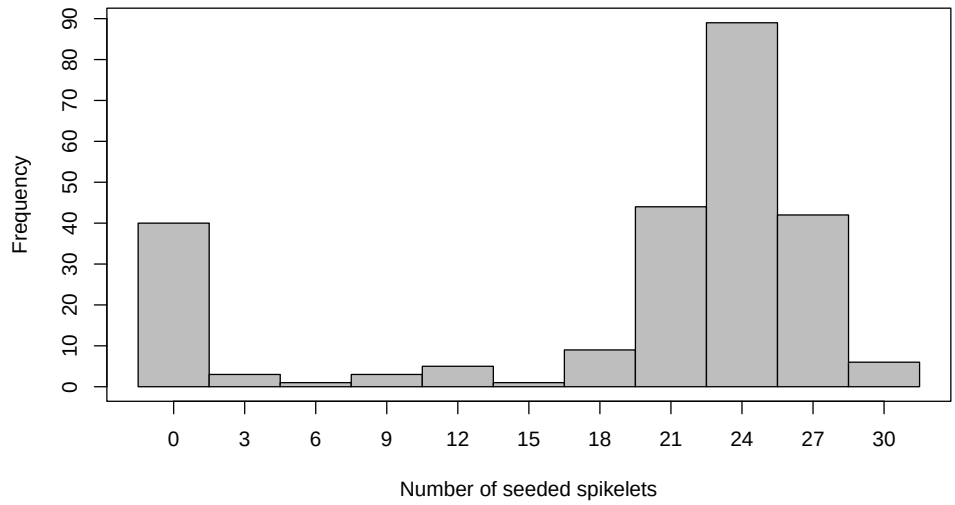


Figure 4.4: Frequency distribution of the number of seeded spikelets of the F_2 wheat population. Among the 243 plants, 39 of them have no seeds (zero category).

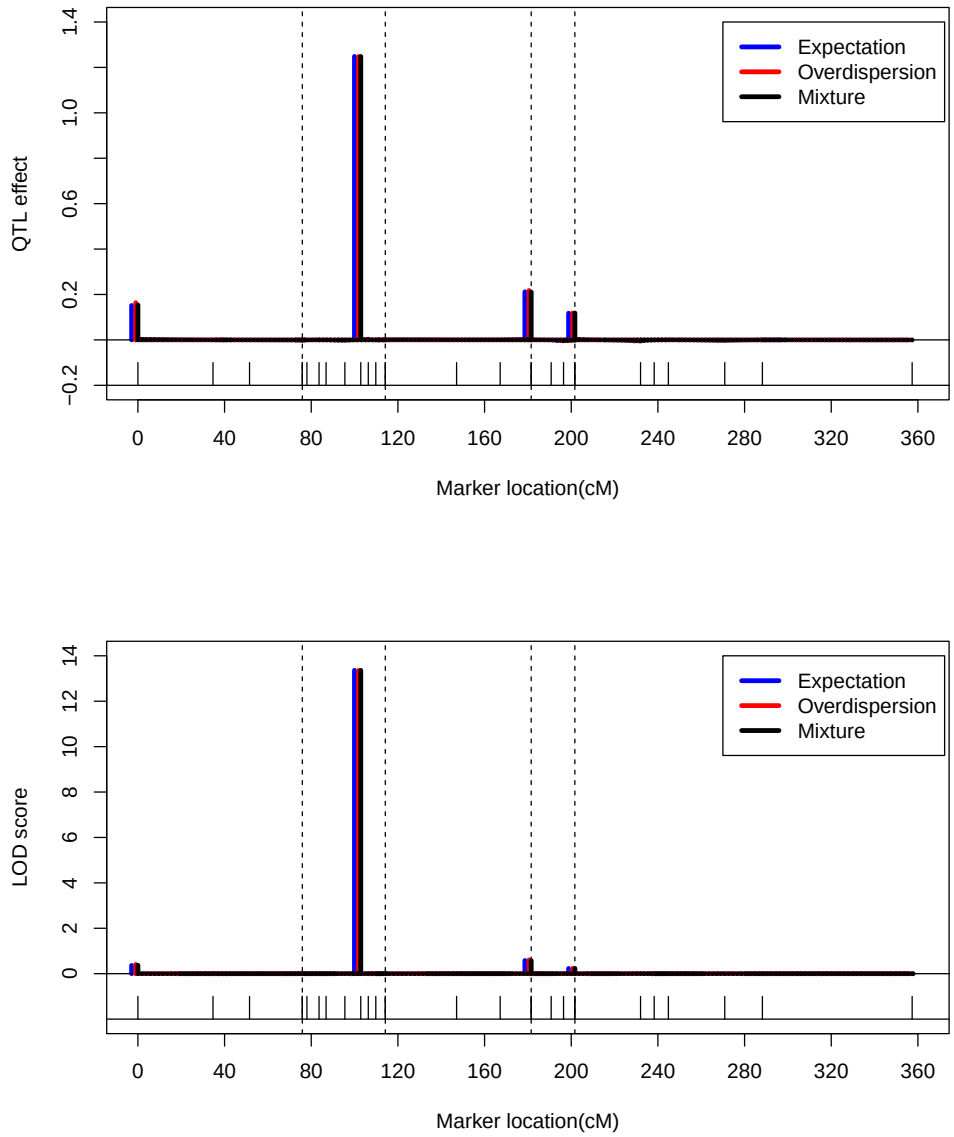


Figure 4.5: Binary trait (seed presence/absence) analysis using the expectation model (blue), overdispersion model (red) and mixture model (black). The top panel shows the estimated QTL effects and the bottom panel shows the LOD scores. Chromosomes are separated by the dotted vertical lines. Positions of true markers are indicated by the inward ticks on the horizontal axis.

4.3.2.3 Truncated binomial trait

We deleted the 39 individuals with zero seeds and only analyzed the 204 seeded plants. This analysis would detect QTL only responsible for the variation of the number of seeded spikelets. Since the trait had a zero-truncated binomial distribution, the modified log likelihood is

$$L(\beta,\gamma)=\sum_{j=1}^{n}\left\{y_j\ln(\mu_j)+(t_j-y_j)\ln(1-\mu_j)-\ln\left[1-(1-\mu_j)^{t_j}\right]\right\} \tag{4.23}$$

The last term, $1-(1-\mu_j)^{t_j}$,is the probability of $y_j>0$.The truncated binomial trait was only analyzed using the expectation and overdispersion models because the mixture model failed to converge. The results are depicted in Figure 6a (the top panel) for the estimated QTL effects and the Figure 6b (the bottom panel) for the LOD scores. The expectation model detected six QTL and the overdispersion model detected seven. Five QTL were detected by both models. The largest QTL on chromosome II at position 28.71 cM (cumulative position 104.60 cM) detected by both the binomial and binary trait analyses remain significant for the zero-truncated binomial trait.

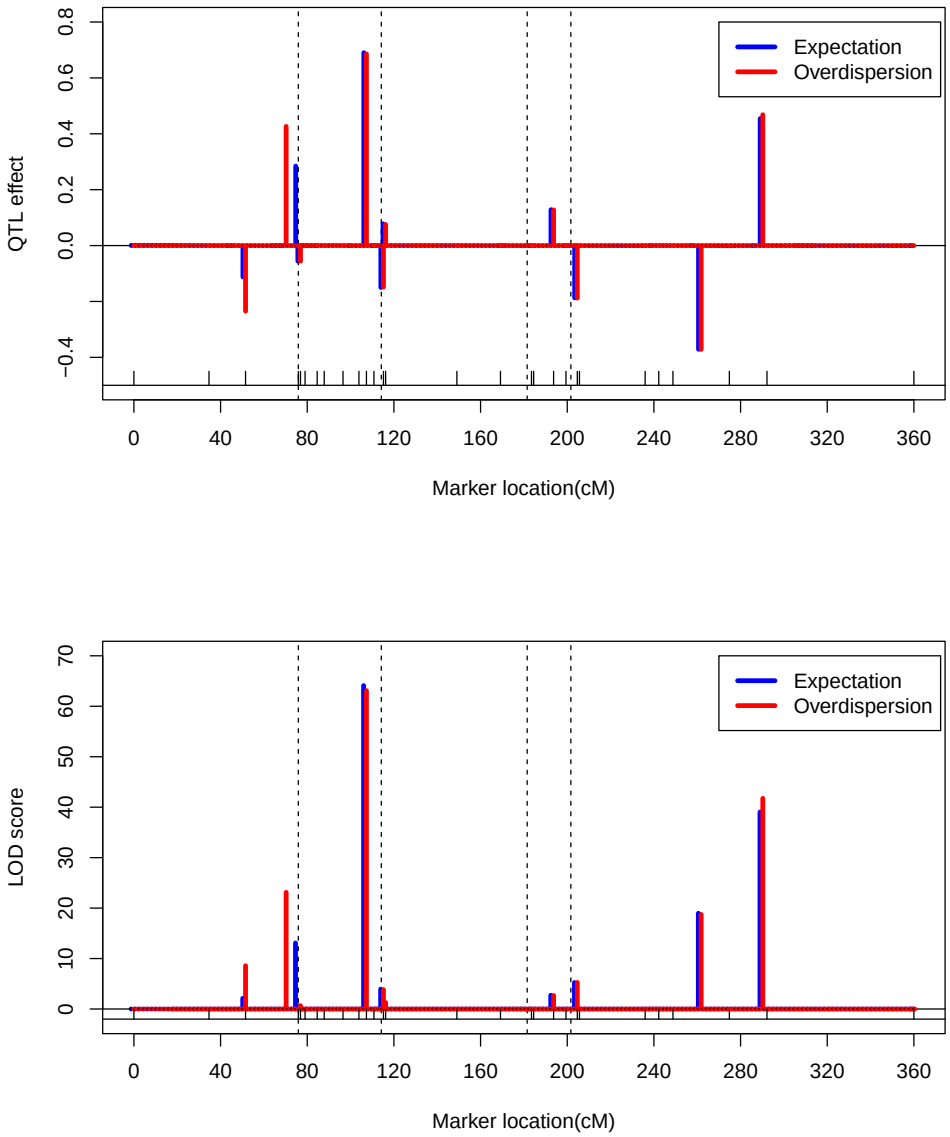


Figure 4.6: Zero-truncated binomial trait (excluding plants with no seeds) analysis using the expectation model (blue) and overdispersion model (red). The top panel shows the estimated QTL effects and the bottom panel shows the LOD scores. Chromosomes are separated by the dotted vertical lines. Positions of true markers are indicated by the inward ticks on the horizontal axis.

4.4 Discussion

We proposed three algorithms to handle the missing genotype problem in multiple QTL mapping. The overdispersion algorithm appears to be optimal over the expectation and mixture model algorithm in terms of less bias, small MSE and high predictability. The advantage of the overdispersion algorithm over the expectation algorithm will diminish as the marker density increases. In the situation where the entire genome is sequenced, all three algorithms would converge to the same result because genotypes of all markers will be observed. However, full genome sequences for most species are not expected soon. In addition, missing genotypes may still exist due to human and technical errors in experiments. Therefore, the missing genotype handling algorithms remain useful for the foreseeable future.

The generalized linear mixed model is sufficiently general so that it can handle traits with any distributions as long as a likelihood function is programmable. Normal distribution is included as a special case. The Newton-Raphson step is on the likelihood function to infer the parameters using information from the data. Combining the prior information and the data, the posterior distribution is inferred. An alternative approach is to perform the Newton-Raphson step on the log posterior, as done by McGilchrist (1994). For the normal priors of QTL effects, this approach is the same as what we did. The advantage of the Newton-Raphson on the log posterior is that we can choose any arbitrary priors for the QTL effects other than the normal priors. The Newton-Raphson step does not need to have an explicit form because the first and second partial derivatives of the log posterior can be easily found numerically through efficient subroutines of computer programs.

The generalized linear mixed model proposed differs from the usual GLMM in that

the variance components are assigned a scaled inverse chi-square distribution to further shrink the variance components towards zero. An obvious question is how to select the shrinkage parameters. In the simulated data analysis, we chose $(\tau, \omega) = (-1, 0)$ as the hyper parameters. For the real data analysis, $(\tau, \omega) = (-2, 0)$ was used. The optimal way of choosing the hyper parameters is to use a cross validation scheme. There are two hyper parameters in the scaled inverse chi-square prior. This means that the search is two-dimensional. Using the grid search scheme, the range of the hyper parameters may take $-2 \leq \tau \leq 0$ and $0 \leq \omega \leq 0.5$. Our past experience showed that we may set $\omega = 0$ and only vary τ (Xu 2010). This will leave the search as one dimensional. A general guideline is to choose a large τ value for large models. In the extreme case where $\tau = 0$, the scaled inverse chi-square is proportional to $1/\sigma_k^2$, the Jeffreys' prior (Jeffreys 1946). This prior leads to is a strong shrinkage. When $\tau = -2$, the prior is uniform and the model becomes a standard generalized linear mixed model if genotypes of putative QTL are fully observed.

In the wheat fertility QTL mapping experiment, the binomial trait appears to be inflated for the zero category. Such a zero-inflated binomial data can be analyzed using the zero inflated binomial distribution (Hall 2000) assuming that QTL genotypes are observed and all effects are fixed. A generalized linear mixed model for zero-inflated binomial traits has not been available yet and it is an interesting topic to be developed. We used an ad hoc approach to analyzing such a zero-inflated binomial trait by performing two separate analyses. One analysis was the binary trait generalized linear mixed model by treating seed presence and absence as 1 and 0, respectively, regardless how many seeds carried by each plant. The other analysis was a zero-truncated binomial trait analysis where only plants with at least one seeded spikelets were subject to the analysis. The log likelihood function was modified to take into account the zero truncation.

Interestingly, we found one QTL responsible for both the binary trait and the truncated binomial trait. Many QTL were found only affected the zero-truncated binomial trait. This discovery is novel and useful to plant breeders for further investigation. We do not expect much difference between this ad hoc analysis and the actual zero-inflated binomial trait analysis. Of course, this is only a speculation and an accurate answer will not be achieved unless a zero-inflated binomial trait analysis is actually done.

Bibliography

[1] Y. Amit. On rates of convergence of stochastic relaxation for gaussian and non-gaussian distributions. *Journal of Multivariate Analysis*, 38(1):82–99, 1991.

[2] CI Amos, RC Elston, GE Bonney, BJ Keats, and GS Berenson. A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. *American journal of human genetics*, 47(2):247, 1990.

[3] Kannan R. Polson N. Applegate, D. Random polynomial time algorithms for sampling from joint distribution. *Technical report, School of Computer Science, Carnegie Mellon University*.

[4] D. Ashby. Bayesian statistics in medicine: a 25 year review. *Statistics in medicine*, 25(21):3589–3631, 2006.

[5] M. Bayes and M. Price. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs. *Philosophical Transactions*, 53:370, 1763.

[6] M.A. Beaumont, W. Zhang, and D.J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025, 2002.

[7] P. Beerli. Comparison of bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, 22(3):341, 2006.

[8] J.O. Berger and J.M. Bernardo. Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association*, pages 200–207, 1989.

[9] J.M. Bernardo. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979.

[10] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, pages 3–41, 1995.

[11] M.P. Boer, D. Wright, L. Feng, D.W. Podlich, L. Luo, M. Cooper, and F.A. Van Eeuwijk. A mixed-model quantitative trait loci (qtl) analysis for multiple-environment trial data using environmental covariables for qtl-by-environment interactions, with an example in maize. *Genetics*, 177(3):1801, 2007.

[12] J. Bollback. Simmap: stochastic character mapping of discrete traits on phylogenies. *Bmc Bioinformatics*, 7(1):88, 2006.

[13] L. Breiman. *Probability*. MA:Addison-Wesley, 1968.

[14] N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, pages 9–25, 1993.

[15] K.W. Broman and T.P. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):641–656, 2002.

[16] S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, pages 434–455, 1998.

[17] B.P. Carlin, K. Chaloner, T. Church, T.A. Louis, and J.P. Matts. Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. *The Statistician*, pages 355–367, 1993.

[18] G. Casella and E.I. George. Explaining the gibbs sampler. *American Statistician*, pages 167–174, 1992.

[19] KS Chan. Asymptotic behavior of the gibbs sampler. *Journal of the American Statistical Association*, pages 320–326, 1993.

[20] M.H. Chen, Q.M. Shao, and J.G. Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer Verlag, 2000.

[21] S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *American Statistician*, pages 327–335, 1995.

[22] S. Chib and E. Greenberg. Markov chain monte carlo simulation methods in econometrics. *Econometric theory*, 12(03):409–431, 1996.

[23] G.A. Churchill and R.W. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963, 1994.

[24] P. Congdon. *Applied bayesian modelling*, volume 394. John Wiley & Sons Inc, 2003.

[25] P. Congdon. *Bayesian models for categorical data*, volume 626. John Wiley & Sons Inc, 2005.

[26] P. Congdon. *Bayesian statistical modelling*, volume 670. Wiley, 2006.

[27] G. Consonni and P. Veronese. Conjugate priors for exponential families having quadratic variance functions. *Journal of the American Statistical Association*, pages 1123–1127, 1992.

[28] M.J. Crowder. Beta-binomial anova for proportions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(1):34–37, 1978.

[29] Y. Cui, D.Y. Kim, and J. Zhu. On the generalized poisson regression mixture model for mapping quantitative trait loci with count data. *Genetics*, 174(4):2159, 2006.

[30] Y. Cui and W. Yang. Zero-inflated generalized poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. *Journal of theoretical biology*, 256(2):276–285, 2009.

[31] F.M. Demenais, A.E. Laing, and G.E. Bonney. Numerical comparisons of two formulations of the logistic regressive models with the mixed model in segregation analysis of discrete traits. *Genetic epidemiology*, 9(6):419–435, 1992.

[32] W. Deng, H. Chen, and Z. Li. A logistic regression mixture model for interval mapping of genetic trait loci affecting binary phenotypes. *Genetics*, 172(2):1349, 2006.

[33] C.M. Desouza. An empirical bayes formulation of cohort models in cancer epidemiology. *Statistics in Medicine*, 10(8):1241–1256, 1991.

[34] D. Dey, S.K. Ghosh, and B.K. Mallick. *Generalized linear models: A Bayesian perspective*, volume 5. CRC, 2000.

[35] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, pages 269–281, 1979.

[36] B. DOU, B. HOU, H. XU, X. LOU, X. CHI, J. YANG, F. WANG, Z. NI, and Q. SUN. Efficient mapping of a female sterile gene in wheat (triticum aestivum l.). *Genetics Research*, 91(05):337–343, 2009.

[37] MD Edwards, CW Stuber, and JF Wendel. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. i. numbers, genomic distribution and types of gene action. *Genetics*, 116(1):113, 1987.

[38] B. Efron and R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86, 2002.

[39] RC Elston, KK Namboodiri, CJ Glueck, R. Fallat, R. Tsang, and V. Leuba. Study of the genetic transmission of hypercholesterolemia and hypertriglyceridemia in a 195 member kindred. *Annals of human genetics*, 39(1):67–87, 1975.

[40] W. Feller. *An introduction to probability theory and its applications*. John Wiley & Sons, 1968.

[41] A.E. Gelfand, S.E. Hills, A. Racine-Poon, and A.F.M. Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, pages 972–985, 1990.

[42] A. Gelman. *Bayesian data analysis*. CRC press, 2004.

[43] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[44] A. Gelman, I. Van Mechelen, G. Verbeke, D.F. Heitjan, and M. Meulders. Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, 61(1):74–85, 2005.

[45] S. German and D. German. Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell*, 6(9):721–741, 1984.

[46] Bernardo J. M. Berger J.O. Dawiv A.P. Geweke, J. and eds. Smith, A. F. M. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, 4, 1992.

[47] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1996.

[48] L.R. Goldin, R.C. Elston, J.B. Graham, C.H. Miller, and E.A. Murphy. Genetic analysis of von willebrand's disease in two large pedigrees: a multivariate approach. *American Journal of Medical Genetics*, 6(4):279–293, 1980.

[49] C.A. Hackett and J.I. Weller. Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics*, pages 1252–1263, 1995.

[50] D.B. Hall. Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics*, pages 1030–1039, 2000.

[51] M. Halperin, D.L. Demets, and J.H. Ware. Early methodological developments for clinical trials at the national heart, lung and blood institute. *Statistics in medicine*, 9(8):881–892, 1990.

[52] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97, 1970.

[53] P. Heidelberger and P.D. Welch. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24(4):233–245, 1981.

[54] D.F. Heitjan. Bayesian interim analysis of phase ii cancer clinical trials. *Statistics in Medicine*, 16(16):1791–1802, 1997.

[55] C. R. Henderson. Estimation of genetic parameters (abstract). *Ann. Math. Statist*, 21:309–310, 1950.

[56] I. Hoeschele and PM VanRaden. Bayesian analysis of linkage between genetic markers and quantitative trait loci. ii. combining prior knowledge with experimental evidence. *TAG Theoretical and Applied Genetics*, 85(8):946–952, 1993.

[57] K.E. Holsinger and L.E. Wallace. Bayesian approaches for the analysis of population genetic structure: an example from platanthera leucophaea (orchidaceae). *Molecular Ecology*, 13(4):887–894, 2004.

[58] N.J. Horton and N.M. Laird. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, 8(1):37, 1999.

[59] G.J. Hunt, AM Collins, R. Rivera, RE Page, and E. Guzman-Novoa. Brief communication. quantitative trait loci influencing honeybee alarm pheromone levels. *Journal of Heredity*, 90(5):585, 1999.

[60] J.G. Ibrahim, M.H. Chen, and S.R. Lipsitz. Bayesian methods for generalized linear models with missing covariates. *The Canadian Journal of Statistics*, 30:55–78, 2002.

[61] J.G. Ibrahim, M.H. Chen, S.R. Lipsitz, and A.H. Herring. Missing-data methods for generalized linear models. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[62] SAS Institute Inc. *The GLIMMIX Procedure, SAS/STAT 9.2 User's Guide*. 2008.

[63] R.C. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135(1):205, 1993.

[64] SD Jayakar. On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. *Biometrics*, pages 451–464, 1970.

[65] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, pages 453–461, 1946.

[66] H. S. Jeffreys. *Theory of probability*. Oxford university press, 1939.

[67] C. Jiang and Z.B. Zeng. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica*, 101(1):47–58, 1997.

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

[69] R.E. Kass, B.P. Carlin, A. Gelman, and R.M. Neal. Markov chain monte carlo in practice: a roundtable discussion. *American Statistician*, pages 93–100, 1998.

[70] A. Kopp, R.M. Graze, S. Xu, S.B. Carroll, and S.V. Nuzhdin. Quantitative trait loci responsible for variation in sexually dimorphic traits in drosophila melanogaster. *Genetics*, 163(2):771, 2003.

[71] L. Kruglyak and E.S. Lander. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, 57(2):439, 1995.

[72] L. Kruglyak and E.S. Lander. A nonparametric approach for mapping quantitative trait loci. *Genetics*, 139(3):1421, 1995.

[73] E.S. Lander and D. Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185, 1989.

[74] C. Lange and J.C. Whittaker. Mapping quantitative trait loci using generalized estimating equations. *Genetics*, 159(3):1325, 2001.

[75] S.H. Lee, J.H.J. van der Werf, B.J. Hayes, M.E. Goddard, and P.M. Visscher. Predicting unobserved phenotypes for complex traits from whole-genome snp data. *PLoS genetics*, 4(10):e1000231, 2008.

[76] A. Legarra, C. Robert-Granié, E. Manfredi, and J.M. Elsen. Performance of genomic selection in mice. *Genetics*, 180(1):611, 2008.

[77] D.V. Lindley. *Bayesian statistics, a review(CBMS-NSF regional conference series in applied mathematics)*. Society for Industrial Mathematics, 1987.

[78] J.S. Liu. *Monte Carlo strategies in scientific computing*. Springer Verlag, 2001.

[79] Wong W. H. and Kong A. Liu, C. Correlation structure and convergence rate of the gibbs sampler (i): Application to the comparison of estimators and augmentation scheme. *Technical report, School of Statistics, University of Chicago*, 1991b.

[80] Wong W. H. and Kong A. Liu, C. Correlation structure and convergence rate of the gibbs sampler (ii): Applications to various scans. *Technical report, Department of Statistics, University of Chicago*, 1991b.

[81] Y.C. Loke, S.B. Tan, Y.Y. Cai, and D. Machin. A bayesian dose finding design for dual endpoint phase i trials. *Statistics in medicine*, 25(1):3–22, 2006.

[82] O. Loudet, S. Chaillou, C. Camilleri, D. Bouchez, and F. Daniel-Vedele. Bay-0× shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in arabidopsis. *TAG Theoretical and Applied Genetics*, 104(6):1173–1184, 2002.

[83] ZW Luo, E. Potokina, A. Druka, R. Wise, R. Waugh, and MJ Kearsey. Sfp genotyping from affymetrix arrays is robust but largely detects cis-acting expression regulators. *Genetics*, 176(2):789, 2007.

[84] A. Manichaikul, J. Dupuis, Ś. Sen, and K.W. Broman. Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. *Genetics*, 174(1):481, 2006.

[85] A. Manichaikul, J.Y. Moon, Ś. Sen, B.S. Yandell, and K.W. Broman. A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics*, 181(3):1077, 2009.

[86] CA Matos, DL Thomas, D. Gianola, RJ Tempelman, and LD Young. Genetic analysis of discrete reproductive traits in sheep using linear and nonlinear models: I. estimation of genetic parameters. *Journal of Animal Science*, 75(1):76, 1997.

[87] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & Hall/CRC, 1989.

[88] C.E. McCulloch and J.M. Neuhaus. *Generalized Linear Mixed Model. Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd, 2005.

[89] CA McGilchrist. Estimation in generalized mixed models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 61–69, 1994.

[90] N. Meteopolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[91] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, et al. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087, 1953.

[92] TH Meuwissen, BJ Hayes, and ME Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819, 2001.

[93] S.P. Meyn, R.L. Tweedie, and P.W. Glynn. *Markov chains and stochastic stability*. Springer London et al., 1993.

[94] G. Milliken and D. E. Johnson. *Analysis of Messy Data, Volume 1: Designed Experiments*. Chapman & Hall/CRC, 2009.

[95] EA Murphy and GS Mutalik. The application of bayesian methods in genetic counselling. *Human Heredity*, 19(2):126–151, 1969.

[96] R.C. Elston C.J. Glueck R. Fallat C.R. Buncher Namboodiri, K.K. and R. Tsangs. Bivariate analysis of cholesterol and triglyceride levels in families in which probands have type tib lipoprotein phenotype. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 454–471, 1994.

[97] J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 370–384, 1972.

[98] I. Ntzoufras. *Bayesian modeling using WinBUGS*, volume 698. John Wiley & Sons Inc, 2009.

[99] J.L. Palmer and P. Müller. Bayesian optimal design in population models for haematologic data. *Statistics in medicine*, 17(14):1613–1622, 1998.

[100] A. Racine-Poon, AP Grieve, H. Fluhler, and AFM Smith. Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Applied Statistics*, 35:93–150, 1986.

[101] S. Rao and XU Shizhong. Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity*, 81(2):214–224, 1998.

[102] A. Rebai. Comparison of methods for regression interval mapping in qtl analysis with non-normal traits. *Genetical research*, 69(01):69–74, 1997.

[103] C.P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Verlag, 2004.

[104] J. M. Robins and A Rotnitzky. On double robustness. *Statistica Sinica*, 11:920–936, 2001.

[105] J.S. Rosenthal. Rates of convergence for data augmentation on finite sample spaces. *Technical report, Department of Mathematics, Harvard University*, 1991a.

[106] J.S. Rosenthal. Rates of convergence for gibbs sampling for variance component models. *Technical report, Department of Mathematics, Harvard University*, 1991b.

[107] D.B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 519. Wiley Online Library, 1987.

[108] J.M. Satagopan, B.S. Yandell, M.A. Newton, and T.C. Osborn. A bayesian approach to detect quantitative trait loci using markov chain monte carlo. *Genetics*, 144(2):805, 1996.

[109] K. Sax. The association of size differences with seed-coat pattern and pigmentation in phaseolus vulgaris. *Genetics*, 8:552, 1923.

[110] M.J. Schervish and B.P. Carlin. On the convergence of successive substitution sampling. *Journal of Computational and Graphical statistics*, pages 111–127, 1992.

[111] R. Schlaifer and H. Raiffa. Applied statistical decision theory. 1961.

[112] D. Schluter, T. Price, A.Ø. Mooers, and D. Ludwig. Likelihood of ancestor states in adaptive radiation. *Evolution*, pages 1699–1711, 1997.

[113] S.R. Searle. *Linear models*. John Wiley & Sons, Inc., 1997.

[114] M.J. Sillanpää and E. Arjas. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, 148(3):1373, 1998.

[115] TR Solberg, R. Shepherd, and JA Woolliams. A fast algorithm for bayesb type of prediction of genome-wide estimates of genetic value. *Genetics, Selection, Evolution*, 41(2), 2009.

[116] M. Soller, 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. *TAG Theoretical and Applied Genetics*, 47(1):35–39, 1976.

[117] D. Sorensen and D. Gianola. *Likelihood, Bayesian and MCMC methods in quantitative genetics*. Springer, 2002.

[118] D.J. Spiegelhalter, K.R. Abrams, and J.P. Myles. *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. Wiley, 2004.

[119] M. Stephens and D.J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.

[120] S. Sturtz, U. Ligges, and A. Gelman. R2winbugs: a package for running winbugs from r. *Journal of Statistical Software*, 12(3):1–16, 2005.

[121] S.D. Tanksley, H. Medina-Filho, and C.M. Rick. Use of naturally-occurring enzyme variation to detect and map genes controlling quantitative traits in an interspecific backcross of tomato. *Heredity*, 49(1):11–26, 1982.

[122] Martin Abba Tanner. *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*. Springer Verlag, 1993.

[123] B. Taran, T.E. Michaels, and K.P. Pauls. Genetic mapping of agronomic traits in common bean. *Crop Sci*, 42:544–556, 2002.

[124] C.J.F. Ter Braak, M.P. Boer, and M.C.A.M. Bink. Extending xu's bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics*, 170(3):1435, 2005.

[125] J.M. Thoday. Effects of disruptive selection. iii. coupling and repulsion. *Heredity*, 14:35–49, 1960.

[126] R. Tibshirani. Noninformative priors for one parameter of many. *Biometrika*, 76(3):604, 1989.

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

[128] L. Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[129] R.M. Turner, R.Z. Omar, and S.G. Thompson. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in medicine*, 20(3):453–472, 2001.

[130] J. Wakefield and A. Racine-Poon. An application of bayesian population pharmacokinetic/pharmacodynamic models to dose recommendation. *Statistics in medicine*, 14(9):971–986, 1995.

[131] H. Wang, Y.M. Zhang, X. Li, G.L. Masinde, S. Mohan, D.J. Baylink, and S. Xu. Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics*, 170(1):465, 2005.

[132] R.W.M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika*, 61(3):439, 1974.

[133] R. Wolfinger and M. O'CONNELL. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, 48(3-4):233–243, 1993.

[134] C. Xu, Z. Li, and S. Xu. Joint mapping of quantitative trait loci for multiple binary characters. *Genetics*, 169(2):1045, 2005.

[135] C. Xu, YM Zhang, and S. Xu. An em algorithm for mapping quantitative resistance loci. *Heredity*, 94(1):119–128, 2004.

[136] S. Xu. Estimating polygenic effects using markers of the entire genome. *Genetics*, 163(2):789, 2003.

[137] S. Xu. An expectation–maximization algorithm for the lasso estimation of quantitative trait locus effects. *Heredity*, 105(5):483–494, 2010.

[138] S. Xu and W.R. Atchley. Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics*, 143(3):1417, 1996.

[139] S. Xu and Z. Hu. Generalized linear model for interval mapping of quantitative trait loci. *TAG Theoretical and Applied Genetics*, 121(1):47–63, 2010.

[140] S. Xu and N. Yi. Mixed model analysis of quantitative trait loci. *Proceedings of the National Academy of Sciences*, 97(26):14542, 2000.

[141] S. Xu, N. Yi, D. Burke, A. Galecki, and R.A. Miller. An em algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. *Genetical research*, 82(2):127–138, 2003.

[142] M. Yano, Y. Harushima, Y. Nagamura, N. Kurata, Y. Minobe, and T. Sasaki. Identification of quantitative trait loci controlling heading date in rice using a high-density linkage map. *TAG Theoretical and Applied Genetics*, 95(7):1025–1032, 1997.

[143] N. Yi and S. Banerjee. Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics*, 181(3):1101, 2009.

[144] N. Yi and D. Shriner. Advances in bayesian multiple quantitative trait loci mapping in experimental crosses. *Heredity*, 100(3):240–252, 2007.

[145] N. Yi and S. Xu. Mapping quantitative trait loci for complex binary traits in outbred populations. *Heredity*, 82(6):668–676, 1999.

[146] N. Yi and S. Xu. A random model approach to mapping quantitative trait loci for complex binary traits in outbred populations. *Genetics*, 153(2):1029, 1999.

[147] N. Yi and S. Xu. Bayesian mapping of quantitative trait loci under the identity-by-descent-based variance component model. *Genetics*, 156(1):411, 2000.

[148] N. Yi and S. Xu. Bayesian mapping of quantitative trait loci under complicated mating designs. *Genetics*, 157(4):1759, 2001.

[149] N. Yi and S. Xu. Bayesian lasso for quantitative trait loci mapping. *Genetics*, 179(2):1045, 2008.

[150] N. Yi, S. Xu, and D.B. Allison. Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics*, 165(2):867, 2003.

[151] N. Yi, S. Xu, V. George, and D.B. Allison. Mapping multiple quantitative trait loci for ordinal traits. *Behavior genetics*, 34(1):3–15, 2004.

[152] N. Yi, B.S. Yandell, G.A. Churchill, D.B. Allison, E.J. Eisen, and D. Pomp. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, 170(3):1333, 2005.

[153] Z.B. Zeng. Precision mapping of quantitative trait loci. *Genetics*, 136(4):1457, 1994.

[154] F. Zou, J.P. Fine, J. Hu, and DY Lin. An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics*, 168(4):2307, 2004.