Chapter 15
Bayesian Multiple QTL Mapping

So far we have learned the least-squares method, the weighted least squares method, and the maximum likelihood method for QTL mapping. These methods share a common problem in handling multiple QTL, that is, the problem of multicollinearity. Therefore, a model can include only a few QTL. Recently, Bayesian method has been developed for mapping multiple QTL (Satagopan et al. 1996; Heath 1997; Sillanpää and Arjas 1998; Sillanpää and Arjas 1999; Xu 2003; Yi 2004; Wang et al. 2005b; Yi and Shriner 2008). Under the Bayesian framework, the model can tolerate a much higher level of multicollinearity than the maximum likelihood method. As a result, the Bayesian method can handle highly saturated model. This chapter is focused on the Bayesian method via the Markov chain Monte Carlo (MCMC) algorithm. Before introducing the methods of Bayesian mapping, it is necessary to review briefly the background knowledge of Bayesian statistics.

15.1 Bayesian Regression Analysis

We will learn the basic principle and method of Bayesian analysis using a simple regression model as an example. The simple regression model has the following form:

$$y_j = X_j\beta + \epsilon_j, \forall j = 1, \ldots, n \tag{15.1}$$

where y_j is the response (dependent) variable, X_j is the regressor (independent variable), β is the regression coefficient, and ϵ_j is the residual error with an assumed $N(0, \sigma^2)$ distribution. This model is a special case of

$$y_j = \alpha + X_j\beta + \epsilon_j, \forall j = 1, \ldots, n \tag{15.2}$$

with $\alpha = 0$, i.e., regression through the origin. We use this special model to derive the Bayesian estimates of parameters. In subsequent sections, we will extend

S. Xu, *Principles of Statistical Genomics*, DOI 10.1007/978-0-387-70807-2_15,

the model to the usual regression with nonzero intercept and also regression with multiple explanatory variables (multiple regression). The log likelihood function is

$$L(\theta) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - X_j\beta)^2 \tag{15.3}$$

where $\theta = \{\beta, \sigma^2\}$. The MLEs of θ are

$$\hat{\beta} = \left(\sum\nolimits_{j=1}^{n} X_j^2\right)^{-1}\left(\sum\nolimits_{j=1}^{n} X_j y_j\right) \tag{15.4}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - X_j\hat{\beta})^2 \tag{15.5}$$

In the maximum likelihood analysis, parameters are estimated from the data. Sometimes investigators have prior knowledge of the parameters. This prior knowledge can be incorporated into the analysis to improve the estimation of parameters. This is the primary purpose of Bayesian analysis. The prior knowledge is formulated as a prior distribution of the parameters. Let $p(\beta, \sigma^2)$ be the joint prior density of θ. Usually, we assume that β and σ^2 are independent so that

$$p(\beta, \sigma^2) = p(\beta)p(\sigma^2) \tag{15.6}$$

The choice of $p(\beta)$ and $p(\sigma^2)$ depends on investigator's knowledge on the problem and mathematical attractiveness. In the simple regression analysis, the following priors are both legitimate and attractive, which are

$$p(\beta) = N(\beta|\mu_\beta, \sigma_\beta^2) \tag{15.7}$$

and

$$p(\sigma^2) = \text{Inv} - \chi^2(\sigma^2|\tau, \omega) \tag{15.8}$$

where $N(\beta|\mu_\beta, \sigma_\beta^2)$ is the notation for the normal density of variable β with mean μ_β and variance σ_β^2, and $\text{Inv} - \chi^2(\sigma^2|\tau, \omega)$ is the probability density for the scaled inverse chi-square distribution of variable σ^2 with degree of freedom τ and scale parameter ω. The notation for a distribution and the notation for the probability density of the distribution are now consistent. For example, $x \sim N(\mu, \sigma^2)$ means that x is normally distributed with mean μ and variance σ^2, which is equivalently described as $p(x) = N(x|\mu, \sigma^2)$. The exact forms of these distributions are

$$p(\beta) = N(\beta|\mu_\beta, \sigma_\beta^2) = \frac{1}{\sqrt{2\pi\sigma_\beta^2}}\exp\left[-\frac{1}{2\sigma_\beta^2}(\beta - \mu_\beta)^2\right] \tag{15.9}$$

and

$$p(\sigma^2) = \text{Inv} - \chi^2(\sigma^2|\tau,\omega) = \frac{(\tau\omega/2)^{\tau/2}}{\Gamma(\tau/2)}(\sigma^2)^{-(\tau/2+1)}\exp\left(-\frac{\tau\omega}{2\sigma^2}\right) \quad (15.10)$$

where $\Gamma(\tau/2)$ is the gamma function with argument $\tau/2$. Conditional on the parameter θ, the data vector y has a normal distribution with probability density

$$p(y|\theta) = \prod_{j=1}^{n} N(y_j|\mu,\sigma^2) \propto \frac{1}{(\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - X_j\beta)^2\right] \quad (15.11)$$

We now have the probability density of the data and the density of the prior distribution of the parameters. We treat both the data and the parameters as random variables and formulate the joint distribution of the data and the parameters,

$$p(y,\theta) = p(y|\theta)p(\theta) \quad (15.12)$$

where $p(\theta) = p(\beta)p(\sigma^2)$. The purpose of Bayesian analysis is to infer the conditional distribution of the parameters given the data and draw conclusion about the parameters from the conditional distribution. The conditional distribution of the parameters has the form of

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} \propto p(y,\theta) \quad (15.13)$$

which is also called the posterior distribution of the parameters. The denominator, $p(y)$, is the marginal density of the data, which is irrelevant to the parameters and can be ignored because we are only interested in the estimation of parameters. Note that the above conditional density is rewritten as

$$p(\beta,\sigma^2|y) = \frac{p(y,\beta,\sigma^2)}{p(y)} \propto p(y,\beta,\sigma^2) \quad (15.14)$$

which is still a joint posterior density with regard to the two components of the parameter vector. The ultimate purpose of the Bayesian analysis is to infer the marginal posterior distribution for each component of the parameter vector. The marginal posterior density for β is obtained by integrating the joint posterior distribution over σ^2,

$$p(\beta|y) = \int_0^\infty p(\beta,\sigma^2|y)\mathrm{d}\sigma^2 \quad (15.15)$$

The integration has an explicit form, which turns out to be the kernel of a t-distribution with $n + \tau - 1$ degrees of freedom (Sorensen and Gianola 2002). The β itself is not a t-distributed variable. It is $(\beta - \tilde{\beta})/\sigma_{\tilde{\beta}}$ that has a t-distribution, where

$$\mathrm{E}(\beta|y) = \tilde{\beta} = \left(\frac{1}{\sigma^2_{\hat{\beta}}} + \frac{1}{\sigma^2_{\beta}}\right)^{-1} \left(\frac{\hat{\beta}}{\sigma^2_{\hat{\beta}}} + \frac{\mu_{\beta}}{\sigma^2_{\beta}}\right) \tag{15.16}$$

is the marginal posterior mean of β and

$$\mathrm{var}(\beta|y) = \sigma^2_{\tilde{\beta}} = \left(\frac{1}{\sigma^2_{\hat{\beta}}} + \frac{1}{\sigma^2_{\beta}}\right)^{-1} \tag{15.17}$$

is the marginal posterior variance of β. Both the mean and the variance contain $\hat{\beta}$ and $\hat{\sigma}^2$, the MLEs of β and σ^2, respectively. The role that $\hat{\sigma}^2$ plays in the above equations is through

$$\sigma^2_{\tilde{\beta}} = \left(\sum\nolimits_{j=1}^{n} X_j^2\right)^{-1} \hat{\sigma}^2 \tag{15.18}$$

The density of the t-distributed variable with mean $\tilde{\beta}$ and variance $\sigma^2_{\tilde{\beta}}$ is denoted by

$$p(\beta|y) = t_{n+\tau-1}(\beta|\tilde{\beta}, \sigma^2_{\tilde{\beta}}) \tag{15.19}$$

The marginal posterior density for σ^2 is obtained by integrating the joint posterior over β,

$$p(\sigma^2|y) = \int_{-\infty}^{\infty} p(\beta, \sigma^2|y)\mathrm{d}\beta \tag{15.20}$$

which happens to be a scaled inverse chi-square distribution with

$$\tau^* = n + \tau - 1 \tag{15.21}$$

degrees of freedom and a scale parameter (Sorensen and Gianola 2002)

$$\omega^* = \frac{\tau\omega + \sum_{j=1}^{n}(y_j - X_j\tilde{\beta})^2}{\tau + n - 1} \tag{15.22}$$

The density of the new scaled inverse chi-square variable is denoted by

$$p(\sigma^2|y) = \mathrm{Inv} - \chi^2(\sigma^2|\tau^*, \omega^*) \tag{15.23}$$

The mean and variance of the above distribution are

$$\mathrm{E}(\sigma^2|y) = \tilde{\sigma}^2 = \frac{\tau\omega + \sum_{j=1}^{n}(y_j - X_j\tilde{\beta})^2}{\tau + n - 3} \tag{15.24}$$

and

$$\text{var}(\sigma^2|y) = \frac{2[\tau\omega + \sum_{j=1}^{n}(y_j - X_j\tilde{\beta})^2]^2}{(\tau + n - 3)^2(\tau + n - 5)} \tag{15.25}$$

respectively (Sorensen and Gianola 2002).

The marginal posterior distribution of each parameter contains all the information we have gathered for that parameter. The Bayesian estimate of that parameter can be either the posterior mean, the posterior mode, or the posterior median, depending on the preference of the investigator. The marginal posterior distribution of a parameter itself can also be treated as an estimate of the parameter. Assume that the marginal posterior mean of a parameter is considered as the Bayesian estimate of that parameter. The Bayesian estimates of β and σ^2 are $\tilde{\beta}$ and $\tilde{\sigma}^2$, respectively.

The simple regression analysis (regression through origin) discussed above is the simplest case of Bayesian analysis where the marginal posterior distribution of each parameter is known. In most situations, especially when the dimensionality of the parameter θ is high, the marginal posterior distribution of a single parameter involves high-dimensional multiple integration, and often the integration does not have an explicit expression. Therefore, the posterior distribution of a parameter often has an unknown form, which makes the Bayesian inference difficult. Thanks to the ever-growing computing power, we can perform multiple numerical integrations very efficiently. We can even utilize Monte Carlo integration by repeatedly simulating multivariate random variables. For extremely high-dimensional problems, Monte Carlo integration is perhaps the only way to implement the Bayesian method.

Let us now discuss the relationship between the joint distribution and the marginal distribution. Let $\theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$ be an m dimensional multiple variables. Let $p(\theta) = p(\theta_1, \ldots, \theta_m|y)$ be the joint posterior distribution. The marginal posterior distribution for the kth component is

$$p(\theta_k|y) = \int \ldots \int p(\theta_1, \ldots, \theta_m|y)\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{k-1}\mathrm{d}\theta_{k+1} \ldots \mathrm{d}\theta_m \tag{15.26}$$

If the multiple integration has an explicit form and we can recognize the marginal distribution of θ_k, i.e., $p(\theta_k|y)$ is the density of a well-known distribution, then the expectation (or mode) of this distribution is what we want to know in the Bayesian analysis. Suppose that we know neither the joint posterior distribution nor the marginal posterior distribution, but somehow we have a joint posterior sample of multivariate θ with size N. In other words, we are only given N joint observations of θ. The sample is denoted by $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}\}$. We can imagine that the data in the sample are arranged in a $N \times m$ matrix. Each row represents an observation, while each column represents a variable. What is the estimated marginal expectation of θ_k drawn from this sample? Remember that this sample is supposed to be generated from the joint posterior distribution. The answer is simple; we only need to calculate the algebraic mean of variable θ_k from this sample, i.e.,

$$\bar{\theta}_k = \frac{1}{N}\sum_{j=1}^{N}\theta_k^{(j)} \tag{15.27}$$

This average value of θ_k is an empirical marginal posterior mean of θ_k, i.e., a Bayesian estimate of θ_k. We can see that as long as we have a joint sample of θ, we can infer the marginal mean of a single component of θ simply by calculating the mean of that component from the sample. While calculating the mean only requires knowledge learned from elementary school, generating the joint sample of θ becomes the main focus of the Bayesian analysis.

15.2 Markov Chain Monte Carlo

There are many different ways to generate a sample of θ from the joint distribution. The classical method is to use the following sequential approach to generate the first observation, denoted by $\theta^{(1)}$:

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

The process is simply repeated N times to simulate an entire sample of θ. Observations generated this way are independent. We can see that we still need the marginal distribution for θ_1 and various levels of marginality of other components. Only θ_m is generated from a fully conditional posterior, which does not involve any integration. Therefore, this sequential approach of generating random sample is not what we want.

The MCMC approach draws all variables from their fully conditional posterior distributions. To draw a variable from a conditional distribution, we must have some values of the variables that are conditioned on. For example, to draw y from $p(y|x)$, the value of x must be known. Let $\theta^{(0)}$ be the initial value of multivariate θ. The first observation of θ is drawn using the following process:

- Simulate $\theta_1^{(1)}$ from $p(\theta_1|\theta_{-1}^{(0)}, y)$
- Simulate $\theta_2^{(1)}$ from $p(\theta_2|\theta_{-2}^{(0)}, y)$
- Simulate $\theta_3^{(1)}$ from $p(\theta_3|\theta_{-3}^{(0)}, y)$
- .....
- Simulate $\theta_m^{(1)}$ from $p(\theta_m|\theta_{-m}^{(0)}, y)$

where $\theta_{-k}^{(0)}$ is a subset of vector $\theta^{(0)}$ that excludes the kth element, i.e.,

$$\theta_{-k}^{(0)} = \{\theta_1^{(0)}, \ldots, \theta_{k-1}^{(0)}, \theta_{k+1}^{(0)}, \ldots, \theta_m^{(0)}\}$$

This special notation (negative subscript) has tremendously simplified the expressions of the MCMC sampling algorithm. The above process concludes the

simulation for the first observation. The process is repeated N times to generate a sample of θ with size N. The sampled $\theta^{(t)}$ depends on $\theta^{(t-1)}$, i.e., the sampled θ in the current cycle only depends on the θ in the previous cycle. Therefore, the sequence

$$\{\theta^{(0)} \to \theta^{(1)} \to \cdots \to \theta^{(N)}\}$$

forms a Markov chain, which explains why the method is called Markov chain Monte Carlo. Because of the Markov chain property, the observations are not independent, and the first few hundred (or even thousand) observations highly depend on the initial value $\theta^{(0)}$ used to start the chain. Once the chain is stabilized, i.e., the sampled θ does not depend on the initial value, we say that the chain has reached its stationary distribution. The period from the beginning to the time when the stationary distribution is reached is called the burn-in period. Observations in the burn-in period should be deleted. After the burn-in period, the observations are presumably sampled from the joint distribution. The observations may still be correlated; such a correlation is called serial correlation or autocorrelation. We can save one observation in every sth cycle to remove the serial correlation, where $s = 20$ or $s = 50$ or any other integers, depending on the particular problem. This process is called trimming or thinning the Markov chain. After burn-in deleting and chain trimming, we collect N^* observations from the total of N observations simulated. The sample of θ with N^* observations is the posterior sample (sampled from the $p(\theta|y)$ distribution). Any Bayesian statistics can be inferred empirically from this posterior sample.

Recall that the marginal posterior for β is a t-distribution and the marginal posterior for σ^2 is a scaled inverse chi-square distribution. Both distributions have complicated forms of expression. The MCMC sampling process only requires the conditional posterior distribution, not the marginal posterior. Let us now look at the conditional posterior distribution of each parameter of the simple regression analysis.

As previously shown, the MLE of β is

$$\hat{\beta} = \left(\sum_{j=1}^{n} X_j^2\right)^{-1} \left(\sum_{j=1}^{n} X_j y_j\right) \tag{15.28}$$

and the variance of the estimate is

$$\sigma_{\hat{\beta}}^2 = \left(\sum_{j=1}^{n} X_j^2\right)^{-1} \sigma^2 \tag{15.29}$$

Note that $\sigma_{\hat{\beta}}^2$ differs from that defined in (15.18) in that σ^2 is used here in place of $\hat{\sigma}^2$. So, just from the data without any prior information, we can infer β. The estimated β itself is a variable, which follows a normal distribution denoted by

$$\beta \sim N_1(\hat{\beta}, \sigma_{\hat{\beta}}^2) \tag{15.30}$$

The subscript 1 means that this is an estimate drawn from the first source of information. Before we observed the data, the prior information about β is considered the second source of information, which is denoted by

$$\beta \sim N_2(\mu_\beta, \sigma^2_{\tilde{\beta}}) \tag{15.31}$$

The posterior distribution of β is obtained by combining the two sources of information (Box and Tiao 1973), which remains normal and is denoted by

$$\beta \sim N(\bar{\beta}, \sigma^2_{\bar{\beta}}) \tag{15.32}$$

where

$$\bar{\beta} = \left(\frac{1}{\sigma^2_{\hat{\beta}}} + \frac{1}{\sigma^2_{\tilde{\beta}}}\right)^{-1} \left(\frac{\hat{\beta}}{\sigma^2_{\hat{\beta}}} + \frac{\mu_\beta}{\sigma^2_{\tilde{\beta}}}\right) \tag{15.33}$$

and

$$\sigma^2_{\bar{\beta}} = \left(\frac{1}{\sigma^2_{\hat{\beta}}} + \frac{1}{\sigma^2_{\tilde{\beta}}}\right)^{-1} \tag{15.34}$$

We now have the conditional posterior distribution for β denoted by

$$p(\beta|\sigma^2, y) = N(\beta|\bar{\beta}, \sigma^2_{\bar{\beta}}) \tag{15.35}$$

from which a random β is sampled.

Given β, we now evaluate the conditional posterior distribution of σ^2. The prior for σ^2 is a scaled inverse chi-square distribution with τ degrees of freedom and a scale parameter ω, denoted by

$$p(\sigma^2) = \text{Inv} - \chi^2(\sigma^2|\tau, \omega) \tag{15.36}$$

The posterior distribution remains a scaled inverse chi-square with a modified degree of freedom and a modified scale parameter, denoted by

$$p(\sigma^2|\beta, y) = \text{Inv} - \chi^2(\sigma^2|\tau^*, \omega^*) \tag{15.37}$$

where

$$\tau^* = \tau + n \tag{15.38}$$

and

$$\omega^* = \frac{\tau\omega + \sum_{j=1}^{n} (y_j - X_j\beta)^2}{\tau + n} \tag{15.39}$$

Note that ω^* defined here differs from that defined in (15.22) in that β is used here while $\tilde{\beta}$ is used in (15.22). The conditional posterior of β is normal, which belongs to the same distribution family as the prior distribution. Similarly, the

conditional posterior of σ^2 remains a scaled inverse chi-square, also the same type of distribution as the prior. These priors are called conjugate priors because they lead to the conditional posterior distributions of the same type.

The MCMC sampling process is summarized as:

1. Initialize $\beta = \beta^{(0)}$ and $\sigma^2 = \sigma^{2(0)}$
2. Simulate $\beta^{(1)}$ from $N(\beta|\bar{\beta}, \sigma^2_{\bar{\beta}})$
3. Simulate $\sigma^{2(1)}$ from $\text{Inv} - \chi^2(\sigma^2|\tau^*, \omega^*)$
4. Repeat Steps (2) and (3) until N observations of the posterior sample are collected.

It can be seen that the MCMC sampling-based regression analysis only involves two distributions, a normal distribution and a scaled inverse chi-square distribution. Most software packages have built-in functions to generate random variables from some simple distributions, e.g., $N(0, 1)$ and $\chi^2(\tau)$. Let $Z \sim N(0, 1)$ be a realized value drawn from the standardized normal distribution and $X \sim \chi^2(\tau^*)$ be a realized value drawn from a chi-square distribution with τ^* degrees of freedom. To sample β from $N(\bar{\beta}, \sigma^2_{\bar{\beta}})$, we sample Z first and then take

$$\beta = \sigma_{\bar{\beta}} Z + \bar{\beta} \tag{15.40}$$

To sample σ^2 from $\text{Inv} - \chi^2(\tau^*, \omega^*)$, we first sample X and then take

$$\sigma^2 = \frac{\tau^* \omega^*}{X} \tag{15.41}$$

In summary, the MCMC process requires sampling a parameter only from the fully conditional posterior distribution, which usually has a simple form, e.g., normal or chi-square, and it draws one variable at a time. This type of MCMC sampling is also called Gibbs sampling (Geman and Geman 1984). With the MCMC procedure, we turn ourselves into experimentalists. Like plant breeders who plant seeds, let the seeds grow into plants, and measure the average plant yield, we plant the seeds of parameters in silico, let the parameters "grow," and measure the average of each parameter. The Bayesian posterior mean of a parameter simply takes the algebraic mean of a parameter in the posterior sample collected from the in silico experiment. Once the Bayesian method is implemented via the MCMC algorithm, it is no longer owned by a few "Bayesians"; rather, it has become a popular tool that can be used by people in all areas, including engineers, biologists, plant and animal breeders, social scientists, and so on.

Before we move on to the next section, let us demonstrate the MCMC sampling process using the simple regression as an example. The values of x and y for 20 observations are given in Table 15.1.

The model is

$$y_j = X_j \beta + \epsilon_j, \quad \forall j = 1, \ldots, 20$$

Table 15.1 Data used in the text to demonstrate the MCMC sampling process

x	y	x	y
1	2.95	−1	−1.23
1	0.61	1	1.06
1	4.61	1	0.41
1	3.46	−1	−3.09
1	1.12	−1	−2.08
1	4.15	−1	−1.55
−1	−2.46	1	1.07
1	4.49	−1	−5.39
1	3.34	−1	−1.26
−1	−1.44	−1	−4.46

The sample size is $n = 20$. Before introducing the prior distributions, we provide the MLEs of the parameters, which are

$$\hat{\beta} = \left(\sum\nolimits_{j=1}^{n} X_j^2\right)^{-1} \sum\nolimits_{j=1}^{n} X_j y_j = 2.5115$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - X_j\hat{\beta})^2 = 2.3590$$

The variance of $\hat{\beta}$ is

$$\sigma_{\hat{\beta}}^2 = \left(\sum\nolimits_{j=1}^{n} X_j^2\right)^{-1} \hat{\sigma}^2 = 0.1180$$

Let us choose the following prior distributions:

$$p(\beta) = N(\beta|\mu_\beta, \sigma_\beta^2) = N(\beta|0.1, 1.0)$$

and

$$p(\sigma^2) = \text{Inv} - \chi^2(\sigma^2|\tau, \omega) = \text{Inv} - \chi^2(\sigma^2|3, 3.5)$$

The marginal posterior mean and posterior variance of β are

$$\text{E}(\beta|y) = \tilde{\beta} = \left(\frac{1}{\sigma_{\hat{\beta}}^2} + \frac{1}{\sigma_\beta^2}\right)^{-1} \left(\frac{\hat{\beta}}{\sigma_{\hat{\beta}}^2} + \frac{\mu_\beta}{\sigma_\beta^2}\right) = 2.2571$$

and

$$\text{var}(\beta|y) = \sigma_{\tilde{\beta}}^2 = \left(\frac{1}{\sigma_{\hat{\beta}}^2} + \frac{1}{\sigma_\beta^2}\right)^{-1} = 0.1055$$

respectively. The marginal poster mean and posterior variance of σ^2 are

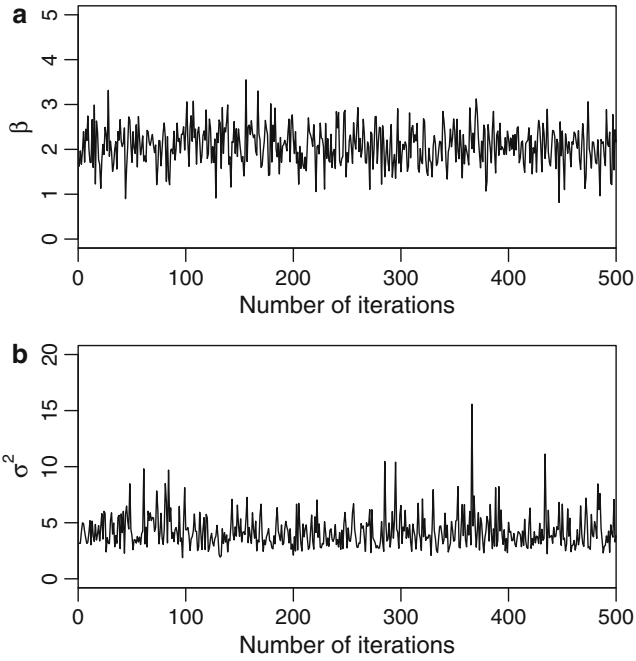

Fig. 15.1 Changes of the sampled parameters over the number of iterations since the MCMC starts. The *top panel* is the change for β, and the *bottom panel* is that for σ^2

$$\mathrm{E}(\sigma^2|y) = \tilde{\sigma}^2 = \frac{\tau\omega + \sum_{j=1}^{n}(y_j - X_j\tilde{\beta})^2}{\tau + n - 3} = 2.8308$$

and

$$\mathrm{var}(\sigma^2|y) = \frac{2[\tau\omega + \sum_{j=1}^{n}(y_j - X_j\tilde{\beta})^2]^2}{(\tau + n - 3)^2(\tau + n - 5)} = 0.8904$$

respectively.

We now use the MCMC sampling approach to generating the joint posterior sample for β and σ^2 and calculate the empirical marginal posterior means and posterior variances for the two parameters. For a problem as simple as this, the burn-in period can be very short or even without burn-in. Figure 15.1 shows the first 500 cycles of MCMC sampler (including the burn-in period) for the two parameters, β and σ^2. The chains converge immediately to the stationary distribution. To be absolutely sure that we actually collect samples from the stationary distribution, we set the burn-in period to 1,000 iterations (very safe), and the chain was subsequently trimmed to save one observation in every 50 iterations after the burn-in. The posterior sample size was 10,000. The total number of MCMC cycles was

Table 15.2 Empirical marginal posterior means and posterior variances for the two parameters, β and σ^2

Parameter	Posterior mean	Posterior variance
β	2.2171	0.1320
σ^2	2.8489	0.9497

$1,000 + 50 \times 10,000 = 5,01,000$. The empirical marginal posterior means and marginal posterior variances for β and σ^2 are given in Table 15.2, which are very close to the theoretical values given before.

15.3 Mapping Multiple QTL

Although interval mapping (under the single QTL model) can detect multiple QTL by evaluating the number of peaks in the test statistic profile, it cannot provide accurate estimates of QTL effects. The best way to handle multiple QTL is to use a multiple QTL model. Such a model requires knowledge of the number of QTL. Most QTL mappers consider that the number of QTL is an important parameter and should be estimated in QTL mapping experiments. Therefore, model selection is often conducted to determine the number of QTL (Broman and Speed 2002). Under the Bayesian framework, model selection is implemented through the reversible jump MCMC algorithm (Sillanpää and Arjas 1998). Xu (2003) and Wang et al. (2005b) had a quite different opinion, in which the number of QTL is not considered as an important parameter. According to Wang et al. (2005b), we can propose a model that includes as many QTL as the model can handle. Such a model is called an oversaturated model. Some of the proposed QTL may be real, but most of them are spurious. As long as we can force the spurious QTL to have zero or close to zero estimated effects, the oversaturated model is considered satisfactory. The selective shrinkage Bayesian method can generate the result of QTL mapping exactly the same as we expect, that is, spurious QTL effects are shrunken to zero while true QTL have effects subject to no shrinkage.

15.3.1 Multiple QTL Model

The multiple QTL model can be described as

$$y_j = \sum_{i=1}^{q} X_{ji}\beta_i + \sum_{k=1}^{p} Z_{jk}\gamma_k + \epsilon_j \tag{15.42}$$

where y_j is the phenotypic value of a trait for individual j for $j = 1, \ldots, n$ and n is the sample size. The non-QTL effects are included in vector $\beta = \{\beta_1, \ldots, \beta_q\}$ with matrix $X_j = \{X_{j1}, \ldots, X_{jq}\}$ being the design matrix to connect β and y_j. The effect of the kth QTL is denoted by γ_k for $k = 1, \ldots, p$ where p is the

proposed number of QTL in the model. Vector $Z_j = \{Z_{j1}, \ldots, Z_{jp}\}$ is determined by the genotypes of the proposed QTL in the model. The residual error ϵ_j is assumed to be i.i.d. $N(0, \sigma^2)$. Let us use a BC population as an example. For the kth QTL, $Z_{jk} = 1$ for one genotype and $Z_{jk} = -1$ for the alternative genotype. Extension to F_2 population and adding the dominance effects are straightforward (only requires adding more QTL effects and increasing the model dimension). The proposed number of QTL is p, which must be larger than the true number of QTL to make sure that large QTL will not be missed. The optimal strategy is to put one QTL in every d cM of the genome, where d can be any value between 5 and 50. If $d < 5$, the model will be ill conditioned due to multicollinearity. If $d > 50$, some genome regions may not be visited by the proposed QTL even if there are true QTL located in those regions. Of course, a larger sample size is required to handle a larger model (more QTL).

15.3.2 Prior, Likelihood, and Posterior

The data involved in QTL mapping include the phenotypic values of the trait and marker genotypes for all individuals in the mapping population. Unlike Wang et al. (2005b) who expressed marker genotypes explicitly as data in the likelihood, here we suppress the marker genotypes from the data to simplify the notation. The linkage map of markers and the marker genotypes only affect the way to calculate QTL genotypes. We first use the multipoint method to calculate the genotype probabilities for all putative loci of the genome. These probabilities are then treated as the prior probabilities of QTL genotypes, from which the posterior probabilities are calculated by incorporating the phenotype and the current parameter values. Therefore, the data used to construct the likelihood are represented by $y = \{y_j, \ldots, y_n\}$. The vector of parameters is denoted by θ, which consists of the positions of the proposed QTL denoted by $\lambda = \{\lambda_1, \ldots, \lambda_p\}$, the effects of the QTL denoted by $\gamma = \{\gamma_1, \ldots, \gamma_p\}$, the non-QTL effects denoted by $\beta = \{\beta_1, \ldots, \beta_q\}$, and the residual error variance σ^2. Therefore, $\theta = \{\lambda, \beta, \gamma, \psi, \sigma^2\}$, where $\psi = \{\sigma_1^2, \ldots, \sigma_p^2\}$, will be defined later. The QTL genotypes $Z_j = \{Z_{j1}, \ldots, Z_{jp}\}$ are not parameters but missing values. The missing genotypes can be redundantly expressed as $\delta_j = \{\delta_{j1}, \ldots, \delta_{jp}\}$ where

$$\delta_{jk} = \delta(G_{jk}, \kappa)$$

is the δ function. If $G_{jk} = \kappa$, then $\delta(G_{jk}, \kappa) = 1$, else $\delta(G_{jk}, \kappa) = 0$, where G_{jk} is the genotype of the kth QTL for individual j and $\kappa = 1, 2, 3$ for an F_2 population (three genotypes per locus). The probability density of δ is

$$p(\delta_j|\lambda) = \prod_{k=1}^{p} p(\delta_{jk}|\lambda_k) \tag{15.43}$$

The independence of the QTL genotype across loci is due to the fact that they are the conditional probabilities given marker information. So, the marker information has entered here to infer the QTL genotypes. The prior for the β is

$$p(\beta) = \prod_{i=1}^{q} p(\beta_i) = \text{constant} \tag{15.44}$$

This is a uniform prior or, more appropriately, uninformative prior. The reason for choosing uninformative prior for β is that the dimensionality of β is usually very low so that β can be precisely estimated from the data alone without resorting to any prior knowledge. The prior for the QTL effects is

$$p(\gamma|\psi) = \prod_{k=1}^{p} p(\gamma_k|\sigma_k^2) = \prod_{k=1}^{p} N(\gamma_k|0, \sigma_k^2) \tag{15.45}$$

where σ_k^2 is the variance of the prior distribution for the kth QTL effect. Collectively, these variances are denoted by $\psi = \{\sigma_1^2, \ldots, \sigma_p^2\}$. This is a highly informative prior because of the zero expectation of the prior distribution. The variance of the prior distribution determines the relative weights of the prior information and the data. If σ_k^2 is very small, the prior will dominate the data, and thus, the estimated γ_k will be shrunken toward the prior expectation, that is, zero. If σ_k^2 is large, the data will dominate the prior so that the estimated γ_k will be largely unaltered (subject to no shrinkage). The key difference between this prior and the prior commonly used in Bayesian regression analysis is that different regression coefficient has a different prior variance and thus different level of shrinkage. Therefore, this method is also called the selective shrinkage method (Wang et al. 2005b). The classical Bayesian regression method, however, often uses a common prior for all regression coefficients, i.e., $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2 = \sigma_\gamma^2$, which is also called ridge regression (Hoerl and Kennard 1970). The problem with this selective shrinkage method is that there are too many prior variances and it is hard to choose the appropriate values for the variances. There are two approaches to choosing the prior variances, empirical Bayesian (Xu 2007) and hierarchical modeling (Gelman 2006). The empirical Bayesian approach attempts to estimate the prior variances under the mixed model methodology by treating each regression coefficient as a random effect. The hierarchical modeling treats the prior variances as parameters and assigns a higher level prior to each variance component. By treating the variances as parameters, rather than as hyperparameters, we can estimate the variances along with the regression coefficients. Here, we take the hierarchical model approach and assign each σ_k^2 a prior distribution. The empirical Bayesian method will be discussed in the next chapter. The scaled inverse chi-square distribution is chosen for each variance component,

$$p(\sigma_k^2) = \text{Inv} - \chi^2(\sigma_k^2|\tau, \omega), \quad \forall k = 1, \ldots, p \tag{15.46}$$

The degree of freedom τ and the scale parameter ω are hyperparameters, and their influence on the estimated regression coefficients is much weaker because the influence is through the σ_k^2's. It is now easy to choose τ and ω. The degree of freedom τ is also called the prior belief. Although the proper prior should have $\tau > 0$ and $\omega > 0$, our past experience showed that an improper prior works better than the proper prior. Therefore, we choose $\tau = \omega = 0$, which leads to

$$p(\sigma_k^2) \propto \frac{1}{\sigma_k^2}, \quad \forall k = 1, \ldots, p \tag{15.47}$$

The joint prior for all the σ_k^2 is

$$p(\psi) = \prod_{k=1}^{p} p(\sigma_k^2) \tag{15.48}$$

The residual error variance is also assigned to the improper prior,

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{15.49}$$

The positions of the QTL depend on the number of QTL proposed, the number of chromosomes, and the size of each chromosome. Based on the average coverage per QTL (e.g., 30 cM per QTL), the number of QTL allocated to each chromosome can be calculated. Let p_c be the number of QTL proposed for the cth chromosome. These p_c QTL should be placed evenly along the chromosome. We can let the positions fixed throughout all the MCMC process so that the positions are simply constants (not parameters of interest). In this case, more QTL should be proposed to make sure that the genome is well covered by the proposed QTL. The alternative and also more efficient approach is to allow QTL position to move along the genome during the MCMC process. There is a restriction for the moving range of each QTL. The positions are disjoint along the chromosome. The first QTL must move between the first marker and the second QTL. The last QTL must move between the last marker and the second last QTL. All other QTL must move between the QTL in the left and the QTL in the right of the current QTL, i.e., the QTL that flank the current QTL. Based on this search strategy, the joint prior probability is

$$p(\lambda) = p(\lambda_1)p(\lambda_2|\lambda_1)\ldots p(\lambda_{p_c}|\lambda_{p_c-1}) \tag{15.50}$$

Given the positions of all other QTL, the conditional probability of the position of QTL k is

$$p(\lambda_k) = \frac{1}{\lambda_{k+1} - \lambda_{k-1}} \tag{15.51}$$

If QTL k is located at either end of a chromosome, the above prior needs to be modified by replacing either λ_{k-1} or λ_{k+1} by the position of the nearest end marker.

We now have a situation where the prior probability of one variable depends on values of other variables. This type of prior is called adaptive prior.

Since marker information has been used to calculate the prior probabilities of QTL genotypes, they are no longer expressed as data. The only data appearing explicitly in the model are the phenotypic values of the trait. Conditional on all parameters and the missing values, the probability density of y_j is normal. Therefore, the joint probability density of all the y_j's (called the likelihood) is

$$\begin{aligned} p(y|\theta,\delta) &= \prod_{j=1}^{n} p(y_j|\theta,\delta_j) \\ &= \prod_{j=1}^{n} N\left(y_j \left| \sum_{i=1}^{q} X_{ji}\beta_i + \sum_{k=1}^{p} Z_{jk}\gamma_k, \sigma^2\right.\right) \end{aligned} \tag{15.52}$$

The fully conditional posterior of each variable is defined as

$$p(\theta_i|\theta_{-i},\delta,y) \propto p(\theta_i,\theta_{-i},\delta,y) \tag{15.53}$$

where θ_i is a single element of the parameter vector θ and θ_{-i} is the collection of the remaining elements. The symbol $\propto$ means that a constant factor (not a function of parameter θ_i) has been ignored. The joint probability density $p(\theta_i,\theta_{-i},\delta,y) = p(\theta,\delta,y)$ is expressed as

$$\begin{aligned} p(\theta,\delta,y) &\propto p(y|\theta,\delta)p(\delta|\theta)p(\theta) \\ &= p(y|\theta,\delta)p(\beta|\psi)p(\psi)p(\delta|\lambda)p(\lambda)p(\sigma^2) \end{aligned} \tag{15.54}$$

The fully conditional posterior probability density for each variable is simply derived by treating all other variables as constants and comparing the kernel of the density with a standard distribution. After some algebraic manipulation, we obtain the fully conditional distribution for most of the unknown variables (including parameters and missing values).

The fully conditional posterior for the non-QTL effect is

$$p(\beta_i|\dots) = N(\beta_i|\hat{\beta}_i,\sigma^2_{\hat{\beta}_i}) \tag{15.55}$$

The special notation $p(\beta_i|\dots)$ is used to express the fully conditional probability density. The three dots $(\dots)$ after the vertical bar mean everything else except the variable of interest. The posterior mean and posterior variance are calculated using (15.58) and (15.59) given below:

$$\hat{\beta}_i = \left(\sum_{j=1}^{n} X_{ji}^2\right)^{-1} \sum_{j=1}^{n} X_{ji}\left(y_j - \sum_{i'\neq i}^{q} X_{ji'}\beta_{i'} - \sum_{k=1}^{p} Z_{jk}\gamma_k\right) \tag{15.56}$$

and

$$\sigma^2_{\hat{\beta}_i} = \left(\sum_{j=1}^{n} X_{ji}^2 \right)^{-1} \sigma^2 \tag{15.57}$$

The fully conditional posterior for the kth QTL effect is

$$p(\gamma_k | \dots) = N(\gamma_k | \hat{\gamma}_k, \sigma^2_{\hat{\gamma}_k}) \tag{15.58}$$

where

$$\hat{\gamma}_k = \left(\sum_{j=1}^{n} Z_{jk}^2 + \frac{\sigma^2}{\sigma_k^2} \right)^{-1} \sum_{j=1}^{n} Z_{ji} \left(y_j - \sum_{i=1}^{q} X_{ji} \beta_i - \sum_{k' \neq k}^{p} Z_{jk'} \gamma_{k'} \right) \tag{15.59}$$

and

$$\sigma^2_{\hat{\gamma}_k} = \left(\sum_{j=1}^{n} Z_{jk}^2 + \frac{\sigma^2}{\sigma_k^2} \right)^{-1} \sigma^2 \tag{15.60}$$

Comparing the conditional posterior distributions of β_i and γ_k, we notice the difference between a normal prior and a uniform prior with respect to the effects on the posterior distributions. When a normal prior is used, a shrinkage factor, $\frac{\sigma^2}{\sigma_k^2}$, is added to $\sum_{j=1}^{n} Z_{jk}^2$. If σ_k^2 is very large, the shrinkage factor disappears, meaning no shrinkage. On the other hand, if σ_k^2 is small, the shrinkage factor will dominate over $\sum_{j=1}^{n} Z_{jk}^2$, and in the end, the denominator will become infinitely large, leading to zero expectation and zero variance for the conditional posterior distribution γ_k. As such, the estimated γ_k is completely shrunken to zero. The conditional posterior distribution for each of the variance component σ_k^2 is a scaled inverse chi-square variable with probability density

$$p(\sigma_k^2 | \dots) = \text{Inv} - \chi^2 \left(\sigma_k^2 \middle| \tau + 1, \frac{\tau\omega + \gamma_k^2}{\tau + 1} \right) \tag{15.61}$$

where $\tau = \omega = 0$. The conditional posterior density for the residual error variance is

$$p(\sigma^2 | \dots) = \text{Inv} - \chi^2 \left(\sigma^2 \middle| \tau + n, \frac{\tau\omega + nS_e^2}{\tau + n} \right) \tag{15.62}$$

where

$$S_e^2 = \frac{1}{n} \sum_{j=1}^{n} \left(y_j - \sum_{i=1}^{q} X_{ji} \beta_i + \sum_{k=1}^{p} Z_{jk} \gamma_k \right)^2 \tag{15.63}$$

The next step is to sample QTL genotypes, which determine the values of the Z_j variables. Let us again use a BC population as an example and consider sampling the kth QTL genotype given that every other variable is known. There are two sources

of information available to infer the probability for each of the two genotypes of the QTL. One information comes from the markers denoted by $p_j(+1)$ and $p_j(-1)$, respectively, for the two genotypes, where $p_j(+1) + p_j(-1) = 1$. These two probabilities are calculated from the multipoint method (Jiang and Zeng 1997). The other source of information comes from the phenotypic value. The connection between the phenotypic value and the QTL genotype is through the probability density of y_j given the QTL genotype. For the two alternative genotypes of the QTL , i.e., $Z_{jk} = 1$ and $Z_{jk} = -1$, the two probability densities are

$$\begin{aligned} p(y_j|Z_{jk} = +1) &= N\left(y_j \left| \sum_{i=1}^{q} X_{ji}\beta_i + \sum_{k' \neq k}^{p} Z_{jk'}\gamma_{k'} + \gamma_k, \sigma^2 \right.\right) \\ p(y_j|Z_{jk} = -1) &= N\left(y_j \left| \sum_{i=1}^{q} X_{ji}\beta_i + \sum_{k' \neq k}^{p} Z_{jk'}\gamma_{k'} - \gamma_k, \sigma^2 \right.\right) \end{aligned} \quad (15.64)$$

Therefore, the conditional posterior probabilities for the two genotypes of the QTL are

$$\begin{aligned} p_j^*(+1) &= \frac{p_j(+1)p(y_j|Z_{jk} = +1)}{p_j(+1)p(y_j|Z_{jk} = +1) + p_j(-1)p(y_j|Z_{jk} = -1)} \\ p_j^*(-1) &= \frac{p_j(-1)p(y_j|Z_{jk} = -1)}{p_j(+1)p(y_j|Z_{jk} = +1) + p_j(-1)p(y_j|Z_{jk} = -1)} \end{aligned} \quad (15.65)$$

where $p_j^*(+1) = p(Z_{jk} = +1|\ldots)$ and $p_j^*(-1) = p(Z_{jk} = -1|\ldots)$ are the posterior probabilities of the two genotypes. The genotype of the QTL is $Z_{jk} = 2u - 1$, where u is sampled from a Bernoulli distribution with probability $p_j^*(+1)$. So far we have completed the sampling process for all variables except the QTL positions. If we place a large number of QTL evenly distributed along the genome, say one QTL in every 10 cM, we can let the positions fixed (not moving) across the entire MCMC process. Although this fixed-position approach does not generate accurate result, it does provide some general information about the ranges where the QTL are located. Suppose that the trait of interest is controlled by only 5 QTL and we place 100 QTL evenly distributed on the genome, then majority of the assumed QTL are spurious. The Bayesian shrinkage method allows the spurious QTL to be shrunken to zero. This is why the Bayesian shrinkage method does not need variable selection. A QTL with close to zero estimated effect is equivalent to being excluded from the model. When the assumed QTL positions are fixed, investigators actually prefer to put the QTL at marker positions because marker positions contain the maximum information. This multiple-marker analysis is recommended before conducting detailed fully Bayesian analysis with QTL positions moving. Result of the detailed analysis is more or less the same as that of the multiple-marker analysis. Further detailed analysis is only conducted after the investigators get a general picture of the result.

We now discuss several different ways to allow QTL positions to move across the genome. If our purpose of QTL mapping is to find the regions of the genome

that most likely carry QTL, the number of QTL is irrelevant and so are the QTL identities. If we allow QTL positions to move, the most important information we want to capture is how many times a particular segment (position) of the genome is hit or visited by nonspurious QTL. A position can be visited many times by different QTL, but all these QTL have negligible effects; such a position is not of interest. We are interested in positions that are visited repeatedly by QTL with large effects. Keeping this in mind, we propose the first strategy of QTL moving, the random walking strategy. We start with a "sufficient" number of QTL evenly placed on the genome. How sufficient is sufficient enough? This perhaps depends on the marker density and sample size of the mapping population. Putting one QTL in every 10 cM seems to work well. Each QTL is allowed to travel freely between the left and the right QTL, i.e., the QTL are distributed along the genome in a disjoint manner. The positions of the QTL are moving but the order of the QTL is preserved. This is the simplest method of QTL traveling. Let us take the kth QTL for example; the current position of the QTL is denoted by λ_k. The new position can be sampled from the following distribution:

$$\lambda_k^* = \lambda \pm \Delta\lambda \tag{15.66}$$

where $\Delta\lambda \sim U(0, \delta)$ and δ is the maximum distance (in cM) that the QTL is allowed to move away from the current position. The following restriction $\lambda_{k-1} < \lambda_k^* < \lambda_{k+1}$ is enforced to preserve the current order of the QTL. Empirically, $\delta = 2$ cM seems to work well. The new position is always accepted, regardless whether it is more likely or less likely to carry a true QTL relative to the current position. The Markov chain should be sufficiently long to make sure that all putative positions are visited a number of times. Theoretically, there is no need to enforce the disjoint distribution for the QTL positions. The only reason for such a restriction is the convenience of programming if the order is preserved. With the random walk strategy of QTL moving, the frequency of hits by QTL at a position is not of interest; instead, the average effect of all the QTL hitting that position is the important information. The random walk approach does not distinguish "hot regions" (regions containing QTL) and "cold regions" (regions without QTL) of the genome. All regions are visited with equal frequency. The hot regions, however, are supposed to be visited more often than the cold regions to get a more accurate estimate of the average QTL effects for those regions. The random walk approach does not discriminate against the cold regions and thus needs a very long Markov chain to ensure that the hot regions are sufficiently visited for accurate estimation of the QTL effects.

The optimal strategy for QTL moving is to allow QTL to visit the hot regions more often than the cold regions. This sampling strategy cannot be accomplished using the Gibbs sampler because the conditional posterior of the position of a QTL does not have a well-known form of the distribution. Therefore, the Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970) is adopted here to sample the QTL positions. Again, the new position is randomly generated in the neighborhood of the old position using the same approach as used in the random walk approach, but the new position λ_k^* is only accepted with a certain probability.

The acceptance probability is determined based on the Metropolis–Hastings rule, denoted by $\min\left[1, \alpha(\lambda_k^*, \lambda_k)\right]$. The new position λ_k^* has an $1 - \min\left[1, \alpha(\lambda_k^*, \lambda_k)\right]$ chance to be rejected, where

$$\alpha(\lambda_k^*, \lambda_k) = \frac{\prod_{j=1}^{n}\left[\sum_{l=-1,+1} \Pr(Z_{jk} = l|\lambda_k^*) p(y_j|Z_{jk} = l)\right]}{\prod_{j=1}^{n}\left[\sum_{l=-1,+1} \Pr(Z_{jk} = l|\lambda_k) p(y_j|Z_{jk} = l)\right]} \frac{q(\lambda_k|\lambda_k^*)}{q(\lambda_k^*|\lambda_k)} \tag{15.67}$$

If it is rejected, the QTL remains at the current position, i.e., $\lambda_k^* = \lambda_k$. If the new position is accepted, the old position is replaced by the new position, i.e., $\lambda_k^* = \lambda \pm \Delta\lambda$. Whether the new position is accepted or not, all other variables are updated based on the information from position λ_k^*, where $\Pr(Z_{jk} = -1|\lambda_k)$ and $\Pr(Z_{jk} = +1|\lambda_k)$ are the conditional probabilities that $Z_{jk} = -1$ and $Z_{jk} = +1$, respectively, calculated from the multipoint method. These probabilities depend on position λ_k. Previously, these probabilities were denoted by $p_j(-1) = \Pr(Z_{jk} = -1|\lambda_k)$ and $p_j(+1) = \Pr(Z_{jk} = +1|\lambda_k)$, respectively. For the new position λ_k^*, these probabilities are $\Pr(Z_{jk} = -1|\lambda_k^*)$ and $\Pr(Z_{jk} = +1|\lambda_k^*)$, respectively. The proposal probabilities $q(\lambda_k^*|\lambda_k)$ and $q(\lambda_k|\lambda_k^*)$ are usually equal to $\frac{1}{2\delta}$ and thus are canceled out each other. However, once λ_k and λ_k^* are near the boundaries, these two probabilities may not be the same. Since the new position is always restricted to the interval where the old position occurs, the proposal density $q(\lambda_k^*|\lambda_k)$ and its reverse partner $q(\lambda_k|\lambda_k^*)$ may be different. Let us denote the positions of the left and right QTL by λ_{k-1} and λ_{k+1}, respectively. If λ_k is close to the left QTL so that $\lambda_k - \lambda_{k-1} < \delta$, then the new position must be sampled from $\lambda_k^* \sim U(\lambda_k - \lambda_{k-1}, \lambda_k + \delta)$ to make sure that the new position is within the required sample space. Similarly, if λ_k is close to the right QTL so that $\lambda_{k+1} - \lambda_k < \delta$, then the new position must be sampled from $\lambda_k^* \sim U(\lambda_k - \delta, \lambda_{k+1})$. In either case, the proposal density should be modified. The general formula of the proposal density after incorporating the modification is

$$q(\lambda_k|\lambda_k^*) = \begin{cases} \frac{1}{\delta + (\lambda_k - \lambda_{k-1})} & \text{if } \lambda_k - \lambda_{k-1} < \delta \\ \frac{1}{\delta + (\lambda_{k+1} - \lambda_k)} & \text{if } \lambda_{k+1} - \lambda_k < \delta \\ \frac{1}{2\delta} & \text{otherwise} \end{cases} \tag{15.68}$$

The assumption of using the above proposal density is that the distance between any two QTL must be larger than δ. The reverse partner of this proposal density is

$$q(\lambda_k^*|\lambda_k) = \begin{cases} \frac{1}{\delta + (\lambda_k^* - \lambda_{k-1})} & \text{if } \lambda_k^* - \lambda_{k-1} < \delta \\ \frac{1}{\delta + (\lambda_{k+1} - \lambda_k^*)} & \text{if } \lambda_{k+1} - \lambda_k^* < \delta \\ \frac{1}{2\delta} & \text{otherwise} \end{cases} \tag{15.69}$$

The differences between sampling λ_k and sampling other variables are the following: (1) The proposed new position may or may not be accepted, while the new values of all other variables are always accepted, and (2) when calculating the

acceptance probability for a new position, the likelihood does not depend on the QTL genotype, while the conditional posterior probabilities of all other variables depend on sampled QTL genotypes.

15.3.3 Summary of the MCMC Process

The MCMC process is summarized as follows:

1. Choose the number of QTL to be placed in the model, p.
2. Initialize parameters and missing values, $\theta = \theta^{(0)}$ and $Z_j = Z_j^{(0)}$.
3. Sample β_i from $N(\beta_i|\hat{\beta}_i, \sigma^2_{\hat{\beta}_i})$.
4. Sample γ_k from $N(\gamma_k|\hat{\gamma}_k, \sigma^2_{\hat{\gamma}_k})$.
5. Sample σ^2_k from $\text{Inv} - \chi^2(\sigma^2_k|1, \gamma^2_k)$.
6. Sample σ^2 from $\text{Inv} - \chi^2(\sigma^2|n, S^2_e)$.
7. Sample Z_{jk} from its conditional posterior distribution.
8. Sample λ_k using the Metropolis–Hastings algorithm.
9. Repeat Step (3) to Step (8) until the chain reaches a desired length.

The length of the chain should be sufficiently long to make sure that, after burn-in deleting and chain trimming, the posterior sample size is large enough to allow accurate estimation of the posterior means (modes or medians) of all QTL parameters. Methods and computer programs are available to check whether the chain has converged to the stationary distribution (Gelfand et al. 1990; Gilks et al. 1996). Our past experience showed that the burn-in period may only contain a few thousand observations. The trimming frequency of saving one in every 20 observations is sufficient. The posterior sample size of 1,000 usually works well. However, if the model is not very large, it is always a good practice to delete more observations for the burn-in and trim more observations to make the chain thinner.

15.3.4 Post-MCMC Analysis

The MCMC process is much like doing an experiment. It only generates data for further analysis. The Bayesian estimates will only be available after summarizing the data (posterior sample). The parameter vector θ is very long, but not all parameters are of interest. Unlike other methods in which the number of QTL is an important parameter, the Bayesian shrinkage method uses a fixed number of QTL, and thus, p is not a parameter of interest. Although the variance component for the kth QTL, σ^2_k, is a parameter, it is also not a parameter of interest. It only serves as a factor to shrink the estimated QTL effect. Since the marginal posterior of σ^2_k does not exist, the empirical posterior mean or mode of σ^2_k does not have any biological meaning. In some observations, the sampled σ^2_k can be very large, and in others,

it may be very small. The residual error variance σ^2 is meaningful only if the number of QTL placed in the model is small to moderate. When p is very large, the residual error variance will be absorbed by the very large number of spurious QTL. The only parameters that are of interest are the QTL effects and QTL positions. However, the QTL identity, k, is also not something of interest. Since the kth QTL may jump all of places over the chromosome where it is originally placed, the average effect γ_k does not have any meaningful biological interpretation. The only things left are the positions of the genome that are hit frequently by QTL with large effects. Let us consider a fixed position of a genome. A position of a genome is only a point or a locus. Since the QTL position is a continuous variable, a particular point of the genome that is hit by a QTL has a probability of zero. Therefore, we define a genome position by a bin with a width of d cM, where d can be 1 or 2 or any other suitable value. The middle point value of the bin represents the genome location. For example, if $d = 2$ cM, the genome location 15 cM actually represents the bin covering a region of the genome from 14 cM to 16 cM, where $14 = 15 - \frac{1}{2}d$ and $16 = 15 + \frac{1}{2}d$. Once we define the bin width of a genome location, we can count the number of QTL that hit the bin. For each hit, we record the effect of that hit. The same location may be hit many times by QTL with the same or different identities. The average effect of the QTL hitting the bin is the most important parameter in the Bayesian shrinkage analysis. Each and every bin of the genome has an average QTL effect. We can then plot the effect against the genome location to form a QTL (effect) profile. This profile represents the overall result of the Bayesian mapping. In the BC example of Bayesian analysis, the kth QTL effect is denoted by γ_k. Since the QTL identity k is irrelevant, it is now replaced by the average QTL effect at position λ, which is a continuous variable. The λ without a subscript indicates a genome location. The average QTL effect at position λ can be expressed as $\gamma(\lambda)$ to indicate that the effect is a function of the genome location. The QTL effect profile is now represented by $\gamma(\lambda)$. If we use $\gamma(\lambda)$ to denote the posterior mean of QTL effect at position λ, we may use $\sigma^2(\lambda)$ to denote the posterior variance of QTL effect at position λ. If QTL moving is not random but guided by the Metropolis–Hastings rule, the posterior sample size at position λ should be a useful piece of information to indicate how often position λ is hit by a QTL. Let $n(\lambda)$ be the posterior sample size at λ; the standard error of the QTL effect at λ should be $\sigma(\lambda)/\sqrt{n(\lambda)}$. Therefore, another useful profile is the so-called t-test statistic profile expressed as

$$t(\lambda) = \sqrt{n(\lambda)}\frac{\gamma(\lambda)}{\sigma(\lambda)} \tag{15.70}$$

The corresponding F-test statistic profile is

$$F(\lambda) = n(\lambda)\frac{\gamma^2(\lambda)}{\sigma^2(\lambda)} \tag{15.71}$$

The t-test statistic profile is more informative than the F-test statistic profile because it also indicates the direction of the QTL effect (positive or negative) while

the F-test statistic profile is always positive. On the other hand, the F-test statistic can be extended to multiple effects per locus, e.g., additive and dominance in an F_2 design. Both the t-test and F-test statistic profiles can be interpreted as kinds of weighted QTL effect profiles because they incorporated the posterior frequency of the genome location.

Before moving on to the next section, let us use a simulated example to demonstrate the behavior of the Bayesian shrinkage mapping and its difference from the maximum likelihood interval mapping. The mapping population was a simulated BC family with 500 individuals. A single chromosome of 2,400 cM in length was evenly covered by 121 markers (20 cM per marker interval). The positions and effects of 20 simulated QTL are demonstrated in Fig. 15.2 (top panel). In the Bayesian model, we placed one QTL in every 25 cM to start the search. The QTL positions constantly moved according to the Metropolis–Hastings rule. The burn-in period was set at 2,000, and one observation was saved in every 50 iterations after the burn-in. The posterior sample size was 1,000. We also analyzed the same data set using the maximum likelihood interval mapping procedure. The QTL effect profiles for both the Bayesian and ML methods are demonstrated in Fig. 15.2 also (see the panels in the middle and at the bottom). The Bayesian shrinkage estimates of the QTL effects are indeed smaller than the true values, but the resolution of the signal is much clearer that the maximum likelihood estimates. The Bayesian method has separated closely linked QTL in several places of the genome very well, which is clearly in contrast to the maximum likelihood method. The ML interval mapping provides exaggerated estimates of the QTL effects across the entire genome.

15.4 Alternative Methods of Bayesian Mapping

15.4.1 Reversible Jump MCMC

Reversible jump Markov chain Monte Carlo (RJMCMC) was originally developed by Green (1995) for model selection. It allows the model dimension to change during the MCMC sampling process. Most people believe that QTL mapping is a model selection problem because the number of QTL is not known a priori. Sillanpää and Arjas (1998, 1999) are the first people to apply the RJMCMC algorithm to QTL mapping. They treated the number of QTL, denoted by p, as an unknown parameter and infer the posterior distribution of p. The assumption is that p is a small number for a quantitative trait and thus can be assigned a Poisson prior distribution with mean ρ. Sillanpää and Arjas (1998) used the Metropolis–Hastings algorithm to sample all parameters, even though most QTL parameters have known forms of the fully conditional posterior distributions. The justification for use of M–H sampling strategy is that it is a general sampling approach while the Gibbs sampling is only a special case of the M–H sampling. The M–H sampler does not require derivation of the conditional posterior distribution for a parameter. However,

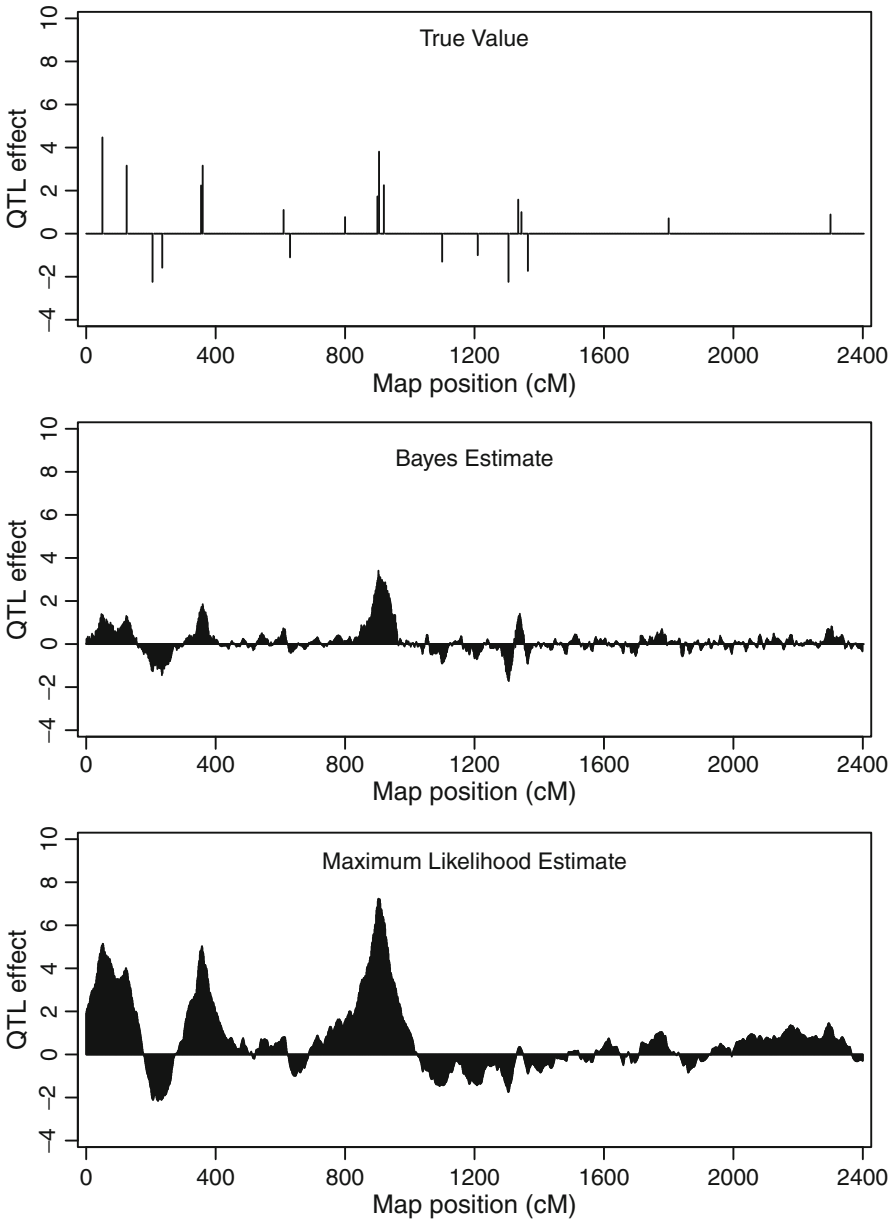

Fig. 15.2 Plots of QTL effect against genome location (QTL effect profiles) for the simulated BC population. The *top panel* shows the true locations and effects of the simulated QTL. The panel in the *middle* shows the Bayesian shrinkage estimates of the QTL effects. The panel at the *bottom* gives the maximum likelihood estimates of the QTL effects

the acceptance probability for a proposed new value of a parameter is usually less than unity because the proposal distribution from which the new value is sampled is a uniform distribution in the neighborhood of the old value and not from the conditional posterior distribution. Therefore, the M–H sampler is computationally less efficient. Yi and Xu (1999, 2000, 2001) extended RJMCMC to QTL mapping for binary traits in line crosses and random mating populations using Gibbs sampler for all parameters except the number of QTL and the location of QTL. In this section, we only introduce the RJMCMC for sampling the number of QTL. All other variables are sampled using the same method as described in the Bayesian shrinkage analysis. Another difference between the RVJMCMC and the Bayesian shrinkage method is that γ_k is assigned a uniform prior distribution for the RVJMCMC method while a $N(0, \sigma_k^2)$ prior is chosen for the shrinkage method. The conditional posterior distribution of γ_k remains normal but with mean and variance defined as

$$\hat{\gamma}_k = \left(\sum_{j=1}^{n} Z_{jk}^2 \right)^{-1} \sum_{j=1}^{n} Z_{jk} \left(y_j - \sum_{i=1}^{q} X_{ji} \beta_i - \sum_{k' \neq k}^{p} Z_{jk'} \gamma_{k'} \right) \tag{15.72}$$

and

$$\sigma_{\hat{\gamma}_k}^2 = \left(\sum_{j=1}^{n} Z_{jk}^2 \right)^{-1} \sigma^2 \tag{15.73}$$

respectively.

We now introduce the reversible jump MCMC. The prior distribution for p is assumed to be a truncated Poisson with mean ϕ and maximum P. The probability distribution function of p is

$$\Pr(p) = \left(\frac{\Gamma(P+1, \phi)}{P!} \right)^{-1} \left(\frac{\phi^p \mathrm{e}^{-\phi}}{p!} \right) \propto \frac{\phi^p \mathrm{e}^{-\phi}}{p!} \tag{15.74}$$

where $\Gamma(P+1, \phi)$ is an incomplete Gamma function and

$$\frac{\Gamma(P+1, \phi)}{P!} = \sum_{p=0}^{P} \Pr(p) \tag{15.75}$$

is the cumulative Poisson distribution up to P, which is irrelevant to p and thus a constant. We make a random choice among three move types of the dimensionality change: (1) Do not change the dimension, but update all other parameters except p with probability p_0; (2) add a QTL to the model with probability p_a; and (3) delete a QTL from the model with probability p_d. The three probabilities of move types sum to one, i.e., $p_0 + p_a + p_d = 1$. The following values of the probabilities may be chosen, $p_0 = p_a = p_d = \frac{1}{3}$. If no change is proposed, all other parameters are sampled from their conditional posterior distributions. If adding a QTL is proposed,

we choose a chromosome to place the QTL. The probability of each chromosome being chosen is proportional to the size of the chromosome. Once a chromosome is chosen, we place the proposed new QTL randomly on the chromosome. All parameters associated with this new QTL are sampled from their prior distributions. The new QTL is then accepted with a probability determined by $\min[1, \alpha(p+1, p)]$, where

$$\alpha(p+1, p) = \frac{\prod_{j=1}^{n} p(y_j|p+1)}{\prod_{j=1}^{n} p(y_j|p)} \times \frac{\phi}{p+1} \times \frac{p_d}{(p+1)p_a} \tag{15.76}$$

There are three ratios occurring in the above equation. The first ratio is the likelihood ratio, the second one is the prior ratio of the number of QTL, and the third ratio is the proposal ratio. The likelihood is defined as

$$p(y_j|p+1) = N\left(y_j \left| \sum_{i=1}^{q} X_{ji}\beta_i + \sum_{k=1}^{p} Z_{jk}\gamma_k + Z_{j(p+1)}\gamma_{(p+1)}, \sigma^2 \right.\right) \tag{15.77}$$

and

$$p(y_j|p) = N\left(y_j \left| \sum_{i=1}^{q} X_{ji}\beta_i + \sum\nolimits_{k=1}^{p} Z_{jk}\gamma_k, \sigma^2 \right.\right) \tag{15.78}$$

The prior probability for p is

$$\Pr(p) = \frac{\phi^p \mathrm{e}^{-\phi}}{p!} \tag{15.79}$$

and the prior probability for $p+1$ is

$$\Pr(p+1) = \frac{\phi^{p+1} \mathrm{e}^{-\phi}}{(p+1)!} \tag{15.80}$$

Therefore, the prior ratio is

$$\frac{\Pr(p+1)}{\Pr(p)} = \frac{\phi^{p+1}\mathrm{e}^{-\phi}}{(p+1)!} \frac{p!}{\phi^p \mathrm{e}^{-\phi}} = \frac{\phi}{p+1} \tag{15.81}$$

The proposal probability for adding a QTL is $\xi(p+1, p) = p_a$. The reverse partner is $\xi(p, p+1) = \frac{p_d}{p+1}$. It is easy to understand $\xi(p+1, p) = p_a$ because we already defined that p_a is the probability of adding a QTL. However, the reverse partner is not p_d but $p_d/(p+1)$, which is hard to understand if we do not understand the Hastings' adjustment for the proposal probability. This probability says that if a deletion has occurred (with probability p_d) given that we have $p+1$ QTL in the model, the probability that the newly added QTL (not any other QTL) is deleted is $1/(p+1)$ due to the fact that each QTL has an equal chance to be deleted. Therefore,

the probability that the newly added QTL (not others) is deleted is $p_d/(p+1)$. As a result, the proposal ratio is

$$\frac{\xi(p, p+1)}{\xi(p+1, p)} = \frac{p_d/(p+1)}{p_a} = \frac{p_d}{(p+1)p_a} \tag{15.82}$$

Note that the proposal ratio is the probability of deleting a QTL to the probability of adding a QTL, not the other way around. This Hastings' adjustment is important to prevent the Markov chain from being trapped at a particular QTL number. This is the very reason for the name "reversible jump." The dimension of the model can jump in either direction without being stuck at a local value of p.

If deleting a QTL is proposed, we randomly select one of the p QTL to be deleted. Suppose that the kth QTL happens to be the unlucky one. The number of QTL would change from p to $p-1$. The reduced model with $p-1$ QTL is accepted with probability $\min[1, \alpha(p-1, p)]$, where

$$\alpha(p-1, p) = \frac{\prod_{j=1}^{n} p(y_j|p)}{\prod_{j=1}^{n} p(y_j|p-1)} \times \frac{p}{\phi} \times \frac{p_a p}{p_d} \tag{15.83}$$

where

$$p(y_j|p-1) = N\left(y_j \left| \sum_{i=1}^{q} X_{ji}\beta_i + \sum_{k' \neq k}^{p} Z_{jk'}\gamma_{k'}, \sigma^2 \right.\right) \tag{15.84}$$

The prior ratio is

$$\frac{\Pr(p-1)}{\Pr(p)} = \frac{\phi^{p-1}\mathrm{e}^{-\phi}}{(p-1)!} \frac{p!}{\phi^p \mathrm{e}^{-\phi}} = \frac{p}{\phi} \tag{15.85}$$

The proposal ratio is

$$\frac{\xi(p, p-1)}{\xi(p-1, p)} = \frac{p_a}{p_d/p} = \frac{p_a p}{p_d} \tag{15.86}$$

The reversible jump MCMC requires more cycles of simulations because of the frequent change of model dimension. When a QTL is deleted, all parameters associated with this QTL are gone. The chain does not memorize this QTL. In the future, if a new QTL is added to the neighborhood of this deleted QTL, the parameter associated to this added QTL must be sampled anew from the prior distribution. Even if the newly added QTL occupies exactly the same location as a previously deleted QTL, the information of the previously deleted QTL is gone permanently and cannot be reused. An improved RJMCMC may be developed to memorize the information associated with deleted QTL. If the position of a deleted QTL is sampled again later in the MCMC process (a new QTL is added to a previously deleted QTL), the parameters associated with that deleted QTL can be used again to facilitate the sampling for the newly added QTL. The improved

method can substantially improve the mixing of the Markov chain and speed up the MCMC process. The tradeoff is the increased computer memory requirement for the improved method.

With the RJMCMC, the QTL number is a very important parameter. Its posterior distribution is always reported. Each QTL occurring in the model is deemed to be important and counted. In addition, the positions of QTL are usually determined by the so-called QTL intensity profile, which is simply the plot of a scaled posterior sample at a particular location $n(\lambda)$ against the genome location λ.

15.4.2 Stochastic Search Variable Selection

Stochastic search variable selection (SSVS) is a variable selection strategy for large models. The method was originally developed by George and McCulloch (1993, 1997) and applied to QTL mapping for the first time by Yi et al. (2003). The difference between this method and many other methods of model selection is that the model dimension is fixed at a predetermined value, just like the Bayesian shrinkage analysis. Model selection is actually conducted by introducing a series of binary variables, one for each model effect, i.e., the QTL effect. For p QTL effects, p indicator variables are required. Let η_k be the indicator variable for the kth QTL. If $\eta_k = 1$, the QTL is equivalent to being included in the model, and the effect will not be shrunken. If $\eta_k = 0$, the effect will be forced to take a value closed to, but not exactly equal to, zero. Essentially, the prior distribution of the kth QTL takes one of two normals. The switching button is variable η_k, as given below:

$$p(\gamma_k) = \eta_k N(\gamma_k|0, \Delta) + (1 - \eta_k)N(\gamma_k|0, \delta) \tag{15.87}$$

where δ is a small positive number closed to zero, say 0.0001, and Δ is a large positive value, say 1,000. The two variances (δ and Δ) are constant hyperparameters. The indicator variable is not known, and thus, the above distribution is a mixture of two normal distributions. Let $p(\eta_k = 1) = \rho$ be the probability that γ_k comes from the first distribution; the mixture distribution is

$$p(\gamma_k) = \rho N(\gamma_k|0, \Delta) + (1 - \rho)N(\gamma_k|0, \delta) \tag{15.88}$$

The mixture proportion ρ is unknown and is treated as a parameter. When the indicator variable (η_k) is known, the posterior distribution of γ_k is $p(\gamma_k|\cdots) = N(\gamma_k|\hat{\gamma}_k, \sigma^2_{\hat{\gamma}_k})$. The mean and variance of this normal are

$$\hat{\gamma}_k = \left(\sum_{j=1}^{n} Z_{jk}^2 + \frac{\sigma^2}{\upsilon_k}\right)^{-1} \sum_{j=1}^{n} Z_{jk} \left(y_j - \sum_{i=1}^{q} X_{ji}\beta_i - \sum_{k' \neq k}^{p} Z_{jk'}\gamma_{k'}\right) \tag{15.89}$$

and

$$\sigma^2_{\hat{\gamma}_k} = \left(\sum_{j=1}^{n} Z^2_{jk} + \frac{\sigma^2}{\upsilon_k} \right)^{-1} \sigma^2 \tag{15.90}$$

respectively, where

$$\upsilon_k = \eta_k \Delta + (1 - \eta_k)\delta \tag{15.91}$$

is the actual variance of the posterior distribution. Let the prior distribution for η_k be

$$p(\eta_k) = \text{Bernoulli}(\eta_k|\rho) \tag{15.92}$$

The conditional posterior distribution of $\eta_k = 1$ is

$$p(\eta_k = 1|\cdots) = \frac{\rho N(\gamma_k|0, \Delta)}{\rho N(\gamma_k|0, \Delta) + (1 - \rho)N(\gamma_k|0, \delta)} \tag{15.93}$$

There is another parameter ρ involved in the conditional posterior distribution. Yi et al. (2003) treated ρ as a hyperparameter and set $\rho = \frac{1}{2}$. This prior works well for small models but fails most often for large models. The optimal strategy is to assign another prior to ρ so that ρ can be estimated from the data. Xu (2007) took a beta prior for ρ , i.e.,

$$p(\rho) = \text{Beta}(\rho|\zeta_0, \zeta_1) = \frac{\Gamma(\zeta_0 + \zeta_1)}{\Gamma(\zeta_0)\Gamma(\zeta_1)} \rho^{\zeta_1 - 1}(1 - \rho)^{\zeta_0 - 1} \tag{15.94}$$

Under this prior, the conditional posterior distribution for ρ remains beta,

$$p(\rho|\cdots) = \text{Beta}\left(\rho \left| \zeta_0 + p - \sum\nolimits_{k=1}^{p} \eta_k, \zeta_1 + \sum\nolimits_{k=1}^{p} \eta_k \right.\right) \tag{15.95}$$

The values of the hyperparameters were chosen by Xu (2007) as $\zeta_0 = 1$ and $\zeta_1 = 1$, leading to an uninformative prior for ρ, i.e.,

$$p(\rho) = \text{Beta}(\rho|1, 1) = \text{constant} \tag{15.96}$$

The Gibbs sampler for σ^2_k in the Bayesian shrinkage analysis is replaced by sampling η_k from

$$p(\eta_k|\cdots) = \text{Bernoulli}\left(\eta_k \left| \frac{\rho N(\gamma_k|0, \Delta)}{\rho N(\gamma_k|0, \Delta) + (1 - \rho)N(\gamma_k|0, \delta)} \right. \right) \tag{15.97}$$

and sampling ρ from

$$p(\rho|\cdots) = \text{Beta}\left(\rho \left| 1 + p - \sum\nolimits_{k=1}^{p} \eta_k, 1 + \sum\nolimits_{k=1}^{p} \eta_k \right.\right) \tag{15.98}$$

in the SSVS analysis.

The additional information extracted from SSVS is the probabilistic statement about a QTL. If the marginal posterior mean of η_k is large, say $p(\eta_k|\text{data})>95\,\%$, the evidence of locus k being a QTL is strong. If the QTL position is allowed to move, η_k does not have any particular meaning. Instead, the number of hit of a particular location of the genome by QTL with $\eta(\lambda) = 1$ is more informative.

15.4.3 Lasso and Bayesian Lasso

Lasso

Lasso refers to a method called least absolute shrinkage and selection operator (Tibshirani 1996). The method can handle extremely large models by minimizing the residual sum of squares subject to a predetermined constraint, the constraint that the sum of absolute values of all regression coefficients is smaller than a predetermined shrinkage factor. Mathematically, the solution of regression coefficients is obtained by

$$\min_{\gamma} \sum_{j=1}^{n} \left(y_j - \sum\nolimits_{k=1}^{p} Z_{jk}\gamma_k \right)^2 \tag{15.99}$$

subject to constraint

$$\sum_{k=1}^{p} |\gamma_k| \leq t \tag{15.100}$$

where $t > 0$. When $t = 0$, all regression coefficients must be zero. As t increases, the number of nonzero regression coefficients progressively increases. As $t \to \infty$, the Lasso estimates of the regression coefficients are equivalent to the ordinary least-squares estimates. Another expression of the problem is

$$\min_{\gamma} \left[\sum_{j=1}^{n} \left(y_j - \sum\nolimits_{k=1}^{p} Z_{jk}\gamma_k \right)^2 + \lambda \sum_{k=1}^{p} |\gamma_k| \right] \tag{15.101}$$

where $\lambda \geq 0$ is a Lagrange multiplier (unknown) which relates implicitly to the bound t and controls the degree of shrinkage. The effect of λ on the level of shrinkage is just the opposite of t, with $\lambda = 0$ being no shrinkage and $\lambda \to \infty$ being the strongest shrinkage where all γ_k are shrunken down to zero. Note that the Lasso model does not involve $X_j\beta$, the non-QTL effect described earlier in the chapter. The non-QTL effect in the original Lasso refers to the population mean. For simplicity, Tibshirani (1996) centered y_j and all the independent variables. The centered y_j is simply the original y_j subtracted by $\bar{y}$, the population mean. The corresponding centered independent variables are also obtained by subtraction of $\bar{Z}_k$ from Z_{jk}. The Lasso estimates of regression coefficients can be efficiently computed via quadratic programming with linear constraints. An efficient algorithm

called LARS (least angle regression) was developed by Efron et al. (2004) to implement the Lasso method. The Lagrange multiplier λ or the original t is called the Lasso parameter. The original Lasso estimates λ using the fivefold cross validation approach. One can also use any other fold cross validations, for example, the n-fold (leave-one-out) cross validation. Under each λ value, the fivefold cross validation is used to calculate the prediction error (PE),

$$\mathrm{PE} = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \sum\nolimits_{k=1}^{p} Z_{jk}\hat{\gamma}_k\right)^2 \tag{15.102}$$

This formula appears to be the same as the estimated residual error variance. However, the prediction error differs from the residual error in that the individuals predicted do not contribute to parameter estimation. With the fivefold cross validation, we use $\frac{4}{5}$ of the sample to estimate γ_k and then use the estimated γ_k to predict the errors for the remaining $\frac{1}{5}$ sample. In other words, when we calculate $\left(y_j - \sum_{k=1}^{p} Z_{jk}\hat{\gamma}_k\right)^2$, the γ_k is estimated from $\frac{4}{5}$ of the sample that excludes y_j. Under each λ, the PE is calculated, denoted by $\mathrm{PE}(\lambda)$. We vary λ from 0 to large value. The λ value that minimizes $\mathrm{PE}(\lambda)$ is the optimal value of λ.

Bayesian Lasso

Lasso can be interpreted as Bayesian posterior mode estimation of regression coefficients when each regression coefficient is assigned an independent double-exponential prior (Tibshirani 1996; Yuan and Lin 2005; Park and Casella 2008). However, Lasso provides neither the estimate for the residual error variance nor the interval estimate for a regression coefficient. These deficiencies of Lasso can be overcome by the Bayesian Lasso (Park and Casella 2008). The double-exponential prior for γ_k is

$$p(\gamma_k|\lambda) = \frac{\lambda}{2}\exp(-\lambda|\gamma_k|) \tag{15.103}$$

where λ is the Lagrange multiplier in the classical Lasso method (see (15.101)). This prior can be derived from a two-level hierarchical model. The first level is

$$p(\gamma_k|\sigma_k^2) = N(\gamma_k|0,\sigma_k^2) \tag{15.104}$$

and the second level is

$$p(\sigma_k^2|\lambda) = \frac{\lambda^2}{2}\exp\left(-\sigma_k^2\frac{\lambda^2}{2}\right) \tag{15.105}$$

Therefore,

$$p(\gamma_k|\lambda) = \int_0^\infty p(\gamma_k|\sigma_k^2)p(\sigma_k^2|\lambda)\mathrm{d}\sigma_k^2 = \frac{\lambda}{2}\exp(-\lambda|\gamma_k|) \tag{15.106}$$

The Bayesian Lasso method uses the same model as the Lasso method. However, centralization of independent variables is not required, although it is still recommended. The model is described as follows:

$$y_j = \sum_{i=1}^{q} X_{ji}\beta_i + \sum_{k=1}^{p} Z_{jk}\gamma_k + \epsilon_j \tag{15.107}$$

where β_i remains in the model and can be estimated along with the residual variance σ^2 and all QTL effects. Bayesian Lasso provides the posterior distributions for all parameters. The marginal posterior mean of each parameter is the Bayesian Lasso estimate, which is different from the posterior mode estimate obtained from the Lasso analysis. The Bayesian Lasso differs from the Bayesian shrinkage analysis only in the prior distribution for σ_k^2. Under the Bayesian Lasso, the prior for σ_k^2 is

$$p(\sigma_k^2|\lambda) = \frac{\lambda^2}{2}\exp\left(-\sigma_k^2\frac{\lambda^2}{2}\right) \tag{15.108}$$

The Lasso parameter λ needs a prior distribution so that we can estimate λ from the data rather than choosing an arbitrary value a priori. Park and Casella (2008) choose the following gamma prior for λ^2 (not λ):

$$p(\lambda^2|a,b) = \mathrm{Gamma}(\lambda^2|a,b) = \frac{b^a}{\Gamma(a)}(\lambda^2)^{a-1}\exp\left(-b\lambda^2\right) \tag{15.109}$$

The reason for choosing such a prior is to enjoy the conjugate property. The hyperparameters, a and b, are sufficiently remote from σ_k^2 and γ_k, and thus, their values can be chosen in an arbitrary fashion. Yi and Xu (2008) used several different sets of values for a and b and found no significant differences among those values. For convenience, we may simply set $a = b = 1$, which is sufficiently different from 0. Note that $a = b = 0$ produces an improper prior for λ^2. Once a and b values are chosen, everything else can be estimated from the data.

The fully conditional posterior distributions for most variables remain the same as the Bayesian shrinkage analysis except that the following variables must be sampled using the posterior distribution derived under the Bayesian Lasso prior distribution. For the kth QTL variance, it is better to deal with $\alpha_k = \frac{1}{\sigma_k^2}$. The conditional posterior for α_k is an inverse Gaussian distribution,

$$p(\alpha_k|\cdots) = \mathrm{Inv-Gassian}\left(\alpha_k \left| \sqrt{\frac{\lambda^2\sigma^2}{\gamma_k^2}}, \lambda^2\right.\right) \tag{15.110}$$

Algorithm for sampling a random variable from an inverse Gaussian is available. Once α_k is sampled, σ_k^2 simply takes the inverse of α_k. The fully conditional posterior distribution for λ^2 remains gamma because of the conjugate property of the gamma prior,

$$p(\lambda^2|\cdots) = \text{Gamma}\left(\lambda^2 \middle| p + a, \frac{1}{2}\sum\nolimits_{k=1}^{p} \sigma_k^2 + b\right) \tag{15.111}$$

The Bayesian Lasso can potentially improve the estimation of regression coefficients for the following reasons: (1) It assigns an exponential prior, rather than a scaled inverse chi-square prior, distribution to σ_k^2, and (2) it increases the hierarchy of the prior to another level so that the hyperparameters do not have strong influence on the Bayesian estimates of the regression coefficients.

15.5 Example: Arabidopsis Data

The first example 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 $\mathtt{F}_7$ progeny of single-seed descendants of the $\mathtt{F}_2$ plants. Flowering time was recorded for each line in two environments: long day (16-h photoperiod) and short day (8-h 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 inbred 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 the QTL mapping. These markers are more or less evenly distributed along five chromosomes with an average 10.8 cM per marker interval. The marker names and positions are given in the original article (Loudet et al. 2002). We inserted a pseudomarker in every 5 cM of the genome. Including the inserted pseudomarkers, the total number of loci subject to analysis was 74 (38 true markers plus 36 pseudomarkers). All the 74 putative loci were evaluated simultaneously in a single model. Therefore, the model for the short-day flowering time trait is

$$y = X\beta + \sum_{k=1}^{74} Z_k \gamma_k + \epsilon$$

where X is a 420×1 vector of unity, Z_k coded as 1 for one genotype and 0 for the other genotype for locus k. If locus k is a pseudomarker, $Z_k = \Pr(\text{genotype} = 1)$, which is the conditional probabilities of marker k being of genotype 1. Finally, γ_k is the QTL effect of locus k. For the original data analysis, the burn-in period was 1,000. The thinning rate was 10. The posterior sample size was 10,000, and thus, the total number of iterations was $1,000 + 10,000 \times 10 = 101,000$.

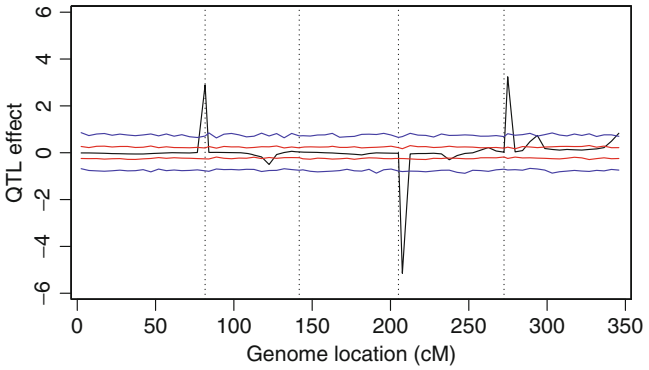

Fig. 15.3 The estimated QTL effects (*black*) and the permutation generated 1 % (*blue*) and 5 % (*red*) confidence intervals for the Arabidopsis short-time flowering time trait. The *dotted reference lines* separate the five chromosomes

We also performed a permutation analysis (Che and Xu 2010) to generate empirical quantiles of the QTL effects under the null model. The posterior sample size in permutation analysis was 80,000. The total number of iterations was $1,000 + 80,000 \times 10 = 801,000$. The estimated QTL effects and the permutation generated 0.5 % and 99.5 % (corresponding to a type I error of 0.01) and 2.5 % and 97.5 % (corresponding to a type I error of 0.05) are shown in Fig. 15.3. Based on the 0.01 criterion, a total of five QTL were detected on four chromosomes (1, 3, 4, and 5).