Chapter 5

QTL Mapping for Discrete Traits

10

Interval Mapping for Ordinal Traits

Many disease resistance traits in agricultural crops are measured in ordered categories. The generalized linear model (GLM) methodology (Nelder and Wedderburn, 1972; Wedderburn, 1974; McCullagh and Nelder, 1999) is an ideal tool to analyze these traits. Ordinal traits are usually controlled by the segregation of multiple QTL and environmental factors. The genetic architecture of such traits can be studied using linkage analysis. One can analyze the association of each marker with the disease phenotype. If the marker information is fully observable, i.e., marker genotypes can be observed, the standard GLM methodology can be directly applied to the association study by screening markers of the entire genome for their association with the disease trait. Many statistical software packages, e.g., SAS (SAS Institute, 2008b), have built-in functions or procedures to perform the standard GLM analysis. One can simply execute the built-in procedures many times, one for each marker, to scan the entire genome without developing a new computer program. In any genetic experiments, missing marker genotypes are unavoidable. In addition, interval mapping requires detection of association between the trait phenotype and loci that are not necessarily located at marker positions. Genotypes of these additional loci are never observed. Therefore, GLM with missing values must be applied. There is a rich literature on the missing value GLM analysis (Ibrahim, 1990; Horton and Laird, 1999; Ibrahim et al., 2002, 2005). The most popular method is the maximum likelihood (ML) method implemented via the EM algorithm (Horton and Laird, 1999). Other methods are also available, such as multiple imputation (MI, Rubin (1987)), fully Bayesian (FB, Ibrahim et al (2002)) and weighted estimating equations (WEE, Ibrahim et al. (2005)). A complete review on the methods can be found in Ibrahim et al. (2005). Hackett and Weller (1995) first applied the ML method to mapping ordinal trait QTL. They took advantage of an existing software package named GeneStat for the standard GLM analysis (without missing covariates) and modified the software by incorporating a weight variable. The modified GLM for missing data duplicates the data by the number of genotypes per locus, e.g., two for a backcross population and three for an $\mathtt{F}_2$ population. The weight variable

is simply the posterior probabilities of the missing genotypes. The weight variable is updated iteratively until the iteration converges. The modified GLM program is not necessarily simpler than a program written anew. Furthermore, the variance-covariance matrix of estimated parameters is not available for the modified GML algorithm. Xu et al. (2003) developed an explicit EM algorithm using the posterior probability of missing covariates as the weight variable and further provided the variance-covariance matrix of the estimated parameters by using the Louis' (1982) adjustment for the information matrix. Standard deviations (square roots of the variances) of estimated parameters represent the precisions of the estimates, which are required in the final report for publication. The variance-covariance matrix of the estimated QTL effects can also be used to calculate the Wald-test statistic (Wald, 1943), which is an alternative test that can replace the likelihood ratio test statistic. Although using the large sample distribution for the likelihood ratio test gives more accurate approximation for small and moderate-sized samples, the latter has a computational advantage since it does not require calculation of the likelihood function under the null model (McCulloch and Searle, 2001). A missing QTL genotype usually has partial information, which can be extracted from linked markers. This information can be used to infer the QTL genotypes using several different ways (McCulloch and Searle, 2001). In QTL mapping for continuously distributed traits, mixture model (Lander and Botstein, 1989) is the most efficient way to take advantage of marker information. The least squares method of Haley and Knott (1992) is the simplest way to incorporate linked markers. Performances of the weighted least squares method of Xu (1998a,b) and estimating equations (EE) algorithm of Feenstra et al. (2006) are usually between the least squares and mixture model methods. These methods have been successfully applied to QTL mapping for continuous traits, but they have not been investigated for ordinal trait QTL mapping. This chapter will introduce several alternative GLM methods for mapping quantitative trait loci of ordinal traits.

10.1 Generalized linear model

Suppose that a disease phenotype of individual j ($j = 1, ..., n$) is measured by an ordinal variable denoted by $S_j = 1, ..., p+1$, where $p+1$ is the total number of disease classes and n is the sample size. Let $Y_j = \{Y_{jk}\}, \forall k = 1, ..., p+1$, be a $(p+1) \times 1$ vector to indicate the disease status of individual j. The kth element of Y_j is defined as

$$Y_{jk} = \begin{cases} 1 & \text{if } S_j = k \\ 0 & \text{if } S_j \neq k \end{cases} \tag{10.1}$$

Using the probit link function, the expectation of Y_{jk} is defined as

$$\mu_{jk} = E(Y_{jk}) = \Phi(\alpha_k + X_j\beta + Z_j\gamma) - \Phi(\alpha_{k-1} + X_j\beta + Z_j\gamma) \tag{10.2}$$

where α_k ($\alpha_0 = -\infty$ and $\alpha_{p+1} = +\infty$) is the intercept, β is a $q \times 1$ vector for some systematic effects (not related to the effects of quantitative trait loci), and γ is an $r \times 1$ vector for the effects of a quantitative trait locus. The symbol $\Phi(.)$ is the standardized cumulative normal function. The design matrix X_j is assumed to be known, but Z_j may not be fully observable because it is determined by the genotype of j for the locus of interest. Because the link function is probit, this type of analysis is called probit analysis. Let $\mu_j = \{\mu_{jk}\}$ be a $(p+1) \times 1$ vector. The expectation for vector Y_j is $E(Y_j) = \mu_j$ and the variance matrix of Y_j is

$$V_j = \text{var}(Y_j) = \psi_j + \mu_j \mu_j^T \tag{10.3}$$

where $\psi_j = \text{diag}(\mu_j)$. The method to be developed requires the inverse of matrix V_j . However, V_j is not of full rank. We can use a generalized inverse of V_j, such as $V_j^- = \psi_j^{-1}$, in place of V_j^{-1}. The parameter vector is $\theta = \{\alpha, \beta, \gamma\}$ with a dimensionality of $(p+q+r) \times 1$. Binary data is a special case of ordinal data in that $p = 1$ so that there are only two categories, $S_j = \{1, 2\}$. The expectation of Y_{jk} is

$$\mu_{jk} = \begin{cases} \Phi(\alpha_1 + X_j\beta + Z_j\gamma) - \Phi(\alpha_0 + X_j\beta + Z_j\gamma) & \text{for } k = 1 \\ \Phi(\alpha_2 + X_j\beta + Z_j\gamma) - \Phi(\alpha_1 + X_j\beta + Z_j\gamma) & \text{for } k = 2 \end{cases} \tag{10.4}$$

Because $\alpha_0 = -\infty$ and $\alpha_2 = +\infty$ in the binary case, we have

$$\mu_{jk} = \begin{cases} \Phi(\alpha_1 + X_j\beta + Z_j\gamma) & \text{for } k = 1 \\ 1 - \Phi(\alpha_1 + X_j\beta + Z_j\gamma) & \text{for } k = 2 \end{cases} \tag{10.5}$$

We can see that $\mu_{j2} = 1 - \mu_{j1}$ and

$$\Phi^{-1}(\mu_{j1}) = \alpha_1 + X_j\beta + Z_j\gamma \tag{10.6}$$

The link function is $\Phi^{-1}(.)$ and thus it is called the probit link function. Once we take the probit transformation, the model becomes a linear model. Therefore, this type of model is called a generalized linear model (GLM). The ordinary linear model we learned before for continuous traits is a special case of the GLM because the link function is simply the identity, i.e.,

$$I^{-1}(\mu_{j1}) = \alpha_1 + X_j\beta + Z_j\gamma \tag{10.7}$$

or simply

$$\mu_{j1} = \alpha_1 + X_j\beta + Z_j\gamma \tag{10.8}$$

Most techniques we learned for the linear model applies to the generalized linear model.

10.2 ML under homogeneous variance

Let us first assume that the genotypes of the QTL are observed for all individuals. In this case, variable Z_j is not missing. The log likelihood function under the probit model is

$$L(\theta) = \sum_{j=1}^{n} L_j(\theta) \tag{10.9}$$

where

$$L_j(\theta) = \sum_{k=1}^{p+1} Y_{jk} \ln [\Phi(\alpha_k + X_j\beta + Z_j\gamma) - \Phi(\alpha_{k-1} + X_j\beta + Z_j\gamma)] \tag{10.10}$$

and $\theta = \{\alpha, \beta, \gamma\}$ is the vector of parameters. This is the simplest GLM problem and the classical iteratively reweighted least squares approach for GLM (Nelder and Wedderburn, 1972; Wedderburn, 1974) can be used without any modification. The iterative equation under the classical GLM is given below,

$$\theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta^{(t)})S(\theta^{(t)}) \tag{10.11}$$

where $\theta^{(t)}$ is the parameter value in the current iteration, $I(\theta^{(t)})$ is the information matrix and $S(\theta^{(t)})$ is the score vector, both evaluated at $\theta^{(t)}$. We can interpret

$$\Delta\theta = I^{-1}(\theta^{(t)})S(\theta^{(t)}) \tag{10.12}$$

in equation (10.11) as the adjustment for $\theta^{(t)}$ to improve the solution in the direction that leads to the ultimate maximum likelihood estimate of θ . Equation (10.3) shows that the variance of Y_j is a function of the expectation of Y_j . This special relationship leads to a convenient way to calculate the information matrix and the score vector, as given by Wedderburn (1974),

$$I(\theta) = \sum\nolimits_{j=1}^{n} D_j^T W_j D_j \tag{10.13}$$

and

$$S(\theta) = \sum\nolimits_{j=1}^{n} D_j^T W_j (Y_j - \mu_j) \tag{10.14}$$

where $W_j = \psi_j^{-1}$. Therefore, the increment (adjustment) of the parameter can be estimated using the following iteratively reweighted least squares approach,

$$\Delta\theta = \left[\sum\nolimits_{j=1}^{n} D_j^T W_j D_j\right]^{-1} \left[\sum\nolimits_{j=1}^{n} D_j^T W_j (Y_j - \mu_j)\right] \tag{10.15}$$

where D_j is a $(p+1) \times (p+q+r)$ matrix for the first partial derivatives of μ_j with respect to the parameters and $W_j = V_j^- = \psi_j^{-1}$ is the weight matrix. Matrix D_j can be partitioned into three blocks,

$$D_j = \frac{\partial \mu_j}{\partial \theta^T} = \left[\, \frac{\partial \mu_j}{\partial \alpha^T} \; \frac{\partial \mu_j}{\partial \beta^T} \; \frac{\partial \mu_j}{\partial \gamma^T} \,\right] \tag{10.16}$$

The first block $\partial\mu_j/\partial\alpha^T = \{\partial\mu_{jk}/\partial\alpha_l\}$ is a $(p+1) \times p$ matrix with

$$\frac{\partial \mu_{jk}}{\partial \alpha_{k-1}} = -\phi(\alpha_{k-1} + X_j\beta + Z_j\gamma)$$
$$\frac{\partial \mu_{jk}}{\partial \alpha_k} = \phi(\alpha_k + X_j\beta + Z_j\gamma)$$
$$\frac{\partial \mu_{jk}}{\partial \alpha_l} = 0,\ \forall l \neq \{k-1\ , k\} \tag{10.17}$$

The second block $\partial \mu_j / \partial \beta^T = \{\partial \mu_{jk} / \partial \beta\}$ is a $(p+1) \times q$ matrix with

$$\frac{\partial \mu_{jk}}{\partial \beta} = X_j^T \left[\phi(\alpha_k + X_j\beta + Z_j\gamma) - \phi(\alpha_{k-1} + X_j\beta + Z_j\gamma)\right] \tag{10.18}$$

The third block $\partial \mu_j / \partial \gamma^T = \{\partial \mu_{jk} / \partial \gamma\}$ is a $(p+1) \times r$ matrix with

$$\frac{\partial \mu_{jk}}{\partial \gamma} = Z_j^T \left[\phi(\alpha_k + X_j\beta + Z_j\gamma) - \phi(\alpha_{k-1} + X_j\beta + Z_j\gamma)\right] \tag{10.19}$$

In all the above partial derivatives, the range of k is $k = 1, \ldots, p+1$. The sequence of parameter values during the iteration process converges to a local maximum likelihood estimate, denoted by $\hat{\theta}$. The variance-covariance matrix of $\hat{\theta}$ is approximately equal to $\text{var}(\hat{\theta}) = I^{-1}(\hat{\theta})$, which is a by-product of the iteration process. Here, we are actually dealing with a situation where the QTL overlaps with a fully informative marker because observed marker genotypes represent the genotypes of the disease locus. If the QTL of interest does not overlap with any markers, the genotype of the QTL is not observable, i.e., Z_j is missing. The classical GLM does not apply directly to such a situation. The missing value Z_j still has some information due to linkage with some markers. Again, we use an $\mathtt{F}_2$ population as an example to show how to handle the missing value of Z_j. The ML estimation of parameters under the homogeneous variance model is obtained simply by substituting Z_j with the conditional expectation of Z_j given flanking marker information. Let

$$p_j(2-g) = \Pr(Z_j = H_g | \text{marker}), \forall g = 1, 2, 3 \tag{10.20}$$

be the conditional probability of the QTL genotype given marker information, where the marker information can be either drawn from two flanking markers (interval mapping, Lander and Botstein (1989)) or multiple markers (multipoint analysis, Jiang and Zeng (1997)). Note that $p_j(2-g)$ is not p_j multiplied by $(2-g)$; rather, it is a notation for the probabilities of the three genotypes. For $g = 1, 2, 3$, we have $p_j(-1)$, $p_j(0)$ and $p_j(+1)$, respectively, where $p_j(-1)$ etc. are defined early in Chapter 9. Vector H_g for $g = 1, 2, 3$ are also defined in Chapter 9 as genotype indicator variables.

Using marker information, we can calculate the expectation of Z_j , which is

$$U_j = E(Z_j) = \sum_{g=1}^{3} p_j(2-g) H_g \tag{10.21}$$

The method is called ML under the homogeneous residual variance because when we substitute Z_j by U_j, the residual error variance is no longer equal to unity; rather it is inflated and the inflation varies across individuals. However, the homogeneous variance model here assumed the residual variance is constant across individuals. This method is the ordinal trait analogy of the Haley and Knott's (1992) method of QTL mapping.

10.3 ML under heterogeneous variance

The homogeneous variance model is only a first moment approximation because the uncertainty of the estimated Z_j has been ignored. Let

$$\Sigma_j = \text{var}(Z_j) = \sum_{g=1}^{3} p_j(2-g)H_g^T H_g - U_j^T U_j \tag{10.22}$$

be the conditional covariance matrix for Z_j . Note that model (10.2) with Z_j substituted by U_j is

$$\mu_{jk} = E(Y_{jk}) = \Phi(\alpha_k + X_j\beta + U_j\gamma) - \Phi(\alpha_{k-1} + X_j\beta + U_j\gamma) \tag{10.23}$$

An underlying assumption for this probit model is that the residual error variance for the "underlying liability" of the disease trait is unity across individuals. Once U_j is used in place of Z_j , the residual error variance becomes

$$\sigma_j^2 = \gamma^T \Sigma_j \gamma + 1 \tag{10.24}$$

This is an inflated variance and it is heterogeneous across individuals. In order to apply the probit model, we need to rescale the model effects as follows (Xu and Hu, 2010),

$$\mu_{jk} = \Phi\left[\tfrac{1}{\sigma_j}(\alpha_k + X_j\beta + U_j\gamma)\right] - \Phi\left[\tfrac{1}{\sigma_j}(\alpha_{k-1} + X_j\beta + U_j\gamma)\right] \tag{10.25}$$

This modification leads to a change in the partial derivatives of μ_j with respect to the parameters. Corresponding changes in the derivatives are given below.

$$\begin{aligned}
\frac{\partial \mu_{jk}}{\partial \alpha_{k-1}} &= -\tfrac{1}{\sigma_j}\phi\left[\tfrac{1}{\sigma_j}(\alpha_{k-1} + X_j\beta + U_j\gamma)\right] \\
\frac{\partial \mu_{jk}}{\partial \alpha_k} &= \tfrac{1}{\sigma_j}\phi\left[\tfrac{1}{\sigma_j}(\alpha_k + X_j\beta + U_j\gamma)\right] \\
\frac{\partial \mu_{jk}}{\partial \alpha_l} &= 0,\ \forall l \neq \{k-1\ , k\}
\end{aligned} \tag{10.26}$$

$$\begin{aligned}
\frac{\partial \mu_{jk}}{\partial \beta} =& \tfrac{1}{\sigma_j}\phi\left[\tfrac{1}{\sigma_j}(\alpha_k + X_j\beta + U_j\gamma)\right] X_j^T \\
&- \tfrac{1}{\sigma_j}\phi\left[\tfrac{1}{\sigma_j}(\alpha_{k-1} + X_j\beta + U_j\gamma)\right] X_j^T
\end{aligned} \tag{10.27}$$

and

$$\frac{\partial \mu_{jk}}{\partial \gamma} = \frac{1}{\sigma_j}\phi\left[\frac{1}{\sigma_j}(\alpha_k + X_j\beta + U_j\gamma)\right]\left[U_j^T - \frac{1}{\sigma_j^2}(\alpha_k + X_j\beta + U_j\gamma)\Sigma_j\gamma\right] - \frac{1}{\sigma_j}\phi\left[\frac{1}{\sigma_j}(\alpha_{k-1} + X_j\beta + U_j\gamma)\right]\left[U_j^T - \frac{1}{\sigma_j^2}(\alpha_{k-1} + X_j\beta + U_j\gamma)\Sigma_j\gamma\right] \tag{10.28}$$

The iteration formula remains the same as (10.11) except that the modified weight and partial derivatives are used under the heterogeneous residual variance model.

10.4 ML under mixture distribution

The mixture model approach defines genotype specific expectation, variance matrix and all derivatives for each individual. Let

$$\mu_{jk}(g) = E(Y_{jk}) = \Phi(\alpha_k + X_j\beta + H_g\gamma) - \Phi(\alpha_{k-1} + X_j\beta + H_g\gamma) \tag{10.29}$$

be the expectation of Y_{jk} if j takes the gth genotype for $g = 1, 2, 3$. The corresponding variance-covariance matrix is

$$V_j(g) = \psi_j(g) - \mu_j(g)\mu_j^T(g) \tag{10.30}$$

where $\psi_j(g) = \text{diag}[\mu_j(g)]$. Let $D_j(g)$ be the partial derivatives of the expectation with respect to the parameters. The corresponding values of $D_j(g)$ are

$$\begin{aligned}\frac{\partial \mu_{jk}(g)}{\partial \alpha_{k-1}} &= -\phi(\alpha_{k-1} + X_j\beta + H_g\gamma)\\ \frac{\partial \mu_{jk}(g)}{\partial \alpha_k} &= \phi(\alpha_k + X_j\beta + H_g\gamma)\\ \frac{\partial \mu_{jk}(g)}{\partial \alpha_l} &= 0,\ \forall l \neq \{k-1\ , k\}\end{aligned} \tag{10.31}$$

$$\frac{\partial \mu_{jk}(g)}{\partial \beta} = X_j^T\left[\phi(\alpha_k + X_j\beta + H_g\gamma) - \phi(\alpha_{k-1} + X_j\beta + H_g\gamma)\right] \tag{10.32}$$

and

$$\frac{\partial \mu_{jk}(g)}{\partial \gamma} = H_g^T\left[\phi(\alpha_k + X_j\beta + H_g\gamma) - \phi(\alpha_{k-1} + X_j\beta + H_g\gamma)\right] \tag{10.33}$$

Let us define the posterior probability of QTL genotype after incorporating the disease phenotype for individual j as

$$p_j^*(2-g) = \frac{p_j(2-g)Y_j^T\mu_j(g)}{\sum_{g'=1}^{3} p_j(2-g')Y_j^T\mu_j(g')} \tag{10.34}$$

The increment for parameter updating under the mixture model is

$$\Delta\theta = \left[\sum_{j=1}^{n} E\left(D_j^T W_j D_j\right)\right]^{-1} \left[\sum_{j=1}^{n} E\left(D_j^T W_j (Y_j - \mu_j)\right)\right] \tag{10.35}$$

where

$$E\left(D_j^T W_j D_j\right) = \sum_{g=1}^{3} p_j^*(g) D_j^T(g) W_j(g) D_j(g) \tag{10.36}$$

$$E\left(D_j^T W_j (Y_j - \mu_j)\right) = \sum_{g=1}^{3} p_j^*(2-g) D_j^T(g) W_j(g)(Y_j - \mu_j(g)) \tag{10.37}$$

and

$$W_j(g) = \psi_j^{-1}(g) \tag{10.38}$$

This is actually an EM algorithm where calculating the posterior probabilities of QTL genotype and using the posterior probabilities to calculate $E\left(D_j^T W_j D_j\right)$ and $E\left(D_j^T W_j (Y_j - \mu_j)\right)$ constitute the E-step and calculating the increment of the parameter using the weighted least square formula makes up the M-step. A problem with this EM algorithm is that $\text{var}(\hat{\theta})$ is not a by-product of the iteration process. For simplicity, if the markers are sufficient close to the trait locus of interest, we can use

$$\text{var}(\hat{\theta}) \approx \left[\sum_{j=1}^{n} E\left(D_j^T W_j D_j\right)\right]^{-1} \tag{10.39}$$

to approximate the covariance matrix of estimated parameters. This is an underestimated variance matrix. A more precise method to calculate $\text{var}(\hat{\theta})$ is to adjust the above equation by the information loss due to uncertainty of the QTL genotype. Let

$$S(\hat{\theta}|Z) = \sum_{j=1}^{n} D_j^T W_j (Y_j - \mu_j) \tag{10.40}$$

be the score vector as if Z were observed. Louis (1982) showed that the information loss is due to the variance-covariance matrix of the score vector, which is

$$\text{var}\left[S(\hat{\theta}|Z)\right] = \sum_{j=1}^{n} \text{var}\left[D_j^T W_j (Y_j - \mu_j)\right] \tag{10.41}$$

The variance is taken with respect to the missing value Z using the posterior probability of QTL genotype. The information matrix after adjusting for the information loss is

$$I(\hat{\theta}) = \sum_{j=1}^{n} E\left(D_j^T W_j D_j\right) - \sum_{j=1}^{n} \text{var}\left[D_j^T W_j (Y_j - \mu_j)\right] \tag{10.42}$$

The variance-covariance matrix for the estimated parameters is then approximated by $\text{var}(\hat{\theta}) = I^{-1}(\hat{\theta})$. Details of $\texttt{var}[D_j^T W_j (Y_j - \mu_j)]$ are given by Xu and Hu (2010).

10.5 ML via the EM algorithm

The EM algorithm to be introduced here is different from the EM under the mixture model described in the previous section. We now use a liability model (Xu et al., 2003) to derive the EM algorithm. Xu et al. (2003) hypothesizes that there is an underlying liability that controls the observed phenotype. The liability is a continuous variable and has exactly the same behavior as a quantitative trait. The only difference is that the liability is not observable

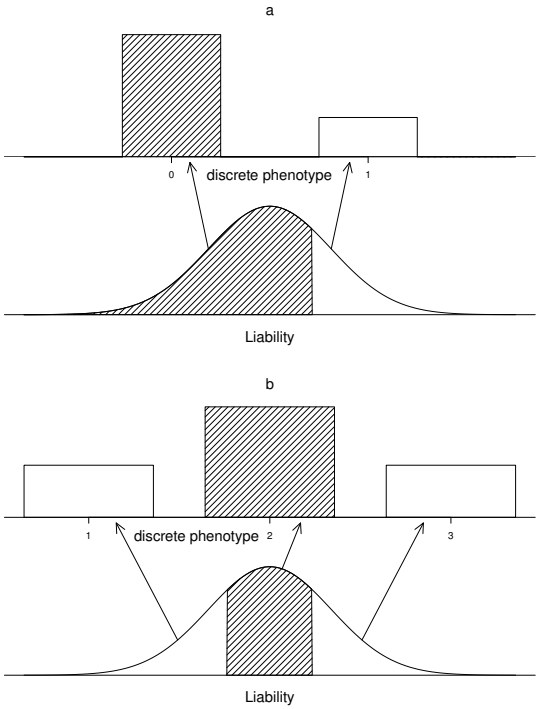

Fig. 10.1. Connection between the unobserved continuous liability and the observed discrete phenotype. The top panel shows the connection for an ordinal trait with two categories and the bottom panel shows the connection for an ordinal trait with three categories.

while the quantitative trait can be measured in experiments. The observed ordinal trait phenotype is connected with the liability by a series of thresholds, as demonstrated in Figure 10.1. In the generalized linear model under the mixture distribution, the EM algorithm treats the QTL genotype as missing value. Here, we treat the liability as missing value as well. Let y_j be the liability for the jth individual. This is different from $Y_j = \{Y_{jk}\}$, the multivariate representation of the ordered categorical phenotype in the generalized linear model. The liability can be described by the following linear model,

$$y_j = X_j\beta + Z_j\gamma + \varepsilon_j \tag{10.43}$$

where $\varepsilon_j \sim N(0, \sigma^2)$ is assumed. Under the liability model, σ^2 cannot be estimated and thus we set $\sigma^2 = 1$. This arbitrary scale will not affect the significance test because the estimated parameters $\theta = \{\alpha, \beta, \gamma\}$ are defined relative to σ^2 . The connection between y_j and the observed phenotype is

$$S_j = k \text{ , for } \alpha_{k-1} < y_j \leq \alpha_k \tag{10.44}$$

where $k = 1, \ldots, p+1$. The thresholds α do not appear in the linear model explicitly, but serve as converters from y_j to S_j. Xu et al. (2003) developed an EM algorithm for ordinal trait QTL mapping by using this liability model. They used a three step approach, where the first step is to estimate the non-QTL effects (β), the second step is to estimate the QTL effects (γ) and the third step is to estimate the thresholds (α). The method does not have a simple way to calculate the variance-covariance matrix of the estimated parameters. Xu and Xu (2006) extended the method using a multivariate version of the GLM. This method gives a way to calculate the variance-covariance matrix of the estimated parameters. Both methods (Xu et al., 2003; Xu and Xu, 2006) are quite complicated in the E-step. When the number of categories is two (the binary case), both methods can be simplified. This section will deal with the simplified binary trait QTL mapping where only one threshold is applied. In this case, the single threshold is set to zero so that it is not a parameter for estimation, and thus we only estimate β and γ. In the binary situation, $S_j = \{1, 2\}$ and

$$Y_{j1} = \begin{cases} 1 & \text{for } S_j = 1 \\ 0 & \text{for } S_j = 2 \end{cases} \tag{10.45}$$

and

$$Y_{j2} = \begin{cases} 0 & \text{for } S_j = 1 \\ 1 & \text{for } S_j = 2 \end{cases} \tag{10.46}$$

The liability model remains the same as that given in equation (10.43). The derivation of the EM algorithm starts with the complete-data situation. If both Z_j and y_j were observed, the ML estimates of β and γ would be

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T Z_j \\ \sum_{j=1}^{n} Z_j^T X_j & \sum_{j=1}^{n} Z_j^T Z_j \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} X_j^T y_j \\ \sum_{j=1}^{n} Z_j^T y_j \end{bmatrix} \tag{10.47}$$

This is simply the ordinary least squares estimates of the parameters. The EM algorithm takes advantage of this explicit solution in the maximization step. If we had observed y_j but still not been able to estimate Z_j , the maximization step of the EM algorithm would be

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T E(Z_j) \\ \sum_{j=1}^{n} E(Z_j^T) X_j & \sum_{j=1}^{n} E(Z_j^T Z_j) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} X_j^T y_j \\ \sum_{j=1}^{n} E(Z_j^T) y_j \end{bmatrix} \tag{10.48}$$

The problem here is that we observe neither Z_j nor y_j. Intuitively, the maximization step of the EM should be

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T E(Z_j) \\ \sum_{j=1}^{n} E(Z_j^T) X_j & \sum_{j=1}^{n} E(Z_j^T Z_j) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} X_j^T E(y_j) \\ \sum_{j=1}^{n} E(Z_j^T y_j) \end{bmatrix} \tag{10.49}$$

where the expectations are taken with respect to both Z_j and y_j using the posterior probabilities of QTL genotypes. We now present the method for calculating these expectation terms. We first address $E(Z_j)$ and $E(Z_j^T Z_j)$ using the posterior probabilities of the QTL genotypes.

$$p_j^*(2-g) = \frac{p_j(2-g)\left[\Phi(X_j\beta + H_g\gamma)\right]^{Y_{j1}}\left[1-\Phi(X_j\beta + H_g\gamma)\right]^{Y_{j2}}}{\sum_{g'=1}^{3} p_j(2-g')\left[\Phi(X_j\beta + H_{g'}\gamma)\right]^{Y_{j1}}\left[1-\Phi(X_j\beta + H_{g'}\gamma)\right]^{Y_{j2}}} \tag{10.50}$$

Given the posterior probabilities, we have

$$E(Z_j) = \sum_{g=1}^{3} p_j^*(2-g) H_g \tag{10.51}$$

and

$$E(Z_j^T Z_j) = \sum_{g=1}^{3} p_j^*(2-g) H_g^T H_g \tag{10.52}$$

The expectations for terms that involve y_j can be expressed as

$$E(y_j) = \underset{Z}{E}\left[\underset{y}{E}(y_j|Z_j)\right] = \sum_{g=1}^{3} p_j^*(2-g)\, \underset{y}{E}(y_j|H_g) \tag{10.53}$$

and

$$E(Z_j^T y_j) = \underset{Z}{E}\left[Z_j^T\, \underset{y}{E}(y_j|Z_j)\right] = \sum_{g=1}^{3} p_j^*(2-g) H_g^T\, \underset{y}{E}(y_j|H_g) \tag{10.54}$$

where

$$\underset{y}{E}(y_j|H_g) = X_j\beta + H_g\gamma + \frac{(Y_{j2} - Y_{j1})\phi(X_j\beta + H_g\gamma)}{[\Phi(X_j\beta + H_g\gamma)]^{Y_{j1}}\,[1 - \Phi(X_j\beta + H_g\gamma)]^{Y_{j2}}} \quad (10.55)$$

Therefore, the EM algorithm can be summarized as

1. Initialize parameters $\theta^{(0)} = \{\beta^{(0)}, \gamma^{(0)}\}$
2. Calculate $E(Z_j)$, $E(Z_j^T Z_j)$, $E(y_j)$ and $E(Z_j^T y_j)$
3. Update β and γ using equation (10.49)
4. Repeat Step 2 to Step 3 until convergence is reached

Once the EM algorithm converges, we obtain the estimated parameters and are ready to calculate the Louis (1982) information matrix. The variance-covariance matrix of the estimated parameters simply take the inverse of the information matrix. Let

$$H(\theta, Z, y) = -\begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T Z_j \\ \sum_{j=1}^{n} Z_j^T X_j & \sum_{j=1}^{n} Z_j^T Z_j \end{bmatrix} \quad (10.56)$$

be the Hessian matrix of the complete-data log likelihood function and and

$$S(\theta, Z, y) = \begin{bmatrix} \sum_{j=1}^{n} X_j^T (y_j - X_j\beta - Z_j\gamma) \\ \sum_{j=1}^{n} Z_j^T (y_j - X_j\beta - Z_j\gamma) \end{bmatrix} \quad (10.57)$$

be the score vector of the complete-data log likelihood function. The Louis information matrix is

$$I(\theta) = -E\left[H(\theta, Z, y)\right] - E\left[S(\theta, Z, y)S^T(\theta, Z, y)\right] \quad (10.58)$$

where the expectations are taken with respect to the missing values of Z and y. Note that

$$\mathrm{var}\left[S(\theta, Z, y)\right] = E\left[S(\theta, Z, y)S^T(\theta, Z, y)\right] - E\left[S(\theta, Z, y)\right] E\left[S^T(\theta, Z, y)\right] \quad (10.59)$$

and $E\left[S(\theta, Z, y)\right] = 0$ at $\theta = \hat{\theta}$. This leads to

$$E\left[S(\theta, Z, y)S^T(\theta, Z, y)\right] = \mathrm{var}\left[S(\theta, Z, y)\right] \quad (10.60)$$

Therefore, the Louis information matrix is also expressed as

$$I(\theta) = -E\left[H(\theta, Z, y)\right] - \mathrm{var}\left[S(\theta, Z, y)\right] \quad (10.61)$$

The first term is easy to obtain, as shown below,

$$-E\left[H(\theta, Z, y)\right] = \begin{bmatrix} \sum_{j=1}^{n} X_j^T X_j & \sum_{j=1}^{n} X_j^T E(Z_j) \\ \sum_{j=1}^{n} E(Z_j^T) X_j & \sum_{j=1}^{n} E(Z_j^T Z_j) \end{bmatrix} \quad (10.62)$$

The second term can be expressed as

$$\text{var}\left[S(\theta, Z, y)\right] = \sum_{j=1}^{n} \text{var}\left[S_j(\theta, Z, y)\right] \tag{10.63}$$

where

$$S_j(\theta, Z, y) = \begin{bmatrix} X_j^T(y_j - X_j\beta - Z_j\gamma) \\ Z_j^T(y_j - X_j\beta - Z_j\gamma) \end{bmatrix} \tag{10.64}$$

Explicit form of $\text{var}\left[S_j(\theta, Z, y)\right]$ can be derived. This matrix is a 2×2 block matrix, denoted by

$$\text{var}\left[S_j(\theta, Z, y)\right] = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{10.65}$$

we now provide detailed expressions of the blocks.

$$\begin{aligned} \Sigma_{11} &= E\left[X_j^T \text{var}(y_j - X_j\beta - Z_j\gamma)X_j\right] \\ \Sigma_{22} &= E\left[Z_j^T \text{var}(y_j - X_j\beta - Z_j\gamma)Z_j\right] \\ \Sigma_{12} &= E\left[X_j^T \text{var}(y_j - X_j\beta - Z_j\gamma)Z_j\right] \end{aligned} \tag{10.66}$$

where $\text{var}(y_j - X_j\beta - Z_j\gamma)$ is the variance of a truncated normal variable (the truncation point being zero) conditional on $Y_j = \{Y_{j1}, Y_{j2}\}$ and Z_j. Let

$$\varphi(Z_j) = \text{var}(y_j - X_j\beta - \text{Z}_j\gamma) \tag{10.67}$$

be the short notation for the variance of the truncated normal variable. With some manipulation on Cohen (1991) formula, we get

$$\varphi(Z_j) = 1 - \psi(X_j\beta + Z_j\gamma)\left[\psi(X_j\beta + Z_j\gamma) - (Y_{j1} - Y_{j2})(X_j\beta + Z_j\gamma)\right] \tag{10.68}$$

where

$$\psi(X_j\beta + Z_j\gamma) = \frac{\phi(X_j\beta + Z_j\gamma)}{\left[1 - \Phi(X_j\beta + Z_j\gamma)\right]^{Y_{j1}}\left[\Phi(X_j\beta + Z_j\gamma)\right]^{Y_{j2}}} \tag{10.69}$$

Therefore,

$$\begin{aligned} \Sigma_{11} &= \sum_{g=1}^{3} p_j^*(2-g)X_j^T\varphi(H_g)X_j \\ \Sigma_{12} &= \sum_{g=1}^{3} p_j^*(2-g)X_j^T\varphi(H_g)H_g \\ \Sigma_{22} &= \sum_{g=1}^{3} p_j^*(2-g)H_g^T\varphi(H_g)H_g \end{aligned} \tag{10.70}$$

Further manipulation on the information matrix, we get

$$I(\theta) = \begin{bmatrix} \sum_{j=1}^{n} E\left[X_j^T\left(1 - \varphi(Z_j)\right)X_j\right], & \sum_{j=1}^{n} E\left[X_j^T\left(1 - \varphi(Z_j)\right)Z_j\right] \\ \sum_{j=1}^{n} E\left[Z_j^T\left(1 - \varphi(Z_j)\right)X_j\right], & \sum_{j=1}^{n} E\left[Z_j^T\left(1 - \varphi(Z_j)\right)Z_j\right] \end{bmatrix} \tag{10.71}$$

which is a 2×2 matrix.

Xu and Xu (2003) proposed an alternative method to calculate the Louis information matrix via Monte Carlo simulations. The method does not involve the above complicated derivation; instead, it simply simulates the QTL genotype (Z_j) using the posterior distribution for each individual and the liability (y_j) conditional on the genotype using the truncated normal distribution for the individual. The method directly uses the following information matrix,

$$I(\theta) = -E\left[H(\theta, Z, y)\right] - E\left[S(\theta, Z, y)S^T(\theta, Z, y)\right] \tag{10.72}$$

with $E\left[S(\theta, Z, y)S^T(\theta, Z, y)\right]$ obtained via Monte Carlo simulations. Let $Z^{(t)}$ and $y^{(t)}$ be simulated Z and y at the tth sample so that $S(\theta, Z^{(t)}, y^{(t)})$ is the score vector given $Z^{(t)}$, $y^{(t)}$ and $\theta = \hat{\theta}$. The Monte Carlo approximation of $E\left[S(\theta, Z, y)S^T(\theta, Z, y)\right]$ is

$$E\left[S(\theta, Z, y)S^T(\theta, Z, y)\right] \approx \frac{1}{T}\sum_{t=1}^{T} S(\theta, Z^{(t)}, y^{(t)})S^T(\theta, Z^{(t)}, y^{(t)}) \tag{10.73}$$

where T is a large number, say 10000. The liability for the jth individual, y_j , is simulated from a truncated normal distribution. We adopt the inverse transformation method that has an acceptance rate of 100% (Rubinstein, 1981). With this method, we first defined

$$v = 1 - \Phi(X_j\beta + Z_j\gamma) \tag{10.74}$$

and then simulated a variable u from $U(0, 1)$. Finally, we took the inverse function of the standardized normal distribution to obtain

$$y_j = Y_{j1}\Phi^{-1}(u\,v) + Y_{j2}\Phi^{-1}[v + u(1 - v)] \tag{10.75}$$

Intrinsic functions for both $\Phi(.)$ and $\Phi^{-1}(.)$ are available in many computer software packages. For example, in the SAS package (SAS Institute, 2008a), $\Phi(x)$ is coded as $\Phi(x) = \text{probnorm}(x)$ and $\Phi^{-1}(u)$ is coded as $\Phi^{-1}(u) = \text{probit}(u)$. The Monte Carlo approximation is time consuming so that we cannot calculate the information matrix for every point of the genome scanned. Instead, we only calculate the information matrix at the points where evidences of QTL are strong.

10.6 Logistic analysis

Similar to the probit link function, we may also use the logit link function to perform the generalized linear model analysis. Let

$$\zeta_{jk} = \frac{\exp(\alpha_k + X_j\beta + Z_j\gamma)}{1 + \exp(\alpha_k + X_j\beta + Z_j\gamma)} \tag{10.76}$$

be the cumulative distribution function of $\alpha_k + X_j\beta + Z_j\gamma$. Under the logistic model, the mean of Y_{jk} is modelled by

$$\mu_{jk} = E(Y_{jk}) = \zeta_{jk} - \zeta_{j(k-1)} \tag{10.77}$$

The logistic model for the binary data is

$$\mu_{jk} = \begin{cases} \zeta_{j1} & \text{for } k = 1 \\ 1 - \zeta_{j1} & \text{for } k = 2 \end{cases} \tag{10.78}$$

From $\mu_{j1} = \zeta_{j1}$ we obtain

$$\text{logit}(\mu_{j1}) = \ln\left(\frac{\mu_{j1}}{1-\mu_{j1}}\right) = \alpha_1 + X_j\beta + Z_j\gamma \tag{10.79}$$

Both the probit and logit transformations of the expectation of Y_{j1} lead to a linear model. Note that the linear model obtained here only shows the property of the transformation. In the actual theory development and data analysis, the linear transformations in equations (10.6) and (10.79) are never used. By showing the linear transformations may potentially cause confusion to students because, by intuition, they may try to transform the ordinal data (Y_{jk}) first and then conduct the usual linear regression on the transformed data, which is not appropriate and certainly not the intension of the GLM developers. The maximum likelihood analysis under the homogeneous variance, heterogeneous variance and mixture model and the EM algorithm described previously in the probit analysis apply to the logistic analysis. We only show the logistic analysis under the homogeneous variance model as an example. Note that under this model, we only need to substitute Z_j by U_j to define the expectation, i.e.,

$$\zeta_{jk} = \frac{\exp(\alpha_k + X_j\beta + U_j\gamma)}{1 + \exp(\alpha_k + X_j\beta + U_j\gamma)} \tag{10.80}$$

and

$$\mu_{jk} = E(Y_{jk}) = \zeta_{jk} - \zeta_{j(k-1)} \tag{10.81}$$

Once μ_j is defined, the weight W_j is also defined. The only items left is D_j, which is

$$D_j = \frac{\partial \mu_j}{\partial \theta^T} = \left[\, \frac{\partial \mu_j}{\partial \alpha^T} \; \frac{\partial \mu_j}{\partial \beta^T} \; \frac{\partial \mu_j}{\partial \gamma^T} \,\right] \tag{10.82}$$

The first block $\partial\mu_j/\partial\alpha^T = \{\partial\mu_{jk}/\partial\alpha_l\}$ is a $(p+1)\times p$ matrix with

$$\begin{aligned} \frac{\partial \mu_{jk}}{\partial \alpha_{k-1}} &= -\zeta_{j(k-1)}(1-\zeta_{j(k-1)}) \\ \frac{\partial \mu_{jk}}{\partial \alpha_k} &= \zeta_{jk}(1-\zeta_{jk}) \\ \frac{\partial \mu_{jk}}{\partial \alpha_l} &= 0, \; \forall l \neq \{k-1\,, k\} \end{aligned} \tag{10.83}$$

The second block $\partial\mu_j/\partial\beta^T = \{\partial\mu_{jk}/\partial\beta\}$ is a $(p+1)\times q$ matrix with

$$\frac{\partial\mu_{jk}}{\partial\beta} = X_j^T\zeta_{jk}(1-\zeta_{jk}) - X_j^T\zeta_{j(k-1)}(1-\zeta_{j(k-1)}) \tag{10.84}$$

The third block $\partial\mu_j/\partial\gamma^T = \{\partial\mu_{jk}/\partial\gamma\}$ is a $(p+1)\times r$ matrix with

$$\frac{\partial\mu_{jk}}{\partial\gamma} = U_j^T\zeta_{jk}(1-\zeta_{jk}) - U_j^T\zeta_{j(k-1)}(1-\zeta_{j(k-1)}) \tag{10.85}$$

In the above partial derivatives, the range of k is $k = 1,\ldots,p+1$.

10.7 Example

The experiment was conducted by Dou et al. (2009). A female sterile line of wheat XND126 and an elite wheat cultivar Gaocheng8901 with normal fertility were crossed for genetic analysis of female sterility measured as the number of seeded spikelets per plant. The parents, their $\mathtt{F}_1$ and $\mathtt{F}_2$ progeny were planted at the Huaian experimental station in China for the 2006-2007 growing season under the normal autumn sowing condition. The mapping

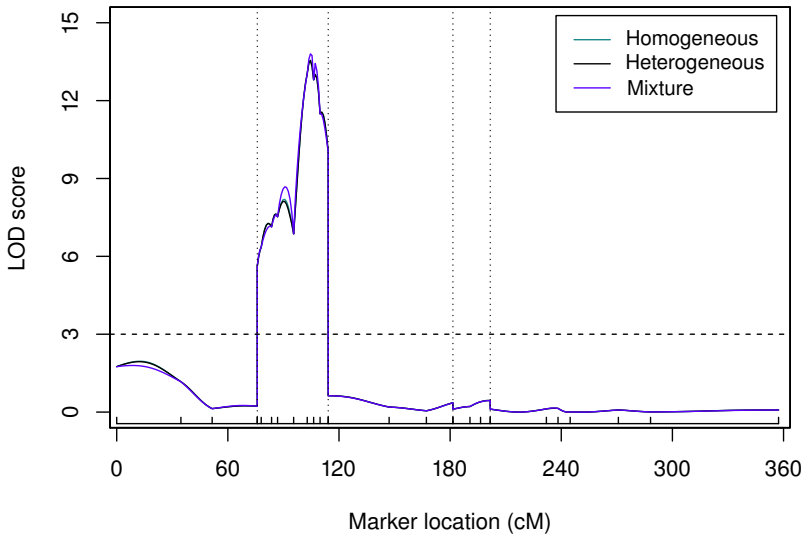

Fig. 10.2. The LOD test statistic profiles for three methods of interval mapping (HOMOGENEOUS, HETEROGENEOUS and MIXTURE). The data were obtained from Dou et al. (2009). The trait investigated is the female fertility of wheat measured as a binary trait (seed presence and absence). The five chromosomes (part of the wheat genome) are separated by the vertical dotted lines. The unevenly distributed black ticks on the horizontal axis indicate the marker locations.

population was an F_2 family consisting of 243 individual plants. About 84% of the F_2 progeny had seeded splikelets and the remaining 16% plants did not have any seeds at all. Among the plants with seeded spikelets, the number of seeded spikelets varied from one to as many as 31. The phenotype is the count data point and can be modeled using the Poisson distribution. The phenotype can also be treated as a binary data point and analyzed using the Bernoulli distribution. In this example, we treated the phenotype as a binary data (seed presence and absence) and analyzed it using the Bernoulli distribution. A total of 28 SSR markers were used in this experiment. These markers covered five chromosomes of the wheat genome with an average genome marker density of 15.5cM per marker interval. The five chromosomes are only part of the wheat genome. These chromosomes were scanned for QTL of the binary data. Let A_1 and A_2 be the alleles carried by Gaocheng8901 and XDN128, respectively. Let A_1A_1, A_1A_2 and A_2A_2 be the three genotypes for the QTL of interest. The genotype is numerically coded as 1, 0 and -1, respectively for the three genotypes. The genome was scanned with 1 cM increment. All the three methods described in this chapter were used for the interval mapping. They are the homogeneous variance model (HOMOGENEOUS), the heterogeneous variance model (HETEROGENEOUS) and the mixture model (MIXTURE). The LOD score profiles are depicted in Figure 10.2. When LOD = 3 is used as the threshold value, all three methods detected two major QTL on chromosome 2. The LOD score for the mixture model appears to be higher than the other two models, but the difference is very small and can be safely ignored.

The estimated QTL effect profiles are given in Figure 10.3. Again the three methods are almost the same for the estimated QTL effects except that the mixture model and the heterogeneous model give slightly higher estimates than the homogeneous model. In practice, we recommend the heterogeneous model because it produces almost the same result as the mixture model but with much less computing time than the mixture model.

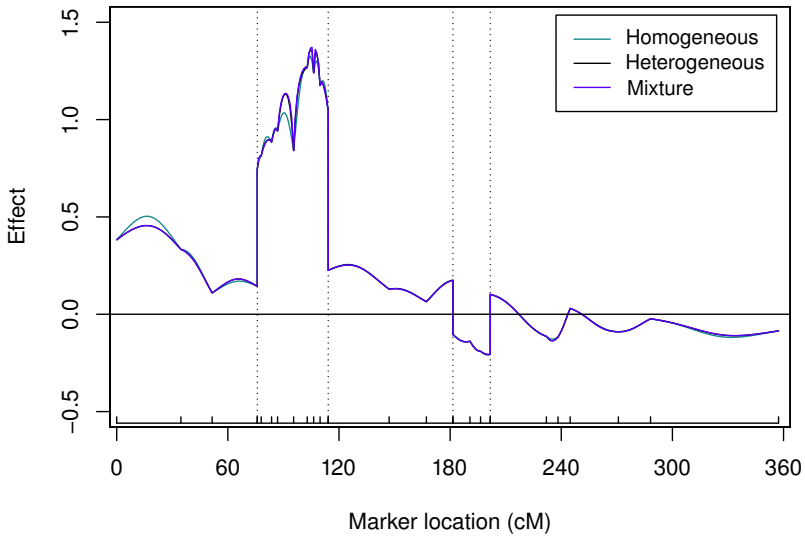

Fig. 10.3. The QTL effect profiles for three methods of interval mapping (HOMOGENEOUS, HETEROGENEOUS and MIXTURE). The data were obtained from Dou et al. (2009). The trait investigated is the female fertility of wheat measured as a binary trait (seed presence and absence). The five chromosomes (part of the wheat genome) are separated by the vertical dotted lines. The unevenly distributed black ticks on the horizontal axis indicate the marker locations.