

### Chapter 3. Statistical methods for QTL detection and parameter estimation.

#### 3.1 Introduction

The main parameters that will be considered are means and variances of QTL genotypes, and recombination frequencies between the genetic markers and the QTL. From the previous chapter it should already be clear that estimation of QTL parameters is not trivial.

The first problem encountered is that in nearly all cases of interest the variance due to the segregating QTL will be only a small fraction of the total variance. Also, since we will generally be dealing with field data there will usually be confounding "nuisance" variables, such as herd or block. Parameter estimates may be biased if these factors are not included in the analysis model. As considered in detail in the previous chapter, linkage between the genetic markers and QTL will be incomplete. Thus, recombination frequency must be included in the analysis, and consequently the analysis model will not be a linear function of the parameters. In the analysis of segregating populations it is also necessary to account for polygenic variance not linked to the genetic markers. Finally, most analyses will include

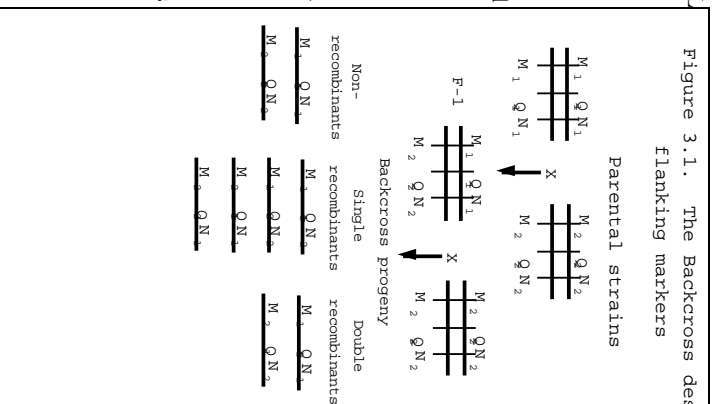


Figure 3.1. The Backcross design. Flanking markers

```
proc nlin ;
R = 0.3 ;
if m = 1 and n = 1 then p = r1*r2/(1-R) ;
if m = 1 and n = 2 then p = r1*(1-r2)/R ;
if m = 2 and n = 2 then p = (1-r1*r2)/(1-R) ;
if m = 2 and n = 1 then p = r2*(1-r1)/R ;
r2 = (r-r1)/(1-2*r1) ;
parameters mu1 = -1 mu2 = .1 r1 = 0 to 0.3 by 0.05 ;
model trait = mu1 + (mu2-mu1)*p ;
bounds -0.3 < mu1 < 0.3 ;
bounds -0.3 < mu2 < 0.3 ;
bounds 0 < r1 < 0.3 ;
run ;
```

Figure 3.2. The SAS code for nonlinear regression interval mapping for the BC design

multiple traits and markers, which creates further complications.

In Section 3.2 we will consider the desired properties of estimators. In Section 3.3 we will describe least squared estimation, with focus on nonlinear models. In Section 3.4 we will describe single parameter maximum likelihood estimation, which is the main techniques that has been applied to estimate QTL parameters. In Section 3.5 we will discuss the principles of ML multiparameter estimation. In Section 3.6 we will describe ML models for QTL parameter estimation in crosses between inbred lines. In Section 3.7 we describe methods to maximize likelihood functions. In Section 3.8 we will consider more complicated models that are amenable to solution by ML. In Section 3.9 we will briefly discuss Bayesian QTL parameter estimation including Gibbs sampling. Finally in Section 3.10 we will consider methods that have been proposed to deal with repeat records, nuisance effects, and polygenic variance.

#### 3.2 Desired properties of QTL parameter estimates

For the general question of parameter estimation, there are four main desired properties of estimators: unbiasedness, minimum estimation error variance, estimates within the parameter space, and consistency. For simple situations, it is possible to derive estimators with all of these properties, but for more complicated cases, it will not be possible to obtain estimates with all the desired properties, and there will be a question of trade-offs. We will now describe these properties in detail.

**Unbiasedness.** assume that is an estimator of  $\theta$ . Then is unbiased if  $E\theta = \theta$ , that is the expectation of the estimator is equal to the parameter value. As an example, in estimating the variance based on the sample mean we divide by  $n-1$ , where  $n$  is the sample size. If instead we divide by  $n$  then the estimator will be a biased estimate of the variance.

**Minimum estimation error variance** is defined as the value of  $\theta$  for which:  $E[(\theta - \theta)^2]$  is minimal. This property is the basis of least squares estimation. The estimator with minimum estimation error variance is also called the "best" estimate.

**Estimates within the parameter space.** Although the requirement of estimates within the parameter space may appear trivial, this is often not the case. In many situations it is not possible to obtain an estimate that is both unbiased and within the parameter space. Simple examples of estimators outside the parameter space are negative variance component estimates, or estimators for recombination frequency less than zero or greater than 0.5. Maximum likelihood estimates are always within the parameter space by definition.

**Consistency:** an estimator,  $\hat{\theta}$ , is considered "consistent" if  $\hat{\theta}$  tends to  $\theta$  as the sample size tends toward infinity. An estimator can be consistent even if it is biased. Consider the example given above of estimating the variance of a sample. If we divide by  $n$  instead of  $n-1$ , the estimator is biased, but consistent; because as  $n$  tends to infinity,  $n$  tends to  $n-1$ . Although this property also appears trivial, it is especially important for QTL detection. Consider the example in the previous chapter for a BC or F-2 design. The contrasts between the marker genotype classes is not a consistent estimator of the contrast between the QTL genotype means. This estimate is biased by  $1-2r$  regardless of the sample size.

### 3.3 Least squares estimation of QTL parameters

We will use matrix notation to briefly describe least squares estimation. Vectors and matrices will be denoted in **bold type**. Vectors will be denoted by Greek symbols or lower case letters, matrices by uppercase Latin letters. The transpose of a vector or matrix will be denoted by an apostrophe.

Assume that there is a series of observations for some variable,  $y$ , which we wish to model in terms of other variables for which data is also available. We will denote  $y$  as the "dependent variable" and the other variables as the "independent variables". The objective is to "explain" the dependent variable in terms of a series of parameter estimates linking the dependent variables to the independent variable. That is to derive a function of the independent variables that approximates the observations for  $y$ . Generally it will not be possible to completely explain  $y$  in terms of the dependent variables. The difference between the estimates of  $y$ , based on the independent variables and the parameter estimates is denoted the "error" or residual of the model.

Least squares estimation is based on deriving the parameter estimates that minimize the expectation of the sum of squared errors. Thus, by definition this method has minimum estimation error variance. Define  $\theta$  = vector of parameters,  $y$  = vector of observations, and  $e$  = vector of residuals.  $y$  will also be noted the dependent variable. In matrix form a completely general model can be written as follows:

$$y = f(\theta) + e \quad \{3.1\}$$

where  $f(\theta)$  is some function of  $\theta$ . The least squares solution,  $\hat{\theta}$ , is the vector that minimizes  $[y-f(\hat{\theta})]^2 = e^2$ . For a linear model, equation {3.1} can be written as follows:

$$y = X\theta + e \quad \{3.2\}$$

where,  $X$  is a matrix of coefficients of  $\theta$ . Effects in linear models can take one of two forms, class or continuous. Discrete effects such as a specific herd, block, or sex are denoted "class effects". Although the levels of these effects can be numbered, there is no relationship between the number of a specific herd and effect associated with it. For continuous effects a linear relationship is assumed between the value for the independent variable and the dependent variable. Each row of  $X$  corresponds to the coefficients of  $\theta$  for a specific record in  $y$ . For class effects the elements in  $X$  will be either zero or one. For continuous effects, each element in  $X$  corresponds to the observed value for the independent variable. For the linear model it is easy to demonstrate that the least square solution for  $\hat{\theta}$  can be derived from the normal equations:  $\theta = (X'X)^{-1}X'y$ . For a linear model the parameter estimates will also be unbiased, consistent, and within the parameter space. If  $y$  is not a linear function of  $\theta$ , then the least squares solution can generally not be derived analytically, although various

iterative methods have been developed. Only effects on the mean of  $y$  are included in the model, thus effects on the variance of  $y$  or higher order moments cannot not be estimated by least squares.

As demonstrated in the previous chapter, the genotype means and variances, and recombination frequencies cannot be described by a linear model of the trait values. Furthermore, as noted in the previous chapter, within the context of least squares estimation, genotype means and the recombination frequency between a QTL and a single marker are completely confounded and cannot be estimated separately. With two markers flanking a QTL it is possible to derive separate estimates of genotype means and recombination frequencies, but it is not possible to construct a linear model that accurately describes the relationship between the observations and the QTL parameters, as will be explained below.

The nonlinear least square method of QTL parameter estimation with two flanking markers was developed independently in 1992 by Haley and Knott and Martinez and Curnow. We will illustrate the method using the BC design, although the method has been adapted to most of the designs considered in the previous chapter. The BC design with two flanking markers is illustrated in Figure 3.1. For the BC progeny only the chromosome from the F-1 parent is shown. There are eight possible gametic haplotypes (including the QTL); two nonrecombinants, four single recombinants, and two double recombinants. Recombination frequency between the two markers,  $M$  and  $N$  will be denoted  $R$ . Recombination frequencies between  $M$  and  $Q$  and  $Q$  and  $N$  will be denoted  $r_1$  and  $r_2$ , respectively. Zero interference will be assumed, thus  $R = r_1 + r_2 - 2r_1r_2$ . The following model can then be defined:

$$Y_{ij} = \mu_i(1-p) + \mu_2p_i + e_{ij} \quad \{3.3\}$$

where  $Y_{ij}$  is the production record of the  $j^{\text{th}}$  individual with marker genotype  $i$ ,  $\mu_1$  is the mean for individuals with genotype  $Q_1Q_2$ ,  $\mu_2$  is the mean for individuals with genotype  $Q_2Q_2$ ,  $p_i$  is the probability that an individual with marker genotype  $i$  has genotype  $Q_2Q_2$ ,  $e_{ij}$  is the residual, and the other terms are as defined previously. This model can be simplified as follows:

$$Y_{ij} = \mu_1 + (\mu_2 - \mu_1)p_i + e_{ij} \quad \{3.3a\}$$

$p_i$  is a function of the recombination parameters, and can be estimated for each of the four marker haplotypes as follows:

$$P_{m1n1} = r_1r_2/(1-R) \quad \{3.4\}$$

$$P_{m1n2} = r_1(1-r_2)/R \quad \{3.5\}$$

$$P_{m2n1} = r_2(1-r_1)/R \quad \{3.6\}$$

$$P_{m2n2} = 1 - r_1r_2/(1-R) \quad \{3.7\}$$

If  $r_1$  was known, it would be possible to substitute these values into equation {3.3a}, and then solve as a simple linear regression, with  $\mu_1$  as the y-intercept and  $\mu_2 - \mu_1$  and the slope. Since  $r_1$  is not known, equation {3.3a} can be considered as four separate equations, one for each marker haplotype. Although assuming zero interference, it is possible to solve for  $r_2$  in terms of  $R$ , which is assumed known, and  $r_1$ , we are left with four equations which are complicated functions of the QTL means and  $r_1$ . Thus this model is nonlinear in  $r_1$ . The least squares solution for all three parameters will be the values that minimize the residual sum of squares. This model readily can be solved by a nonlinear least squares algorithm, such as PROC NLIN of SAS (SAS, 1988). The SAS code is given in Figure 3.2.

The main advantages of this method are that it can be performed by more statistical packages, and significance of the QTL effect can be tested by an F-test of the model mean squares against the residual mean squares, which is more familiar to most researchers than a likelihood ratio test. The disadvantage of this method is that it is applicable only in certain situations. It cannot be applied to estimate recombination frequency between a QTL and a single marker, used to estimate QTL variance effects, or applied to segregating populations. All of these questions have been addressed by ML (Weller, 1986; Bovenhuis and Weller, 1994).

### 3.4 Maximum likelihood estimation for a single parameter

Maximum likelihood (ML) is much more flexible than least squares estimation, but requires rather complex programming, except for models which can be analyzed by available software, such as program LE of BMDDP (Elkind et al., 1994). There are three steps in ML parameter estimation:

1. Defining the assumptions on which the statistical model is based. In addition to the "usual" assumptions listed above, zero crossing-over interference between the segments MQ and QN is assumed.
2. Constructing the likelihood function, which is the joint density of the observations conditional on the parameters.
3. Maximizing the likelihood function with respect to the parameters.

Since ML estimation is less familiar than least squares, the basic methodology for ML estimation of a single parameter will be illustrated using an example from a binomial distribution. Assume that from a sample of ten observations, three are "successes" and the other seven are "failures". We wish to derive the ML estimate (MLE) of  $p$ , the probability of success. This is done by writing the binomial probability of obtaining this result as a function of  $p$ :

$$L = \frac{10!}{p^3(1-p)^7} \quad \{3.8\}$$

3171

where  $L$  is the probability of obtaining this result, conditional on  $p$ .  $L$  is denoted the likelihood function. The MLE for  $p$  is that value of  $p$  which maximizes  $L$ . The MLE is computed by differentiating  $L$  with respect to  $p$ , and solving for  $p$  with this derivative set equal to zero. In practice it is usually easier to compute and differentiate the log of  $L$ . This is equivalent to differentiating  $L$ , since a function and its log will have maximum value for the same variable value. It is thus possible to derive the MLE of  $p$  as follows:

$$\text{Log } L = \log(10!) - \log(3!7!) + 3(\log p) + 7[\log(1-p)] \quad \{3.9\}$$

$$d(\text{Log } L)/dp = 3/p - 7/(1-p) = 0 \quad \{3.10\}$$

$$p = 3/10 \quad \{3.11\}$$

This is of course the proportion of successes derived in the sample. Thus, for this simple case, the MLE is the intuitive estimate value. From the above discussion, it should be clear why MLE must lie within the parameter space. A parameter estimate outside the parameter space will by definition have a likelihood of zero, and can therefore not be the MLE.

For a continuous distribution, the likelihood is computed as the statistical density of the distribution, conditional on the sample. Statistical density,  $f(y)$ , for a continuous variable,  $y$ , is defined as the ordinate of the distribution function for a given value of  $y$ . For example, assume that a sample was taken from a normal distribution. To obtain the MLE for the mean, it is necessary to compute the joint statistical density of the sample. For a single observation the likelihood will be:

$$L = \frac{e^{-(y-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \quad \{3.12\}$$

where  $\sigma$  is the standard deviation,  $e$  is the base for natural logarithms and is approximately equal to 2.72,  $\mu$  is the mean, and  $y$  is the variable value. For a sample of  $N$  observations, the likelihood will be the product of the likelihoods for each individual observation. As in the previous case, the MLE for  $\mu$  can be derived by taking the derivative of the log of the likelihood, with respect to the mean, and setting this function equal to zero. The derivative of log  $L$  for a sample from a normal distribution is computed as follows:

$$d(\text{Log } L)/d\mu = \sum (y_i - \mu) \quad \{3.13\}$$

Setting this function equal to zero, we find that the MLE of the population mean is the sample mean, which is again the intuitively correct result.

The MLE for the variance could be derived in the same manner, and would again yield the

intuitive result of the sample variance. Although in the two examples given so far, ML has been used to derive estimates that could have been derived by other methods, it will be demonstrated below that for more complicated problems, ML is the only method of estimation that can utilize all the available data.

### 3.5 Maximum likelihood multi-parameter estimation

ML can also be used to estimate several parameters simultaneously, for example, to estimate both the mean and variance in a normal distribution. In that case it is necessary to maximize the likelihood with respect to both parameters. This can be done by taking the partial derivatives of the log likelihood with respect to each parameter, and setting each derivative equal to zero. It is then necessary to solve a system of equations equal to the number of parameters being estimated. In general the likelihood function for estimation of  $m$  parameters,  $(\theta_1, \theta_2, \dots, \theta_m)$ , from a sample of  $n$  observations  $(Y_1, Y_2, \dots, Y_n)$  can be written as follows:

$$\begin{aligned} L &= p(Y_1, Y_2, \dots, Y_n | \theta_1, \theta_2, \dots, \theta_m) = \\ &= p(Y_1 | \theta_1, \theta_2, \dots, \theta_m) p(Y_2 | \theta_1, \theta_2, \dots, \theta_m) \dots p(Y_n | \theta_1, \theta_2, \dots, \theta_m) \quad \{3.14\} \\ &= \Pi p(Y_i | \theta_1, \theta_2, \dots, \theta_m) \end{aligned}$$

Where  $p(Y_i | \theta)$  represents the probability of obtaining  $y_i$ , conditional on  $\theta$ , and  $\Pi$  signifies the product  $p(Y_1 | \theta_1, \theta_2, \dots, \theta_m)$  from  $Y_1$  through  $Y_n$ . If the distribution is continuous, then  $p(Y_i | \theta)$  will be replaced by  $f(y_i | \theta)$ , i. e., the density of  $y_i$ , conditional on  $\theta$ . Thus any problem that can be phrased in terms of equation {3.15} can be solved by ML.

### 3.6 Prediction error variances for MLE

In addition to deriving parameter estimates, it is also important to know how accurate the estimates are. Generally the standard error of the estimate is used for this purpose. The square of the standard error is denoted the prediction error variance. The following equation can generally be used to derive the prediction error variance for MLE of a single parameter:

$$\text{Var}() = \frac{-1}{E[d^2(\log L)/d\theta^2]} \quad \{3.15\}$$

Where is the MLE of  $\theta$ , and  $E[d^2(\log L)/d\theta^2]$  is the expectation of the second derivative of  $L$  with respect to  $\theta$ . Equation {3.15} will be correct if the first derivative of  $\theta$  is a multiple of the difference between the true parameter value and its estimate. Otherwise the prediction error variance will be slightly greater than the right-hand-side of equation {3.15}. Under a

wide range of conditions, equation {3.15} will be asymptotically correct; that is, as the sample size increases the difference between the left-hand and right-hand sides of the equation tends towards zero. The square root of the prediction error variance, the standard error of the estimate, can be used to determine the confidence interval of the estimate.

The prediction error variances for the multi-parameter estimation problem can be derived in a manner parallel to that described in equation {3.14}. The parameter estimates and the first derivatives will each consist of a vector with the number of elements equal to  $m$ . The second derivatives and the prediction error variances will both be square  $m$  by  $m$  matrices. Using brackets to denote matrices and vectors, the matrix of prediction error variances can be computed with the following equations:

$$\text{Var} [ ] = - \left[ \frac{\partial^2 \log L}{\partial \theta^2} \right]^{-1} \quad \{3.16\}$$

Where the right-hand-side is the inverse of the matrix of second partial derivatives with respect to  $\{\theta\}$ . The diagonal elements will be the prediction error variances of the estimates, and the off-diagonal elements will be the prediction error-covariances between the elements. These are needed to test hypotheses based on linear functions of the parameters.

Even if the prediction error variance is not computed, ML can still be used to test hypothesis, by a "likelihood ratio test". In a likelihood ratio test the maximum likelihoods obtained under two alternative hypotheses are compared. In the null hypothesis, one or more of the parameters that are maximized in the alternative hypothesis are fixed. For example setting a mean equal to zero. Provided that the null hypothesis is "nested" within the alternative hypothesis, under the assumption of no difference, the natural log of the likelihood ratio will be asymptotically distributed as:  $(-1/2)\chi^2$ , where  $\chi^2$  is the Chi-squared statistic. By "nested" hypothesis we mean that, although some parameters that are fixed in the null hypothesis are maximized in the alternative hypothesis, all parameters that are fixed in the alternative hypothesis are also fixed in the null hypothesis. The number of degrees of freedom will be equal to the number of parameters that are maximized in the alternative hypothesis, but fixed in the null hypothesis.

Although it is generally possible to write the likelihood function, and differentiate  $\log L$  with respect to the different parameters, for QTL detection models it will not be possible to derive analytical solutions to the resultant system of equations. Iterative methods to derive solutions will be described below.

### 3.7 Maximum likelihood QTL parameter estimation for crosses between inbred lines and a single marker

We will illustrate ML first for the BC design and a single QTL linked to a single genetic, as illustrated in Figure 3.1. We will assume that the quantitative trait has a normal distribution and that the two QTL genotypes have equal variances. As will be seen below, ML is a much

more flexible technique than least squares, and can handle many situations in which these assumptions do not hold. The statistical density function for a single individual of genotype  $M_1M_2$  will be:

$$f(y) = \frac{(1-r)e^{-(y-\mu_1)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} + \frac{(r)e^{-(y-\mu_2)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \quad \{3.17\}$$

Where  $y$  is the trait value,  $\sigma$  is the standard deviation,  $\mu_1$  is the mean of individuals with the  $Q_1Q_2$  genotype, and  $\mu_2$  is the mean of individuals with the  $Q_2Q_2$  genotype. Individuals with the  $M_2M_2$  genotype will have the same likelihood, except that the QTL mean values will be reversed. The complete likelihood for a sample of individuals can be written as follows:

$$L = \prod_{i=1}^{n_1} [f(y_i, M_1M_2)] \prod_{j=1}^{n_2} [f(y_j, M_2M_2)] \quad \{3.18\}$$

where  $\Pi$  represents the product of a series,  $f(y_i, M_1M_2)$  and  $f(y_j, M_2M_2)$  are the statistical densities for  $i^{\text{th}}$  and  $j^{\text{th}}$  observations with genotypes  $M_1M_2$  and  $M_2M_2$ , respectively; and  $n_1$  and  $n_2$  are the number of individuals with the two genotypes, respectively.

To obtain the ML parameter estimates, the log of this function must be differentiated with respect to four parameters,  $\mu_1$ ,  $\mu_2$ ,  $\sigma$ , and  $r$ . The partial derivatives must then be equated to zero, and this system of four equations must be solved. This system of equations can not be solved analytically. Iterative methods to derive solutions will be described below.

Linkage of the genetic marker to a segregating QTL can be tested by comparing the likelihood in equation {3.18} at convergence to the maximum likelihood obtained with  $r$  fixed to 0.5, i. e., no linkage between the QTL and the genetic marker. As note about, under the null hypothesis of no linkage, with one parameter fixed, the ratio of the natural log of the likelihood ratio will be asymptotically distributed as:  $(-1/2)\chi^2$ , where  $\chi^2$  is the Chi-squared statistic with one degree of freedom.

Alternatively, equation {3.17} can be readily modified so that a different residual variance is assumed for each QTL genotype. In this case it is necessary to estimate five parameters, instead of four. The hypothesis of heterogeneous variance can also be tested against the null hypothesis of homogeneous variance by the log likelihood ratio of the heterogeneous and homogeneous variance maximum likelihoods.

For the F-2 design described in Figure 2.2, each genotype will consist of a mixture of three normal distributions for the two QTL homozygotes and the QTL heterozygote. The probabilities of each of the three QTL genotypes within each marker genotype are given in Table 2.2. Thus, it will be necessary to estimate at least five parameters, the three QTL means,  $r$ , and the residual variance. This model can also be modified so that a separate residual variance is assumed for each QTL genotype. In this case is necessary to estimate seven parameters: three means, three variances, and  $r$  (Weller, 1986). These examples give some indication of the flexibility of ML estimation

### 3.8 Maximum likelihood QTL parameter estimation for crosses between inbred lines and two flanking markers

The likelihood function for the BC design with two flanking markers is described as follows (Lander and Botstein, 1989):

$$L = \prod_{i=1}^{n_{M1N1}} \prod_{j=1}^{n_{M1N2}} \prod_{k=1}^{n_{M2N1}} \prod_{l=1}^{n_{M2N2}} f_{M1N1} f_{M1N2} f_{M2N1} f_{M2N2} \quad \{3.19\}$$

where  $n_{M1N1}$ ,  $n_{M1N2}$ ,  $n_{M2N1}$ ,  $n_{M2N2}$  are the number of individuals with genotypes  $M_1N_1/M_2N_2$ ,  $M_1N_2/M_2N_2$ ,  $M_2N_1/M_2N_2$ , and  $M_2N_2/M_2N_2$ , respectively, and  $f_{M1N1}$ ,  $f_{M1N2}$ ,  $f_{M2N1}$ , and  $f_{M2N2}$  are the density functions for the four possible marker genotypes. Since all individuals received an mn chromosomal segment from the recurrent parent, only the chromosomal segment received from the F-1 is indicated. The density functions for the possible marker genotypes are computed as follows:

$$f_{M1N1} = (1-\alpha)f(Q_1) + \alpha f(Q_2) \quad \{3.20\}$$

$$f_{M1N2} = (1-\beta)f(Q_1) + \beta f(Q_2) \quad \{3.21\}$$

$$f_{M2N1} = (1-\beta)f(Q_2) + \beta f(Q_1) \quad \{3.22\}$$

$$f_{M2N2} = (1-\alpha)f(Q_2) + \alpha f(Q_1) \quad \{3.23\}$$

where  $\alpha = r_1r_2/(1-R)$ ,  $\beta = r_1(1-r_2)/R$ , and  $f(Q_1)$  and  $f(Q_2)$  are the normal density functions for each observation with standard deviations of  $\sigma$ , and means of  $\mu_1$  for the  $Q_1Q_2$  genotype and  $\mu_2$  for the  $Q_2Q_2$  genotype. Thus, the likelihood can be computed by calculating the appropriate density function for each individual, depending on its marker genotype, and multiplying. As for the nonlinear regression model in Section 3.3, zero interference will be assumed, thus  $R = r_1 + r_2 - 2r_1r_2$ , and it is necessary to derive ML estimates for only four parameters,  $r_1$ ,  $\sigma$ ,  $\mu_1$  and  $\mu_2$ .

### 3.9 Methods to maximize likelihood functions

Numerous iterative methods have been proposed to maximize multiparameter likelihood function. The initial solutions for all methods are selected arbitrary. These methods will be compared based on ease of application, speed of convergence, and probability of convergence. Of all the methods that will be considered below, only expectation-maximization (EM) is guaranteed to converge to a maximum, provided a maximum exists within the parameter space.

However, even for EM, the convergence point is a only a local maximum. It is possible that another set of solutions yield the global likelihood maximum. Generally the problem of multiple maxima is addressed by iterating from several different sets of initial values. If all runs converge to the same parameter estimates, then it is likely, but not certain, that this parameter set is a global maximum.

Iterative maximization methods be divided into derivative free methods, methods based on computation of first derivatives, and methods based on computation of second derivatives. For all derivative based methods, the parameter estimates of the  $i^{\text{th}}$  iterate are computed by solving a system of equations equal in number to the number of parameters being estimated. These reduced equations are themselves functions of the parameter estimates from the previous iteration. Thus, it is necessary to continue iteration until changes between rounds fall below a sufficiently small value. Convergence is generally most rapid for second derivative methods, but convergence is not guaranteed, even if there is a maximum within the parameter space. We will consider first derivative-free methods, then methods based on computation of second derivatives, and finally methods based of computation of first derivatives.

**Derivative free methods:** Several general purpose algorithms that find the maximum of a function without computing derivatives have been devised. These methods are available in many software packages, and can be applied to virtually any continuous function. Derivative free methods tend to be inefficient for large samples or many parameters, and are not guaranteed to converge.

**Second derivative based methods:** In Newton-Raphson (Dahlquist and Björck, 1974), both the first derivatives and the matrix of second derivatives are computed analytically. Solutions for the  $i^{\text{th}}$  round of iteration are computed by solving the following system of equations:

$$[I_i] = [I_{i-1}] - \frac{\partial(\log L)}{\partial[\theta]} \left| \frac{\partial^2 \log L}{\partial[\theta^2]} \right|^{-1} \quad \{3.24\}$$

Where  $[I_i]$  is the estimate of  $\theta$  for the  $i^{\text{th}}$  iterate,  $[I_{i-1}]$  is the previous estimate of  $[\theta]$ , and the other terms are as defined above, with derivatives computed for the  $i-1$  estimate of  $[\theta]$ . The main advantage of Newton-Raphson is that convergence is generally rapid. The disadvantages are that the algorithm may not converge, even if the likelihood does have a maximum within the parameter space, and computation of the matrix of second derivatives is often a non-trivial task. This problem is alleviated somewhat if Fisher's method of scoring (Bailey, 1961) is used instead of Newton-Raphson. In this method numerical methods are used to estimate the differentials (Jensen, 1989). Thus, the algebra is simplified somewhat, but there is some sacrifice in both efficiency, in terms of computing time, and the accuracy of the estimates and the prediction error variances. However, as shown above, this matrix can be used to derive estimates of the standard errors of the estimates, which is itself an important objective.

**First derivative based methods (EM):** Expectation-maximization (EM) is based on computation of first derivatives. EM is generally considered the method of choice, because it

is guaranteed to converge to a maximum, provided that one exists within the parameter space. The rate of convergence, however, may be very slow. The principle behind EM is to consider two sampling densities, one based on the complete-data specification (unknown), and the second based on the incomplete-data specification (known).

The EM algorithm consists of two steps: the estimation step, in which the "sufficient statistics" are estimated for the complete-data density function; and the maximization step, in which this function is maximized with respect to the parameters. Lander and Bostein (1989) employed a partial EM algorithm that solved for QTL means and variances for a fixed recombination frequency. This procedure was then repeated for a range of recombination values to obtain the  $r$  value which resulted in ML.

Jansen (1992) derived complete EM equations, which are suitable for a wide range of QTL models. For each individual  $i$ , the likelihood function can be written as  $f(y_i; m_i)$ , where  $y_i$  and  $m_i$  are the quantitative trait value and marker genotype for individual  $i$ . The joint likelihood over all individuals will be  $\Pi f(y_i; m_i)$  as given above. By the general product rule of probability:

$$L = \Pi f(y_i; m_i) = \Pi p(m_i) \Pi f(y_i | m_i) \quad \{3.25\}$$

where  $I$  = total number of individuals,  $p(m_i)$  = probability of genotype  $m_i$ , and  $f(y_i | m_i)$  = density of the trait value given the marker genotype for individual  $i$ . Setting the log of  $L$  to zero gives:

$$\frac{\partial(\log L)}{\partial \theta} = \sum \frac{\partial[\log p(m_i)]}{\partial \theta} + \sum \frac{\partial[\log f(y_i | m_i)]}{\partial \theta} \quad \{3.26\}$$

After algebra equation {3.26} can be expressed as follows:

$$0 = \frac{\partial(\log L)}{\partial \theta} = \sum \frac{I Q}{\partial \theta} \frac{\partial[\log p(q|y_i; m_i)]}{\partial \theta} + \sum \frac{I Q}{\partial \theta} \frac{\partial[\log f(y_i | q)]}{\partial \theta} \quad \{3.27\}$$

Where:

$Q$  = the total number of QTL genotypes

$p(q|m_i)$  = probability of QTL genotype  $q$  conditional on marker genotype  $m_i$ .

$f(y_i | q)$  = density of trait value  $y_i$  conditional on QTL genotype  $q$ .

$p(q|y_i; m_i)$  = probability of QTL genotype  $q$  conditional on trait value  $y_i$  and marker genotype  $m_i$ .

The differential in first expression on the right-hand side of equation {3.27} is only a function of the recombination parameters  $r$  or  $r_1$ , and the differential in the second expression is only a function of QTL genotype means and variances. Using the values of  $p(q|y_i; m_i)$  from

the current iteration, solutions can be derived for the parameters by setting each term equal to zero. In many of the designs considered, these equations can now be easily solved for the current values of  $p(q|y_i, m_j)$ . The "estimation" step consists of computing  $p(q|y_i, m_j)$  using the current values of  $\theta$ . for each individual. For example, for the BC design and a single marker,  $p(q=Q_1|y_i, m_j)$  for the  $M_1 M_2$  genotype is computed as follows:

$$p(q=Q_1|y_i, m_j) = \frac{e^{-(y_i - \mu_1)^2 / 2\sigma^2}}{-(y_i - \mu_1)^2 / 2\sigma^2 + (r)e^{-(y_i - \mu_2)^2 / 2\sigma^2}} \quad \{3.28\}$$

$p(q=Q_2|y_i, m_j)$  is similarly computed with  $\mu_2$  in the numerator instead of  $\mu_1$ . Thus the sum of  $p(q=Q_2|y_i, m_j)$  and  $p(q=Q_1|y_i, m_j)$  of each individual are equal to unity, and these can be considered weighting factors for the differentials in equation {3.27}.

The maximization step consists of solving equation {3.27}. The first term is a weighted non-linear regression, and is a function only of the recombination parameter (r). For the BC design,  $\log [p(q|m_j)]$  is equal to either  $\log r$  or  $\log (1-r)$ , with derivatives of  $1/r$  or  $-1/(1-r)$ , respectively. The second term is a weighted linear regression, and is only a function of QTL means and variances. For the BC design, assuming a normal distribution, the QTL means and variances can be estimated as the trait means and variances weighted by  $p(q|y_i, m_j)$  for each combination of individual by genotype. Other "nuisance" factors, such as block or herd can readily be incorporated as part of the second term, which can be solved as a general linear model for traits with a normal distribution.

With marker brackets the first term is somewhat more complicated, but will still be a function of only a single parameter,  $r$ , assuming zero interference, and that R is known. This method can also be readily applied to analyze multiple brackets. Algorithms are described in Jansen (1992). This method can also be applied even if the within-QTL genotype distribution of the quantitative trait is not normal, provided it is possible to compute the differential of  $\log f(y_i|q)$ , as will be described below for the case of discrete traits.

### 3.10 Analysis of more complicated models by maximum likelihood

We will consider briefly several more complex models that are amenable to solution by ML. Mackinnon and Weller (1995) used ML to estimate QTL parameters for the daughter design, under the assumption that only two QTL alleles were segregating in the population. In addition to the parameters given above for the BC design, it is also necessary to solve for the relative frequency of the two QTL alleles. The likelihood for the model of Mackinnon and Weller (1995) is as follows:

$$L = \prod_{k=1}^K \prod_{v=1}^3 \prod_{i=1}^{L_1} \prod_{j=1}^3 c_{j|iv} f(y_{k|im} - \mu_j) \quad \{3.29\}$$

Where K is the number of sires,  $P_v$  is the probability of sire QTL genotype  $v$ ,  $c_{j|iv}$  is the probability of progeny QTL genotype  $j$  conditional on the combination of sire QTL genotype  $v$  and progeny marker genotype  $i$ ,  $L_1$  is the number of daughters with marker genotype  $i$ ,  $x_{k|im}$  is the trait value for progeny  $l$  of sire  $k$ , with marker genotype  $i$ ,  $\mu_j$  is the mean for progeny QTL genotype  $j$ , and  $f(y_{k|im} - \mu_j)$  the normal density function for progeny  $m$  of sire  $k$ , conditional on QTL genotype  $j$ . Mackinnon and Weller (1995) assumed that only two QTL alleles were segregating in the population, and all sires are heterozygous for the genetic marker. They further assumed that only sires heterozygous for the genetic marker were included in the analysis. Thus, four sire genotypes are determined with respect to the QTL and the genetic marker: homozygotes  $Q_1 Q_1$  and  $Q_2 Q_2$ , the heterozygote with  $Q_1$  linked to  $M_1$ , and the heterozygote with  $Q_1$  linked to  $M_2$ .  $P_v$  can be computed assuming a Hardy-Weinberg distribution of genotypes. It is necessary to define only three marker genotype classes for the progeny; those that receive  $M_1$ , but not  $M_2$ , those that receive  $M_2$ , but not  $M_1$ , and those that receive both paternal alleles. Formula for  $c_{j|iv}$  are given in Mackinnon and Weller (1995). For progeny with the paternal marker genotype, it is not known which allele was received from the sire, and which allele was received from the dam. Thus  $c_{j|iv}$  is a function of marker allele frequency among the dams. Assuming that QTL genotype does not affect the variance, it is necessary to estimate six parameters, the three QTL genotype means, the residual variance, recombination frequency between the marker and the QTL, and the frequency of the  $Q_1$  allele. With large samples, reasonable estimates can be derived for all parameters (Mackinnon and Weller, 1995).

Mackinnon and Weller (1995) modified this model to include a polygenic sire effect. In this case the normal density function becomes  $f(y_{k|im} - \mu_j - g_k)$ , where  $g_k$  is the polygenic effect of sire  $k$  on his daughters. This increases the number of parameters that must be estimated by the number of sires. Bovenhuis and Weller (1994) modified this model to include a direct effect of the marker genotype in addition to a linked QTL.

Several studies have derived ML formula for the granddaughter design. Most following Hoesechele and VanRaden (1993a) have analyzed the sons' Daughter Yield Deviations (DYD) rather than the individual granddaughter production records. DYD are the daughter record means of each son adjusted for systematic environmental effects and merits of mates. The advisability of this will be discussed below. Hackett and Weller (1995) derived an ML model suitable for categorical traits. They assumed a threshold model with an underlying normal distribution, and solved using the method of Jansen (1992).

### 3.11 Bayesian estimation of QTL effects and Gibbs sampling

Bayesian estimation differs from ML in that instead of maximizing the likelihood function, the "posterior probability" of  $\theta$ ,  $p(\theta|y)$  is maximized as a function of the likelihood function and the prior distribution of  $\theta$ . Bayes theorem in general terms for multiple parameters and observations is computed as follows:

$$p(\theta_1, \theta_2, \dots, \theta_m | y_1, y_2, \dots, y_n) = p(\theta_1, \theta_2, \dots, \theta_m) p(y_1, y_2, \dots, y_n | \theta_1, \theta_2, \dots, \theta_m) \quad \{3.30\}$$

Where  $p(\theta_1, \theta_2, \dots, \theta_m)$  is the "prior probability" of the parameters, and  $p(y_1, y_2, \dots, y_n | \theta_1, \theta_2, \dots, \theta_m)$  is the likelihood function. Assuming that prior information of the parameters is available, Bayesian estimation, which makes use of this information should be preferable to ML, which ignores any prior information on the parameters. There are two major drawbacks to Bayesian estimation. First, prior information on the parameters is often vague, and it is not possible to mathematically represent this information without additional assumptions which may not be true. Second, if many records are included in the analysis, then the likelihood function tends to "overwhelm" the prior distribution of  $\theta$ , and the Bayesian estimates tend to the ML estimates. Since in nearly all cases of QTL estimation, the sample sizes will be quite large, little is to be gained by Bayesian parameter estimation. Hoeschle and VanRaden (1993) derived a prior distribution of parameters and the likelihood function for a granddaughter design analysis. They assumed a prior exponential distribution of QTL effects, with an imposed lower limit on the magnitude of QTL effect. They further assumed only two alleles for each QTL segregating in the population, and a uniform distribution of allele frequencies over the complete range of zero to unity. Given the distribution of QTL effects and allele frequencies, and the total additive genetic variance, they estimated the expected number of detectable QTL. Distribution of these QTL throughout the genome was assumed uniform. The Bayesian and ML analyses gave very similar results, and both were quite close to the simulated values (Hoeschle and VanRaden, 1993a).

In Gibbs sampling, a value is generated for each unknown parameter and missing data point from its distribution conditional on the observed data and on all other sampled values. After many repeat samples empirical posterior distributions of the parameters are derived, which can be used to estimate parameter values and confidence limits. Generally very large samples, on the order of 10,000, are required to obtain results that are independent of the starting values (Hoeschle, 1994).

### 3.12 Repeat records, nuisance effects, and polygenic variance

As noted previously, in the analysis of most experimental data, and all field data it will be necessary to account for systematic environmental effects and other "nuisance" effects, such as age or sex. In addition individuals may have multiple records which are partially correlated. Finally, for models with complicated pedigree structure, such as daughter or granddaughter designs, individuals with common marker genotypes can also have a common polygenic component of variance. In the previous chapter, we discussed the model of Fernando and Grossman (1989) that can potentially handle all of these factors, but has several significant deficiencies. Thus, alternative methods have been proposed.

For the granddaughter design, which includes all of these problems, several studies have suggested analyzing either estimate breeding values (EBV) (Andersson-Elkund and Rendel, 1990, Cowan et al., 1992) or DYD (Hoeschle and vanRaden, 1993) based on mixed models that include repeat records and fixed nuisance effects. The EBV or DYD are then analyzed by

a linear model including only the effects associated directly with the genetic markers. EBV derived from a mixed model will be regressed, and therefore estimates of QTL effects derived as described will be biased. In addition, the variances of either BV or DYD will be a function of the quantity of information on the son. Thus, these studies have proposed to weight the evaluations by some function of their reliabilities. In the mixed model equations the coefficient matrix is multiplied by the inverse of the residual variance matrix. Therefore, for DYD, for which the variance decreases as the number of daughters increases, weighting by the repeatabilities is approximately correct. However, for mixed model EBV, which *increase* in variance as the number of progeny increases, the effect of weighting by the repeatability has the opposite effect desired. Sons with few daughters are multiplied by a smaller factor, even though their variance is less. Furthermore, our results, based on simulated data, indicate that QTL effects derived by analysis of DYD also underestimate the simulated effects.

Other alternatives have also been considered, for example iterative solution of polygenic effects as progeny means after correction for QTL effects (Maekinnon and Weller, 1995). In this model the sire polygenic effect was assumed to be fixed. If polygenic effects are assumed random, then the correct solution for maximum likelihood estimation is to integrate the likelihood function with respect to these effects (Elsen et al, 1988). This, of course, tremendously complicates maximizing the likelihood, especially if the number of polygenic effects that must be estimated is large. Approximate integration has also been proposed and tested on simulated full sib data (Knot and Haley, 1992). At present there is no completely satisfactory method for QTL parameter estimation of complex pedigrees.

### 3.13 Summary

In this chapter we considered various methods for QTL parameter estimation, emphasizing maximum likelihood, which although not trivial to applied, can be applied to many models which are not amenable to solution by other methods. A number of recent studies have proposed nonlinear least square models and Bayesian methods as alternative. Least square models give nearly identical results to ML, but are much more limited as to possibilities of model specification. Application of Bayesian methodology requires rather specific knowledge about the prior distribution of QTL effects, which is generally lacking. There is no complete satisfactory method at present for analysis of QTL data from large complex families, such as commercial dairy cattle populations.

### References

- Andersson-Elkund, L., Danell, B. and Rendel, J. (1990) Associations between blood groups, blood protein polymorphisms and breeding values for production traits in Swedish Red & White dairy bulls. *Anim. Genet.* **21**: 361-76.
- Bailey, N. T. (1961). *Introduction to the Mathematical Theory of Genetical Linkage.*

Clarendon Press, Oxford.

Bovenhuis, H. and Weller J. I. (1994) Mapping and analysis of dairy cattle quantitative trait loci by maximum likelihood methodology using milk protein genes as genetic markers. *Genetics* **137**: 267-280.

Cowan, C. M., Dentine, M. R., and Colye, T. (1992) Chromosome substitution effects associated with  $\kappa$ -Casein and  $\beta$ -Lactoglobulin in Holstein cattle. *J. Dairy Sci.* **75**: 1097-1104.

Dahlquist, G. and Björck, A. (1974) *Numerical methods*. Prentice Hall, Englewood Cliffs.

Elsen, J. M., Khang, J. V. T. and Le Roy, P. (1988) A statistical model for genotype determination at a major locus in a progeny test design. *Genet. Sel. Evol.* **20**: 211-226.

Fernando, R., Grossman, M. (1989) Marker assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.* **21**: 467-477.

Hackett, C. A. and Weller J. I. (1995) Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**: 1252-1263.

Haley, C. S. and Knott, S. A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315-324.

Hoeschele, I. (1994) Bayesian QTL mapping via the Gibbs sampler. *Proc. 5th World Cong. Genet. Appl. Livest. Prod. Guelph* **21**: 241-244.

Hoeschele, I., and VanRaden, P. M. (1993) Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. *Theor. Appl. Genet.* **85**: 953-960.

Hoeschele, I., and VanRaden, P. M. (1993a) Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. *Theor. Appl. Genet.* **85**: 946-952.

Jansen, R. C. (1992) A general mixture model for mapping quantitative trait loci by using molecular markers. *Theor. Appl. Genet.* **85**: 252-260.

Jenson, J. (1989) Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. *Theor. Appl. Genet.* **78**: 613-618.

Knott, S. A. and Haley, C. S. (1992) Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics* **132**: 1211-1222.

Lander, E. S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative

traits using RFLP linkage maps. *Genetics* **121**: 185-199.

Maekinnon, M. J. and Weller, J. I. (1995). Methodology and accuracy of estimation of quantitative trait loci parameters in a half-sib design using maximum likelihood. *Genetics* **141**: 755-770.

Martinez, O. and Curnow, R. N. (1992) Estimating the locations and the sized of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet.* **85**: 480-488.

Simpson, S. P. (1989) Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theor. Appl. Genet.* **77**: 815-819.

Weller, J. I. (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**: 627-640.