Regression Based Robust QTL Analysis using Flanking Marker with Intercross (F2) Population

Md. Jahangir Alam  
Statistical programming in clinical trials  
Shafi Consultancy Bangladesh  
Sylhet, Bangladesh  
jahangir_statru63@yahoo.com

Md. Alamin  
Department of Agronomy  
Zhejiang University  
Hangzhou, China  
alamin_stat76@yahoo.com

Department of Statistics  
University of Rajshahi  
Rajshahi, Bangladesh  
*sonirustat10@gmail.com, **aaman.ru@gmail.com and ***mollah.stat.bio@ru.ac.bd

Abstract—Quantitative trait loci (QTL) analysis is a most popular and widely used statistical approach for the detection of important genes, responsible for a quantitative trait, in the chromosomes. Maximum likelihood (ML) based interval mapping (IM) is one of the most popular approaches for QTL analysis. It is relatively complex and computationally slow than regression based interval mapping approach. Haley-Knott (HK) and extended Haley-Knott (eHK) regression based IM can save time in computation and produce similar results to those obtained by ML-IM method. However, most of the IM approaches including ML-IM, HK-IM, eHK-IM are not robust against phenotypic outliers. In this paper, an approach has been developed to robustify the regression based IM by maximizing beta-likelihood function. The proposed method reduces to the traditional HK method when beta tends to zero. The tuning parameter beta plays a key role on the performance of the proposed method for QTL mapping. It controls the trade-off between the robustness and efficiency of the estimators. The analysis results of simulated data show that the proposed method improves performance over the existing IM approaches in the case of data contaminations; otherwise, it shows almost same results as classical IM approaches.

Index Terms—QTL analysis, robust regression, maximum beta-likelihood estimation, beta-LRT criterion and robustness.

I. INTRODUCTION

The rapid increase in availability of fine-scale genetic markers due to the rapid advancement in molecular biology has led to the intensive use of QTL mapping in the genetic study of quantitative traits in bioinformatics. The simplest method to discover QTL using data on an experimental cross is to perform analysis of variance (ANOVA) at each of marker loci (Soller et al., 1976). Thoday (1960) first proposed the idea of using two markers to bracket a region for testing QTLs. Lander and Botstein (1989) implemented a similar, but much improved, method to use two adjacent markers to test the existence of a QTL in the interval by performing a likelihood ratio test (LRT) at every position in the interval. This is known as interval mapping (IM) approach. There are several interval mapping approaches like maximum likelihood (ML) based IM (Lander and Botstein, 1989) and regression based IM (Haley and Knott, 1992).

In practice, QTL effects are either treated as fixed or random (Xu, 1998). In fixed QTL model, allelic substitution effects are usually estimated and tested, and QTL variance is calculated from estimated allelic effects. In random QTL model, the QTL effects and QTL variance are directly estimated and tested. In MLE based IM model, the conditional expectations of the QTL genotype given the flanking marker genotypes are unknown (Lander and Botstein, 1989). This QTL effect model can be treated as a random effect model. So, we can call the MLE based interval mapping model as a random effect model (REM). In REM we can’t apply least square estimate. On the other hands, in the HK regression based IM model the conditional expectation of the QTL genotype given the flanking marker genotypes is considered as fixed (Kao, 2000), and this model can be treated as a fixed effect model (FEM).

The existing QTLs interval mapping based on REM (Lander and Botstein, 1989) and FEM (Haley and Knott, 1992) are two most popular methods for QTL analysis. But these methods are not robust against outliers. In this paper, by using the proposed robust method (Alamin, 2011), we want to analysis the FEM for intercross (F2) population. We also show a simulation study to investigate the performance of the proposed method with the existing random effect QTL model and fixed effect QTL model for F2 population.
II. A QTL MAPPING FOR INTERCROSS (F2) POPULATION USING REGRESSION APPROACH

Let us consider no epistasis between two QTLs, no interference in crossing over, and only one QTL in the testing interval. A QTL mapping fixed effect model for intercross population (F2) for testing a QTL in a marker interval is define as

\[ y_j = \mu + ax_{ji} + dz_{ji} + u_j, \quad i=1, 2, 3 \text{ and } j=1, 2 \]

where \( y_j \) is the phenotypic value of the j-th individual, \( x_{ji} = p_{j1} - p_{j2}, z_{ji} = p_{j2}, \mu \) is the general mean effect, \( a \) is the QTL additive effect, \( d \) is the QTL dominance effect, and \( u_j \) is a random error. We assume that \( u_j \sim N(0, \sigma^2) \).

The conditional probabilities for QTL genotypes \( QQ, Qq, qQ, qq \) given the flanking marker genotypes are denoted by \( p_{j1}, p_{j2} \) and \( p_{j3} \) respectively. The conditional probabilities of the QTL genotypes given marker genotypes are given in TABLE I for the intercross (F2) population. We extract the conditional probabilities from this table for F2 population.

**TABLE I. CONDITIONAL PROBABILITIES OF A PUTATIVE QTL GENOTYPE GIVEN THE FLANKING MARKER GENOTYPES FOR AN F2 POPULATION**

<table>
<thead>
<tr>
<th>Marker Genotypes</th>
<th>Expected Frequency</th>
<th>QTL Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN/MN</td>
<td>((1-r)^2/4)</td>
<td>(Q(p_{j1}))</td>
</tr>
<tr>
<td>MN/Mn</td>
<td>(r(1-r)^2)</td>
<td>((1-p))</td>
</tr>
<tr>
<td>Mn/MN</td>
<td>(r^2/4)</td>
<td>((1-p)^2)</td>
</tr>
<tr>
<td>MN/mn</td>
<td>((1-r)^2)</td>
<td>(p)</td>
</tr>
<tr>
<td>Mn/nm</td>
<td>([((1-r)^2 + r^2)/2]</td>
<td>(cp(1-p))</td>
</tr>
<tr>
<td>Mn/mn</td>
<td>(r(1-r)^2)</td>
<td>(0)</td>
</tr>
<tr>
<td>mn/mn</td>
<td>(r^2/4)</td>
<td>(p^2)</td>
</tr>
<tr>
<td>mn/mn</td>
<td>((1-r)^2/4)</td>
<td>(0)</td>
</tr>
<tr>
<td>mn/mn</td>
<td>((1-r)^2/4)</td>
<td>(0)</td>
</tr>
</tbody>
</table>

Here \( p = r_{MQ}r_{MN} \), where \( r_{MQ} \) is the recombination fraction between the left marker M and the putative QTL and \( r_{MN} \) is the recombination fraction between two flanking markers M and N. Here \( c = r_{MN}^2 + r_{MN}^2 + (1 - r_{MN}^2) \). The possibility of a double recombination event in the interval is ignored.

To investigate the existence of a QTL at a given position in a marker interval, we want to test the following statistical hypothesis:

\[ H_0: \alpha = 0 \text{ and } d = 0 \] (i.e., there is no QTL at a given position) vs \( H_1: H_0 \) is not true. (2)

In estimation, both ordinary least-squares (OLSE) and maximum-likelihood (MLE) techniques can be implemented to estimate the parameters \( \mu, a, d \) and \( \sigma^2 \). Under the normality assumption, the normal probability density of the trait value within each QTL genotype class is defined as

\[ f(y_j | \theta) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{1}{2} \frac{(y_j - \mu - ax_{ji} - dz_{ji})^2}{\sigma^2} \right) \]

(3)

where \( \theta = (\mu, a, d, \sigma^2) \). The likelihood function for parameter \( \theta = (\mu, a, d, \sigma^2) \) is given below:

\[ L(\theta | Y) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{1}{2} \frac{(y_j - \mu - ax_{ji} - dz_{ji})^2}{\sigma^2} \right) \]

(4)

To test \( H_0 \) against \( H_1 \), the likelihood ratio test (LRT) statistic is defined as

\[ LRT = -2 \log \left( \frac{\sup_{\theta_0} L(\theta | Y)}{\sup_{\Theta} L(\theta | Y)} \right) = 4.608295 \times \text{LOD} \]

(5)

where \( \Theta_0 \) and \( \Theta \) are the restricted and unrestricted parameter spaces. The threshold value to reject the null hypothesis can't be simply chosen from a chi-square distribution because of the violation of regularity conditions of asymptotic theory under \( H_0 \). The number and size of intervals should be considered in determining the threshold value. Since multiple tests are performed in mapping, the hypotheses are usually tested at every position of an interval and for all intervals of the genome to produce a continuous LRT statistic profile. At every position, the position parameter \( p \) is predetermined and only \( \mu, a, d \) and \( \sigma^2 \) are involved in estimation and testing. If the tests are significant in a chromosome region, the position with the largest LRT statistic is inferred as the estimate of the QTL position, and the MLEs at this position are the estimates of \( \mu, a, d \) and \( \sigma^2 \) obtained by iterative way.

Setting the first derivative of the log-likelihood, with respect to each parameter, equal to zero and solving the log likelihood equations, we obtain the MLEs of the parameters. We have

\[ \hat{\gamma} = (X^TX)^{-1}(X^TY) \]

(6)

\[ \hat{\sigma}^2 = \frac{1}{n}(Y - \hat{X}\hat{\gamma})^T(Y - \hat{X}\hat{\gamma}) \]

(7)

where \( Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \hat{X} = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix} \) and \( \gamma = \begin{bmatrix} \mu \\ \alpha \\ d \end{bmatrix} \).

Note that OLSE and MLE of \( \gamma \) are exactly same, but for \( \sigma^2 \) these are not exactly same. The OLSE of \( \sigma^2 \) has a divisor \( n-3 \), whereas the MLE of \( \sigma^2 \) has a divisor \( n \) instead of \( (n-3) \). However, obviously these estimates are very much sensitive to
outliers. Therefore, regression analysis by OLSE or MLE produces misleading results in presence of outliers.

III. ROBUST QTL MAPPING FOR INTERCROSS POPULATION USING REGRESSION APPROACH

The β-likelihood function (for details about β-likelihood, see (Mollah et al., 2007)) for \( \theta = (\mu, a, d, \sigma^2) \) is given by

\[
L_\beta(\theta | Y) = \frac{1}{n} \prod_{i=1}^{n} f^\beta_y(y_i) = 1
\]

Setting the first derivative of Eq. 8, with respect to each parameter, equal to zero and solving the equations, we obtain the proposed estimators of the parameters \( \theta = (\mu, a, d, \sigma^2) \). We get

\[
\delta \frac{\partial L(\theta | Y)}{\partial \theta} = 0, \quad i = 1, 2
\]

\[
\Rightarrow \sum_{j=1}^{n} \left[ (y_j - \mu - ax_{j1} - dx_{j2})w_jx_q \right] = 0, \quad k = 0, 1, 2
\]

where \( x_q = 1 \) for all \( j = 1, 2, \ldots, n \) and \( w_j \) is the weight function defined as

\[
w_j = w(y_j | \theta, x_q) = \exp \left[ -\frac{\beta}{2a^2} \left( y_j - \mu - ax_{j1} - dx_{j2} \right)^2 \right], \quad i = 1, 2.
\]

The weight function \( w_j \) produces almost zero weight for outlying observations.

Solving Eq. 9, we can get the normal equations which can be written in matrix as follows:

\[
\sum_{j=1}^{n} w_j \sum_{j=1}^{n} x_{j1} - \sum_{j=1}^{n} w_j x_{j2} = 0
\]

\[
\sum_{j=1}^{n} w_j x_{j1} - \sum_{j=1}^{n} w_j x_{j2} = 0
\]

\[
\sum_{j=1}^{n} w_j x_{j1} x_{j2} - \sum_{j=1}^{n} w_j x_{j2} = 0
\]

\[
\sum_{j=1}^{n} w_j x_{j1} x_{j3} - \sum_{j=1}^{n} w_j x_{j3} = 0
\]

\[
\Rightarrow (X^T X) \hat{\beta} = X^T Y
\]

Hence, the proposed estimates of the parameters are as follows:

\[
\hat{\gamma} = (X^T X)^{-1} X^T Y \quad (10)
\]

\[
\hat{\sigma}^2 = \frac{1}{n} (Y - X \hat{\gamma})^T (Y - X \hat{\gamma}) \quad (11)
\]

where \( X_{11} = X \otimes (W_{11}, 1) \). The notation \( \otimes \) denotes the Hademerd product.

To investigate the existence of a QTL at a given position in a marker interval, we want to test the hypothesis \( H_0 \). To test \( H_0 \) against \( H_1 \), the proposed test criterion is defined as

\[
\hat{\lambda}_p = 2n \left[ L_\beta(\hat{\theta}_1 | Y) - L_\beta(\hat{\theta}_0 | Y) \right]
\]

where \( \hat{\theta}_0 = (\mu, \sigma^2) \) and \( \hat{\theta}_1 = (\mu, a, d, \sigma^2) \).

By the permutation test, we compute the p-value for testing \( H_0 \) against \( H_1 \) using the following formula

\[
P = \frac{N_p}{\sum_{k=1}^{N_p} \lambda^*_p(k)}
\]

where \( N_p \) is the number of permutation under \( H_0 \) and \( \hat{\lambda}_p(k) \) is the estimate of \( \lambda_p \) for the original dataset and \( \hat{\lambda}_p(k) \) is the estimate of \( \lambda_p \) for the k-th permutation of the values of the response variable. Note that, for \( \beta \to 0 \), \( \hat{\lambda}_p \) reduces to the approximate \( \chi^2 \) distribution.

IV. SIMULATION RESULTS

To illustrate the performance of the proposed method in a comparison of REM and FEM for QTL mapping with intercross population, let us first consider total 8 chromosomes and 11 equally spaced markers in each of chromosomes, where any two successive marker interval size is 5cM. The true QTL position is located in chromosome 1 with marker 5. The true QTL and the remaining 80% is subject to the environmental effects (random error).

Figure 1(a) represents the scatter plot of 250 trait values. To investigate the robustness of the proposed method in a comparison of the REM and FEM methods, we contaminated 10% trait values in this dataset by outliers. Figure 1(b) shows the scatter plot of contaminated dataset. Then we computed LOD scores by REM, FEM and the proposed methods for both types of data sets. It should be noted here that the name ‘LOD scores’ is used in this paper for convenience of presentation instead of both LRT scores of REM method, FEM method and the β-LOD scores of proposed method, respectively. Figure 1(c) shows the LOD scores profile for the uncontaminated dataset, where dotted, two dash and solid lines represents the LOD scores at every 1cM position in the chromosomes for REM, FEM and the proposed method with \( \beta = 0.2 \), respectively. Figure 1(d) shows the LOD scores profile for the contaminated data set, where dotted, two dash and solid lines represents the LOD scores at every 1cM position in the chromosomes as before for REM, FEM and the proposed method with \( \beta = 0.2 \), respectively. It is seen that the highest LOD score peak occurs in the true QTL position of the true chromosome 1 with marker 5 by all three methods for the intercross population, let us first consider total 8 chromosomes and 11 equally spaced markers in each of chromosomes, where any two successive marker
The true QTL position is located in chromosome 2, 4 and 6 with marker 5. The true values for the parameters in the fixed effect model are assumed as $\mu = 0.05$, $a=0.8$, $d=0.4$ and $\sigma^2 = 0.5$. To test the null hypothesis $H_0: a = 0$ and $d = 0$ against the existence of a QTL ($a \neq 0$, $d \neq 0$), we generated 250 trait values with heritability $h^2 = 0.2$. Figure 2(a) represents the scatter plot of 250 trait values. To investigate the robustness of the proposed method in a comparison of the REM and FEM methods, we contaminated 10% trait values in this dataset by outliers. Figure 2(b) shows the scatter plot of contaminated dataset.

Then we computed LOD scores by REM, FEM and the proposed methods for both types of datasets. Figure 2(c) shows the LOD scores profile for the uncontaminated dataset, where dotted, two dash and solid lines represents the LOD scores at every 1cM position in the chromosomes for REM, FEM and the proposed method with $\beta = 0.2$, respectively. Figure 2(d) shows the LOD scores profile for the contaminated dataset, where dotted, two dash and solid lines represents the LOD scores at every 1cM position in the chromosomes as before for REM, FEM and the proposed method with $\beta = 0.2$, respectively. It is seen that the highest LOD score peak occurs in the true QTL position of the true chromosome 2, 4, and 6 with marker 5 by all three methods for the uncontaminated dataset. However, in presence of outliers, the highest LOD score peak occurs in the true QTL position by the proposed method only [see Fig. 2(d)].

V. CONCLUSION

In this paper, we have discussed a new robust regression based interval mapping approach for QTL analysis by maximum $\beta$-likelihood estimation. The value of the tuning parameter $\beta$ plays a key role on the performance of the proposed method. An appropriate value for the tuning parameter $\beta$ can be selected by cross validation. The proposed method with tuning parameter $\beta = 0$ reduces to the traditional interval mapping approach. Simulation results show that the proposed method significantly improves the performance over the classical interval mapping approaches in presence of phenotypic outliers. Otherwise, it keeps equal performance to the classical methods.

ACKNOWLEDGMENT

We would like to thank the learned reviewers whose valuable comments helped to strengthen this paper.

REFERENCES