
Robust QTL analysis by minimum β -divergence method

Md. Nurul Haque Mollah*

The Institute of Statistical Mathematics,
4-6-7, Minami-Azabu,
Minato-Ku, Tokyo 106-8569, Japan

E-mail: nhmollah@ism.ac.jp

E-mail: mnhmollah@yahoo.co.in

*Corresponding author

Shinto Eguchi

The Institute of Statistical Mathematics,
4-6-7, Minami-Azabu,
Minato-Ku, Tokyo 106-8569, Japan

E-mail: eguchi@ism.ac.jp

Abstract: Robustness has received too little attention in Quantitative Trait Loci (QTL) analysis in experimental crosses. This paper discusses a robust QTL mapping algorithm based on Composite Interval Mapping (CIM) model by minimising β -divergence using the EM like algorithm. We investigate the robustness performance of the proposed method in a comparison of Interval Mapping (IM) and CIM algorithms using both synthetic and real datasets. Experimental results show that the proposed method significantly improves the performance over the traditional IM and CIM methods for QTL analysis in presence of outliers; otherwise, it keeps equal performance.

Keywords: QTL; Quantitative trait loci; Gaussian mixture distribution; CIM; Composite interval mapping; Minimum β -divergence method; β -LOD scores; robustness.

Reference to this paper should be made as follows: Mollah, M.N.H. and Eguchi, S. (xxxx) 'Robust QTL analysis by minimum β -divergence method', *Int. J. Data Mining and Bioinformatics*, Vol. x, No. x, pp.xxx-xxx.

Biographical notes: Md. Nurul Haque Mollah received his PhD in Statistics in 2005 from the Graduate University for Advanced Studies, Tokyo. He is an Associate Professor in the Department of Statistics, University of Rajshahi, Bangladesh. Currently, he is a project researcher at the Institute of Statistical Mathematics (ISM), Tokyo. His research interests include robust estimation, multivariate statistics, statistical inference on genome science and signal processing.

Shinto Eguchi received his PhD in Statistics at the University of Hiroshima, Japan. He is a professor at the Graduate University for

Advanced Studies, Tokyo and the Institute of Statistical Mathematics (ISM). His research interests includes robust estimation, multivariate statistics, machine learning theory and statistical inference on genome science.

1 Introduction

The basic methodology for mapping QTLs involves arranging a cross between two inbred strains differing substantially in a quantitative trait: segregating progeny are scored both for the trait and for a number of genetic markers. A cross between two parental inbred lines P_1 and P_2 is performed to produce an F_1 population. The F_1 progeny are all heterozygote's with the same genotype. Typically, the segregating progeny are produced by a backcross ($B_1 = F_1 \times \text{parent}$) or an intercross ($F_2 = F_1 \times F_1$).

With the rapid advances in molecular biology, it has become possible to gain fine-scale genetic maps for various organisms by determining the genomic positions of a number of genetic markers (RFLP, isozymes, RAPDs, AFLP, VNTRs, etc.) and to obtain a complete classification of marker genotypes by using codominant markers. These advances greatly facilitate the mapping and analysis of QTLs. Thoday (1960) first introduced 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 termed as IM. However, IM can bias identification and estimation of QTLs when multiple QTLs are located in the same linkage group (Lander and Botstein, 1989; Haley and Knott, 1992; Jansen, 1992, 1993). It is also not efficient to use only two markers at a time for mapping analysis. In view of these problems, QTL mapping combining IM with the multiple marker regression analysis is discussed by Jansen (1993), Zeng (1993). Zeng (1994) named this combination as CIM. It avoids the use of multiple marker intervals to deal with the problems of mapping multiple QTL by conditioning a test for a QTL on some linked or unlinked markers that diffuse the effects of other potential QTLs. Kao and Zeng (1997) generalise the CIM model for QTL analysis by maximising likelihood function using EM algorithm. However, the QTL analysis algorithms mentioned above are not robust against outliers. Mollah and Eguchi (2008) have discussed robust QTL analysis for F_2 intercross population using CIM model by minimising β -divergence (Minami and Eguchi, 2002). In this paper, an attempt is made to extend the discussion of Mollah and Eguchi (2008) for robust QTL analysis using CIM model with a backcross population.

In Section 2, we discuss the genetic model and its extension to statistical CIM model. Section 3 introduce the proposed method for robust QTL analysis based on CIM model. We demonstrate the performance of the proposed method using both simulated and real datasets in Section 4 and make a conclusion of our study in Section 5.

2 Genetic model

Let us consider a QTL in the backcross population in which the frequencies of genotypes QQ and Qq are $1/2$ and $1/2$, respectively. The genetic model for a QTL is as follows:

$$\mathbf{G} = \begin{bmatrix} G_2 \\ G_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1/2 \\ -1/2 \end{bmatrix} [a] = \mathbf{1}_{2 \times 1} \mu + \mathbf{D} \mathbf{E}.$$

It was proposed to model the relation between a genotypic value G and the genetic parameters μ and a . Here G_2 and G_1 are the genotypic values of genotypes QQ and Qq . We call D the genetic design matrix. The unique solutions of the genetic parameters in terms of genotypic values and frequencies are $\mu = (G_2 + G_1)/2$ and $a = G_2 - G_1$.

Let loci M, with alleles M and m , and N with alleles N and n , denote two flanking markers for an interval where a putative QTL is being tested. Let the unobserved QTL locus Q with alleles Q and q be located in the interval flanked by markers M and N. The distribution of unobserved QTL genotypes can be inferred from the observed flanking marker genotypes according to the recombination frequencies between them. To infer the distribution of QTL genotype, we assume that there is no crossover interference and also that double recombination events within the interval are very rare and can be ignored to simplify the analysis. The conditional probabilities of the QTL genotypes given marker genotypes are given in Table 1 for the backcross population. We extract the conditional probabilities from this table to form a matrix \mathbf{Q} for backcross population.

Table 1 Conditional probabilities of a putative QTL genotype given the flanking marker genotypes for a backcross population

Marker genotypes	Expected frequency	QTL genotypes	
		$QQ(p_{j1})$	$Qq(p_{j2})$
MN/MN	$(1 - r_{MN})/4$	1	0
MN/Mn	$r_{MN}/2$	$1 - p$	p
MN/mN	$r_{MN}/4$	p	$1 - p$
MN/mn	$(1 - r_{MN})/2$	0	1

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. The possibility of a double recombination event in the interval is ignored.

2.1 Statistical model for QTL mapping

We assume no epistasis between QTLs, no interference in crossing over, and only one QTL in the testing interval. QTL mapping data consists of two parts, $y_j (j = 1, \dots, n)$ for the quantitative trait value and $X_j (j = 1, \dots, n)$ for the genetic markers and other explanatory variables, for example sex or diet. A CIM

statistical model based on the genetic model for testing a QTL in a marker interval is proposed as

$$y_j = ax_j^* + X_j\gamma + \epsilon_j \quad (1)$$

where

$$x_j^* = \begin{cases} 1/2, & \text{for } QQ \\ -1/2, & \text{for } Qq \end{cases}$$

y_j is the phenotypic value of the j th individual; X_j , a subset of \mathbf{X}_j , may contain some chosen markers and other explanatory variables; γ is the partial regression coefficient vector including the mean μ ; and ϵ_j is a random error. We assume $\epsilon_j \sim N(0, \sigma^2)$. The advantages of using X_j in QTL mapping have been discussed in Kao and Zeng (1997), Zeng (1993, 1994). Basically, it could control for the confounding effect of linked QTLs and reduce the residual variance in the analysis.

2.2 QTL analysis by CIM model based on maximum likelihood estimators

Given the data with n individuals, the likelihood function for $\theta = (p, a, \gamma, \sigma^2)$ is

$$L(\theta | Y, X) = \prod_{j=1}^n \left[\sum_{i=1}^2 p_{ji} \phi\left(\frac{y_j - \mu_{ji}}{\sigma}\right) \right] \quad (2)$$

where $\phi(\cdot)$ is a standard normal probability density function, $\mu_{j1} = a/2 + X_j\gamma$ and $\mu_{j2} = -a/2 + X_j\gamma$. The density of each individual is assumed as a mixture of three normal densities with different means and mixing proportions. The mixing proportions p_{ji} 's which are functions of the QTL position parameter p , are conditional probabilities of QTL genotypes given marker genotypes. The EM algorithm is used to obtain MLEs of the likelihood treating the normal mixture model as an incomplete-data problem.

In QTL mapping, a statistical test is performed whether there is a QTL at a given position within a marker interval. The statistical hypothesis are

$$\begin{aligned} H_0 : a &= 0 \quad (\text{i.e., there is no QTL at a given position}), \\ H_1 : a &\neq 0 \quad (\text{i.e., there is a QTL at that position}). \end{aligned}$$

To test the hypothesis, the LRT statistic

$$\begin{aligned} \text{LRT} &= -2 \log \left[\frac{\sup_{\Theta_0} L(\theta | Y, X)}{\sup_{\Theta} L(\theta | Y, X)} \right] \\ &= 2 \left[\log \sup_{\Theta} L(\theta | Y, X) - \log \sup_{\Theta_0} L(\theta | Y, X) \right] \end{aligned} \quad (3)$$

is used as the test statistic, where Θ_0 and Θ are the restricted and unrestricted parameter spaces. The threshold value to reject the null hypothesis can not be

simply chosen from a χ^2 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 hypothesis 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 a, γ and σ^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 p , and the MLEs at this position are the estimates of a, γ and σ^2 obtained by EM algorithm (Kao and Zeng, 1997). Note that EM algorithm has been also used to obtain MLEs in several studies of QTL mapping analysis (Lander and Botstein, 1989; Carbonell et al., 1992; Jansen, 1992; Zeng, 1994).

3 Robust QTL analysis by CIM model based on minimum β -divergence estimators

The β -divergence between two probability density functions $p(u)$ and $q(u)$ is defined as

$$D_\beta(p, q) = \int \left[\frac{1}{\beta} \{p^\beta(u) - q^\beta(u)\} p(u) - \frac{1}{\beta+1} \{p^{\beta+1}(u) - q^{\beta+1}(u)\} \right] du,$$

for $\beta > 0$. It is non-negative, that is $D_\beta(p, q) \geq 0$, equality holds iff $p = q$, (Basu et al., 1998; Minami and Eguchi, 2002). We note that β -divergence reduces to Kullback Leibler (KL) divergence when $\beta \rightarrow 0$, that is

$$\lim_{\beta \downarrow 0} D_\beta(p, q) = \int p(u) \log \frac{p(u)}{q(u)} du = D_{\text{KL}}(p, q).$$

The minimum β -divergence estimators are defined by the minimisation of the β -divergence between the empirical distribution $p(y)$ and the parametric distribution $f_\theta(y)$ with respects to the parameter $\theta = \{\theta_1, \theta_2, \dots, \theta_k\}$. That is

$$\begin{aligned} \theta_\beta &= \underset{\theta}{\operatorname{argmin}} D_\beta(p(y), f_\theta(y)) \\ &= \underset{\theta}{\operatorname{argmax}} L_\beta(\theta) \end{aligned} \tag{4}$$

where,

$$L_\beta(\theta) = \frac{1}{\beta} \int p(y) f_\theta^\beta(y) dy - b_\beta(\theta) \tag{5}$$

with

$$b_\beta(\theta) = \frac{1}{\beta+1} \int f_\theta^{\beta+1}(y) dy$$

which is independent of y . The empirical version of equation (5) with respect to the CIM model (1) can be written as

$$L_\beta(\theta | Y, X) = \frac{1}{n\beta} \sum_{j=1}^n f_\theta^\beta(y_j | X_j) - b_\beta(\theta, X) \quad (6)$$

which we call β -likelihood function for convenience of presentation. In our current context

$$f_\theta(y_j | X_j) = \sum_{i=1}^2 p_{ji} \phi\left(\frac{y_j - \mu_{ji}}{\sigma}\right) \quad (7)$$

be the normal mixture model, where the notations $\phi(\cdot)$, p_{ji} , μ_{ji} , X_j and $\theta = (p, a, \gamma, \sigma^2)$ are defined in equation (2). Then the minimum β -divergence estimators of a, γ and σ^2 are obtained maximising β -likelihood function using EM like algorithm treating the normal mixture model as an incomplete-data density as discussed below. Let

$$g(x_j^*) = \begin{cases} p_{j1}, & \text{if } x_j^* = 1/2 \\ p_{j2}, & \text{if } x_j^* = -1/2 \end{cases} \quad (8)$$

is the distribution of QTL genotype specified by x_j^* . Let us treat the unobserved QTL genotype x_j^* as missing data, denoted by $y_{j(mis)}$, and treat trait (y_j) and selected markers and explanatory variables (X_j) as observed data, denoted by $y_{j(obs)}$. Then, the combination of $y_{j(mis)}$ and $y_{j(obs)}$ is the complete data, denoted by $y_{j(com)}$. The conditional distribution of observed data, given missing data, is considered as an independent sample from a population such that

$$y_j | (\theta, X_j, x_j^*) \sim N(ax_j^* + X_j\gamma, \sigma^2).$$

Thus the complete-data density model in this problem is regarded as a two-stage hierarchical model. First the value of random variable x_j^* is sampled by a binomial experiment to decide QTL genotype, and then a normal variate for that genotype is generated. The values of random variable x_j^* of individual j are $1/2$ and $-1/2$ for QTL genotype QQ and Qq with probability p_{j1} and p_{j2} , respectively. Thus the complete-data density function is given by

$$f(y_{j(com)} | \theta) = \left\{ p_{j1} \phi\left(\frac{y_j - \mu_{j1}}{\sigma}\right) \right\}^{(\frac{1}{2} + x_j^*)} \times \left\{ p_{j2} \phi\left(\frac{y_j - \mu_{j2}}{\sigma}\right) \right\}^{(\frac{1}{2} - x_j^*)}.$$

To compute the mixing proportions p_{j1} and p_{j2} with respect to QTL genotypes QQ and Qq for each individual, p is determined at a given position using the ratio of two recombination fractions as defined in Table 1 based on flanking markers. The recombination fractions can be computed using Haldane's map function or any other map function also. To obtain the minimum β -divergence estimators of a, γ

and σ^2 by maximising β -likelihood function using EM like algorithm, the iteration of the $(t + 1)$ EM-step is as follows:

E-step: The conditional expected complete-data β -likelihood with respect to the conditional distribution of Y_{mis} given Y_{obs} and the current estimated parameter value $\theta^{(t)}$ is given by

$$\begin{aligned} Q_\beta(\theta | \theta^{(t)}) &= \int L_\beta(\theta | Y_{com}) h(Y_{mis} | Y_{obs}, \theta = \theta^{(t)}) dY_{mis} \\ &= \frac{1}{n\beta} \sum_{j=1}^n \int f^\beta(y_{j(com)} | \theta) \times h(y_{j(mis)} | y_{j(obs)}, \theta = \theta^{(t)}) dy_{j(mis)} - l_\beta(\theta) \\ &= \frac{1}{n\beta} \sum_{j=1}^n \sum_{i=1}^2 \left[\phi\left(\frac{y_j - \mu_{ji}}{\sigma}\right) p_{ji} \right]^\beta \times \pi_{ji}^{(t)} - l_\beta(\theta) \end{aligned}$$

where $l_\beta(\theta) = (1 + \beta)^{-3/2} (2\pi\sigma^2)^{-\beta/2}$ and

$$\pi_{ji} = \frac{p_{ji} \phi\left(\frac{y_j - \mu_{ji}}{\sigma}\right)}{\sum_{i=1}^2 p_{ji} \phi\left(\frac{y_j - \mu_{ji}}{\sigma}\right)} \quad (9)$$

is the posterior probability of i th QTL genotype with respect to the j th individual ($i = 1, 2; j = 1, 2, \dots, n$).

M-step: Find $\theta^{(t+1)}$ to maximise the conditional expected β -likelihood by taking the derivatives of $Q_\beta(\theta | \theta^{(t)})$ with respect to each parameter. The solutions of parameters in closed form are as follows.

$$a^{(t+1)} = (Y - X\gamma^{(t)})^T \Pi_\beta^{(t)} D [\mathbf{1}^T \Pi_\beta^{(t)} (D \# D)]^{-1} \quad (10)$$

$$\gamma^{(t+1)} = [X^T \{X \# (\Pi_\beta^{(t)} \mathbf{1})\}]^{-1} [X^T \{Y \# (\Pi_\beta^{(t)} \mathbf{1}) - \Pi_\beta^{(t)} D a^{(t+1)}\}] \quad (11)$$

$$\begin{aligned} \sigma^{2(t+1)} &= [(Y - X\gamma^{(t+1)})^T \{(Y - X\gamma^{(t+1)}) \# (\Pi_\beta \mathbf{1})\} \\ &\quad - 2(Y - X\gamma^{(t+1)})^T \Pi_\beta^{(t)} D a^{(t+1)} + V^{(t)} a^{2(t+1)}] \\ &\quad [\mathbf{1}^T \Pi_\beta^{(t)} \mathbf{1} - \beta(1 + \beta)^{-3/2}]^{-1} \end{aligned} \quad (12)$$

where

$$\Pi_\beta = \left\{ \left[\exp \left\{ -\frac{1}{2} \left(\frac{y_j - \mu_{ji}}{\sigma} \right)^2 \right\} p_{ji} \right]^\beta \pi_{ji} \right\}_{n \times 2}, \quad (13)$$

$V = \mathbf{1}^T \Pi_\beta (D \# D)$ and the notation $\#$ denotes Hadamards product, which is the element-by-element product of corresponding elements of two same-order matrices. The coefficient of π_{ji} in Π_β is known as β -weight for j th individual in the i th QTL genotype. For $\beta = 0$, the matrix Π_β reduces to the matrix of standard posterior probabilities. It should be noted here that each element of matrices $\mathbf{1}$'s around equations (10)–(12) is 1 with appropriate orders for matrix operation. The E and M steps are iterated until a convergent criterion is satisfied. The converged values of a, γ and σ^2 are the values of minimum β -divergence

estimators. Note that minimum β -divergence estimators (10), (11) and (12) with $\beta = 0$ reduce to Maximum Likelihood Estimators (MLE) proposed by Kao and Zeng (1997) for QTL mapping with backcross population.

Under null hypothesis $H_0: a = 0$, the minimum β -divergence estimators for the parameters γ and σ^2 are obtained iteratively as follows

$$\gamma^{(t+1)} = [X^T \{X \# (W_\beta^{(t)} \mathbf{1})\}]^{-1} \{X \# (W_\beta^{(t)} \mathbf{1})\}^T Y \quad (14)$$

$$\sigma^{2(t+1)} = (Y - X\gamma^{(t+1)})^T [(Y - X\gamma^{(t+1)}) \# W_\beta^{(t)}] \\ [1^T W_\beta^{(t)} - \beta(1 + \beta)^{-3/2}]^{-1} \quad (15)$$

where

$$W_\beta = \left[\exp \left\{ -\frac{\beta}{2} \left(\frac{y_j - X_j \gamma}{\sigma} \right)^2 \right\} \right]_{n \times 1} \quad (16)$$

which is the vector of β -weights under H_0 . Thus the β -LOD score for the evidence of a QTL is given by

$$\text{LOD}_\beta = 2n \left\{ \sup_{\Theta} L_\beta(\theta | Y, X) - \sup_{\Theta_0} L_\beta(\theta | Y, X) \right\} \quad (17)$$

where Θ_0 and Θ are the restricted and unrestricted parameter spaces as before. For $\beta \rightarrow 0$, the LOD_β reduces to the Log-likelihood Ratio Test (LRT) criterion as defined by equation (3). During iteration, first component of γ should be initialised by the median of the phenotypic observations.

3.1 Robustness

The minimum β -divergence estimators for $\theta = \{a, \gamma, \sigma^2\}$ as defined in equations (10)–(12) under full model (H_1) and (14)–(15) under reduced model (H_0) are all weighted estimators. All estimators are weighted by β -weights described in Π_β and W_β under H_1 and H_0 , respectively. In both Π_β and W_β as defined in equations (13) and (16), a common scaling factor

$$\exp \left\{ -\frac{\beta}{2} \left(\frac{y - \mu}{\sigma} \right)^2 \right\}$$

produces larger β -weights with the usual phenotypic observations and smaller β -weights with the outlying observations those are far from the mean μ during parameter estimation with $\beta > 0$. For a wide range of $\beta > 0$ denoted by R_β , β -weights corresponding to only outlying observations reduce to almost zero. Thus outlying observations cannot influence the minimum β -divergence estimators for $\beta \in R_\beta$. A larger β increase the robustness but decrease the efficiency of the estimators and vice versa for the smaller β . Therefore, a smaller $\beta \in R_\beta$ is better than the larger $\beta \in R_\beta$ for robust and efficient estimation. However, R_β should satisfy $0 \leq R_\beta \leq 1$ (Basu et al., 1998; Minami and Eguchi, 2002).

4 Simulation study

To illustrate the performance of the proposed method in a comparison of traditional IM (Lander and Botstein, 1989) and CIM (Kao and Zeng, 1997) algorithms for QTL mapping, we consider backcross population for simulation study. Let us first assume only one QTL on a chromosome with 10 equally spaced markers, where any two successive marker interval size is 5 cM. The QTL position is located in the middle of chromosome 10. The true values for the parameters in the CIM model are assumed as $\mu = 0.05$, $a = 0.5$, $\gamma = 0.5$ and $\sigma^2 = 1$. To test the null hypothesis $H_0: a = 0$ against the existence of a QTL ($a \neq 0$), we generated 250 trait values with heritability $h^2 = 0.1$ using the CIM model as defined in equation (1). Figure 1(a) represent the scatter plot of 250 trait values and a covariate. To investigate the robustness of each of the three methods, we contaminated 15% trait values in this dataset by outliers. Figure 1(b) shows the scatter plot of contaminated dataset. Then we computed LOD scores by IM, CIM and the proposed methods for both types of datasets. 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 CIM method and the β -LOD scores of proposed method, respectively. Figure 1(c) shows the LOD scores profile for the uncontaminated dataset, where dotted, dot dash and solid lines represents the LOD scores at every 2 cM position in the chromosomes for IM, CIM and the proposed method with $\beta = 0.2$, respectively. Figure 1(d) shows the LOD scores profile for the contaminated dataset, where dotted, dot dash and solid lines represents the LOD scores at every 2 cM position in the chromosomes as before for IM, CIM 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 10 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.

We also investigate the performance of the proposed method with interval size 15 cM between two successive markers with the previous setting. We observe that the performance of the proposed method in a comparison of the IM and CIM methods are almost same as previous. We also observe that the performance of these three methods are good in presence of smaller phenotypic outliers and high leverage points (outliers with covariates). The performance of IM and CIM are not good in presence of larger phenotypic outliers, while in this case performance of the proposed method is good only. It should be noted here that the mixing proportions (π_{ji} ; $i = 1, 2$) of Gaussian mixture distribution for each individual ($j = 1, 2, \dots, n$) are computed based on QTL position parameter p using R/qtl package (Broman et al. (2003), homepage: <http://www.rqtl.org/>), where p is determined based on pairwise marker genotypes using Haldane’s map function. Other parameters for CIM model are estimated iteratively by R programming. Results for IM are obtained by R/qtl software.

To investigate the performance of the proposed method in a comparison of traditional IM and CIM algorithms in presence of multiple unlinked QTLs based on simulated data for 200 backcross progeny in an organism with 19 chromosomes of 50 cM each, we generated the quantitative phenotype for each individual by summing individual allelic effects at 3 QTLs and adding a covariate

with random environmental normal noise. For each individual, crossovers were generated assuming no interference and genotypes recorded with 10 equally spaced markers throughout the genome. The QTLs positions were located on the middle of chromosomes 10, 13 and 16, respectively. Figure ??(a) represent the scatter plot of 200 trait values and a covariate. To discuss the robustness of each of the three methods in presence of multiple unlinked QTLs, we contaminated around 20% trait values in this dataset by outliers (+). Figure ??(b) shows the scatter plot of contaminated dataset. Then we compute LOD scores by IM, CIM and the proposed methods for both types of datasets. Figure ??(c) shows the LOD scores profile for the uncontaminated dataset, where dotted, dot dash and solid lines represents the LOD scores at every 2 cM position in the chromosomes for IM, CIM and the proposed method with $\beta = 0.2$, respectively. Figure ??(d) shows the LOD scores profile for the contaminated dataset, where dotted, dot dash and solid lines represents the LOD scores at every 2 cM position in the chromosomes as before for IM, CIM and the proposed method with $\beta = 0.2$, respectively. It is seen that the higher and statistically significant LOD score peak occurs in the true QTL positions of chromosomes 10, 13 and 16 by all three methods for the uncontaminated dataset. However, in presence of outliers, the higher and significant LOD score peak occurs in the true 3 unlinked QTL positions by the proposed method only.

4.1 *An example of real data analysis*

To investigate the performance of the proposed method for real data analysis in a comparison of traditional IM and CIM algorithms, we consider the dataset of Sugiyama et al. (2001) which is available in R/qtl package (Broman et al. (2003), homepage: <http://www.rqtl.org/>). This dataset was analysed to investigate the genetic control of salt-induced hypertension on male mice from a reciprocal backcross between the salt-sensitive c57BL/6J and the non-salt-sensitive A/J(A) inbred mouse strains. Figure ??(a) represent the high blood pressure of 250 male progeny backcross to B6. To discuss the robustness of each of the three methods in the case real data analysis, we contaminated around 15% high blood pressure in this dataset by outliers (+). Figure ??(b) shows the scatter plot of contaminated dataset. Then we computed LOD scores by IM, CIM and the proposed methods for both types of datasets. Figure ??(c) shows the LOD scores profile for the uncontaminated dataset, where dotted, dot dash and solid lines represents the LOD scores at every 2 cM position in the chromosomes for IM, CIM and the proposed method with $\beta = 0.2$, respectively. Figure ??(d) shows the LOD scores profile for the contaminated dataset, where dotted, dot dash and solid lines represents the LOD scores at every 2 cM position in the chromosomes as before for IM, CIM and the proposed method with $\beta = 0.2$, respectively.

Genome-wide LOD thresholds are obtained by permutation tests (Churchill and Doerge, 1994), using 10,000 permutation replicates. For comparisons with the existing results of Sugiyama et al. (2001), we estimated 95% (63%) genome-wide LOD thresholds for IM, CIM and the proposed methods those are 2.9 (1.8), 6.8 (5.7) and 6.5 (5.6), respectively. Because of the larger differences in the LOD thresholds for the three methods, we subtracted $(6.8 - 2.9) = 3.9$ and $(6.5 - 2.9) = 3.6$ from the LOD scores of CIM and the proposed methods

respectively, for convenience of presentation and comparison of three methods using the same decision boundary at a LOD threshold 2.9. In both Figures ??(c) and (d), the long-dash lines parallel to the x -axis indicate genome-wide suggestive (LOD threshold 1.8 with p -value ≤ 0.37) and significant (LOD threshold 2.9 with p -value ≤ 0.05) QTLs associated with the blood pressure, respectively. Figure ??(c) shows that two QTLs on chromosome 1 and two QTLs on chromosome 4 are statistically highly significant genome-wide, and one on each of chromosomes 2, 5, 6 and 15 are genome-wide suggestive by all three methods for the uncontaminated real dataset. The same results were also reported by Sugiyama et al. (2001) using one-way ANOVA. However, in presence of outliers, almost similar results are obtained by the proposed method only as shown in Figure ??(d) using solid line. Therefore, the proposed method significantly improves the performance over the traditional IM and CIM methods in presence of outliers; it keeps equal performance otherwise.

5 Conclusion

This paper discusses a new robust QTL mapping algorithm based on CIM model in an experimental organisms by minimising β -divergence using the EM like algorithm. The proposed method with $\beta = 0$ reduces to the traditional CIM method. The value of the tuning parameter β plays a key role on the performance of the proposed method. An appropriate value for the tuning parameter β may be selected by cross validation. Based on our experience, we can select an appropriate β within 0.1–0.5 such that maximum 50% components of W_β reduces to zero. It should be noted here that smaller β is better than the larger β for robust and efficient estimation. Therefore, our suggestion is to use $\beta = 0.2$ for data analysis. However, we would like to discuss an adaptive selection procedure for the tuning parameter β in the extended version of this paper in near future. Simulation studies including real data analysis show that the proposed method significantly improves the performance over the traditional IM and CIM methods in presence of outliers; otherwise, it keeps equal performance.

Acknowledgements

The authors would like to acknowledge the many helpful suggestions of three anonymous reviewers and the participants of the 2008 BIBM Conference on earlier versions of this paper. We also thank the Editor of this Journal.

References

- Basu, A., Harris, I.R., Hjort, N.L. and Jones, M.C. (1998) ‘Robust and efficient estimation by minimising a density power divergence’, *Biometrika*, Vol. 85, pp.549–559.
- Broman, K.W., Sen, W.H. and Churchill, G.A. (2003) ‘R/qtl: QTL mapping in experimental crosses’, *Bioinformatics*, Vol. 19, pp.889–890.

- Carbonell, E.A., Gerig, T.M., Balansard, E. and Asin, M.J. (1992) 'Interval mapping in the analysis of non-additive quantitative trait loci', *Biometrics*, Vol. 48, pp.305–315.
- Churchill, G.A. and Doerge, R.W. (1994) 'Empirical threshold values for quantitative trait mapping', *Genetics*, Vol. 138, pp.963–971.
- Haley, C.S. and Knott, S.A. (1992) 'A simple regression method for mapping quantitative trait in line crosses using flanking markers', *Heredity*, Vol. 69, pp.315–324.
- Jansen, R.C. (1992) 'A general mixture model for mapping quantitative trait loci by using molecular markers', *Theor. Appl. Genet.*, Vol. 85, pp.252–260.
- Jansen, R.C. (1993) 'Interval mapping of multiple quantitative trait loci', *Genetics*, Vol. 135, pp.205–211.
- Kao, C.H. and Zeng, Z.B. (1997) 'General formulae for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm', *Biometrics*, Vol. 53, pp.653–665.
- Lander, E.S. and Botstein, D. (1989) 'Mapping mendelian factors underlying quantitative traits using RFLP linkage maps', *Genetics*, Vol. 121, pp.185–199.
- Minami, M. and Eguchi, S. (2002) 'Robust blind source separation by beta-divergence', *Neural Computation*, Vol. 14, pp.1859–1886.
- Mollah, M.N.H. and Eguchi, S. (2008) 'Robust composite interval mapping for QTL analysis by minimum β -divergence method', *IEEE International Conference 2008 on Bioinformatics and Biomedicine (IEEE BIBM 2008)*, Philadelphia, USA, pp.115–120.
- Thoday, J.M. (1960) 'Effects of disruptive selection. III. coupling and repulsion', *Heredity*, Vol. 14, pp.35–49.
- Sugiyama, F., Churchill, G.A., Higgins, D.C., Johns, C., Makaritsis, K.P., Gavras, H. and Paigen, B. (2001) 'Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci', *Genomics*, Vol. 71, No. 1, pp.70–77.
- Zeng, Z.B. (1993) 'Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci', *Proc. Nat. Acad. Sci. USA*, Vol. 90, Washington DC, USA, pp.10972–10976.
- Zeng, Z.B. (1994) 'Precision mapping of quantitative trait loci', *Genetics*, Vol. 136, pp.1457–1468.

Bibliography

- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) 'Maximum likelihood from incomplete data via the EM algorithm', *J. Roy. Statist. Soc. B*, Vol. 39, pp.1–38.
- Fujisawa, H. and Eguchi, S. (2006) 'Robust estimation in the normal mixture model', *Journal of Statistical Planning and Inference*, Vol. 136, pp.3989–4011.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.
- Louis, T.A. (1982) 'Finding the observed information matrix when using the EM algorithm', *Journal of Royal Statistical Society, Series B*, Vol. 44, pp.226–233.
- Mollah, M.N.H., Minami, M. and Eguchi, S. (2006) 'Exploring latent structure of mixture ICA models by the minimum β -divergence method', *Neural Computation*, Vol. 18, No. 1, pp.166–190.