Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 6423 – 6434

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Locating multiple interacting quantitative trait loci using robust model selection

Andreas Baierl[a,*], Andreas Futschik[a], Małgorzata Bogdan[b], Przemysław Biecek[b]

[a] *Institute of Statistics and Decision Support Systems, University of Vienna, Universitätsstrasse 5/9, A-1010 Vienna, Austria*
[b] *Institute of Mathematics and Computer Science, Wrocław University of Technology, 50-370 Wrocław, Poland*

Received 31 March 2006; received in revised form 8 February 2007; accepted 9 February 2007
Available online 20 February 2007

Abstract

One of the most popular criteria for model selection is the Bayesian Information Criterion (BIC). It is based on an asymptotic approximation using Bayes rule when the sample size tends to infinity and the dimension of the model is fixed. Although it works well in classical applications, it performs less satisfactorily for high dimensional problems, i.e. when the number of regressors is very large compared to the sample size. For this reason, an alternative version of the BIC has been proposed for the problem of mapping quantitative trait loci (QTLs) considered in genetics. One approach is to locate QTLs by using model selection in the context of a regression model with an extremely large number of potential regressors. Since the assumption of normally distributed errors is often unrealistic in such settings, we extend the idea underlying the modified BIC to the context of robust regression.

Keywords: QTL mapping; BIC; Model selection; M-estimates

1. Introduction

Although the Bayesian Information Criterion (BIC) is a very popular criterion for model selection, its behavior is less satisfactory in high dimensional regression models. This is maybe not unexpected, since the BIC proposed by Schwarz (1978) is based on an asymptotic approximation using the Bayes rule to derive posterior probabilities for all the competing submodels of a regression model. The BIC cannot be expected to provide a good approximation in cases where the number of potential regressors is large compared to n. Such problems are common in the context of mapping quantitative traits (quantitative trait locus (QTL) mapping) in genetics, where the number of potential regressors is often much larger than the number of observations. In such a setting it has been observed, for instance by Broman (1997) and Broman and Speed (2002), that the classical BIC has a strong tendency to overestimate model size. To understand this phenomenon, notice that the arguments regarding the asymptotics, which lead to the BIC, imply that the prior is negligible and as a consequence all models are taken to be equally probable by the BIC. However, if the number of regressors is very large, then there are many more high dimensional models than low dimensional ones (when there are p^* potential regressors there are actually $\binom{p^*}{k}$ submodels of dimension k). As a consequence, it is likely that some of these higher dimensional models lead to a low value of the BIC just by chance.

[*] Corresponding author. Tel.: +43 1 4277 38601; fax: +43 1 4277 9386.
E-mail address: andreas.baierl@univie.ac.at (A. Baierl).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.02.010

For these reasons, a modified version of the BIC, the mBIC, has been proposed by Bogdan et al. (2004) and Baierl et al. (2006) in the context of QTL mapping. The method is based on supplementing the BIC with a correction based on a more realistic prior distribution over the set of possible models and leading to a higher penalty for larger models.

The mBIC is based on standard L_2 regression, assuming that the conditional distribution of the trait given the marker genotypes is normal. In practice this assumption is rarely satisfied. While the Central Limit Theorem shows that moderate deviations from normality have little influence on the mBIC, the properties of this criterion deteriorate drastically when the distribution of the trait has a heavy tail or the data include a certain proportion of outliers. Thus we consider an alternative approach based on robust regression techniques and construct robust versions of the mBIC. For this purpose, we use several well-known contrast functions and investigate the resulting versions of the mBIC both analytically and using computer simulations. It turns out that the robust versions of the mBIC perform consistently well for all the distributions analyzed and much better than the standard version of the mBIC in situations when the distribution of the trait is heavy tailed. A possible exception to this rule is Huber's contrast function with a very small k, which we considered as a close approximation to L_1 regression. The corresponding procedure was outperformed by the other considered procedures under several models describing the error.

While the basic idea of including a prior that penalizes high dimensional models more heavily should be of interest in a more general context, the modification of the BIC must be adapted to the structure of the model. Thus our investigations are carried out in an ANOVA setting with one-way interactions, as encountered in the context of QTL mapping. Consequently, our regressors are taken to be the dummy variables which describe the genotypes of the markers, as well as products of pairs of these dummy variables.

The contents of this paper are as follows. In Section 2, we discuss potential applications of our approach in genetics. Section 3 contains a detailed discussion of the underlying statistical model and mBIC. Section 4 introduces robust methods of model selection and the associated versions of the mBIC. Results from simulations and analysis of real data are included in Sections 5 and 6, respectively. Section 7 contains some concluding remarks.

2. Genetic background

A QTL is a genome position hosting genes that influence a quantitative trait of interest (e.g. height, yield). Locating QTLs is one of the major tasks of quantitative genetics. Apart from broadening general biological knowledge, QTL mapping is often used to detect genes influencing economically important traits in domesticated animals and crops (see e.g. Chardon et al., 2004; Khatkar et al., 2004; Walling et al., 2000). In humans, QTL mapping is applied to locate genes responsible for the development of certain medical conditions (see e.g. Dick and Foroud, 2002, as well as examples and references in Lynch and Walsh, 1998).

In order to locate QTLs, geneticists use molecular markers. These are pieces of DNA that exhibit variation between individuals. Their characteristics (i.e. genotypes) can be determined experimentally. In organisms where chromosomes occur in pairs (i.e. diploid organisms), the genotype at a particular locus is specified by two pieces of DNA that are potentially different. From a statistical point of view, marker genotypes can be treated as qualitative explanatory variables. If a QTL is located close to a given marker, we expect to see an association between the genotype of the marker and the value of the trait.

Locating QTLs in natural, outbred, populations is relatively difficult. This is due to the fact that as a result of crossover events, which occur every time gametes are produced, the association between a QTL and a neighboring marker may be very weak. Therefore, to control the number of crossovers, scientists usually use data from families with more than one offspring or extended pedigrees (see e.g. Lynch and Walsh, 1998; Thompson, 2000). Locating QTLs is easier and usually more precise when the data come from experimental populations. Such populations consist of inbred lines of individuals, who are homozygous at every locus on the genome (i.e. have identical pairs of chromosomes). By crossing individuals from such inbred lines, scientists can control the number of meioses (the process in which gametes are produced) and produce large experimental populations for which the correlation structure between the genotypes at different markers is easy to predict. Inbred lines have been created in many species of plant, as well as animal species (e.g. mice). Results from research on experimental populations can often be used to predict biological phenomena in an outbred population, due to the similarity of genomes in the two populations (see e.g. Philips, 2002).

A summary of standard methods of QTL mapping used in experimental populations can be found, for example, in Lynch and Walsh (1998), Doerge et al. (1997) or Doerge (2002). This article is devoted to detecting QTLs in backcross populations. According to such a design, two inbred lines are crossed to produce offspring, who are crossed with one of the parental lines. The resulting offspring exhibits one of only two possible genotypes at any particular locus: either an individual is homozygous for the allele (version of the gene) from the parental line used at the second level of crossing, or heterozygous (has alleles from both parental lines).

One of the methods of locating QTLs relies on choosing markers closely linked to a QTL by fitting a multiple regression model relating the values of the trait values to the genotypes of the marker. The most difficult part of this process is to estimate the number of QTLs. For this purpose, one could use one of many model selection criteria, such as Akaike's Information Criterion (Akaike, 1974) or the Bayesian Information Criterion (BIC) proposed by Schwarz (1978). These criteria have been used or discussed in the context of QTL mapping e.g. by Broman (1997), Piepho and Gauch (2001), Ball (2001), Broman and Speed (2002), Bogdan et al. (2004), Siegmund (2004) and Baierl et al. (2006). In particular Broman (1997) and Broman and Speed (2002) observed that the usually conservative BIC has a strong tendency to overestimate the number of QTLs. Bogdan et al. (2004) explain this phenomenon by observing that a serious problem related to multiple testing arises from the use of a large number of markers, as is normally the case in typical genome searches. This requires a modification of the standard criteria for model selection. The problem of overestimation is exacerbated when one also looks for epistatic QTLs, which influence the trait only by interacting with other genes. The modification (mBIC) of the BIC proposed by Bogdan et al. (2004) uses an additional penalty for larger models taking prior knowledge on the number of QTLs into account. In the case where no prior knowledge on the number of QTLs is available, the penalty is calibrated so that the overall probability of a type I error is controlled at a level close to 0.05 for typically used sample sizes ($n > 200$). Bogdan et al. (2004) and Baierl et al. (2006) report the results of an extensive range of simulations demonstrating the properties of the mBIC under a wide range of possible genetic models.

Finally, we wish to mention that linear models have also been used to explore gene regulatory networks (see for instance Bussemaker et al., (2001) Keles et al., (2004)). From a statistical point of view, this may be viewed as a problem of selection from a massive range of models which also involve interactions. Thus our ideas might also be of interest in this setting.

3. The statistical model

We start by briefly reviewing the multiple regression model considered. It is an ANOVA model with one-way interactions and can be used for QTL mapping within the context of the backcross design. Let X_{ij} denote the genotype of the ith individual at the jth marker. We set $X_{ij} = -\frac{1}{2}$, if the ith individual is homozygous at the jth marker and $X_{ij} = \frac{1}{2}$, if it is heterozygous. We fit a multiple regression model of the form

$$Y_i = \mu + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv} + \varepsilon_i, \tag{1}$$

where I is a subset of the set $N = \{1, \ldots, n_m\}$, n_m denotes the number of markers available, U is a subset of $N \times N$ and ε_i is a random error term. In order to identify markers which are close to a QTL, we need to select an appropriate submodel. For this purpose, we allow the inclusion of interaction terms in the model, even when the corresponding main effects are not included. This approach is justified by the recent discoveries of genes that do not have their own additive effects, but only influence a trait by interacting with other genes (see e.g. Fijneman et al., 1996, 1998).

One of the most popular tools for choosing influential regressor variables is the BIC, which recommends choosing the simplest model for which

$$n \log RSS + k \log n \tag{2}$$

obtains its minimal value. Here RSS is the residual sum of squares from regression, k is the number of regressor variables and n is the sample size (number of individuals considered). As mentioned above, the BIC has a tendency to overestimate the number of QTLs. The modified version of the BIC, the mBIC, proposed by Bogdan et al. (2004), exploits the Bayesian context of the BIC and supplements this criterion with additional terms taking into account a realistic prior distribution on the number of QTLs. The resulting mBIC is based on minimizing the following quantity:

$$mBIC = n \log RSS + (p + q) \log n + 2p \log(l - 1) + 2q \log(u - 1),$$

where p is the number of main effects in the model and q is the number of epistatic terms. The coefficients of the additional penalty terms, $l := n_m/\mathrm{E}N_m$ and $u := n_e/\mathrm{E}N_e$, depend on the number of possible regressors n_m and

$n_e = n_m(n_m - 1)/2$ (corresponding to the number of main and interaction effects, respectively), as well as the expected numbers $\mathrm{E}N_v$ and $\mathrm{E}N_e$ of main and interaction effects, where these expectations are chosen according to given prior distributions. A prior can sometimes be obtained by looking at similar experiments that have been carried out previously. In the absence of prior information, Bogdan et al. (2004) propose the use of $\mathrm{E}N_m = \mathrm{E}N_e = 2.2$. This choice guarantees that the probability of a type I error under H_0 using the appropriate procedure (i.e. detecting at least one QTL when there are none) is smaller than 0.07 for sample sizes $n \geqslant 200$ and a moderate number of markers ($M > 30$).

Bogdan et al. (2004) and Baierl et al. (2006) present the results of an extensive range of simulations, which confirm the good properties of the standard and extended version of mBIC when applied to QTL mapping.

For normally distributed errors, ε_i, classical least squares regression is well justified, for instance, by the Gauss–Markov Theorem. As a result of the Central Limit Theorem, if the sample is large enough, then the standard test procedures for the significance of regression coefficients are resistant to moderate deviations from the assumption of normality. It can be seen that this property is shared by the BIC and its modification for mapping QTLs, the mBIC. However, the estimates and tests derived under the assumption of normality cannot be expected to work well in the cases of skewed and heavy tailed distributions. In some situations, as in the case of the Cauchy distribution, the Central Limit Theorem does not even hold. It is also well known that standard L_2 regression is highly sensitive to outliers.

Methods of robust regression provide an alternative in such situations. They perform well under a wide range of error distributions without losing too much power when normality holds. One approach to obtaining robust regression estimates is to use M-estimates, i.e. to minimize another measure of distance instead of the residual sum of squares. Another approach would be to use MM-estimates, which have the additional property of also being robust with respect to leverage points (see Yohai, 1985). However, the model considered only contains dummy variables and, due to appropriate randomization, the design is close to being balanced. Therefore, leverage points should not be expected. We thus focus on M-estimates but mention here that our simulations gave nearly identical results for both M-estimates and the corresponding MM-estimates.

M-estimates of the regression parameters are based on the minimization of $\sum_{i=1}^{n} \rho(r_i)$, where the r_i are the residuals standardized using a robust scale estimator and $\rho(x)$ is a contrast function. We consider the following popular contrast functions:

$$\rho_{\text{Huber}}(x) := \begin{cases} k|x| - k^2/2 & \text{for } |x| > k, \\ x^2/2 & \text{for } |x| \leqslant k, \end{cases} \tag{3}$$

$$\rho_{\text{Bisquare}}(x) := \begin{cases} k^2/6 & \text{for } |x| > k, \\ \dfrac{k^2}{6}\left[1 - \left(1 - \left(\dfrac{x}{k}\right)^2\right)^3\right] & \text{for } |x| \leqslant k, \end{cases} \tag{4}$$

$$\rho_{\text{Hampel}}(x) := \begin{cases} a(b - a + c)/2 & \text{for } |x| > c, \\ a(b - a + c)/2 - \dfrac{a(|x| - c)^2}{2(c - b)} & \text{for } b < |x| \leqslant c, \\ a|x| - a^2/2 & \text{for } a < |x| \leqslant b, \\ x^2/2 & \text{for } |x| \leqslant a. \end{cases} \tag{5}$$

The calculation of regression coefficients based on these contrast functions requires an iterative method, such as iteratively reweighted least squares. In our simulations we standardized residuals using the median absolute deviation (MAD) from the median. For details on this and other aspects of robust regression (e.g. confidence regions and tests for M-estimates), see Chapter 7 in Huber (1981). Applications of robust regression have been discussed, for instance, in Carroll (1980).

Notice that for small k, Huber's contrast function is very close to the objective function $\rho(x) = |x|$ used in L_1 regression. Among others, we will consider such a version of Huber's M-estimate and expect it to provide some insight concerning the performance of L_1 regression. We refer to Bassett and Koenker (1978) for a more detailed discussion of L_1 regression.

In the next section of the paper we discuss the problem of model selection in the context of robust regression.

4. Robust model selection and the modified BIC

A natural way to obtain a robust version of the BIC (or the mBIC) is to replace the residual sum of squares in the criterion for model selection by a sum of contrasts. However, unlike in the L_2 case, a sum of contrasts will usually not be scale invariant. We therefore propose to standardize the Y_i in a robust way and work with standardized observations, $Y_i^{(\mathrm{s})}$, obtained by subtracting the median and dividing by the MAD as defined in Ronchetti et al. (1997). Notice that it is necessary to use the same estimate of the MAD in all the models considered, in order to make comparisons between models possible. We therefore propose to use the MAD calculated under the null model of no effect for the purposes of model selection. For the purpose of estimating the parameters of the various regression models, we additionally rescale the residuals separately for each model. In our simulations, we used the robust rescaling provided by the function 'rlm' in the library MASS of the R package, which is available under http://www.R-project.org.

This leads to the following robust version of the BIC (see (2)):

$$BIC_{\rho}^{*} := n \log \sum_{i=1}^{n} \rho(Y_i^{(\mathrm{s})} - x_i'\hat{\theta}) + k \log(n). \tag{6}$$

Here, x_i' denotes the vector of regressors in the model (see (1)) and $\hat{\theta}$ contains the regression coefficients estimated using $\rho(\cdot)$ as the contrast function.

An alternative approach proposed by Ronchetti et al. (1997) in the context of model selection is to rescale the contrast function instead of standardizing the Y_i's. As before, the same rescaling factors have to be used in all models. Ronchetti et al. (1997) propose estimation of the rescaling factors based on the largest possible model. This is not possible in our setup, since the largest possible model usually contains many more variables than observations. We therefore modified their approach and estimated rescaling constants under the null model. This way of rescaling the contrast function led to results that were almost identical to those obtained after standardizing Y using the MAD from the null model.

It has been shown by Machado (1993) that the robust BIC (6) is still consistent under quite general conditions on the error distribution. Martin (1980), as well as Ronchetti (1985), used similar ideas, in order to make the AIC robust. However, consistency is a minimal requirement, since the actual performance of BIC_{ρ}^{*} will depend both on $\rho(x)$ and the error distribution. Indeed this dependence becomes apparent from results in Jurečková and Sen (1996, p. 410), who derived the limiting distribution of

$$\sum_{i=1}^{n} (\rho(Y_i^{(\mathrm{s})} - x_i'\hat{\theta}_1) - \rho(Y_i^{(\mathrm{s})} - x_i'\hat{\theta}_2))$$

for a fixed null model M_1 versus a higher dimensional model M_2.

In order to obtain more reliable performance of the robust BIC for different contrast functions $\rho(x)$ and error distributions, it seems natural to renormalize BIC_{ρ}^{*} by taking the limiting distribution of

$$D_n = n \left(\log \sum \rho(Y_i^{(\mathrm{s})} - x_i'\hat{\theta}_1) - \log \sum \rho(Y_i^{(\mathrm{s})} - x_i'\hat{\theta}_2) \right)$$

into account. Ideally, $BIC(M_2) - BIC(M_1)$ should have the same asymptotic distribution for different ρ and error models. For this purpose, we will derive the asymptotic distribution of D_n using results from Jurečková and Sen (1996), as well as the delta method. Since the asymptotic distribution of D_n depends not only on ρ, but also on the unknown error distribution, the required normalization constant needs to be estimated.

We compare model M_1 with parameter vector θ_1 of dimension $p_1 + q_1$ and model M_2 with parameter vector θ_2 of dimension $p_2 + q_2$. M_1 is assumed to be a submodel of M_2.

Theorem 1. *Let ρ be a contrast function satisfying the regularity conditions specified in Chapter* 5.5 *of* Jurečková and Sen (1996). *Define $\tilde{Y}_i^{(s)}$ to be the observations standardized according to the population median and the population MAD. Furthermore, define the score function $\psi(x) = \rho'(x)$. Let $\gamma = \int \psi'(x) f(x)\,\mathrm{d}x$, $\sigma_{\psi}^2 = \int \psi(x)^2 f(x)\,\mathrm{d}x$ and $\delta = \int \rho(x) f(x)\,\mathrm{d}x$, where $f(x)$ denotes the density function of the error distribution under the true regression model based on the observations $\tilde{Y}_i^{(\mathrm{s})}$. Moreover, let us define the constant $c_e = 2\gamma\delta/\sigma_{\psi}^2$.*

Then under model M_1

$$c_e D_n \xrightarrow{\mathrm{d}} \chi^2_{(p_2+q_2)-(p_1+q_1)} \tag{7}$$

as the sample size $n \to \infty$.

Proof. Let $l_n(\theta) = \sum \rho(Y_i^{(\mathrm{s})} - x_i'\theta)$, where θ is the true parameter vector. Note that

$$\frac{1}{n} l_n(\theta) = \frac{1}{n} \sum \rho(Y_i^{(\mathrm{s})} - x_i'\theta) \xrightarrow{\mathrm{p}} \mathrm{E}[\rho(\tilde{Y}_i^{(\mathrm{s})} - x_i'\theta)] = \delta.$$

From the consistency of M-estimates (see e.g. Huber, 1981) and the uniform continuity of l_n in a neighborhood of θ, it also holds that both

$$\frac{1}{n} l_n(\hat{\theta}_1) \xrightarrow{\mathrm{p}} \delta \quad \text{and} \quad \frac{1}{n} l_n(\hat{\theta}_2) \xrightarrow{\mathrm{p}} \delta, \tag{8}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the M-estimates under models M_1 and M_2, respectively. Approximating D_n by the first term of its Taylor series expansion leads to

$$n(\log l_n(\hat{\theta}_1) - \log l_n(\hat{\theta}_2)) = n\left(\log\left(1 + \frac{l_n(\hat{\theta}_1) - l_n(\hat{\theta}_2)}{l_n(\hat{\theta}_2)}\right)\right) \tag{9}$$

$$= \frac{l_n(\hat{\theta}_1) - l_n(\hat{\theta}_2)}{(1/n) l_n(\hat{\theta}_2)} + R_n, \tag{10}$$

where

$$R_n = \mathrm{O}\left(\frac{(l_n(\hat{\theta}_1) - l_n(\hat{\theta}_2))^2}{(1/n) l_n^2(\hat{\theta}_2)}\right).$$

Jurečková and Sen (1996, pp. 408–416, note their discussion of the extension of their results to studentized observations) proved that

$$\frac{2\gamma}{\sigma_\psi^2}(l_n(\hat{\theta}_1) - l_n(\hat{\theta}_2)) \xrightarrow{\mathrm{D}} \chi^2_{(p_2+q_2)-(p_1+q_1)}. \tag{11}$$

Thus to conclude (7), it is enough to observe that from (8) and (11)

$$R_n \xrightarrow{\mathrm{p}} 0$$

as $n \to \infty$. □

For least squares regression, $\rho(Y_i^{(\mathrm{s})} - x_i'\theta)$ is equal to the residual sum of squares. In this situation, asymptotically D_n has a χ^2-distribution with $(p_2 + q_2) - (p_1 + q_1)$ degrees of freedom (Serfling, 1980). Hence, the normalization constant c_e is equal to 1 in this case.

As can be seen from Theorem 1, specific normalization constants (c_e) depend on the error distribution and the ρ-function considered. If the error distribution is assumed to be known, c_e can be derived analytically. The appropriate values for chosen distributions is shown in Table 1. In practice, the error distribution and c_e have to be estimated. For this purpose, we first carry out model selection with c_e equal to 1, i.e. the normalizing constant for Gaussian errors. The empirical distribution of the resulting residuals is then used to approximate the expected values, defining γ, σ_ψ^2 and δ by the corresponding averages (see e.g. p. 409 of Jurečková and Sen, 1996). Plugging in these quantities leads to the estimate $\hat{c}_e$.

The discussion above leads us finally to the following robust version of the mBIC:

$$mBIC = \hat{c}_e n \log \sum \rho(Y_i^{(\mathrm{s})} - x_i'\hat{\theta}) + (p+q)\log n + 2p\log(l-1) + 2q\log(u-1), \tag{12}$$

where ρ is a given contrast function and $\hat{\theta}$ is the corresponding M-estimate of the $(p+q)$-dimensional parameter vector of the model considered.

Table 1
Values for normalization constants

Error distr.	$\text{Huber}_{k=0.05}$	$\text{Huber}_{k=1.345}$	Bisquare	Hampel
Normal	1.267	1.079	1.095	1.025
Laplace	1.967	1.397	1.387	1.254
Cauchy	[a]	[a]	2.242	2.428
Tukey	1.770	1.952	1.408	1.653
χ^2	1.199	1.153	1.164	1.125
χ^2_{med}	1.295	1.255	1.248	1.165

The definitions of the error distributions are given in Section 4.2.

[a] In the case of the Cauchy distribution, the integral for ρ_{Huber} leading to δ is infinite.

5. Comparison of performance under different error models

5.1. Design of the simulations

Simulations are carried out to compare the performance of least squares regression and robust methods for QTL mapping under a variety of error distributions. We consider M-estimates for robust models based on the following contrast functions: ρ_{Huber}, ρ_{Bisquare} and ρ_{Hampel}. The parameters in the ρ-functions are set to $a = 2$, $b = 4$ and $c = 8$ for Hampel's function, $k = 1.345$ for Huber's function, and $c = 4.685$ for Tukey's bisquare M-estimator. These are the default parameter values used in the R-package MASS, which was used to obtain robust regression estimates. We also chose $k = 0.05$ for Huber's M-estimator as a close, smooth approximation to the L_1 contrast function $\rho(x) = |x|$. Notice that due to the smoothness of ρ_{Huber}, Theorem 1 still applies.

The model selection process was carried out using the standard version of the mBIC (12) with $l = n_m/2.2$ and $u = n_e/2.2$. To solve the problem of searching over a large class of possible models, we use forward selection.

Three arrangements of marker genotypes with a backcross population of 200 individuals were simulated.

Arrangement 1: One chromosome of length 100 cM with five equally spaced markers.

Arrangement 2: Two chromosomes of length 100 cM both with 11 equally spaced markers.

Arrangement 3: Five chromosomes of length 100 cM each with 11 equally spaced markers.

Two scenarios were considered for each arrangement: a null model with no effects and a "3 QTL" model with one main effect of size $\beta = 0.55$ and one interaction effect (involving two loci) of size $\gamma = 1.2$. We assumed all the QTLs to be located at marker positions.

All the methods were applied to each of the arrangements under each of six different error distributions. 1000 replications were used for Arrangements 1 and 2, whereas we carried out 500 simulations for Arrangement 3, which is computationally quite demanding. The performance of each method was measured by the average number of correctly identified main and epistasis effects, as well as the false discovery rate defined as

$$FDR = \frac{1}{n} \sum_{i=1}^{n} \frac{fp.m_i + fp.e_i}{fp.m_i + fp.e_i + c.m_i + c.e_i},$$

where the quantities $fp.m_i$ and $fp.e_i$ denote the number of false positive main and epistasis effects that were detected in replication i, and $c.m_i$ and $c.e_i$ are the number of correctly identified main and epistasis effects, respectively. According to the definition of the FDR (see Benjamini and Hochberg, 1995), the terms in the sum corresponding to replicates with no detections are set to be equal to zero. Under the null model, the false discovery rate is equivalent to the multiple type I (or familywise) error of detecting at least one incorrect effect.

An inferred main effect was classified as being a false positive, if it was more than 15 cM away from the true QTL or the QTL had already been detected. An epistatic effect was classified as being a false positive, if at least one of the two QTLs involved was more than 15 cM from the true QTL. Notice that this definition is fairly strict, since effects that are not very strong often lead to the detection of markers that are further than 15 cM away from the true QTLs (see Bogdan and Doerge, 2005, for a discussion).

Table 2
Multiple type I errors for Arrangement 1

Estimate	Error distributions					
	Normal	Laplace	Cauchy	Tukey	Chisq	Chisq-med
Huber$_{k=0.05}$ theor.	10.3	11.6	[a]	11.7	7.3	9.6
Huber$_{k=0.05}$ est.	10.2	12.0	9.7	10.3	7.6	10.5
Huber$_{k=1.34}$ theor.	11.1	10.1	[a]	15.5	10.7	13.0
Huber$_{k=1.34}$ est.	12.8	12.1	10.6	10.9	10.6	11.9
Bisquare theor.	11.1	10.6	12.0	13.6	10.5	12.3
Bisquare est.	12.2	12.2	11.4	12.0	10.4	10.6
Hampel theor.	10.8	11.1	12.4	14.6	12.3	12.8
Hampel est.	11.9	11.3	9.5	12.0	11.0	12.3
L_2^{mBIC}	12.9	10.6	4.6	7.3	11.1	12.3
L_2^{BIC}	26.4	22.1	14.9	18.9	23.5	24.0

[a] Comparison of the probability of type I errors under the null model when the distribution of residuals is assumed to be known (theor., c_e) or has to be estimated (est., $\hat{c}_e$). For L_2 regression, c_e is always equal to 1. Definitions of the error distributions are given in Section 4.2.

5.2. Error distributions

We considered the following error distributions, which were all centered around the origin and standardized such that the inter-quartile range ($IQR = Q_{75} - Q_{25}$) was 1.5.

(1) Normal: $1.11 \times N(0, \sigma^2)$ with $\sigma = 1$.
(2) Laplace (double exponential): $1.08 \times (1/2\lambda)\exp(-|\lambda t|)$ with $\lambda = 1$.
(3) *Cauchy* ($scale = 0.75$) with $scale = 0.5 \times IQR$.
(4) Tukey's gross error model: $1.081 \times (\lambda N(0, \sigma^2) + (1 - \lambda)N(0, \tau\sigma^2))$ with $\lambda \sim$ *Bernoulli* ($p = 0.95$), $\sigma = 1$ and $\tau = 100$.
(5) χ^2 centered around the mean with 6 d.f.: $0.342 \times (\chi^2_6 - 6)$.
(6) χ^2 centered around the median with 6 d.f.: $0.342 \times (\chi^2_6 - \tilde{x}_6)$ with $\tilde{x}_6 = 5.348$.

5.3. Results of the simulations and discussion

We focus on two points in particular. The first issue is to investigate whether our approach of using estimated normalization constants leads to similar values for the multiple type I error and the false discovery rate under various models for the error. We call this property error-robustness. The second issue is whether the power of our procedure for model selection remains high under non-normal errors and when outliers are present. We call this property power-robustness.

We start by investigating the error-robustness. The probability of type I errors under the rules considered for selecting a model can be found in Tables 2–4. The two rows associated with each of the procedures present the results in the cases when the proper theoretical constant was used and when the constant was estimated (separately for each replicate), respectively. These tables show that the multiple type I errors using the mBIC and estimated normalizing constants are comparable for all of the procedures. As expected from the formulae given in Bogdan et al. (2004), the probability of multiple type I errors for Arrangement 1 (only five markers) are approximately 0.1 and lower for Arrangements 2 and 3. For most of the examples there is only a slight difference between the results obtained using theoretical and estimated constants. The slightly larger difference obtained for the procedure using the Huber contrast under both the Laplace and heavy-tailed distributions results from the problem of estimating the density of the Laplace distribution close to 0 (the Huber contrast with $k = 0.05$ assigns a relatively large weight to the corresponding residuals) and the tails of the Tukey distribution (the Huber contrast function tends to infinity as $x \to \pm\infty$). Also, note that the multiple type I error of the procedure for model selection based on the original BIC is much larger than desired and rapidly increases as the number of regressors becomes larger. This demonstrates the advantage of using the mBIC rather than the BIC.

Table 3
Multiple type I errors for Arrangement 2

Estimate	Error distributions					
	Normal	Laplace	Cauchy	Tukey	Chisq	Chisq-med
Huber$_{k=0.05}$ theor.	3.4	6.8	[a]	3.5	2.2	4.3
Huber$_{k=0.05}$ est.	4.2	5.3	6.2	4.6	5.6	5.6
Huber$_{k=1.34}$ theor.	4.4	5.1	[a]	7.2	4.7	7.1
Huber$_{k=1.34}$ est.	3.6	5.5	4.5	5.8	6.0	6.1
Bisquare theor.	4.1	4.1	4.6	4.8	4.5	6.7
Bisquare est.	3.4	5.0	5.4	5.2	5.4	5.8
Hampel theor.	4.2	5.6	5.9	6.3	5.1	6.5
Hampel est.	4.2	4.6	4.7	5.0	6.5	6.7
L_2^{mBIC}	4.4	4.6	8.8	3.4	6.1	5.7
L_2^{BIC}	90.4	89.8	81.6	86.9	88.4	88.6

[a] Comparison of the probability of type I errors under the null model when the distribution of residuals is assumed to be known (theor., c_e) or has to be estimated (est., $\hat{c}_e$). For L_2 regression, c_e is always equal to 1. Definitions of the error distributions are given in Section 4.2.

Table 4
Multiple type I errors for Arrangement 3

Estimate	Error distributions					
	Normal	Laplace	Cauchy	Tukey	Chisq	Chisq-med
Huber$_{k=0.05}$ theor.	2.6	5.6	[a]	3.8	1.4	2.6
Huber$_{k=0.05}$ est.	5.6	6.8	6.2	3.4	6.4	6.0
Huber$_{k=1.34}$ theor.	6.4	5.2	[a]	9.7	2.0	6.6
Huber$_{k=1.34}$ est.	6.0	5.1	4.6	4.3	4.0	5.0
Bisquare theor.	3.8	4.4	3.6	4.0	3.4	5.0
Bisquare est.	4.8	2.6	3.8	4.6	5.0	5.0
Hampel theor.	5.0	7.2	5.6	7.2	4.4	6.2
Hampel est.	6.4	5.7	4.8	4.8	5.2	6.2
L_2^{mBIC}	5.5	2.0	6.5	3.0	3.5	3.5
L_2^{BIC}	100	100	99.0	100	100	100

[a] Comparison of the probability of type I errors under the null model when the distribution of residuals is assumed to be known (theor., c_e) or has to be estimated (est., $\hat{c}_e$). For L_2 regression, c_e is always equal to 1. Definitions of the error distributions are given in Section 4.2.

The false discovery rates under the "3 QTL" model can be found in Figs. 1–3. They provide a similar picture, showing that our estimated normalizing constants lead to similar false discovery rates for several models for the error term. Indeed, the false discovery rates when applying the mBIC are approximately 15% for Arrangement 1 and 10% for Arrangements 2 and 3 (recall that a proportion of these "false" positives are due to the problem of localizing a QTL accurately). Least squares regression, in combination with standard BIC, results in false discovery rates above 20% for Arrangement 1, around 60% for Arrangement 2 and close to 100% for Arrangement 3, which again demonstrates the necessity of modifying the original BIC.

Figs. 1–3 also provide information regarding the power-robustness. They present the average detection power (averaged over additive and interaction effects) and the estimated false discovery rate for the analyzed procedures under different error distributions. In the case of normal errors, model selection based on least squares regression and M-estimation performs comparably for all three arrangements. The only robust regression method which is significantly worse than L_2 regression under normal errors is the one based on the Huber contrast with $k = 0.05$. This confirms the low efficiency of L_1 regression under normality. The Huber contrast function with $k = 0.05$ also performed significantly worse than the other robust methods in the case of the Tukey and χ^2 error distributions. Our simulations demonstrate that standard L_2 regression performs relatively well under the Laplace and χ^2 distributions, while it is inferior to some of the robust methods. The largest difference between L_2 regression and robust methods is observed for the heavy-tailed Cauchy distribution (for which L_2 regression fails completely) and the Tukey distribution, according to which an outlier occurs with some given probability. Tukey's bisquare estimate performed well in all the problems considered.

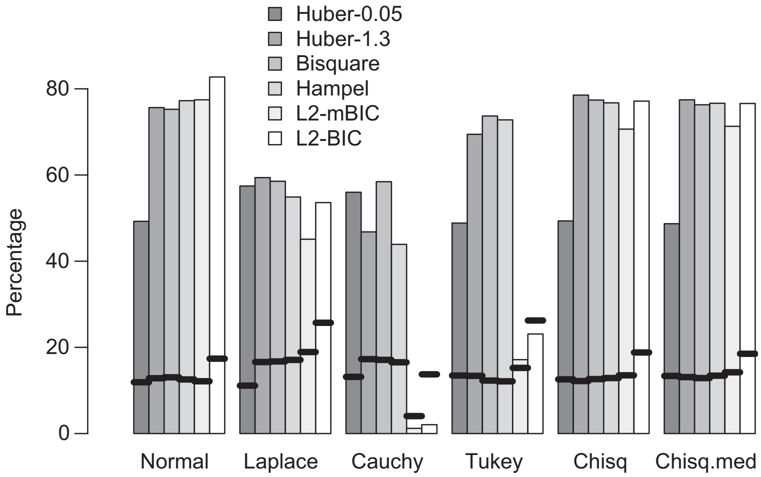

Fig. 1. Percentage of correctly identified main and epistatic effects (shaded bars) and false discovery rates (horizontal black lines) for Arrangement 1. Definitions of the error distributions are given in Section 4.2.

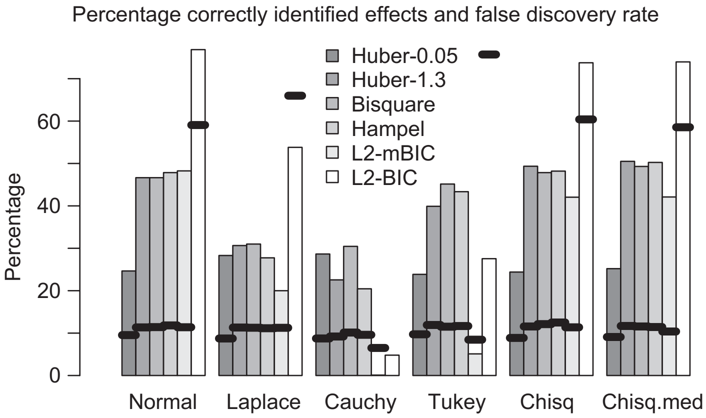

Fig. 2. Percentage of correctly identified main and epistatic effects (shaded bars) and false discovery rates (horizontal black lines) for Arrangement 2. Definitions of the error distributions are given in Section 4.2.

6. Application to real data

We apply our method to a data set obtained from QTL experiments on mice. Mähler et al. (2002) analyzed the susceptibility to colitis in strains that carry a deficient IL-10 gene, which is important in controlling the response of the immune system to intestinal antigenes. We consider their data obtained from a backcross to the less susceptible B6 strain and the quantitative traits MidPC1 and CecumPC1, which are the first two principal components of four scores measuring the severity and type of lesions in middle colon and cecum, respectively.

The data set contains 203 individuals and 12 markers from nine chromosomes, which were selected from a preliminary genome scan of 40 individuals and 67 markers spread over all 20 chromosomes.

Mähler et al. (2002) found one significant main effect for the MidPC1 trait on chromosome 12, which explains 6.7% of the variance and one possible main effect for CecumPC1 on chromosome 13, which explains 3.4% of the variance, but no epistatic effects. The distributions of the residuals under the selected model clearly deviated from normality by being bimodal in the case of both traits.

The modified BIC based on both least squares and M-estimation confirmed the main effect for ModPC1. In addition, both methods found an epistatic effect between the marker on chromosome 4 at 71 cM and the marker on chromosome 7

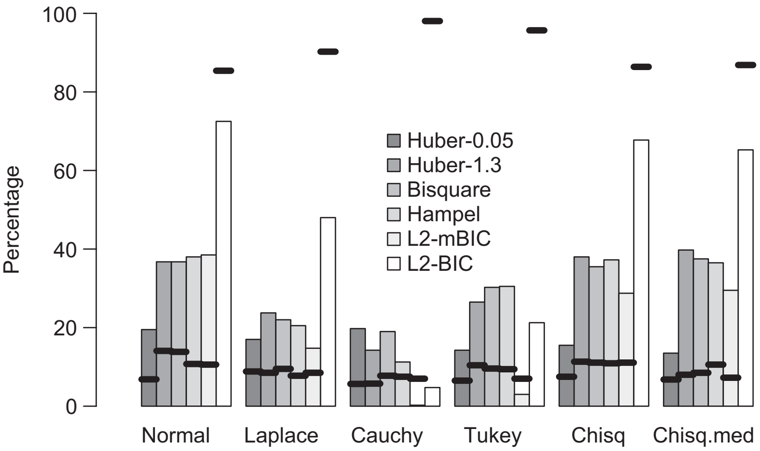

Fig. 3. Percentage of correctly identified main and epistatic effects (shaded bars) and false discovery rates (horizontal black lines) for Arrangement 3. Definitions of the error distributions are given in Section 4.2.

at 46 cM. In the case of CecumPC1, model selection using least squares regression found no QTL. Using M-estimation with a bi-square, Huber or Hampel contrast function, we detected two QTL, one on each of chromosome 5 and chromosome 13. The effect on chromosome 5 is slightly stronger and has a different sign to the effect on chromosome 13, as suggested by Mähler et al. (2002).

This example illustrates that our robust methods of model selection are capable of finding additional effects when the distribution of errors is not normal.

7. Conclusions

Overall, the performance of the M-estimators considered is superior to least squares regression in the context of QTL-mapping under various conditions. Considering the wide spectrum of possible error distributions, M-estimates based on the contrast functions considered also prove to be more flexible than L_1 regression. Among the robust methods considered, Tukey's bisquare estimate showed particularly good overall performance.

Acknowledgment

We wish to thank David Ramsey and two anonymous reviewers for comments and suggestions that led to a substantial improvement of the manuscript.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19, 716–723.

Baierl, A., Bogdan, M., Frommlet, F., Futschik, A., 2006. On locating multiple interacting quantitative trait loci in intercross designs. Genetics 173, 1693–1703.

Ball, R., 2001. Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. Genetics 159, 1351–1364.

Bassett, G., Koenker, R., 1978. Asymptotic theory of least absolute error regression. J. Amer. Statist. Assoc. 73, 618–622.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. B 57, 289–300.

Bogdan, M., Doerge, R.W., 2005. Biased estimators of genetic effects and heritability in interval mapping. Heredity 95, 476–484.

Bogdan, M., Ghosh, J.K., Doerge, R.W., 2004. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics 167, 989–999.

Broman, K.W., 1997. Identifying quantitative trait loci in experimental crosses. Ph.D. Dissertation, Department of Statistics, University of California, Berkeley, CA.

Broman, K.W., Speed, T.P., 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. J. Roy. Statist. Soc. B 64, 641–656.

Bussemaker, H.J., Li, H., Siggia, E.D., 2001. Regulatory element detection using correlation with expression. Nature Genetics 27, 167-171.

Carroll, R.J., 1980. Robust methods for factorial experiments with outliers. Appl. Statist. 29, 246–251.

Chardon, F., Virlon, B., Moreau, L., Falque, M., Joets, J., Decousset, L., Murigneux, A., Charcosset, A., 2004. Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. Genetics 168, 2169–2185.

Dick, D.M., Foroud, T., 2002. Genetic strategies to detect genes involved in alcoholism and alcohol-related traits. Alcohol Res. Health 26, 172–180.

Doerge, R.W., 2002. Mapping and analysis of quantitative trait loci in experimental populations. Natur. Rev. Genet. 3, 43–52.

Doerge, R.W., Zeng, Z.-B., Weir, B.S., 1997. Statistical issues in the search for genes affecting quantitative traits in experimental populations. Statist. Sci. 12, 195–219.

Fijneman, R.J.A., de Vries, S.S., Jansen, R.C., Demant, P., 1996. Complex interactions of new quantitative trait loci, *Sluc1*, *Sluc2*, *Sluc3*, and *Sluc4*, that influence the susceptibility to lung cancer in the mouse. Natur. Genet. 14, 465–467.

Fijneman, R.J.A., Jansen, R.C., Van der Valk, M.A., Demant, P., 1998. High frequency of interactions between lung cancer susceptibility genes in the mouse: mapping of Sluc5 to Sluc14. Cancer Res. 58, 4794–4798.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Jurečková, J., Sen, P.K., 1996. Robust Statistical Procedures: Asymptotics and Interrelations. Wiley, New York.

Keles, S. , van der Laan, M.J., Vulpe, C., 2004. Regulatory motif finding by logic regression. Bioinformatics 20, 2799-2811.

Khatkar, M.S., Thomson, P.C., Tammen, I., Raadsma, H.W., 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genet. Sel. Evol. 36, 163–190.

Lynch, M., Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Sinauer, Sunderland, MA.

Machado, J.A.F., 1993. Robust model selection and M-estimation. Econometric Theory 9, 478–493.

Mähler, M., Most, C., Schmidtke, S., Sundberg, J.P., Li, R., Hendrich, H.J., Churchill, G.A., 2002. Genetics of colitis susceptibility in IL-10-deficient mice: backcross versus F2 results contrasted by principal component analysis. Genomics 80, 274–282.

Martin, R.D., 1980. Robust estimation of autoregressive models. In: Brillinger, D.R., Tiao, G.C. (Eds.), Directions in Time Series. Institute of Mathematical Statistics, Hayward, CA, pp. 228–262.

Philips, T., 2002. Animal models for the genetic study of human alcohol phenotypes. Alcohol Res. Health 26, 202–207.

Piepho, H.-P., Gauch Jr., H.G., 2001. Marker pair selection for mapping quantitative trait loci. Genetics 157, 433–444.

Ronchetti, E., 1985. Robust model selection in regression. Statist. Probab. Lett. 3, 21–23.

Ronchetti, E., Field, Ch., Blanchard, W., 1997. Robust model selection by cross-validation. J. Amer. Statist. Assoc. 92, 1017–1023.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Serfling, R., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.

Siegmund, D., 2004. Model selection in irregular problems: applications to mapping quantitative trait loci. Biometrika 91, 785–800.

Thompson, E.A., 2000. Statistical inferences from genetic data on pedigress. In: NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 6. IMS, Beachwood, OH.

Walling, G.A., Visscher, P.M., Andersson, L., Rotschild, M.F., Wang, L., Moser, G., Groenen, M.A.M., Bidanel, J.-P., Cepica, S., Archibald, A.L., Geldermann, H., de Koning, D.J., Milan, D., Haley, C.S., 2000. Combining analyses of data from quantitative trait loci mapping studies: chromosome 4 effects of porcine growth and fatness. Genetics 155, 1369–1378.

Yohai, V.J., 1985. High breakdown-point and high efficiency robust estimates for regression. Ann. Statist. 15, 642–656.