Theor Appl Genet (2010) 121:363–372
DOI 10.1007/s00122-010-1315-8

ORIGINAL PAPER

# A two-phase procedure for QTL mapping with regression models

**Zehua Chen · Wenquan Cui**

Received: 24 February 2009 / Accepted: 27 February 2010 / Published online: 25 March 2010


**Abstract** It is typical in QTL mapping experiments that the number of markers under investigation is large. This poses a challenge to commonly used regression models since the number of feature variables is usually much larger than the sample size, especially, when epistasis effects are to be considered. The greedy nature of the conventional stepwise procedures is well known and is even more conspicuous in such cases. In this article, we propose a two-phase procedure based on penalized likelihood techniques and extended Bayes information criterion (EBIC) for QTL mapping. The procedure consists of a screening phase and a selection phase. In the screening phase, the main and interaction features are alternatively screened by a penalized likelihood mechanism. In the selection phase, a low-dimensional approach using EBIC is applied to the features retained in the screening phase to identify QTL. The two-phase procedure has the asymptotic property that its positive detection rate (PDR) and false discovery rate (FDR) converge to 1 and 0, respectively, as sample size goes to infinity. The two-phase procedure is compared with both traditional and recently developed approaches by simulation studies. A real data analysis is presented to demonstrate the application of the two-phase procedure.

**Keywords** Extended Bayes information criterion · False discovery rate · Positive detection rate · QTL mapping · Two-phase procedure

Communicated by D. Mather.

The research leading to this article is supported by the National University of Singapore research grant R-155-000-065-112.

Z. Chen (✉)
Department of Statistics and Applied Probability,
National University of Singapore, 3 Science Drive 2,
Singapore 117543, Singapore
e-mail: stachenz@nus.edu.sg

W. Cui
Department of Statistics and Finance,
University of Science and Technology of China, Hefei, China

## Introduction

The past few decades have witnessed the great development of statistical methodologies for quantitative trait loci (QTL) mapping. Starting with the naive single marker locus analysis (Soller et al. 1976) and the ordinary regression models (Cowen 1989, Moreno-Gonzalez 1992), many advanced methods have been developed. Those advanced methods can be roughly classified into two major categories: ones based on single-locus models and ones based on multiple-locus models. Methods based on single-locus models examine putative QTL one at a time with or without adjustment to other QTL or background effects, and do a one-dimensional search across the genome, see, e.g., Lander and Botstein (1989), Jansen (1993), Jansen and Stam (1994) and Zeng (1994). Methods based on multiple-locus models examine the effects of multiple putative QTL simultaneously and do a much more complicated high-dimensional search over the space of all possible multiple-locus models [see, e.g., Kao et al. (1999), Kao and Zeng (2002), Chen (2004), Chen and Liu (2009) and Li and Chen (2009)]. The appeal of the methods based on single-locus models is their relatively easy implementation. But it is at the price of an induced bias in the estimation of the number of QTL, the QTL effects, and QTL positions. On the contrast, methods based on multiple-locus models rectify this bias and are more efficient in terms of the power for

the detection of true QTL and the false discovery rate. Methods based on multiple-locus models are more promising when there exist multiple QTL with complicated epistasis effects.

The methods based on multiple-locus models amount to variable selection from the statistical point of view. There are a host of traditional variable selection approaches such as all-subsets method, forward, backward, or stepwise procedures, etc. [see Miller (2002)]. Some more advanced approaches based on various penalized likelihoods have been developed in recent years such as the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996), the smoothly clipped absolute deviation (SCAD) penalty considered by Fan and Li (2001), the elastic net advocated by Zou and Hastie (2005), etc. In genetic studies, only the traditional approaches have been commonly used so far, partly because there is a wide choice of computing softwares for the traditional approaches and partly because the newly developed approaches need time to get the familiarity of the genetic society.

However, the traditional approaches have serious drawbacks which have been argued by many authors, see, e.g., Breiman (1996). The all-subsets method and the backward procedure are infeasible with large or moderate number of variables. The forward or stepwise procedures have a well-known greedy nature; that is, the virtue of a variable is assessed only against the variables already included in the model, not considered in its synergetic role among all the variables. As a consequence, relatively unimportant variables might be selected but more important variables might be missed. This problem becomes even more prominent if the number of variables is large and the sample size is relatively small, which is typically the case in QTL mapping when epistasis effects are considered.

All the variable selection approaches, except the all-subsets method, share a common feature; that is, they assess only a few subsets generated by their respective mechanisms. Whether an approach is satisfactory depends on whether among the subsets it generates there is one which is the exact set of causal variables. By causal variables, we mean the ones that truly cause the variation of the response variable. The forward, backward, and stepwise procedures by no means guarantee the inclusion of the exact set among the subsets they assess. But, as to the penalized likelihood approaches, there is a so-called oracle property; that is, asymptotically, the exact set of causal variables is among the subsets generated by the penalized likelihood mechanisms, see, e.g., Fan and Li (2001) and Zhao and Yu (2006). Because of the oracle property, the penalized likelihood approaches have an edge over the traditional variable selection approaches.

In this article, we propose a two-phase procedure for QTL mapping based on the idea of penalized likelihoods. In the first phase of the procedure, the putative QTL effects are screened alternatively for main effects and epistasis effects by an iterative LASSO procedure. Through this phase, the number of the variables corresponding to main and epistasis effects of putative QTL is reduced to a tractable one. In the second phase, a sequential process using LASSO is applied to rank the variables retained from the first phase. Nested subsets of these variables are then formed according to their ranks, and a model selection criterion is used to assess these subsets for the final variable selection.

There are various model selection criteria such as the Akaike information criterion (AIC, Akaike 1973), the Bayes information criterion (BIC, Schwarz 1978), the cross-validation (CV), and generalized cross-validation (GCV) [see Stone (1974) and Craven and Wahba (1979)]. But these criteria are usually too liberal; that is, they tend to select many spurious covariates. In genetic studies, some ramifications of BIC have been used [see Broman and Speed (2002), Bogdan et al. (2004), Baierl et al. (2006), Zak et al. (2007) and Li and Chen (2009)]. Recently, an extended Bayes information criterion (EBIC) was developed, and its favorable asymptotic properties was shown by Chen and Chen (2008). In the two-phase procedure, the EBIC is used as the model selection criterion.

The proposed two-phase procedure can be applied in a variety of contexts such as regression models with quantitative trait as response and either marker genotypes as covariates (Broman and Speed 2002, Bogdan et al. 2004) or expected putative QTL genotypes conditioning on marker genotypes as covariates (Haley and Knott 1992), and rank regression models with the ranks of the trait values as responses (Zak et al. 2007). The theme of this article is to advocate the proposed two-phase procedure and demonstrate its advantage over the other variable selection approaches. To this end, we compare the two-phase procedure by simulation studies with a traditional forward procedure considered by Bogdan et al. (2004) and Zak et al. (2007), and with a more advanced approach developed by Fan and Lv (2007).

The remainder of the article is arranged as follows. The two-phase procedure is described in "Two-phase procedure" . The simulation studies are presented in "Simulation studies". An application of the two-phase procedure to a real data set is reported in "Analysis of DBH of Ridiata Pine data". The article is concluded with a discussion in "Discussion".

## Two-phase procedure

We now describe the two-phase procedure in the context of a generic regression model. Let $z_i$ be the response of the

$i$-th individual which is the individual's trait value or its rank among all the individuals in the sample. Let $x_{ij}$ be the value of the covariates $x_j$, $j = 1,\ldots,p$, of individual $i$. The covariates can be either marker genotypes or expected putative QTL genotypes. The model we consider is as follows:

$$z_i = \beta_0 + \sum_{j}^{p} \beta_j x_{ij} + \sum_{1 \le j < k \le p} \tau_{jk} x_{ij} x_{ik} + \epsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

In this model, both the main effects and interaction effects of the covariates are taken into account. For convenience, we refer the $x_j$'s as main features and $x_j x_k$'s as interaction features. In a typical QTL mapping problem, the number of covariates $p$ is usually bigger than the sample size $n$, and hence the total number of features $p(p + 1)$ is huge and much bigger than $n$. This situation is referred to as small-$n$-large-$p$ in the statistical literature. In addition, the number of true QTL in a species is usually not too big; that is, among the main and interaction features, there are only a small number of them which are truly associated with the response variable. This implies that in model (1) only a few coefficients are non-zero. This is referred to as sparsity. The nature of small-$n$-large-$p$ and sparsity in model (1) makes the two-phase procedure a natural way to proceed. We can roughly classify the features into three groups: the truly associated features (which will be referred to as causal features hereafter), the features of ambiguity, i.e, the ones which cannot be distinguished from the causal features easily, and the obviously un-associated features. The first phase of the two-phase procedure will screen out those obviously un-associated features and the second phase will distinguish the causal ones from the ambiguous ones. We refer to the first phase as the screening phase and the second one the selection phase.

## Screening phase

In the screening phase, main features and interaction features are screened alternatively by the LASSO mechanism which minimizes a penalized sum of squares with $L_1$ penalty $\lambda \sum_j |\theta_j|$ where $\theta_j$'s are either main feature coefficients or interaction feature coefficients. The LASSO has a special nature that the parameter $\lambda$ can be tuned such that the minimization of the penalized sum of squares produces a specified number of zero-estimated $\theta_j$'s. (This nature is shared by any penalty function which is singular at the origin). Let $N_{\mathrm{M}}$ and $N_{\mathrm{I}}$ be rough upper bounds for the numbers of causal main and interaction features, respectively (the bounds are taken much larger than the possible number of causal features). Let $\mathcal{V}_M$ denote the index set of the main features and $\mathcal{V}_I$ the index set of interaction features. Denote by $\mathcal{S}_M$ and $\mathcal{S}_I$ the index sets of tentatively selected features from $\mathcal{V}_M$ and $\mathcal{V}_I$, respectively. At the beginning, $\mathcal{S}_M$ and $\mathcal{S}_I$ are taken as empty sets. The screening phase iterates the following steps:

**Main feature screening:** Minimize with respect to $\{\beta_j : j \in \mathcal{V}_M\}$ and $\{\tau_{jk} : (j,k) \in \mathcal{S}_I\}$ the following penalized sum of squares:

$$\sum_{i=1}^{n} \left[ z_i - \sum_{j \in \mathcal{V}_M} \beta_j x_{ij} - \sum_{(j,k) \in \mathcal{S}_I} \tau_{jk} x_{ij} x_{ik} \right]^2 + \lambda_M \sum_{j \in \mathcal{V}_M} |\beta_j|$$

by tuning the parameter $\lambda_{\mathrm{M}}$ to a specific value such that in the minimizer only $N_{\mathrm{M}}$ of the $\beta_j$'s are non-zero, and all the other $\beta_j$'s are oppressed to zero. The $x_j$'s corresponding to the non-zero $\beta_j$'s are retained and $\mathcal{S}_m$ is updated as the index set of these retained $x_j$'s.

**Interaction feature screening:** Minimize with respect to $\{\tau_{jk} : (j,k) \in \mathcal{V}_{\mathrm{I}}\}$ and $\{\beta_j : j \in \mathcal{S}_{\mathrm{M}}\}$ the following penalized sum of squares:

$$\sum_{i=1}^{n} \left[ z_i - \sum_{j \in \mathcal{S}_{\mathrm{M}}} \beta_j x_{ij} - \sum_{(j,k) \in \mathcal{V}_{\mathrm{I}}} \tau_{jk} x_{ij} x_{ik} \right]^2 + \lambda_{\mathrm{I}} \sum_{(j,k) \in \mathcal{V}_{\mathrm{I}}} |\tau_{jk}|$$

by tuning the parameter $\lambda_I$ to a specific value such that in the minimizer only $N_I$ of the $\tau_{jk} : (i,k) \in \mathcal{V}_{\mathrm{I}}$ are non-zero. The set $\mathcal{S}_{\mathrm{I}}$ is updated as the set of those indices in $\mathcal{V}_I$ corresponding to nonzero $\tau_{jk}$'s.

Usually, after a few iterations, the selected set will become stable; that is, most of the selected terms will not change. In our simulation studies, the number of iterations never exceeds 10. Therefore, a fixed number of iterations can be set accordingly from the outset. The above procedure can be easily implemented by using the R function glmpath in the glmpath package developed by Park and Hastie (2007). The function glmpath can carry out the minimization when the number of covariates is much larger than the number of observations. It can handle any number of covariates subject to the capacity of the computing facility. If the number of intervals, $p$, is so large that the number of interaction features exceeds the capacity of the computing facility, the second screening step above can be modified as follows. Divide the interaction features into subgroups. The single step for the interaction feature screening is then replaced by several sub steps. In each sub step, only one subgroup is subjected to screening; that is, while all the features in this subgroup and those already selected features are included in the penalized sum of squares, only the coefficients corresponding to the current subgroup are subjected to the penalty.

Selection phase

In this phase, the features with indices in $\mathcal{S}_M$ and $\mathcal{S}_I$ are put together. For convenience, these features are denoted by $w_j$'s with a single index. Let $\theta_j$ be the coefficient associated with $w_j$. The penalized sum of squares

$$\sum_{i=1}^{n}\left[z_i-\sum_j \theta_j w_{ij}\right]^2+\lambda\sum_j|\theta_j|$$

is minimized at a sequence of decreasing $\lambda$ values: $\lambda_1, \lambda_2,\ldots$, where $\lambda_1$ is the value corresponding to the minimizer with the least number of non-zero components, and for $k>1$, $\lambda_k$ is the largest value which is smaller than $\lambda_{k-1}$ and corresponds to a minimizer with more non-zero components than the minimizer corresponding to $\lambda_{k-1}$. This process produces a sequence of feature sets. Let $\mathcal{C}_l$ denote the $l$th set in the sequence. Then the extended Bayes information criterion for $\mathcal{C}_l$ is given by

$$\mathrm{EBIC}(l)=n\log \mathrm{RSS}(l)+(m+q)\ln n + 2\gamma\ln\binom{p}{m}\binom{p(p-1)/2}{q},$$

where $\mathrm{RSS}(l)=(1/n)\sum_{i=1}^{n}\left[z_i-\sum_{j\in\mathcal{C}_l}\hat{\theta}_j w_{ij}\right]^2$, $\hat{\theta}_j$ 's are the least square estimates (without penalty), $m$ is the number of main features in $\mathcal{C}_l$, $q$ is the number of interaction features in $\mathcal{C}_l$ and $\gamma$ is a positive parameter whose value is to be determined by the user. We delay the consideration on the choice of $\gamma$ to "Simulation studies". For more details on EBIC, the reader is referred to Chen and Chen (2008). Eventually, the set with the smallest EBIC value is selected. Any putative QTL appearing in this set either as a main feature or an interaction feature is considered as a detected QTL.

The two-phase procedure described above has the following properties: if sample size is large enough, (1) the screening phase is able to retain almost all the causal features, and (2) the selection phase is able to select quite accurately the exact set of causal features. The first property is guaranteed by the sure screening property of the LASSO mechanism proved in Chen and Chen (2009). Let $\mathcal{C}_0$ be the set of all causal features. The sure screening property states that, if a set of features, say $\mathcal{C}$, is subjected to screening by the LASSO mechanism, and $\mathcal{C}^*$ is the set of selected features with size bigger than $\mathcal{C}_0$, then the probability that $\mathcal{C}^*$ contains $\mathcal{C}_0\cap\mathcal{C}$ converges to 1 as sample size goes to infinity. The second property follows from three results: the sure screening property of the screening phase given above, the oracle property of LASSO, and the selection-consistency of the EBIC proved by Chen and Chen (2008). The oracle property states that, when the sample size is large enough, there is a $\lambda$ value such that the LASSO produces the exact set $\mathcal{C}_0$. The selection-consistency states that if the number of features is of order $O(n^\kappa)$ and $\gamma>1-1/(2\kappa)$, then the probability that the EBIC evaluated at $\mathcal{C}_0$ is smaller than the EBIC evaluated at any other model which can be formed from the features under consideration converges to 1 as sample size goes to infinity. Thus, in an asymptotic sense, the sure screening property guarantees that $\mathcal{C}_0$ is contained in the set of features retained in the selection phase, the oracle property guarantees that there is a $\lambda_l$ which will produce $\mathcal{C}_0$ in the ranking process, and finally the selection-consistency of EBIC guarantees that $\mathcal{C}_0$ will be eventually selected.

In other feature selection strategies such as multiple testing, one tries to control the overall type-I error by a tight threshold value for the test statistics. However, the concept of overall type-I error is not much relevant in feature selection. A more appropriate measure, the false discovery rate (FDR), was proposed by Benjamini and Hochberg (1995) and advocated by many authors [see, e.g., Efron et al. (2001), Efron and Tibshirani (2002) and Storey (2002)]. We adapt the concept of FDR in the context of QTL mapping and define the FDR as the ratio of the number of loci which have been falsely claimed as QTL and the total number of loci claimed as QTL. Besides FDR, another concern in QTL mapping is how many true QTL have been claimed as QTL. Therefore, we define the positive detection rate (PDR) as the ratio of the number of true QTL which has been detected and the total number of true QTL. In terms of FDR and PDR, the argument of the previous paragraph is that, when the number of variables under study is of order $O(n^\kappa)$ and $n$ goes to infinity, if $\gamma$ in EBIC is chosen greater than $1-1/(2\kappa)$, then the FDR and PDR of the two-phase procedure converge to 0 and 1, respectively.

## Simulation studies

We carried out two simulation studies. In the first study, we compared the two-phase procedure with a traditional forward procedure considered by Bogdan et al. (2004) and Zak et al. (2007) under the context that the ranks of the trait values are taken as the response values and the expected genotypes of putative QTL are taken as covariates. In the second study, we compared the two-phase procedure with an approach using sure independent screening followed by ordinary LASSO (SIS-LASSO) considered by Fan and Lv (2007) under the context that the trait values themselves are taken as response values and marker genotypes are taken as covariates.

Simulation study I

In this study, $z_i$ is the rank of the trait value of individual $i$ within the sample and $x_{ij}$ is the expected putative QTL

genotype on interval $j$ of individual $i$. The $x_{ij}$'s are derived by using the flanking marker genotypes of interval $j$. For convenience, we only consider a backcross design for the study. We code the combination of the two flanking maker genotypes as follows. The code is 1, if both markers are homozygous; 2, if the left marker is homozygous and the right one is heterozygous; 3, if the left one is heterozygous and the right one is homozygous; 4, if both markers are heterozygous. Denote the code for individual $i$ on interval $j$ by $c_{ij}$. Let $\gamma_j$ be the recombination fraction between the two flanking markers of interval $j$, $r_j$ the recombination fraction between the left marker and the putative QTL on interval $j$. The $x_{ij}$ is determined as follows:

| $c_{ij}$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $x_{ij}$ | $\frac{(1-r_j)(1-s_j)}{1-\gamma_j}$ | $\frac{(1-r_j)s_j}{\gamma_j}$ | $\frac{(1-s_j)r_j}{\gamma_j}$ | $\frac{r_j s_j}{1-\gamma_j}$ |

where $s_j = (\gamma_j - r_j)/(1 - 2r_j)$. The $x_{ij}$ is in fact the conditional probability that the putative QTL on interval $j$ is homozygous for individual $i$ given its marker genotypes. For each interval, we let the middle point be the surrogate of the putative QTL on that interval. This is equivalent to taking $r_j = s_j$, i.e., $r_j = (1 - \sqrt{1 - 2\gamma_j})/2 \approx \gamma_j/2$, for small $\gamma_j$.

In the forward procedure considered by Zak et al. (2007), a different model selection criterion, rBIC, is used for assessing models. The rBIC is the mBIC developed by Bogdan et al. (2004) applied to the rank regression model. In fact, the rBIC and the EBIC with $\gamma = 1$ are asymptotically equivalent (see "Discussion"). In their forward procedure, at each consecutive step, they consider all the main and interaction features and choose the one whose presence in the model yields the lowest value of rBIC. The procedure is stopped after 30 steps and the resulting 31 models are evaluated on the basis of minimizing the rBIC. The number of steps is set at 30 because the rBIC attains its minimum never beyond 20 steps. To make the comparison fair, we use the same criterion in both procedures in the simulation study. Both the rBIC and the EBIC with $\gamma = 1$ are used as the model selection criterion. Three settings are considered. The details of the settings and the results are given in the following.

The first setting is exactly the same as Setup 1 of Zak et al. (2007). In this setting, two chromosomes each of length 100 cM with markers equally spaced every 10 cM are considered. Three QTL are fixed throughout the simulation at 20 cM on chromosome 1, 20, and 70 cM on chromosome 2. The quantitative trait is generated according to the model

$$Y_i = 0.55\delta_{i1} + 1.2\delta_{i2}\delta_{i3} + \epsilon_i,$$

where $\delta_{ij}$, $j = 1, 2, 3$, are the genotype indicators of the $i$th individual at the three QTL, $\varepsilon_i$ are i.i.d. errors with mean zero. The sample size is taken as $n = 200$ and 3,000 replicates are simulated.

To make the simulation setting more realistic, in the second and third settings, we mimicked the structure of the *Radiata Pine* genome studied by Kao and Zeng (1999). The *Radiata Pine* genome consists of 12 chromosomes. Except one short chromosome with length 30 cM, the lengths of the other chromosomes range from 70 to 270 cM. In the two settings, we generated the lengths of the chromosomes as random numbers uniformly distributed between 80 and 280 cM. Then, on each chromosome, the marker positions are generated such that the lengths of intervals are uniformly distributed between 5 and 20 cM. This results in a pseudo-genome with total length 2,063.5 cM and 164 intervals for the second setting and a pseudo-genome with total length 1,947.6 cM and 161 intervals for the third setting. In these two settings, we considered 12 QTL with effect structure given below:

$$Y_i = \sum_{j=1}^{8} \beta_j \delta_{ij} + \tau_{59}\delta_{i5}\delta_{i9} + \tau_{6,10}\delta_{i6}\delta_{i10} + \tau_{7,11}\delta_{i7}\delta_{i11} + \tau_{8,12}\delta_{i8}\delta_{i12} + \epsilon_i.$$

In the second setting, the parameters take the same value 2. In the third setting, the parameters take the following values:

$$\begin{aligned}&(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \tau_{59}, \tau_{6,10}, \tau_{7,11}, \tau_{8,12})\\ &\quad = (3.98, 3.49, 3.35, 3.24, 2.85, 1.43,\\ &\qquad -1.27, -1.24, 1.07, 2.32, 2.54, 2.48)\end{aligned}$$

Unlike in the first setting where the QTL positions are fixed for all simulation replicates, in the second and third setting, the positions of the 12 QTL are randomly chosen in each replicate. The sample size is the same as in the first setting. But because of the huge amount of computation involved in the forward procedure, only 200 replicates are simulated in the second and third setting.

The same distributions considered by Zak et al. (2007) are assumed for the errors in the above settings. For the reader's convenience, we quote from Zak et al. (2007) the description of the distributions as follows:

1. Normal: 1.11 × N(0, 1).
2. Laplace: 1.08 × Laplace(0, 1).
3. Cauchy: Cauchy(0, 0.75).
4. Tukey's gross error model: 1.081 × Tukey(0.95, 100, 1).
5. $\chi^2$ with 6 d.f. centered around the mean: $0.342 \times (\chi^2_6 - 6)$.

In Tukey's gross error model, the error distribution is a mixture of two normal distributions leading to a certain

percentage of outliers. More specifically, Tukey($\alpha$, $\tau$, $\sigma$) = $\lambda \times N(0, \sigma^2) + (1 - \lambda) \times N(0, \tau \times \sigma^2)$, where $\lambda \sim$ Binomial (1, $\alpha$).

For the choice of $N_M$ and $N_I$, we found that whenever $N_M$ is not less than three times the number of true causal main features and $N_I$ is not less than five times the number of true causal interaction features, the final selection result is not much affected. In the reported simulation results, ($N_M$, $N_I$) is taken as (30, 20) throughout.

We counted positive detections and false discoveries in the following way. A locus is claimed a discovered QTL if it is involved either in a main effect term or an interaction term in the selected model. A QTL is considered detected if there is at least one claimed locus which is less than 15 cM from the QTL. A claimed locus is considered as a false discovery if it is beyond 15 cM from any QTL. The average PDR and FDR of the two-phase procedure and the forward procedure with both EBIC and rBIC over the simulation replicates in the above three settings are presented in Tables 1, 2 and 3, respectively. The numbers within the parentheses in the tables are standard deviations of average PDRs and FDRs.

The EBIC and rBIC produce similar results in all the three settings as to the comparison between the two-phase procedure and the forward procedure. The performance of EBIC is comparable vis-à-vis rBIC. We now discuss the comparison with EBIC as the model selection criterion. From Tables 1—3, we can see clearly the following. In the first setting, the forward procedure has a slightly higher PDR than the two-phase procedure, but the inroad is made with a much higher FDR. In fact, averaging over the five error distributions, the relative increase in PDR of the forward procedure is just 3.1%, but the relative increase in FDR of the forward procedure is 52%. In the second setting, the two-phase procedure dominates the forward procedure in the sense that it has both higher PDR and lower FDR for all the error distributions. The average relative decrease in PDR of the forward procedure is 3.8% and the average relative increase in FDR of the forward procedure is 20.8%. In the third setting, both procedures have a comparable PDR, but the forward procedure has a uniformly higher FDR. The average relative increase in FDR of the stepwise procedure is 20.7%.

## Simulation study II

In the second simulation study, we compare the two-phase procedure with the SIS-LASSO procedure. In a nutshell, the SIS-LASSO procedure goes as follows. First, the response is fitted to a single feature model for each of all the features, the $p$-value for testing the feature effect is obtained, the $p$-values are ordered in ascending order, and finally a pre-determined number of features with $p$-values in the lower end are retained. Second, for the retained features, an ordinary LASSO procedure is applied with cross-validation. Since interaction effects have not been considered using the SIS-LASSO approach elsewhere, we take a similar strategy in the SIS step to that used in the two-phase procedure; that is, we order the $p$-values of the main features and those of the interaction features separately and retain a pre-determined number of main features and a pre-determined number of interaction features. For the two-phase procedure, two values of $\gamma$ in EBIC are considered: 0 and 1. The EBIC with $\gamma = 0$ is indeed the original BIC.

The second simulation study is based on the *Radiata Pine* data. A more detailed description of the data is quoted from Kao and Zeng (1999): *Radiata pine is one of the most widely planted forestry species in the Southern Hemisphere. Two elite parents were crossed to produce 134 progeny. For each progeny, random amplified polymorphic DNA (RAPD) markers were generated, and traits measured included annual brown cone number at eight years of age, diameter of stem at breast height, and branch quality score…. The RAPD marker data contained 120 markers in 12 linkage groups and covered* $\sim$1679.3 cM. This data set

**Table 1** Average PDR and FDR of the two-phase (TP) procedure and the forward (FW) procedure using both EBIC and rBIC as criterion with the first simulation setting (numbers in parentheses are standard deviations)

| Error | PDR | | | | FDR | | | |
|---|---|---|---|---|---|---|---|---|
| | TP | | FW | | TP | | FW | |
| | EBIC | rBIC | EBIC | rBIC | EBIC | rBIC | EBIC | rBIC |
| Normal | 0.607 (0.0049) | 0.639 (0.0048) | 0.640 (0.0053) | 0.665 (0.0051) | 0.090 (0.0038) | 0.091 (0.0037) | 0.127 (0.0045) | 0.130 (0.0044) |
| Laplace | 0.466 (0.0047) | 0.498 (0.0046) | 0.482 (0.0051) | 0.518 (0.0050) | 0.102 (0.0046) | 0.110 (0.0046) | 0.144 (0.0053) | 0.149 (0.0052) |
| Cauchy | 0.365 (0.0045) | 0.406 (0.0045) | 0.356 (0.0049) | 0.396 (0.0048) | 0.090 (0.0045) | 0.101 (0.0046) | 0.153 (0.0058) | 0.167 (0.0058) |
| Tukey | 0.509 (0.0047) | 0.540 (0.0047) | 0.526 (0.0051) | 0.558 (0.0051) | 0.093 (0.0042) | 0.098 (0.0042) | 0.144 (0.0051) | 0.148 (0.0050) |
| $\chi^2_6$ | 0.636 (0.0049) | 0.664 (0.0048) | 0.661 (0.0051) | 0.686 (0.0050) | 0.079 (0.0035) | 0.083 (0.0035) | 0.124 (0.0043) | 0.129 (0.0043) |
| Average | 0.517 | 0.549 | 0.533 | 0.565 | 0.091 | 0.097 | 0.138 | 0.145 |

**Table 2** Average PDR and FDR of the two-phase (TP) procedure and the forward (FW) procedure using both EBIC and rBIC as criterion with the second simulation setting (numbers in parentheses are standard deviations)

| Error | PDR | | | | FDR | | | |
|---|---|---|---|---|---|---|---|---|
| | TP | | FW | | TP | | FW | |
| | EBIC | rBIC | EBIC | rBIC | EBIC | rBIC | EBIC | rBIC |
| Normal | 0.687 (0.0087) | 0.675 (0.0087) | 0.670 (0.0092) | 0.660 (0.0092) | 0.081 (0.0069) | 0.076 (0.0065) | 0.112 (0.0091) | 0.099 (0.0082) |
| Laplace | 0.628 0.0093) | 0.603 (0.0088) | 0.614 (0.0092) | 0.606 (0.0090) | 0.090 (0.0072) | 0.080 (0.0071) | 0.103 (0.0083) | 0.098 (0.0079) |
| Cauchy | 0.393 (0.0089) | 0.384 (0.0087) | 0.356 (0.0107) | 0.353 (0.0106) | 0.112 (0.0095) | 0.105 (0.0095) | 0.146 (0.0133) | 0.142 (0.0134) |
| Tukey | 0.499 (0.0102) | 0.485 (0.0102) | 0.465 (0.0105) | 0.461 (0.0104) | 0.123 (0.0092) | 0.115 (0.0092) | 0.139 (0.0114) | 0.141 (0.0115) |
| $\chi^2_6$ | 0.668 (0.0085) | 0.648 (0.0084) | 0.661 (0.0091) | 0.650 (0.0095) | 0.099 (0.0075) | 0.088 (0.0068) | 0.110 (0.0089) | 0.104 (0.0085) |
| Average | 0.575 | 0.559 | 0.553 | 0.546 | 0.101 | 0.092 | 0.122 | 0.117 |

**Table 3** Average PDR and FDR of the two-phase (TP) procedure and the forward (FW) procedure using both EBIC and rBIC as criterion with the third simulation setting (numbers in parentheses are standard deviations)

| Error | PDR | | | | FDR | | | |
|---|---|---|---|---|---|---|---|---|
| | TP | | FW | | TP | | FW | |
| | EBIC | rBIC | EBIC | rBIC | EBIC | rBIC | EBIC | rBIC |
| Normal | 0.591 (0.0082) | 0.587 (0.0079) | 0.596 (0.0080) | 0.593 (0.0077) | 0.058 (0.0061) | 0.054 (0.0059) | 0.063 (0.0069) | 0.061 (0.0065) |
| Laplace | 0.680 (0.0073) | 0.671 (0.0072) | 0.687 (0.0071) | 0.682 (0.0073) | 0.067 (0.0065) | 0.062 (0.0060) | 0.088 (0.0078) | 0.082 (0.0074) |
| Cauchy | 0.476 (0.0098) | 0.458 (0.0097) | 0.459 (0.0102) | 0.452 (0.0102) | 0.122 (0.0103) | 0.112 (0.0103) | 0.145 (0.0108) | 0.143 (0.0105) |
| Tukey | 0.555 (0.0096) | 0.536 (0.0096) | 0.538 (0.0093) | 0.527 (0.0095) | 0.111 (0.0090) | 0.093 (0.0080) | 0.120 (0.0092) | 0.111 (0.0086) |
| $\chi^2_6$ | 0.718 (0.0075) | 0.703 (0.0071) | 0.718 (0.0067) | 0.712 (0.0065) | 0.058 (0.0056) | 0.046 (0.0049) | 0.086 (0.0073) | 0.078 (0.0065) |
| Average | 0.604 | 0.591 | 0.600 | 0.593 | 0.083 | 0.073 | 0.100 | 0.095 |

has been analyzed using mixture normal models by Kao and Zeng (1999). In our simulation, the data set is used in the following way. The genotypes at the 120 markers of the 134 progeny are used in all replicates of the simulation. Let $p_M$ and $p_I$ denote the numbers of causal main and interaction features, respectively. At each replicate, $p_M$ markers are randomly selected from the 120 markers as main causal features, $p_I$ additional markers are randomly selected from the remaining ones and paired with $p_I$ markers chosen at random from the $p_M$ main causal features to form $p_I$ interaction features. The quantitative trait values are generated by a linear model with normal errors. The coefficients of the features in the linear model are generated in each replicate as follows: Let $a = 5\log(n)/\sqrt{(n)}$, Each coefficient is then independently generated as $(-1)^{I\{u > 0.4\}}$ $(a + |z|)$, where $u$ is a uniform random variable on [0, 1] and $z$ is a standard normal variable.

The numbers (4, 1) and (5, 3) are considered for $(p_M, p_I)$. Different standard deviations are considered for the error terms such that they correspond roughly to specified levels of heritability 0.6, 0.8, and 0.9. For the two-phase procedure, the $N_M$ and $N_I$ in the screening phase are taken as 25 and 15, respectively. For the SIS-LASSO procedure, the numbers of main and interaction features retained in the SIS step are taken as 30 and 20, respectively. The mean and standard deviation of FDR and PDR of the two procedures over 2000 replicates are presented in Tables 4 and 5. As demonstrated in Table 4, in the case that there are four causal main features and one causal interaction feature, the two-phase procedure with both $\gamma = 0$ and 1 dominates the

**Table 4** Average PDR and FDR of the two-phase procedure with $\gamma$ = 0 (TP0) and $\gamma = 1$ (TP1) and the SIS-LASSO (SL) procedure when the numbers of causal main interaction features are 4 and 1, respectively (numbers in parentheses are standard deviations)

| $\sigma$ | PDR | | | FDR | | |
|---|---|---|---|---|---|---|
| | TP0 | TP1 | SL | TP0 | TP1 | SL |
| 0.44 | 0.992 (0.041) | 0.964 (0.109) | 0.942 (0.101) | 0.496 (0.180) | 0.370 (0.197) | 0.721 (0.069) |
| 0.65 | 0.993 (0.039) | 0.959 (0.115) | 0.941 (0.101) | 0.507 (0.175) | 0.370 (0.198) | 0.720 (0.068) |
| 1.00 | 0.991 (0.044) | 0.938 (0.141) | 0.939 (0.103) | 0.523 (0.168) | 0.369 (0.199) | 0.722 (0.066) |

SIS-LASSO procedure in the sense that it has higher PDR and lower FDR compared with the SIS-LASSO procedure. In particular, its FDR is significantly much lower than that of the SIS-LASSO procedure. In the case that there are five causal main features and three causal interaction features, the two-phase procedure with $\gamma = 0$ still dominates the SIS-LASSO procedure, and the two-phase procedure with $\gamma = 1$ has much smaller FDR while the PDR is comparable, though slightly lower, compared with the SIS-LASSO procedure, as shown in Table 5. The results in Tables 4 and 5 again shed light on the general nature of the two-phase procedure: it is able to control FDR at lower level without compromising PDR too much compared with other approaches.

## Analysis of DBH of Ridiata Pine data

In this section, we demonstrate how the two-phase procedure is applied in real problems by analyzing the Ridiata Pine data for the mapping of the trait: diameter of stem at breast height (DBH).

The mapping of this trait has been considered by Kao and Zeng (1999) in the context of multiple interval mapping in an ad hoc way as follows. A stepwise procedure was used to select intervals. It was first found that no interval, when considered one at a time, is significant enough (the likelihood ratio test statistic exceeding the Bonferroni-adjusted overall critical value with level 0.05) to enter the model. Then a "chunkwise" procedure was adopted. In the chunkwise procedure, at a forward step, intervals are first considered one at a time. If the most significant new interval qualifies to enter the model, the procedure switches to the backward step; otherwise, intervals are considered a pair at a time. If none of the pairs is significant enough, the whole procedure stops; if there are significant pairs, the most significant one enters the model and the procedure switches to the backward step. At a backward step, each of the intervals included in the model is tested; if any of them are non-significant, the non-significant ones are deleted and the procedure switches to the forward step. By this way, three pairs of intervals were selected. By using the notation [$c$: $m_1$, $m_2$] to denote the interval on chromosome $c$ flanked by the $m_1$-th and $m_2$-th marker on this chromosome, the pairs of intervals are ([1: 4, 5], [1: 5, 6]), ([5: 6, 7], [10: 6, 7]), and ([2: 3, 4], [12: 6, 7]).

In the application of the two-phase procedure to the mapping of DBH, we used the $\gamma$ values in the EBIC ranging from 0 to 1 with points equally spaced 0.01 apart. With the $\gamma$ values ranging from 0 to 0.04, the same model with the following seven interaction features is selected: (1:4, 3:5), (3:7, 12:3) , (3:7, 12:5), (5:5, 10:5), (5:5, 10:6), (6:5, 10:6), and (6:6, 10:7), where ($c_1$: $m_1$, $c_2$:$m_2$) denotes the interaction term formed by the $m_1$-th marker on chromosome $c_1$ and the $m_2$-th marker on chromosome $c_2$. There are 11 markers involved in this model. With the $\gamma$ values ranging from 0.05 to 0.56, the same model with the following two interaction features is selected: (5:5, 10:6) and (6:5, 10:6). There are three markers involved in this model. With $\gamma$ values bigger than 0.56, no feature at all is selected. A bootstrap-like procedure, which will be briefly described in the next section, was used to estimate the FDR of these two models. The estimated FDR for the first model is 0.82 and for the second model is 0.32. In other words, the number of correctly identified markers can be estimated about three in the first model and about two in the second model. Since the markers in the second model are also contained in the first model, taking into account the two estimated FDR, we can claim with high confidence that the three markers in the second model are in linkage disequilibrium (LD) with true QTL. It is interesting to note that all the selected markers are in six blocks: (1) 1:4; (2) 3:5, 3:7; (3) 5:5; (4) 6:5, 6:6; (5) 10:5, 10:6, 10:7; (6) 12:4, 12:5. The lengths of blocks (2), (4), (5), and (6) are, respectively, 51.8, 16.7, 21, and 16.7 cM. The three markers in the second model fall into blocks (3), (4), and (5). We can consider that these three blocks are identified as being in LD with true QTL.

Comparing with the mapping results of Kao and Zeng (1999), in our second selected model, block (3) is adjacent to interval [5:6,7] detected by Kao and Zeng (1999), block (5) covers interval [10:6,7] detected by Kao and Zeng (1999). In our first selected model, in addition to the blocks just mentioned, block (1) and interval [1:4,5] are identical and block (6) is adjacent to interval [12:6,7]. If we take the immediate adjacent regions of the identified markers or

**Table 5** Average PDR and FDR of the two-phase procedure with $\gamma = 0$ (TP0) and $\gamma = 1$ (TP1) and the SIS-LASSO (SL) procedure when the numbers of causal main interaction features are five and three, respectively (numbers in parentheses are standard deviations)

| $\sigma$ | PDR | | | FDR | | |
|---|---|---|---|---|---|---|
| | TP0 | TP1 | SL | TP0 | TP1 | SL |
| 0.55 | 0.910 (0.112) | 0.716 (0.280) | 0.794 (0.123) | 0.485 (0.114) | 0.322 (0.182) | 0.670 (0.070) |
| 0.80 | 0.905 (0.112) | 0.703 (0.284) | 0.792 (0.123) | 0.490 (0.111) | 0.320 (0.182) | 0.671 (0.072) |
| 1.35 | 0.891 (0.120) | 0.637 (0.289) | 0.785 (0.123) | 0.508 (0.111) | 0.291 (0.190) | 0.670 (0.073) |

intervals as identified regions in LD with QTL, we can see that blocks (3) and (5) are identified both by our second model (with high confidence) and by the ad hoc approach of Kao and Zeng (1999). In addition to these blocks, blocks (1) and (6) are identified both by our first model and by the ad hoc approach of Kao and Zeng (1999). Blocks (2) and (4) are not identified by Kao and Zeng (1999), and the interval [2:3,4] is not identified by our two-phase procedure.

The advantage of the analysis by the two-phase procedure is that (a) it provides estimated FDR for the selected sets of markers and (b) it provides a rank of importance for the selected markers, for example, the makers selected by the second model are more important than other markers selected by the first model in terms of their effects. These two together give some ground for inference, which other approaches do not offer.

## Discussion

We end this paper by some discussion regarding the application of EBIC. First we show that the EBIC with $\gamma = 1$ is asymptotically equivalent to rBIC (or mBIC) proposed by Bogdan et al. (2004) and Zak et al. (2007). Note that the EBIC and rBIC differ by the term $2\ln\binom{p}{m}\binom{p(p-1)/2}{q}$ in EBIC and the term $2m\ln(p/2.2-1)+2q\ln[p(p-1)/(2\times 2.2)-1]$ in rBIC. Here, for consistency, we use our notation in the term of rBIC. For large $p$, $\binom{p}{m}\sim p^m$, $\binom{p(p-1)/2}{q}\sim[p(p-1)/2]^q$, $\ln(p/2.2-1)\sim\ln(p)$ and $\ln[p(p-1)/(2\times 2.2)-1]\sim\ln[p(p-1)/2])$, where $a\sim b$ meaning $a/b\to 1$. Thus, $2\ln\binom{p}{m}\binom{p(p-1)/2}{q}\sim 2m\ln(p)+2q\ln[p(p-1)/2]$ and $2m\ln(p/2.2-1)+2q\ln[p(p-1)/(2\times 2.2)-1]\sim 2m\ln(p)+2q\ln[p(p-1)/2]$. The equivalence between the EBIC with $\gamma = 1$ and the rBIC then follows. In our simulation studies and applications in other problems (not reported here) [see, Chen and Chen (2008, 2009), and Zhao (2008)], we found that in all the cases where the number of covariates is bigger than the sample size $n$, the choice of $\gamma = 1$ yields desirable low FDR and reasonable PDR. Compared with other approaches, its FDR is always lower and its PDR does not differ too much from the other approaches.

In connection with the use of EBIC for controlling FDR or making inference, the two-phase procedure can be applied more sophisticatedly as follows. At first, the procedure is applied with the $\gamma$ values in EBIC equally spaced in an interval [0, G]. Usually, only a few different models will be selected, each model corresponding to a sub-interval of the $\gamma$ values. For each model, the FDR of the procedure with $\gamma$ taking the upper bound of the corresponding sub-interval is estimated by a bootstrap-like procedure. Then the inference can be made based on the estimated FDR and the selected features, as illustrated in "Analysis of DBH of Ridiata Pine data". The bootstrap-like procedure is briefly described in the following. For a given $\gamma$ value, at the first step, the two-phase procedure is applied to the original data with this $\gamma$ value. The estimates of the effects of the selected features and their variances are obtained. At the second step, the same number of features with the same inter-relations as those which have been selected in the first step are randomly sampled from the original marker data. The corresponding effects are generated as normally distributed with the estimated effects as means and their estimated variances as variances and the trait values are generated using these pseudo-features and effects. The two-phase procedure using the same $\gamma$ value is then applied for mapping the generated trait values with the original marker data and the FDR for the mapping is derived. This step is then repeated for a large number of times; in the real data analysis presented in "Analysis of DBH of Ridiata Pine data", the step is repeated 1,000 times. The average FDR over the replicates is taken as the estimate of the FDR for the given $\gamma$ value. The details and the properties of the bootstrap-like procedure are beyond the scope of this paper. They will be reported elsewhere.

**Acknowledgments** The authors would like to express their appreciation to the editor and the anonymous referees for their valuable comments and suggestions which have led to a great deal of improvement on the paper.

## References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrox BN, Caski F (eds) Second Int Symp Info Theory. Akademiai Kiado, Budapest, pp 267–281

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

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300

Bogdan M, Ghosh JK, Doerge RW (2004) Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics 167:989–999

Breiman L (1996) Bagging predictors. Mach Learn 26:123–140

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J Roy Stat Soc Ser B 64:641–656

Chen Z (2004) The full EM algorithm for the MLEs of QTL effects and positions and their estimated variances in multiple interval mapping. Biometrics 61:474–480

Chen J, Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95:759–771

Chen Z, Chen J (2009) Tournament screening cum EBIC for feature selection with high dimensional feature spaces. Sci China Ser A Math Phys Astron 52:1327–1341

Chen Z, Liu J (2009) Mixture generalized linear models for multiple interval mapping of quantitative trait loci in experimental crosses. Biometrics 65:470–477

Cowen NM (1989) Multiple linear regression analysis of RELP data sets used in mapping QTLs. In: Helentjaris T, Burr B (eds) Development and application of molecular markers to problems in plant genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, pp 113–116

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Appl Numer Math 31:377–403

Efron B, Tibshirani R (2002) Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol 23:70–86

Efron B, Tibshirani R, Storey JD and Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160

Fan J, Li R (2001) Variable selection via non-concave penalized likelihood and its oracle properties. J Am Stat Assoc 96: 1348–1360

Fan J, Lv J (2007) Sure independence screening for ultra-high dimensional feature space. Ann Stat 70:849–911

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line cross using flanking markers. Heredity 69:315–324

Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics 135:205–211

Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455

Li W, Chen Z (2009) Multiple interval mapping for quantitative trait loci with a spike in the trait distribution. Genetics 182:337–342

Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152:1203–1216

Kao CH, Zeng ZB (2002) Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics 160:1243–1261

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199

Miller A (2002) Subset selection in regression. Chapman & Hall/CRC, Boca Raton

Moreno-Gonzalez J (1992) Genetic models to estimate additive and non-additive effects of marker-associated QTL using multiple regression techniques. Theor Appl Genet 85:435–444

Park MY, Hastie T (2007) An $L_1$ regularization path algorithm for generalized linear models. J Roy Stat Soc B Ser 69:659–677

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Soller M, Brody T, Genizi A (1976) On the power of experimental design for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor Appl Genet 47:35–39

Stone M (1974) Cross-validatory choice and assessment of statistical predictions (with Discussion). J Roy Stat Soc B Ser 39:111–147

Storey JD (2002) A direct approach to false discovery rates. J Roy Stat Soc B Ser 64:479–498

Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J Roy Stat Soc B Ser 58:267–288

Zak M, Baierl A, Bogdan M, Futschik A (2007) Locating multiple interacting quantitative trait loci using rank-based model selection. Genetics 176:1845–1854

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468

Zhao J (2008) Model selection methods and their applications in genome-wide association studies. Dissertation, National University of Singapore

Zhao P, Yu B (2006) On model selection consistency of LASSO. J Mach Learn Res 7:2541–2567

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Roy Stat Soc B Ser 67:301–320