Published in final edited form as:

Comput Stat Data Anal. 2009 March 15; 53(5): 1850–1860. doi:10.1016/j.csda.2008.01.016.

Deviance Information Criterion (DIC) in Bayesian Multiple QTL Mapping

Daniel Shriner* and Nengjun Yi

Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, AL 35294

Abstract

Mapping multiple quantitative trait loci (QTL) is commonly viewed as a problem of model selection. Various model selection criteria have been proposed, primarily in the non-Bayesian framework. The deviance information criterion (DIC) is the most popular criterion for Bayesian model selection and model comparison but has not been applied to Bayesian multiple QTL mapping. A derivation of the DIC is presented for multiple interacting QTL models and calculation of the DIC is demonstrated using posterior samples generated by Markov chain Monte Carlo (MCMC) algorithms. The DIC measures posterior predictive error by penalizing the fit of a model (deviance) by its complexity, determined by the effective number of parameters. The effective number of parameters simultaneously accounts for the sample size, the cross design, the number and lengths of chromosomes, covariates, the number of QTL, the type of QTL effects, and QTL effect sizes. The DIC provides a computationally efficient way to perform sensitivity analysis and can be used to quantitatively evaluate if including environmental effects, gene-gene interactions, and/or gene-environment interactions in the prior specification is worth the extra parameterization. The DIC has been implemented in the freely available package R/qtlbim, which greatly facilitates the general usage of Bayesian methodology for genome-wide interacting QTL analysis.

Keywords

complex trait; deviance; DIC; model selection and comparison; quantitative trait loci

INTRODUCTION

The aims of mapping quantitative trait loci (QTL) are to infer the number of QTL affecting the phenotype of interest, to determine the locations of the QTL, and to estimate the types and sizes of QTL effects. If a traditional regression model is used to relate genotypes to phenotypes, then one is attempting to optimize the fit of the specified model to the data and minimize error in predicting phenotypes. The fit of a model can be increased by simply increasing the number of predictors. However, increasing the complexity of the model beyond the complexity of the data is called overfitting and violates Occam's razor (also known as the parsimony principle). Hence, model selection is concerned with selecting "good" models that are not over-parameterized from a set of potential models. In the context of QTL mapping, model selection

*Corresponding author. RPHB 327, 1530 3rd Ave. S. Birmingham, AL 35294, Email: dshriner@ms.soph.uab.edu, Telephone: +1 (205) 975-9218, Fax: +1 (205) 975-2540.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

involves identifying genomic locations most likely to have replicable, non-zero effects on the phenotype (Broman and Speed, 2002).

One of the main difficulties with multiple QTL mapping is that the number of potential effects can be quite large. One way to address this problem is with variable selection. The aim of variable selection is to decide if potential effects have sizes sufficiently close to zero to justify excluding them from the model. Variable selection approaches to QTL mapping have been developed using forward selection, backward selection, and stepwise selection in the frequentist framework (Zeng et al., 1999) and using stochastic search variable selection in the Bayesian framework (Yi et al., 2003a). Another way to address this problem is with shrinkage estimators. The least absolute shrinkage and selection operator (Tibshirani, 1996; Xu, 2007) and shrinkage priors (Xu, 2003; Wang et al., 2005; Yi et al., 2005) have been used to shrink estimates of QTL effects toward zero.

An alternative approach to multiple QTL mapping is to perform model selection using an information criterion, which is a function of model adequacy penalized for model complexity. The goal of this approach is to select a single “good” model. In the maximum-likelihood framework, the two most commonly used information criteria are the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978). The AIC is given by

$$AIC = -2\ln(p(\text{data}|\hat{\theta})) + 2k, \quad (1)$$

with $p(\cdot|\cdot)$ being the likelihood function, $\hat{\theta}$ representing the maximum-likelihood estimates of the parameters in the model, and k representing an asymptotic bias correction term for the number of parameters (Burnham and Anderson, 2002). A second-order bias correction should be applied for small sample sizes n , yielding the AIC_c :

$$AIC_c = -2\ln(p(\text{data}|\hat{\theta})) + 2k + \frac{2k(k+1)}{n-k+1}, \quad (2)$$

(Burnham and Anderson, 2002). The BIC is given by

$$BIC = -2\ln(p(\text{data}|\hat{\theta})) + k\ln(n), \quad (3)$$

(Schwarz, 1978). The BIC tends to favor models with fewer predictors than the AIC since the BIC penalizes parameters more heavily.

In the context of QTL mapping, several modified forms of the BIC have been developed for non-Bayesian model selection. Broman and Speed (2002) noted that the original BIC, despite its tendency to be conservative, tended to include too many QTL. Ball (2001) and Broman and Speed (2002) developed a form of the BIC called BIC_δ .

$$BIC_\delta = -2\ln(p(\text{data}|\hat{\theta})) + \delta k\ln(n), \quad (4)$$

with $\delta > 1$ representing a way to increase the penalization for the number of parameters. Bogdan et al. (2004) developed another modification of the BIC to address models with gene-gene interaction terms in backcross designs by including an extra penalty depending upon the number of markers as well as the prior expected numbers of main effects and gene-gene interactions. Baierl et al. (2006) further extended the modified BIC of Bogdan et al. (2004) for intercross designs. The properties of the BIC for models including environmental effects or gene-environment interactions have not yet been reported.

One problem with model selection using the AIC or the BIC is that inference ignores uncertainty in the model. Bayesian model averaging accounts for model uncertainty by averaging the posterior distributions under each model considered, weighted by the posterior model probabilities (Madigan and Raftery, 1994). One of the main difficulties in implementing Bayesian model averaging is specifying the prior distribution over competing models (Hoeting et al., 1999).

Bayes factors are analogous to likelihood ratio statistics; the likelihood of the data is averaged over the parameters rather than being maximized (Kass and Raftery, 1995). Suppose the data are assumed to have been generated under one of two hypotheses, H_1 and H_2 , with prior probabilities $Pr(H_1)$ and $Pr(H_2)$, respectively. According to Bayes' theorem,

$$\frac{Pr(H_1|\text{data})}{Pr(H_2|\text{data})} = \frac{Pr(\text{data}|H_1)}{Pr(\text{data}|H_2)} \frac{Pr(H_1)}{Pr(H_2)}, \quad (5)$$

with the Bayes factor (Jeffreys, 1961) being defined as

$$K = \frac{Pr(\text{data}|H_1)}{Pr(\text{data}|H_2)}. \quad (6)$$

In the Bayesian framework, the densities in Equation 6 are obtained through integration over the parameter space for the corresponding hypothesis. Thus, the penalty for model complexity is implicit for Bayes factors and explicit for the BIC. Under certain regularity conditions, Kass and Wasserman (1995) showed that

$$BIC \approx -2\ln(K). \quad (7)$$

Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC), which is a generalization of the AIC and is intended for use with MCMC output and for hierarchical linear models. van der Linde (2005) incorporated the idea of the DIC into variable selection in conventional regression linear models. The DIC has been implemented in the general Bayesian software Bugs as a standard criterion for Bayesian model selection and model comparison (Spiegelhalter et al., 2002). The DIC has been used in genetics in the context of evaluating hierarchical models of gene-environment interactions (Fikse et al., 2003; Rekaya et al., 2003) and genetic evaluations in animal breeding (Legarra et al., 2005). However, the DIC has not been applied to multiple QTL mapping.

In the past decade, several Bayesian methods for QTL mapping have been developed (Satagopan and Yandell, 1996; Satagopan et al., 1996; Heath, 1997; Sillanpää and Arjas,

1998; Stephens and Fisch, 1998; Gaffney, 2001; Hoeschele, 2001; Sen and Churchill, 2001; Xu, 2003; Wang et al., 2005; Zhang et al., 2005). Markov chain Monte Carlo (MCMC) approaches have been developed to map multiple interacting QTL (Yi and Xu, 2002; Yi et al., 2003b; Narita and Sasaki, 2004; Yi et al., 2005; Yi et al., 2007). We have implemented our Bayesian methods in a freely available package R/qlbim (Yandell et al., 2007). Yi et al. (2005) performed posterior analysis by summarizing the MCMC output using Bayesian model averaging.

To calibrate Bayesian model averaging with respect to prior specification, we could use either Bayes factors or the DIC. In this study, we derive the DIC for multiple interacting QTL models and show how to calculate the DIC by using posterior samples generated by MCMC algorithms. We explore the usefulness of the DIC in choosing among models with main genetic effects, environmental effects, gene-gene interactions, and gene-environment interactions in the context of complex trait analysis. We show that the DIC provides a computationally simple way to examine the effect of specification of the prior distribution on posterior inference as part of Bayesian MCMC analysis.

MATERIALS AND METHODS

The Multiple Interacting QTL Model

We follow the linear regression framework of Yi et al. (2007) for multiple interacting QTL models. The phenotypes $\mathbf{y} = (y_1, \dots, y_n)^T$ are expressed as

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}_G \boldsymbol{\beta}_G + \mathbf{X}_{GG} \boldsymbol{\beta}_{GG} + \mathbf{X}_E \boldsymbol{\beta}_E + \mathbf{X}_{GE} \boldsymbol{\beta}_{GE} + \mathbf{e}, \quad (8)$$

in which $\boldsymbol{\mu} = (\mu, \dots, \mu)^T$ is the overall mean; $\mathbf{X}_G$, $\mathbf{X}_{GG}$, $\mathbf{X}_E$, and $\mathbf{X}_{GE}$ are the design matrices for main effects, gene-gene interactions, environmental effects, and gene-environment interactions, respectively; $\boldsymbol{\beta}_G$, $\boldsymbol{\beta}_{GG}$, $\boldsymbol{\beta}_E$, and $\boldsymbol{\beta}_{GE}$ are vectors of main effects, gene-gene interactions, environmental effects, and gene-environment interactions, respectively; and $\mathbf{e}$ is a vector of independent normal errors with mean zero and variance $\sigma^2 \mathbf{I}_n$. In Equation 8, $\mathbf{X}_G \boldsymbol{\beta}_G$ represents the sum of all main effects, $\mathbf{X}_{GG} \boldsymbol{\beta}_{GG}$ represents the sum of all gene-gene interactions, $\mathbf{X}_E \boldsymbol{\beta}_E$ represents the sum of all environmental effects, and $\mathbf{X}_{GE} \boldsymbol{\beta}_{GE}$ represents the sum of all gene-environment interactions. It is important to note that each of these four summations represents an arbitrary number of terms. This means, for example, that our focus is not on the number of gene-gene interactions. Rather, we focus on the simpler question of whether we should include or exclude gene-gene interactions from the model.

Bayesian methods can treat the unknown number of QTL as a random variable, which results in the complication that model space has variable dimensionality. A reversible jump MCMC algorithm offers a way to sample a posterior distribution of variable dimensionality (Green, 1995). Implementation of these algorithms is challenging due to the need to ensure reversible jumping. A composite model space approach circumvents this problem by augmenting variable dimensionality to fixed dimensionality by placing an upper bound on the dimensionality (in this case, the unknown number of QTL) (Yi, 2004). In the composite model space approach, a vector of latent indicator variables $\boldsymbol{\gamma}$ is used to indicate which effects (main effects, gene-gene interactions, and gene-environment interactions) across the possible loci are included in or excluded from the model (Yi 2004; Yi et al., 2005; Yi et al., 2007). The indicator vector $\boldsymbol{\gamma}$ determines the number of included QTL and the associated genetic effects. We denote the positions of the included QTL by $\boldsymbol{\lambda}$. The vectors $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ thus determine the genetic architecture, i.e., the number and positions of QTL and their gene actions. The goal of our Bayesian approach is to infer the posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ and estimate the associated effects. Since the

individual effects included in the model change during MCMC sampling, it is not easy to calculate the DIC at the level of individual effects.

The DIC for Multiple Interacting QTL Models

Following Spiegelhalter et al. (2002) and Gelman et al. (2004), for data $\mathbf{y}$ and model parameters $\boldsymbol{\theta}$, statistical deviance $D(\mathbf{y}, \boldsymbol{\theta})$ is defined as

$$D(\mathbf{y}, \boldsymbol{\theta}) = -2\ln(p(\mathbf{y}|\boldsymbol{\theta})). \quad (9)$$

The deviance at the posterior mean of $\boldsymbol{\theta}$, denoted by $\bar{\boldsymbol{\theta}}$, is

$$D_{\bar{\boldsymbol{\theta}}}(\mathbf{y}) = D(\mathbf{y}, \bar{\boldsymbol{\theta}}). \quad (10)$$

The posterior mean deviance is

$$\bar{D}(\mathbf{y}, \boldsymbol{\theta}) = E(D(\mathbf{y}, \boldsymbol{\theta})|\mathbf{y}). \quad (11)$$

The difference between the posterior mean deviance and the deviance at the posterior means is an estimate of the effective number of parameters, pD ,

$$pD = \bar{D}(\mathbf{y}, \boldsymbol{\theta}) - D(\mathbf{y}, \bar{\boldsymbol{\theta}}), \quad (12)$$

(Spiegelhalter et al., 2002).

Spiegelhalter et al. (2002) then derived the deviance information criterion, DIC, as

$$DIC = 2\bar{D}(\mathbf{y}, \boldsymbol{\theta}) - D(\mathbf{y}, \bar{\boldsymbol{\theta}}). \quad (13)$$

Substituting Equation 12 into Equation 13 and simplifying,

$$DIC = \bar{D}(\mathbf{y}, \boldsymbol{\theta}) + pD. \quad (14)$$

Equivalently,

$$DIC = D(\mathbf{y}, \bar{\boldsymbol{\theta}}) + 2pD. \quad (15)$$

The DIC in Equation 15 can be seen as analogous to the AIC in Equation 1, with two essential differences. One, the plug-in deviance in the AIC is calculated at maximum-likelihood estimates of parameters whereas the plug-in deviance in the DIC is calculated at posterior means (or, alternatively, posterior medians or posterior modes). Two, the penalty function for

model complexity in the AIC is determined by the nominal number of parameters in the model whereas the penalty function for model complexity in the DIC is determined by an estimate of the effective number of parameters in the model. The nominal number of parameters assumes that parameters are independent, whereas the effective number of parameters allows for nonzero covariance among parameters. Furthermore, the nominal number of parameters can be difficult to ascertain, especially in hierarchical models, whereas the effective number of parameters is estimated from the data.

Consider the t^{th} MCMC sample, $t = \{1, \dots, T\}$. Further consider the i^{th} individual, $i = \{1, \dots, n\}$. The deviance for the i^{th} individual at the t^{th} MCMC sample is

$$D_i^{(t)} = -2\ln(L_i^{(t)}). \quad (16)$$

Let $\mathbf{X}$ represent the design matrix combining $\mathbf{X}_G$, $\mathbf{X}_{GG}$, $\mathbf{X}_E$, and $\mathbf{X}_{GE}$. Similarly, let $\boldsymbol{\beta}$ represent the combination of $\boldsymbol{\beta}_G$, $\boldsymbol{\beta}_{GG}$, $\boldsymbol{\beta}_E$, and $\boldsymbol{\beta}_{GE}$. Then, let

$$\mathbf{G} = \mathbf{X}\boldsymbol{\beta}, \quad (17)$$

which represents a vector with a single value per individual reflecting the combination of all genetic effects, environmental effects, and interactions, which greatly simplifies the calculation of the DIC. Let $\boldsymbol{\theta}$ represent the vector of the mean, μ , the vector of effects, $\mathbf{G}$, and the variance, σ^2 . For the t^{th} MCMC sample, $\boldsymbol{\theta}^{(t)} = \mu^{(t)}, \mathbf{G}^{(t)}, \sigma^{2(t)}$. Substituting into the kernel of a normal distribution,

$$L_i^{(t)} = p(y_i | \boldsymbol{\theta}^{(t)}) = \frac{1}{\sigma^{(t)} \sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^{2(t)}} (y_i - \mu^{(t)} - G_i^{(t)})^2\right). \quad (18)$$

Substituting Equation 18 into Equation 16 and simplifying,

$$D_i^{(t)} = \ln(2\pi\sigma^{2(t)}) + \frac{(y_i - \mu^{(t)} - G_i^{(t)})^2}{\sigma^{2(t)}}. \quad (19)$$

Then, the deviance over all individuals is given by

$$D^{(t)} = \sum_{i=1}^n \left[\ln(2\pi\sigma^{2(t)}) + \frac{(y_i - \mu^{(t)} - G_i^{(t)})^2}{\sigma^{2(t)}} \right], \quad (20)$$

which simplifies to

$$D^{(t)} = n \ln(2\pi\sigma^{2(t)}) + \frac{1}{\sigma^{2(t)}} \sum_{i=1}^n (y_i - \mu^{(t)} - G_i^{(t)})^2. \quad (21)$$

The posterior mean deviance across the MCMC samples can be estimated by

$$\bar{D}(\mathbf{y}, \boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^T D^{(t)}. \quad (22)$$

Similarly, the posterior means of the parameters are given by

$$\bar{\mu} = \frac{1}{T} \sum_{t=1}^T \mu^{(t)}, \quad (23)$$

$$\bar{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \sigma^{2(t)}, \quad (24)$$

and

$$\bar{G}_i = \frac{1}{T} \sum_{t=1}^T G_i^{(t)}. \quad (25)$$

The deviance at the posterior means of the parameters can then be estimated by substituting Equations 23, 24, and 25 into Equation 10,

$$D(\mathbf{y}, \bar{\boldsymbol{\theta}}) = n \ln(2\pi\bar{\sigma}^2) + \frac{1}{\bar{\sigma}^2} \sum_{i=1}^n (y_i - \bar{\mu} - \bar{G}_i)^2. \quad (26)$$

The pD and DIC in Equations 12 and 13 can be easily calculated from Equations 22 and 26.

Implementation in R/qtlbim

The freely available QTL mapping package R/qtlbim (www.qtlbim.org) is an extensible, interactive environment for Bayesian analysis of multiple interacting QTL in experimental crosses (Yandell et al., 2007). It provides several efficient MCMC algorithms for evaluating the posterior of genetic architectures, i.e., the number and locations of QTL, their main and epistatic effects, and gene-environment interactions. R/qtlbim provides tools to monitor mixing behavior and convergence of the simulated Markov chain, and provides extensive informative graphical and numerical summaries of the MCMC output to infer and interpret the genetic architecture of complex traits. Code implementing DIC calculations has been included in R/qtlbim, which greatly facilitates the general usage of Bayesian model selection and model comparison for genome-wide interacting QTL analysis.

Simulations

We simulated and analyzed data using R/qtlbim (Yandell et al., 2007). The different experimental conditions are detailed in Table 1. We simulated an F₂ design. The genome

consisted of four autosomes, each of length 100 cM. On each chromosome, ten fully observed markers were equally spaced. Pseudomarkers were inserted at 1 cM intervals and genotype probabilities were estimated using the multipoint method. The phenotypic trait was assumed to be normally distributed. For each experimental condition, 1,000 independent data sets were replicated.

RESULTS

Testing main effects vs. gene-gene interactions

We compared the fit of a model of main effects (the true model) to the fit of a model of main effects and gene-gene interactions (the false model). For data generated under the conditions of Experiment 1, we compared the DIC for a prior specification of main effects vs. a prior specification of main effects and gene-gene interactions, with a prior expectation of one QTL. As shown in Table 2, in the absence of gene-gene interactions, a true model of main effects was fit almost equally well by a model of main effects and a model of gene-gene interactions, despite the model of gene-gene interactions being more heavily penalized for having more parameters.

To perform the reciprocal test, we simulated data consisting of one gene-gene interaction between two QTL. We compared the fit of a model of main effects (the false model) to the fit of a model of main effects and gene-gene interactions (the true model). For data generated under the conditions of Experiment 2, we compared the DIC for a prior specification of main effects vs. a prior specification of main effects and gene-gene interactions, with two expected QTL. In the presence of gene-gene interactions, a true model of gene-gene interactions was fit better by a model of gene-gene interactions than by a model of main effects.

Regardless of which model was the true model, the probability of choosing the correct model increased with increasing the sample size to 500 (Experiments 3 and 4) and 1,000 (Experiments 5 and 6). The results suggest that for effect sizes typical of complex traits, sample sizes of thousands are required to credibly classify the type of genetic effect. The results also suggest that a main effect can often be well explained by a model of gene-gene interactions but that it is difficult to explain a gene-gene interaction by a model of main effects.

Testing environmental effects

In the Bayesian framework, a fixed effect is treated as a random effect with infinite prior variance. We simulated a fixed effect for a binary covariate, such as sex. We compared data generated for Experiments 1 and 7. The prior expectation was for one QTL. As shown in Table 3, in the absence of fixed effects, a model excluding fixed effects provided a better fit than a model including fixed effects. In the presence of fixed effects, a model including fixed effects provided a better fit than a model excluding fixed effects.

Testing gene-environment interactions

We investigated how well a true model of main and fixed effects (Experiment 7) was estimated by a model of main effects compared to a model of main and fixed effects or a model of main and fixed effects and gene-environment interaction terms. The prior expectation was for one QTL. As shown in Table 4, in the absence of gene-environment interactions, a model including fixed effects provided a better fit than a model without fixed effects, and was chosen with marginal probability $0.525 + 0.225 = 0.750$. The model without the gene-environment interaction terms provided a slightly better fit than the model with the gene-environment interaction terms. In this three-way comparison, the correct model was chosen with probability 0.525.

In the presence of gene-environment interactions (Experiment 8), a model including fixed effects provided a better fit than a model without fixed effects, and was chosen with marginal probability $0.097 + 0.835 = 0.932$. The model with gene-environment interactions provided the best fit. In this three-way comparison, the correct model was chosen with probability 0.835.

Testing multiple effects

We simulated data containing a main effect, a gene-gene interaction, a fixed environmental effect, and a gene-environment interaction (Experiment 9). The prior expectation was for four QTL. We compared the fits of six different models: 1) a false model of main effects; 2) a false model of main effects and gene-gene interactions; 3) a false model of main and fixed effects; 4) a false model of main effects, fixed effects, and gene-environment effects; 5) a false model of main effects, fixed effects, and gene-gene interactions; and 6) the true model of main effects, gene-gene interactions, fixed effects, and gene-environment interactions. As shown in Table 5, the correct model provided the best fit and was chosen with probability 0.601. Models containing fixed effects were chosen with marginal probability $0.026 + 0.237 + 0.072 + 0.601 = 0.936$. Models containing gene-environment interactions were chosen with marginal probability $0.237 + 0.601 = 0.838$. Models containing gene-gene interactions were chosen with marginal probability $0.052 + 0.072 + 0.601 = 0.715$.

DISCUSSION

In this study, we illustrated the use of the deviance information criterion (DIC) in multiple QTL mapping. We simulated main effects, gene-gene interactions, environmental effects, and gene-environment interactions. We intentionally simulated small effect sizes (each effect explained <6% of the phenotypic variance) in order to test the DIC in difficult situations such as those regularly encountered during analysis of complex traits. For hierarchical models, the user may choose the level at which to evaluate the DIC (Spiegelhalter et al., 2002). In the context of QTL mapping, this means that we may choose to evaluate the DIC at the narrower level of individual effects or at the broader level of classes of effects. In the composite model space approach, the set of parameters that are included in the model varies during MCMC iteration (Yi, 2004). We propose to address this problem by focusing on classes of effects rather than on individual effects. We show that the DIC can be used to provide evidence for including each of the four classes of effects in the model.

The DIC has been derived from a decision-theoretic basis (Spiegelhalter et al., 2002; van der Linde, 2004) and from an information-theoretic basis (Spiegelhalter et al., 2002; van der Linde 2005). In the context of QTL mapping, complex traits are thought to be influenced by many small effects. The goal of the DIC is to allow selection of the model that not only best explains the observed data but also minimizes uncertainty about future observations. We interpret our results to indicate that simpler models (say, models containing just main effects) and more complex models (say, models containing main effects and gene-gene interactions) are often indistinguishably “good” at predicting future observations with moderate sample sizes (~250 individuals). Furthermore, it is more difficult to distinguish simple vs. complex models when the data-generating model is simple than when the data-generating model is complex. The implication is that while moderate sample sizes are usually sufficient to provide adequate power to detect a genetic effect in complex trait analysis, much larger sample sizes (>1,000 individuals) will be needed to credibly differentiate between types of genetic effects.

The Akaike information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz, 1978) are both based on maximum likelihood. For output of Monte Carlo Markov chain (MCMC) algorithms, maximum-likelihood estimates of parameters are difficult or impossible to obtain. The DIC makes use of posterior central measures that are easily obtained from MCMC output (Spiegelhalter et al., 2002). Also, both the AIC and the BIC penalize for

model complexity based on a known number of estimable parameters. For complicated models, such as hierarchical models, or in the case of correlated parameters, the number of effective parameters is at best poorly defined. The DIC estimates the number of effective parameters directly from the data (Spiegelhalter et al., 2002). As the DIC is a measure of posterior predictive error, comparison of non-nested models, accounting for different prior distributions, is straightforward (van der Linde, 2005).

The regression model of Yi et al. (2007) consists of a phenotypic mean, a vector of effects, and residual variance. The vector of effects is a complicated data structure that represents the product of the combined design matrix and the combined vector of all effect sizes. In the Bayesian framework, the joint posterior distribution of all unknowns simultaneously takes into account the sample size, the cross design, the number of chromosomes, the lengths of chromosomes, covariates, the number of QTL, the type of QTL effects, and QTL effect sizes. The DIC provides a computationally simple and efficient way to estimate the number of effective parameters while simultaneously accounting for all of these factors. The composite space model (Yi, 2004) allows for model selection on a fixed dimensional space, with the upper bound on the dimension constrainable by the user such that the “large p , small n ” problem is mitigated. Thus, there should be no need for modifications of the DIC similar to modified versions of the BIC (Bogdan et al., 2004; Baierl et al., 2006).

The AIC, the BIC, and the DIC all belong to a class of tools for model selection based on model adequacy and a penalty for model complexity. We can include Bayes factors in this class as well, based on the asymptotic equivalence of Bayes factors and the BIC (Kass and Raftery, 1995; Kass and Wasserman, 1995). Future directions include extending the DIC calculations for other types of traits, such as binary and ordinal traits, and investigating the usefulness of Bayes factors in sensitivity analysis in the same way we used the DIC (i.e., by summing over classes of effects in a hierarchical linear model). Importantly, the usefulness of the DIC in sensitivity analysis with respect to the prior expectation of the number of QTL should also be examined.

As an alternative approach, we are investigating posterior predictive model checking (Gelman et al., 2004). The basic idea of posterior predictive model checking is to simulate data under the selected model and assess how well the simulated data match the original data. This technique requires a user-defined test statistic in order to compare the predictive distribution with the original data. Uses of posterior predictive model checking include analysis of outliers and assessment of overparameterization (Congdon, 2006). Compared to model selection techniques, posterior predictive model checking is generally more flexible but more difficult to implement and evaluate.

Acknowledgments

This work was supported in part by National Institutes of Health grants and DK062710 (DS) and GM069430 (NY). We thank the Enabling Technology Laboratory in the University of Alabama at Birmingham Department of Mechanical Engineering for assistance with high performance computing. We thank the anonymous reviewers for comments that helped to improve this manuscript.

References

- Akaike H. A new look at the statistical model identification. *IEEE T Automat Contr* 1974;AC-19:716–723.
- Baierl A, Bogdan M, Frommlet F, Futschik A. On locating multiple interacting quantitative trait loci in intercross designs. *Genetics* 2006;173:1693–1703. [PubMed: 16624924]

- Ball RD. Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics* 2001;159:1351–1364. [PubMed: 11729175]
- Bogdan M, Ghosh JK, Doerge RW. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* 2004;167:989–999. [PubMed: 15238547]
- Broman KW, Speed TP. A model selection approach for the identification of quantitative trait loci in experimental crosses. *J R Stat Soc B* 2002;64:641–656.
- Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociol Method Res* 2002;33:261–304.
- Congdon, P. *Bayesian Statistical Modelling*. 2. John Wiley & Sons, Ltd; West Sussex, England: 2006.
- Fikse WF, Rekaya R, Weigel KA. Genotype $\times$ environment interaction for milk production in Guernsey cattle. *J Dairy Sci* 2003;86:1821–1827. [PubMed: 12778593]
- Gaffney, PJ. PhD Dissertation. University of Wisconsin; Madison, WI: 2001. An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses.
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DS. *Bayesian Data Analysis*. 2. Chapman & Hall/CRC; Boca Raton, Florida: 2004.
- Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995;82:711–732.
- Heath SC. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* 1997;61:748–760. [PubMed: 9326339]
- Hoeschele, I. Mapping quantitative trait loci in outbred pedigrees. In: Balding, DJ.; Bishop, M.; Cannings, C., editors. *Handbook of Statistical Genetics*. Wiley; New York: 2001. p. 599–644.
- Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Stat Sci* 1999;14:382–417.
- Jeffreys, H. *Theory of Probability*. 3. Oxford University Press; Oxford, UK: 1961.
- Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc* 1995;90:773–795.
- Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J Am Stat Assoc* 1995;90:928–934.
- Legarra A, López-Romero P, Ugarte E. Bayesian model selection of contemporary groups for BLUP genetic evaluation in Latxa dairy sheep. *Liv Prod Sci* 2005;93:205–212.
- Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J Am Stat Assoc* 1994;89:1535–1546.
- Narita A, Sasaki Y. Detection of multiple QTL with epistatic effects under a mixed inheritance model in an outbred population. *Genet Sel Evol* 2004;36:415–433. [PubMed: 15231232]
- Rekaya R, Weigel KA, Gianola D. Bayesian estimation of parameters of a structural model for genetic covariances between milk yield in five regions of the United States. *J Dairy Sci* 2003;86:1837–1844. [PubMed: 12778595]
- Satagopan, JM.; Yandell, BS. Estimating the number of quantitative trait loci via Bayesian model determination. Special Contributed Paper Session on Genetic Analysis of Quantitative Traits and Complex Disease, Biometric Section, Joint Statistical Meetings; Chicago. 1996.
- Satagopan JM, Yandell BS, Newton MA, Osborn TC. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* 1996;144:805–816. [PubMed: 8889541]
- Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978;6:461–464.
- Sen S, Churchill GA. A statistical framework for quantitative trait mapping. *Genetics* 2001;159:371–387. [PubMed: 11560912]
- Sillanpää MJ, Arjas E. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* 1998;148:1373–1388. [PubMed: 9539450]
- Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J R Statist Soc B* 2002;64:583–616.
- Stephens DA, Fisch RD. Bayesian analysis of quantitative trait loci data using reversible jump Markov chain Monte Carlo. *Biometrics* 1998;54:1334–1347.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Statist Soc B* 1996;58:267–288.

- van der Linde A. On the association between a random parameter and an observable. *Test* 2004;13:85–111.
- van der Linde A. DIC in variable selection. *Stat Neerl* 2005;59:45–56.
- Wang H, Zhang YM, Li X, Masinde GL, Mohan S, Baylink DJ, Xu S. Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* 2005;170:465–480. [PubMed: 15781696]
- Xu S. Estimating polygenic effects using markers of the entire genome. *Genetics* 2003;163:789–801. [PubMed: 12618414]
- Xu S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* 2007;63:513–521. [PubMed: 17688503]
- Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R, Moon JY, Neely WW, Wu H, von Smith R, Yi N. R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics* 2007;23:641–643. [PubMed: 17237038]
- Yi N, Xu S. Mapping quantitative trait loci with epistatic effects. *Genet Res* 2002;79:185–198. [PubMed: 12073556]
- Yi N, George V, Allison DB. Stochastic search variable selection for identifying quantitative trait loci. *Genetics* 2003a;164:1129–1138. [PubMed: 12871920]
- Yi N, Xu S, Allison DB. Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics* 2003b;165:867–883. [PubMed: 14573494]
- Yi N. A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* 2004;167:967–975. [PubMed: 15238545]
- Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* 2005;170:1333–1344. [PubMed: 15911579]
- Yi N, Shriner D, Banerjee S, Mehta T, Pomp D, Yandell BS. An efficient Bayesian model selection approach for interacting QTL models with many effects. *Genetics* 2007;176:1865–1877. [PubMed: 17483424]
- Zeng ZB, Kao CH, Basten CJ. Estimating the genetic architecture of quantitative traits. *Genet Res* 1999;74:279–289. [PubMed: 10689805]
- Zhang M, Montooth KL, Wells MT, Clark AG, Zhang D. Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* 2005;169:2305–2318. [PubMed: 15520261]

Table 1

Description of experimental conditions.

Experiment	Sample Size	Main effects	Gene-gene interactions	Environmental effects	Gene-environment interactions	Total percent variance explained
1 ^a	250	a=0.25, d=0.25	n/a	n/a	n/a	~4.5
2 ^b	250	a=0, d=0	aa=0.5	n/a	n/a	~6
3 ^a	500	a=0.25, d=0.25	n/a	n/a	n/a	~4.5
4 ^b	500	a=0, d=0	aa=0.5	n/a	n/a	~6
5 ^a	1000	a=0.25, d=0.25	n/a	n/a	n/a	~4.5
6 ^b	1000	a=0, d=0	aa=0.5	n/a	n/a	~6
7 ^a	250	a=0.25, d=0.25	n/a	0.25	n/a	~6
8 ^a	250	a=0, d=0	n/a	0.25	a=0.5, d=0.5	~6
9 ^c	250	a=0.25, d=0.25	aa=0.5	0.25	a=0.5, d=0.5	~15

^aThe QTL was located at 45 cM on chromosome 1. The QTL had both additive (a) and dominance (d) effects.

^bThe QTL were located at 45 cM on chromosome 1 and 60 cM on chromosome 2. The QTL had an additive-additive effect only.

^cThe main effect QTL was located at 45 cM on chromosome 1. The gene-gene interacting QTL were located at 60 cM on chromosome 2 and 45 cM on chromosome 3. The gene-environment interacting QTL was located at 45 cM on chromosome 4.

Table 2

Model selection between main effects and gene-gene interactions.

True Model	Sample size	ΔDIC^a	% chosen ^b
Main effect	250	0.29 (2.55)	78.6
Gene-gene interaction	250	6.29 (10.23)	76.7
Main effect	500	1.35 (1.91)	90.5
Gene-gene interaction	500	16.23 (13.09)	90.1
Main effect	1000	1.51 (1.74)	92.5
Gene-gene interaction	1000	40.27 (16.84)	99.6

^a“ ΔDIC ” indicates the difference between the DIC from the false model and the DIC from the true model. Shown are the averages (standard deviations) of the differences over 1,000 simulation replicates.

^b“% chosen” indicates the percentage of simulation replicates for which the true model had the better (i.e., smaller) DIC.

Table 3

Model selection between main effects and environmental effects.

True Model	ΔDIC^a	% chosen ^b
Main effect	2.92 (4.27)	70.8
Main and fixed effect	1.01 (1.31)	83.0

^a“ ΔDIC ” indicates the difference between the DIC from the false model and the DIC from the true model. Shown are the averages (standard deviations) of the differences over 1,000 simulation replicates.

^b“% chosen” indicates the percentage of simulation replicates for which the true model had the better (i.e., smaller) DIC.

Table 4
Model selection among main effects, environmental effects, and gene-environment interactions.

True Model ^c	ΔDIC^a			% chosen ^b		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Model 2	2.93 (4.26)	0 (0)	0.24 (2.25)	25.0	52.5	22.5
Model 3	11.7 (9.22)	9.02 (8.31)	0 (0)	6.8	9.7	83.5

^a“ ΔDIC ” indicates the difference between the DIC from the model in the column and the DIC from the true model in the row. Shown are the averages (standard deviations) of the differences over 1,000 simulation replicates.

^b“% chosen” indicates the percentage of simulation replicates for which the model in the column had the best (i.e., smallest) DIC.

^cModel 1 includes only main effects. Model 2 includes main effects and environmental effects. Model 3 includes main effects, environmental effects, and gene-environment interactions.

Table 5

Model selection among main effects, environmental effects, gene-gene interactions, and gene-environment interactions.

Model ^a	Δ DIC ^b	% chosen ^c
Model 1	16.84 (14.12)	1.2
Model 2	14.14 (13.62)	2.6
Model 3	6.10 (11.32)	23.7
Model 4	11.66 (11.11)	5.2
Model 5	8.46 (9.31)	7.2
Model 6	0 (0)	60.1

^aModel 1 includes only main effects. Model 2 includes main effects and environmental effects. Model 3 includes main effects, environmental effects, and gene-environment interactions. Model 4 includes main effects and gene-gene interactions. Model 5 includes main effects, environmental effects, and gene-gene interactions. Model 6 includes main effects, environmental effects, gene-gene interactions, and gene-environment interactions.

^b" Δ DIC" indicates the difference between the DIC from the model and the DIC from the true model 6. Shown are the averages (standard deviations) of the differences over 1,000 simulation replicates.

^c"% chosen" indicates the percentage of simulation replicates for which the model had the best (i.e., smallest) DIC.