Extended Bayesian Information Criteria for Model Selection with Large Model Spaces

By JIAHUA CHEN

Department of Statistics, University of British Columbia, Vancouver,

British Columbia, V6T 1Z2 Canada

jhchen@stat.ubc.ca

and ZEHUA CHEN

Department of Statistics and Applied Probability, National University of Singapore,

Singapore 117546

stachenz@nus.edu.sg

Summary

The ordinary Bayes information criterion is too liberal for model selection when the model space is large. In this article, we re-examine the Bayesian paradigm for model selection and propose an extended family of Bayes information criteria. The new criteria take into account both the number of unknown parameters and the complexity of the model space. Their consistency is established, in particular allowing the number of covariates to increase to infinity with the sample size. Their performance in various situations is evaluated by simulation studies. It is demonstrated that the extended Bayes information criteria incur a small loss in the positive selection rate but tightly control the false discovery rate, a desirable property in many applications. The extended Bayes information criteria are extremely useful for variable selection in problems with a moderate sample size but a huge number of covariates, especially in genome-wide association studies, which are now an active area in genetics research.

Some keywords: Bayesian paradigm; Consistency; Genome-wide association study; Tournament approach; Variable selection.

1. Introduction

In many applications a variable of interest is influenced by a number of unidentified covariates among a large collection of potential covariates, whose number is much larger than the number of observations. For example, in genome-wide association studies, geneticists type tens or hundreds of thousands of single nucleotide polymorphisms spreading over the whole genome to identify a handful of them that are responsible for the genetic variation of a quantitative trait or a disease status; see Marchini et al. (2005). In principle, the statistical issue involved is simply a variable selection problem. However the sheer number of covariates P and the comparatively small sample size n make the variable-selection problem a great statistical challenge. In such situations, classical criteria such as the Akaike information criterion or AIC (Akaike, 1973), the Bayes information criterion or BIC (Schwarz, 1978), and other methods such as cross validation and generalized cross validation (Stone, 1974; Craven & Wahba, 1979), are usually too liberal; that is, they tend to select a model with many spurious covariates. This phenomenon has been observed by Broman & Speed (2002), Siegmund (2004) and Bogdan et al. (2004), in their use of BIC for quantitative trait loci mapping, and will also be shown later in this article.

Variable selection with large model spaces has drawn increasing attention recently. Meinshausen & Bühlmann (2006) and Zhao & Yu (2006) investigated consistency properties, while Zhang & Huang (2008) studied the sparsity and bias properties of the Lasso-based variable-selection methods (Tibshirani, 1996). To ensure consistency of the Lasso-based variable-selection procedure, the tuning parameter must be set to an appropriate asymptotic order, and the design matrix must satisfy a sparse Riesz condition.

In this paper, we propose a class of extended Bayes information criteria to better

meet the needs of variable selection for large model spaces. The original BIC is an approximate Bayes approach, see Berger & Pericchi (2001) and some details later in this article. The simplicity and effectiveness of the BIC have made it very attractive, even when the regularity conditions are not satisfied. More recently, in unpublished work, J. O. Berger has developed a more rigorous Bayes approach called the generalized Bayes information criterion, which sticks more to the Bayes paradigm and refines the choice of prior distributions for various parametric models. However, Berger's criterion still deals mostly with the case where P is not large compared with n.

The extended Bayes information criterion family that we propose is particularly suitable for model selection for large model spaces. It includes the original BIC as a special case and retains its simplicity. Under some mild conditions, these new criteria are shown to be consistent. The result is particularly useful even when the covariates are heavily collinear. Furthermore, unlike competitors such as that of Meinshausen & Bühlmann (2006), the extended Bayes information criterion family does not require a data adaptive tuning parameter procedure in order to be consistent, and hence is easy to use in applications.

2. An extended family of Bayes information criteria

Let $\{(y_i, x_i) : i = 1, \ldots, n\}$ be independent observations. Suppose that the conditional density function of y_i given x_i is $f(y_i|x_i, \theta)$, where $\theta \in \Theta \subset R^P$, P being a positive integer. The likelihood function of θ is given by

$$L_n(\theta) = f(x; \theta) = \prod_{i=1}^{n} f(y_i|x_i, \theta),$$

where $Y = (y_1, \ldots, y_n)$. Let s be a subset of $\{1, \ldots, P\}$. Denote by $\theta(s)$ the parameter θ with those components outside s being set to 0 or some prespecified values. The

BIC proposed by Schwarz (1978) selects the model that minimizes

$$\text{BIC}(s) = -2\log L_n\{\hat{\theta}(s)\} + \nu(s)\log n,$$

where $\hat{\theta}(s)$ is the maximum likelihood estimator of $\theta(s)$, and $\nu(s)$ is the number of components in s. Let $\mathcal{S}$ be the model space under consideration and let $p(s)$ be the prior probability of model s. Assume that, given s, the prior density of $\theta(s)$ is given by $\pi\{\theta(s)\}$. The posterior probability of s is obtained as

$$p(s|Y) = \frac{m(Y|s)p(s)}{\sum_{s\in\mathcal{S}} p(s)m(Y|s)},$$

where $m(Y|s)$ is the likelihood of model s, given by

$$m(Y|s) = \int f\{Y;\theta(s)\}\pi\{\theta(s)\}d\theta(s).$$

Under the Bayes paradigm, a model s^* that maximizes the posterior probability is selected. Since $\sum_{s\in\mathcal{S}} p(s)m(Y|s)$ is a constant, $s^* = \text{argmax}_{s\in\mathcal{S}} m(Y|s)p(s)$. Under some regularity conditions on $f(Y;\theta)$ such as the requirement that s must contain all the nonzero components of θ and have constant dimension, the maximum likelihood estimator of $\theta(s)$ is root-n consistent, and $-2\log\{m(Y|s)\}$ has a Laplace approximation given by $\text{BIC}(s)$ up to an additive constant. That is, the BIC is an approximate Bayes approach as mentioned in introduction. An implicit assumption underlying BIC is that $p(s)$ is constant for s over $\mathcal{S}$.

It is well known that BIC is consistent (Rao & Wu, 1989) under some standard conditions such as P is fixed. In nonregular problems such as change-point analysis, the root-n consistency of $\hat{\theta}(s)$ may be violated, yet BIC is still consistent (Yao, 1988; Csörgö & Horváth, 1997). Nevertheless, BIC is not without drawbacks. The precision of the Laplace approximation is influenced by the specific form of the prior density on $\theta(s)$ and the correlation structure between observations. The latter affects the

interpretation of the sample size n in the definition of BIC(s). The recent unpublished work of Berger, and Clyde et al. (2007) have concentrated on these issues. They have focused on the marginal likelihood $m(Y|s)$ and rectified the problems caused by the Laplace approximation. However, they have not targeted the problems that could be caused by large model spaces.

In a typical genome-wide association study with single nucleotide polymorphisms, the number of covariates is of the order of tens or hundreds of thousands while the sample size is only in hundreds. Suppose the number of covariates under consideration is $P = 1000$. The class of models containing a single covariate, $\mathcal{S}_1$, has size 1000, while the class of models containing two covariates, $\mathcal{S}_2$, has size $1000 \times 999/2$. The constant prior behind BIC amounts to assigning probabilities to the S_j proportional to their sizes. Thus the probability assigned to S_2 is $999/2$ times that assigned to S_1. The size of S_j increases as j increases to $j = P/2 = 500$, so that the probability assigned to S_j by the prior increases almost exponentially. Models with a larger number of covariates, 50 or 100 say, receive much higher probabilities than models with fewer covariates. This is obviously unreasonable, being strongly against the principle of parsimony.

This re-examination of BIC prompts us naturally to consider other reasonable priors over the model space in the Bayes approach. Assume that the model space $\mathcal{S}$ is partitioned into $\cup_{j=1}^{P} \mathcal{S}_j$, such that models within each $\mathcal{S}_j$ have equal dimension. Let $\tau(\mathcal{S}_j)$ be the size of $\mathcal{S}_j$. For example, if $\mathcal{S}_j$ is the collection of all models with j covariates, $\tau(\mathcal{S}_j) = \binom{P}{j}$. We assign the prior distribution over $\mathcal{S}$ as follows. For each s in the same subspace $\mathcal{S}_j$, assign an equal probability, i.e., $\text{pr}(s|\mathcal{S}_j) = 1/\tau(\mathcal{S}_j)$ for any $s \in \mathcal{S}_j$. This is reasonable since all the models in $\mathcal{S}_j$ are equally plausible. Then, instead of assigning probabilities $\text{pr}(\mathcal{S}_j)$ proportional to $\tau(\mathcal{S}_j)$, as in the ordinary BIC,

we assign $\mathrm{pr}(\mathcal{S}_j)$ proportional to $\tau^{\xi}(\mathcal{S}_j)$ for some ξ between 0 and 1. This results in the prior probability $p(s)$ for $s \in \mathcal{S}_j$ being proportional to $\tau^{-\gamma}(\mathcal{S}_j)$ where $\gamma = 1 - \xi$. This type of prior distribution on the model space gives rise to an extended BIC family as follows:

$$\mathrm{BIC}_\gamma(s) = -2 \log L_n\{\hat{\theta}(s)\} + \nu(s) \log n + 2\gamma \log \tau(\mathcal{S}_j), \quad 0 \le \gamma \le 1,$$

where $\hat{\theta}(s)$ is the maximum likelihood estimator of $\theta(s)$ given model s. The first two terms in $\mathrm{BIC}_\gamma(s)$ are the Laplace approximation to $-2\log\{m(Y|s)\}$ and the last term is indeed $-2\log\{p(s)\}$ up to a common constant. The criterion BIC_γ is referred to as an extended Bayes information criterion. In the targeted application, P can be very large but the cardinality of the candidate models is small. Thus, the Laplace approximation is still valid. However, the consistency of the extended Bayes information criteria does not depend on the validity of the Laplace approximation, as will be seen later.

If some of the covariates are heavily collinear, the effective number of different models might be smaller than that indicated by $\tau(\mathcal{S}_j)$. Will this cause any serious detrimental effect on our method? Let us consider an extreme case in which half of the covariates are duplicates. Thus, in considering $\tau(\mathcal{S}_j)$, P should be replaced by $P/2$. However, it is easy to see that, when P is replaced by $P/2$ the change in $\gamma \log\{\tau(\mathcal{S}_j)\}$ is of a smaller order than the order $\log n + \log P$ of the leading terms. Thus, some adjustment might be helpful but the effect will not be important when either n or P is large.

With a similar motivation, Bogdan et al. (2004) proposed adding to the BIC penalty an additional term $\nu(s)\log(l-1)$, where l is to be chosen to reflect the prior knowledge on the number of quantitative trait loci in the context of genetic interval mapping. They showed via simulation that their modified BIC performs well

but did not investigate its consistency properties. As will be seen, adding a term that increases with P is crucial in ensuring the consistency when P increases to infinity with n.

3. The consistency of the extended Bayes information criteria

We consider the consistency of the extended Bayes information criteria in the following setting. Let $P = p_n = O(n^\kappa)$ as $n \to \infty$ for some $\kappa > 0$. Note that this does not require that, in practice, we must have a process such that P and n go to infinity. Also, it does not require, but allows, P to be large. In applications, the values of n and P are constants. The asymptotic result provides insight into properties of the statistical method when n and P are large and of the sizes indicated by these orders. This is also a common setting in the model selection literature; see Shao (1997), Meinshausen & Bühlmann (2006), and Zhang & Huang (2008).

Let y_n be the vector of n observations on the response variable, let X_n be the corresponding design matrix with all the covariates of concern, and let β be the vector of regression coefficients. Assume that

$$y_n = X_n\beta + e_n, \tag{1}$$

where $e_n \sim N(0, \sigma^2 I_n)$ and I_n is the identity matrix of size n. Let s_0 be the smallest subset of $\{1, \ldots, p_n\}$ such that $\mu_n = Ey_n = X_n(s_0)\beta(s_0)$, where $X_n(s_0)$ and $\beta(s_0)$ are respectively the design matrix and the coefficients corresponding to s_0. In general, a model is considered to be identifiable when no two sets of parameter values specify the same distribution. However this notion of identifiablity is not appropriate for the small-n-large-P problem of the linear models. In the following, we introduce an identifiability condition suitable for this problem.

We call s_0 the true submodel and denote $\nu(s_0)$ by K_0. Let the projection matrix

of $X_n(s)$ be $H_n(s) = X_n(s)\{X_n^{\mathrm{T}}(s)X_n(s)\}^{-1}X_n^{\mathrm{T}}(s)$. Define

$$\Delta_n(s) = \|\mu_n - H_n(s)\mu_n\|^2,$$

with $\|\cdot\|$ being the Euclidean norm. Clearly, if $s_0 \subset s$, we have $\Delta_n(s) = 0$. The new identifiability condition is as follows.

Condition 1: asymptotic identifiability. Model (1) with true submodel s_0 is asymptotically identifiable if

$$\lim_{n\to\infty} \min\{(\log n)^{-1}\Delta_n(s) : s \neq s_0, \nu(s) \leq K_0\} = \infty.$$

In other words, the model is identifiable if no model other than the true submodel of comparable size can predict the response almost equally well. This identifiability requirement is probably the weakest possible. It is weaker than the commonly assumed so-called sparse Riesz condition that will be discussed below. Nevertheless, it provides a sufficient condition for consistency of extended Bayes information criteria.

The sparse Riesz condition is an identifiability condition assumed by Zhang & Huang (2008) and others. It requires that $n^{-1}X_n^{\mathrm{T}}(s)X_n(s)$ be uniformly positive definite for any s of size $2K$, where K is an upper bound for K_0. The sparse Riesz condition implies the foregoing asymptotic identifiability condition. A simple proof is given as follows. For any submodel s, we have

$$\begin{aligned}
(\log n)^{-1}\Delta_n &= (\log n)^{-1}\|\mu_n - H_n(s)\mu_n\|^2 \\
&= (\log n)^{-1}\inf_{\alpha}\|X_n(s_0 - s)\beta(s_0 - s) - X(s)\alpha\|^2 \\
&= (\log n)^{-1}\inf_{\alpha}[\{\beta^{\mathrm{T}}(s_0 - s), \alpha\}\{X_n^{\mathrm{T}}(s_0 \cup s)X_n(s_0 \cup s)\}\{\beta^{\mathrm{T}}(s_0 - s), \alpha\}^{\mathrm{T}}] \\
&\geq n(\log n)^{-1}\lambda(s_0 \cup s)\|\beta(s_0 - s)\|^2,
\end{aligned}$$

where $\lambda(s_0 \cup s)$ is the smallest eigenvalue of $n^{-1}X_n^{\mathrm{T}}(s_0 \cup s)X_n(s_0 \cup s)$. By the sparse Riesz condition, $\lambda(s_0 \cup s)$ is uniformly larger than 0 over s of size K. The asymptotic identifiability is hence apparent.

In an earlier version of this article, we proposed the sparse Riesz condition independently albeit under a different name. As pointed out by an anonymous referee this condition is void, for instance, when any two columns of X_n are completely collinear. This possibility cannot be ignored in applications where $P >> n$.

We now provide a more useful sufficient condition. Let $s_0 = \{1, \ldots, K_0\}$ and let s_{-k} be the set with the kth element of s_0 removed. Let $H_n(s_{-k} \cup s)$ be the projection matrix of $X_n(s_{-k} \cup s)$.

Lemma 1. *The asymptotic identifiability condition is satisfied when*

$$\lim_{n\to\infty} \min_{s\neq s_0, \nu(s)\leq K_0} \max_k [(\log n)^{-1} \|\{I - H_n(s_{-k} \cup s)\} X_n(\{k\})\|] = \infty.$$

A proof will be given in the Appendix. In other words, this condition requires that at least one column of $X_n(s_0)$ be not contained in the linear space of the remaining columns of $X_n(s_0 \cup s)$. For example, if $X_n(\{1\})$ is orthogonal to $X_n(\{k\})$ for $k = 2, \ldots, P$, this condition is satisfied even if the covariates are heavily collinear. In particular, duplicating columns not in $X_n(s_0)$ does not affect the asymptotic identifiability. This is not true for the sparse Riesz condition.

The result in Lemma 1 serves as a guideline rather than as a condition to be directly verified because s_0 is unknown. In principle, we could examine this condition for all s_0 with given size $K_0 \leq K$ for some K. Then we would have verified, whether or not the sparse Riesz condition holds. However, this is impractical because of the size of P, and it is unnecessary since the sparse Riesz condition is much more stringent. A better solution is to use our method to identify a candidate $\hat{s}_0$ first, and then verify the condition of Lemma 1 with this $\hat{s}_0$. If none of $\max_k \|\{I - H_n(s_{-k} \cup s)\} X_n(\{k\})\|$ is small, the condition could be practically regarded as satisfied. If a few submodels $s \neq \hat{s}_0$ are found to have small $\max_k \|\{I - H_n(s_{-k} \cup s)\} X_n(\{k\})\|$, scientific knowledge

can be used to judge their plausibility. If many such models are found, additional data must be collected to resolve the nonidentifiability.

We now state the consistency result as follows.

Theorem 1. *Assume that $p_n = O(n^\kappa)$ for some constant κ. If $\gamma > 1 - 1/(2\kappa)$, then, under the asymptotic identifiability condition,*

$$\mathrm{pr}[\min\{\textsc{bic}_\gamma(s) : \nu(s) = j, s \neq s_0\} > \textsc{bic}_\gamma(s_0)] \to 1$$

for $j = 1, \ldots, K$, as $n \to \infty$.

The proof is given in the Appendix. The theorem implies that, as $n \to \infty$, the probability that any model other than the true model will be selected tends to zero. Consequently, the false discovery rate, when $K_0 > 0$, goes to 0.

In addition, our proof reveals that when $p_n > \sqrt{n}$, the ordinary BIC is likely inconsistent. A non-technical explanation is as follows. Consider the situation in which the true model is an empty s_0. A one-covariate model will be selected over the true model when any one of its loglikelihoods exceeds the likelihood of s_0 by $0.5 \log n$. Using the classical result (Wilks, 1938), we can show that the maximum inflation in the loglikelihood is of order $0.5 \log n$ when $P \doteq \sqrt{n}$. Thus, a wrong one-covariate model will be selected with positive probability, and the ordinary BIC loses consistency. Clearly, since there are many more 2-covariate models $P(P-1)/2$, the false discovery rate increases when they are included in the model space.

4. Simulation studies

4.1. *Numerical consideration*

In the extended Bayes information criteria $\textsc{bic}_\gamma$, three values of γ are of special interest, namely, $\gamma = 0, 0.5$ and 1. The value 0 corresponds to the original BIC. The value 1 ensures consistency of extended Bayes information criteria when $p_n = O(n^\kappa)$

for any $\kappa \geq 0$ not depending on n, and the value 0.5 ensures consistency when $\kappa < 1$.

When P is large, we cannot afford to calculate $\text{BIC}_\gamma(s)$ for all possible s. Instead, we combine BIC_γ with a penalized likelihood technique developed by Tibshirani (1996) and others. Let

$$l_\lambda(\beta) = -2\log\{L_n(\beta)\} + n\sum_k p(|\beta_k|;\lambda),$$

where $p(|\beta_k|;\lambda)$ is a penalty function of $|\beta_k|$, the components of β, with λ a tuning parameter. If we choose a $p(\cdot)$ that has a spike at 0, the penalized negative loglikelihood function attains its minimum at some $\hat{\beta}$ with many zero components. Two common choices of $p(\cdot)$ are the L_1 penalty considered by Tibshirani (1996) and the smoothly clipped absolute deviation penalty proposed by Fan & Li (2001).

When $\lambda \to \infty$, the penalized negative loglikelihood function attains its minimum with β being 0 except for the intercept term. When $\lambda = 0$, the function attains its minimum at the least squares estimator. When λ increases, the number of zero components in $\hat{\beta}$ increases. We increase λ gradually so that the number of zero components of $\hat{\beta}$ gradually increases, and obtain a sequence of nested models. The values of BIC_γ are computed for this sequence of models and a model is selected.

When $P >> n$, the above approach is still computationally infeasible. Instead, a tournament described in by Z. Chen, J. Chen and J. Liu is used, so that the original set of all the covariates is randomly partitioned into disjoint subsets, each containing about $n/2$ covariates. The foregoing penalized likelihood is applied to each subset to obtain a $\hat{\beta}$ with a specific number of nonzero components, and therefore a much reduced subset of covariates. These reduced subsets are pooled and the procedure is repeated if necessary. Otherwise, the penalized likelihood procedure together with the extended Bayes information criteria is applied to the reduced pooled subset to obtain the final selected model.

4.2. *The model and the covariate structure*

Our simulation study examines the performance of the extended Bayes information criteria in three cases.

In Case 1, we set $P = 50$ and $n = 200$ with randomly generated covariates. In this situation in which $P << n$, the original BIC is applicable.

In Case 2, we set $P = 1000$ and $n = 200$ with randomly generated covariates. In this situation $P >> n$ and we expect the original BIC to fail in some way and the extended Bayes information criteria to excel.

In Case 3, covariates from a real data set were used with $n = 233$ and $P = 1414$. In this situation, EBIC_γ with $\gamma = 1$ should work best. The covariate structure in Case 3 is completely natural and somewhat unknown. It is crucial that the extended Bayes information criteria work in this situation before it is used in real data analysis.

The following linear model is generally employed in all three cases:

$$y_i = x_i^{\mathrm{T}}(s)\boldsymbol{\beta}(s) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

for some s, where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma = 1$ but is assumed unknown in the analysis.

In Case 1, the covariates are generated from $N(0, 1)$ with one of the following covariance structures:

(i) $\mathrm{cov}(x_{ij}, x_{ik}) = \rho$, for all pairs of j and k;

(ii) $\mathrm{cov}(x_{ij}, x_{ik}) = \rho$, for j and $k = j \pm 1$, and 0 otherwise;

(iii) $\mathrm{cov}(x_{ij}, x_{ik}) = \rho^{|k-j|}$ for all pairs of k and j.

Two values of ρ are explored, 0.2 and 0.4.

In Case 2, the covariates are generated in 20 groups of size 50. The first ten groups of covariates are generated as above. The other ten groups are generated from

a discrete distribution over $(-2, 0, 4)$ with probabilities $(0.5, 0.25, 0.25)$, and then scaled to have standard deviation 1. Once covariates are created, they remain fixed throughout the simulation studies. In each replicate, a submodel s of pre-spedified size is selected at random, and the response values are generated according to model (2) with pre-determined $\beta(s)$ which is kept unchanged throughout the simulation.

In each replicate, we compare the set s^* selected by the extended Bayes information criteria with the real model s. The average number of covariates correctly selected and the average number incorrectly selected are computed. We define the positive selection rate as the ratio $\nu(s \cap s^*)/\nu(s)$, and the false discovery rate as the ratio $\nu(s^* - s)/\nu(s^*)$. These mimic the power and type I error in hypothesis testing. The concept of a false discovery rate was first introduced in problems involving a huge number of multiple tests; see Benjamini & Hochberg (1995). It is a much more sensible measure for the problems considered in this article.

4.3. *Simulation results*

Case 1. We put $\beta(s) = (0, 0.7, 0.9, 0.4, 0.3, 1.0, 0.2, 0.2, 0.1)^{\mathrm{T}}$, which resemble a set of fitted values in a real data analysis. The results are reported in Table 1. The standard deviations based on the 200 repetitions are presented in parentheses, thereby providing information on simulation errors.

It turns out that BIC_1 has selected slightly fewer true covariates, but also fewer false covariates compared to BIC_0. To see this more clearly, we pooled outcomes with a different covariate structure and computed the positive selection rates and false discovery rates for $P = 50$. The pooling is sensible because the correlation structures do not seem to cause an appreciable difference in the trend of the positive selection rate and the false discovery rate. The outcomes are given in Table 2.

As expected, both the positive selection rate and the false discovery rate have

a declining trend from BIC_0 to $\text{BIC}_{0.5}$ to BIC_1, since BIC_1 is the most stringent and BIC_0 is the least stringent. The striking point is that the decline in the positive selection rate is inconspicuous but the difference in the false discovery rate is quite significant. In general, the positive selection rate of BIC_0 is slightly higher than those of $\text{BIC}_{0.5}$ and BIC_1, but BIC_0 suffers a much higher false discovery rate than the other two. It appears that the positive selection rate is affected by the correlation among the covariates, but the effects on the false discovery rate do not seem to show an appreciable difference.

We also conducted simulations for $P = 20$, $\sigma = 2$, and so on. The results are similar and are omitted.

Case 2. We let $\beta = (0, 1.0, 0.7, 0.5, 0.3, 0.2)^{\mathrm{T}}$ to generate the response variable according to model (2). The tournament approach was applied. The set of 1000 covariates was randomly partitioned into subsets of equal size 100 in the first round. From each subset, 12 covariates were selected by the penalized likelihood method to enter the second round of competition. In the second round, these 120 selected covariates were pooled and 20 of them were selected as the final candidate covariates. Three final models were selected according to BIC_0, $\text{BIC}_{0.5}$ and BIC_1. The average numbers of correctly and incorrectly selected covariates and the corresponding positive selection rate and false discovery rate are presented in Table 3.

As shown in Table 3, the false discovery rate of BIC_0 is intolerably high: the lowest is 61.6% and the highest is 76%. The high false discovery rate makes BIC_0 an inappropriate model selection criterion for small-n-large-P problems. However, BIC_1 effectively controls the false discovery rate, which is about 5% in all the cases considered, while it retains a high positive selection rate. Again, the performance of $\text{BIC}_{0.5}$ falls between these two.

Case 3. We use covariates of a real dataset from a genetic genome-wide association study in this simulation. In the genetic study, B lymphocytes from blood samples of individuals are transformed into immortalized lymphoblastoid cell lines by the Epstein-Barr virus. In the transformation, the Epstein-Barr virus genes are expressed in the lymphoblastoid cell lines. Of interest is whether the mRNA expression level of a particular gene, the EBNA-3A, is associated with any single nucleotide polymorphisms over the human genome. Data were collected from 16 pedigrees with a total of 233 individuals. For each individual, the mRNA expression level of the EBNA-3A gene were obtained together with the genotypes at 2155 single nucleotide polymorphisms spread over 23 chromosomes. The dataset was originally analyzed in the report by Z. Chen, J. Chen, and J. Liu. As the result of a preliminary analysis, only 1414 single nucleotide polymorphisms are retained in the analysis because the others either are uninformative or have a large proportion of missing genotypes.

In the simulation studies, the dataset is used in the following ways.

Setting 1. The observed response values are randomly permuted and reassigned to the individuals so that the possible association between the response variable and the single nucleotide polymorphism genotypes is de-linked.

Setting 2. The response values are randomly generated from model (2) under the assumption of no association, while the observed response values are ignored.

Setting 3. The response values are generated from model (2) with three single nucleotide polymorphisms having major effects, where, in each simulation replicate, these three single nucleotide polymorphisms are randomly selected from the 1414 that are available:

$$\beta(s) = (0, -1.56, -1.09, 1.22, -0.06, -0.08, -0.012, 0.067, -0.047, -0.07, 0.05)^{\mathrm{T}}.$$

The extra seven nonzero coefficients are not considered useful and are not used to

compute positive selection rates and so on.

Setting 4. The trait values are generated as in Setting 3 but with ten single nucleotide polymorphisms having minor effects:

$$\beta(s) = (0, -0.31, 0.23, 0.42, -0.32, -0.33, -0.26, 0.41, 0.29, -0.35, -0.69)^{\mathrm{T}}.$$

In Settings 1 and 2, we assess the performance of the model selection criteria when there is no covariate associated with the response variable. In Settings 3 and 4, we assess how the model selection criteria perform when the effects of the covariates are at different levels. The model selection criteria are applied with the tournament procedure, and the number of simulation replicates is 200.

The simulation results are given in Table 4. When there is no single nucleotide polymorphism associated with the trait, BIC_1 tightly controls false discovery, virtually no single nucleotide polymorphisms being wrongly discovered, but BIC_0 has little control over false discovery; in both settings 1 and 2, about 6 single nucleotide polymorphisms are wrongly discovered. The criterion $\text{BIC}_{0.5}$ reasonably controls false discovery but less satisfactorily than BIC_1.

Under Setting 3, the positive selection rates of the three criteria are the same but their false discovery rates are very different: BIC_0 suffers a false discovery rate as high as 54.1% while $\text{BIC}_{0.5}$ and BIC_1 effectively control the false discovery rate, at rates 5.9% and 0.7% respectively.

Under Setting 4, although there is a decline in the positive selection rate from BIC_0 to BIC_1, the decline is not dramatic; the difference in the false discovery rate is again large between BIC_0 and BIC_1. The general pattern in the positive selection rate and the false discovery rate among the three criteria is essentially the same as that of Setting 2. The results of Setting 3 reveal an additional feature: if the effects of the truly associated covariates are prominent, the difference in the positive selection rate

among the three criteria seems small.

5. Further discussion

The choice of the extended Bayes information criteria to use in a particular problem is an important issue. The version BIC_1 is consistent as long as P does not increase with n exponentially. The version $\text{BIC}_{0.5}$ is consistent when $\kappa < 1$. In the genome-wide association study, one can be highly confident about the single nucleotide polymorphisms selected by BIC_1, while those selected by $\text{BIC}_{0.5}$ are subject to further investigation. Another way of choosing γ is to solve for κ from $P = n^\kappa$ and then to set $\gamma = 1 - 1/(2\kappa)$.

Although consistency is proved under the normality assumption and a linear regression model, the criteria are obviously applicable without these assumptions. Furthermore, the consistency result is probably valid much more widely. Development of such a theory will be a topic of future research.

Our result implies that, when $p_n = O(n^\kappa)$ with $\kappa < 0.5$, the original BIC is consistent. Shao (1997), Li (1987) and Rao & Wu (1989) have discussed many consistent variable-selection procedures, yet our consistency result has extended our understanding of the original BIC for the small-n-large-P problem.

Acknowledgment

The research was partially supported by the Natural Science and Engineering Research Council of Canada, and by Research Grant R-155-000-065-112 of the National University of Singapore. The authors are grateful to the referee, associate editor and editor for their constructive suggestions.

Appendix

Proofs

Proof of Lemma 1. Note that, for each $k \in s_0$ and s, we have

$$\begin{aligned}\Delta_n(s) &= \|X_n(s_0)\beta(s_0) - H_n(s)X_n(s_0)\beta(s_0)\|^2 \\ &= \inf_{\alpha} \|X_n(s_0)\beta(s_0) - X_n(s)\alpha\|^2 \\ &\geq \inf_{\alpha} \|X_n(k)\beta(k) - X_n(s_{-k} \cup s)\alpha\|^2 \\ &= |\beta(k)|^2 \|\{I - H_n(s_{-k} \cup s)\}X_n(k)\|^2.\end{aligned}$$

In the foregoing derivation, α represents a regression coefficient of proper dimension. Let $c = \min\{|\beta(k)| : k \in s_0\} > 0$. We obtain that

$$(\log n)^{-1}\Delta_n(s) \geq c(\log n)^{-1} \max_k \{\|\{I - H_n(s_{-k} \cup s)\}X_n(k)\|^2\},$$

which goes to infinity uniformly in s under the condition. This completes the proof of asymptotic identifiability.

Proof of Theorem 1. Without loss of generality, we assume that $\sigma^2 = 1$.

We consider the case $s_0 \not\subset s$ first. Note that

$$y_n^{\mathrm{T}}\{I_n - H_n(s_0)\}y_n = e_n^{\mathrm{T}}\{I_n - H_n(s_0)\}e_n = \sum_{i=1}^{n-\nu(s_0)} Z_j^2 = n\{1 + o_p(1)\},$$

where the Z_j are independent standard normal variables. Furthermore,

$$\begin{aligned}&y_n^{\mathrm{T}}\{I_n - H_n(s)\}y_n - e_n^{\mathrm{T}}\{I_n - H_n(s_0)\}e_n \\ &\quad = \mu_n^{\mathrm{T}}\{I_n - H_n(s)\}\mu_n + 2\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}e_n - e_n^{\mathrm{T}}H_n(s)e_n + e_n^{\mathrm{T}}H_n(s_0)e_n.\end{aligned}$$

By asymptotic identifiability, uniformly over s such that $\nu(s) \leq K$, we have

$$(\log n)^{-1}\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}\mu_n \to \infty.$$

Write

$$\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}e_n = \sqrt{[\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}\mu_n]}Z(s),$$

where

$$Z(s) = \frac{\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}e_n}{\sqrt{[\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}\mu_n]}} \sim N(0,1).$$

We hence arrive at

$$\begin{aligned}
\max[\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}e_n : s \in \mathcal{S}_j] &\leq \sqrt{[\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}\mu_n]}\max\{Z(s) : s \in \mathcal{S}_j\} \\
&\leq \sqrt{[\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}\mu_n]}O_p\{\sqrt{(2\log p_n)}\} \\
&= o_p[\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}\mu_n],
\end{aligned}$$

where the last inequality follows the Bonferoni inequality.

Since $H_n(s)$ is a projection matrix, we have

$$e_n^{\mathrm{T}}H_n(s)e_n = Z_1^2(s) + \cdots + Z_j^2(s)$$

for some independent standard normal random variables. Thus, by Bonferoni inequality,

$$\max\{e_n^{\mathrm{T}}H_n(s)e_n : s \in \mathcal{S}_j\} = O_p(\log p_n) = O_p(\log n).$$

The term $e_n^{\mathrm{T}}H_n(s_0)e_n$ is a χ^2-distributed statistic with a fixed degrees of freedom K_0.

In summary, $\mu_n^{\mathrm{T}}\{I_n - H_n(s)\}\mu_n$, which goes to infinity faster than $\log n$, is the dominating term in $y_n^{\mathrm{T}}\{I_n - H_n(s)\}y_n - e_n^{\mathrm{T}}\{I_n - H_n(s_0)\}e_n$. Thus,

$$\frac{y_n^{\mathrm{T}}\{I_n - H_n(s)\}y_n - e_n^{\mathrm{T}}\{I_n - H_n(s)\}e_n}{e_n^{\mathrm{T}}\{I_n - H_n(s)\}e_n} \geq \frac{C\log n}{n},$$

for any large constant C in probability, and

$$\begin{aligned}
\log\left[\frac{y_n^{\mathrm{T}}\{I_n - H_n(s)\}y_n}{y_n^{\mathrm{T}}\{I_n - H_n(s_0)\}y_n}\right] &= \log\left[1 + \frac{y_n^{\mathrm{T}}\{I_n - H_n(s)\}y_n - e_n^{\mathrm{T}}\{I_n - H_n(s_0)\}e_n}{e_n^{\mathrm{T}}\{I_n - H_n(s_0)\}e_n}\right] \\
&\geq \log\{1 + C(\log n)/n\}.
\end{aligned}$$

Note that the second and the third terms in BICγ are of order $\log n$. Hence, choosing $C > K(1 + 2\gamma\kappa)$, we obtain

$$\text{BIC}\gamma(s) - \text{BIC}\gamma(s_0) \geq n\log\{1 + C(\log n)/n\} - K(1 + 2\gamma\kappa)\log n \to \infty,$$

as $n \to \infty$ uniformly in $s \in \mathcal{S}_j$ for all $j = 1, \ldots, K$.

We now turn to the case $s_0 \subset s$, for which we have $\{I_n - H_n(s)\}X_n(s_0) = 0$. Hence, $y_n^{\mathrm{T}}\{I_n - H_n(s)\}y_n = e_n^{\mathrm{T}}\{I_n - H_n(s)\}e_n$ and

$$e_n^{\mathrm{T}}\{I_n - H_n(s_0)\}e_n - e_n^{\mathrm{T}}\{I_n - H_n(s)\}e_n = e_n^{\mathrm{T}}\{H_n(s) - H_n(s_0)\}e_n = \sum_{i=1}^{j} Z_i^2(s),$$

where $j = \nu(s) - \nu(s_0)$ and the $Z_i(s)$ are some independent standard normal random variables depending on s. Let $\hat{e}_n = \{I_n - H_n(s_0)\}e_n$. We obtain that

$$\begin{aligned} n\big(\log[e_n^{\mathrm{T}}\{I - H_n(s_0)\}e_n] - \log[e_n^{\mathrm{T}}\{I - H_n(s)\}e_n]\big) &= n\log\left\{1 + \frac{\sum_{i=1}^{j} Z_i^2(s)}{\hat{e}_n^{\mathrm{T}}\hat{e}_n - \sum_{i=1}^{j} Z_i^2(s)}\right\} \\ &\leq \frac{n\sum_{i=1}^{j} Z_i^2(s)}{\hat{e}_n^{\mathrm{T}}\hat{e}_n - \sum_{i=1}^{j} Z_i^2(s)}. \end{aligned}$$

As $n \to \infty$, $n^{-1}\hat{e}_n^{\mathrm{T}}\hat{e}_n \to \sigma^2 = 1$. It follows that

$$\max\{\sum_{i=1}^{j} Z_i^2(s) : s \in \mathcal{S}_{\nu(s)}\} = 2j\log p_n\{1 + o_p(1)\}.$$

Thus,

$$\begin{aligned} \max\left\{\frac{n\sum_{i=1}^{j} Z_i^2(s)}{\hat{e}_n^{\mathrm{T}}\hat{e}_n - \sum_{i=1}^{j} Z_i^2(s)} : s \in \mathcal{S}_{\nu(s)}\right\} &\leq \frac{2nj\log p_n\{1 + o_p(1)\}}{(n - 2j\log p_n)\{1 + o_p(1)\}} \\ &= 2j\log p_n\{1 + o_p(1)\}. \end{aligned}$$

Consequently, uniformly in s such that $\nu(s) = \nu(s_0) + j$,

$$\textsc{bic}\gamma(s) - \textsc{bic}\gamma(s_0) \geq j\{\log n + (2\gamma - 2)\log p_n\} = j\{1 + 2\kappa(\gamma - 1)\}\log n.$$

When $\gamma > 1 - 1/(2\kappa)$, we have, as $n \to \infty$, that $\textsc{bic}\gamma(s) - \textsc{bic}\gamma(s_0) \to \infty$. The conclusion hence follows.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second Int. Symp. Info. Theory*, Ed. B.N. Petrox and F. Caski, pp: 267-81. Budapest: Akademiai Kiado.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate – A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289-300.

Berger, J. O. & Pericchi, L. R. (2001). Objective Bayesian method for model selection: Introduction and comparison. In *Model Selection*, Ed. P. Lahiri. pp 135-207. Hayward, CA: Inst. Math. Statist. Lecture Notes Monograph Series Volume 38.

Bogdan, M., Doerge, R. & Ghosh, J. K. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**, 989-99.

Broman, K. W. & Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. R. Statist. Soc.* B **64**, 641-56.

Clyde, M. A., Berger, J. O., Bullard, F., Ford, E. B., Jefferys, W. H., Luo, R., Paulo, R. & Loredo, T. (2007). Current challenges in Bayesian model choice. In *Statistical Challenges in Modern Astronomy IV*, Ed. G. F. Babu and E. D. Feigelson, pp 224-40. Astronomical Society of the Pacific Conference Series Volume 371.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-

validation. *Numer. Math.* **31**, 377-403.

Csörgö, M. & Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. New York: John Wiley & Sons.

Fan, J. & Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348-60.

Li, K.-C. (1987). Asymptotic optimality for *Cp, CL*, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15**, 958-75.

Marchini, J., Donnelly, P. & Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413-7.

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34**, 1436-62.

Rao, C. R. & Wu, Y. H. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76**, 369-74.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-4.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 221-64.

Siegmund, D. (2004). Model selection in irregular problems: Application to mapping quantitative trait loci. *Biometrika* **91**, 785-800.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with Discussion). *J. R. Statist. Soc.* B **39**, 111-47.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc.* B **58**, 267-88.

Wilks, S. S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Statist.* **9**, 60-2.

Yao, Y. C. (1988). Estimating the number of change-points via Schwartz criterion. *Statist. Prob. Lett.* **6**, 181-9.

Zhang, C. H. & Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* To Appear.

Zhao, P. & Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541-67.

Table 1: Case 1: Mean correct(C) and incorrect(IC) selections with $P = 50$ and $\sigma = 1$, standard deviations are in brackets

	$\rho = 0.2$			$\rho = 0.4$		
	BIC_0	$\text{BIC}_{0.5}$	BIC_1	BIC_0	$\text{BIC}_{0.5}$	BIC_1
	Correlation structure (i)					
C	3.94(0.23)	3.90(0.30)	3.82(0.40)	3.06(0.72)	2.48(0.81)	2.00(0.89)
IC	1.17(1.20)	0.34(0.63)	0.12(0.39)	1.15(1.12)	0.28(0.52)	0.08(0.30)
	Correlation structure (ii)					
C	3.96(0.18)	3.95(0.22)	3.92(0.28)	3.09(0.74)	2.75(0.75)	2.45(0.81)
IC	0.46(0.76)	0.24(0.54)	0.14(0.37)	0.42(0.73)	0.20(0.51)	0.08(0.33)
	Correlation structure (iii)					
C	3.98(0.12)	3.90(0.31)	3.74(0.47)	2.96(0.74)	2.52(0.83)	2.08(0.85)
IC	1.16(1.27)	0.34(0.68)	0.14(0.37)	0.95(1.02)	0.29(0.56)	0.06(0.24)

Table 2: Case 1: Pooled positive selection rates(PSR) and false discovery rates(FDR)

	$\rho = 0.2$			$\rho = 0.4$		
	BIC_0	$\text{BIC}_{0.5}$	BIC_1	BIC_0	$\text{BIC}_{0.5}$	BIC_1
PSR	0.495	0.490	0.478	0.380	0.323	0.272
FDR	0.190	0.073	0.034	0.217	0.090	0.033

Table 3: Simulation results for Case 2 with $P = 1000$, standard deviations are in brackets

σ		$\rho = 0.2$			$\rho = 0.4$		
		BIC_0	$BIC_{0.5}$	BIC_1	BIC_0	$BIC_{0.5}$	BIC_1
		Correct(C) *and incorrect*(IC) *selections*					
1	C	4.05(0.69)	3.72(0.66)	3.45(0.56)	2.83(0.68)	2.34(0.72)	1.82(0.67)
	IC	7.11(2.43)	0.72(1.01)	0.09(0.30)	8.98(2.36)	0.85(1.21)	0.05(0.24)
2	C	4.03(0.70)	3.71(0.67)	3.43(0.63)	2.87(0.78)	2.26(0.72)	1.74(0.66)
	IC	6.47(2.57)	0.76(1.06)	0.16(0.43)	8.72(2.66)	1.00(1.30)	0.10(0.31)
		Positive selection rates(PSR) *and false discovery rates*(FDR)					
1	PSR	0.810	0.744	0.690	0.566	0.468	0.364
2		0.806	0.742	0.686	0.574	0.452	0.348
1	FDR	0.637	0.162	0.025	0.760	0.266	0.027
2		0.616	0.170	0.045	0.752	0.307	0.054

Table 4: Simulation results for Case 3, standard deviations are in brackets

Setting	BIC_0	$\text{BIC}_{0.5}$	BIC_1
	Incorrect selection number		
1	6.19 (3.08)	0.23 (0.50)	0.00 (0.00)
2	6.61 (2.63)	0.34 (0.59)	0.01 (0.10)
3	3.51 (2.17)	0.19 (0.45)	0.02 (0.14)
4	3.97 (1.54)	0.94 (0.99)	0.47 (0.73)
	Correct selection number		
3	2.98 (0.14)	2.98 (0.14)	2.98 (0.14)
4	6.38 (1.39)	5.80 (1.73)	5.15 (2.02)
	False discovery rate		
3	0.541	0.060	0.007
4	0.384	0.139	0.084
	Positive selection rate		
3	0.993	0.993	0.993
4	0.638	0.580	0.515