Extended BIC for small-n-large-P sparse GLM

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 small-n-large-P situation has become common in genetics research, medical studies, risk management, and other fields. Feature selection is crucial in these studies yet poses a serious challenge. The traditional criteria such as AIC, BIC, and cross-validation choose too many features. To overcome the difficulties caused by the small-n-large-P situation, Chen and Chen (2008) developed a family of extended Bayes information criteria (EBIC). Under normal linear models, EBIC is found to be consistent with nice finite sample properties. Proving consistency for non-normal and nonlinear models poses serious technical difficulties. In this paper, through a number of novel techniques, we establish the consistency of EBIC under generalized linear models in the small-n-large-P situation. We also report simulation results and a real-data analysis to illustrate the effectiveness of EBIC for feature selection.

Keywords: Consistency; Exponential family; Extended Bayes information criterion; Feature selection; Generalized linear models; Small-n-large-P.

1 Introduction

The small-n-large-P situation has become common in genetics research, medical studies, risk management, and other fields. In these studies, researchers are interested not only in a relationship between a response variable and some explanatory features but also in the identification of causal features. Examples of such features include disease genes and quantitative trait loci in the human genome, biomarkers responsible for disease pathways, and stocks generating profits in investment portfolios. The selection of causal features is a crucial aspect of the abovementioned studies. The situation, where the sample size n is relatively small but the number of features P under consideration is extremely large, poses a serious challenge to the selection of causal features. Feature selection in the sense of identifying causal features is different from but often interwoven with model selection; the latter involves two operational components: a procedure for selecting candidate models, and a criterion for assessing the candidate models. In this article, we concentrate on the issue of model selection criteria.

Traditional model selection criteria such as Akaike's information criterion (AIC) (Akaike, 1973), cross-validation (CV) (Stone, 1974) and generalized cross-validation (GCV) (Craven and Wahba, 1979) essentially address the prediction accuracy of selected models. The popular Bayes information criterion (BIC) (Schwarz, 1978) was developed from the Bayesian paradigm in a different vein. BIC approximates the posterior model probability while the prior is uniform over all models. When the number of features is small and fixed, as in classical problems, these criteria all perform well because the goal of prediction accuracy does not conflict with feature selection. However, in the small-n-large-P situation, these criteria become overly liberal and fail to serve the purpose of feature selection. This phenomenon has been observed by Broman & Speed (2002), Siegmund (2004), and Bogdan et al. (2004) in genetic studies. Wang et al. (2007) also noticed that BIC is liberal in general. Small-n-large-P situations are abundant, see Donoho (2000), Singh et al. (2002), Marchini et al. (2005), Clayton et al. (2005), Fan and Li (2006), Zhang and Huang (2008), and Hoh et al. (2008).

Recently, Chen and Chen (2008) have developed a family of extended Bayes information criteria (EBIC) especially for feature selection in small-n-large-P situations. They have established the consistency of EBIC under normal linear models. Under a normal linear model, the least square estimates of the regression coefficients and the residuals have simple analytical forms and a multivariate normal distribution. It is relatively easy to assess the order of the tail probabilities of the maximums of various quantities over all possible models. Under general models, assessing the tail probabilities poses serious technical difficulties although EBIC is readily applied to non-normal and nonlinear models. In this paper, we have tailor-designed a tail probability inequality for linear combinations of random variables with exponential family distributions (Lemma 1) and a geometric relationship between an arbitrary hyper-ball and a fixed number of hyperplanes (Lemma 2). These technical results are of interest in themselves. Moreover, they are particularly useful in proving the uniform consistency of the maximum likelihood estimates of the coefficients in the linear predictor of all generalized linear models (GLM) containing causal features (Theorem 1), and the consistency of EBIC under GLM with canonical links (Theorem 2).

We have investigated the performance of EBIC under the logistic regression model with canonical links in simulation. Since the logistic regression model is valid in both prospective and retrospective studies, see McCullagh and Nelder (1989, Chapter 4), it plays an important role in case-control genome-wide association study which is a major approach in genetic research, see for example The Wellcome Trust Case-

Control Consortium (2007). Because it is impossible to evaluate all possible models, a computationally feasible procedure for selecting candidate models is needed. The traditional stepwise procedure in regression analysis is not appropriate because of its well-known greedy nature and the prohibitively huge amount of computation when P is large. Following a rather distinct line of thinking, a class of penalized likelihood-based methods have been developed in recent years, including the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and Elastic Net (Zou and Hastie, 2005). All these methods choose a penalty function that is singular at zero, and induce sparsity in regression models by increasing the penalty. Implicitly, candidate models are selected using different levels of penalty and then assessed by cross-validation. Among these methods, the LASSO is the most appealing to us because of an efficient R algorithm called `glmpath` recently developed by Park and Hastie (2007). In our simulation studies, we use LASSO to select candidate models, but instead of cross-validation, we use EBIC to assess models. The procedure is implemented by using the R function `glmpath`.

The remainder of the paper is arranged as follows. In Section 2 the GLM is briefly reviewed and its properties in the small-n-large-P framework are investigated. In Section 3, EBIC for GLM is introduced and its consistency is established. Simulation studies are reported in Section 4, and a real-data example is analyzed in Section 5.

2 The small-n-large-P sparse GLM and its asymptotic properties

Let Y be a response variable and $\boldsymbol{x}$ a vector of feature variables (hereafter, for convenience, the variables will be called simply features). A GLM consists of three components. The first component is an exponential family distribution assumed for

Y. The density function of the exponential family is of the form:

$$f(y;\theta) = \exp\{\theta^\tau y - b(\theta)\} \tag{1}$$

with respect to a σ-finite measure ν. The parameter θ is called the natural parameter and the set

$$\Theta = \left\{\theta : \int \exp\{\theta^\tau y\} d\nu < \infty\right\}.$$

is called the natural parameter space. The exponential family has the following properties: (a) The natural parameter space Θ is convex; (b) At any interior point of Θ, $b(\theta)$ has all derivatives and $b'(\theta) = E(Y) \equiv \mu$, $b''(\theta) = \text{Var}(Y) \equiv \sigma^2$; (c) At any interior point of Θ, the moment generating function of the family exists and is given by $M(t) = \exp\{b(\theta + t) - b(\theta)\}$. The second component of the GLM is a linear predictor given by $\eta = \boldsymbol{x}^\tau \boldsymbol{\beta}$; that is, the GLM assumes that the features affect the distribution of Y through this linear form. The third component of the GLM is a link function g that relates the mean μ to the linear predictor by $g(\mu) = \eta = \boldsymbol{x}^\tau \boldsymbol{\beta}$. For more details of the GLM, the reader is referred to McCullagh and Nelder (1989).

In this paper, we are particularly interested in feature selection problems under models with two special characteristics: (i) small-n-large-P, i.e., the number of features is much larger than the sample size; and (ii) sparsity, i.e., only a few unidentified features affect Y. We refer to a GLM with these two characteristics as the small-n-large-P sparse GLM. We investigate issues of feature selection in this framework.

Let $\mathcal{X}$ be the set of all features under consideration. Let $s_0 \in \mathcal{X}$ be the subset that contains and only contains all the features affecting Y. These features are called causal features. Let s denote any subset of $\mathcal{X}$. Denote by $\nu(s)$ the number of features in s, and let $\boldsymbol{\beta}(s)$ be the vector of the components in $\boldsymbol{\beta}$ that correspond to the features in s. Let $\boldsymbol{\beta}_0$ be the unknown true value of the parameters. Note that the components of $\boldsymbol{\beta}_0$ other than those corresponding to s_0 are all zero. Let $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ be the observations. Denote by $\boldsymbol{x}_i(s)$ the vector of the components of $\boldsymbol{x}_i$ that correspond to $\boldsymbol{\beta}(s)$. Let

$$\mathcal{B} = \{\boldsymbol{\beta} : \boldsymbol{x}_i^\tau \boldsymbol{\beta} \in \Theta, i = 1, 2, \ldots, n\}.$$

Note that $\mathcal{B}$ is convex. Let l_n be the log likelihood function; that is,

$$l_n(\beta) = \sum_{i=1}^{n} \log f(y_i; \theta_i),$$

where θ_i depends on $\boldsymbol{x}_i$ through the relationship $g(\mu_i) = \boldsymbol{x}_i^\tau \boldsymbol{\beta}$. Here g is the link function. In this article, we consider only the canonical link, i.e., $g(\mu_i) = \theta_i$, noting that θ_i is implicitly a function of μ_i. Let

$$\boldsymbol{s}_n(\boldsymbol{\beta}) = \frac{\partial l_n}{\partial \boldsymbol{\beta}}, \quad H_n(\boldsymbol{\beta}) = -\frac{\partial^2 l_n}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\tau}.$$

With the canonical link, we have

$$\begin{aligned} l_n(\beta) &= \sum_{i=1}^{n} [y_i \boldsymbol{x}_i^\tau \boldsymbol{\beta} - b(\boldsymbol{x}_i^\tau \boldsymbol{\beta})], \\ \boldsymbol{s}_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n} [y_i - b'(\boldsymbol{x}_i^\tau \boldsymbol{\beta})] \boldsymbol{x}_i, \\ H_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{x}_i b''(\boldsymbol{x}_i^\tau \boldsymbol{\beta})] \boldsymbol{x}_i^\tau = \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^\tau. \end{aligned}$$

Here β is a generic reduced $\beta(s)$, and $\boldsymbol{x}_i$, $\boldsymbol{s}_n$, and H_n are corresponding reduced quantities.

Let K be a constant larger than $\nu(s_0)$. Define

$$A_0 = \{s : s_0 \subset s; \nu(s) \leq K\},$$

and

$$A_1 = \{s : s_0 \not\subset s; \nu(s) \leq K\}.$$

We assume the following conditions:

A1 As $n \to \infty$, $P = O(n^{\kappa})$ for some $\kappa > 0$.

A2 The causal feature set s_0 is fixed.

A3 The interior of $\mathcal{B}$ is not empty and $\boldsymbol{\beta}_0 \in \mathcal{B}$.

A4 There exist positive constants c_1, c_2 such that for all sufficiently large n,

$$c_1 \leq \lambda_{\min}(n^{-1} H_n(\boldsymbol{\beta}_0(s))) \leq \lambda_{\max}(n^{-1} H_n(\boldsymbol{\beta}_0(s))) \leq c_2,$$

for all s such that $\nu(s) \leq K$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote respectively the smallest and the largest eigenvalues.

A5 For any given $\epsilon > 0$, there exists a constant $\delta > 0$ such that, when n is sufficiently large,

$$(1-\epsilon) H_n(\boldsymbol{\beta}_0(s)) \leq H_n(\boldsymbol{\beta}(s)) \leq (1+\epsilon) H_n(\boldsymbol{\beta}_0(s))$$

for all s and $\boldsymbol{\beta}(s)$ such that $\nu(s) \leq K$ and $\|\boldsymbol{\beta}(s) - \boldsymbol{\beta}_0(s)\| \leq \delta$.

A6 Denote by x_{ij} the jth component of $\boldsymbol{x}_i$.

$$\max_{1 \leq j \leq P} \max_{1 \leq i \leq n} \left\{ \frac{x_{ij}^2}{\sum_{i=1}^n x_{ij}^2 \sigma_i^2} \right\} = o((\log n)^{-1}).$$

Let $\hat{\boldsymbol{\beta}}(s)$ be the MLE of $\boldsymbol{\beta}(s)$ in the GLM with features in s. In the remainder of this section, we explore the uniform consistency of $\hat{\boldsymbol{\beta}}(s)$ for $s \in A_0$. The following lemma is used to establish the uniform consistency.

Lemma 1. *Let Y_i, $i = 1, 2, \ldots, n$, be independent random variables following exponential family distributions of form (1) with natural parameters θ_i. Let μ_i and σ_i^2 denote the mean and variance of Y_i respectively. Suppose that $\{\theta_i; i = 1, 2, \ldots, n\}$ is* *contained in a compact subset of the natural parameter space Θ. Let a_{ni}, $i = 1, \dots, n$, be real numbers such that*

$$\sum_{i=1}^{n} a_{ni}^2 \sigma_i^2 = 1, \quad \max_{1 \le i \le n} \{a_{ni}^2\} = o((\log(n))^{-1}).$$

Then, for any $m > 0$, as $n \to \infty$, we have

$$P\left(\sum_{i=1}^{n} a_{ni}(Y_i - \mu_i) > \sqrt{2m \log n}\right) = o(n^{-m}).$$

Proof: For convenience, denote $\sqrt{2m \log n}$ by q_n. It is trivial to see that

$$I(\sum_{i=1}^{n} a_{ni}(Y_i - \mu_i) > q_n) \le \exp\{t[\sum_{i=1}^{n} a_{ni}(Y_i - \mu_i) - q_n]\}$$

for any $t > 0$. Then

$$\begin{aligned} P(\sum_{i=1}^{n} a_{ni}(Y_i - \mu_i) > q_n) &\le E[\exp\{t[\sum_{i=1}^{n} a_{ni}(Y_i - \mu_i) - q_n]\}] \\ &= \exp\{\sum_{i=1}^{n} [b(\theta_i + a_{ni}t) - b(\theta_i) - \mu_i a_{ni} t] - q_n t\}, \end{aligned}$$

where the last equality follows from the moment generating function of the exponential family. For any t_n such that $\max_{1 \le i \le n}\{|a_{ni}t_n|\} \to 0$, we have

$$\sum_{i=1}^{n} [b(\theta_i + a_{ni}t_n) - b(\theta_i) - \mu_i a_{ni} t] = \frac{t_n^2}{2} \sum_{i=1}^{n} a_{ni}^2 b''(\theta_i + a_{ni}\tilde{t}_n) = \frac{t_n^2}{2}\{1 + o(1)\},$$

noting that, because of the uniform continuity of $b''(\theta)$ in the compact subset, $b''(\theta_i + a_{ni}\tilde{t}_n)$ converges to σ_i^2 uniformly. In particular, letting $t_n = q_n$, we have

$$\sum_{i=1}^{n} [b(\theta_i + a_{ni}t_n) - b(\theta_i) - \mu_i a_{ni} t] - q_n t_n = -\frac{1}{2} q_n^2 \{1 + o(1)\}.$$

Hence,

$$P(\sum_{i=1}^{n} a_{ni}(Y_i - \mu_i) > q_n) \le \exp\left[-\frac{1}{2} q_n^2 \{1 + o(1)\}\right] = o(n^{-m}).$$

This completes the proof.

We now state and prove the uniform consistency of $\hat{\boldsymbol{\beta}}(s)$ for $s \in A_0$.

Theorem 1. *Under conditions A1-A6, with probability tending to 1 as $n \to \infty$,*

$$||\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)|| = O(n^{-1/3}),$$

uniformly for $s \in A_0$.

Proof: For any unit vector $\boldsymbol{u}$, let $\boldsymbol{\beta}(s) = \boldsymbol{\beta}_0(s)+n^{-1/3}\boldsymbol{u}$. Clearly, when n is sufficiently large, this $\boldsymbol{\beta}(s)$ falls into the neighborhood of $\boldsymbol{\beta}_0(s)$ so that conditions A4 and A5 become applicable. Thus, for all $s \in A_0$,

$$\begin{aligned} l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s)) &= n^{-1/3}\boldsymbol{u}^\tau \boldsymbol{s}_n(\boldsymbol{\beta}_0(s)) - \frac{1}{2}n^{1/3}\boldsymbol{u}^\tau\{n^{-1}H_n(\tilde{\boldsymbol{\beta}}(s))\}\boldsymbol{u} \\ &\leq n^{-1/3}\boldsymbol{u}^\tau \boldsymbol{s}_n(\boldsymbol{\beta}_0(s)) - c_1(1-\epsilon)n^{1/3}. \end{aligned}$$

Hence, for some generic positive constant c,

$$\begin{aligned} & P\{l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s)) > 0: \text{ for some } \boldsymbol{u}\} \\ \leq\ & P\{\boldsymbol{u}^\tau \boldsymbol{s}_n(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}; \text{ for some } \boldsymbol{u}\} \\ \leq\ & P\{\boldsymbol{s}_n^\tau(\boldsymbol{\beta}_0(s))\boldsymbol{s}_n(\boldsymbol{\beta}_0(s)) \geq cn^{4/3}\} \\ \leq\ & \sum_{j\in s} P(s_{nj}^2(\boldsymbol{\beta}_0(s)) \geq cn^{4/3}) \\ =\ & \sum_{j\in s} P(s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}) + \sum_{j\in s} P(-s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}). \end{aligned}$$

Note that

$$s_{nj}(\boldsymbol{\beta}_0(s)) = \sum_{i=1}^n [Y_i - b'(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0(s))]x_{ij} = \sum_{i=1}^n (Y_i - \mu_i)x_{ij},$$

and that condition A4 implies $\sum_{i=1}^n x_{ij}^2\sigma_i^2 = O(n)$, uniformly for all j. Thus we have

$$[\sum_{i=1}^n x_{ij}^2\sigma_i^2]^{1/2}q_n = O((n\log n)^{1/2}).$$

Let $a_{ni} = x_{ij}/[\sum_{i=1}^n x_{ij}^2\sigma_i^2]^{1/2}$. Applying Lemma 1, we obtain

$$\begin{aligned} P(s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}) &\leq P(s_{nj}(\boldsymbol{\beta}_0(s)) \geq c[\sum_{i=1}^n x_{ij}^2\sigma_i^2]^{1/2}q_n) \\ &= P(\sum_{i=1}^n a_{ni}(Y_i - \mu_i) > q_n) \\ &= o(n^{-m}), \end{aligned}$$

for any $m > 0$ uniformly for all $s \in A_0$. By choosing $m > \kappa K$, we have $O(n^{-m}) = o(P^{-K})$, and hence, by the Bonferoni inequality,

$$\sum_{s\in A_0}\sum_{j\in s} P[s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}] = o(1).$$

By replacing $Y_i - \mu_i$ with $-(Y_i - \mu_i)$ in the above argument, we also have

$$\sum_{s\in A_0}\sum_{j\in s} P[-s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}] = o(1).$$

Because $l_n(\boldsymbol{\beta}(s))$ is a concave function for any s, the above result implies that with probability tending to 1 as $n \to \infty$, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}(s)$ exists and falls within a $n^{-1/3}$-neighborhood of $\boldsymbol{\beta}_0(s)$ uniformly for $s \in A_0$. The theorem is proved.

3 The EBIC and its consistency under small-n-large-P sparse GLM

In the small-n-large-P setting, the traditional Bayes information criterion (BIC) is inappropriate for feature selection. It tends to select too many features and not just the causal features. BIC was obtained from the Bayesian paradigm which selects models based on their posterior probabilities. The prior probabilities behind BIC are uniform over all the possible models. As a consequence, BIC tends to select models with a larger number of features, see Chen and Chen (2008). When P is large, this tendency becomes prominent and disables BIC for feature selection. Other model selection criteria, such as AIC and CV, essentially select models based on their prediction accuracy. They do not address the issue of feature selection and are unsuitable in the small-n-large-P situation.

Chen and Chen (2008) have recently proposed a family of extended Bayes information criteria (EBIC) by considering non-constant priors. In EBIC, models are classified according to the number of features they contain, and the prior probability assigned to a model is somehow inversely proportional to the size of the model class to which the model belongs. EBIC for a model consisting of the features in a subset s is defined as

$$\mathrm{EBIC}(s) = -2l_n(\hat{\boldsymbol{\beta}}(s)) + \nu(s)\log n + 2\nu(s)\gamma\log P,\ \ \gamma \geq 0.$$

The consistency of EBIC has been proved under normal linear models when $P = O(n^\kappa)$ and $\gamma > 1-1/(2\kappa)$. The false discovery rate (FDR) (Benjamini and Hochberg, 1995) is the proportion of falsely selected features among all the selected features, and the positive selection rate (PSR) is the proportion of selected causal features among all the causal features. Simulation results indicate that EBIC with γ in the above consistency range can effectively keep the FDR low while achieving a reasonable PSR.

In this section, we establish the consistency of EBIC under generalized linear models with canonical links. First we present a lemma that is of interest in itself and will be used in the proof of consistency.

Lemma 2. *Let $\boldsymbol{z} = (z_1, \ldots, z_K)^\tau$ be any K-dimensional vector. For any $\delta > 0$ and $C > 0$, there exists a finite set of unit vectors $\mathcal{U}$ independent of C such that*

$$\{\boldsymbol{z} : \boldsymbol{z}^\tau\boldsymbol{z} \geq C\} \subset \cup_{\boldsymbol{u}\in\mathcal{U}}\{\boldsymbol{z} : \boldsymbol{u}^\tau\boldsymbol{z} \geq (1-\delta)\sqrt{C}\}.$$

Proof: We first note that the space of unit vector $\mathbf{U}$ of dimension K is compact. For each $\boldsymbol{u} \in \mathbf{U}$, and any positive constant v, define

$$\boldsymbol{Z}(\boldsymbol{u}, v) = \{\boldsymbol{z} : \boldsymbol{u}^{\tau}\boldsymbol{z} \geq \max(v\sqrt{C}, \|\boldsymbol{z}\|/\sqrt{2})\}.$$

The above definition specifies a set of $\boldsymbol{u}$ such that the angle θ between $\boldsymbol{z}$ and $\boldsymbol{u}$ is at most $\pi/4$, and their inner product is at least $v\sqrt{C}$. These two properties also imply that the norm of any vector in $\boldsymbol{Z}(\boldsymbol{u}, v)$ is at least $v\sqrt{C}$, and the angle restriction implies that $\cos(\theta) \geq \sin(\theta) \geq 0$.

It is easily seen that

$$\{\boldsymbol{z} : \boldsymbol{z}^{\tau}\boldsymbol{z} \geq C\} = \cup_{\boldsymbol{u}\in\mathbf{U}}\boldsymbol{Z}(\boldsymbol{u}, v = 1).$$

We aim to show that $\mathbf{U}$ can be replaced by a set containing only a finite number of unit vectors when $v = 1$ is replaced with $v = 1 - \delta$. We first show that for any sufficiently small $\delta > 0$ and unit vector $\boldsymbol{u}_0$ such that

$$\boldsymbol{u}^{\tau}\boldsymbol{u}_0 \geq \sqrt{1 - \delta^2/4}, \tag{2}$$

we have

$$\boldsymbol{Z}(\boldsymbol{u}, v = 1) \subset \boldsymbol{Z}(\boldsymbol{u}_0, v = 1 - \delta). \tag{3}$$

Let α be the angle between $\boldsymbol{u}$ and $\boldsymbol{u}_0$. When (2) is satisfied, we have

$$\cos(\alpha) \geq 1 - \delta/2 \quad \text{and} \quad \sin(\alpha) \leq \delta/2. \tag{4}$$

Consequently, for any $\boldsymbol{z} \in \boldsymbol{Z}(\boldsymbol{u}, v = 1)$, let θ_0 be the angle between $\boldsymbol{u}_0$ and $\boldsymbol{z}$, and θ

the angle between $\boldsymbol{u}$ and $\boldsymbol{z}$, then $\theta_0 \le \theta + \alpha \le \pi/2$. Therefore,

$$\begin{aligned}
\boldsymbol{u}_0^\tau \boldsymbol{z} &= \|\boldsymbol{z}\| \cos(\theta_0) \ge \|\boldsymbol{z}\| \cos(\theta + \alpha) \\
&= \|\boldsymbol{z}\| [\cos(\theta)\cos(\alpha) - \sin(\theta)\sin(\alpha)] \\
&\ge \|\boldsymbol{z}\| [\cos(\theta)\cos(\alpha) - \cos(\theta)\sin(\alpha)] \\
&= \|\boldsymbol{z}\| \cos(\theta)[\cos(\alpha) - \sin(\alpha)] \\
&\ge \sqrt{C}[\cos(\alpha) - \sin(\alpha)] \\
&\ge (1-\delta)\sqrt{C}.
\end{aligned}$$

where the second-last inequality follows from the definition of $\boldsymbol{Z}(\boldsymbol{u}, 1)$ and the last inequality follows from (4). This implies that $\boldsymbol{z} \in \boldsymbol{Z}(\boldsymbol{u}_0, 1-\delta)$. Thus, for any sufficiently small $\delta > 0$ and any $\boldsymbol{u} \in \mathbf{U}$, there is an open neighborhood of $\boldsymbol{u}$, $A_{\boldsymbol{u}} \subset \mathbf{U}$, such that

$$\boldsymbol{Z}(\boldsymbol{u}; 1-\delta) \supset \cup_{\tilde{\boldsymbol{u}} \in A_{\boldsymbol{u}}} \boldsymbol{Z}(\tilde{\boldsymbol{u}}, 1).$$

Clearly,

$$\cup_{\boldsymbol{u} \in \mathbf{U}} A_{\boldsymbol{u}} \supset \mathbf{U}$$

By the compactness of $\mathbf{U}$, there exists a finite set $\mathcal{U}$ such that

$$\cup_{\boldsymbol{u} \in \mathcal{U}} A_{\boldsymbol{u}} \supset \mathbf{U}$$

and hence

$$\cup_{\boldsymbol{u} \in \mathcal{U}} \boldsymbol{Z}(\boldsymbol{u}; 1-\delta) \supset \cup_{\boldsymbol{u} \in \mathcal{U}} \cup_{\tilde{\boldsymbol{u}} \in A_{\boldsymbol{u}}} \boldsymbol{Z}(\tilde{\boldsymbol{u}}; 1) = \cup_{\boldsymbol{u} \in \mathbf{U}} \boldsymbol{Z}(\mathbf{u}, 1) = \{\boldsymbol{z} : \boldsymbol{z}^\tau \boldsymbol{z} \ge C\}.$$

Note that the finite set $\mathcal{U}$ does not depend on C. This completes the proof.

We now tackle the issue of the consistency of EBIC. The consistency follows from the following theorem.

Theorem 2. *Under conditions A1-A6, as $n \to \infty$, we have*

$$P\{\min_{s \in A_1} EBIC(s) \le EBIC(s_0)\} \to 0, \tag{5}$$

for any $\gamma > 0$;

$$P\{\min_{s \in A_0, s \neq s_0} EBIC(s) \le EBIC(s_0)\} \to 0, \tag{6}$$

for $\gamma > 1 - \frac{1}{2\kappa}$.

Proof of (5) : Note that for any s, EBIC(s) $\le$ EBIC(s_0) implies that

$$l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_0)) \ge -2K(1+\gamma)\log n. \tag{7}$$

In turn, (7) implies that

$$l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\boldsymbol{\beta}_0(s_0)) \ge -2K(1+\gamma)\log n. \tag{8}$$

Consequently, it suffices to show that the probability of (8) occurring at any $s \in A_1$ goes to 0.

For any $s \in A_1$, let $\tilde{s} = s \cup s_0$. Now consider those $\boldsymbol{\beta}(\tilde{s})$ near $\boldsymbol{\beta}_0(\tilde{s})$. We have

$$\begin{aligned} & l_n(\boldsymbol{\beta}(\tilde{s})) - l_n(\boldsymbol{\beta}_0(\tilde{s})) \\ = \ & [\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})]^\tau \boldsymbol{s}_n(\boldsymbol{\beta}_0(\tilde{s})) - \frac{1}{2}[\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})]^\tau H_n(\boldsymbol{\beta}^*(\tilde{s}))[\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})] \end{aligned}$$

for some $\boldsymbol{\beta}^*(\tilde{s})$ between $\boldsymbol{\beta}(\tilde{s})$ and $\boldsymbol{\beta}_0(\tilde{s})$. By conditions A4 and A5,

$$[\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})]^\tau H_n(\boldsymbol{\beta}^*(\tilde{s}))[\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})] \ge c_1 n(1-\epsilon)\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|^2.$$

Therefore,

$$l_n(\boldsymbol{\beta}(\tilde{s})) - l_n(\boldsymbol{\beta}_0(\tilde{s})) \le [\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})]^\tau \boldsymbol{s}_n(\boldsymbol{\beta}_0(\tilde{s})) - \frac{c_1}{2} n(1-\epsilon)\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|^2.$$

Hence, for any $\boldsymbol{\beta}(\tilde{s})$ such that $\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\| = n^{-1/4}$, we have

$$l_n(\boldsymbol{\beta}(\tilde{s})) - l_n(\boldsymbol{\beta}_0(\tilde{s})) \le n^{-1/4}\|\boldsymbol{s}_n(\boldsymbol{\beta}_0(\tilde{s}))\| - \frac{c_1}{2} n^{1/2}(1-\epsilon).$$

Lemma 1 implies that $\max_{s\in A_1}\|\boldsymbol{s}_n(\boldsymbol{\beta}_0(\tilde{s}))\| = O_p(n^{1/2}\log n)$ under conditions A4 and A5. Therefore,

$$l_n(\boldsymbol{\beta}(\tilde{s})) - l_n(\boldsymbol{\beta}_0(\tilde{s})) \leq c(n^{1/4}\log n - n^{1/2}) \leq -cn^{1/2}$$

uniformly over $\tilde{s}$ and $\boldsymbol{\beta}(\tilde{s})$ such that $\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\| = n^{-1/4}$ for a generic positive constant c.

Because $l_n(\boldsymbol{\beta}(\tilde{s}))$ is concave in $\boldsymbol{\beta}(\tilde{s})$, the above result implies that the maximum of $l_n(\boldsymbol{\beta}(\tilde{s}))$ is attained inside $\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\| \leq n^{-1/4}$. The concavity also implies that

$$\begin{aligned}&\sup\{l_n(\boldsymbol{\beta}(\tilde{s})) - l_n(\boldsymbol{\beta}_0(\tilde{s})) : \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\| \geq n^{-1/4}\}\\ &\qquad\leq \sup\{l_n(\boldsymbol{\beta}(\tilde{s})) - l_n(\boldsymbol{\beta}_0(\tilde{s})) : \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\| = n^{-1/4}\} \leq -cn^{1/2} \qquad (9)\end{aligned}$$

uniformly over $s \in A_1$.

Now let $\breve{\boldsymbol{\beta}}(\tilde{s})$ be $\hat{\boldsymbol{\beta}}(s)$ augmented with zeros corresponding to the elements in $\tilde{s}/s$. It can be seen that

$$\|\breve{\boldsymbol{\beta}}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\| \geq \|\boldsymbol{\beta}_0(s_0/s)\| > n^{-1/4},$$

because $\|\boldsymbol{\beta}_0(s_0/s)\|$ has a positive lower bound depending on neither n nor s. Therefore,

$$l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\boldsymbol{\beta}_0(s_0)) = l_n(\breve{\boldsymbol{\beta}}(\tilde{s})) - l_n(\boldsymbol{\beta}_0(\tilde{s})) \leq -cn^{1/2}.$$

This implies that the probability that (8) holds goes to zero and hence (5) is proved. *Proof of (6)*: For $s \in A_0$, let $m = \nu(s) - \nu(s_0)$. It suffices to consider a fixed m since m takes only the values $1, 2, \ldots, K - \nu(s_0)$. By definition, EBIC(s) $\leq$ EBIC(s_0) if and only if

$$l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_0)) \geq m[0.5\log n + \gamma\log P].$$

Since $P = O(n^{-\kappa})$, the number of models in A_0 with $\nu(s) - \nu(s_0) = m$ does not exceed $P^m = O(n^{\kappa m})$. By the Bonferoni inequality, it suffices to show that

$$P(l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_0)) \geq m[0.5 \log n + \gamma \log P]) = o(n^{-\kappa m}) \tag{10}$$

uniformly for $s \in A_0$.

By condition A4, when n is sufficiently large, we have

$$\begin{aligned} & l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_0)) \\ \leq\ & l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\boldsymbol{\beta}_0(s)) \\ \leq\ & [\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)]^\tau \boldsymbol{s}_n(\boldsymbol{\beta}_0(s)) - \frac{1-\epsilon}{2}[\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)]^\tau H_n(\boldsymbol{\beta}_0(s))[\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)] \\ \leq\ & \frac{1}{2(1-\epsilon)} \boldsymbol{s}_n(\boldsymbol{\beta}_0(s))^\tau [H_n(\boldsymbol{\beta}_0(s))]^{-1} \boldsymbol{s}_n(\boldsymbol{\beta}_0(s)) = \frac{1}{2}(1+\epsilon) \boldsymbol{z}_n^\tau \boldsymbol{z}_n \end{aligned}$$

with $\boldsymbol{z}_n = [H_n(\boldsymbol{\beta}_0(s))]^{-1/2} \boldsymbol{s}_n(\boldsymbol{\beta}_0(s))$. Since $\epsilon > 0$ can be made arbitrarily small, we will regard it as 0 in the following argument. This does not invalidate the proof, but saves the trouble of tedious notation.

By Lemma 2, we have

$$P(\boldsymbol{z}_n^\tau \boldsymbol{z}_n \geq 2m\kappa \log n) \leq \sum_{\boldsymbol{u} \in \mathcal{U}} P(\boldsymbol{u}^\tau \boldsymbol{z} \geq \sqrt{2m\kappa \log n}),$$

where $\mathcal{U}$ is a finite set of unit vectors independent of n.

Note that $\boldsymbol{u}^\tau \boldsymbol{z}_n$ is a linear combination of y_i, $i = 1, \ldots, n$. We can write $\boldsymbol{u}^\tau \boldsymbol{z}_n = \sum_{i=1}^n a_{ni} y_i$. Since $H_n(\boldsymbol{\beta}_0(s))$ is the variance matrix of $\boldsymbol{s}_n(\boldsymbol{\beta}_0(s))$ and hence the variance matrix of $\boldsymbol{z}_n$ is the identity matrix, we have $\mathrm{Var}(\boldsymbol{u}^\tau \boldsymbol{z}_n) = \sum_{i=1}^n a_{ni}^2 \sigma_i^2 = 1$. Then by Lemma 1, we obtain

$$P(\boldsymbol{z}_n^\tau \boldsymbol{z}_n \geq 2m\kappa \log n) \leq o(n^{-\kappa m}).$$

Thus we have

$$P(l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_0)) \geq m\kappa \log n) \leq P(\boldsymbol{z}_n^\tau \boldsymbol{z}_n \geq 2m\kappa \log n) \leq o(n^{-\kappa m}). \tag{11}$$

When $\gamma > 1 - 1/(2\kappa)$, we have $\kappa < 0.5 + \gamma\kappa$. Hence, (11) implies (10) and the proof is complete.

4 Simulations

In this section, we present our simulation studies for assessing the performance of EBIC under GLM. The simulation studies are conducted in the framework of a case-control study with an equal number of cases and controls. The disease status y, the response variable, takes value 1 for cases and 0 for controls. The features x_{ij} under study are generated as single nucleotide polymorphisms (SNPs) in the human genome. Let s_0 be the index set of SNPs that are causally associated with the disease status. For $j \notin s_0$, the x_{ij} values are generated under the assumption of Hardy-Weinberg equilibrium; that is, they are simulated from a binomial distribution with parameters $(2, p_j)$ where p_j is the allele frequency of one of the alleles for the jth SNP. The allele frequency p_j is not fixed but is generated from a Beta distribution with parameters $(\alpha = 2, \beta = 2)$, independently for each j and in each simulation run. This choice was made after some simple computer experiments. The outcomes of generated p_j concentrate mainly around 0.5, but moderately spread out to 0.1 and 0.9. Given p_j, $x_{ij} : i = 1, \ldots, n = n_1 + n_2$ are independently generated. For $j \in s_0$, the x_{ij} values are generated such that

$$\mathrm{logit} P(Y = 1 | X(s_0) = \boldsymbol{x}(s_0)) = \alpha + \boldsymbol{x}^{\tau}(s_0)\boldsymbol{\beta}_0$$

for a pre-specified s_0 and a specified $\boldsymbol{\beta}_0$. It can be seen that

$$P(X(s_0) = \boldsymbol{x}(s_0) | Y = 1) = P(X(s_0) = \boldsymbol{x}(s_0) | Y = 0) \exp(\alpha^* + \boldsymbol{x}^{\tau}(s_0)\boldsymbol{\beta}_0)$$

where α^* is the normalization factor. If $\nu(s_0) = m$ then there are 3^m possible $\boldsymbol{x}(s_0)$ vector-values. The corresponding conditional probability mass function of $X(s_0)$ is given as above. We hence sample from 3^m vectors of length m.

In the simulation studies, we set the number of both cases and controls to be $n = 500$. Because of the extensive computational effort required, we did not increase n, but instead used a number of different $\boldsymbol{\beta}_0$ vectors, which has the same effect on the detectability of the causal features. The choices of m, P, and $\boldsymbol{\beta}_0$ used in the simulation studies are given in the following table.

Model	m	P	$\boldsymbol{\beta}_0$
1	2	500	(.5, .7)
2	2	500	(.3, .5)
3	3	500	(.5, .6, .7)
4	3	500	(.3, .4, .5)
5	5	500	(.3, .4, .5, .6, .7)
6	8	500	(.2, .3, .4, .5, .6, .7, .8, .9)
7	2	1000	(.5, .7)
8	2	1000	(.3, .5)
9	3	1000	(.5, .6, .7)
10	3	1000	(.3, .4, .5)
11	5	1000	(.3, .4, .5, .6, .7)
12	8	1000	(.2, .3, .4, .5, .6, .7, .8, .9)
13	3	10000	(.6, .7, .8)
14	3	10000	(.4, .5, .6)
15	5	10000	(.3, .4, .5, .6, .7)
16	8	10000	(.2, .3, .4, .5, .6, .7, .8, .9)
17	3	100000	(1.3, 1.4, 1.5)
18	3	100000	(.5, .6, .7)

When $P = 500$, the `glmpath` function is directly applied to identify a sequence of ordered SNPs, denoted by s, of length no more than $K = 40$. We then use `glm.fit` to evaluate the submodels (at most 40 of them) formed by the first k variables in the sequence for $k = 1, 2, \ldots, K$. The BIC and EBIC values are computed and the submodels minimizing BIC or EBIC are selected.

When $P = 1000$, we randomly divide the SNPs into two groups and obtain s_j,

$j = 1, 2$ in the same way as when $P = 500$. We then pool the SNPs in s_j, and apply the `glmpath` function to the pooled set to obtain an ordered set s. The same strategy is used when $P = 10,000$ and 100,000 except that more rounds of the above procedure are applied.

The procedure of using several rounds of selection when P is very large was designed by Chen and Chen (2008) and called a tournament procedure. It has been found to be a competitive alternative to several computational strategies recently published in the literature, e.g., the sure independent screening considered by Fan and Lv (2008).

To shed light on the appropriate size of γ, we present the results obtained by taking $\gamma = 0, 0.25, .5, 1$. Note that the ordinary BIC is a special form of EBIC with $\gamma = 0$. A data-driven choice of γ would be preferred, but the issue is beyond the scope of this paper. We plan to discuss this issue more thoroughly in a separate work.

The number of simulation replicates is $N = 500$ for Models 1-16, and $N = 200$ for Models 17, 18. The simulation results in terms of average positive selection and false discovery rates (PSR and FDR) as well as the average number of selected SNPs are summarized in Table 1. The average PSR and FDR are defined as follows. Let s_0 be the set of causal features, and s_j^* the features selected in the jth replicate, $j = 1, 2, \ldots, N$. Then

$$\text{PSR} = \frac{\sum_{j=1}^{N} \nu(s_j^* \cap s_0)}{N\nu(s_0)}, \qquad \text{FDR} = \frac{\sum_{j=1}^{N} \nu(s_j^*/s_0)}{\sum_{j=1}^{N} \nu(s_j^*)}.$$

In Table 1, $\nu^* = (1/N)\sum_{j=1}^{N} \nu(s_j^*)$.

The simulation results confirm the inadequacy of BIC for feature selection when P is large. The FDR with BIC is high under all models and increases as P gets larger. On the other hand, EBIC with $\gamma = 1$ tightly controls the FDR in all cases except

Table 1: Simulation results for EBIC under GLM (FDR, PSR, ν^*)

Model	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 1.0$
1	(.704, .996, 6.73)	(.256, .985, 2.65)	(.056, .965, 2.04)	(.015, .918, 1.86)
2	(.690, .996, 6.44)	(.299, .985, 2.81)	(.069, .966, 2.07)	(.018, .922, 1.88)
3	(.590, .984, 7.21)	(.204, .954, 3.59)	(.044, .893, 2.80)	(.010, .800, 2.43)
4	(.653, .765, 6.60)	(.326, .570, 2.54)	(.142, .428, 1.50)	(.103, .360, 1.20)
5	(.503, .848, 8.53)	(.177, .729, 4.43)	(.046, .617, 3.23)	(.011, .520, 2.63)
6	(.383, .766, 9.93)	(.117, .686, 6.21)	(.028, .621, 5.11)	(.006, .571, 4.59)
7	(.807, .998, 10.3)	(.409, .992, 3.36)	(.098, .974, 2.16)	(.019, .933, 1.90)
8	(.834, .849, 10.2)	(.480, .740, 2.85)	(.159, .651, 1.55)	(.045, .574, 1.20)
9	(.734, .999, 11.3)	(.302, .994, 4.28)	(.064, .978, 3.14)	(.011, .953, 2.89)
10	(.751, .901, 10.9)	(.357, .777, 3.63)	(.099, .635, 2.12)	(.021, .496, 1.52)
11	(.636, .920, 12.6)	(.232, .817, 5.32)	(.052, .686, 3.61)	(.008, .558, 2.81)
12	(.540, .814, 14.2)	(.167, .686, 6.59)	(.035, .578, 4.79)	(.006, .490, 3.94)
13	(.918, .954, 35.1)	(.665, .905, 8.12)	(.200, .835, 3.13)	(.032, .753, 2.34)
14	(.922, .928, 35.6)	(.692, .858, 8.37)	(.208, .738, 2.79)	(.034, .612, 1.90)
15	(.888, .790, 35.3)	(.600, .713, 8.91)	(.146, .610, 3.57)	(.020, .523, 2.67)
16	(.822, .783, 35.3)	(.465, .713, 10.7)	(.083, .624, 5.45)	(.007, .548, 4.41)
17	(.983, .071, 13.1)	(.867, .069, 1.54)	(.834, .066, 1.20)	(.813, .066, 1.06)
18	(.989, .080, 20.8)	(.870, .080, 1.83)	(.813, .074, 1.18)	(.795, .072, 1.05)

under Models 17 and 18. At the same time, its PSR remains competitive with that of BIC. This is particularly important because the latter yields substantially smaller average model sizes $\nu*$.

In practical problems, one often needs to make a trade-off between PSR and FDR. If the FDR is of less concern, a value of γ less than 1 can be used. The simulation results indicate that $\gamma = 0.5$ is worth considering. It keeps the FDR at reasonably low levels but achieves a higher PSR than $\gamma = 1$. $\gamma = 0.25$ could also be an appropriate choice if P is not too large, say $P < 1000$. But BIC is not a good choice because of its high FDR and its liberal nature as indicated by the noticeably larger average model sizes in Table 1.

The results under Models 17 and 18 are worth a special remark. In these two models, the few causal features are submerged in a sea of non-causal features. It is hard for any method to identify the causal features unless the sample size is very large. As indicated by the simulation results, neither BIC nor EBIC is successful in the identification of the causal SNPs. This might be partially due to the screening procedure failing to retain the causal SNPs in the final group, or due to the ranking procedure failing to rank the SNPs in the final group in line with their magnitude of effect because of high spurious correlations. Nevertheless, EBIC with all three γ-values tightly controls the average number of selected features, and BIC does not. Hence, EBIC does not mislead us by suggesting a large number of false causal features as BIC does. When the signals are too weak to be detected, though EBIC does not help in identifying the causal features, it serves as a good control on the number of false discoveries.

We conducted simulations more extensive than presented here. For some models, we used both $K = 30, 40$ and found that the simulation results were not sensitive to the choice of K when $P \leq 1000$. When $P = 10,000$, the average number of features selected by BIC is larger when $K = 40$ than when $K = 30$, but the other results are not affected by the choice of K. Hence we have reported the results with $K = 40$ for all models.

5 An example

In Singh et al. (2002), the researchers measured $P = 6033$ genes on each of $n = 102$ men with $n_1 = 50$ controls and $n_2 = 52$ prostate cancer patients. The purpose of the study was to build a model for predicting the disease status of a man given the microarray measurement of the same 6033 genes of the man. Efron (2008) proposed an empirical Bayes approach. In a nutshell, this approach puts the discriminate power of each gene into a single t-type statistic, $\hat{W}_i$, and a linear combination $\hat{S}_\lambda = \sum \hat{\delta}_i \hat{W}_i$ is used for prediction. The Bayes component is introduced through a shrunken centroids algorithm (Tibshirani et al., 2002). The algorithm shrinks the $\hat{\delta}_i$ values by increasing the value of a tuning parameter (shrinkage value λ). The non-zero $\hat{\delta}_i$ values corresponding to a particular value of the tuning parameter are actually used for prediction. In particular, when the shrinkage value $\lambda = 2.16$, 377 genes are chosen, which achieves the lowest cross-validation error rate of 9%. When $\lambda = 4.32$, 4 genes are chosen with a cross-validation error rate of 41%.

In this section, we re-analyze the data by building a generalized linear model

$$\text{logit}\{P(Y = 1|\boldsymbol{x})\} = \boldsymbol{x}^{\tau}\boldsymbol{\beta}$$

where Y is the status of prostate cancer, and $\boldsymbol{x}$ is the vector of $P = 6033$ gene expression levels. The feature selection method with EBIC is used for the analysis.

We first examined the correlation structure of the gene expression data. We randomly selected each time a sample of 20 out of the 6033 genes and computed the eigenvalues of the matrix $\boldsymbol{x}^{\tau}(s)\boldsymbol{x}(s)$. We found that in only about 8% of these samples was the smallest eigenvalue of the above matrix below 10. Therefore, we are confident that the identifiability conditions A4 and A5 can be satisfied with $K = 20$. We chose $\gamma = 0.5$ for EBIC, because it offers a good trade-off between the FDR and the PSR as our simulation study suggested.

The tournament procedure as in the simulation study was used. The 6033 genes were divided at random into groups of about 50 and the R-function `glmpath` was applied to choose 15 genes from each group. These genes were then pooled and the process was repeated until a final group of about 50 was obtained. Then 15 genes were selected from the final group by `glmpath`. The tournament procedure was repeated

100 times, and 100 sets of 15 genes were produced. Eventually, the 50 genes that appeared most often in these 100 sets were selected, and they were subjected to `glmpath` selection once more. The genes are given in the first line of the following table in the order produced by `glmpath`. The generalized linear models including the first gene, the first two genes, and so on, were fitted and the deviances (given in the second row) and EBIC values (given in the third row) were computed. The delete-one cross-validation errors were also computed and are given in the fourth row.

Gene No.	610	1720	332	364	1068	914	3940	1077	4331	579
Deviance	113.6	94.2	80.0	74.8	64.3	58.0	50.0	31.9	25.1	21.3
$\text{EBIC}_{0.5}$	113.6	107.6	106.7	114.8	117.6	124.6	130.0	125.2	131.7	141.2
CV-error	27.5	19.6	16.7	14.7	14.7	8.8	11.8	7.8	9.8	10.8

Using EBIC, the first three genes were selected. When these genes were used for classification, the cross-validation error rate was 16.7%. In comparison, the Bayes method chooses around 80 genes to attain a similar cross-validation error rate, see Efron (2008). If cross-validation was used as the criterion, an eight-gene model was selected with a cross-validation error rate of 7.8%. The delete-one cross-validation is widely known to be too liberal. EBIC selects a more parsimonious model but retains a low cross-validation error rate.

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.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second Int. Symp. Info. Theory*, B. N. Petrox and F. Caski, eds., 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.

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.

Chen, J. & Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika*, **94**, 759-771.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D., & Todd, J. A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, **30**, 1243-1246.

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.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century.

Efron, B. (2008). Empirical Bayes estimates for large-scale prediction problems. Manuscript.

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

Fan, J. & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians*, M. Sanz-Sole, J. Soria, J. L. Varona, J. Verdera, eds. Vol. III, 595-622.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. Roy. Statist. Soc.* B **70**, 849-911.

Hoh, J., Wille, A., & Ott, J. (2008). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research* **11**, 2115-9.

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

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models.* 2nd ed., Chapman and Hall.

Park, M. Y. & Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. R. Statist. Soc.* B **69**, 659-77.

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

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

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kanto, P. W., Golub, T. R., & Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-9.

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

The Wellcome Trust Case-Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678.

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

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567-6572.

Wang, H., Li, R., & Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Zhang, C. H. & Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-94.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B **67**, 301-320.