Statistics and Its Interface Volume 6 (2013) 275–284

Selection consistency of EBIC for GLIM with non-canonical links and diverging number of parameters[*]

Shan Luo and Zehua Chen[†]

In this article, we investigate the properties of the EBIC in variable selection for generalized linear models with non-canonical links and a diverging number of parameters in ultra-high dimensional feature space. The selection consistency of the EBIC in this situation is established under moderate conditions. The finite sample performance of the EBIC coupled with a forward selection procedure is demonstrated through simulation studies and a real data analysis.

AMS 2000 subject classifications: Primary 62F99; secondary 62J12.

Keywords and phrases: Extended BIC, Generalized linear model, Non-canonical links, Selection consistency, Variable selection.

1. INTRODUCTION

Variable selection is a primary concern in many important contemporary scientific fields such as signal processing, medical research and genetic studies etc. In these fields, usually, a relatively small set of relevant variables need to be selected from a huge collection of available variables. For example, in genetic genome-wide association studies (GWAS), to identify loci or genes that affect a quantitative trait or a disease status, thousands of thousands, even millions, of single nucleotide polymorphisms (SNP) are under consideration. The number of variables is much larger than the sample size in such studies. This phenomenon is referred to as small-n-large-p. Variable selection in small-n-large-p problems poses a great challenge.

A major approach for variable selection is model based; that is, a model is formulated to describe the relationship between a response variable (e.g., the measurement of a quantitative trait) and a set of predictor variables or covariates (e.g., the genotypes of SNPs), and the covariates are selected by a certain variable selection criterion. A variable selection criterion is crucial in model based variable selection. Traditional variable selection criteria such as Akaike's Information Criterion (AIC) (Akaike, 1973), Bayes Information Criterion (BIC) (Schwarz, 1978) and Cross Validation (CV) (Stone, 1974) are no longer appropriate for variable selection in small-n-large-p problems. These traditional criteria tend to select too many irrelevant covariates because they are generally not selection consistent. Recently, some BIC-type criteria have been proposed for small-n-large-p problems. Bogdan et al. (2004) considered a criterion called modified BIC (mBIC) for QTL mapping models. Wang et al. (2009) studied another modified BIC for models with a diverging number of parameters. Chen and Chen (2008) extended the original BIC to a family called extended BIC (EBIC) governed by a parameter γ.

The criterion considered by Wang et al. (2009) modifies the original BIC by multiplying the second term of BIC with a diverging parameter and is somehow ad hoc. To achieve selection consistency, it requires $p/n^{\xi} < 1$ for some $0 < \xi < 1$, and hence is not applicable when $p > n$. The mBIC and EBIC considered by Bogdan et al. (2004) and Chen and Chen (2008) respectively are developed from a Bayesian framework. For the mBIC, a binomial prior on the number of covariates is imposed on each model. For EBIC, the prior on a model is proportional to a power of the size of model class which the model belongs. Asymptotically, mBIC is a special case of EBIC corresponding to $\gamma = 1$. The selection consistency of EBIC for linear models with a fixed number of parameters is established in Chen and Chen (2008). The result is then extended to generalized linear models (GLIM) with canonical links in Chen and Chen (2012). The EBIC has been used for choosing tuning parameters in penalized likelihood approaches, see Huang et al. (2010), for feature selection procedures, see Wang (2009) and Luo and Chen (2011), and for QTL mapping and disease gene mapping studies, see Li and Chen (2009) and Zhao and Chen (2012).

In GLIMs, canonical links do not always provide the best fit. Generally, there is no reason apriori why a canonical link should be used, and in many cases a non-canonical link is preferable, see McCullagh and Nelder (1989) and Czado and Munk (2000). In many conventional scientific fields such as those mentioned at the beginning of this article, it becomes a norm that the number of covariates under consideration is

[*] The authors would like to thank Professor Peter Song for the invitation of the submission of this paper to *Statitstics and Its Interface*. They would also like to express their appreciation to the two anonymous referees for their valuable comments and suggestions which made the improvement of the original draft possible.

[†] Corresponding author.

so large that it can be considered as having an exponential order of the sample size. This is referred to as the case of ultra-high dimensional feature space. In problems such as QTL and disease gene mapping, a quantitative trait or disease status is usually affected by many loci. Except a few so-called major genes, most of the loci have only a small effect which cannot be detected when the sample size is small. As the sample size increases, so does the number of detectable such effects. This phenomenon is mathematically well modeled by diverging number of parameters, i.e., the number of truly relevant covariates diverges as the sample size increases. Therefore the GLIMs with non-canonical links and diverging number of parameters in the case of ultra-high dimensional feature space become appealing. In this article, we investigate the properties of EBIC for such models and establish its selection consistency. The selection consistency of EBIC for GLIMs with canonical links does not trivially pass to the case of non-canonical links. The selection consistency in the case of non-canonical links is established under more general conditions than those in Chen and Chen (2012). The conditions, though general, are naturally satisfied by many popular examples as given in Wedderburn (1976). We also present a forward selection procedure with EBIC for the GLIMs. This procedure is applied in simulation studies and a real data analysis to evaluate its validity.

The remainder of this article is organized as follows. In section 2, the main results are presented and discussed. In section 3, simulation studies are reported and analyzed. In section 4, the forward selection procedure with EBIC is applied to analyze a well known Leukemia data set published in Golub et al. (1999). All the technical proofs are provided in the Appendix.

2. SELECTION CONSISTENCY OF EBIC

Let $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, be the observations where y_i is a response variable and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip_n})^\tau$ is a p_n-vector of covariates. We consider the generalized linear model (GLIM) below:

$$y_i \sim f(y_i; \theta_i) = \exp\{\theta_i y_i - b(\theta_i)\} \text{ w.r.t. } \nu, \quad i = 1, \ldots, n,$$

where ν is a σ-finite measure. From the properties of exponential family, we have

$$\mu(\theta_i) = E(y_i) = b'(\theta_i), \qquad \sigma^2(\theta_i) = \text{Var}(y_i) = b''(\theta_i),$$

where b' and b'' are the first and the second derivatives of b respectively. The θ_i is related to $\boldsymbol{x}_i$ through the relationship:

$$g(\mu(\theta_i)) = \eta_i = \boldsymbol{x}_i^\tau \boldsymbol{\beta},$$

where g is a monotone function called link function and $\boldsymbol{\beta}$ is p_n-dimensional parameter vector. If $g(\mu(\theta_i)) = \theta_i$, i.e., $g = \mu^{-1}$, the link is called the canonical link. In this article, we consider general link functions including the canonical link. Because of the one-to-one correspondence between θ_i and η_i, there is a function h such that $\theta_i = h(\eta_i) = h(\boldsymbol{x}_i^\tau \boldsymbol{\beta})$. Thus the probability density function of y_i can be expressed as

$$f(y_i; h(\boldsymbol{x}_i^\tau \boldsymbol{\beta})) = \exp\{y_i h(\boldsymbol{x}_i^\tau \boldsymbol{\beta}) - b(h(\boldsymbol{x}_i^\tau \boldsymbol{\beta}))\}.$$

In the above GLIM, we assume that $p_n = O(\exp\{n^\kappa\})$, $0 < \kappa < 1$, and that only a relatively small number of components of $\boldsymbol{\beta}$ are nonzero. Throughout the article, the following notation and convention are used. Denote by s any subset of the index set $\mathcal{S} = \{1, 2, \ldots, p_n\}$ and $|s|$ its cardinality. For convenience, s is used exchangeably to denote both an index set and the set of covariates with indices in the index set, and is also referred to as a model, i.e., the GLIM consisting of the covariates in s. Let $s_{0n} = \{j : \beta_j \neq 0, j = 1, \ldots, p_n\}$ and $p_{0n} = |s_{0n}|$. The covariates belonging to s_{0n} are called relevant features and the others irrelevant features. s_{0n} is also referred to as the true model. Let $X = (\boldsymbol{x}_1^\tau, \ldots, \boldsymbol{x}_n^\tau)^\tau$. Denote by $X(s)$ the sub matrix formed by the columns of X whose indices fall into s. Let $\boldsymbol{x}_i(s)$ be the vector consisting of the components of $\boldsymbol{x}_i$ whose indices belong to s, and let $\boldsymbol{\beta}(s)$ be the corresponding sub vector of $\boldsymbol{\beta}$. Let S_j denote the set of $\binom{p_n}{j}$ combinations of j indices from $\mathcal{S}$. Denote $\tau(S_j) = \binom{p_n}{j}$.

The EBIC of a model s, as defined in Chen and Chen (2008), is

$$\text{EBIC}_\gamma(s) = -2 \ln L_n\big(\hat{\boldsymbol{\beta}}(s)\big) + |s| \ln n + 2\gamma \ln \tau(S_{|s|}), \quad \gamma \geq 0,$$

where $L_n(\hat{\boldsymbol{\beta}}(s))$ is the maximum likelihood of model s and $\hat{\boldsymbol{\beta}}(s)$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}(s)$.

Denote by $l_n(\boldsymbol{\beta}(s)), \boldsymbol{s}_n(\boldsymbol{\beta}(s))$ and $H_n(\boldsymbol{\beta}(s))$ the log likelihood function, the score vector and the Hessian matrix of the model s respectively. Suppose that b and g are thrice and twice differentiable respectively, which is usually the case in practical GLIMs, then h is twice differentiable. Thus, we have

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

$$\begin{aligned}
&H_n(\boldsymbol{\beta}(s)) \\
&\quad = -\frac{\partial^2 l_n(\boldsymbol{\beta}(s))}{\partial \boldsymbol{\beta}(s) \partial \boldsymbol{\beta}^{s\tau}} \\
&\quad = \sum_{i=1}^n \{b^{''}(h(\boldsymbol{x}_i^\tau(s)\boldsymbol{\beta}(s)))[h^{'}(\boldsymbol{x}_i^\tau(s)\boldsymbol{\beta}(s))]^2 \\
&\qquad - [y_i - b^{'}(h(\boldsymbol{x}_i^\tau(s)\boldsymbol{\beta}(s)))] h^{''}(\boldsymbol{x}_i^\tau(s)\boldsymbol{\beta}(s))\} \boldsymbol{x}_i(s) \boldsymbol{x}_i^\tau(s) \\
&\quad = H_{n1}(\boldsymbol{\beta}(s)) - H_{n0}(\boldsymbol{\beta}(s)), \quad \text{say.}
\end{aligned}$$

When $s_{0n} \subset s$, we simply denote $\mu_i = b'(h(\boldsymbol{x}_i^\tau(s)\boldsymbol{\beta}(s)))$ and $\sigma_i^2 = b''(h(\boldsymbol{x}_i^\tau(s)\boldsymbol{\beta}(s)))$. The major difference between the case of canonical links and the case of non-canonical links is as follows. If g is the canonical link, $h' \equiv 1$ and $h'' \equiv 0$, hence $H_{n0} \equiv 0$ and $H_n(\boldsymbol{\beta}(s))$ is positive definite when $X(s)$ is of full column rank. Therefore, $l_n(\boldsymbol{\beta}(s))$ is a strictly concave function of $\boldsymbol{\beta}(s)$ and the MLE exists in the interior of the parameter space. But, if g is a non-canonical link, $H_n(\boldsymbol{\beta}(s))$ is not necessarily positive definite. As a consequence, $l_n(\boldsymbol{\beta}(s))$ is not necessarily concave, and the maximum likelihood estimate of $\boldsymbol{\beta}(s)$ is not necessarily in the interior of the parameter space if it exists at all. We will show that $H_{n0}(\boldsymbol{\beta}(s))$ is asymptotically negligible (Lemma 2.1) for $\boldsymbol{\beta}(s)$ in a neighborhood of the true parameter value of the GLIM. Thus $H_n(\boldsymbol{\beta}(s))$ is asymptotically locally positive definite. To guarantee the existence of the MLE of $\boldsymbol{\beta}(s)$ for finite samples, we assume that the link function g is chosen such that $l_n(\boldsymbol{\beta}(s))$ is concave. For example, for Poisson distribution, we can choose $g(\mu)$ as $\mu, \ln\mu$ or $\mu^r (0 < r < 1)$; for the binomial distribution, we can choose $g(\mu)$ as $\ln(\mu/(1-\mu)), \Phi^{-1}(\mu)$ or $\ln(-\ln(1-\mu))$. For details and more examples, the reader is referred to Wedderburn (1976). We now state the conditions required for the selection consistency of the EBIC.

C1 $\ln(p_n) = O(n^\kappa), p_{0n} = O(n^b)$ where $0 \le b < 1/6, \kappa > 0$;

C2 $\min_{j\in s_{0n}} |\beta_j| \ge Cn^{-1/4}$ for some constant $C > 0$;

C3 For any s, the interior of

$$\mathcal{B}(s) = \left\{\boldsymbol{\beta} : \int \exp(h(\boldsymbol{x}_i^\tau(s)\boldsymbol{\beta})y)d\nu < \infty, i = 1, 2, \ldots, n\right\}$$

is not empty. Let $\boldsymbol{\beta}_0$ denote the true parameter of the GLIM. If $|s| \le kp_{0n}$, where $k > 1$, then $\boldsymbol{\beta}_0(s)$ is in the interior of $\mathcal{B}(s)$;

C4 There exist positive c_1 and c_2 such that for all sufficiently large n,

$$\begin{aligned} c_1 n \le \lambda_{\min}(H_{n1}(\boldsymbol{\beta}_0(s\cup s_{0n}))) \\ \le \lambda_{\max}(H_{n1}(\boldsymbol{\beta}_0(s\cup s_{0n}))) \le c_2 n \end{aligned}$$

for all s with $|s| \le kp_{0n}$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote respectively the smallest and largest eigenvalues;

C5 For any given $\xi > 0$, there exists a $\delta > 0$ such that when n is sufficiently large, for $j = 0, 1$,

$$\begin{aligned} (1-\xi)H_{nj}(\boldsymbol{\beta}_0(s\cup s_{0n})) \le H_{nj}(\boldsymbol{\beta}(s\cup s_{0n})) \\ \le (1+\xi)H_{nj}(\boldsymbol{\beta}_0(s\cup s_{0n})), \end{aligned}$$

whenever $\|\boldsymbol{\beta}(s\cup s_{0n}) - \boldsymbol{\beta}_0(s\cup s_{0n})\|_2 \le \delta$ for all s with $|s| \le kp_{0n}$;

C6 For $i = 1, \ldots, n; j = 1, \ldots, p_n$, the quantities $|x_{ij}|$, $|h'(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)|$, $|h''(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)|$ are bounded from above, and $\sigma_i^2, i = 1, \ldots, n$ are bounded both from above and below away from zero. Furthermore,

$$\max_{1\le j\le p_n; 1\le i\le n} \frac{x_{ij}^2[h'(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)]^2}{\sum_{i=1}^n \sigma_i^2 x_{ij}^2[h'(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)]^2} = o(n^{-1/3}),$$

$$\max_{1\le i\le n} \frac{[h''(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)]^2}{\sum_{i=1}^n \sigma_i^2[h''(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)]^2} = o(n^{-1/3}).$$

Conditions C2 and C3 are the same as conditions A2 and A3 in Chen and Chen (2012). Conditions C4–C5 reduce to conditions A4–A5 in Chen and Chen (2012) for canonical links. When A6 in Chen and Chen (2012) is satisfied, C6 is satisfied by commonly used GLIMs such as Poisson distribution with log and power function links, Binary distribution with identity, arcsin, complementary log-log and probit links, Gamma distribution with log and inverse power function links. These GLIMs are thoroughly studied in Wedderburn (1976). The verification of C6 for these GLIMs is given in a complementary document at website: http://www.intlpress.com/SII/p/2013/6-2/SII-6-2-luo-supplement.pdf.

We now state our main results as follows. Define $\mathcal{A}_0 = \{s : s_{0n} \subset s, s_{0n} \ne s, |s| \le kp_{0n}\}$ and $\mathcal{A}_1 = \{s : s_{0n} \not\subset s, |s| \le kp_{0n}\}$. We have

Theorem 2.1. *Under assumptions C1–C6, as $n \to +\infty$,*

(1) For any $\gamma > 0$,

$$P\left(\min_{s\in\mathcal{A}_1} EBIC_\gamma(s) \le EBIC_\gamma(s_{0n})\right) \to 0.$$

(2) For any $\gamma > \frac{1}{1-\epsilon}(1 - \frac{\log n}{2\log p_n})$, where ϵ is an arbitrarily small positive constant,

$$P\left(\min_{s\in\mathcal{A}_0} EBIC_\gamma(s) \le EBIC_\gamma(s_{0n})\right) \to 0.$$

The following result is needed in the proof of Theorem 2.1.

Lemma 2.1. *Under conditions C1–C6, whenever $\|\boldsymbol{\beta}(s\cup s_{0n}) - \boldsymbol{\beta}_0(s\cup s_{0n})\|_2 \le \delta$,*

(1)
$$\boldsymbol{u}^\tau H_n(\boldsymbol{\beta}(s\cup s_{0n}))\boldsymbol{u} = \boldsymbol{u}^\tau H_{n1}(\boldsymbol{\beta}(s\cup s_{0n}))\boldsymbol{u}(1 + o_p(1)),$$

uniformly in s with $|s| \le kp_{0n}$, where $\boldsymbol{u}$ is any unit vector of dimension $|s\cup s_{0n}|$.

Equation (1) is satisfied when $\boldsymbol{\beta}(s\cup s_{0n})$ is replaced by $\boldsymbol{\beta}_0(s\cup s_{0n})$ because of C4 and the fact that $\boldsymbol{u}^\tau H_{n0}(\boldsymbol{\beta}_0(s\cup s_{0n}))\boldsymbol{u} = o_p(n)$, see the proof of Lemma 2.1 in the Appendix. Lemma 2.1 indicates that equation (1) in fact holds in a neighbourhood of $\boldsymbol{\beta}_0(s\cup s_{0n})$. By using Lemma 2.1, Theorem 2.2 below will be proved. This theorem provides the convergence rate of the L_2-consistency of the MLE of $\boldsymbol{\beta}(s)$ when $s_{0n} \subset s$, which is of its own interest.

Theorem 2.2. *Under conditions C1–C6, as $n \to \infty$, $\|\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\|_2 = O_p(n^{-1/3})$, uniformly for $s \in \mathcal{A}_0$.*

The technical details of the proof for the above results are given in the Appendix. Theorem 2.1 implies that if we confine to the models with cardinality less than or equal to kp_{0n} and select the model with the smallest EBIC among all those models then, with probability converging to 1, the selected model, say, s_n^*, will be the same as the true model s_{0n}. This property is what is called selection consistency. The constraint that $|s| \leq kp_{0n}$ is natural since we do not need to consider any models with cardinality much larger than that of the true model in practical problems. However, in practice, the evaluation of all models with cardinality up to kp_{0n} is computationally impossible. Like any other model selection criteria, the EBIC is to be used in a certain model selection procedure. In addition to the traditional forward selection procedures, a variety of procedures based on penalized likelihood approach have been developed within the last twenty years such as the LASSO (Tibishirani, 1996), SCAD (Fan and Li, 2001), Elastic Net (Zou and Hastie, 2005), and so on. A model selection criterion can be used in these procedures to choose the penalty parameter, which corresponds to choosing a model. However, though some desirable properties such as the so-called oracle property have been established for these penalized likelihood approaches under certain conditions, the asymptotic properties of these approaches with GLIM and ultra-high dimensional feature space have not been thoroughly studied yet to our knowledge. The traditional forward selection methods have been criticized for its greedy nature. But, recently, it is discovered that the greedy nature might not be bad especially when the model selection is for the selection of relevant variables rather than for a prediction model, see, e.g., Tropp (2004), Tropp and Gilbert (2007) and Wang (2009). In this article, we consider the application of the EBIC with the traditional forward regression procedure for GLIM in our simulation studies and real data analysis.

3. SIMULATION STUDY

In our simulation studies, we consider a GLIM with binary response and the complementary log-log link. We take the divergent pattern $(n, p_n, p_{0n}) = (n, [40e^{n^{0.2}}], [5n^{0.1}])$ for $n = 100, 200, 500$. The settings for the covariates, which are adapted from Fan and Song (2010), are described below.

Setting 1. Let $q = 15$, $s_1 = \{1, \ldots, q\}$, $s_2 = \{q+1, \ldots, [\frac{p_n}{3}]\}$, $s_3 = \{[\frac{p_n}{3}]+1, \ldots, [\frac{2p_n}{3}]\}$ and $s_4 = \{[\frac{2p_n}{3}]+1, \ldots, p_n\}$. Let the covariate vector $\boldsymbol{x}$ be decomposed into $\boldsymbol{x} = (\boldsymbol{x}(s_1), \boldsymbol{x}(s_2), \boldsymbol{x}(s_3), \boldsymbol{x}(s_4))$, where $\boldsymbol{x}(s_1)$ is generated from $N(\mathbf{0}, \Sigma_\rho)$, Σ_ρ having diagonal elements 1 and off-diagonal elements ρ, $\boldsymbol{x}(s_2)$ is generated from $N(\mathbf{0}, I)$, the components of $\boldsymbol{x}(s_3)$ are generated independently from a double exponential distribution with location 0 and scale 1, and the components of $\boldsymbol{x}(s_4)$ are generated independently from the normal mixture $\frac{1}{2}[N(-1,1) + N(1,0.5)]$. The covariates $\boldsymbol{x}_i(s_k), i = 1, \ldots, n$, are i.i.d. copies of $\boldsymbol{x}(s_k)$, $k = 1, 2, 3, 4$. Four values of ρ: 0, 0.3, 0.5 and 0.7, are considered. $s_{0n} = \{L \times t, t = 1, \ldots, p_{0n}\}$, where $L = 10$. $\beta_j = 1$, if $j = L \times t$ with odd t, 1.3, if $j = L \times t$ with even t, 0, otherwise.

Setting 2. The same as setting 1 except $L = 5$. The essential difference between setting 1 and this setting is that, in setting 1, all the relevant features are independent while, in this setting, three of them have pairwise correlation ρ. Two values of ρ: 0.3 and 0.5, are considered in this setting.

Setting 3. $L = 10, q = 50$. In all the settings for (n, p_n, p_{0n}), this q is much smaller than p_n and $p_n - q$ is much bigger than Lp_{0n}. The distribution of the covariate vector $\boldsymbol{x}$ is specified as follows. For $j = 1, \ldots, p_n - q$, the components x_j's are i.i.d. standard normal variables. For $p_n - q < j \leq p_n$,

$$x_j = \frac{1}{5}\left[\sum_{t=1}^{p_{0n}} (-1)^{t+1} x_{Lt} + \sqrt{25 - p_{0n}}\xi_j\right],$$

where the ξ_j's are i.i.d. standard normal variables. $\boldsymbol{x}_i$'s are generated as i.i.d. copies of $\boldsymbol{x}$. The specification for s_{0n} and $\boldsymbol{\beta}$ is the same as in setting 1. In this setting, all the relevant features are independent, the last q irrelevant features, which are highly pairwise correlated, have a weak marginal correlation with each of the relevant features but a strong overall correlation with the totality of the relevant features.

In our simulation studies, we apply the EBIC in the forward selection procedure for model selection. The forward selection procedure starts by fitting the GLIMs with one covariate, the covariate corresponding the model with the largest maximum likelihood is the first selected variable. Then GLIMs with two covariates including the first selected variable are considered, the additional covariate corresponding to the two-covariate model with the largest maximum likelihood is the second selected variable. The procedure continues this way and, at each step, one more variable is selected. The EBIC is used as a stopping rule. At each step, the EBIC is computed for the model consisting of the selected variables. The selection procedure stops when EBIC reaches a minimum. To reduce the amount of computation, when p_n is bigger than 1,000, the sure independence screening procedure based on the maximum marginal estimator (MME) (Fan and Song (2010)) is used to reduce the dimension of the feature to 400 before the forward selection procedure is invoked. We consider four γ values in EBIC, i.e., $\gamma_1 = 0, \gamma_2 = \frac{1}{2}(1 - \frac{\ln n}{2\ln p_n}), \gamma_3 = 1 - \frac{\ln n}{4\ln p_n}$ and $\gamma_4 = 1$. We choose these values because γ_1 corresponds to the original BIC, γ_4 corresponds to mBIC, γ_2 is halfway between 0 and $1 - \frac{\ln n}{2\ln p_n}$, the lower bound of the consistent range of γ, and γ_3 is halfway between $1 - \frac{\ln n}{2\ln p_n}$ and 1. Thus we can evaluate the asymptotic behavior of EBIC when the γ value is below and above the lower bound of the consistent range and also make a comparison with BIC and mBIC. The performance of the procedure is evaluated by positive discovery rate (PDR) and false discovery rate (FDR) defined

Table 1. The PDR and FDR of the forward selection procedure with EBIC under simulation setting 1 (the PDR and FDR are averaged over 200 replicates, the numbers in parenthesis are standard errors)

		γ_1		γ_2		γ_3		γ_4	
ρ	n	PDR	FDR	PDR	FDR	PDR	FDR	PDR	FDR
0	100	0.736	0.375	0.735	0.362	0.646	0.193	0.481	0.074
		(0.281)	(0.292)	(0.284)	(0.291)	(0.382)	(0.228)	(0.453)	(0.141)
	200	0.930	0.272	0.918	0.223	0.879	0.127	0.862	0.078
		(0.220)	(0.252)	(0.253)	(0.215)	(0.311)	(0.147)	(0.337)	(0.108)
	500	0.971	0.408	0.963	0.371	0.939	0.079	0.936	0.026
		(0.135)	(0.181)	(0.163)	(0.152)	(0.231)	(0.119)	(0.238)	(0.062)
0.3	100	0.708	0.407	0.708	0.398	0.621	0.196	0.471	0.081
		(0.298)	(0.296)	(0.298)	(0.306)	(0.384)	(0.230)	(0.442)	(0.152)
	200	0.933	0.281	0.924	0.239	0.889	0.143	0.855	0.083
		(0.202)	(0.248)	(0.232)	(0.212)	(0.303)	(0.161)	(0.344)	(0.111)
	500	0.969	0.428	0.959	0.354	0.938	0.047	0.933	0.014
		(0.130)	(0.169)	(0.177)	(0.138)	(0.238)	(0.091)	(0.247)	(0.048)
0.5	100	0.712	0.401	0.711	0.383	0.632	0.201	0.451	0.080
		(0.293)	(0.295)	(0.294)	(0.292)	(0.385)	(0.223)	(0.447)	(0.146)
	200	0.929	0.281	0.923	0.243	0.881	0.128	0.858	0.084
		(0.219)	(0.257)	(0.236)	(0.223)	(0.313)	(0.130)	(0.343)	(0.110)
	500	0.967	0.434	0.959	0.371	0.939	0.043	0.933	0.006
		(0.142)	(0.166)	(0.168)	(0.147)	(0.235)	(0.085)	(0.249)	(0.031)
0.7	100	0.674	0.432	0.674	0.414	0.606	0.244	0.430	0.092
		(0.291)	(0.289)	(0.291)	(0.287)	(0.365)	(0.241)	(0.432)	(0.144)
	200	0.931	0.292	0.926	0.248	0.888	0.148	0.874	0.112
		(0.196)	(0.246)	(0.218)	(0.207)	(0.295)	(0.146)	(0.314)	(0.125)
	500	0.970	0.427	0.966	0.365	0.937	0.032	0.934	0.010
		(0.134)	(0.173)	(0.150)	(0.150)	(0.234)	(0.072)	(0.240)	(0.038)

as follows:

$$\mathrm{PDR}_n = \frac{\nu(s^* \cap s_{0n})}{\nu(s_{0n})}, \qquad \mathrm{FDR}_n = \frac{\nu(s^* \backslash s_{0n})}{\nu(s^*)},$$

where s^* is the set of selected features. The selection consistency is equivalent to $P(\mathrm{PDR}_n = 1, \mathrm{FDR}_n = 0) \to 1$, as $n \to \infty$, which implies $\mathrm{PDR}_n \to 1$ and $\mathrm{FDR}_n \to 0$, in probability.

The PDR and FDR are averaged over 200 replications. The results under Settings 1–3 are reported in Tables 1–3 respectively.

By examining Tables 1–3, we can find the following common trends: 1) with all the four γ values, the PDR increases as n gets larger, 2) with γ_1 and γ_2 (which are below the lower bound of the consistent range), the FDR does not show a trend to decrease while, with γ_3 and γ_4 (which are within the consistent range), the FDR reduces rapidly towards zero, 3) though the PDRs with γ_3 and γ_4 are lower than those with γ_1 and γ_2 when sample size is small, but they become comparable as the sample size increases, and 4) the FDR with γ_4 is lower than that with γ_3 when sample size is small, however, the PDR is also lower, as sample size gets larger, both the PDR and FDR with γ_3 and those with γ_4 become comparable. These findings demonstrate that the selection consistency of EBIC is well realized in a finite sample case.

4. REAL DATA ANALYSIS

In this section, we apply the forward selection procedure with EBIC to analyze a Leukemia data set. The data consists of the expression levels of 7,129 genes obtained from 47 patients with acute lymphoblastic leukemia (ALL) and 25 with acute myeloid leukemia (AML). The data set is available in the R packages *Biobase* and *golubEsets*. The initial version of this data set is described and analyzed by a method called "neighborhood analysis" in Golub et al. (1999). The data set is later analyzed using GLIM with probit link in Lee et al. (2003) and using GLIM with logit link in Liao and Chin (2007). 50 genes are identified as important ones affecting the types of leukemia in Golub et al. (1999), 27 genes are identified in Lee et al. (2003), and 19 genes are identified in Liao and Chin (2007). There are only a few overlapped genes among the three identified sets.

We analyzed the data by the forward selection procedure with four different link functions: *logit, probit, cauchit* and *cloglog*. First, with each link function, the procedure was carried out until 50 genes were selected. The identified genes are reported in Table 4. These 50 genes are compared with three identified sets mentioned above. Those which were identified in Golub et al. (1999), Lee et al. (2003) and Liao and Chin (2007) are indicated by $\star$, $\triangle$ and $*$ respectively. There are three genes: 1834,1882, 6855, which are in all the three

Table 2. The PDR and FDR of the forward selection procedure with EBIC under simulation setting 2 (the PDR and FDR are averaged over 200 replicates, the numbers in parenthesis are standard errors)

		γ_1		γ_2		γ_3		γ_4	
ρ	n	PDR	FDR	PDR	FDR	PDR	FDR	PDR	FDR
0.3	100	0.662	0.424	0.660	0.409	0.594	0.233	0.492	0.132
		(0.272)	(0.287)	(0.276)	(0.286)	(0.350)	(0.237)	(0.392)	(0.195)
	200	0.931	0.256	0.926	0.231	0.891	0.111	0.881	0.068
		(0.199)	(0.245)	(0.212)	(0.222)	(0.281)	(0.137)	(0.295)	(0.101)
	500	0.973	0.401	0.967	0.339	0.946	0.041	0.941	0.018
		(0.127)	(0.173)	(0.149)	(0.134)	(0.209)	(0.089)	(0.217)	(0.055)
0.5	100	0.571	0.489	0.570	0.478	0.521	0.304	0.442	0.189
		(0.259)	(0.274)	(0.261)	(0.276)	(0.303)	(0.265)	(0.337)	(0.230)
	200	0.918	0.272	0.910	0.239	0.888	0.121	0.869	0.081
		(0.204)	(0.256)	(0.230)	(0.231)	(0.267)	(0.148)	(0.293)	(0.122)
	500	0.970	0.402	0.964	0.351	0.946	0.056	0.942	0.021
		(0.129)	(0.183)	(0.148)	(0.153)	(0.199)	(0.115)	(0.212)	(0.062)

Table 3. The PDR and FDR of the forward selection procedure with EBIC under simulation setting 3 (the PDR and FDR are averaged over 200 replicates, the numbers in parenthesis are standard errors)

	γ_1		γ_2		γ_3		γ_4	
n	PDR	FDR	PDR	FDR	PDR	FDR	PDR	FDR
100	0.586	0.506	0.586	0.484	0.524	0.332	0.387	0.198
	(0.258)	(0.252)	(0.258)	(0.253)	(0.316)	(0.252)	(0.366)	(0.239)
200	0.796	0.414	0.791	0.386	0.767	0.285	0.746	0.221
	(0.261)	(0.282)	(0.274)	(0.273)	(0.311)	(0.247)	(0.334)	(0.228)
500	0.946	0.479	0.936	0.416	0.912	0.195	0.896	0.171
	(0.167)	(0.165)	(0.197)	(0.150)	(0.248)	(0.185)	(0.269)	(0.176)

identified sets are selected by the forward selection procedure. They are all among the selected genes with logit and cloglog links. Two of them, i.e., 1834, 1882, are only among the selected genes with probit and cauchit links. The other selected genes except two of them are in only one of the identified sets. Note that the selected genes and their ordering are different among the four different links. This indicates that the link function does matter in the selection procedure. Second, we used 8-fold cross validation to select the optimal link function among the four links. The optimal link is the logit link. Finally, we made a final selection using EBIC with $\gamma = 1 - \frac{\ln n}{3 \ln p_n}$ which is slightly bigger than the lower bound of the consistent range. The final selected variables together with the maximum log likelihood of the corresponding model are reported in Table 5. To compare the final selection of the logit link with the other links, the selected results with all the four links are reported. The genes selected by the logit link are 1834 and 4438. The maximum log likelihood of the selected model with the logit link is the largest among all the four links. Note that, the same two genes are also selected by probit link and the gene 4438 is selected by cloglog link. We thus can conclude quite confidently that the two genes selected by logit link are the most important genes for studying the etiology of leukemia.

APPENDIX A. TECHNICAL PROOFS

Proof of Lemma 2.1. First consider $s \in \mathcal{A}_1$, let $\tilde{s} = s \cup s_{0n}$. Let a_{ni} in Lemma 1 of Chen and Chen (2012) be $h''(\boldsymbol{x}_i^\tau(\tilde{s})\boldsymbol{\beta}_0(\tilde{s}))\mathrm{sign}(y_i - \mu_i)/\sqrt{\sum_{i=1}^n \sigma_i^2 (h''(\boldsymbol{x}_i^\tau(\tilde{s})\boldsymbol{\beta}_0(\tilde{s})))^2}$. Since $\boldsymbol{x}_i^\tau(\tilde{s})\boldsymbol{\beta}_0(\tilde{s}) = \boldsymbol{x}_i^\tau \boldsymbol{\beta}_0$, from Condition C6, we have

$$(2)\quad P\left(\sum_{i=1}^n |(y_i - \mu_i)h''\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right)| \geq Cn^{2/3}\right) \leq 2\exp(-Cn^{1/3}).$$

For any unit vector $\boldsymbol{u}$ with dimension $|\tilde{s}|$,

$$(3)\quad \begin{aligned} \boldsymbol{u}^\tau H_{n0}(\boldsymbol{\beta}_0(\tilde{s}))\boldsymbol{u} &= \sum_{i=1}^n (y_i - \mu_i)h^{''}\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right)\left(\boldsymbol{u}^\tau \boldsymbol{x}_i(\tilde{s})\right)^2 \\ &\leq \sum_{i=1}^n |(y_i - \mu_i)h^{''}\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right)| \|\boldsymbol{x}_i(\tilde{s})\|_2^2 \\ &\leq C(k+1)p_{0n} \sum_{i=1}^n |(y_i - \mu_i)h^{''}\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right)|. \end{aligned}$$

The last inequality is true because all x_{ij}'s are bounded, as assumed in C6. (2) and (3) together with C5 imply that,

Table 4. Analysis of Leukemia Data: the top 50 genes selected by the forward selection procedure with the four links: logit (lo), probit (pr), cauchit (ca) and cloglog (cl)

	Rank and Gene ID									
	1	2	3	4	5	6	7	8	9	10
lo	$1834^{*\triangle\star}$	4438	4951	$6539^{\star}$	155	2181	$1882^{*\triangle\star}$	6472	65	1953
pr	$1834^{*\triangle\star}$	4438	4951	155	5585	5466	706	$7119^{\star}$	3119	4480
ca	$1882^{*\triangle\star}$	4951	$6281^{\star}$	4499	4443	$6539^{\star}$	5107	$1834^{*\triangle\star}$	4480	6271
cl	$1834^{*\triangle\star}$	$6855^{*\triangle\star}$	4377	5122	2830	4407	4780	6309	$4973^{\star}$	715
	11	12	13	14	15	16	17	18	19	20
lo	3692	706	1787	$5191^{\star}$	1239	3119	2784	1078	3631	6308
pr	$6201^{\triangle}$	490	6895	$1882^{*\triangle\star}$	1809	2855	3123	4211^{*}	$2020^{*\star}$	3631
ca	6378	3631	$2111^{\star}$	$6201^{\triangle}$	$6373^{\star}$	1800	4780	321	$4107^{\triangle}$	$1779^{\triangle}$
cl	5376	930	1800	$1882^{*\triangle\star}$	5794	4399	$4389^{\star}$	922	1962	4267
	21	22	23	24	25	26	27	28	29	30
lo	$6373^{\star}$	$1909^{\star}$	4153	$1685^{\triangle}$	$6855^{*\triangle\star}$	7073	5539	2830	4819	6347
pr	5823	1953	$1745^{\triangle\star}$	65	997	$1928^{\star}$	3307	1787	538	5539
ca	6277	1544	$5254^{\star}$	$1928^{\star}$	$1745^{\triangle\star}$	3163	7073	310	$4389^{\star}$	5146
cl	1926	4229	$5254^{\star}$	770	2141	6923	7073	2828	$4847^{\star}$	698
	31	32	33	34	35	36	37	38	39	40
lo	1081	1095	5328	4279	4373	5737	4366	5280	3307	284
pr	4107	2385	1087	$1909^{\star}$	5376	5552	6005	1604	3391	5442
ca	1927	885	3137	2258	4334	6657	2733	5336	5972	6167
cl	1779	$1928^{\star}$	4049	876	6857	6347	$6376^{\star}$	2361	4664	758
	41	42	43	44	45	46	47	48	49	50
lo	6676	4291	1945	4079	3722	668	782	$4196^{\star}$	25	$4389^{\star}$
pr	6702	6309	$2348^{\star}$	4282	4925	6167	2323	1779	5122	$3847^{\star}$
ca	4229	$4328^{\star}$	715	4149	$5191^{\star}$	6283	200	6702	5794	4190
cl	3631	6308	4499	4480	5971	6510	5300	3475	3932	6801

Table 5. Analysis of Leukemia Data: the final selected genes by EBIC

Link Function	Selected Genes	Maximum Likelihood
logit	1834, 4438	−2.296e-08
probit	1834, 4438	−3.022e-08
cauchit	1882, 4951	−2.122e-06
cloglog	1834, 6855	−6.908e-08

for any $\xi > 0$, there exists a $\delta > 0$ such that under the constraint $\|\boldsymbol{\beta}(s \cup s_{0n}) - \boldsymbol{\beta}_0(s \cup s_{0n})\|_2 \leq \delta$,

$$\begin{aligned} P&\left(\max_{s\in\mathcal{A}_1,\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\tau H_{n0}\left(\boldsymbol{\beta}(s\cup s_{0n})\right)\boldsymbol{u} \geq C p_{0n} n^{2/3}\right) \\ &\leq P\left(\max_{s\in\mathcal{A}_1,\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\tau H_{n0}\left(\boldsymbol{\beta}_0(s\cup s_{0n})\right)\boldsymbol{u} \right. \\ &\qquad\qquad \left. \geq \frac{C}{1+\xi} p_{0n} n^{2/3}\right) \\ &\leq |\mathcal{A}_1| P\left(\sum_{i=1}^n |(y_i - \mu_i) h''\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right)| \geq \tilde{C} n^{2/3}\right) \\ &\leq 2\exp\left(k p_{0n} \ln p_n - \frac{C}{1+\xi} n^{1/3}\right) = o(1); \end{aligned}$$

that is, $\boldsymbol{u}^\tau H_{n0}(\boldsymbol{\beta}(s \cup s_{0n}))\boldsymbol{u} = O_p(p_{0n} n^{2/3}) = o_p(n)$, since $p_{0n} = o(n^{1/3})$ by C1. By C4 and C5, $\boldsymbol{u}^\tau H_{n1}(\boldsymbol{\beta}(\tilde{s}))\boldsymbol{u}$ is of order n. Thus the lemma is proved for $s \in \mathcal{A}_1$. For $s \in \mathcal{A}_0$, since $s_{0n} \subset s$, by replacing $\tilde{s}$ with s in the above argument, the lemma is also proved for $s \in \mathcal{A}_0$. □

Proof of Theorem 2.1. According to the definition of EBIC, for any model s, $\mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})$ if and only if

$$\begin{aligned} (4)\qquad & \ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \\ & \qquad \geq (|s| - p_{0n}) \ln n/2 + \gamma\left(\ln \tau(S_{|s|}) - \ln \tau(S_{p_{0n}})\right). \end{aligned}$$

To prove the selection consistency of EBIC, or mathematically, as $n \to +\infty$,

$$P\left(\min_{s:|s|\leq k p_{0n}, s\neq s_{0n}} \mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})\right) \to 0,$$

it suffices to show that inequality (4) holds with a probability converging to 0 as the sample size goes to infinity uniformly for all $s \in \mathcal{A}_0 \cup \mathcal{A}_1$. This is completed by dealing with $s \in \mathcal{A}_0$ and $\mathcal{A}_1$ separately.

(I) *Case 1:* $s \in \mathcal{A}_1$. Since $\tau(S_{p_{0n}}) < p_n^{p_{0n}}$ and $|s| \ln n/2 + \gamma \ln \tau(S_{|s|}) > 0$, inequality (4) implies that

$$(5)\quad \ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \geq -p_{0n}(\ln n/2 + \gamma \ln p_n).$$

Therefore, define

$$D = \sup_{s \in \mathcal{A}_1} \ln L_n\big(\hat{\boldsymbol{\beta}}(s)\big) - \ln L_n\big(\hat{\boldsymbol{\beta}}(s_{0n})\big),$$

if we can show

$$(6) \qquad P\left(D \geq -p_{0n}(\ln n/2 + \gamma \ln p_n)\right) \to 0 \quad \text{as } n \to +\infty,$$

then we will have

$$P\left(\min_{s: s \in \mathcal{A}_1} \mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})\right) \to 0 \quad \text{as } n \to +\infty.$$

The key becomes to assess the order for $\sup_{s \in \mathcal{A}_1} \ln L_n(\hat{\boldsymbol{\beta}}(s)) - \ln L_n(\hat{\boldsymbol{\beta}}(s_{0n}))$. For any $s \in \mathcal{A}_1$, let $\tilde{s} = s \cup s_{0n}$ and $\breve{\boldsymbol{\beta}}(\tilde{s})$ be $\hat{\boldsymbol{\beta}}(s)$ augmented with zeros corresponding to the elements in $\tilde{s} \backslash s$. It can be seen that

$$\ln L_n\big(\boldsymbol{\beta}_0(\tilde{s})\big) = \ln L_n\big(\boldsymbol{\beta}_0(s_{0n})\big) \leq \ln L_n\big(\hat{\boldsymbol{\beta}}(s_{0n})\big),$$
$$\ln L_n\big(\hat{\boldsymbol{\beta}}(s)\big) = \ln L_n\big(\breve{\boldsymbol{\beta}}(\tilde{s})\big),$$

which lead to

$$(7) \qquad \begin{aligned} &\sup_{s \in \mathcal{A}_1} \ln L_n\big(\hat{\boldsymbol{\beta}}(s)\big) - \ln L_n\big(\hat{\boldsymbol{\beta}}(s_{0n})\big) \\ &\quad \leq \sup_{s \in \mathcal{A}_1} \ln L_n\big(\breve{\boldsymbol{\beta}}(\tilde{s})\big) - \ln L_n\big(\boldsymbol{\beta}_0(\tilde{s})\big). \end{aligned}$$

And also

$$\|\breve{\boldsymbol{\beta}}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 \geq \|\boldsymbol{\beta}(s_{0n} \backslash s)\|_2 > \min_{j \in s_{0n}} \{|\boldsymbol{\beta}_j|\} \geq Cn^{-1/4}.$$

To simplify the left-hand side of (7), we firstly investigate $\sup\{\ln L_n(\boldsymbol{\beta}(\tilde{s})) - \ln L_n(\boldsymbol{\beta}_0(\tilde{s})) : \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 = n^{-1/4}, s \in \mathcal{A}_1\}$:

To derive the order of the right-hand side in the above inequality, we take the Taylor Expansion of $\ln L_n(\boldsymbol{\beta}(\tilde{s})) - \ln L_n(\boldsymbol{\beta}_0(\tilde{s}))$ as follows:

$$(8) \qquad \begin{aligned} &\ln L_n\left(\boldsymbol{\beta}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \\ &\quad = \left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau s_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \\ &\quad\quad - \frac{1}{2}\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau H_{n1}\left(\boldsymbol{\beta}^*(\tilde{s})\right)\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right) \\ &\quad\quad + \frac{1}{2}\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau H_{n0}\left(\boldsymbol{\beta}^*(\tilde{s})\right)\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right) \end{aligned}$$

where $\boldsymbol{\beta}^*(\tilde{s})$ is between $\boldsymbol{\beta}(\tilde{s})$ and $\boldsymbol{\beta}_0(\tilde{s})$ component-wise. By condition C4 and C5,

$$\begin{aligned} &\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau H_{n1}\left(\boldsymbol{\beta}^*(\tilde{s})\right)\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right) \\ &\quad \geq c_1 n(1-\xi)\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2^2. \end{aligned}$$

Lemma 2.1 implies that, for any $\boldsymbol{\beta}(\tilde{s})$ such that $\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 = n^{-1/4}$, uniformly, the third term in equation (8) is positive and has order $o_p(n\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2^2)$, which is $o_p(n^{1/2})$. Thus, when $\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 = n^{-1/4}$, there exists $0 < c < c_1$ such that, with probability tending to 1 as n goes to $+\infty$,

$$(9) \qquad \begin{aligned} &\ln L_n\left(\boldsymbol{\beta}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \\ &\quad \leq \|\boldsymbol{\beta}(s) - \boldsymbol{\beta}^{s_{0n}}\|_1 \|s_n(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} - \frac{c}{2}n^{1/2}(1-\xi) \\ &\quad \leq \sqrt{|s|}\|\boldsymbol{\beta}(s) - \boldsymbol{\beta}^{s_{0n}}\|_2 \|s_n(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} - \frac{c}{2}n^{1/2}(1-\xi) \\ &\quad \leq Cp_{0n}^{1/2} n^{-1/4}\|s_n(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} - \frac{c}{2}n^{1/2}(1-\xi). \end{aligned}$$

In the following, we show that

$$(10) \qquad Cp_{0n}^{1/2} n^{-1/4}\|s_n(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} = o_p(n^{1/2}),$$

which then implies that

$$(11) \qquad \ln L_n\left(\boldsymbol{\beta}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \leq -Cn^{1/2},$$

for some positive constant C. We first show that

$$(12) \qquad P\left(\max_{1 \leq j \leq p_n} s_{nj}\left(\boldsymbol{\beta}_0\right) \geq Cn^{2/3}\right) = o(1).$$

For fixed j, let

$$a_{ni} = x_{ij} h'(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0) \Big/ \sqrt{\sum_{i=1}^n \sigma_i^2 x_{ij}^2 (h'(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0))^2}.$$

Under C6, by applying Lemma 1 of Chen and Chen (2012), we have

$$\begin{aligned} &P\big(s_{nj}\big(\boldsymbol{\beta}_0\big) \geq Cn^{2/3}\big) \\ &\quad = P\left(\sum_{i=1}^n a_{ni}(y_i - \boldsymbol{\mu}_i)\right. \\ &\quad\quad\quad \left. > Cn^{2/3} \Big/ \sqrt{\sum_{i=1}^n \sigma_i^2 x_{i,j}^2 \left(h'(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0)\right)^2}\right) \\ &\quad \leq P\left(\sum_{i=1}^n a_{ni}(y_i - \boldsymbol{\mu}_i) > Cn^{1/6}\right) \leq \exp(-Cn^{1/3}). \end{aligned}$$

Under C1, $\ln p_n = o(n^{1/3})$, hence

$$\sum_{j=1}^{p_n} P\big(s_{nj}\big(\boldsymbol{\beta}_0\big) \geq Cn^{2/3}\big) = \exp(\ln p_n - Cn^{1/3}) = o(1),$$

which implies (12). Equality (10) then follows from the fact that $p_{0n} = o(n^{1/6})$ under C1. Equality (11) indicates that the maximum likelihood estimate $\hat{\boldsymbol{\beta}}(\tilde{s})$ is an interior point of $\mathcal{N}_{\boldsymbol{\beta}_0(\tilde{s})} = \{\boldsymbol{\beta}(\tilde{s}) : \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 \leq n^{-1/4}\}$. By its definition, $\breve{\boldsymbol{\beta}}(\tilde{s}) \notin \mathcal{N}_{\boldsymbol{\beta}_0(\tilde{s})}$. Therefore,

$$(13) \qquad \sup_{s \in \mathcal{A}_1} \ln L_n\big(\breve{\boldsymbol{\beta}}(\tilde{s})\big) - \ln L_n\big(\boldsymbol{\beta}_0(\tilde{s})\big)$$

$$
\begin{aligned}
&\le \sup_{s\in\mathcal{A}_1} \{\ln L_n(\boldsymbol{\beta}(\tilde{s})) - \ln L_n(\boldsymbol{\beta}_0(\tilde{s})) : \\
&\qquad\qquad \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 \ge n^{-1/4}\} \\
&= \sup_{s\in\mathcal{A}_1} \{\ln L_n(\boldsymbol{\beta}(\tilde{s})) - \ln L_n(\boldsymbol{\beta}_0(\tilde{s})) : \\
&\qquad\qquad \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 = n^{-1/4}\}.
\end{aligned}
$$

The last equation follows from the concavity of $l_n(\boldsymbol{\beta})$. Inequalities (7) and (13) together lead to

$$
\sup_{s\in\mathcal{A}_1} \ln L_n\big(\hat{\boldsymbol{\beta}}(s)\big) - \ln L_n\big(\hat{\boldsymbol{\beta}}(s_{0n})\big) \le -Cn^{1/2}.
$$

Since under C1, $p_{0n}\ln n = o(n^{1/3}), p_{0n}\ln p_n = o(n^{1/3})$, inequality (6) then follows.

(II) *Case 2:* $s \in \mathcal{A}_0$. When $p_{0n} = O(n^b), b < 1/6, \ln p_n = O(n^\kappa), \kappa > 0$, and $p_{0n} < |s| < kp_{0n}$, we have $\dfrac{\ln|s|}{\ln p_n} \to 0$. It follows from Lemma 1 of Luo and Chen (2011) that $\ln\tau(S_{|s|}) \approx |s|\ln p_n, \ln\tau(S_{p_{0n}}) \approx p_{0n}\ln p_n$. Hence, asymptotically, $\mathrm{EBIC}_\gamma(s) \le \mathrm{EBIC}_\gamma(s_{0n})$ if and only if

$$
\ln L_n\big(\hat{\boldsymbol{\beta}}(s)\big) - \ln L_n\big(\hat{\boldsymbol{\beta}}(s_{0n})\big) \ge m[0.5\ln n + \gamma\ln p_n], \tag{14}
$$

where $m = |s| - \nu(s_{0n})$. Therefore, it suffices to show

$$
\begin{aligned}
P\bigg(&\sup_{s\in\mathcal{A}_0} \ln L_n\big(\hat{\boldsymbol{\beta}}(s)\big) - \ln L_n\big(\hat{\boldsymbol{\beta}}(s_{0n})\big) \\
&\ge m[0.5\ln n + \gamma\ln p_n]\bigg) \to 0 \quad \text{as } n\to\infty.
\end{aligned} \tag{15}
$$

Note that

$$
\begin{aligned}
&\ln L_n\big(\hat{\boldsymbol{\beta}}(s)\big) - \ln L_n\big(\hat{\boldsymbol{\beta}}(s_{0n})\big) \\
&\quad \le \ln L_n\big(\hat{\boldsymbol{\beta}}(s)\big) - \ln L_n\big(\boldsymbol{\beta}_0(s_{0n})\big)) \\
&\quad = (\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau s_n(\boldsymbol{\beta}_0(s)) \\
&\qquad - \frac{1}{2}(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau H_n(\tilde{\boldsymbol{\beta}}(s))(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)) \\
&\quad = (\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau s_n(\boldsymbol{\beta}_0(s)) \\
&\qquad - \frac{1}{2}(1+o_p(1))(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau H_{n1}(\tilde{\boldsymbol{\beta}}(s))(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)) \\
&\quad \le (\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau s_n(\boldsymbol{\beta}_0(s)) \\
&\qquad - \frac{1-\xi}{2}(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau H_{n1}(\tilde{\boldsymbol{\beta}}(s))(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)),
\end{aligned} \tag{16}
$$

for an arbitrarily small positive ξ when n is large, where $\tilde{\boldsymbol{\beta}}(s)$ is between $\hat{\boldsymbol{\beta}}(s)$ and $\boldsymbol{\beta}_0(s)$ component-wise. The first equality follows from the Taylor expansion and the second equality follows from Lemma 1. We are going to apply C5 to simplify the right-hand side of the above inequality. The applicability of C5 requires that $\sup_{s\in\mathcal{A}_0}\|\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\|_2 \to 0$ as n goes to infinity. In fact, we have, under conditions C1–C6, uniformly for $s \in \mathcal{A}_0$,

$$
\|\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\|_2 = O_p(n^{-1/3}).
$$

The above claim is shown in the following. For any unit vector u, let $\boldsymbol{\beta}(s) = \boldsymbol{\beta}_0(s) + n^{-1/3}\boldsymbol{u}$. For convenience, let $\mathcal{T}$ denote the event

$$
\Big\{\max_{s\in\mathcal{A}_0, \|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\tau H_{n0}(\boldsymbol{\beta}(s))\,\boldsymbol{u} \le Cp_{0n}n^{2/3}\Big\}.
$$

From the proof of Lemma 2.1, we have $P(\mathcal{T}) \to 1$. Thus,

$$
\begin{aligned}
&P\left(\ln L_n(\boldsymbol{\beta}(s)) - \ln L_n(\boldsymbol{\beta}_0(s)) > 0,\ \ s\in\mathcal{A}_0\right) \\
&\quad = P\left(\ln L_n(\boldsymbol{\beta}(s)) - \ln L_n(\boldsymbol{\beta}_0(s)) > 0,\ \ s\in\mathcal{A}_0|\mathcal{T}\right)P(\mathcal{T}) \\
&\qquad + P\left(\ln L_n(\boldsymbol{\beta}(s)) - \ln L_n(\boldsymbol{\beta}_0(s)) > 0,\right. \\
&\qquad \left. s\in\mathcal{A}_0|\mathcal{T}^c\right)P(\mathcal{T}^c) \\
&\quad \le P\left(\ln L_n(\boldsymbol{\beta}(s)) - \ln L_n(\boldsymbol{\beta}_0(s)) > 0,\ \ s\in\mathcal{A}_0|\mathcal{T}\right) + o(1).
\end{aligned}
$$

On $\mathcal{T}$, when n is large enough, for all $s \in \mathcal{A}_0$, uniformly, we have

$$
\begin{aligned}
&\ln L_n(\boldsymbol{\beta}(s)) - \ln L_n(\boldsymbol{\beta}_0(s)) \\
&\quad = n^{-1/3}\boldsymbol{u}^\tau s_n(\boldsymbol{\beta}_0(s)) - \frac{1}{2}n^{1/3}\boldsymbol{u}^\tau\left(n^{-1}H_{n1}\left(\tilde{\boldsymbol{\beta}}(s)\right)\right)\boldsymbol{u} \\
&\qquad - \frac{1}{2}n^{-2/3}\left(\boldsymbol{u}^\tau H_{n0}\left(\tilde{\boldsymbol{\beta}}(s)\right)\boldsymbol{u}\right) \\
&\quad = n^{-1/3}\boldsymbol{u}^\tau s_n(\boldsymbol{\beta}_0(s)) - c_1(1-\xi)n^{1/3}/2 + O(p_{0n}) \\
&\quad \le n^{-1/3}\boldsymbol{u}^\tau s_n(\boldsymbol{\beta}_0(s)) - cn^{1/3},
\end{aligned}
$$

where $\tilde{\boldsymbol{\beta}}(s)$ is between $\boldsymbol{\beta}(s)$ and $\boldsymbol{\beta}_0(s)$ component-wise. The first equality is the Taylor expansion and the second equality follows from C5. Hence, for some positive constant c, we have

$$
\begin{aligned}
&P\big(\ln L_n\big(\boldsymbol{\beta}(s)\big) - \ln L_n\big(\boldsymbol{\beta}_0(s)\big) > 0 :\ \text{for some}\ \boldsymbol{u}\big) \\
&\quad \le P\big(\boldsymbol{u}^\tau s_n\big(\boldsymbol{\beta}_0(s)\big) \ge cn^{2/3} :\ \text{for some}\ \boldsymbol{u}\big) \\
&\quad \le \sum_{j\in s} P\big(s_{n,j}\big(\boldsymbol{\beta}_0(s)\big) \ge cn^{2/3}\big) \\
&\qquad + \sum_{j\in s} P\big(-s_{n,j}\big(\boldsymbol{\beta}_0(s)\big) \ge cn^{2/3}\big)
\end{aligned}
$$

From (12), we know that

$$
\sum_{i\in\mathcal{A}_0}\sum_{j\in s} P\big(s_{n,j}\big(\boldsymbol{\beta}_0(s)\big) \ge cn^{2/3}\big) = o(1).
$$

The same for the second term. Therefore,

$$
\begin{aligned}
&P\left(\ln L_n(\boldsymbol{\beta}(s)) - \ln L_n(\boldsymbol{\beta}_0(s)) > 0 :\right. \\
&\qquad \left.\text{for some}\ \boldsymbol{u}, s\in\mathcal{A}_0\right) = o(1).
\end{aligned} \tag{17}
$$

Because $\ln L_n(\boldsymbol{\beta}(s))$ is a concave function for any $\boldsymbol{\beta}(s)$, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}(s)$ exists and falls within a $n^{-1/3}$ neighborhood of $\boldsymbol{\beta}_0(s)$ uniformly for $s \in \mathcal{A}_0$. Thus, we have $P(\|\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\|_2 = O(n^{-1/3})) \to 1$.

Now applying C5, the right-hand side of (16) is bounded by

$$
\begin{aligned}
&(\hat{\boldsymbol{\beta}}(s)-\boldsymbol{\beta}_0(s))^\tau s_n(\boldsymbol{\beta}_0(s)) - \frac{(1-\xi)(1-\epsilon)}{2}(\hat{\boldsymbol{\beta}}(s)-\boldsymbol{\beta}_0(s))^\tau \\
&\quad \times H_{n1}(\boldsymbol{\beta}_0(s))(\hat{\boldsymbol{\beta}}(s)-\boldsymbol{\beta}_0(s)) \\
&\quad \le \frac{1}{2(1-\epsilon)} s_n^\tau(\boldsymbol{\beta}_0(s))\{H_{n1}(\boldsymbol{\beta}_0(s))\}^{-1} s_n(\boldsymbol{\beta}_0(s))
\end{aligned}
$$

where ϵ is an arbitrarily small positive value. Hence,

$$
\begin{aligned}
&P\Bigg(\sup_{s\in\mathcal{A}_0} \ln L_n(\hat{\boldsymbol{\beta}}(s)) - \ln L_n(\hat{\boldsymbol{\beta}}(s_{0n})) \\
&\quad \ge m[0.5\ln n + \gamma \ln p_n]\Bigg) \\
&\quad \le P\Bigg(\frac{s_n^\tau(\boldsymbol{\beta}_0(s))\{H_{n1}(\boldsymbol{\beta}_0(s))\}^{-1} s_n(\boldsymbol{\beta}_0(s))}{2(1-\epsilon)} \\
&\qquad\qquad \ge m[0.5\ln n + \gamma \ln p_n]\Bigg) \\
&\quad \le |\mathcal{A}_0| \exp(-m(1-\epsilon)[0.5\ln n + \gamma \ln p_n]) \\
&\quad \le \exp\left(m\left[(\ln(p_n - p_{0n}) - (1-\epsilon)\gamma \ln p_n - \frac{(1-\epsilon)}{2}\ln n\right]\right) \\
&\quad \to 0, \quad \text{if } \gamma > \frac{1}{1-\epsilon}\left[1 - \frac{\ln n}{2\ln p_n}\right].
\end{aligned}
$$

That is, (15) is proved. Hence, the theorem is proved. □

ACKNOWLEDGEMENTS

The research of the authors is supported by the AcRF Tier 1 Grant R-155-000-091-112 of the National University of Singapore.

Received 13 December 2011

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, Akademiai Kiado, 267–281. MR0483125

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

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica* **22** 555–574.

Czado, C. and Munk, A. (2000). Noncanonical links in generalized linear models – when is the effort justified? *Journal of Statistical Planning and Inference* **87** 317–345. MR1771122

Fan, J. Q. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc. B* **96** 1348–1360. MR1946581

Fan, J. Q. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

Huang, J., Horowitz, J. L., and Wei, F. R. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38**(4) 2282–2313. MR2676890

Lee, K. E., Sha, N. J., Dougherty, E. R., Vannucci, M. and Mallick, B. K. (2003). Gene selection: A Bayesian variable selection approach. *Bioinformatices* **19** 90–97.

Li, W. and Chen, Z. (2009). Multiple interval QTL mapping for trait distribution with a spike. *Genetics* **182** 337–342.

Liao, J. G. and Chin, K. V. (2007). Logistic regression for disease classification using microarray data: Model selection in a large p small n case. *Bioinformatics* **23** 1945–1951.

Luo, S. and Chen, Z. (2011). Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. arXiv:1107.2502v1 [math.ST].

Luo, S. and Chen, Z. (2011). Sequential Lasso for feature selection with ultra-high dimensional feature space. arXiv:1107.2734v1 [math.ST].

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, Second Edition. Chapman and Hall, London. MR0727836

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc. B* **39** 111–147. MR0356377

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B* **58** 267–288. MR1379242

Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory* **50** 2231–2242. MR2097044

Tropp, J. A. and Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory* **53** 4655–666. MR2446929

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Am. Statist. Assoc.* **104** 1512–1524. MR2750576

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc. B* **71** 671–683. MR2749913

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63** 27–32. MR0408092

Zhao, J. and Chen, Z. (2012). A two-stage penalized logistic regression approach to case-control genome-wide association studies. *Journal of Probability and Statistics*. doi:10.1155/2012/642403. MR2862471

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**(2) 301–320. MR2137327

Shan Luo
Department of Statistics and Applied Probability
National University of Singapore
Singapore
E-mail address: luoshan08@nus.edu.sg

Zehua Chen
Department of Statistics and Applied Probability
National University of Singapore
Singapore
E-mail address: stachenz@nus.edu.sg