EXTENDED BIC FOR SMALL-n-LARGE-P SPARSE GLM

Jiahua Chen and Zehua Chen

University of British Columbia and National University of Singapore

Abstract: 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. In this paper, we examine the variable selection problem under the generalized linear models. We study the approach where a prior takes specific account of the small-n-large-P situation. The criterion is shown to be variable selection consistent under generalized linear models. We also report simulation results and a data analysis to illustrate the effectiveness of EBIC for feature selection.

Key words and phrases: Consistency, exponential family, extended Bayes information criterion, feature selection, generalized linear model, small-n-large-P.

1. Introduction

In many scientific investigations, researchers explore the relationship between a response variable and some explanatory features through a random sample. 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 in this. When the sample size n is relatively small but the number of features P under consideration is extremely large, there is 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 when the prior is

uniform on the model space. 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 and Speed (2002), Siegmund (2004), and Bogdan, Doerge and Ghosh (2004) in genetic studies. See also Donoho (2000), Singh et al. (2002), Marchini, Donnelly, and Cardon (2005), Clayton et al. (2005), Fan and Li (2006), Zhang and Huang (2008), and Hoh, Wille, and Ott (2008). Some recent BIC related model selection procedures in new situations can be found in Wang, Li, and Tsai (2007), Jiang et al. (2008) and many others.

Recently, Chen and Chen (2008) pointed out that the uniform prior on the model space is the cause of BIC's liberality in small-n-large-P situation. Correction of this problem leads to a family of extended Bayes information criteria (EBIC). Bogdan, Doerge and Ghosh (2004) made the same observation but provided slightly different correction measures. Mathematically, the EBIC is the classical BIC with an additional penalty term $2\gamma \log P$ with a positive γ. Interestingly, Foster and George (1994) found that, instead of adding the $\log P$ term, simply replacing $\log n$ with $2 \log P$ in BIC gives empirically optimal results in view of risk inflation, and their finding was echoed in Abramovich et al. (2006).

The EBIC is shown to be selection consistent in the small-n-large-P framework under the normal linear model. Its validity under a wide class of regression models remains an unsolved problem. In this paper, we have tailor-developed technical results for exponential family distributions which are of interest in themselves. 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 selection consistency of EBIC under GLM with canonical links (Theorem 2). In the implementation, we need to place an upper bound K on the number of causal features. If K is chosen too small in an application, the selection consistency of the procedure with EBIC may not be realized. To tackle this issue, we show that if K is chosen too small the EBIC will select a model exhausting all K features. If the selected model exhausts all the K features allocated, reanalyzing the data with a larger K is suggested.

We have also investigated the performance of EBIC by simulation under the logistic regression model and the Poisson log-linear model. The logistic regression model is valid in both prospective and retrospective studies, see McCullagh and Nelder (1989, Chap. 4), and is a major approach in genetic research, see for example The Wellcome Trust Case-Control Consortium (2007). In principle, EBIC is an all subsets method which is computationally infeasible. Our implementation strategy for EBIC follows that of Wang, Li, and Tsai (2007) and Zhang, Li, and Tsai (2010). We use regularization methods such as LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)) or Elastic Net (Zou and Hastie (2005)) to

obtain regression models with various levels of sparsity. Because they only determine the order of the penalty level for selection consistency, some cross validation procedure is ultimately used to select the final model. Replacing the computer intensive cross validation procedure by EBIC creates a promising new approach. In simulation, we used R packages (R Development Core Team (2010)) `glmpath` (Park and Hastie (2007)) and `glmnet` (Friedman, Hastie, and Tibshirani (2010)). designed for LASSO and Elastic Net.

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. A data example is analyzed in Section 5, and we put some technical proofs and other information in an Appendix.

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 are called features). A GLM consists of three components. The first component is an exponential family distribution assumed for Y, with density function

$$f(y; \theta) = \exp\{\theta^\tau y - b(\theta)\} \qquad (2.1)$$

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 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 $g(\mu) = \eta = \boldsymbol{x}^\tau \boldsymbol{\beta}$.

We investigate the feature selection problems given a random sample $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ with two characteristics: (i) small-n-large-P, the number of features is much larger than the sample size; and (ii) sparsity, only a few un-identified features affect Y. We refer to a GLM with these two characteristics as the small-n-large-P sparse GLM.

The GLM implies a restrictive relationship between μ and σ^2. A more flexible model is $f(y; \theta, \phi) = \exp\{\phi^{-1}[\theta^\tau y - b(\theta)] + c(y, \phi^{-1})\}$ for some function $c(y, \phi^{-1})$ and dispersion parameter ϕ. When the value of ϕ is assumed, this model reduces to the usual GLM. Otherwise, ϕ must be estimated together with μ or θ. Nevertheless, the EBIC procedure discussed in the next section can be directly implemented. The asymptotic properties of the GLM and the selection consistency of EBIC must be re-established for the new model. For the special normal linear regression model where $\phi = \sigma^2$, selection consistency and other properties of the EBIC are given in Chen and Chen (2008).

Let $\mathcal{X}$ be the set of all features under consideration. Let s be a subset of $\mathcal{X}$, $\nu(s)$ the number of features in s, and $\boldsymbol{\beta}(s)$ the vector of the components in $\boldsymbol{\beta}$ that corresponds to the features in s. Let $s_0 \in \mathcal{X}$ be the subset of causal features that contains and only contains all the features affecting Y. Let $\boldsymbol{\beta}_0$ be the unknown true value of the parameters. The components of $\boldsymbol{\beta}_0$ other than those in s_0 are zero. Let $\boldsymbol{x}_i(s)$ be the vector of the components of $\boldsymbol{x}_i$ that correspond to $\boldsymbol{\beta}(s)$. Let

$$\mathcal{B}(s) = \{\boldsymbol{\beta} : \boldsymbol{x}_i(s)^\tau \boldsymbol{\beta}(s) \in \Theta, i = 1, \dots, n\}.$$

Note that $\mathcal{B}(s)$ is convex. Let l_n be the log likelihood function $l_n(\beta) = \sum_{i=1}^n \log f(y_i; \theta_i)$, 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. We consider only the canonical link $g(\mu_i) = \theta_i$. 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

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i \boldsymbol{x}_i^\tau \boldsymbol{\beta} - b(\boldsymbol{x}_i^\tau \boldsymbol{\beta})],$$

$$\boldsymbol{s}_n(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i - b'(\boldsymbol{x}_i^\tau \boldsymbol{\beta})] \boldsymbol{x}_i,$$

$$H_n(\boldsymbol{\beta}) = \sum_{i=1}^n b''(\boldsymbol{x}_i^\tau \boldsymbol{\beta}) \boldsymbol{x}_i \boldsymbol{x}_i^\tau.$$

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

Let $A_0 = \{s : s_0 \subset s; \nu(s) \leq K\}$, and $A_1 = \{s : s_0 \not\subset s; \nu(s) \leq K\}$. We allow the composition of s_0 and therefore A_0 and A_1 vary in the limiting process. The asymptotic results are established when $n \to \infty$, with s_0, $\boldsymbol{\beta}_0$ and other subjects evolving in an orderly fashion as specified in the following conditions. We assume that features are standardized. Our results are rigorously established for fixed

K, while providing insight on what happens when $K = O(\log n)$ to avoid overly tedious technical specifications.

A1. As $n \to \infty$, $P = O(\exp(n^{\kappa}))$ for some constant $0 < \kappa < 1/3$.

A2. $\inf \min\{|\beta_0(j)| : j \in s_0\} > n^{-1/4}$.

A3. The interior of $\mathcal{B}(s)$ is not empty, and $\boldsymbol{\beta}_0(s) \in \mathcal{B}(s)$ for all $s \in A_0 \cup A_1$.

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

$$c_1 \leq \lambda_{\min}(n^{-1}H_n(\boldsymbol{\beta}_0(s \cup s_0))) \leq \lambda_{\max}(n^{-1}H_n(\boldsymbol{\beta}_0(s \cup s_0))) \leq c_2,$$

for all $s \in A_1$, 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,

$$(1 - \epsilon)H_n(\boldsymbol{\beta}_0(s \cup s_0)) \leq H_n(\boldsymbol{\beta}(s \cup s_0)) \leq (1 + \epsilon)H_n(\boldsymbol{\beta}_0(s \cup s_0))$$

for all $s \in A_1$ and $\boldsymbol{\beta}(s \cup s_0)$, and $\|\boldsymbol{\beta}(s \cup s_0) - \boldsymbol{\beta}_0(s \cup s_0)\| \leq \delta$.

A6. With x_{ij} the jth component of $\boldsymbol{x}_i$, there exists a positive constant C such that $|x_{ij}| \leq C$ and

$$\max_{1 \leq i \leq n} \left\{ \frac{x_{ij}^2}{\sum_{i=1}^n x_{ij}^2 \sigma_i^2} \right\} \leq \frac{Cn^{-1/6}}{\log n}$$

for all $1 \leq j \leq P$, all n sufficiently large.

Remark. Condition A4 is similar to the UUP condition given by Candes and Tao (2007), which might be too restrictive. For instance, if two features in s but not in s_0 are collinear, A4 is violated. The selection consistency is likely valid under weaker conditions than A4 but we have yet to identify such. We also suspect that the proof under weaker conditions would be lengthy. Condition A5 extends A4 to a small neighborhood of β_0. These two require the true model to stay at some distance from wrong models as n increases. A6 can be violated only if the square of a feature has a severely skewed distribution, for instance, when a binary feature has less than $\log n$ 1's in n observations. However, such features would have readily been screened out before a variable selection procedure is applied. The above conditions are placed on $\boldsymbol{x}$'s as if they were not random. For random $\boldsymbol{x}$, these properties are usually satisfied in probability due to the Law of Large Numbers.

Let $\hat{\boldsymbol{\beta}}(s)$ be the MLE of $\boldsymbol{\beta}(s)$ in the GLM with features in s. We first have a uniform consistency result of $\hat{\boldsymbol{\beta}}(s)$ for $s \in A_0$. The proof is given in the Appendix.

Theorem 1. *Under A1-A6 with $n \to \infty$,*

$$\max_{s \in A_0} \|\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\| = O_p(n^{-1/3}).$$

Remark: (a) The result is unaffected if β_0 depends on n. (b) Condition A4 is crucial for the validity of this result. For, suppose there exist two completely collinear features, say $\boldsymbol{x}_1 = \boldsymbol{x}_2$, then the likelihood values at $\boldsymbol{\beta}_0(s)$ and $\boldsymbol{\beta}_1(s) = \boldsymbol{\beta}_0(s) + (\alpha, -\alpha, 0, \ldots 0)$ are identical for any α and this clearly invalidates our result. Since $\boldsymbol{\beta}_1(s)$ is not as sparse as $\boldsymbol{\beta}_0(s)$, the EBIC likely screens it out of consideration. Hence, the selection consistency of the EBIC (to be established) may still be true though the proof is beyond our reach at the moment. (c) As long as $Kn^\kappa = o(n^{1/3})$, the proof of this result remains valid, that is, the theoretical result remains true when K increases with n albeit at a low rate, for instance $K = O(\log n)$. However, Condition A4 with very large K becomes more restrictive.

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 that are not necessarily causal. Chen and Chen (2008) have recently proposed a family of extended Bayes information criteria (EBIC). In EBIC, models are classified according to the number of features they contain, and the prior probability assigned to a model is inversely proportional to the size of the model class to which the model belongs. EBIC for model s is

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

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. The selection consistency leads to that FDR converges to 0 and PSR converges 1 simultaneously as n goes to infinity. Simulation results indicate that EBIC with γ in the above consistency range effectively keeps the FDR low while achieving a reasonable PSR.

We now state the consistency of EBIC under generalized linear models with canonical links. Its proof is deferred to the Appendix.

Theorem 2. *Under A1−A6 with $n \to \infty$, we have*

$$P\{\min_{s \in A_1} EBIC(s) \leq EBIC(s_0)\} \to 0, \tag{3.1}$$

for any $\gamma > 0$;

$$P\{\min_{s \in A_0, s \neq s_0} EBIC(s) \leq EBIC(s_0)\} \to 0, \qquad (3.2)$$

for $\gamma > 1 - \log n/(2 \log P)$.

As long as $Kn^\kappa = o(n^{1/3})$, for instance when $K = O(\log n)$, the above results remain valid. In applications, one must choose a K to start the process. If K is less than the cardinality of s_0, Theorem 3 below shows that the EBIC selects almost surely the model that exhausts all K retained features.

We make some preparation. Let

$$G_n(k) = \sup\{\sum_{i=1}^n [\theta_i b'(\theta_{0i}) - b(\theta_i)] : \theta_i = x_i(s)^\tau \beta(s); \ \nu(s) = k\},$$

where $\theta_{0i} = x_i^\tau \beta_0$ is the true parameter value and $\theta_i b'(\theta_{0i}) - b(\theta_i) = E\{\log f(y_i; \theta_i)\}$. Because of the convexity of $b(\cdot)$, the supremum $G_n(k)$ is attained at some β_k that has k non-zero components. We replace conditions $A4$ and $A5$ by

> $A4'$ $c_1 < \lambda_{\min}(n^{-1}H_n(\beta(s)))$ for all $\nu(s) < 2\nu(s_0)$ and some $c_1 > 0$.
>
> $A5'$ $n^{1/3}\|\beta_k - \beta_{k+1}\|^2 \to \infty$ as $n \to \infty$.

Condition $A4'$ is slightly more restrictive than $A4$. Both have potential to be weakened but we have yet to find substitutes. Condition $A5'$ is not restrictive though hard to grasp. If $n^{-1}\sum_{i=1}^n [\theta_i b'(\theta_{0i}) - b(\theta_i)]$ has a non-degenerate and smooth limit, then β_k should converge to the maximum point of this limit. Having different number of non-zero elements, β_k and β_{k+1} should be different, which leads to $A5'$. The technical conditions to ensure these results could be lengthy.

Theorem 3. *Under A1−A3, A4', A5' and A6 with $n \to \infty$,*

$$P\{EBIC(k) > EBIC(k + 1)\} \to 1, \qquad (3.3)$$

where $EBIC(k) = \min\{EBIC(s) : \nu(s) = k\}$ for any $\gamma > 0$ and $k < \nu(s_0)$.

The proof is given in the Appendix. Theorem 3 states that EBIC(k) is monotone decreasing when $k < \nu(s_0)$ and provides an adaptive strategy for the choice of K. We start with a generous upper bound of the number of causal features according to the scientific background and the feasibility of the current sample size. If the selected model does not exhaust all K retained features, K is then a right choice. Otherwise, we choose a larger K and repeat the procedure until K is large enough that the selected final model does not exhaust all the retained features.

4. Simulations

We present simulation studies for assessing the performance of EBIC under GLM. They were conducted under the logistic regression model and the Poisson log-linear model.

We first consider simulations 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 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 $j \notin s_0$, x_{ij} values were generated under the assumption of Hardy-Weinberg equilibrium; that is, they were simulated from a binomial distribution with parameters $(2, p_j)$, p_j the allele frequency of an allele for the jth SNP. The allele frequency p_j was not fixed but was 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 concentrated mainly around .5, but moderately spread out to .1 and .9. Given p_j, $x_{ij} : i = 1, \ldots, n = n_1 + n_2$ were independently generated.

For $j \in s_0$, the x_{ij} values were generated in the same way as in control group. At the same time, from

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

we get

$$P(X(s_0) = \boldsymbol{x}(s_0)|Y = 1) = P(X(s_0) = \boldsymbol{x}(s_0)|Y = 0)\exp(\alpha^* + \boldsymbol{x}^\tau(s_0)\boldsymbol{\beta}_0),\quad (4.1)$$

where α^* is the normalization factor.

The conditional probability $P(X(s_0) = \boldsymbol{x}(s_0)|Y = 0)$ can be reasonably specified such that it decreases as the number of disease alleles of the SNPs involved increases, or simply taken as a constant. Once $P(X(s_0) = \boldsymbol{x}(s_0)|Y = 0)$ is specified, the conditional probabilities given by (4.1) up to the normalization factor are computed. After having been normalized, these probabilities are used in an R program to sample from the set of all possible $\boldsymbol{x}(s_0)$'s. When s_0 contains m variables, there are 3^m possible $\boldsymbol{x}(s_0)$ distinct vectors of dimension m with $-1, 0, 1$ entries. The vectors for the n_2 cases are sampled with replacement.

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 Table 1.

With $P \leq 10,000$, the `glmnet` function was directly applied to identify a sequence of ordered SNPs, denoted by s, of length no more than $K = 40$.

Table 1. Model specifications.

Model	m	P	β_0
1	2	500	(0.5, 0.7)
2	2	500	(0.3, 0.5)
3	3	500	(0.5, 0.6, 0.7)
4	3	500	(0.3, 0.4, 0.5)
5	5	500	(0.3, 0.4, 0.5, 0.6, 0.7)
6	8	500	(0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)
7	2	1,000	(0.5, 0.7)
8	2	1,000	(0.3, 0.5)
9	3	1,000	(0.5, 0.6, 0.7)
10	3	1,000	(0.3, 0.4, 0.5)
11	5	1,000	(0.3, 0.4, 0.5, 0.6, 0.7)
12	8	1,000	(0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)
13	3	10,000	(0.6, 0.7, 0.8)
14	3	10,000	(0.4, 0.5, 0.6)
15	5	10,000	(0.3, 0.4, 0.5, 0.6, 0.7)
16	8	10,000	(0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)
17	3	100,000	(10.3, 10.4, 10.5)
18	3	100,000	(0.5, 0.6, 0.7)

We then used `glm.fit` to evaluate the submodels (at most 40 of them) formed by the first k variables in the sequence for $k = 1, \ldots, K$. The BIC and EBIC values were computed and the submodels minimizing BIC or EBIC were selected. Because `glmnet` does not give complete sample path, some intermediate models were not included. In comparison, `glmpath` provides complete sample path but is computationally less efficient. The simulation results based on `glmpath` for $P \leq 10,000$ were similar and are omitted here because our purpose is to compare information criteria, not numerical methods.

With $P = 100,000$, the computer does not have large enough memory to store the design matrix to be screened by `glmnet`. We randomly divided them into groups of size 1,000. We used `glmnet` to select 30+ features from each group, and pooled them together to be further screened. Using several rounds of selection when P is very large was proposed and investigated in Chen and Chen (2008).

To shed light on the appropriate size of γ, we present the results obtained by taking $\gamma = 0, .25, .5, 1$. Note that the ordinary BIC is a special form of EBIC with $\gamma = 0$. The number of simulation replicates is $N = 500$. 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 2. The average PSR and FDR are defined as follows. Let s_0 be the set of causal

Table 2. Simulation results under Logistic Regression Model (FDR, PSR, ν^*)

Model	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 1.0$
1	(0.60, 0.95, 6.73)	(0.27, 0.88, 3.05)	(0.13, 0.80, 2.10)	(0.03, 0.67, 1.42)
2	(0.60, 0.76, 5.58)	(0.26, 0.61, 2.11)	(0.07, 0.44, 1.06)	(0.02, 0.29, 0.62)
3	(0.53, 0.95, 7.75)	(0.24, 0.89, 4.17)	(0.10, 0.83, 3.00)	(0.02, 0.70, 2.20)
4	(0.56, 0.78, 6.82)	(0.22, 0.64, 2.87)	(0.07, 0.49, 1.71)	(0.01, 0.29, 0.90)
5	(0.45, 0.86, 9.25)	(0.18, 0.75, 5.05)	(0.07, 0.66, 3.71)	(0.01, 0.48, 2.47)
6	(0.39, 0.80, 11.96)	(0.15, 0.71, 7.27)	(0.07, 0.64, 5.74)	(0.01, 0.52, 4.25)
7	(0.73, 0.96, 1.06)	(0.32, 0.88, 3.50)	(0.13, 0.78, 2.08)	(0.03, 0.63, 1.35)
8	(0.74, 0.76, 9.08)	(0.31, 0.57, 2.33)	(0.10, 0.42, 1.08)	(0.01, 0.26, 0.54)
9	(0.67, 0.96, 11.68)	(0.28, 0.89, 4.48)	(0.11, 0.80, 2.93)	(0.02, 0.67, 2.10)
10	(0.70, 0.79, 1.37)	(0.26, 0.60, 2.90)	(0.09, 0.46, 1.67)	(0.01, 0.26, 0.80)
11	(0.59, 0.87, 12.64)	(0.24, 0.76, 5.70)	(0.08, 0.63, 3.61)	(0.01, 0.46, 2.33)
12	(0.51, 0.81, 15.14)	(0.20, 0.71, 7.77)	(0.08, 0.62, 5.64)	(0.02, 0.49, 4.07)
13	(0.84, 1.00, 2.51)	(0.60, 1.00, 1.37)	(0.34, 0.96, 6.03)	(0.15, 0.89, 3.73)
14	(0.90, 1.00, 31.84)	(0.61, 0.99, 11.20)	(0.34, 0.96, 5.95)	(0.13, 0.89, 3.57)
15	(0.82, 1.00, 3.28)	(0.51, 0.99, 13.22)	(0.29, 0.95, 8.30)	(0.10, 0.86, 5.25)
16	(0.66, 0.99, 25.19)	(0.37, 0.97, 14.50)	(0.19, 0.92, 1.12)	(0.06, 0.83, 7.36)
17	(0.89, 1.00, 29.00)	(0.76, 0.99, 16.16)	(0.53, 0.97, 9.12)	(0.26, 0.90, 4.69)
18	(0.90, 0.99, 3.22)	(0.77, 0.99, 16.15)	(0.47, 0.96, 7.54)	(0.15, 0.86, 3.60)

features, and s_j^* the features selected in the jth replicate, $j = 1, \ldots, N$. Then

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

In Table 2, $\nu^* = (1/N) \sum_{j=1}^{N} \nu(s_j^*)$.

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. 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. The 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 2.

We used $K = 30$ and recorded the number of times all K features were exhausted. This was 181 in Models 1-16 and 534 in Models 17-18 for BIC. For

Table 3. Frequency of $\nu(\hat{s}) = K$ in 500 repetitions

K	m	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 1.0$
5	10	357	297	212	75
5	15	419	380	325	185
15	10	55	2	0	0
15	15	119	13	1	0
30	10	0	0	0	0
30	15	0	0	0	0

Table 4. Simulation results under the Poisson Log-linear Model (FDR, PSR, ν^*).

Model	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.50$	$\gamma = 1.0$
8	(0.72, 0.97, 7.97)	(0.30, 0.95, 3.06)	(0.07, 0.90, 2.01)	(0.00, 0.79, 1.58)
10	(0.63, 0.98, 8.88)	(0.22, 0.96, 3.97)	(0.06, 0.94, 3.03)	(0.00, 0.86, 2.57)
12	(0.36, 0.95, 12.56)	(0.10, 0.93, 8.35)	(0.02, 0.90, 7.36)	(0.00, 0.85, 6.82)
14	(0.85, 0.99, 21.16)	(0.49, 0.98, 6.42)	(0.10, 0.97, 3.38)	(0.00, 0.93, 2.78)
16	(0.65, 0.93, 22.24)	(0.24, 0.90, 9.77)	(0.05, 0.86, 7.29)	(0.00, 0.80, 6.43)
18	(0.89, 0.99, 28.30)	(0.65, 0.99, 9.26)	(0.17, 0.98, 3.77)	(0.00, 0.96, 2.89)

EBIC with $\gamma = 0.25$ the corresponding totals were $2, 57$, with $\gamma = 0.5, 1, 4$, and none with $\gamma = 1$. Apparently, the average model size would be even larger if BIC is accompanied with $K > 30$. At the same time, Theorem 3 correctly predicts that there is no need of increasing K for EBIC.

To examine the relevance of the asymptotic result in Theorem 3, we generated data from models with $m = \nu(s_0) = 10, 15$, $P = 500$, and let β be replicates of $(0.2, 0.3, 0.4, 0.5, 0.6)$. We applied EBICs with $K = 5, 15, 30$. The numbers of times all K features were exhausted are reported in Table 3. The simulation results clearly conform to the theoretical conclusions.

In Table 4, we report simulation results under the Poisson log-linear model. In this case, the response Y has Poisson distribution with mean in the form $x(s_0)\beta_0(s_0)$. We generated x the same way as in the logistic regression model for the control group. We used the same β_0, but divided by 2 because the task of identifying casual features was found much less challenging, in the pilot study. We only included models with $P \geq 1{,}000$ and smaller β_0.

The results were impressive for EBIC. Take the outcome of Model 8 as an example, EBIC$_{0.5}$ cut the false positive rate from 72% to 7% while the positive selection rate was only reduced from 97% to 90%. The most impressive case was for Model 18 where $P = 100{,}000$. EBIC$_{0.5}$ practically achieved the same positive selection rate but substantially reduced the false positive rate from 89% to 17%. In all cases, the average model sizes selected by EBIC$_{0.5}$ were close to the true model sizes, while BIC persistently selected too many features.

5. An Example

In Singh et al. (2002), the researchers measured $P = 6{,}033$ genes on each of $n = 102$ men, $n_1 = 50$ 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 6,033 genes. Efron (2008) proposed an empirical Bayes approach that puts the discriminant 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)); it shrinks the $\hat{\delta}_i$ values by increasing the value of a tuning parameter (shrinkage value λ) and the non-zero $\hat{\delta}_i$ values corresponding to a particular value of the tuning parameter are 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%.

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 $P = 6{,}033$ 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 samples of 20 of the 6,033 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.

With the powerful R software package `glmnet`, we easily identified the 10 most important genes. These genes are then ordered in importance 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), BIC and EBIC values (given in the third and fourth row) were computed. The delete-one cross-validation errors were also computed and are given in the fifth 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
BIC	113.6	98.8	89.2	88.7	82.8	81.1	77.7	64.3	62.1	62.9
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). When cross-validation was used as the criterion, an eight-gene model was selected with a cross-validation error rate of 7.8%. If the ordinary BIC is used, a nine-gene model was selected with a cross-validation error rate of 9.8%. The delete-one cross-validation is widely known to be too liberal, and the BIC is even worse in this example. EBIC selects a more parsimonious model but retains a low cross-validation error rate.

Acknowledgement

A. Appendix

We state and prove one technical lemma, provide proofs of Theorems 1−3 and give a pseudo-code in R for the convenience of potential users.

A.1. Lemma 1 and its proof

Lemma 1. *Let Y_i, $i = 1, \ldots, n$, be independent random variables following exponential family distributions of form (2.1) with natural parameters θ_i. Let μ_i and σ_i^2 denote the mean and variance of Y_i, respectively. Suppose that $\{\theta_i; i = 1, \ldots, n\}$ is contained in a compact subset of the natural parameter space Θ. Let a_{ni}, $i = 1, \ldots, n$, be real numbers such that $\sum_{i=1}^n a_{ni}^2 \sigma_i^2 = 1$, and $\max_{1 \le i \le n} \{|a_{ni}|\} = o(n^{-1/6})$. Then, for any $m = O(n^{1/3})$ and positive ϵ,*

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

when n is sufficiently large.

Remark. The constant ϵ does not depend on a particular $\{a_{ni}\}_{i=1}^n$ whenever a_{ni} have the same upper bound.

Proof. Let $q_n = \sqrt{2m}$. For any $t > 0$

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

$$= \exp\{\sum_{i=1}^{n} [b(\theta_i + a_{ni}t) - b(\theta_i) - \mu_i a_{ni}t] - q_n t\}$$

$$= \exp\left[-\frac{t^2}{2}\{1 + \sum_{i=1}^{n} a_{ni}^2(b''(\theta_i + a_{ni}\tilde{t}) - b''(\theta_i))\}\right]$$

for some $0 < \tilde{t} < t$. Let $t = q_n$ and, by the compactness of Θ, $|a_{ni}\tilde{t}| = o(1)$. Hence, there exists a generic constant ϵ such that

$$\sum_{i=1}^{n} a_{ni}^2 |b''(\theta_i + a_{ni}\tilde{t}) - b''(\theta_i)| \leq \epsilon.$$

Consequently,

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

Hence,

$$P(\sum_{i=1}^{n} a_{ni}(Y_i - \mu_i) > \sqrt{2m}) \leq \exp\{-m(1 - \epsilon)\}$$

for any $\epsilon > 0$ and sufficiently large n. This completes the proof.

A.2. Proof of Theorem 1

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, $\boldsymbol{\beta}(s)$ falls into the neighborhood of $\boldsymbol{\beta}_0(s)$ so that A4 and A5 are applicable. Thus, for all $s \in A_0$,

$$l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s)) = n^{-1/3}\boldsymbol{u}^\tau \boldsymbol{s}_n(\boldsymbol{\beta}_0(s)) - \frac{1}{2}n^{1/3}\boldsymbol{u}^\tau\{n^{-1}H_n(\tilde{\boldsymbol{\beta}}(s))\}\boldsymbol{u}$$

$$\leq n^{-1/3}\boldsymbol{u}^\tau \boldsymbol{s}_n(\boldsymbol{\beta}_0(s)) - c_1(1 - \epsilon)n^{1/3}.$$

Hence, for some generic positive constant c,

$$P\{l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s)) > 0 : \text{for some } \boldsymbol{u}\}$$

$$\leq P\{\boldsymbol{u}^\tau \boldsymbol{s}_n(\boldsymbol{\beta}_0(s)) \geq cn^{2/3} : \text{ for some } \boldsymbol{u}\}$$

$$\leq \sum_{j \in s} P(s_{nj}^2(\boldsymbol{\beta}_0(s)) \geq cn^{4/3})$$

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

Note that

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

Let $B_n^2 = \sum_{i=1}^{n} x_{ij}^2 \sigma_i^2$ and $a_{ni} = B_n^{-1}x_{ij}$. By A6, we have $B_n^2 = O(n)$ and $\max|a_{ni}| = o(n^{-1/6})$. Under these conditions, Lemma 1 provides the inequality:

$$P(s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}) \leq P(\sum_{i=1}^{n} a_{ni}(Y_i - \mu_i) \geq \sqrt{2cn^{1/3}}) \leq \exp\{-cn^{1/3}\}.$$

The number of models in A_0 is no more than $P^K = \exp\{Kn^{\kappa}\} = \exp\{o(n^{1/3})\})$. Therefore,

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

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

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

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

A.3. Proof of Theorem 2

Proof of (3.1). Note that for any s, $\text{EBIC}(s) \leq \text{EBIC}(s_0)$ implies that

$$l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_0)) \geq \{\nu(s) - \nu(s_0)\}(\log n + 2\gamma \log P)$$
$$\geq -\nu(s_0)(\log n + 2\gamma \log P). \tag{A.1}$$

We show that the probability that (A.1) occurs at any $s \in A_1$ goes to 0. For any $s \in A_1$, let $\tilde{s} = s \cup s_0$. Consider those $\boldsymbol{\beta}(\tilde{s})$ near $\boldsymbol{\beta}_0(\tilde{s})$. We have

$$l_n(\boldsymbol{\beta}(\tilde{s})) - l_n(\boldsymbol{\beta}_0(\tilde{s})) = \{\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\}^{\tau} \boldsymbol{s}_n(\boldsymbol{\beta}_0(\tilde{s}))$$
$$-\frac{1}{2}\{\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\}^{\tau} H_n(\boldsymbol{\beta}^*(\tilde{s}))\{\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\}$$

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

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

Therefore,

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

Hence, for any $\beta(\tilde{s})$ such that $\|\beta(\tilde{s}) - \beta_0(\tilde{s})\| = n^{-1/4}$, we have

$$l_n(\beta(\tilde{s})) - l_n(\beta_0(\tilde{s})) \leq n^{-1/4}\|s_n(\beta_0(\tilde{s}))\| - \frac{c_1}{2}(1-\epsilon)n^{1/2}.$$

By Lemma 1, $\max_{s \in A_1} \|s_n(\beta_0(\tilde{s}))\| = O_p(\sqrt{nm})$, where $m = o(n^{1/3})$. Therefore,

$$\max\{l_n(\beta(\tilde{s})) - l_n(\beta_0(\tilde{s})) : s \in A_1, \|\beta(\tilde{s}) - \beta_0(\tilde{s})\| = n^{-1/4}\}$$
$$\leq c\{n^{-1/4}(nm)^{1/2} - n^{1/2}\} \leq c(n^{5/12} - n^{1/2}) \leq -cn^{1/2}$$

for a generic positive constant c in probability.

Because $l_n(\beta(\tilde{s}))$ is concave in $\beta(\tilde{s})$, the above result implies that the maximum of $l_n(\beta(\tilde{s}))$ is attained inside $\|\beta(\tilde{s}) - \beta_0(\tilde{s})\| \leq n^{-1/4}$. The concavity also implies that

$$\sup\{l_n(\beta(\tilde{s})) - l_n(\beta_0(\tilde{s})) : s \in A_1, \|\beta(\tilde{s}) - \beta_0(\tilde{s})\| \geq n^{-1/4}\}$$
$$\leq \sup\{l_n(\beta(\tilde{s})) - l_n(\beta_0(\tilde{s})) : s \in A_1, \|\beta(\tilde{s}) - \beta_0(\tilde{s})\| = n^{-1/4}\}$$
$$\leq -cn^{1/2}. \tag{A.2}$$

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

$$\|\breve{\beta}(\tilde{s}) - \beta_0(\tilde{s})\| \geq \|\beta_0(s_0 - s)\| > n^{-1/4},$$

by A2. Therefore, uniformly over $s \in A_1$ and with probability tending to 1,

$$l_n(\hat{\beta}(s)) - l_n(\beta_0(s_0)) = l_n(\breve{\beta}(\tilde{s})) - l_n(\beta_0(\tilde{s})) \leq -cn^{1/2}.$$

Hence, the probability that (A.1) occurs at any $s \in A_1$ tends to 0 which is (3.1).

Proof of (3.2). For $s \in A_0$, let $k = \nu(s) - \nu(s_0)$. It suffices to consider a fixed k, since k takes only the values $1, \ldots, K - \nu(s_0)$. By definition, $\text{EBIC}(s) \leq \text{EBIC}(s_0)$ if and only if

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

We show that, uniformly in $s \in A_0$ with $\nu(s) = k$, this inequality does not occur. For large n,

$$l_n(\hat{\beta}(s)) - l_n(\hat{\beta}(s_0)) \leq l_n(\hat{\beta}(s)) - l_n(\beta_0(s))$$
$$\leq \{\hat{\beta}(s) - \beta_0(s)\}^{\tau} s_n(\beta_0(s)) - \frac{1-\epsilon}{2}\{\hat{\beta}(s) - \beta_0(s)\}^{\tau} H_n(\beta_0(s))\{\hat{\beta}(s) - \beta_0(s)\}$$
$$\leq \frac{1}{2(1-\epsilon)} s_n(\beta_0(s))^{\tau}\{H_n(\beta_0(s))\}^{-1} s_n(\beta_0(s))$$

Hence we show that, uniformly over $s \in A_0$ with $\dot{\nu}(s) = k$,

$$s_n(\beta_0(s))^{\tau}\{H_n(\beta_0(s))\}^{-1} s_n(\beta_0(s)) \geq 2(1-\epsilon)k\{0.5 \log n + \gamma \log P\}$$

occurs with diminishing probability. Note that $H_n^{-1/2} s_n(\beta_0(s))$ is a linear combination of $Y_i - \mu_i$ as specified in Lemma 1. Thus, applying Lemma 1, we have for each $s \in A_0$,

$$P\left[s_n(\beta_0(s))^\tau \{H_n(\beta_0(s))\}^{-1} s_n(\beta_0(s)) \geq 2k(1-\epsilon)(0.5 \log n + \gamma \log P) \right]$$
$$\leq \exp\{-k(1-\epsilon)(0.5 \log n + \gamma \log P)\}$$

with a generic but arbitrarily small $\epsilon > 0$. With the choice of γ as given, we have

$$\exp\{-k(1-\epsilon)(0.5 \log n + \gamma \log P)\} \leq P^{-k(1+\epsilon)}$$

with a different generic $\epsilon > 0$ on the right hand side. Since the number of models in A_0 is lower than P^k, we have shown that

$$P(s_n(\beta_0(s))^\tau \{H_n(\beta_0(s))\}^{-1} s_n(\beta_0(s)) \geq 2k(1-\epsilon)\{0.5 \log n + \gamma \log P\},$$
$$\text{any } s \in A_0) \to 0,$$

This completes the proof.

A.4. Proof of Theorem 3

Proof. Write $\theta_{0i} = x_i^\tau \beta_0$. For any $\theta_i = x_i^\tau \beta$, it is seen that

$$\sum_{i=1}^n \log\{f(y_i; \theta_i)\} = \sum_{i=1}^n \theta_i \{b(\theta_{0i}) - b(\theta_i)\} + \sum_{i=1}^n \theta_i \{y_i - b(\theta_{0i})\}$$
$$= \sum_{i=1}^n \theta_i \{b(\theta_{0i}) - b(\theta_i)\} + \left(\sum_{i=1}^n \theta_i^2 \sigma_i^2 \right)^{1/2} o_p(n^{1/3})$$

uniformly in $o_p(n^{1/3})$ over all β with at most K non-zero components, with the order assessment from Lemma 1. In particular, the probability of violation is smaller than $n^{-\kappa K}$.

Consequently, with $a_k = \sup_\theta \{\sum_{i=1}^n \theta_i^2 \sigma_i^2\}^{1/2}$, we have

$$\text{EBIC}(k) - \text{EBIC}(k+1) \geq 2\{G_n(k+1) - G_n(k)\} - (1 + 2\gamma\kappa + a_k + a_{k+1})o_p(n^{1/3}).$$

Under the compactness assumption on Θ, we have $a_k = O(n^{1/2})$. Hence, the inequality reveals that it suffices to show that

$$\frac{G_n(k+1) - G_n(k)}{n^{2/3}} \to \infty. \tag{A.3}$$

By the Mean Value Theorem, we easily obtain

$$G_n(k+1) - G_n(k) = (\beta_{k+1} - \beta_k)^\tau H_n(\tilde{\beta})(\beta_{k+1} - \beta_k)$$

for some $\tilde{\beta}$ which has at most $2k + 1$ nonzero components. Hence, (A.3) is valid under conditions A4' and A5'.

A.5. Pseudo R-code for EBIC

We provide a pseudo-code for those interested in using EBIC.

1. input/create design matrix xx; response vector y;

2. identify a sequence of models with at most K features:

   `output=glmnet(xx, y, family="binomial", alpha=0.99, pmax=K)`

3. identify features from the output, `aa[[k]]` contains features in kth model:

   `bb=abs(output$beta); bb[,1]=1:P; aa[[k]]=bb[bb[ ,k]>0, 1]`

4. re-calcuate the deviance of the kth model:

   `dev[k] = glm.fit(xx[, aa[[k]] ], y, family= binomial(),`
   `intercept=T)$deviance`

5. add $\nu[k](\log(n) + 2\gamma \log(P))$ to `dev[k]` with user's choice of γ to get $\text{EBIC}_\gamma[k]$. The recommended value is $\gamma = 0.5$.

6. Select the model and corresponding features that minimizes $\text{EBIC}_\gamma[k]$.

In most practical situations, one would examine the outcomes based on several choices of γ. After all, the subject matter has the final say.

References

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584-653.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second Int. Symp. Info. Theory* (Edited by B. N. Petrox and F. Caski), 267-281. Budapest: Akademiai Kiado.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate – A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-30.

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

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Statist. Soc. Ser. B* **64**, 641-656.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313-2351.

Chen, J. and 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. and 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. and 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. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-6.

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

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947-1975.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Soft.* **33**.

Hoh, J., Wille, A. and Ott, J. (2008). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research* **11**, 2115-2119.

Jiang, J., Rao, J. S., Gu, Z. and Nguyen T. (2008). Fence methods for mixed model selection. *Ann. Statist.* **36**, 1669-1692.

Marchini, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413-417.

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

Park, M. Y. and Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B* **69**, 659-677.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Siegmund, D. (2004). Model selection in irregular problems: Application to mapping quantitative trait loci. *Biometrika* **91**, 785-80.

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. and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-209.

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

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. Roy. Statist. Soc. Ser. B* **58**, 267-288.

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

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generlized information criterions. *J. Amer. Statist. Assoc.* **105**, 312-323.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-32.

Department of Statistics, University of British Columbia, 2329 West Mall Vancouver, B.C., Canada V6T 1Z4.

E-mail: jhchen@stat.ubc.ca

Department of Statistics and Applied Probability, National University of Singapore, 3 Science Drive 2, Singapore 117543, Republic of Singapore.

E-mail: stachenz@nus.edu.sg