Average (E)BIC-like Criteria for Bayesian Model Selection

Jingnan Xue, Ye Luo and Faming Liang*

June 1, 2017

Abstract

Markov chain Monte Carlo (MCMC) has been an indispensable tool for Bayesian analysis of complex statistical models even for high-dimensional problems. However, there still lacks a consistent criterion for selecting models based on the outputs of MCMC. The existing deviance information criterion (DIC) is known to be inconsistent and non-invariant for reparameterization. This paper proposes an *Average BIC-like* (ABIC) model selection criterion and an *Average EBIC-like* (AEBIC) model selection criterion for low and high-dimensional problems, respectively; establishes their consistency under mild conditions; and illustrates their applications using generalized linear models. The proposed criteria overcome shortcomings of DIC. The numerical results indicate that the proposed criteria can significantly outperform DIC as well as the MLE-based criteria, such as AIC, BIC and EBIC, in terms of model selection accuracy.

Keywords: Asymptotic Consistency; Deviance Information Criterion; High-Dimensional Data; Generalized Linear Model; Markov Chain Monte Carlo.

*To whom correspondence should be addressed: F. Liang. Liang is Professor, Department of Biostatistics, University of Florida, Gainesville, FL 32611, email: faliang@ufl.edu; J. Xue is Graduate Student, Department of Statistics, Texas A&M University, College station, TX 77843; email: jnxue@stat.tamu.edu; Y. Luo is Assistant Professor, Department of Economics, University of Florida, Gainesville, FL 32611, email:kurtluo@gmail.com.

1 Introduction

Model selection is a fundamental part of the statistical modeling process, and it has been an active research area since 1970s. The most famous model selection criteria are perhaps AIC(Akaike, 1974) and BIC(Schwarz, 1978). The former is commonly called Akaike Information Criterion after Hirotogu Akaike; and the latter is called Bayesian Information Criterion or Schwarz Information Criterion after Gideon Schwarz. Both criteria are boiled down to a trade-off between goodness-of-fit and model complexity: A good model fits the data well with a minimum number of variables/parameters. To be more precise, AIC is defined by

$$\mathrm{AIC}(S) = -2L_n(X_n|\widehat{\beta}_S) + 2|S|/n,$$

where $X_n = \{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ denote a set of n *iid* observations, S denotes a specific model, $|S|$ denotes the number of parameters of the model, β_s denotes the vector of parameters of the model, $\widehat{\beta}_S$ denotes the maximum likelihood estimate (MLE) of β_S, and $L_n(\cdot)$ denotes the averaged log-likelihood function, i.e.,

$$L_n(X_n|\beta_S) = \frac{1}{n}\sum_{i=1}^{n} \log f(x^{(i)}|\beta_S), \tag{1}$$

where $f(\cdot)$ denote the probability mass/density function of the model. BIC is defined by

$$\mathrm{BIC}(S) = -2L_n(X_n|\widehat{\beta}_S) + |S|\log(n)/n,$$

which penalizes model complexity more than AIC and so will favor simpler models.

AIC and BIC represent two very different approaches for model selection. AIC is essentially a data-driven criterion which is to select a model that explains the data well and will make good short-term predictions. The penalty term $2|S|$ has been shown to be asymptotically equivalent to leave-one-out cross-validation. Based on this observation, different cross-validation-based criteria have been proposed in the literature, see e.g., generalized cross validation (Wahba and Craven, 1979) and WAIC (Watanabe, 2010). AIC and the cross-validation-based criteria are not asymptotically consistent. BIC is derived under the Bayesian framework, for which the differences provide an approximation to the log(Bayes factor) under specific "unit information" prior assumptions (Kass and Raftery, 1995). Under specific conditions, BIC has been shown to be asymptotically consistent, see e.g. Nishii (1984) and Haughton (1988).

Since all these AIC, BIC and cross-validation criteria require an exhaustive evaluation of all possible models for the problem, they are difficult to be applied when the model space is large. When the time entered into 1990s, the use of Markov chain Monte Carlo (MCMC) made it possible to fit arbitrarily complex Bayesian models with an arbitrarily large model space. Note that the AIC, BIC and cross-validation criteria essentially fail for these problems. Hence, there was a pressing need for a model selection criterion that is able to select an appropriate model based on the outputs of MCMC. This need was partially addressed by

Deviance information criterion (Spiegelhalter et al., 2002), which is defined by

$$\mathrm{DIC}(S) = -2E_{\beta_S|X_n,S}L_n(X_n|\beta_S) + [-2\bar{L}(X_n|\beta_S) + 2L_n(X_n|\bar{\beta}_S)],$$

where $E_{\beta_S|X_n,S}$ denotes the expectation with respect to the posterior distribution of β_S for a given model S, and $\bar{\beta}_S = E_{\beta_S|X_n,S}(\beta_S)$ denotes the posterior mean of β_S. When the sample size n is sufficiently large, it can be shown that $[-2\bar{L}(X_n|\beta_S) + 2L_n(X_n|\bar{\beta}_S)] \approx |S|/n$, which measures the model complexity and is called "effective number of parameters" in Spiegelhalter et al. (2002), and $E_{\beta_S|X_n,S}L_n(X_n|\beta_S) \approx L_n(X_n|\widehat{\beta}_S)$. Therefore, DIC can be viewed as an approximation to AIC with posterior samples. Twelve years later, Spiegelhalter et al. (2014) acknowledged that DIC has some problems, e.g., it is not invariant to reparameterization, it is not asymptotically consistency, it is not based on a proper predictive criterion, and it has a weak theoretical justification.

To alleviate the difficulties suffered by DIC, we proposed a new model selection criterion, the so-called average BIC-like Criterion,

$$\mathrm{ABIC}(S) = -2nE_{\beta_S|X_n,S}L_n(X_n|\beta_S) + |S|\log(n), \tag{2}$$

which shares the same penalty for the model complexity with BIC and is abbreviated as "average-BIC" or "ABIC". Under appropriate conditions, we can show that ABIC is asymptotically consistent.

During the past two decades, the research in model selection has been reflourished due to the emergence of high-dimensional data where the sample size n is typically smaller than the dimension p. To address the ill-posed issue in modeling high-dimensional data, a variety of variable selection methods have been proposed, in particular, for linear and generalized linear models. To name a few, they include Lasso(Tibshirani, 1994), SCAD(Fan and Li, 2001), elastic net (Zou and Hastie, 2005), Dantzig selector (Candes and Tao, 2007), MCP (Zhang, 2010), rLasso (Song and Liang, 2015a), and EBIC (Chen and Chen, 2008, 2012; Chen and Luo, 2013). Like AIC and BIC, all these methods require to find the global optimizer for a penalized log-likelihood function. However, as shown by Ge et al. (2016) and Huo and Chen (2010), except for Lasso, finding the global optimizer for the other methods is strongly NP-hard due to the concaveness of their penalty functions. Therefore, the models selected by these methods are often sub-optimal. Although Lasso is computationally attractive, it is often inconsistent as the required representative condition (Zhao and Yu, 2006) is hard to be satisfied. Practically, Lasso tends to select a model including many false variables.

On the other hand, MCMC continue to work as an indispensable tool in Bayesian modeling for high dimensional problems. As indicated by recent publications, see e.g., Johnson and Rossell (2012), Liang et al. (2013) and Song and Liang (2015b), Bayesian methods can often outperform frequentist methods in identifying important variables for high-dimensional problems. Hence, there is now also a need to develop an appropriate criterion for selecting models based on the outputs of MCMC for high-dimensional

problems. To address this issue, we extend EBIC to AEBIC, which is defined by

$$\text{AEBIC}(S) = -2nE_{\beta_S|X_n,S}L_n(X_n|\beta_S) + |S|\log(n) + 2\gamma|S|\log(p). \tag{3}$$

The asymptotic consistency of this criterion can be established under some sparsity and regularity conditions. Here we would like to point out that our proof for the asymptotic consistency is general, which is not restricted to generalized linear models.

The remainder of this article is organized as follows. Section 2 provides an informal derivation of ABIC and AEBIC. Section 3 establishes the consistency properties of ABIC and AEBIC with the proofs deferred to the Appendix. Section 4 evaluates the finite sample performance of ABIC and AEBIC through some simulation studies and a real data example. Section 5 concludes the article with a brief discussion.

2 Intuitions on ABIC and AEBIC

This section gives an informal derivation and intuitive explanation for ABIC and AEBIC. Let $X_n = \{X^{(i)} : i = 1, \ldots, n\}$ denote a dataset of n iid observations. For example, for high-dimensional regression, $X^{(i)} = (y^{(i)}, \boldsymbol{z}^{(i)})$, where y is the response variable, and $\boldsymbol{z}$ is a high-dimensional random vector containing p_n predictors. Under the high dimensional setting, p_n can increase with the sample size n. Let $S \subset \{1, 2, \ldots, p_n\}$ denote a generic model, let S^* denote the true model, and let $\mathcal{S} := \{S : S \subset \{1, 2, ..., p_n\}\}$ denote the space of all possible models. For high-dimensional problems, the true model can be sparse, i.e., $|S^*| = o(n)$ holds. The problem of model selection is to identify S^* from $\mathcal{S}$ based on the observed data.

Let's consider the problem under the Bayesian framework, which is to choose a model with the maximum posterior probability. Let $\pi(S)$ denote the prior distribution imposed on the model space $\mathcal{S}$ and let $\pi(\beta_S)$ denote the prior distribution of β_S. Then the posterior probability of the model S is given by

$$P(S|X_n) = \frac{m(X_n|S)\pi(S)}{\sum_{S\in\mathcal{P}} m(X_n|S)\pi(S)},$$

where

$$m(X_n|S) = \int \exp\left(nL_n(X_n|\beta_S)\right)\pi(\beta_S)d\beta_S.$$

By choosing appropriate priors, it can be shown that $P(S^*|X_n)$ converges to 1 in probability as n goes to infinity (Liang et al., 2013; Shin et al., 2015), which is the so-called global model consistency in Bayesian variable selection (Johnson and Rossell, 2012). This further implies that $\arg\max_S P(S|X_n) = S^*$ holds in probability as $n \to \infty$.

However, in most cases, it is impossible to calculate $P(S|X_n)$ exactly, so we need to approximate it. The approximation can be done using MCMC or the Laplace method. We note that the latter has been

used in deriving BIC and EBIC. Since

$$\arg\max_S P(S|X_n) = \arg\max_S m(X_n|S)\pi(S) = \arg\max_S \log\{m(X_n|S)\pi(S)\},$$

we only need to care about $\log\{m(X_n|s)\pi(S)\}$, which can be re-expressed as

$$\log\{m(X_n|S)\pi(S)\} = \log \int \exp\left(nL_n(X_n|\beta_S)\right)\pi(\beta_S)d\beta_S + \log\pi(S). \tag{4}$$

To deal with the log-integral term in (4), we expand $L_n(X_n|\beta_S)$ at $\widehat{\beta}_S$, the MLE of β_S. That is,

$$\begin{aligned} L_n(X_n|\beta_S) &\approx L_n(X_n|\widehat{\beta}_S) + \frac{1}{2}(\beta_S - \widehat{\beta}_S)'\left[\frac{\partial^2 L_n(X_n|\widehat{\beta}_S)}{\partial\beta_S\partial\beta_S^T}\right](\beta_S - \widehat{\beta}_S) \\ &= L_n(X_n|\widehat{\beta}_S) - \frac{1}{2}(\beta_S - \widehat{\beta}_S)'\left[\widehat{I}(X_n, \widehat{\beta}_S)\right](\beta_S - \widehat{\beta}_S), \end{aligned} \tag{5}$$

where $\widehat{I}(X_n, \widehat{\beta}_S) = -\frac{\partial^2 L_n(X_n|\widehat{\beta}_S)}{\partial\beta_S\partial\beta_S^T}$ is the averaged observed information matrix. Therefore,

$$\begin{aligned} & \int \exp\left(nL_n(X_n|\beta_S)\right)\pi(\beta_S)d\beta_S \\ \approx\ & \exp\left(nL_n(X_n|\widehat{\beta}_S)\right)\pi(\widehat{\beta}_S)\int \exp\left\{-\frac{1}{2}(\beta_S - \widehat{\beta}_S)'\left[n\widehat{I}(X_n, \widehat{\beta}_S)\right](\beta_S - \widehat{\beta}_S)\right\}d\beta_S \\ =\ & \exp\left(nL_n(X_n|\widehat{\beta}_S)\right)\pi(\widehat{\beta}_S)(2\pi/n)^{\frac{|S|}{2}}|\widehat{I}(X_n, \widehat{\beta}_S)|^{-\frac{1}{2}}, \end{aligned}$$

where the approximation is valid provided that the prior $\pi(\beta_S)$ is "flat" over the neighborhood of $\widehat{\beta}_S$ and $L_n(X_n|\beta_S)$ is ignorable outside the neighborhood of $\widehat{\beta}_S$. Finally,

$$\begin{aligned} &\log\{m(X_n|S)\pi(S)\} = \log \int \exp\left(nL_n(X_n|\beta_S)\right)\pi(\beta_S)d\beta_S + \log\pi(S) \\ &\approx nL_n(X_n|\widehat{\beta}_S) + \log\pi(\widehat{\beta}_S) + \frac{|S|}{2}\log(2\pi) - \frac{|S|}{2}\log(n) - \frac{1}{2}\log|\widehat{I}(X_n, \widehat{\beta}_S)| + \log\pi(S) \\ &\approx nL_n(X_n|\widehat{\beta}_S) - \frac{|S|}{2}\log(n) + \log\pi(S), \end{aligned}$$

where the second approximation is justified by the fact that when n is sufficiently large, $\log(2\pi)$ is comparably ignorable than $\log(n)$, $|\widehat{I}(X_n, \widehat{\beta}_S)|$ converges to a constant, and $\log\pi(\widehat{\beta}_S)$ can be controlled at a level of $O(1)$.

For the prior $\pi(S)$, if the setting of Liang et al. (2013) is adopted, then we have

$$\pi(S) = \lambda_n^{|S|}(1 - \lambda_n)^{p_n - |S|}, \tag{6}$$

where λ_n denotes the prior probability of each variable to be included in S and takes a value in the form

$$\lambda_n = \frac{1}{1 + p_n^{\gamma}\sqrt{2\pi}}, \tag{7}$$

for some parameter $\gamma > 0$. If the setting of Chen and Chen (2008) is adopted, then we have

$$\pi(S) \propto \binom{p_n}{|S|}^{-\gamma}. \tag{8}$$

For the case that $|S|$ is far smaller than P_n, it is easy to verify that $\log \pi(S) = -\gamma|S| \log(p_n) + O(1)$ holds for both prior settings. Therefore,

$$\log\{m(X_n|S)\pi(S)\} \approx nL_n(X_n|\widehat{\beta}_S) - \frac{|S|}{2}\log(n) - \gamma|S|\log(p_n), \tag{9}$$

which exactly equals to $-\frac{1}{2}$EBIC.

In this article, we propose a new method to approximate the log-integral term in (4). Starting from (9), we further approximate $L_n(X_n|\widehat{\beta}_S)$ by $E_{\{\beta_S|X_n,S\}}L_n(X_n|\beta_S)+\frac{|S|}{2n}$, where $E_{\{\beta_S|X_n,S\}}$ denotes the expectation with respect to the posterior distribution of β_S. This approximation has been discussed in Spiegelhalter et al. (2002) and can be verified as follows. It follows from (5), for sufficiently large n, we have

$$\begin{aligned} P(\beta_S|X_n, S) &= \frac{\exp\left(nL_n(X_n|\beta_S)\right)\pi(\beta_S)}{m(X_n|S)} \propto \exp\left(nL_n(X_n|\beta_S)\right)\pi(\beta_S) \\ &\approx \exp\left(nL_n(X_n|\widehat{\beta}_S)\right)\pi(\widehat{\beta}_S)\exp\left\{-\frac{1}{2}(\beta_S - \widehat{\beta}_S)'\left[n\widehat{I}(X_n, \widehat{\beta}_S)\right](\beta_S - \widehat{\beta}_S)\right\}, \end{aligned}$$

in the neighborhood of $\widehat{\beta}_S$ where $L_n(X_n|\beta_S)$ is dominant; that is, the posterior of β_s is asymptotically normal with mean $\widehat{\beta}_S$ and covariance matrix $\left[n\widehat{I}(X_n, \widehat{\beta}_S)\right]^{-1}$. Therefore,

$$\begin{aligned} &E_{\{\beta_S|X_n,S\}}L_n(X_n|\beta_S) \\ &\approx E_{\{\beta_S|X_n,S\}}\{L_n(X_n|\widehat{\beta}_S) - \frac{1}{2}(\beta_S - \widehat{\beta}_S)'\left[\widehat{I}(X_n, \widehat{\beta}_S)\right](\beta_S - \widehat{\beta}_S)\} \\ &= L_n(X_n|\widehat{\beta}_S) - \frac{1}{2}E_{\{\beta_S|X_n,S\}}(\beta_S - \widehat{\beta}_S)'\left[\widehat{I}(X_n, \widehat{\beta}_S)\right](\beta_S - \widehat{\beta}_S) \\ &= L_n(X_n|\widehat{\beta}_S) - \frac{1}{2}|S|/n, \end{aligned}$$

and $\log\{m(X_n|S)\pi(S)\}$ can be approximated as follows

$$\begin{aligned}\log\{m(X_n|S)\pi(S)\} &\approx nL_n(X_n|\widehat{\beta}_S) - \frac{|S|}{2}\log(n) - \gamma|S|\log(p_n)\\ &\approx nE_{\{\beta_S|X_n,S\}}L_n(X_n|\beta_S) + \frac{|S|}{2} - \frac{|S|}{2}\log(n) - \gamma|S|\log(p_n)\\ &\approx nE_{\{\beta_S|X_n,S\}}L_n(X_n|\beta_S) - \frac{|S|}{2}\log(n) - \gamma|S|\log(p_n).\end{aligned}$$

Under the large sample setting, $\frac{|S|}{2}$ is negligible because it is of lower order of $\frac{|S|}{2}\log(n)$.

As with EBIC, we define AEBIC as -2 times the approximated $\log\{m(X_n|S)\pi(S)\}$, i.e.,

$$AEBIC(s) = -2nE_{\{\beta_S|X_n,S\}}L_n(X_n|\beta_S) + |S|\log(n) + 2\gamma|S|\log(p_n). \tag{10}$$

For many problems, $E_{\{\beta_S|X_n,S\}}L_n(X_n|\beta_S)$ doesn't have a closed form, so we use MCMC samples to approximate it. This leads to the following model selection procedure based on the outputs of MCMC:

(a) *Run a MCMC algorithm to simulate T samples $\{(\beta_{S^{(t)}}, S^{(t)}) : t = 1, 2, \ldots, T\}$ from the posterior distribution $\pi(\beta_S, S|X_n)$.*

(b) *Let $\mathcal{S}_T$ denote the set of distinct models appeared in $\{S^{(t)} : t = 1, 2, \ldots, T\}$. For each $S \in \mathcal{S}_T$, calculate*

$$\widehat{AEBIC}(S) = \frac{-2n}{\#\{t : S^{(t)} = S\}} \sum_{\{t:S^{(t)}=S\}} L_n(X_n|\beta_{S^{(t)}}, S^{(t)}) + |S|\log(n) + 2\gamma|S|\log(p_n).$$

(c) *Set $\widehat{S} = \arg\min_{S\in\mathcal{S}_T} \widehat{AEBIC}(S)$ as the estimator for the true model S^*.*

Remark:

- Compared to AIC, BIC and EBIC, AEBIC has at least two merits: First, it avoids to calculate MLE, which can be very difficult for some problems. Second, it conducts a systematic search over the model space via MCMC; however, AIC, BIC and EBIC often employ a heuristic method to search for candidate models, and the search is often limited and incomplete.

- Compared to DIC, AEBIC has also at least two merits: First, AEBIC is invariant to reparameterization, for which the penalty term involves only the model size instead of model parameters; while this is not for DIC. Second, AEBIC is asymptotically consistent, while DIC is not. The asymptotic consistency of AEBIC for high-dimensional model selection will be rigorously proved in the next section.

- For low-dimensional problems, the penalty term $2\gamma|S|\log(p_n)$ can be ignored, and AEBIC is reduced to ABIC.

3 Asymptotic Consistency of AEBIC

For any model $S \subset \{1, 2, ..., p_n\}$, we let β_S denote the associated vector of parameters, and let $f(x|\beta_S)$ denote the likelihood function for a single data point x. Let S^* denote the true model, and let β_{S^*} denote the parameters of the true model. Assume $|supp(\beta_{S^*})| = s_0 = o(n)$ holds. Let $X_n = \{X^{(1)}, \ldots, X^{(n)}\}$ denote a set of n iid observations, and let $L_n(X_n|\beta_S) := \mathbb{E}_n[\log(f(x^{(i)}|\beta_S))] = \frac{1}{n}\sum_{i=1}^{n}\log(f(x^{(i)}|\beta_S))$ denote the empirical log-likelihood function, which is an analog of $L(\beta_S) := E[\log(f(X|\beta_S))]$. Denote the maximizer of $L_n(X_n|\beta_S)$ and $L(\beta_S)$ as $\hat{\beta}_S$ and $\beta_{S,0}$, respectively. For any fixed S with size $s := |S|$ such that $s/n \to 0$, we know that under some general consistency conditions of M-estimator, $\sqrt{n}(\hat{\beta}_S - \beta_{S,0}) \longrightarrow N(0, V)$, where $V := \left(E[\frac{\partial^2 \log(f(X|\beta_S))}{\partial\beta_S \partial\beta_S^T}|_{\beta_S=\beta_{S,0}}]\right)^{-1}$ is the inverse of the information matrix for model S. Let q denote the maximum size of the models under consideration, which is allowed to increase with n and p. Also, we assume that the size of the true model is no greater than q, i.e., $s_0 \leq q$. Let $\mathcal{P}(q) := \{S|S \subset \{1, 2, ..., p_n\}, |S| \leq q\}$ denote the collection of all models with size smaller than or equal to q.

To establish consistency of AEBIC, we assume the following set of conditions:

(A1) The information matrix $I_S = E[\frac{\partial^2 \log f(X|\beta_S)}{\partial\beta_S^2}|_{\beta_S=\beta_{S,0}}]$ has bounded minimum and maximum eigenvalues for all $|S| \leq q$. There exists an envelop function $J(x)$ such that $J(x) \geq \max_{i,j\in S, S\in\mathcal{P}(q)} |\frac{\partial^2 \log f(x|\beta_S)}{\partial\beta_i\partial\beta_j}|$ and $E[J^2(x)] \leq C$ for some generic constant $C > 0$.

(A2) $\log(f(x|\beta_S)$ is three times differentiable and locally L_1-Lipschitz. There exists a function $J_f(x)$ such that $J_f(x) \geq \max_{S\in\mathcal{P}(q)} |\log(f(x|\beta_S))|$ and $E[J_f(x)^2] \leq C$ for some generic constant $C > 0$. The third order derivative of $\log(f(x|\beta_S))$ is uniformly continuous and bounded on β_S and for all $S \subset \mathcal{P}(q)$. Define $f_1(x|\beta_S) := \frac{\partial \log(f(x|\beta_S))}{\partial\beta_S}$. There exists a function $J_1(x) > |f_1(x|\beta_{S,0})|$ for all x and $S \in \mathcal{P}(q)$ such that (i) $E[|J_1(x)|^2] < C_1$ for some generic constant $C_1 > 0$, and (ii) $\sup_{x\in\mathcal{X}} |J_1(x)| < C_2 q$, where $\mathcal{X}$ is the domain of observations X_n, and C_2 is some generic constant. The domain $\mathcal{X}$ satisfies $\sup_{i=1,2,...,n;j=1,2,...,p_n} |x_j^{(i)}| < M$ for some generic constant M, where $x_j^{(i)}$ denotes the jth component of $x^{(i)}$.

(A3) For any $S \in \mathcal{P}(q)$, the prior density function $\pi(\beta_S|S)$ is uniformly Lipschitz and bounded from the above, i.e., there exists a generic constant $L > 0$ such that $|\pi(\beta_S|S) - \pi(\beta'_S|S)| \leq L||\beta_S - \beta'_S||$ for any $\beta_S, \beta'_S \in \mathbb{R}^S$. Also, there exists a constant $c > 0$ such that $\min_{S\in\mathcal{P}(q)} \pi(\beta_{S,0}|S) > c$.

(A4) There exists an envelope function $F(x) > 0$ such that $f(x|\beta_S) < F(x)$ for all β_S with $S \in \mathcal{P}(q)$, and $E[F(x)] \leq \infty$.

(A5) The domain of β_S, denoted by $\mathcal{N}(S)$, is compact, i.e., there exists a constant $M > 0$ such that $|\beta_S|_\infty \leq M$ for any model $S \in \mathcal{P}(q)$. There exists an $\epsilon > 0$ such that $B_\epsilon(\beta_{S,0}) \subset \mathcal{N}(S)$, where $B_\epsilon(x)$ denotes the ϵ-ball of x.

(A6) For any model $S \in \mathcal{P}(q)$, there exists a unique maximizer $\beta_{S,0}$ of $L(\beta_S)$. Furthermore, for any $\eta > 0$, $\sup_{\beta_S \notin B_\eta(\beta_{S,0})} L(\beta_S) \leq L(\beta_{S,0}) + (c\eta^2 \wedge 1)$ for some constant $c > 0$, where $a \wedge b = \min(a, b)$.

(A7) $\min_i |\beta_i| / (\frac{q \log(p)}{n})^{\frac{1}{4}} \to \infty$ as $n \to \infty$.

Condition (A1) is primarily an extension of the sparse eigenvalue condition stated in Bickel et al. (2009). Condition (A2) is required for uniform convergence of the sample average of the loglikelihood function over different models $S \in \mathcal{P}(q)$. Here we assume that all the observations in X_n can be included in a compact set. This assumption is more or less a technical assumption, which has often been used in theoretical studies of Bayesian methods, see e.g. Jiang (2007). The conditions (A3)-(A5) are regularity conditions for f, π and the domain of β_S. Condition (A6) imposes a restriction on the behavior of $L(\beta_S)$ which enforces consistency of the maximum likelihood estimator uniformly over all $S \in \mathcal{P}(q)$. Condition (A7) is for model selection property.

First, we establish a uniform law of large numbers (ULLN) for $L_n(X_n|\beta_S)$, which is modified from Lemma 1 of Foygel and Drton (2011). The proofs of all lemmas and theorems are deferred to the Appendix.

Lemma 1. *(ULLN) If conditions (A1)-(A6) hold and* $\frac{q^2 \log(p)}{n} \to 0$ *as* $n \to \infty$*, then with probability going to 1, uniformly for all* $S \in \mathcal{P}(q)$ *and* $\beta_S \in \mathbb{R}^{|S|}$*,*

(i) $L_n(X_n|\beta_S) \to_p L(\beta_S)$*, where* $\to_p$ *denotes convergence in probability.*

(ii) $\hat{\beta}_S \to_p \beta_{S,0}$*.*

(iii) For any $\|\beta_S - \beta_{S,0}\|_2 \leq l_n/\sqrt{q}$*, where* l_n *is a decreasing sequence that converges to* 0*,* $\hat{I}_S(\beta_S)$ *converges to* $I_S(\beta_S)$ *in Frobenius norm. Consequently, the maximum and minimum eigenvalue of* $\hat{I}_S(\beta_S)$ *is bounded away from* 0 *and bounded from the above uniformly for all* $S \in \mathcal{P}(q)$*.*

(iv) $\|\hat{\beta}_S - \beta_{S,0}\|_2 = O\left(\left(\frac{q \log(p)}{n}\right)^{\frac{1}{4}}\right)$*. If* $\frac{q^3 \log(p)}{n} \to 0$*, then* $\hat{I}_S(\hat{\beta}_S)$*'s eigenvalues are bounded away from* 0 *and bounded from the above uniformly for all* $S \in \mathcal{P}(q)$*.*

Together with condition (A1), statement (iv) of Lemma 1 implies that $\hat{I}_S(\hat{\beta}_S)$ has uniformly bounded minimum and maximum eigenvalues for all $S \in \mathcal{P}(q)$.

By Lemma 1, uniformly for any S and for any $\beta_S \notin B_{\epsilon_n}(\hat{\beta}_S)$, $L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S) = L(\beta_S) - L(\beta_{S,0}) + o_p(1)$. Define $\hat{I}_S(\beta_S) := \mathbb{E}_n[\frac{\partial^2 f(x_i|\hat{\beta}_S)}{\partial \hat{\beta}_S^2}]\|_{\beta_S}$.

Lemma 2 (Approximation)**.** *Suppose the conditions (A1)-(A6) hold. If* $\frac{q^3 \log(p)}{n} \to 0$ *as* $n \to \infty$*, then for any sufficiently large* n *and a decreasing sequence* l_n *that converges to* 0*, with probability going to 1, uniformly over all* $S \in \mathcal{P}(q)$*, we have*

(i) $L_n(X_n|\beta_S) - L_n(\hat{\beta}_S) \leq_p L(\beta_S) - L(\beta_{S,0}) + O(\sqrt{\frac{q \log(p)}{n}})$.

(ii) For any β_S such that $||\beta_S - \hat{\beta}_S|| \leq \frac{l_n}{\sqrt{q}}$, then

$$L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S) = -\frac{1}{2}(\beta_S - \hat{\beta}_S)^T \hat{I}_S(\hat{\beta}_S)(\beta_S - \hat{\beta}_S) + O_p\left(\sqrt{q}\|\beta_S - \hat{\beta}_S\|^3 \vee \sqrt{\frac{q^2\log(p)}{n}}\|\beta_S - \hat{\beta}_S\|^2\right).$$

Recall that the posterior density function of β_S, given model S, is given by

$$P(\beta_S|X_n, S) = \exp\{nL_n(X_n|\beta_S)\}\pi(\beta_S)/m(X_n|S),$$

where $m(X_n|S) := \int_{\beta_S} \exp\{nL_n(X_n|\beta_S)\}\pi(\beta_S)d\beta_S$. Without loss of generality, we assume that $\pi(S) = 0$ for any $s \notin \mathcal{P}(q)$. Based on the previous two lemmas, we can now establish our main results. Theorem 1 shows that the posterior mean of $\log P(\beta_S|X_n, S)$ can be approximated by $\log P(\hat{\beta}_S|X_n, S)$, where $\hat{\beta}_S$ denotes the MLE of β_S. This theorem forms a theoretical foundation for the consistency of AEBIC.

Theorem 1 (Posterior mean approximation of the log-likelihood function)**.** *Suppose the conditions (A1)-(A6) hold and $\frac{q^4\log^2(n)\log(p)}{n} \to 0$ as $n \to \infty$. Then for sufficiently large n, with probability going to 1, uniformly over all $S \in \mathcal{P}(q)$, we have*

$$E_{\beta_S|X_n,S}[L_n(X_n|\beta_S)] = L_n(X_n|\hat{\beta}_S) - \frac{|S|}{2n} + O\left(\frac{|S|}{n}\sqrt{\frac{q^4\log^2 n\log(p)}{n}}\right),$$

where $\hat{\beta}_S$ denotes the MLE of β_S.

Lemma 3 establishes an upper bound for the maximum likelihood values of overfitted models.

Lemma 3. *Assume the conditions (A1)-(A6) hold. For any model S such that $|S| \leq q$, $S^* \subset S$ and $\frac{q^4\log^2(p)}{n} \to 0$, the following statement holds uniformly for such models with probability going to 1,*

$$2n(L_n(X_n|\hat{\beta}_S) - L_n(X_n|\hat{\beta}_{S^*})) \leq (1+\eta)|S\backslash S^*|\log(n^\zeta p),$$

where $\eta > 0$ is an arbitrarily small constant, $\zeta > \frac{1}{5}$ is any fixed constant, and $\hat{\beta}_{S^*}$ denotes the MLE of β_{S^*}.

Finally, we have the following theorem for the consistency of AEBIC in model selection.

Theorem 2 (Model Selection Consistency)**.** *Assume that the conditions (A1)-(A7) are satisfied. If $\gamma > 1$, $\frac{q^6\log^2(n)}{n\log(p)} \to 0$, and $\frac{q^4\log^2(p)}{n} \to 0$, then for sufficiently large n, with probability going to 1, it holds that for all models S with $|S| \leq q$ and $S \neq S^*$,*

$$AEBIC(S^*) < AEBIC(S),$$

where $AEBIC(S)$ is as defined in (10).

4 Applications of ABIC and AEBIC to Generalized Linear Models

In this section we consider the applications of ABIC and AEBIC to generalized linear models (GLMs) (McCullagh and Nelder, 1989). Let $X = (y, \boldsymbol{z})$, where y denotes the response variable and $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)$ denote a vector of predictors. Suppose that the distribution function of X is given by

$$f(y|\boldsymbol{z}, \beta) = \exp\left\{a(\theta)y + b(\theta) + c(y)\right\}, \tag{11}$$

where $a(\cdot)$ and $b(\cdot)$ are continuously differentiable functions of θ, $c(y)$ is a constant function of y, θ is the so-called natural parameter that relates y to the predictors via a linear function $\theta = \boldsymbol{z}^T\beta = \beta_1 z_1 + \cdots + \beta_p z_p$, where $\beta_1, \ldots, \beta_p$ are regression coefficients. Here, the intercept term has been treated as a special predictor included in $\boldsymbol{z}$. The mean function $u = E(y|\boldsymbol{z}) = -b'(\theta)/a'(\theta) := \psi(\theta)$, where $\psi(\theta)$ is the inverse of a chosen link function. This class of GLMs includes regression models with the responses that are binary, Poisson, and Gaussian (with known variance), whose respective link functions are logit, log and linear.

For this class of GLMs, we follow Liang et al. (2013) to assume that each model S is subject to the following prior distribution

$$\pi(S) \propto \lambda_n^{|S|}(1-\lambda_n)^{p_n-|S|} I(|S| \le q), \tag{12}$$

where λ_n is chosen according to (7). Further, we let β_S be subject to a Gaussian prior distribution $N(0, \sigma_S^2 I_{|S|})$, where $I_{|S|}$ denotes a $|S| \times |S|$-matrix and

$$\sigma_S^2 = \frac{1}{2\pi} e^{C_0/|S|}, \tag{13}$$

for some positive constant C_0. As argued in Liang et al. (2013), under such prior setting, we have $\log \pi(\beta_S) = O(1)$, which satisfies the requirement of unit prior information in derivation of BIC (Ando, 2010). The unit prior information is also implicitly required in condition (A3). Then, with an appropriate choice of q, which works as a sparsity constraint for the true model, we can verify that the GLM (11) satisfies the conditions (A1)-(A7). We note that under a set of slightly weaker but less general conditions, we have also proved the consistency of AEBIC for GLMs. Refer to the supplementary material of the paper for the detail.

4.1 Low-Dimensional Case

We first work on some low-dimensional problems. In this case, ABIC, which is defined in (2), is used as the model selection criterion.

Logistic Regression Consider a logistic regression with three active predictors

$$\text{logit } P(y = 1|\boldsymbol{z}) = 2z_1 + z_2 + 3z_3, \tag{14}$$

for which the total number of candidate predictors is $p = 50$ and the sample size is $n = 500$. All candidate predictors are generated from the standard Gaussian distribution. For the prior $\pi(S)$, we set $\gamma = 0$ in (7) and ignored the upper bound q; and for the prior $\pi(\beta_S)$, we set $C_0 = 10$ in (13).

To simulate from the posterior distribution of (S, β_S), we implemented the reversible jump MCMC algorithm (Green, 1995), which includes three types of moves, birth, death and parameter updating. In the birth step, we randomly choose a predictor excluded from the current model S, and include it into S to produce a larger model S'. The regression coefficient of the new predictor is generated from the Gaussian distribution $N(0, 10^2)$. In the death step, we randomly choose a predictor included in the current model S, and remove it from S to produce a smaller model S'. In the parameter updating step, we keep S unchanged, but randomly choose one coefficient β_i, $i \in S$, to update with a Gaussian random walk proposal for which the variance is set to 0.5^2. To accelerate the convergence of the simulation, an advanced MCMC algorithm, such as stochastic approximation Monte Carlo (Liang et al., 2007), can be used. The algorithm was run for 10^6 iterations, where the first 10^5 iterations were discarded for the burn-in process and the posterior samples were collected at every 10th iteration from the remaining part of the run. Based on the collected samples, ABIC was calculated for different models according to the procedure given in Section 2. Based on the same set of posterior samples, we also calculated AIC,BIC and DIC for different models. For simplicity, we only considered the models with the sampling frequency exceeding 100.

The simulation was repeated for 100 times, where a different dataset was generated at each time. The results were summarized in Table 1. The performance of the methods was evaluated using three metrics: (i) mean ($\mu_{|\widehat{S}_i|}$) and standard deviation ($\sigma_{|\widehat{S}_i|}$) of $|\widehat{S}_i|$, where $\widehat{S}_i$ denotes the model selected in the ith run; (ii) recall= $[\sum_{i=1}^{100} |S^* \cap \widehat{S}_i|]/[|S^*| \cdot 100]$; and (iii) precision= $[\sum_{i=1}^{100} |S^* \cap \widehat{S}_i|]/[\sum_{i=1}^{100} |\widehat{S}_i|]$.

Table 1: Comparison of AIC, BIC, DIC and ABIC for low dimensional logistic regression.

	AIC	BIC	DIC	ABIC
$\mu_{\|\widehat{S}_i\|}$ ($\sigma_{\|\widehat{S}_i\|}$)	8.86(1.99)	3.72(0.96)	9.04(2.15)	3.33(0.57)
Recall	1.000	1.000	1.000	1.000
Precision	0.339	0.806	0.332	0.901

Table 1 indicates that the simulation results agree well with the theoretical results: BIC and ABIC are consistent for model selection, while AIC and DIC are not and they tend to select larger models. It is remarkable that ABIC also significantly outperforms BIC in terms of model selection accuracy, i.e. $\mu_{|\widehat{S}_i|}$ and precision.

Linear Regression Next we consider a linear regression with five active predictors

$$y = z_1 + 2.2z_2 - 1.6z_3 + 2z_4 - 1.4z_5 + \epsilon, \tag{15}$$

where ϵ follows the standard normal distribution. For this example, we also set $p = 50$ and $n = 500$ and generated all predictors from the standard normal distribution. In addition, we specified the same prior distributions and employed the same MCMC algorithm as for the logistic regression example. Table 2 summarizes the results for 100 independent runs. For this example, similar conclusions can be made as for the logistic example: AIC and DIC tend to select larger models, BIC and ABIC tend to select correct models, and ABIC outperforms all the other criteria.

Table 2: Comparison of AIC, BIC, DIC and ABIC for low dimensional linear regression.

	AIC	BIC	DIC	ABIC
$\mu_{\lvert\widehat{S}_i\rvert}$ $(\sigma_{\lvert\widehat{S}_i\rvert})$	8.70(1.15)	5.59(0.82)	9.05(1.08)	5.35(0.74)
Recall	1.000	1.000	1.000	1.000
Precision	0.575	0.894	0.552	0.934

4.2 High Dimensional Case

For the high-dimensional case, AEBIC is used as the model selection criterion.

Logistic Regression Consider the logistic regression (14) again. Here we increased p from 50 to 2000, while keeping the sample size $n = 500$ unchanged. For the prior $\pi(S)$, we tried four different γ values $0.5, 0.6, 0.7$, and 0.8, and fixed the upper bound q to 50. For the prior $\pi(\beta_S)$, we still set $C_0 = 10$. The simulations were done as for the previous examples. Based on the collected samples, we calculated both EBIC and AEBIC for the models with the sampling frequency greater than 20. AIC, BIC and DIC are not applicable to high-dimensional problems. Table 3 summarizes the results for 100 independent runs.

Table 3: Comparison of EBIC and AEBIC for high-dimensional logistic regression.

	γ	0.5	0.6	0.7	0.8
EBIC	$\mu_{\lvert\widehat{S}_i\rvert}$ $(\sigma_{\lvert\widehat{S}_i\rvert})$	3.60(0.96)	3.21(0.46)	3.15(0.43)	3.07(0.30)
	Recall	1.000	1.000	1.000	1.000
	Precision	0.833	0.935	0.949	0.977
AEBIC	$\mu_{\lvert\widehat{S}_i\rvert}$ $(\sigma_{\lvert\widehat{S}_i\rvert})$	3.39(0.89)	3.11(0.31)	3.06(0.28)	3.02(0.20)
	Recall	1.000	1.000	1.000	0.997
	Precision	0.885	0.965	0.977	0.990

The comparison indicates that as γ increases, both EBIC and AEBIC tend to select more sparse models. This is due to that a larger value of γ imposes a heavier penalty on the model size. It is remarkable that AEBIC is uniformly better than EBIC for all values of γ in terms of model size and true discovery rate.

Linear Regression We reconsidered the linear regression (15) with an enlarged value of $p = 2000$, while keeping other settings unchanged. We adopted the same prior setting as for the high-dimensional logistic regression example. For simulation, the MH algorithm was slightly modified for this example. The

modification is in the birth step, where we no longer choose new predictors uniformly, instead, we choose each predictor with a probability proportional to $|corr(y, x_k)|$. Here $corr(\cdot)$ denotes the Pearson correlation of two random variables. This modification can significantly accelerate the convergence of the simulation. Table 4 summarizes the results for 100 independent runs. The comparison indicates again that AEBIC outperforms EBIC in terms of model size and true discovery rate.

Table 4: Comparison of EBIC and AEBIC for high-dimensional linear regression.

	γ	0.5	0.6	0.7	0.8
EBIC	$\mu_{\lvert\widehat{S}_i\rvert}$ ($\sigma_{\lvert\widehat{S}_i\rvert}$)	5.30(0.54)	5.15(0.39)	5.07(0.26)	5.03(0.17)
	Recall	1.000	1.000	1.000	1.000
	Precision	0.943	0.971	0.986	0.994
AEBIC	$\mu_{\lvert\widehat{S}_i\rvert}$ ($\sigma_{\lvert\widehat{S}_i\rvert}$)	5.16(0.42)	5.06(0.24)	5.05(0.22)	5.01(0.10)
	Recall	1.000	1.000	1.000	1.000
	Precision	0.969	0.988	0.990	0.998

4.3 A Real Data Example

We considered a dataset collected by Singh et al. (2002), which consists of the expression values of 6033 genes on 102 samples, 52 prostate cancer patients and 50 controls. The dataset is downloadable at `http://statweb.stanford.edu/~ckirby/brad/LSI/datasets-and-programs/datasets.html`, and has been analyzed in quite a few papers, see e.g., Chen and Chen (2012), Efron (2009), and Liang et al. (2013). Our aim is to identify important genes that are associated with the prostate cancer via high-dimensional logistic regression.

For the prior $\pi(S)$, we set $\gamma = 0.7$ and fix the upper bound q to 50. For the prior $\pi(\beta_S)$, we set $C_0 = 10$. The MH algorithm was slightly modified from the one used before. The modification is in the birth step, where we no longer choose new predictors uniformly, instead, we choose each predictor with a probability proportional to the weight $w_k = \text{Null Deviance} - \text{Deviance}_k + 0.1$, where k indexes the genes, "Null Deviance" denotes the deviance of the null model which includes the intercept term only, and Deviance_k denotes the deviance of the model which includes the intercept term and gene k only.

The algorithm was run for 10^7 iterations, where the first 10^6 iterations were discarded for the burn-in process and the samples were collected at every 10th iterations from the remaining part of the run. Based on the collected samples, AEBIC were calculated for each model with the sampling frequency greater than 10. For this example, we observed multiple models with very similar AEBIC values. As in Chen and Chen (2012), we ranked genes based on their appearance frequencies in top 100 models. It is interesting to point out that our rank has much overlap with the rank given in Chen and Chen (2012), which was established based on EBIC but with a different model searching procedure. Table 5 gives top 10 genes reported in Chen and Chen (2012) and their corresponding ranks by our method.

Table 5: Top 10 genes ranked by EBIC in Chen and Chen (2012) and their corresponding ranks by AEBIC, where the genes are labeled by their column numbers in the dataset.

Gene	v610	v1720	v332	v364	v1068	v914	v3940	v1077	v4331	v579
EBIC	1	2	3	4	5	6	7	8	9	10
AEBIC	1	15	8	6	10	5	3	7	4	19

5 Conclusion

In this paper, we have proposed ABIC and AEBIC as general Bayesian model selection criteria for low and high-dimensional problems, respectively; established their consistency under mild conditions; and illustrated their applications using generalized linear models. The numerical results indicate that the proposed criteria can significantly outperform the existing ones, such as AIC, BIC, DIC and EBIC, in terms model selection accuracy.

Compared to AIC, BIC and EBIC, the proposed criteria avoid to calculate MLEs, which can be very difficult for some problems. Also, the MCMC simulations required by the proposed criteria provide a systematic way to search for the best ABIC or AEBIC models. While the AIC, BIC or EBIC methods often employ a heuristic method to search over the model space, and the search is often limited and incomplete. ABIC and AEBIC have also overcome many limitations of DIC. As explained previously, they both are invariant to reparameterization and asymptotically consistent, while DIC is not. Given these attractive features, we expect that ABIC and AEBIC will be widely adopted by researchers for Bayesian model selection in the near future.

Acknowledgement

Liang's research was partially supported by grants DMS-1612924 and R01-GM117597.

Appendix A Technical Details

In the appendix, we use $X_n = \{X^{(1)}, \ldots, X^{(n)}\}$ to denote a set of n observations, use $\mathbb{P}_n$ to denote the empirical measure of the observations, use $L_n(X_n|\beta_S)$ to denote the averaged log-likelihood function as defined in (1), use $\mathbf{S}(X_n, \beta_S)$ to denote the score function $\partial L_n(X_n|\beta_S)/\partial\beta_S$, and use $\hat{I}_S(\beta_S)$(or $\hat{I}_S$) to denote the observed information matrix $-\partial^2 L_n(X_n|\beta_S)/\partial\beta_S\partial\beta_S^T$. For simplicity, we also depress the subscript of p_n and replace it by p.

A.1 Proof of Lemma 1

Part (i) Define $\mathcal{F}_S(M) := \{\log f(x|\beta_S)1(F \leq M) : \beta_S \in \mathbb{R}^{|S|}\}$, where F is defined in $(A4)$. Define $\mathcal{F}_q(M) := \cup_{|S|\leq q, S\subset\{1,2,\ldots,p\}}\mathcal{F}_S(M)$. It is easy to see that for any $s \leq s_0$, by the Lipschitz condition

of $\log f(x|\beta)$ and condition (A6), the covering number of $\mathcal{F}_S(M)$ relative to the $L_1(\mathbb{P}_n)$-norm, which is denoted by $N(\epsilon, \mathcal{F}_S(M), L_1(\mathbb{P}_n))$, satisfies

$$N\left(\epsilon, \mathcal{F}_S(M), L_1(\mathbb{P}_n)\right) \leq C_0 \frac{1}{\epsilon^{|S|}},$$

for some generic constant $C_0 > 0$, where $\mathbb{P}_n$ denotes the empirical measure of n observations.

Therefore, the covering number of $\mathcal{F}_q(M)$ satisfies

$$N\left(\epsilon, \mathcal{F}_q(M), L_1(\mathbb{P}_n)\right) \leq \sum_{s=1}^{q} C_p^s C_0 \frac{1}{\epsilon^s},$$

where C_p^s is the combinatorial number of choosing s from p. Thus,

$$\log N\left(\epsilon, \mathcal{F}_s(M), L_1(\mathbb{P}_n)\right) \leq \sum_{s=1}^{q} \log C_p^s + \log(C_0) + q\log(1/\epsilon) \precsim q\log p + q\log(1/\epsilon) = o(n).$$

Therefore, by Uniform Law of Large Numbers (Theorem 2.4.3 of Van der Vaart and Wellner (1997)),

$$|L_n(X_n|\beta_S) - L(\beta_S)| \to_p 0,$$

holds uniformly for all $S \in \mathcal{P}(q)$.

Part (ii) It follows directly from the conclusion of part (i) and condition (A7).

Part (iii) First, by the continuity condition stated in condition (A2), it is easy to see that uniformly over all $S \in \mathcal{P}(q)$, $\|I_S(\beta_S) - I_S(\beta_{S,0})\|_\infty \leq L\|\beta_S - \beta_{S,0}\|$ holds for some constant L. Hence,

$$\|\hat{I}_S(\beta_S) - \hat{I}_S(\beta_{S,0})\|_2 \leq \sqrt{q}\|\hat{I}_S(\beta_S) - \hat{I}_S(\beta_{S,0})\|_\infty \leq L\sqrt{q}||\beta_S - \beta_{S,0}||. \tag{16}$$

For any β_S such that $\frac{\|\beta_S - \beta_{S,0}\|}{q} \to 0$, the eigenvalue of $\hat{I}_S(\beta_S)$ is bounded away from 0 and bounded from the above uniformly if $\hat{I}_S(\beta_{S,0})$'s eigenvalues have the same property. Therefore, it suffices to show a ULLN that $\hat{I}_S(\beta_{S,0})$ converges in probability to $I_S(\beta_{S,0})$ uniformly in Frobenius norm, for any $S \in \mathcal{P}(q)$.

Again, we show it by first examining the convergence of $\hat{I}_S(\beta_{S,0})$ to $I_S(\beta_{S,0})$ under L^∞ norm. By condition (A1), $J(x)$ performs as an envelop function of $\frac{\partial^2 \partial^2 \log f(x|\beta_{S,0})}{\partial\beta_i \partial\beta_j}$. Therefore, by the Lipschitz condition of $\frac{\partial^2 \log(f(x|\beta_{S,0}))}{\partial\beta_i \partial\beta_j}$ and the existence of function J, the covering number of $\mathcal{F}_I(q) := \{\frac{\partial^2 \log(f(x|\beta_{S,0}))}{\partial\beta_i \partial\beta_j} | i, j \in S, S \in \mathcal{P}(q)\}$ satisfies

$$\log N(\epsilon, \mathcal{F}_I(q), L_2(\mathbb{P}_n)) \precsim \log\left(\sum_{s=1}^{q} s^2 C_p^s\right) + q\log(1/\epsilon),$$

for any $\epsilon > 0$.

By Theorem 2.5.2 of Van der Vaart and Wellner (1997), $P^*(\sqrt{n}\max_{S\in\mathcal{P}(q)}\|\hat{I}_S(\beta_{S,0}-I_S(\beta_{S,0})\|_\infty) > x) \leq C\frac{2}{x}\int_{\epsilon=0}^{1}\sqrt{\log N(\epsilon,\mathcal{F}_I(q),L_2(\mathbb{P}_n))} \leq C_1\frac{2}{x}\sqrt{q\log(p)}$, where $P^*(A)$ denotes the outer probability of an event A and C, C_1 are some constant that only depends on the function J. Setting $x = K\sqrt{q\log(p)}$ for any $K > 0$, we have $P^*(\max_{S\in\mathcal{P}(q)}||\hat{I}_S(\beta_{S,0}) - I_S(\beta_{S,0})||_\infty) > K\sqrt{\frac{q\log(p)}{n}}) < \frac{2C_1}{K}$, i.e., $\max_{S\in\mathcal{P}(q)}||\hat{I}_S(\beta_{S,0}) - I_S(\beta_{S,0})||_\infty) = O_p(\sqrt{\frac{q\log(p)}{n}})$.

Since the Frobenius norm of $\hat{I}_S(\beta_{S,0}) - I_S(\beta_{S,0})$ satisfies

$$\|\hat{I}_S(\beta_{S,0}) - I_S(\beta_{S,0})\|_2 \leq \sqrt{q}\|\hat{I}_S(\beta_{S,0}) - I_S(\beta_{S,0})\|_\infty = O_p\left(\sqrt{\frac{q^2\log(p)}{n}}\right), \tag{17}$$

uniformly for all $S \in \mathcal{P}(q)$, we have

$$\|\hat{I}_S(\beta_S) - I_S(\beta_{S,0})\|_2 \precsim \sqrt{q}\|\beta_S - \beta_{S,0}\| + \sqrt{\frac{q^2\log(p)}{n}},$$

by combining inequality (16) and inequality (17). By the assumption that $q^2\log(p)/n \to 0$, the statement (iii) holds.

Part (iv) By condition (A6), $L(\hat{\beta}_S) \geq L(\beta_{S,0}) + c||\hat{\beta}_S - \beta_{S,0}||^2$. On the other hand, $L_n(\hat{\beta}_S) \leq L_n(\beta_{S,0})$.

By the Lipschitz condition of $\log f(x|\beta_S)$ and the existence of function J_f, the covering number of $\mathcal{F}(q) := \{\log f(x|\beta_S) : |S| \leq q\}$ satisfies

$$\log N(\epsilon,\mathcal{F}(q),L_2(\mathbb{P}_n)) \precsim \log(\sum_{s=1}^{q} C_p^s) + q\log(1/\epsilon),$$

for any $\epsilon > 0$.

By Theorem 2.5.2 of Van der Vaart and Wellner (1997), $P^*(\max_{S\in\mathcal{P}(q)}\sqrt{n}(L_n(X_n|\beta_S) - L(\beta_S)| > x) \leq C_2\frac{2}{x}\int_{\epsilon=0}^{1}\sqrt{\log N(\epsilon,\mathcal{F}(q),L_2(\mathbb{P}_n))} \leq C_3\frac{2}{x}\sqrt{q\log(p)}$, where $P^*(A)$ denotes the outer probability of an event A and C_2, C_3 are some constant that only depends on the function J_f. Setting $x = K\sqrt{q\log(p)}$ for any $K > 0$, we have

$$P^*(\max_{S\in\mathcal{P}(q)}|L_n(X_n|\beta_S) - L(\beta_S)| > K\sqrt{\frac{q\log(p)}{n}}) \leq \frac{2C_3}{K},$$

i.e., $\max_{S\in\mathcal{P}(q)}(L_n(X_n|\beta_S) - L(\beta_S) = O(\sqrt{\frac{q\log(p)}{n}})$. Consequently, $L(\hat{\beta}_S) \leq L_n(\hat{\beta}_S) + O_p(\sqrt{\frac{q\log(p)}{n}}) \leq L_n(\beta_{S,0}) + O_p(\sqrt{\frac{q\log(p)}{n}}) \leq L(\beta_{S,0}) + O(\sqrt{\frac{q\log(p)}{n}})$. This implies that

$$c\|\hat{\beta}_S - \beta_{S,0}\|^2 = O_p(\sqrt{\frac{q\log(p)}{n}}), \tag{18}$$

i.e., $||\hat{\beta}_S - \beta_{S,0}|| \leq \left(\frac{q\log(p)}{n}\right)^{\frac{1}{4}}$.

If $\frac{q^3\log(p)}{n} \to 0$, then $\left(\frac{q\log(p)}{n}\right)^{\frac{1}{4}} = o(\frac{1}{\sqrt{q}})$, i.e., the condition for statement (iii) holds. Hence, $\hat{I}_S(\hat{\beta}_S)$'s eigenvalues are bounded away from 0 and bounded from the above uniformly over all $S \in \mathcal{P}(q)$.

A.2 Proof of Lemma 2

Part (i) By the optimality of $\hat{\beta}_S$, we have $L_n(X_n|\beta_S) - L_n(\hat{\beta}_S) \leq L_n(X_n|\beta_S) - L_n(\beta_{S,0})$. By Uniform Central Limit Theorem obtained in the proof of Statement (iv) of Lemma 1, $L_n(X_n|\beta_S) - L_n(\beta_{S,0}) \leq L(\beta_S) - L(\beta_{S,0}) + O(\sqrt{\frac{q\log(p)}{n}})$. Therefore,

$$L_n(X_n|\beta_S) - L_n(\hat{\beta}_S) \leq L(\beta_S) - L(\beta_{S,0}) + O(\sqrt{\frac{q\log(p)}{n}}).$$

Part (ii) Consider the local Taylor expansion of $L_n(X_n|\beta_S)$ around $\hat{\beta}_S$,

$$L_n(X_n|\beta_S) = L_n(X_n|\hat{\beta}_S) - \frac{1}{2}(\beta_S - \hat{\beta}_S)^T \hat{I}_S(\beta_S^*)(\beta_S - \hat{\beta}_S),$$

where β_S^* is a point between β_S and $\hat{\beta}_S$. By statement (iii) of Lemma 1, $\hat{I}_S(\beta_S^*) = \hat{I}_S(\hat{\beta}_S) + o_p(1)$ uniformly with respect to the Frobenius norm when $||\beta_S - \hat{\beta}_S|| = o(\frac{1}{\sqrt{q}})$.

The proof of statement (iii) of Lemma 1 in fact provides a bound for the Frobenius norm of $\hat{I}_S(\beta_S^*) - \hat{I}_S(\hat{\beta}_S)$. Therefore, $L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S) = -\frac{1}{2}(\beta_S - \hat{\beta}_S)^T \hat{I}_S(\beta_S - \hat{\beta}_S) + ||\beta_S - \hat{\beta}_S||^2 O(\sqrt{q}||\beta_S - \hat{\beta}_S|| \vee \sqrt{\frac{q^2\log(p)}{n}})$.

A.3 Proof of Theorem 1

For sufficiently large n, with probability going to 1, all statements in Lemma 1 and Lemma 2 hold and the following proof is based on these statements. First, we note that

$$E_{\beta_S|X_n,S}[L_n(X_n|\beta_S)] - L_n(X_n|\hat{\beta}_S) = \int P(\beta_S|X_n, S)(L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S))d\beta_S, \tag{19}$$

where $P(\beta_S|X_n, S) = \exp(nL_n(X_n|\beta_S))\pi(\beta_S)/m(S)$ and $m(S) = \int_{\beta_S} \exp(nL_n(X_n|\beta_S))\pi(\beta_S)d\beta_S$.

We split the integral domain in (19) into three regions: a small neighborhood around $\hat{\beta}_S$, denoted by $B_{\epsilon_K}(\hat{\beta}_S)$, where $\epsilon_K = \sqrt{\frac{Kq\log(n)}{n}}$; the area between the small neighborhood $B_{\epsilon_K}(\hat{\beta}_S)$ and a larger neighborhood $B_{\epsilon_n}(\hat{\beta}_S)$, denoted by $B_{\epsilon_K}(\hat{\beta}_S) - B_{\epsilon_n}(\hat{\beta}_S)$, where $\epsilon_n = \frac{l_n}{\sqrt{q}}$ and $l_n = \left(\frac{q^3\log(p)}{n}\right)^{\frac{1}{8}} \to 0$; and the rest area $B^c_{\epsilon_n}(\hat{\beta}_S)$.

For the neighborhood $B_{\epsilon_K}(\hat{\beta}_S)$, since $q^3\log(p)/n \to 0$, we have $\epsilon_K = o(\frac{1}{\sqrt{q}})$. Applying Lemma 2, we obtain that, with probability going to 1, for any $\beta_S \in B_{\epsilon_K}(\hat{\beta}_S)$,

$$L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S) = -\frac{1}{2}(\beta_S - \hat{\beta}_S)^T \hat{I}_S(\beta_S - \hat{\beta}_S) + r_n(\beta_S)||\beta_S - \hat{\beta}_S||_2^2,$$

where $|r_n(\beta_S)| \leq C(\sqrt{q}\|\beta_S - \hat{\beta}_S\|_2 \vee \sqrt{\frac{q^2\log(p)}{n}}) = C\sqrt{\frac{q^2\log(p)}{n}}$ uniformly for some constant $C > 0$.

Similarly, we have

$$\exp(nL_n(X_n|\beta_S))\pi(\beta_S) = \exp(nL_n(X_n|\hat{\beta}_S))\exp(-n\frac{1}{2}(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S))\exp(nr_n(\beta_S)||\beta_S-\hat{\beta}_S||_2^2)\pi(\beta_S),$$

where $nr_n(\beta_S)||\beta_S - \hat{\beta}_S||_2^2 \leq n\sqrt{\frac{q^2\log(p)}{n}}\frac{Kq\log(n)}{n} \precsim \sqrt{\frac{q^4\log^2(n)\log(p)}{n}}$ and thus $\exp(nr_n||\beta_S - \hat{\beta}_S||_2^2) = (1 + O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}))$.

Also, for any $\beta_S \in B_{\epsilon_K}$, by condition (A3), we have

$$\pi(\beta_S) = \pi(\hat{\beta}_S) + O(||\beta_S - \hat{\beta}_S||_2) = \pi(\hat{\beta}_S)(1 + O(\sqrt{\frac{q\log(n)}{n}}).$$

Since $\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}$ dominates $\sqrt{\frac{q\log(n)}{n}}$, we finally have

$$\begin{aligned}\exp(nL_n(X_n|\beta_S))\pi(\beta_S) = \pi(\hat{\beta}_S)\exp\left(nL_n(X_n|\hat{\beta}_S) - \frac{n}{2}(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S)\right) \\ \times \left(1 + O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})\right).\end{aligned} \tag{20}$$

For any $\beta_S \in B_{\epsilon_K}(\hat{\beta}_S) - B_{\epsilon_n}(\hat{\beta}_S)$, since $\epsilon_n = \frac{l_n}{\sqrt{q}} = o(\frac{1}{\sqrt{q}})$, we can apply Lemma 2 again and get $|r_n(\beta_S)| \leq \sqrt{q}||\beta_S - \hat{\beta}_S||_2 \to 0$. Therefore

$$\begin{aligned}L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S) &= -\frac{1}{2}(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S) + r_n(\beta_S)||\beta_S - \hat{\beta}_S||_2^2 \\ &= -\frac{1}{2}(\beta_S - \hat{\beta}_S)'\hat{I}_S(\beta_S - \hat{\beta}_S)(1 + o(1)) \leq -\frac{c_1}{4}||\beta_S - \hat{\beta}_S||^2,\end{aligned}$$

where $c_1 > 0$ is the lower bound of the minimum eigenvalue of $\hat{I}_S$ for all $S \in \mathcal{P}(q)$. Hence, for sufficiently large n, we have

$$\exp(nL_n(X_n|\beta_S))\pi(\beta_S) \leq \exp(nL_n(X_n|\hat{\beta}_S))\exp(-\frac{nc_1}{4}||\beta_S - \hat{\beta}_S||^2)\pi(\beta_S). \tag{21}$$

For any $\beta_S \in B_{\epsilon_n}^c(\hat{\beta}_S)$, by statement (i) of Lemma 2, we have

$$L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S) \leq L(\beta_S) - L(\beta_{S,0}) + O(\sqrt{\frac{q\log(p)}{n}}).$$

By condition (A6), we also have

$$L(\beta_S) - L(\beta_{S,0}) \leq -c\min(\inf_{\beta_S \in B_{\epsilon_n}^c(\beta_S)}\|\beta_S - \hat{\beta}_S\|^2, 1) \leq -\frac{cl_n^2}{q}.$$

Note that the quantity l_n^2/q dominates $\sqrt{\frac{q\log(p)}{n}}$ because $l_n = \left(\frac{q^3\log(p)}{n}\right)^{\frac{1}{8}}$.

Hence, for any $\beta_S \in B_\epsilon^c(\hat{\beta}_S)$, we have

$$\begin{aligned}\exp(nL_n(X_n|\beta_S))\pi(\beta_S) &\leq \exp(nL_n(X_n|\hat{\beta}_S))\pi(\beta_S)\exp\{-n(\frac{cl_n^2}{q})(1+o(1)))\} \\ &\leq \pi(\beta_S)\exp(nL_n(X_n|\hat{\beta}_S))\exp(-\frac{c'nl_n^2}{q}),\end{aligned} \tag{22}$$

for some generic constant $c' > 0$.

Now let's calculate $m(S)$ by dividing it into three parts according to the previous partition of the integral region, that is,

$$\begin{aligned}m(S) &:= \int \exp(nL_n(X_n|\beta_S))\pi(\beta_S)d\beta_S \\ &= \int_{\beta_S\in B_{\epsilon_K}(\hat{\beta}_S)} \exp(nL_n(X_n|\beta_S))\pi(\beta_S)d\beta_S + \int_{\beta_S\in B_{\epsilon_n}(\hat{\beta}_S)-B_{\epsilon_K}(\hat{\beta}_S)} \exp(nL_n(X_n|\beta_S))\pi(\beta_S)d\beta_S \\ &+ \int_{\beta_S\in B_\epsilon^c(\hat{\beta}_S)} \exp(nL_n(X_n|\beta_S))\pi(\beta_S)d\beta_S \\ &= (Int1) + (Int2) + (Int3).\end{aligned} \tag{23}$$

The first term of equation (23) satisfies:

$$\begin{aligned}Int1: &= \int_{\beta_S\in B_{\epsilon_K}(\hat{\beta}_S)} \exp(nL_n(X_n|\beta_S))\pi(\beta_S)d\beta_S \\ &\leq \exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(1+O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})) \\ &\quad \int_{(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S)\leq c_2\frac{Kq\log(n)}{n}} \exp(-\frac{n}{2}(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S))d\beta_S,\end{aligned}$$

where c_2 is the upper bound of the maximum eigenvalue of $\hat{I}_S$. Let $\xi = \sqrt{n}\hat{I}^{1/2}(\beta_S - \hat{\beta}_S)$, then we have

$$\begin{aligned}&\int_{(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S)\leq c_2\frac{Kq\log(n)}{n}} \exp(-\frac{n}{2}(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S))d\beta_S \\ = \ &(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\int_{\xi^T\xi\leq c_2Kq\log(n)}(2\pi)^{-\frac{|S|}{2}}\exp(-\frac{1}{2}\xi^T\xi)d\xi \leq (\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}},\end{aligned}$$

where the last inequality is derived from the fact that $(2\pi)^{-\frac{|S|}{2}}\exp(-\frac{1}{2}\xi^T\xi)$ exactly equals the density function of $N(0, I_{|S|})$. Now we obtain the upper bound of $Int1$

$$Int1 \leq \exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(1+O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}))(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}.$$

Similarly, we have

$$\begin{aligned} Int1: \quad &\geq \quad \exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(1-O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})) \\ &\int_{(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S)\leq c_1\frac{Kq\log(n)}{n}} \exp(-\frac{n}{2}(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S))d\beta_S, \end{aligned}$$

where c_1 is the lower bound of the minimum eigenvalue of $\hat{I}_S$. Let $\xi = \sqrt{n}\hat{I}^{1/2}(\beta_S-\hat{\beta}_S)$, this time we have

$$\begin{aligned} &\int_{(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S)\leq c_1\frac{Kq\log(n)}{n}} \exp(-\frac{n}{2}(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S))d\beta_S \\ = \quad &(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\int_{\xi^T\xi\leq c_1Kq\log(n)}(2\pi)^{-\frac{|S|}{2}}\exp(-\frac{1}{2}\xi^T\xi)d\xi \\ \geq \quad &(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\int_{\xi^T\xi\leq 5|S|\log(n)}(2\pi)^{-\frac{|S|}{2}}\exp(-\frac{1}{2}\xi^T\xi)d\xi \geq (\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}(1-\frac{1}{n^{|S|}}), \end{aligned}$$

where the first inequality holds if we choose $K > 5c_1$, and the second inequality is based on the Lemma 1 of Foygel Barber et al. (2015). Now we obtain the lower bound of $Int1$

$$Int1 \geq \exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(1-O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}))(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}(1-\frac{1}{n^{|S|}}).$$

Since $\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}$ dominates $\frac{1}{n^{|S|}}$, therefore we have

$$Int1 = \exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}(1+O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})).$$

The second term of equation (23) satisfies

$$\begin{aligned} Int2: \quad &= \quad \int_{\beta_S\in B_{\epsilon_n}-B_{\epsilon_K}} \exp(nL_n(X_n|\beta_S))\pi(\beta_S)d\beta_S \\ &\leq \quad \exp(nL_n(X_n|\hat{\beta}_S))\int_{\beta_S\in B_{\epsilon_n}-B_{\epsilon_K}}\pi(\beta_S)\exp(-\frac{nc_1}{4}||\beta_S-\hat{\beta}_S||^2)d\beta_S \\ &\leq \quad \exp(nL_n(X_n|\hat{\beta}_S))\exp(-\frac{nc_1}{4}\frac{Kq\log(n)}{n})\int_{\beta_S\in B_{\epsilon_n}-B_{\epsilon_K}}\pi(\beta_S)d\beta_S \\ &\leq \quad C_5\exp(nL_n(X_n|\hat{\beta}_S))\frac{1}{n^{Kqc_1/4}}. \end{aligned}$$

The third term of equation (23) satisfies

$$\begin{aligned} Int3: &= \int_{\beta_S \in B^c_{\epsilon_n}(\hat{\beta}_S)} \exp(nL_n(X_n|\beta_S))\pi(\beta_S)d\beta_S \\ &\leq \exp(nL_n(X_n|\hat{\beta}_S))\exp(-\frac{c'nl_n^2}{q})\int \pi(\beta_S)d\beta_S \\ &\leq \exp(nL_n(X_n|\hat{\beta}_S))\exp(-\frac{c'nl_n^2}{q}). \end{aligned}$$

Combining them together, we have

$$\begin{aligned} m(S) &= Int1 + Int2 + Int3 \\ &= \exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}(1 + O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})) \\ &\quad + \exp(nL_n(X_n|\hat{\beta}_S))O(\frac{1}{n^{Kqc_1/4}}) + \exp(nL_n(X_n|\hat{\beta}_S))O(\exp(-\frac{c'nl_n^2}{q})) \\ &= \exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\Big\{1 + O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}) \\ &\quad + \pi^{-1}(\hat{\beta}_S)(\frac{n}{2\pi})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}O(\frac{1}{n^{Kqc_1/4}}) + \pi^{-1}(\hat{\beta}_S)(\frac{n}{2\pi})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}O(\exp(-\frac{c'nl_n^2}{q}))\Big\}. \end{aligned}$$

By choosing $K > 2/c_1$, it can be verified that

$$\pi^{-1}(\hat{\beta}_S)(\frac{n}{2\pi})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}O(\frac{1}{n^{Kqc_1/4}}) = o(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}),$$

and

$$\pi^{-1}(\hat{\beta}_S)(\frac{n}{2\pi})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}O(\exp(-\frac{c'nl_n^2}{q})) = o(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}).$$

Therefore, we finally have

$$m(S) = \exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\Big\{1 + O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})\Big\}.$$

Now we return to the calculation of $E_{\beta_S|X_n,S}[L_n(X_n|\hat{\beta}_S)] - L_n(X_n|\hat{\beta}_S))$. That is,

$$\begin{aligned} &\int P(\beta_S|X_n,S)(L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S))d\beta_S = -\int_{B_{\epsilon_K}} P(\beta_S|X_n,S)\frac{1}{2}(\beta_S - \hat{\beta}_S)'\hat{I}_S(\beta_S - \hat{\beta}_S)d\beta_S \\ &+ \int_{B_{\epsilon_K}} P(\beta_S|X_n,S)r_n(\beta_S)||\beta_S - \hat{\beta}_S||^2 d\beta_S + \int_{B_{\epsilon_n} - B_{\epsilon_K}} P(\beta_S|X_n,S)(L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S))d\beta_S \\ &+ \int_{B^c_{\epsilon_n}} P(\beta_S|X_n,S)(L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S))d\beta_S \\ &= (I1) + (I2) + (I3) + (I4). \end{aligned} \tag{24}$$

The first component of equation (24) is

$$
\begin{aligned}
(I1) &= \frac{1}{m(S)} \int_{B_{\epsilon_K}} \exp(nL_n(X_n|\beta_S))\pi(\beta_S)\frac{-1}{2}(\beta_S - \hat{\beta}_S)'\hat{I}_S(\beta_S - \hat{\beta}_S)d\beta_S \\
&= \frac{1}{m(S)} \exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(1 + O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})) \\
&\int_{B_{\epsilon_K}} \exp(-n\frac{1}{2}(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S))(-1)\frac{1}{2}(\beta_S - \hat{\beta}_S)'\hat{I}_S(\beta_S - \hat{\beta}_S)d\beta_S.
\end{aligned}
$$

Let $\xi = \sqrt{n}\hat{I}^{1/2}(\beta_S - \hat{\beta}_S)$, then we have

$$
\begin{aligned}
& \int_{B_{\epsilon_K}} \exp(-n\frac{1}{2}(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S))\frac{1}{2}(\beta_S - \hat{\beta}_S)'\hat{I}_S(\beta_S - \hat{\beta}_S)d\beta_S \\
\leq & \int_{(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S) \leq c_2\frac{Kq\log(n)}{n}} \exp(-n\frac{1}{2}(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S))\frac{1}{2}(\beta_S - \hat{\beta}_S)'\hat{I}_S(\beta_S - \hat{\beta}_S)d\beta_S \\
= & (\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}} \int_{\xi^T\xi \leq c_2 Kq\log(n)} (2\pi)^{-\frac{|S|}{2}} \exp(-\frac{1}{2}\xi^T\xi)\frac{1}{2n}\xi^T\xi d\xi \leq (\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\frac{1}{2n}|S|.
\end{aligned}
$$

We also have

$$
\begin{aligned}
& \int_{B_{\epsilon_K}} \exp(-n\frac{1}{2}(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S))\frac{1}{2}(\beta_S - \hat{\beta}_S)'\hat{I}_S(\beta_S - \hat{\beta}_S)d\beta_S \\
\geq & \int_{(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S) \leq c_1\frac{Kq\log(n)}{n}} \exp(-n\frac{1}{2}(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S))\frac{1}{2}(\beta_S - \hat{\beta}_S)'\hat{I}_S(\beta_S - \hat{\beta}_S)d\beta_S \\
= & (\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\frac{1}{2n} \int_{\xi^T\xi \leq c_1 Kq\log(n)} (2\pi)^{-\frac{|S|}{2}} \exp(-\frac{1}{2}\xi^T\xi)\xi^T\xi d\xi \\
= & (\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\frac{1}{2n}\left[|S| - \int_{\xi^T\xi \geq c_1 Kq\log(n)} (2\pi)^{-\frac{|S|}{2}} \exp(-\frac{1}{2}\xi^T\xi)\xi^T\xi d\xi\right].
\end{aligned}
$$

For sufficiently large n, $\exp(\frac{1}{4}\xi^T\xi) \geq \xi^T\xi$, so

$$
\begin{aligned}
& \int_{\xi^T\xi \geq c_1 Kq\log(n)} (2\pi)^{-\frac{|S|}{2}} \exp(-\frac{1}{2}\xi^T\xi)\xi^T\xi d\xi \\
\leq & \int_{\xi^T\xi \geq c_1 Kq\log(n)} (2\pi)^{-\frac{|S|}{2}} \exp(-\frac{1}{4}\xi^T\xi)d\xi \\
\leq & (2\pi)^{-\frac{|S|}{2}}\frac{4(\pi)^{|S|/2}}{\Gamma(|S|/2)}\frac{[c_1 Kq\log(n)]^{|S|-1}}{1/4}\exp(-\frac{1}{4}c_1 Kq\log(n)) = o(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}),
\end{aligned}
$$

where the second inequality is based on the Lemma 2 of Foygel and Drton (2011). Therefore,

$$
\begin{aligned}
& \int_{B_{\epsilon_K}} \exp(-n\frac{1}{2}(\beta_S - \hat{\beta}_S)^T\hat{I}_S(\beta_S - \hat{\beta}_S))\frac{1}{2}(\beta_S - \hat{\beta}_S)'\hat{I}_S(\beta_S - \hat{\beta}_S)d\beta_S \\
= & (\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\frac{|S|}{2n}(1 + o(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})),
\end{aligned}
$$

and

$$
\begin{aligned}
(I1) &= \frac{1}{m(S)}\exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(1+O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}))(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\frac{-|S|}{2n} \\
&= \frac{\exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(1+O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}))(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\frac{-|S|}{2n}}{\exp(nL_n(X_n|\hat{\beta}_S))\pi(\hat{\beta}_S)(\frac{2\pi}{n})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}\left\{1+O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})\right\}} \\
&= -\frac{|S|}{2n}(1+O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})).
\end{aligned} \tag{25}
$$

For the second component of equation (24), we can show

$$
\begin{aligned}
|I2| &= \left|\int_{B_{\epsilon_K}} P(\beta_S|X_n,S)r_n(\beta_S)\|\beta_S-\hat{\beta}_S\|^2 d\beta_S\right| \\
\leq\ & C_6\frac{q\log(n)}{n}\sqrt{\frac{q^2\log(p)}{n}}\int_{B_{\epsilon_K}} P(\beta_S|X_n,S)d\beta_S = \frac{C_6}{n}\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}(1+o(1)),
\end{aligned}
$$

for some generic constant C_6.

For any $\beta_S \in B_{\epsilon_n}(\hat{\beta}_S) - B_{\epsilon_K}(\hat{\beta}_S)$,

$$
\begin{aligned}
L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S) &= -\frac{1}{2}(\beta_S-\hat{\beta}_S)^T\hat{I}_S(\beta_S-\hat{\beta}_S) + r_n(\beta_S)||\beta_S-\hat{\beta}_S||_2^2 \\
&= -\frac{1}{2}(\beta_S-\hat{\beta}_S)'\hat{I}_S(\beta_S-\hat{\beta}_S)(1+o(1)) \geq -c_2||\beta_S-\hat{\beta}_S||^2,
\end{aligned}
$$

where $c_2 > 0$ is the upper bound of the maximum eigenvalue of $\hat{I}_S$ for all $S \in \mathcal{P}(q)$.

Then for the third component of equation (24), we have

$$
\begin{aligned}
|I3| &= \left|\int_{B_{\epsilon_n}-B_{\epsilon_K}} P(\beta_S|X_n,S)(L_n(X_n|\beta_S)-L_n(X_n|\hat{\beta}_S))d\beta_S\right| \\
\leq\ & \sup_{\beta_S\in B_{\epsilon_n}-B_{\epsilon_K}}|L_n(X_n|\beta_S)-L_n(X_n|\hat{\beta}_S)|\int_{B_{\epsilon_n}-B_{\epsilon_K}} P(\beta_S|X_n,S)d\beta_S \\
\leq\ & c_2\frac{l_n^2}{q}\frac{1}{m(S)}\int_{B_{\epsilon_n}-B_{\epsilon_K}}\exp(nL_n(X_n|\beta_S))\pi(\beta_S)d\beta_S \\
=\ & c_2\frac{l_n^2}{q}\pi^{-1}(\hat{\beta}_S)(\frac{n}{2\pi})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}O(\frac{1}{n^{Kqc_1/4}})\left(1+O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})\right).
\end{aligned}
$$

By choosing $K > 4/c_1$, it can be verified that

$$
\pi^{-1}(\hat{\beta}_S)(\frac{n}{2\pi})^{\frac{|S|}{2}}|\hat{I}_S|^{-\frac{1}{2}}O(\frac{1}{n^{Kqc_1/4}}) = o\left(\frac{1}{n}\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}\right).
$$

Therefore

$$|I3| = \left| \int_{B_{\epsilon_n} - B_{\epsilon_K}} P(\beta_S|X_n, S)(L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S))d\beta_S \right| = o\left(\frac{1}{n} \sqrt{\frac{q^4 \log^2(n) \log(p)}{n}} \right).$$

To deal with the last component of equation (24), we should first recall for $\beta_S \in B^c_{\epsilon_n}$,

$$nL_n(X_n|\beta_S) - nL_n(X_n|\hat{\beta}_S) \leq -\frac{c' n l_n^2}{q}.$$

Therefore, for sufficiently large n, we have

$$\exp(0.5|nL_n(X_n|\beta_S) - nL_n(X_n|\hat{\beta}_S)|) > |nL_n(X_n|\beta_S) - nL_n(X_n|\hat{\beta}_S)|.$$

Now the last component of equation (24) becomes

$$\begin{aligned}
& |I4| = \left| \int_{B^c_{\epsilon_n}} P(\beta_S|X_n, S)(L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S))d\beta_S \right| \\
= & \frac{1}{m(S)} \left| \int_{B^c_{\epsilon_n}} \exp(nL_n(X_n|\beta_S))\pi(\beta_S)(L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S))d\beta_S \right| \\
= & \frac{\exp(nL_n(X_n|\hat{\beta}_S))}{nm(S)} \left| \int_{B^c_{\epsilon_n}} \pi(\beta_S)[\exp(nL_n(X_n|\beta_S) - nL_n(X_n|\hat{\beta}_S))](nL_n(X_n|\beta_S) - nL_n(X_n|\hat{\beta}_S))d\beta_S \right| \\
\leq & \frac{\exp(nL_n(X_n|\hat{\beta}_S))}{nm(S)} \left| \int_{B^c_{\epsilon_n}} \pi(\beta_S)[\exp(0.5nL_n(X_n|\beta_S) - 0.5nL_n(X_n|\hat{\beta}_S))]d\beta_S \right| \\
\leq & \frac{\exp(nL_n(X_n|\hat{\beta}_S))}{nm(S)} \exp(-\frac{c' n l_n^2}{2q}) \left| \int_{B^c_{\epsilon_n}} \pi(\beta_S)d\beta_S \right| \\
\leq & \exp(-\frac{c' n l_n^2}{2q}) \frac{1}{n\pi(\hat{\beta}_S)(\frac{2\pi}{n})^{\frac{|S|}{2}} |\hat{I}_S|^{-\frac{1}{2}} \left\{ 1 + O(\sqrt{\frac{q^4 \log^2(n) \log(p)}{n}}) \right\}}.
\end{aligned}$$

Recall we have already verified

$$\pi^{-1}(\hat{\beta}_S)(\frac{n}{2\pi})^{\frac{|S|}{2}} |\hat{I}_S|^{-\frac{1}{2}} O(\exp(-\frac{c' n l_n^2}{q})) = o\left(\sqrt{\frac{q^4 \log^2(n) \log(p)}{n}} \right).$$

Therefore we finally have

$$\left| \int_{B^c_{\epsilon_n}} P(\beta_S|X_n, S)(L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S))d\beta_S \right| = o\left(\frac{1}{n} \sqrt{\frac{q^4 \log^2(n) \log(p)}{n}} \right).$$

Combining all the steps above, we conclude

$$\begin{aligned} E_{\beta_S|X_n,S}[L_n(X_n|\beta_S)] - L_n(X_n|\hat{\beta}_S)] &= \int P(\beta_S|X_n,S)(L_n(X_n|\beta_S) - L_n(X_n|\hat{\beta}_S))d\beta_S \\ &= -\frac{|S|}{2n}\left(1 + O(\sqrt{\frac{q^4\log^2(n)\log(p)}{n}})\right). \end{aligned}$$

A.4 Proof of Lemma 3

For sufficiently large n, with probability going to 1, all statements in Lemma 1 and Lemma 2 hold and the following proof are based on those statements. Without loss of generality, for any $S^* \subset S$, we can extend β_{S^*} to a vector in $\mathbb{R}^S$ by adding zeros in each $j \in S\backslash S^*$. We denote such a vector by $\beta_{S^*,S}$.

Applying Lemma 1, for any $S^* \subset S$ and $|S| \leq q$, we have

$$||\hat{\beta}_S - \beta_{S,0}||_2 = O((\frac{q\log(p)}{n})^{\frac{1}{4}}) = o(\frac{1}{\sqrt{q}}),$$

and

$$||\hat{\beta}_{S^*} - \beta_{S^*,0}||_2 = O((\frac{q\log(p)}{n})^{\frac{1}{4}}) = o(\frac{1}{\sqrt{q}}).$$

Hence,

$$\|\hat{\beta}_S - \hat{\beta}_{S^*,S}\|_2 = o(\frac{1}{\sqrt{q}}).$$

Applying Lemma 2, we obtain

$$\begin{aligned} &n(L_n(X_n|\hat{\beta}_S) - L_n(X_n|\hat{\beta}_{S^*})) \\ &= \frac{1}{2}(\hat{\beta}_S - \hat{\beta}_{S^*,S})'n\hat{I}_S(\hat{\beta}_S)(\hat{\beta}_S - \hat{\beta}_{S^*,S}) + nO(\sqrt{q}||\hat{\beta}_S - \hat{\beta}_{S^*,S}||^3 \vee \sqrt{\frac{q^2\log(p)}{n}}||\hat{\beta}_S - \hat{\beta}_{S^*,S}||^2). \end{aligned} \tag{26}$$

It follows from Lemma 1 that the smallest eigenvalue of $\hat{I}_S(\hat{\beta}_S)$ is uniformly bounded away from 0. Therefore, it is easy to verify

$$nO(\sqrt{q}||\hat{\beta}_S - \hat{\beta}_{S^*,S}||^3 \vee \sqrt{\frac{q^2\log(p)}{n}}||\hat{\beta}_S - \hat{\beta}_{S^*,S}||^2) = o(\frac{1}{2}(\hat{\beta}_S - \hat{\beta}_{S^*,S})'n\hat{I}_S(\hat{\beta}_S)(\hat{\beta}_S - \hat{\beta}_{S^*,S})),$$

which implies the residual term of (26) does not affect our results asymptotically and can thus be ignored.

From the proof of statement (iii) of Lemma 1, we know

$$S_n(\hat{\beta}_S) - S_n(\beta_{S,0}) = -\hat{I}_S(\hat{\beta}_S)(\hat{\beta}_S - \beta_{S,0}) + O(\sqrt{\frac{q^2\log(p)}{n}})||\hat{\beta}_S - \beta_{S,0}||_2. \tag{27}$$

Since the score function $S_n(\beta_S) := \frac{\partial L_n(X_n|\beta_S)}{\partial \beta_S}$ equals to 0 when $\beta_S = \hat{\beta}_S$,

$$-S_n(\beta_{S,0}) = -\hat{I}_S(\hat{\beta}_S)(\hat{\beta}_S - \beta_{S,0}) + O(\sqrt{\frac{q^2\log(p)}{n}})||\hat{\beta}_S - \beta_{S,0}||_2.$$

Recall that the minimum eigenvalue of $\hat{I}_S(\hat{\beta}_S)$ is uniformly bounded away from 0, so

$$O(\sqrt{\frac{q^2\log(p)}{n}})||\hat{\beta}_S - \beta_{S,0}||_2 = o(I_S(\hat{\beta}_S)(\hat{\beta}_S - \beta_{S,0})),$$

which implies that the residual term of (27) does not affect our results asymptotically and can be ignored.

In the following, for notational simplicity, we rewrite $\hat{I}_S(\hat{\beta}_S)$ as $\hat{I}_S$ and rewrite $I_S(\beta_{S,0})$ as I_S. Then we have

$$\sqrt{n}(\hat{\beta}_S - \beta_{S,0}) = (1+o(1))\sqrt{n}\hat{I}_S^{-1}S_n(\beta_{S,0}) = (1+o(1))\sqrt{n}I_S^{-1}S_n(\beta_{S,0}), \tag{28}$$

holds uniformly for all $S \supset S^*$, and $S \in \mathcal{P}(q)$.

The same argument holds for S^* as well, i.e.,

$$\sqrt{n}(\hat{\beta}_{S^*} - \beta_{S^*,0}) = (1+o(1))\sqrt{n}\hat{I}_{S^*}^{-1}S_n(\beta_{S^*,0}) = (1+o(1))\sqrt{n}I_{S^*}^{-1}S_n(\beta_{S^*,0}). \tag{29}$$

For a vector x, denote the subvector of x indexed by S as x^S. Without loss of generality, denote $x = (x^{S^*}, x^{S\setminus S^*})$ for any $x \in \mathbb{R}^S$. By combining equation (28) and equation (29), we have

$$\sqrt{n}(\hat{\beta}_S - \hat{\beta}_{S^*,S}) = (1+o(1))\left(I_S^{-1}\sqrt{n}S_n(\beta_{S,0}) - (I_{S^*}^{-1}\sqrt{n}S_n(\beta_{S^*,0}),0)\right). \tag{30}$$

Denote $v = \sqrt{n}S_n(\beta_{S,0})$ and $v = (v^{S^*}, v^{S\setminus S^*}) := (V_1, V_2)$. Then v has the following properties

$$E(v) = \mathbf{0}, \quad E[vv'] = nI_S, \quad E[v^{S^*}(v^{S^*})'] = nI_{S^*}.$$

Therefore, equation (30) can be re-expressed as

$$\begin{aligned}\sqrt{n}(\hat{\beta}_S - \hat{\beta}_{S^*,S}) &= (1+o(1))I_S^{-1}\left((V_1,V_2) - I_S(I_{S^*}^{-1}V_1, 0)\right) \\ &= (1+o(1))I_S^{-1}\begin{bmatrix} 0 & 0 \\ -E[V_2V_1']E[V_1V_1']^{-1} & I_2 \end{bmatrix}\begin{bmatrix} V_1 \\ V_2 \end{bmatrix}.\end{aligned}$$

Go back to the major component of equation (26): For a matrix M, define $M(S_1, S_2)$ as the submatrix of

M indexed by (S_1, S_2). Then,

$$\begin{aligned}
& \frac{1}{2}(\hat{\beta}_S - \hat{\beta}_{S^*,S})' n\hat{I}_S(\hat{\beta}_S)(\hat{\beta}_S - \hat{\beta}_{S^*,S}) \\
= \; & (1+o(1))\frac{1}{2}(\hat{\beta}_S - \hat{\beta}_{S^*,S})' n I_S(\hat{\beta}_S - \hat{\beta}_{S^*,S}) \\
= \; & \frac{1+o(1)}{2}[V_1', V_2'] \begin{bmatrix} 0 & 0 \\ -E[V_2V_1']E[V_1V_1']^{-1} & I_2 \end{bmatrix}' I_S^{-1} I_S I_S^{-1} \begin{bmatrix} 0 & 0 \\ -E[V_2V_1']E[V_1V_1']^{-1} & I_2 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \\
= \; & \frac{1+o(1)}{2}(V_2 - E[V_2V_1']E[V_1V_1']^{-1}V_1)'\Sigma(S,S^*)(V_2 - E[V_2V_1']E[V_1V_1']^{-1}V_1),
\end{aligned}$$

where $\Sigma(S,S^*) := I_S^{-1}(S\backslash S^*, S\backslash S^*)$ is the submatrix of I_S^{-1} indexed by $(S\backslash S^*, S\backslash S^*)$.

Define

$$H(S,S^*) := \Sigma(S,S^*)^{\frac{1}{2}}(V_2 - E[V_2V_1']E[V_1V_1']^{-1}V_1).$$

Then

$$\frac{1}{2}(\hat{\beta}_S - \hat{\beta}_{S^*,S})' n\hat{I}_S(\hat{\beta}_S)(\hat{\beta}_S - \hat{\beta}_{S^*,S}) = \frac{(1+o(1))}{2}H'(S,S^*)H(S,S^*).$$

To prove the main results, we only need to show that for sufficiently large n and any $\eta > 0$ and $\zeta > 1/5$,

$$\sup_{S\supset S^*, S\in\mathcal{P}(q)} \frac{H'(S,S^*)H(S,S^*)}{|S\backslash S^*|} \leq 2(1+\eta)\log(n^\zeta p),$$

holds with probability going to 1.

Define $A_{S,n} := \sqrt{2(1+\eta)\log(n^\zeta p)|S\backslash S^*|}$, then

$$Pr\left(\frac{H'(S,S^*)H(S,S^*)}{|S\backslash S^*|} \geq 2(1+\eta)\log(n^\zeta p)\right) = Pr\left(||H(S,S^*)||_2 \geq A_{S,n}\right).$$

First, for any $u \in \mathbb{R}^S$, $||u||_2 = 1$, we consider to bound $Pr\left(u'H(S,S^*) \geq \sqrt{(1+\epsilon)2|S\backslash S^*|\log(n^\zeta p)}\right)$.

Rewrite

$$u'H(S,S^*) := \frac{1}{\sqrt{n}}\sum_{i=1}^n u'\Sigma(S,S^*)^{\frac{1}{2}}(\frac{1}{\sqrt{n}}V_{2,i} - E[V_2V_1']E[V_1V_1']^{-1}\frac{1}{\sqrt{n}}V_{1,i}),$$

where $\frac{1}{\sqrt{n}}V_{1,i} = \frac{\partial \log f(x_i|\beta_{S,0})}{\partial \beta_{S^*}}$ and $\frac{1}{\sqrt{n}}V_{2,i} = \frac{\partial \log f(x_i|\beta_{S,0})}{\partial \beta_{S\backslash S^*}}$. For simplicity, we further define

$$Y_i(S) := u'\Sigma(S,S^*)^{\frac{1}{2}}(\frac{1}{\sqrt{n}}V_{2,i} - E[V_2V_1']E[V_1V_1']^{-1}\frac{1}{\sqrt{n}}V_{1,i}).$$

Then it becomes

$$u'H(S,S^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i(S).$$

It is easy to see that

$$E[Y_i(S)] = 0.$$

Moreover, since

$$\Sigma(S, S^*) = n\left(E[V_2V_2'] - E[V_2V_1']E[V_1V_1']^{-1}E[V_1V_2']\right)^{-1},$$

then

$$E[(\frac{1}{\sqrt{n}}V_{2,i} - E[V_2V_1']E[V_1V_1']^{-1}\frac{1}{\sqrt{n}}V_{1,i})(\frac{1}{\sqrt{n}}V_{2,i} - E[V_2V_1']E[V_1V_1']^{-1}\frac{1}{\sqrt{n}}V_{1,i})'] = \Sigma(S, S^*)^{-1},$$

which implies

$$E[Y_i(s)^2] = u'I_{S\setminus S^*}u = 1.$$

By condition (A2), $|Y_i(S)| \leq c_3^{\frac{1}{2}}C_2\sqrt{q^3} := K\sqrt{q^3}$, where c_3 is a generic constant which bounds the maximum eigenvalue of $\Sigma(S, S^*)$ from the above. By condition (A1), such a constant exists.

By applying the Bernstein's Inequality, we have

$$Pr(|\sum_{i=1}^{n} Y_i| > x) \leq 2\exp(-\frac{1}{2}\frac{x^2}{n + K\sqrt{q^3}x/3}),$$

for any $x > 0$. Plugging in $x = \sqrt{n}\sqrt{2(1+\epsilon)|S\setminus S^*|\log(n^\zeta p)}$, we get

$$Pr(u'H(S, S^*) \geq \sqrt{2(1+\epsilon)|S\setminus S^*|\log(n^\zeta p)}) = Pr(|\sum_{i=1}^{n} Y_i| > x) \leq 2\exp(\frac{-1}{2}\frac{2(1+\epsilon)n|S\setminus S^*|\log(n^\zeta p)}{n + K\sqrt{q^3}x}).$$

By the assumption that $\frac{q^5\log(p)}{n} \to 0$, we also have

$$K\sqrt{q^3}x \leq K\sqrt{n}\sqrt{2(1+\epsilon)q^4\log(n^\zeta p)} = o(n).$$

Hence, for sufficiently large n,

$$Pr(u'H(S, S^*) \geq \sqrt{2(1+\epsilon)|S\setminus S^*|\log(n^\zeta p)}) \leq 2\exp(-(1+\epsilon/2)|S\setminus S^*|\log(n^\zeta p)).$$

To link this inequality to our final objective, $Pr(||H(S, S^*)||_2 \geq A_{S,n})$, we rely on Lemma 2 in Chen and Chen (2012), which states that for a given $\delta_n > 0$, there exists a finite set of unit vectors $\mathbf{U}_S(\delta_n) \subset \mathbb{R}^{S\setminus S^*}$ such that for all $\mathbf{v} \in \mathbb{R}^{S\setminus S^*}$, $||\mathbf{v}||_2 \leq (1+\delta_n)\max_{\mathbf{u}\in\mathbf{U}_S(\delta_n)}\mathbf{u}'\mathbf{v}$ holds. Since $\mathbf{U}_S(\delta_n)$ is a finite set, following the proof of Lemma 2 of Chen and Chen (2012), the cardinality of $\mathbf{U}(\delta_n)$ is bounded by $|\mathbf{U}_S(\delta_n)| \leq K(\frac{1}{\delta_n})^{|S\setminus S^*|}$ for some generic constant $K > 0$.

Hence, for a fixed model S with $S \supset S^*$ and $|S| = d \leq q$, we can pick δ_n such that $(1+\delta_n)\sqrt{1+\epsilon_n} =$

$\sqrt{1+\eta}$.

$$\begin{aligned} P\{||H(S,S^*)||_2 \geq A_{S,n}\} &\leq \sum_{u\in U(\delta_n)} P(u'H(S,S^*) \geq \sqrt{2(1+\epsilon)|S\backslash S^*|\log(n^\zeta p)}) \\ &\leq |U(\delta_n)|2\exp(-(1+\epsilon/2)|S\backslash S^*|\log(n^\zeta p)). \end{aligned}$$

Then

$$\begin{aligned} & P\{\exists\ S \text{ with } S\supset S^* \text{ and } |S|\leq q, ||H(S,S^*)||_2 \geq A_{S,n}\} \\ \leq & \sum_{S\supset S^* \text{ and } |S|\leq q} P\{||H(S,S^*)||_2 \geq A_{S,n}\} \\ \leq & \sum_{S\supset S^* \text{ and } |S|\leq q} |U_S(\delta_n)|2\exp(-(1+\epsilon/2)|S\backslash S^*|\log(n^\zeta p)) \\ \leq & \sum_{d'=1}^{q-|S^*|}\binom{p}{d'}|U_S(\delta_n)|2\exp(-(1+\epsilon/2)|S\backslash S^*|\log(n^\zeta p)) \\ \leq & \sum_{d'=1}^{q-|S^*|} 2\exp(d'\log(p)+\log(|U_S(\delta_n)|)-(1+\epsilon/2)d'\log(n^\zeta p))\} \\ \leq & \sum_{d'=1}^{q-|S^*|} 2\exp(-(\epsilon/2)d'\log(p)-d'\log(n^\zeta)+\log(|U_S(\delta_n)|)). \end{aligned}$$

Since $\log(|U_S(\delta_n)|)\leq \log(K)+|S\backslash S^*|\log(\frac{\sqrt{|S\backslash S^*|}}{\delta_n})$, for any fixed δ_n, we have

$$\exp(-(\epsilon/2)d'\log(p)-d'\log(n^\zeta)+\log(|U_S(\delta_n)|)) \leq C\exp(-d'(\epsilon/2\log(p)+\log(n^\zeta)-\log(d'))),$$

for some constant $C>0$. Since $\frac{q^5\log(p)}{n}\to 0$, we have $\log(d')=o(n^{\frac{1}{5}})$. So if $\zeta>\frac{1}{5}$, then

$$C\exp(-d'(\epsilon/2\log(p)+\log(n^\zeta)-\log(d'))) \leq C\exp(-d'\log(n^{\zeta-\frac{1}{5}})).$$

Therefore,

$$\begin{aligned} & P\{\exists\ S \text{ with } S\supset S^* \text{ and } |S|\leq q, ||H(S,S^*)||_2 \geq A_{S,n}\} \\ \leq & \sum_{d'=1}^{q-|S^*|} 2\exp(-(\epsilon/2)d'\log(p)-d'\log(n^\zeta)+\log(|U_S(\delta_n)|)) \\ \leq & \sum_{d'=1}^{\infty} 2C\exp(-d'\log(n^{\zeta-\frac{1}{5}})) \leq Cn^{-(\zeta-\frac{1}{5})} = o(1). \end{aligned}$$

A.5 Proof of Theorem 2

The proof of Theorem 2 is based on the proof of Theorem 2 of Foygel and Drton (2011). For sufficiently large n, with probability going to 1, all statements in Lemma 1, Lemma 2 and Lemma 3 hold.

Let S^* be the true model and define $A_0 = \{S : S^* \subset S, |S| \leq q\}$ and $A_1 = \{S : S^* \nsubseteq S, |S| \leq q\}$, then we consider two cases.

Case 1: $S \in A_0$**.** In this case, we have

$$\begin{aligned} & AEBIC(S) - AEBIC(S^*) \\ = \; & -2nE_{\beta_S|X_n,S}[L_n(X_n|\beta_S)] + |S|\log(n) + 2\gamma|S|\log(p) \\ & -\Big\{ -2nE_{\beta_{S^*}|X_n,S^*}[L_n(X_n|\beta_{S^*})] + |S^*|\log(n) + 2\gamma|S^*|\log(p) \Big\} \\ = \; & -2nE_{\beta_S|X_n,S}[L_n(X_n|\beta_S)] + 2nE_{\beta_{S^*}|X_n,S^*}[L_n(X_n|\beta_{S^*})] + |S\backslash S^*|\log(n) + 2\gamma|S\backslash S^*|\log(p). \end{aligned}$$

From Theorem 1, we know

$$E_{\beta_S|X_n,S}[L_n(X_n|\beta_S)] = L_n(X_n|\hat{\beta}_S) - \frac{|S|}{2n} + O(\frac{|S|}{n}\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}),$$

and

$$E_{\beta_{S^*}|X_n,S^*}[L_n(X_n|\beta_{S^*})] = L_n(X_n|\hat{\beta}_{S^*}) - \frac{|S^*|}{2n} + O(\frac{|S^*|}{n}\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}).$$

Therefore,

$$\begin{aligned} & AEBIC(S) - AEBIC(S^*) \\ = \; & |S\backslash S^*| + O\left(q\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}\right) + |S\backslash S^*|\log(n) + 2\gamma|S\backslash S^*|\log(p) - 2n(L_n(X_n|\hat{\beta}_S) - L_n(X_n|\hat{\beta}_{S^*})). \end{aligned}$$

From Lemma 3, we also know

$$2n(L_n(X_n|\hat{\beta}_S) - L_n(X_n|\hat{\beta}_{S^*})) \leq 2(1+\eta)|S\backslash S^*|\log(n^\zeta p),$$

for any small η and sufficiently large n. Hence,

$$\begin{aligned} & AEBIC(S) - AEBIC(S^*) \\ \geq \; & |S\backslash S^*| + O\left(q\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}\right) + |S\backslash S^*|\log(n) + 2\gamma|S\backslash S^*|\log(p) - 2(1+\eta)|S\backslash S^*|\log(n^\zeta p). \end{aligned}$$

For $\gamma > 1$, pick $\eta = \min(\frac{\gamma-1}{2}, \xi)$, where $\xi > 0$ is a small constant, then $2\gamma|S\backslash S^*|\log(p) - 2(1+\eta)|S\backslash S^*|\log(p) \geq \frac{\gamma-1}{2}|S\backslash S^*|\log(p)$. In addition, it is obvious that $\log(p)$ dominates $O\left(q\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}\right)$, as $\frac{q^6\log^2(n)}{n\log(p)} \to 0$. Hence, for sufficiently large n, we have

$$2\gamma|S\backslash S^*|\log(p) - 2(1+\eta)|S\backslash S^*|\log(p) + O\left(q\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}\right) > 0.$$

Choose $\zeta \in (\frac{1}{5}, \frac{1}{2+2\xi})$, for sufficiently large n, we also have

$$|S\backslash S^*|(1+\log(n)) > 2(1+\eta)|S\backslash S^*|\log(n^\zeta).$$

Combining them together, we finally proved $AEBIC(S) - AEBIC(S^*) > 0$ for sufficiently large n.

Case 2: $S \in A_1$**.** Let $S' = S \cup S^*$ and $\hat{\beta}_{S,S'}$ denote a vector corresponding to model S', which is generated by augmenting $\hat{\beta}_S$ with zeros in $S'\backslash S$. Following a similar procedure as in case 1, we have

$$\begin{aligned}
&AEBIC(S) - AEBIC(S^*) \\
=& -2nE_{\beta_S|X_n,S}[L_n(X_n|\beta_S)] + 2nE_{\beta_{S^*}|X_n,S^*}[L_n(X_n|\beta_S)] + (|S| - |S^*|)\log(n) + 2\gamma(|S| - |S^*|)\log(p) \\
=& -2n(L_n(X_n|\hat{\beta}_S) - L_n(X_n|\hat{\beta}_{S^*})) + (|S| - |S^*|) + (|S| - |S^*|)\log(n) \\
&+ 2\gamma(|S| - |S^*|)\log(p) + O\left(q\sqrt{\frac{q^4\log^2(n)\log(p)}{n}}\right) \\
\geq& -2n(L_n(X_n|\hat{\beta}_S) - L_n(X_n|\hat{\beta}_{S^*})) - q(\log(n) + 2\gamma\log(p) + 1 + o(1)) \\
\geq& -2n(L_n(X_n|\hat{\beta}_{S,S'}) - L_n(X_n|\hat{\beta}_{S^*})) - q(\log(n) + 2\gamma\log(p) + 2). \qquad (31)
\end{aligned}$$

By optimality, $L_n(X_n|\hat{\beta}_{S^*}) \geq L_n(X_n|\beta_{S^*,0}) = L_n(X_n|\beta_{S',0})$, since $\beta_{S^*,0}$ equals to $\beta_{S',0}$ up to components of zeros in $S'\backslash S^*$. Then

$$-2n(L_n(X_n|\hat{\beta}_{S,S'}) - L_n(X_n|\hat{\beta}_{S^*})) \geq -2n(L_n(X_n|\hat{\beta}_{S,S'}) - L_n(X_n|\beta_{S',0})).$$

From the proof of Lemma 1, we have

$$\begin{aligned}
|L_n(X_n|\hat{\beta}_S) - L(\hat{\beta}_S)| &= O(\sqrt{\frac{q\log(p)}{n}}), \\
|L_n(X_n|\beta_{S^*,0}) - L(\beta_{S^*,0})| &= O(\sqrt{\frac{q\log(p)}{n}}).
\end{aligned}$$

By condition (A6) and (A7), we have

$$L(\beta_{S',0}) - L(\hat{\beta}_{S,S'}) \geq \min(1, c||\hat{\beta}_{S,S'} - \beta_{S',0}||_2^2) \geq \min(1, c[\min_j |\beta_{S^*,0}|]^2)$$
$$\succ \min(1, c[(\frac{q\log(p)}{n})^{\frac{1}{4}}]^2) = c\sqrt{\frac{q\log(p)}{n}}.$$

Combining them together, we obtain

$$\begin{aligned} & -2n(L_n(X_n|\hat{\beta}_{S,S'}) - L_n(X_n|\beta_{S',0})) \\ \geq \; & -2n\left(|L_n(X_n|\hat{\beta}_{S,S'}) - L(\hat{\beta}_{S,S'})| + |L_n(X_n|\beta_{S',0}) - L(\beta_{S',0})| - (L(\hat{\beta}_{S',0}) - L(\hat{\beta}_{S,S'})\right) \\ \succ \; & n\sqrt{\frac{q\log(p)}{n}} = \sqrt{nq\log(p)}. \end{aligned}$$

Back to inequality (31), for sufficiently large n, we have

$$-2n(L_n(X_n|\hat{\beta}_{S,S'}) - L_n(X_n|\hat{\beta}_{S^*})) - q(\log(n) + 2\gamma\log(p) + 2) > 0,$$

since the assumption $\frac{q\log(p)}{n} \to 0$ leads to $\frac{\sqrt{nq\log(p)}}{q\log(p)} \to \infty$. Hence, $AEBIC(S) > AEBIC(S^*)$ holds.

References

Akaike, H. (1974), "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, 19, 716–723.

Ando, T. (2010), *Bayesian Model Selection and Statistical Modeling*, New York: Chapman & Hall.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009), "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, 37, 1705–1732.

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 criteria for model selection with large model spaces," *Biometrika*, 95, 759–771.

— (2012), "Extended BIC for small-n-large-p sparse GLM," *Statistica Sinica*, 22, 555–574.

Chen, Z. and Luo, S. (2013), "Selection consistency of EBIC for GLIM with non-canonical links and diverging number of parameters," *Statistics and Its Interface*, 6, 275–284.

Efron, B. (2009), "Empirical Bayes Estimates for Large-Scale Prediction Problems," *Journal of the American Statistical Association*, 104, 1015–1028.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Foygel, R. and Drton, M. (2011), "Bayesian model choice and information criteria in sparse generalized linear models," *ArXiv e-prints*.

Foygel Barber, R., Drton, M., and Tan, K. M. (2015), "Laplace Approximation in High-dimensional Bayesian Regression," *ArXiv e-prints*.

Ge, D., Wang, Z., Ye, Y., and Yin, H. (2016), "Strong NP-hardness result for regularized L_q-minimization problems with concave penalty functions," *arXiv:1501.00622v3*.

Green, P. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Haughton, D. M. A. (1988), "On the Choice of a Model to Fit Data from an Exponential Family," *Ann. Statist.*, 16, 342–355.

Huo, X. and Chen, J. (2010), "Complexity of penalized likelihood estimation," *Journal of Statistical Computation and Simulation*, 80, 747–759.

Jiang, W. (2007), "Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities," *Ann. Statist.*, 35, 1487–1511.

Johnson, V. E. and Rossell, D. (2012), "Bayesian Model Selection in High-Dimensional Settings," *Journal of the American Statistical Association*, 107, 649–660.

Kass, R. and Raftery, A. (1995), "Bayes factors and model uncertainty," *Journal of the American Statistical Association*, 90, 773–795.

Liang, F., Liu, C., and Carroll, R. (2007), "Stochastic Approximation in Monte Carlo Computation," *Journal of the American Statistical Association*, 102, 305–320.

Liang, F., Song, Q., and Yu, K. (2013), "Bayesian Subset Modeling for High-Dimensional Generalized Linear Models," *Journal of the American Statistical Association*, 2013.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models (Second edition)*, London: Chapman & Hall.

Nishii, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *Ann. Statist.*, 12, 758–765.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Ann. Statist.*, 6, 461–464.

Shin, M., Bhattacharya, A., and Johnson, V. E. (2015), "Scalable Bayesian Variable Selection Using Nonlocal Prior Densities in Ultrahigh-Dimensional Settings," *ArXiv e-prints*.

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., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002), "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, 1, 203 – 209.

Song, Q. and Liang, F. (2015a), "High-Dimensional Variable Selection With Reciprocal L1-Regularization," *Journal of the American Statistical Association*, 110, 1607–1620.

— (2015b), "A Split-and-Merge Bayesian Variable Selection Approach for Ultra-high dimensional Regression," *Journal of the Royal Statistical Society, Series B*, 77, 947–972.

Spiegelhalter, D., Best, N., Carlin, B., and Germany, B. (2014), "The deviance information criterion: 12 years on," *Journal of the Royal Statistical Society, Series B*, 76, 485–493.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society Series B*, 64, 583–639.

Tibshirani, R. (1994), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wahba, G. and Craven, P. (1979), "Smoothing Noisy Data with Spline Functions. Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation." *Numerische Mathematik*, 31, 377–404.

Watanabe, S. (2010), "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, 11, 3571–3594.

Zhang, C.-H. (2010), "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, 38, 894–942.

Zhao, P. and Yu, B. (2006), "On model selection consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67, 301–320.