Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2014)
76, *Part* 1, *pp.* 141–167

Model selection principles in misspecified models

Jinchi Lv

University of Southern California, Los Angeles, USA

and Jun S. Liu

Harvard University, Cambridge, USA

[Received April 2010. Final revision January 2013]

Summary. Model selection is of fundamental importance to high dimensional modelling featured in many contemporary applications. Classical principles of model selection include the Bayesian principle and the Kullback–Leibler divergence principle, which lead to the Bayesian information criterion and Akaike information criterion respectively, when models are correctly specified. Yet model misspecification is unavoidable in practice. We derive novel asymptotic expansions of the two well-known principles in misspecified generalized linear models, which give the generalized Bayesian information criterion and generalized Akaike information criterion. A specific form of prior probabilities motivated by the Kullback–Leibler divergence principle leads to the generalized Bayesian information criterion with prior probability, GBIC_p, which can be naturally decomposed as the sum of the negative maximum quasi-log-likelihood, a penalty on model dimensionality, and a penalty on model misspecification directly. Numerical studies demonstrate the advantage of the new methods for model selection in both correctly specified and misspecified models.

Keywords: Akaike information criterion; Bayesian information criterion; Bayesian principle; Generalized Akaike information criterion; Generalized Bayesian information criterion; Generalized Bayesian information criterion with prior probability; Kullback–Leibler divergence principle; Model misspecification; Model selection

1. Introduction

We consider in this paper the type of data $(y_i, x_{i1}, \ldots, x_{ip})_{i=1}^n$, where the y_is are independent observations of the response variable Y conditional on its covariates, or explanatory variables, $(x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. When the dimensionality or the number of covariates p is large compared with the sample size n, it is desirable to produce sparse models that involve only a small subset of predictors. With such models one can improve the prediction accuracy and enhance the model interpretability. See, for example, Fan and Lv (2010) for an overview of recent progress in high dimensional variable selection problems. A natural and fundamental problem that arises from such a task is how to compare models with different sets of predictors.

Two classical principles of model selection are the Kullback–Leibler (KL) divergence principle and the Bayesian principle, which lead to the Akaike information criterion AIC by Akaike (1973, 1974) and Bayesian information criterion BIC by Schwarz (1978) respectively, when the models are correctly specified. AIC and BIC have been proved to be powerful tools for model selection in various settings; see, for example, Burnham and Anderson (1998) for an account

Address for correspondence: Jinchi Lv, Information and Operations Management Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA.
E-mail: jinchilv@marshall.usc.edu

 1369–7412/14/76141

of these developments. Stone (1977) showed heuristically the asymptotic equivalence of AIC and leave-one-out cross-validation. Bozdogan (1987) studied asymptotic properties of AIC and BIC, showing that, in parametric settings, AIC asymptotically has a positive probability of overestimating the true dimension, whereas BIC is asymptotically consistent in estimating the true model. The model selection consistency of BIC has also been shown in Casella *et al.* (2009). Hall (1990) compared AIC with KL cross-validation in the setting of histogram density estimation. Yang and Barron (1998) studied the asymptotic property of model selection criteria that are related to AIC and minimum description length principles in density estimation. Other work on model selection includes the risk inflation criterion (Foster and George, 1994), the generalized information criterion (Konishi and Kitagawa, 1996), a Bayesian measure using the deviance information criterion (Spiegelhalter *et al.*, 2002; Gelman *et al.*, 2004), model evaluation using the absolute prediction error (Tian *et al.*, 2007), tuning parameter selection in a penalization method (Wang *et al.*, 2007; Fan and Tang, 2013), the 'parametricness' index (Liu and Yang, 2011) and many extensions of AIC and BIC (see, for example, Bozdogan (1987, 2000), Bogdan *et al.* (2004), Chen and Chen (2008) and Żak-Szatkowska and Bogdan (2011)).

In parametric modelling, one chooses *a priori* a family of distributions and then finds a model in this family that best fits the data, often on the basis of the maximum likelihood principle. Despite this common practice, it has been a common wisdom in statistics that 'all models are wrong, but some are more useful than others'. Thus, in a broad sense all models are misspecified. Neither AIC nor BIC, however, explicitly accounts for such model misspecification.

When the model family is misspecified, evaluating whether a model selection procedure is good or bad becomes more subtle because we can no longer simply say that a selected model is 'correct' or 'incorrect'. We can mathematically define that a procedure is consistent if it can asymptotically select the model in the family that is theoretically closest to the true model, according to a certain discrepancy measure such as the KL divergence. By the same token, we can define a procedure to be better than others if on average it can select a model that is closer to the true model than others, according to a certain discrepancy measure. In practice, this type of discrepancy computation can, however, be non-trivial. We content ourselves with the scenario of generalized linear models (GLMs) (McCullagh and Nelder, 1989) with p predictors. In this case, a procedure is better than another if on average it can select more relevant predictors than the other. Prediction is also another important measure for model evaluation.

In this paper, we derive asymptotic expansions of the Bayesian and KL divergence principles in misspecified GLMs. The general idea appears also applicable to other model settings. The technical results lead us to propose the generalized BIC GBIC and the generalized AIC GAIC. A specific choice of prior probabilities for the models motivated by the KL divergence principle leads us to propose the generalized BIC with prior probability, GBIC_p, which can be written as the sum of the negative maximum quasi-log-likelihood, a penalty on model dimensionality, and a penalty on model misspecification directly. Our results can also be viewed as an extension of the classical approaches to non-independent and identically distributed (non-IID) settings.

The rest of the paper is organized as follows. In Section 2, we discuss quasi-maximum-likelihood estimation in misspecified GLMs and two classical model selection principles. We investigate asymptotic expansions of these principles in misspecified models and present resulting model selection criteria in Section 3. In Section 4, we provide all the technical conditions and investigate asymptotic properties of the quasi-maximum likelihood estimator (QMLE). We present several numerical examples to illustrate the finite sample performance of these model selection criteria in both correctly specified and misspecified models in Section 5. In Section 6,

we provide some discussions of our results and their implications as well as some possible extensions. All technical details are relegated to Appendix A and Appendix B.

2. Model misspecification and model selection principles

2.1. Quasi-maximum-likelihood estimation

To facilitate the presentation, we introduce some notation and state some basic facts about the QMLE in misspecified GLMs with deterministic design matrices. Conditionally on the covariates, we assume that the n-dimensional random-response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ has a true unknown distribution G_n with density function

$$g_n(\mathbf{y}) = \prod_{i=1}^{n} g_{n,i}(y_i), \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. Each observation y_i of the response variable may depend on all or some of the p predictors x_{ij}. Model (1) entails that all components of $\mathbf{Y}$ are independent but not necessarily identically distributed.

Consider an arbitrary subset $\mathfrak{M} \subset \{1, \ldots, p\}$ with $d = |\mathfrak{M}| \leqslant n \wedge p$. Define $\mathbf{x}_i = (x_{ij} : j \in \mathfrak{M})^{\mathrm{T}}$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, an $n \times d$ deterministic design matrix. Since the true model G_n is unknown, we choose a family of GLMs $F_n(\cdot, \boldsymbol{\beta})$ as our working models, with density function

$$f_n(\mathbf{z}, \boldsymbol{\beta})\, \mathrm{d}\mu_0(\mathbf{z}) = \prod_{i=1}^{n} f_0(z_i, \theta_i)\, \mathrm{d}\mu_0(z_i) \equiv \prod_{i=1}^{n} \exp\{\theta_i z_i - b(\theta_i)\}\, \mathrm{d}\mu(z_i), \tag{2}$$

where $\mathbf{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^{\mathrm{T}} = \mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\beta} \in \mathbb{R}^d$, $b(\theta)$ is a smooth convex function, μ_0 is the Lebesgue measure on Euclidean spaces and μ is some fixed measure on $\mathbb{R}$. Clearly $\{f_0(z, \theta) : \theta \in \mathbb{R}\}$ is a family of distributions in the regular exponential family and may not contain $g_{n,i}$s. Denote by $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1), \ldots, b(\theta_n))^{\mathrm{T}}$ and $\boldsymbol{\mu}(\boldsymbol{\theta}) = (b'(\theta_1), \ldots, b'(\theta_n))^{\mathrm{T}}$, and define a matrix-valued function $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{diag}\{b''(\theta_1), \ldots, b''(\theta_n)\}$. It is well known that, for any n-dimensional random vector $\mathbf{Z}$ with distribution $F_n(\cdot, \boldsymbol{\beta})$ given by equation (2), we have

$$\begin{aligned} E(\mathbf{Z}) &= \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}), \\ \mathrm{cov}(\mathbf{Z}) &= \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{3}$$

Theorem 5 in Section 4.2 shows that, under some regularity conditions, there is a unique distribution $F_n(\cdot, \boldsymbol{\beta}_{n,0})$ in the family of misspecified GLMs in equation (2) that has the smallest KL divergence from the true model G_n, where the KL divergence (Kullback and Leibler, 1951) of the posited GLM $F_n(\cdot, \boldsymbol{\beta})$ in equation (2) from the true model G_n is given by

$$I\{g_n; f_n(\cdot, \boldsymbol{\beta})\} = \int \log\{g_n(\mathbf{z})\}\, g_n(\mathbf{z})\, \mathrm{d}\mathbf{z} - \int \log\{f_n(\mathbf{z}, \boldsymbol{\beta})\}\, g_n(\mathbf{z})\, \mathrm{d}\mathbf{z} = \sum_{i=1}^{n} I\{g_{n,i}; f_0(\cdot, \theta_i)\}, \tag{4}$$

where

$$I\{g_{n,i}; f_0(\cdot, \theta_i)\} = \int \log\{g_{n,i}(z)\}\, g_{n,i}(z)\, \mathrm{d}z - \int \log\{f_0(z, \theta_i)\}\, g_{n,i}(z)\, \mathrm{d}z.$$

Thus, the parameter $\boldsymbol{\beta}$ is identifiable in such misspecified models. This uniqueness entails that $F_n(\cdot, \boldsymbol{\beta}_{n,0}) = G_n$ when the model is correctly specified, i.e. $G_n \in \{F_n(\cdot, \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^d\}$. The QMLE of the d-dimensional parameter vector $\boldsymbol{\beta}_{n,0}$ is defined as

$$\hat{\boldsymbol{\beta}}_n = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^d} l_n(\mathbf{y}, \boldsymbol{\beta}), \tag{5}$$

where

$$l_n(\mathbf{y}, \boldsymbol{\beta}) = \log\{f_n(\mathbf{y}, \boldsymbol{\beta})\} = \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - \mathbf{1}^{\mathrm{T}}\mathbf{b}(\mathbf{X}\boldsymbol{\beta}) + \sum_{i=1}^{n} \log\left\{ \frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(y_i) \right\} \tag{6}$$

is the quasi-log-likelihood function. Clearly $\hat{\boldsymbol{\beta}}_n$ is the solution to the score equation

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}) \equiv \frac{\partial l_n(\mathbf{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\{\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})\} = \mathbf{0}, \tag{7}$$

which becomes the normal equation $\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$ in the linear regression model.

Under some regularity conditions, theorem 6 in Section 4.2 shows that the QMLE $\hat{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\beta}_{n,0}$, and theorem 7 further establishes its asymptotic normality. The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_n$ is $\mathbf{A}_n^{-1}\mathbf{B}_n\mathbf{A}_n^{-1}$, where

$$\mathbf{B}_n \equiv \mathrm{cov}\{\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})\} = \mathrm{cov}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}) = \mathbf{X}^{\mathrm{T}}\,\mathrm{cov}(\mathbf{Y})\mathbf{X}, \tag{8}$$

in view of equation (31) in Section 4.2 with $\mathrm{cov}(\mathbf{Y}) = \mathrm{diag}\{\mathrm{var}(Y_1), \ldots, \mathrm{var}(Y_n)\}$ by the independence assumption, and $\mathbf{A}_n = \mathbf{A}_n(\boldsymbol{\beta}_{n,0})$ with

$$\mathbf{A}_n(\boldsymbol{\beta}) \equiv \frac{\partial^2 I\{g_n; f_n(\cdot, \boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}^2} = -\frac{\partial^2 l_n(\mathbf{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})\mathbf{X} \tag{9}$$

under the smoothness condition. Throughout the paper, the expectation and covariance are taken with respect to the true distribution G_n unless specified otherwise.

In view of expression (3), $\mathbf{A}_n(\boldsymbol{\beta})$ defined in expression (9) is exactly the covariance matrix of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ when $\mathbf{Y}$ has distribution $F_n(\cdot, \boldsymbol{\beta})$, and thus $\mathbf{A}_n$ is the covariance matrix of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ under the best misspecified GLM $F_n(\cdot, \boldsymbol{\beta}_{n,0})$, whereas $\mathbf{B}_n$ defined in expression (8) is the covariance matrix of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ under the true model G_n. It is well known in classical maximum likelihood theory that the identity $\mathbf{A}_n = \mathbf{B}_n$ holds when the model is correctly specified, with $\mathbf{A}_n$ and $\mathbf{B}_n$ being the Hessian and outer product forms of the Fisher information matrix respectively. These two matrices play a pivotal role in quasi-maximum-likelihood estimation of misspecified models.

2.2. Bayesian and Kullback–Leibler divergence principles

Suppose that we have a set of competing models: $\{\mathfrak{M}_m : m = 1, \ldots, M\}$. A typical Bayesian model selection procedure is first to give non-zero prior probability $\alpha_{\mathfrak{M}_m}$ on each model $\mathfrak{M}_m$, and then to prescribe a prior distribution $\mu_{\mathfrak{M}_m}$ for the parameter vector in the corresponding model. Assume that the density function of $\mu_{\mathfrak{M}_m}$ is bounded in $\mathbb{R}^{\mathfrak{M}_m} = \mathbb{R}^{d_m}$ and locally bounded away from zero throughout the domain. The Bayesian principle of model selection is to choose the most probable model *a posteriori*, i.e. to choose model $\mathfrak{M}_{m_0}$ such that $m_0 = \arg\max_{m \in \{1,\ldots,M\}} S(\mathbf{y}, \mathfrak{M}_m; F_n)$, where

$$S(\mathbf{y}, \mathfrak{M}_m; F_n) = \log\left[\int \alpha_{\mathfrak{M}_m} \exp\{l_n(\mathbf{y}, \boldsymbol{\beta})\}\, \mathrm{d}\mu_{\mathfrak{M}_m}(\boldsymbol{\beta}) \right] \tag{10}$$

with the log-likelihood $l_n(\mathbf{y}, \boldsymbol{\beta})$ as in equation (6) and the integral over $\mathbb{R}^{d_m}$. Fix an arbitrary model $\mathfrak{M} = \mathfrak{M}_m$ and let $\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_{n,m}$ be the QMLE. Using the Taylor series expansion, Schwarz (1978) showed that $S(\mathbf{y}, \mathfrak{M}; F_n)$ can be asymptotically expanded as

$$l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \frac{\log(n)}{2}|\mathfrak{M}| + \log(\alpha_{\mathfrak{M}})$$

with $\hat{\boldsymbol{\beta}}_n$ the maximum likelihood estimator of $\boldsymbol{\beta}$, for the case of IID observations with correctly specified regular exponential family of models. This expression allowed Schwarz to introduce the seminal BIC for model selection:

$$\mathrm{BIC}(\mathbf{y}, \mathfrak{M}) = -2 l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + \log(n)|\mathfrak{M}|, \tag{11}$$

where a factor of 2 is included to make it consistent with the usual definition of AIC that is given in equation (15) below. Schwarz's original arguments were generalized later by Cavanaugh and Neath (1999).

For each $m \in \{1, \ldots, M\}$, the marginal distribution ν_m of the response vector $\mathbf{Y}$ conditional on model $\mathfrak{M}_m$ has density function

$$\frac{\mathrm{d}\nu_m}{\mathrm{d}\mu_0}(\mathbf{z}) = \int \exp\{l_n(\mathbf{z}, \boldsymbol{\beta})\}\, \mathrm{d}\mu_{\mathfrak{M}_m}(\boldsymbol{\beta}), \qquad \mathbf{z} \in \mathbb{R}^n. \tag{12}$$

The KL divergence of ν_m from the true model G_n is given by

$$I\left(g_n; \frac{\mathrm{d}\nu_m}{\mathrm{d}\mu_0}\right) = E[\log\{g_n(\mathbf{Y})\}] - E\left(\log\left[\int \exp\{l_n(\mathbf{Y}, \boldsymbol{\beta})\}\, \mathrm{d}\mu_{\mathfrak{M}_m}(\boldsymbol{\beta})\right]\right), \tag{13}$$

which leads to the following proposition.

Proposition 1.

(a) For each $m \in \{1, \ldots, M\}$, we have

$$E\{S(\mathbf{Y}, \mathfrak{M}_m; F_n)\} = -I\left(g_n; \frac{\mathrm{d}\nu_m}{\mathrm{d}\mu_0}\right) + \log(\alpha_{\mathfrak{M}_m}) + E[\log\{g_n(\mathbf{Y})\}], \tag{14}$$

where the expectation is taken with respect to the true distribution G_n.

(b) Assume $\alpha_{\mathfrak{M}_1} = \ldots = \alpha_{\mathfrak{M}_M}$. Then we have

$$\arg \max_{m \in \{1,\ldots,M\}} E\{S(\mathbf{Y}, \mathfrak{M}_m; F_n)\} = \arg \min_{m \in \{1,\ldots,M\}} I\left(g_n; \frac{\mathrm{d}\nu_m}{\mathrm{d}\mu_0}\right).$$

Proposition 1 holds true regardless of whether or not the true model is in the set of candidate models. We see that, for each $m \in \{1, \ldots, M\}$, $-S(\mathbf{y}, \mathfrak{M}_m; F_n) + \log(\alpha_{\mathfrak{M}_m})$ gives, up to a common additive constant, an unbiased estimate of $I(g_n; \mathrm{d}\nu_m/\mathrm{d}\mu_0)$. This shows that the Bayesian principle of model selection can be restated as choosing the model that minimizes the KL divergence of the marginal distribution of the response vector $\mathbf{Y}$ from its true distribution, provided that we assign equal prior probabilities on the M competing models.

Given a set of candidate models $\{\mathfrak{M}_m : m = 1, \ldots, M\}$ of form (2), we can also compare them by using the KL divergence principle of model selection that was introduced by Akaike (1973, 1974), who studied the case of IID observations with correctly specified models. With the sequence of QMLEs $\{\hat{\boldsymbol{\beta}}_{n,m} : m = 1, \ldots, M\}$ based on these models, Akaike's principle of model selection is to choose the model $\mathfrak{M}_{m_0}$ that minimizes the KL divergence $I\{g_n; f_n(\cdot, \hat{\boldsymbol{\beta}}_{n,m})\}$ of the fitted model $F_n(\cdot, \hat{\boldsymbol{\beta}}_{n,m})$ from the true model G_n, i.e. $m_0 = \arg\min_{m \in \{1,\ldots,M\}} I\{g_n; f_n(\cdot, \hat{\boldsymbol{\beta}}_{n,m})\}$. This principle is equivalent to choosing the model $\mathfrak{M}_{m_0}$ that maximizes the expected log-likelihood with the expectation taken with respect to an independent copy of $\mathbf{Y}$. For notational simplicity, we drop the subscript m hereafter when there is no confusion. Using the asymptotic theory of maximum likelihood estimation, Akaike (1973) showed for the case of IID observations that the expected log-likelihood can be asymptotically expanded as $l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - |\mathfrak{M}|$, which leads to the seminal AIC for comparing models:

$$\mathrm{AIC}(\mathbf{y}, \mathfrak{M}) = -2 l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + 2|\mathfrak{M}|. \tag{15}$$

3. Asymptotic expansions of model selection principles in misspecified models

We here derive a novel asymptotic expansion of the Bayesian principle in misspecified GLMs. Our derivation indicates that, in contrast with traditional AIC and BIC, directly incorporating the effect of model misspecification in a model selection criterion can be beneficial. The KL divergence principle has been considered by many researchers in the setting of model misspecification (see Section 3.3); the derivations were, however, mostly for the case of IID observations and often by heuristic arguments. We also give a rigorous asymptotic expansion of this principle in misspecified GLMs with deterministic design matrices.

3.1. Generalized Bayesian information criterion in misspecified models

Let us consider the log-marginal-likelihood $S(\mathbf{y}, \mathfrak{M}; F_n)$, as defined in equation (10), in misspecified models. In GLM (2), the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^n$ for the response vector $\mathbf{Y}$ is given by $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$. Under this working model and some regularity conditions, as in Schwarz (1978), we can apply the classical Laplace approximation (i.e. Taylor series expansion of the exponent to second order and using Gaussian integration to obtain necessary terms) to show that, with probability tending to 1,

$$\begin{aligned} S(\mathbf{y}, \mathfrak{M}; F_n) &= l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + \log\left[\int \exp\left\{ -\frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)^{\mathrm{T}} \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)}{2} \right\} \mathrm{d}\boldsymbol{\beta} \right] + C'_n \\ &= l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \frac{\log(n)}{2} |\mathfrak{M}| - \frac{1}{2} \log |n^{-1} \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n)| + C_n, \end{aligned} \tag{16}$$

where $\hat{\boldsymbol{\beta}}_n$ is the maximizer of $l_n(\mathbf{y}, \boldsymbol{\beta})$, $\mathbf{A}_n(\boldsymbol{\beta}) = -\partial^2 l_n(\mathbf{y}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^2 = \mathbf{X}^{\mathrm{T}} \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta}) \mathbf{X}$ and C'_n and C_n are both bounded quantities. When the model is misspecified, it is, however, no longer clear whether the C_n-term is still bounded. Moreover, the second-order term $-\frac{1}{2} \log |n^{-1} \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n)|$ in equation (16) depends on the scale of the design matrix $\mathbf{X}$ and thus is not very meaningful. We shall derive an asymptotic expansion of the Bayes factor in misspecified models by using the asymptotic theory of the QMLE that is presented in Section 4.2.

Let $\mu_{\mathfrak{M}}$ be the prior distribution on $\boldsymbol{\beta} \in \mathbb{R}^d$ for model $\mathfrak{M}$ and

$$\pi(\boldsymbol{\beta}) = \frac{\mathrm{d}\mu_{\mathfrak{M}}}{\mathrm{d}\mu_0}(\boldsymbol{\beta})$$

its prior density relative to the Lebesgue measure μ_0 on $\mathbb{R}^d$. In high dimensional Bayesian modelling, we need to put some basic constraint on the prior distribution to prevent it from assigning a diminishingly low probability near the theoretically best model under KL divergence. Let δ_n be a sequence of diverging positive numbers. By the proof of consistency of the QMLE $\hat{\boldsymbol{\beta}}_n$ in theorem 6, we have the concentration probability bound

$$P\{\hat{\boldsymbol{\beta}}_n \in N_n(\delta_n)\} \geqslant 1 - 4d/c^2 \delta_n^2,$$

where $N_n(\delta_n) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\mathbf{B}_n^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n,0})\| \leqslant \delta_n\}$ is an asymptotically shrinking neighbourhood of the parameter vector $\boldsymbol{\beta}_{n,0}$ of the best GLM $F_n(\cdot, \boldsymbol{\beta}_{n,0})$ and c is some positive constant that is specified in condition 2 in Section 4.1. We denote by

$$\tilde{N}_n(\delta_n) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\mathbf{B}_n^{1/2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)\| \leqslant \delta_n\}$$

a neighbourhood of the QMLE $\hat{\boldsymbol{\beta}}_n$ and $\tilde{N}_n^{\mathrm{c}}(\delta_n)$ the complement of $\tilde{N}_n(\delta_n)$. Clearly $\tilde{N}_n(\delta_n) \subset N_n(2\delta_n)$ conditional on the above event $\{\hat{\boldsymbol{\beta}}_n \in N_n(\delta_n)\}$. The key idea of deriving the asymptotic expansion of the log-marginal-likelihood $S(\mathbf{y}, \mathfrak{M}; F_n)$ is to expand the quasi-log-likelihood

function $l_n(\mathbf{y}, \boldsymbol{\beta})$ at the QMLE $\hat{\boldsymbol{\beta}}_n$ and to show that the integral with the prior distribution on the neighbourhood $\tilde{N}_n(\delta_n)$ dominates that on $\tilde{N}_n^c(\delta_n)$. Note that the neighbourhood $N_n(2\delta_n)$ of $\boldsymbol{\beta}_{n,0}$ shrinks asymptotically with the sample size n, meaning that the constraint on the prior becomes weaker asymptotically.

For any $\boldsymbol{\beta} \in \mathbb{R}^d$, we define a quantity

$$l_n^*(\mathbf{y}, \boldsymbol{\beta}) = l_n(\mathbf{y}, \boldsymbol{\beta}) - l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n), \tag{17}$$

which is the deviation of the quasi-log-likelihood from its maximum. It follows from equations (10) and (17) that

$$S(\mathbf{y}, \mathfrak{M}; F_n) = l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + \log[E_{\mu_{\mathfrak{M}}}\{U_n(\boldsymbol{\beta})^n\}] + \log(\alpha_{\mathfrak{M}}), \tag{18}$$

where $U_n(\boldsymbol{\beta}) = \exp\{n^{-1} l_n^*(\mathbf{y}, \boldsymbol{\beta})\}$. We present the asymptotic expansion of the term $\log[E_{\mu_{\mathfrak{M}}}\{U_n(\boldsymbol{\beta})^n\}]$ in the following theorem.

Theorem 1. Under conditions 1, 2 and 5 in Section 4.1, we have with probability tending to 1

$$S(\mathbf{y}, \mathfrak{M}; F_n) = l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \frac{\log(n)}{2}|\mathfrak{M}| + \frac{1}{2}\log|\mathbf{H}_n| + \log(\alpha_{\mathfrak{M}}) + \frac{\log(2\pi)}{2}|\mathfrak{M}| + \log(c_n) + o(1), \tag{19}$$

where $\mathbf{H}_n = \mathbf{A}_n^{-1}\mathbf{B}_n$ and $c_n \in [c_1, c_2]$.

Following the asymptotic expansion in theorem 1, we introduce GBIC as follows.

Definition 1. We define GBIC of the competing model $\mathfrak{M}$ by

$$\mathrm{GBIC}(\mathbf{y}, \mathfrak{M}; F_n) = -2 l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + \log(n)|\mathfrak{M}| - \log|\hat{\mathbf{H}}_n|, \tag{20}$$

where $\hat{\mathbf{H}}_n$ is an estimate of the covariance contrast matrix $\mathbf{H}_n$ given in Section 3.4.

Compared with BIC, GBIC takes into account model misspecification explicitly. The second term in GBIC is the same as that in BIC and penalizes the model complexity. The third term $-\log|\hat{\mathbf{H}}_n|$ in GBIC is, however, not always non-negative and thus is not necessarily a penalty term on model misspecification. When the model is correctly specified, we would expect $\hat{\mathbf{H}}_n \approx I_d$ since $\mathbf{H}_n = I_d$ because $\mathbf{A}_n = \mathbf{B}_n$ and thus $\log|\hat{\mathbf{H}}_n| \approx \log|I_d| = 0$. In this sense, GBIC introduced above generalizes BIC. We shall see in Section 5 that the second-order term in GBIC can lead to improved performance in small samples even in correctly specified models.

3.2. $GBIC_p$ in misspecified models

Assigning a reasonable prior probability to each candidate model is an art. One usually chooses to give a higher prior probability to a model with lower complexity. In this section, we consider enlightening this prior choice with the KL divergence principle. For each candidate model $\mathfrak{M}_m$ with $1 \leqslant m \leqslant M$, denote by

$$D_m = E[I\{g_n; f_n(\cdot, \hat{\boldsymbol{\beta}}_{n,m})\} - I\{g_n; f_n(\cdot, \boldsymbol{\beta}_{n,m,0})\}] \tag{21}$$

the expected excess KL divergence of the best fitted model $F_n(\cdot, \hat{\boldsymbol{\beta}}_{n,m})$ from the true model G_n relative to that of the theoretically best model $F_n(\cdot, \boldsymbol{\beta}_{n,m,0})$ from the true model. It follows from the definition of $\boldsymbol{\beta}_{n,m,0}$ in theorem 5 in Section 4.2 that

$$I\{g_n; f_n(\cdot, \hat{\boldsymbol{\beta}}_{n,m})\} - I\{g_n; f_n(\cdot, \boldsymbol{\beta}_{n,m,0})\} \geqslant 0,$$

which entails that D_m is non-negative. It is of interest to consider the prior assignment $\alpha_{\mathfrak{M}_m} \propto \exp(-D_m)$, which prefers models with less variability. In the case when the model family is

correctly specified, this prior probability is asymptotically proportional to $\exp(-d/2)$, where d is the number of covariates in the model. By the proof of theorem 3 in Appendix A, we have $\log(\alpha_{\mathfrak{M}}) = -\frac{1}{2}\mathrm{tr}(\mathbf{H}_n) + o(1)$ up to a common additive constant, which along with theorem 1 leads to the following asymptotic expansion of the Bayes factor.

Theorem 2. Under conditions 1–6 in Section 4.1, if the prior probabilities $\alpha_{\mathfrak{M}_m} = C\exp(-D_m)$ with C some normalization constant, we have, with probability tending to 1,

$$S(\mathbf{y}, \mathfrak{M}; F_n) = l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \frac{\log(n)}{2}|\mathfrak{M}| - \frac{1}{2}\mathrm{tr}(\mathbf{H}_n) + \frac{1}{2}\log|\mathbf{H}_n| + \frac{\log(2\pi)}{2}|\mathfrak{M}| + \log(c_n C) + o(1), \tag{22}$$

where $\mathbf{H}_n = \mathbf{A}_n^{-1}\mathbf{B}_n$ and $c_n \in [c_1, c_2]$.

Following the asymptotic expansion in theorem 2, we introduce the generalized BIC with prior probability, GBIC_p, as follows.

Definition 2. We define GBIC_p of the competing model M by

$$\mathrm{GBIC}_p(\mathbf{y}, \mathfrak{M}; F_n) = -2l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + \log(n)|\mathfrak{M}| + \mathrm{tr}(\hat{\mathbf{H}}_n) - \log|\hat{\mathbf{H}}_n|, \tag{23}$$

where $\hat{\mathbf{H}}_n$ is an estimate of the covariance contrast matrix $\mathbf{H}_n$ that is given in Section 3.4.

Let us consider the case of correctly specified models. As discussed in Section 3.4, we can construct a consistent estimate $\hat{\mathbf{H}}_n$ of the covariance contrast matrix $\mathbf{H}_n = I_d$, which entails that the second-order term $\mathrm{tr}(\hat{\mathbf{H}}_n) - \log|\hat{\mathbf{H}}_n|$ is asymptotically close to $|\mathfrak{M}|$. This term is asymptotically dominated by the penalty term $\log(n)|\mathfrak{M}|$ on model complexity. In view of equation (23), this observation suggests that, under the uniform asymptotic dominance of the second-order term, the model selection criterion GBIC_p is asymptotically equivalent to BIC in correctly specified models, leading to model selection consistency; a similar result also holds for GBIC. GBIC_p also enjoys a representation shown in the following proposition.

Proposition 2. When $\hat{\mathbf{H}}_n$ in equation (23) is chosen to be $\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n$ for some $d \times d$ symmetric positive definite matrices $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$, we have

$$\mathrm{GBIC}_p(\mathbf{y}, \mathfrak{M}; F_n) = -2l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + 1 + \log(n)|\mathfrak{M}| + 2I\{N(\mathbf{0}, \hat{\mathbf{B}}_n); N(\mathbf{0}, \hat{\mathbf{A}}_n)\}, \tag{24}$$

where $I\{N(\mathbf{0}, \hat{\mathbf{B}}_n); N(\mathbf{0}, \hat{\mathbf{A}}_n)\} = \frac{1}{2}\{\mathrm{tr}(\hat{\mathbf{H}}_n) - \log|\hat{\mathbf{H}}_n| - d\}$ is the KL divergence of the d-variate Gaussian distribution $N(\mathbf{0}, \hat{\mathbf{A}}_n)$ from the d-variate Gaussian distribution $N(\mathbf{0}, \hat{\mathbf{B}}_n)$.

This representation shows that GBIC_p in misspecified models admits a natural decomposition of the form

$$\text{goodness of fit} + \text{model complexity} + \text{model misspecification}, \tag{25}$$

where the first term is the negative maximum quasi-log-likelihood and the two penalty terms on model complexity and model misspecification are both non-negative. However, neither the plain GBIC nor GAIC, which will be introduced in Section 3.3, has such a decomposition. The model misspecification term in GBIC_p has a natural interpretation. Under the true model G_n, we have $E(\mathbf{X}^{\mathrm{T}}\mathbf{Y}) = \mathbf{X}^{\mathrm{T}}E(\mathbf{Y})$ and $\mathrm{cov}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}) = \mathbf{X}^{\mathrm{T}}\mathrm{cov}(\mathbf{Y})\mathbf{X} = \mathbf{B}_n$. Under some regularity conditions, an application of the central limit theorem shows that the asymptotic distribution of $\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\mu})$ with $\boldsymbol{\mu} = E_{G_n}(\mathbf{Y})$ is close to $N(\mathbf{0}, \mathbf{B}_n)$. However, if we misspecify the true model as the GLM $F_n(\cdot, \boldsymbol{\beta})$, the QMLE $\hat{\boldsymbol{\beta}}_n$ asymptotically converges to $\boldsymbol{\beta}_{n,0}$ such that $F_n = F_n(\cdot, \boldsymbol{\beta}_{n,0})$ is the best model in the posited GLM family to approximate the true model G_n under the KL divergence. It follows from equation (31) in Section 4.2 and equation (3) that $E_{F_n}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}) = \mathbf{X}^{\mathrm{T}}E_{F_n}(\mathbf{Y}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\mu}$ and $\mathrm{cov}_{F_n}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta}_{n,0})\mathbf{X} = \mathbf{A}_n$. Under some regularity conditions, the asymptotic

distribution of $\mathbf{X}^{\mathrm{T}}(\mathbf{Y}-\boldsymbol{\mu})$ is now close to $N(\mathbf{0},\mathbf{A}_n)$. Thus the KL divergence of $N(\mathbf{0},\mathbf{A}_n)$ from $N(\mathbf{0},\mathbf{B}_n)$ provides a natural measure of model misspecification. When the model is correctly specified, we would expect $\hat{\mathbf{A}}_n \approx \hat{\mathbf{B}}_n$ since $\mathbf{A}_n = \mathbf{B}_n$, and thus $I\{N(\mathbf{0},\hat{\mathbf{B}}_n); N(\mathbf{0},\hat{\mathbf{A}}_n)\} \approx 0$. We shall show in Section 5 that the second-order term in GBIC_p can also lead to improved performance in small samples even in correctly specified models.

3.3. Generalized Akaike information criterion in misspecified models

We now consider the KL divergence principle in misspecified models. AIC has been studied by many researchers, mainly for the case of IID observations. Takeuchi (1976) generalized the derivation of AIC and obtained a term $\mathrm{tr}(\mathbf{A}^{-1}\mathbf{B})$ in place of the dimension of the model, where $\mathbf{B}$ and $\mathbf{A}$ are two matrices that involve the first and second derivatives of the log-likelihood function respectively. This term was also obtained by Stone (1977) in deriving the asymptotic expansion for cross-validation. In fact, $\mathrm{tr}(\mathbf{A}^{-1}\mathbf{B})$ is the well-known Lagrange multiplier test statistic. See also Hosking (1980), Shibata (1989), Konishi and Kitagawa (1996), Burnham and Anderson (1998) and Bozdogan (2000).

In view of equation (4), the KL divergence of the fitted model $F_n(\cdot,\hat{\boldsymbol{\beta}}_{n,m})$ from the true model G_n can be written as

$$I\{g_n; f_n(\cdot,\hat{\boldsymbol{\beta}}_{n,m})\} = E[\log\{g_n(\tilde{\mathbf{Y}})\}] - \eta_n(\hat{\boldsymbol{\beta}}_{n,m}), \tag{26}$$

where $\eta_n(\boldsymbol{\beta}) = E\{l_n(\tilde{\mathbf{Y}},\boldsymbol{\beta})\}$ and $\tilde{\mathbf{Y}}$ is an independent copy of $\mathbf{Y}$. Thus the best model chosen by Akaike's principle of model selection is $m_0 = \arg\max_{m\in\{1,\ldots,M\}} \eta_n(\hat{\boldsymbol{\beta}}_{n,m})$. For notational simplicity, we drop the subscript m. We also drop the last constant term in the quasi-log-likelihood $l_n(\mathbf{y},\boldsymbol{\beta})$ defined in equation (6), which does not depend on $\boldsymbol{\beta}$, and redefine it as

$$l_n(\mathbf{y},\boldsymbol{\beta}) = \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - \mathbf{1}^{\mathrm{T}}\mathbf{b}(\mathbf{X}\boldsymbol{\beta}).$$

In view of equation (26), for the purpose of model comparison, we need to estimate the quantity of the expected log-likelihood $\eta_n(\hat{\boldsymbol{\beta}}_n)$. In practice, we have only a single sample $\mathbf{y}$ available. Clearly the maximum quasi-log-likelihood $l_n(\mathbf{y},\hat{\boldsymbol{\beta}}_n)$ tends to overestimate $\eta_n(\hat{\boldsymbol{\beta}}_n)$ since we would use the same data twice, i.e. in both estimation and model evaluation. We present the asymptotic expansion of $E\{\eta_n(\hat{\boldsymbol{\beta}}_n)\}$ in the following theorem.

Theorem 3. Under conditions 1–4 and 6 in Section 4.1, we have

$$E\{\eta_n(\hat{\boldsymbol{\beta}}_n)\} = E\{l_n(\mathbf{y},\hat{\boldsymbol{\beta}}_n)\} - \mathrm{tr}(\mathbf{H}_n) + o(1), \tag{27}$$

where $\mathbf{H}_n = \mathbf{A}_n^{-1}\mathbf{B}_n$ and both expectations are taken with respect to the true distribution G_n.

As mentioned before, a correction term of form $\mathrm{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n)$ is well known in the literature but has been derived usually for the case of IID observations and often by heuristic arguments. Following the asymptotic expansion in theorem 3, we have GAIC as follows.

Definition 3. GAIC for the competing model $\mathfrak{M}$ is

$$\mathrm{GAIC}(\mathbf{y},\mathfrak{M};F_n) = -2l_n(\mathbf{y},\hat{\boldsymbol{\beta}}_n) + 2\,\mathrm{tr}(\hat{\mathbf{H}}_n), \tag{28}$$

where $\hat{\mathbf{H}}_n$ is an estimate of the covariance contrast matrix $\mathbf{H}_n$ given in Section 3.4.

GAIC shares some common features with GBIC. It incorporates the effects of model complexity and model misspecification in a single term. When $\hat{\mathbf{H}}_n$ in equation (28) is chosen to be $\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n$ for some $d\times d$ symmetric positive definite matrices $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$, we have $\mathrm{tr}(\hat{\mathbf{H}}_n) = \mathrm{tr}(\hat{\mathbf{B}}_n^{1/2}\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n^{1/2}) \geqslant 0$, showing that the term added to the negative maximum quasi-log-likelihood

in GAIC is indeed a penalty term. When the model is correctly specified, it holds that $\mathrm{tr}(\mathbf{H}_n) = \mathrm{tr}(I_d) = d = |\mathfrak{M}|$ since $\mathbf{H}_n = I_d$ because $\mathbf{A}_n = \mathbf{B}_n$, which is the number of predictors in GLM (2). Thus we would expect $\mathrm{tr}(\hat{\mathbf{H}}_n) \approx \mathrm{tr}(I_d) = |\mathfrak{M}|$, which reflects the penalty on model complexity as in AIC. As mentioned before, many extensions of AIC under different settings share the same form as equation (28). We shall also see in Section 5 that the second-order term in GAIC can lead to improved performance in small samples even in correctly specified models. We point out that equation (4.5) in Stone's derivation (Stone, 1977) suggests that, under the same working models, GAIC and cross-validation are asymptotically equivalent under some weak conditions, where cross-validation uses the log-density assessment.

3.4. Estimation of covariance contrast matrix $\mathbf{H}_n$

The model selection criteria GAIC, GBIC and GBIC$_p$ require an estimate $\hat{\mathbf{H}}_n$ of the covariance contrast matrix $\mathbf{H}_n$, which is a contrast between the covariance structures in the misspecified model and in the true model. Specifically, we need to estimate the trace $\mathrm{tr}(\mathbf{H}_n)$ for GAIC, the log-determinant $\log|\mathbf{H}_n|$ for GBIC and both quantities for GBIC$_p$. Estimating these quantities in misspecified models is non-trivial. In this paper, we consider two specific estimators for them and illustrate the performance of the resulting model selection criteria with extensive numerical studies. More accurate estimation of $\mathrm{tr}(\mathbf{H}_n)$ and $\log|\mathbf{H}_n|$, especially in high dimensions, demands rigorous theoretical studies, which we shall not address in the paper. Through our numerical studies, we found that estimating the trace of the covariance contrast matrix $\mathbf{H}_n$ and estimating its log-determinant exhibit different behaviour. To show the utility of the new model selection criteria, we estimate the two quantities separately with two different estimators $\hat{\mathbf{H}}_n$: a simple estimator for $\log|\mathbf{H}_n|$ and a bootstrap estimator for $\mathrm{tr}(\mathbf{H}_n)$.

To estimate $\log|\mathbf{H}_n|$, we consider the estimate $\hat{\mathbf{H}}_n$ of the plug-in form $\hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n$ with $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ some $d \times d$ symmetric positive definite matrices. Since the QMLE $\hat{\boldsymbol{\beta}}_n$ provides a consistent estimator of $\boldsymbol{\beta}_{n,0}$ in the best misspecified GLM $F_n(\cdot, \boldsymbol{\beta}_{n,0})$, a natural estimate of matrix $\mathbf{A}_n$ is given by

$$\hat{\mathbf{A}}_n = \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\mathbf{X}. \tag{29}$$

When the model is correctly specified or more generally $E(\mathbf{Y}) = \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})$, the simple estimator

$$\hat{\mathbf{B}}_n = \mathbf{X}^{\mathrm{T}}\,\mathrm{diag}[\{\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\} \circ \{\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\}]\mathbf{X}, \tag{30}$$

with '$\circ$' representing the Hadamard (componentwise) product, provides an asymptotically unbiased estimator of $\mathbf{B}_n$. In view of expression (8), we adopt the form of the diagonal matrix in equation (30) for estimating $\log|\mathbf{H}_n|$. This form seems effective in ensuring the positive definiteness and well-conditionedness of the matrix, which are crucial for effectively estimating the log-determinant. We refer to the resulting estimator $\hat{\mathbf{H}}_n$ as the simple estimator. The formulae for matrices $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ in equations (29) and (30) are provided in Appendix B, when the working model F_n is chosen to be one of the three commonly used GLMs: the linear regression model, logistic regression model and Poisson regression model. The following theorem justifies the consistency of the above simple estimator $\hat{\mathbf{H}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n$ when the model is correctly specified.

Theorem 4. Assume that the model is correctly specified, conditions 1, 2 and 7 in Section 4.1 hold, and the smallest and largest eigenvalues of $n^{-1}\mathbf{B}_n$ are bounded away from 0 and ∞. Then $\hat{\mathbf{H}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n = I_d + o_P(1)$, where $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ are given in equations (29) and (30) respectively.

In theorem 4, the assumption that the smallest and largest eigenvalues of $n^{-1}\mathbf{B}_n$ are bounded away from 0 and ∞ can be relaxed to converge to 0 or to diverge to ∞, which will affect the consistency rate. Although it is positive definite, the simple estimator $\hat{\mathbf{B}}_n$ in equation (30) may not be most suitable for the trace estimation. The general bias in such an estimator can affect more the performance of GAIC than that of GBIC and GBIC$_p$, because of the presence of the penalty term $\log(n)d$ on model complexity in the latter two information criteria. For example, $\mathrm{tr}\{(1+c)\mathbf{H}_n\} - \mathrm{tr}(\mathbf{H}_n) = c\,\mathrm{tr}(\mathbf{H}_n)$ is of the same order as $\mathrm{tr}(\mathbf{H}_n)$ for some non-zero constant c close to 0, whereas $\log|(1+c)\mathbf{H}_n| - \log|\mathbf{H}_n| = \log(1+c)d$ can be asymptotically dominated by the term $\log(n)d$. To mitigate the bias in estimating $\mathbf{B}_n$, it is sensible to construct the residual vector with a large model. This approach generally gives rise to a less biased estimator for $\mathbf{B}_n$. Motivated by these considerations, to estimate $\mathrm{tr}(\mathbf{H}_n)$ we introduce the estimate $\hat{\mathbf{H}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n$ with $\hat{\mathbf{A}}_n$ given above and $\hat{\mathbf{B}}_n$ constructed by using the following bootstrap procedure (Efron, 1979). To reduce the model bias, we first use a large model, with regularization when needed, and the idea of cross-validation to construct a proxy $\mathbf{d}$ of the deviation vector $\mathbf{Y} - E(\mathbf{Y})$. Then, we draw some bootstrap samples, each of size n, from the n observed vectors of the covariates and constructed deviations. For each bootstrap sample $(\mathbf{X}^*, \mathbf{d}^*)$, we calculate the covariance vector $(\mathbf{X}^*)^{\mathrm{T}}\mathbf{d}^*$. Finally, we obtain the bootstrap estimate $\hat{\mathbf{B}}_n$ as the covariance matrix of the covariance vectors constructed from the bootstrap samples. We refer to the resulting estimator $\hat{\mathbf{H}}_n$ as the bootstrap estimator.

The estimates for $\mathrm{tr}(\mathbf{H}_n)$ and $\log|\mathbf{H}_n|$ that were introduced above may not be the most effective. In practice, we can construct other estimates of both quantities by, for example, treating the squared residual $\{\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\} \circ \{\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\}$ as a function of the corresponding fitted value $\boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)$ and some covariates, and doing some local smoothing. Another possible way of estimating $\mathbf{H}_n$ is by using the bootstrap as suggested by theorem 7. We show in theorem 7 that, under certain regularity conditions, the QMLE $\hat{\boldsymbol{\beta}}_n$ is asymptotically normal with mean $\boldsymbol{\beta}_{n,0}$ and covariance matrix $\mathbf{C}_n^{-1}(\mathbf{C}_n^{-1})^{\mathrm{T}} = \mathbf{A}_n^{-1}\mathbf{B}_n\mathbf{A}_n^{-1}$, where $\mathbf{C}_n = \mathbf{B}_n^{-1/2}\mathbf{A}_n$. In view of $\mathbf{H}_n = (\mathbf{A}_n^{-1}\mathbf{B}_n\mathbf{A}_n^{-1})\mathbf{A}_n$, this asymptotic normality motivates us to construct the estimate $\hat{\mathbf{H}}_n$ as $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_n)\hat{\mathbf{A}}_n$ by using the bootstrap estimate $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_n)$ of the covariance matrix of the QMLE $\hat{\boldsymbol{\beta}}_n$. Regularization methods for large covariance matrix estimation can also be exploited to improve the estimation of both $\mathrm{tr}(\mathbf{H}_n)$ and $\log|\mathbf{H}_n|$.

4. Technical conditions and results

4.1. Technical conditions

Denote by $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ the smallest and largest eigenvalues of a given matrix respectively, and $\|\cdot\|$ the Euclidean norm of a vector as well as the matrix operator norm. We need the following mild regularity conditions.

Condition 1. $b(\theta)$ is twice differentiable with $b''(\theta)$ always positive and $\mathbf{X}$ has full column rank d.

Condition 2. $\lambda_{\min}(\mathbf{B}_n) \to \infty$ as $n \to \infty$ and there is some $c > 0$ such that, for any $\delta > 0$, $\min_{\boldsymbol{\beta} \in N_n(\delta)} \lambda_{\min}\{\mathbf{V}_n(\boldsymbol{\beta})\} \geqslant c$ for all sufficiently large n, where $\mathbf{V}_n(\boldsymbol{\beta}) = \mathbf{B}_n^{-1/2}\mathbf{A}_n(\boldsymbol{\beta})\mathbf{B}_n^{-1/2}$ and $N_n(\delta) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\mathbf{B}_n^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n,0})\| \leqslant \delta\}$.

Condition 3. For any $\delta > 0$, $\max_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d \in N_n(\delta)} \|\tilde{\mathbf{V}}_n(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) - \mathbf{V}_n\| = o(1)$, where $\mathbf{V}_n = \mathbf{V}_n(\boldsymbol{\beta}_{n,0}) = \mathbf{B}_n^{-1/2}\mathbf{A}_n\mathbf{B}_n^{-1/2}$ and $\tilde{\mathbf{V}}_n(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) = \mathbf{B}_n^{-1/2}\tilde{\mathbf{A}}_n(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)\mathbf{B}_n^{-1/2}$ with the jth row of $\tilde{\mathbf{A}}_n(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ the corresponding row of $\mathbf{A}_n(\boldsymbol{\beta}_j)$ for each $j = 1, \ldots, d$.

Condition 4. $\max_{i=1}^n E|Y_i - E(Y_i)|^3 = O(1)$ and $\Sigma_{i=1}^n (\mathbf{x}_i^{\mathrm{T}}\mathbf{B}_n^{-1}\mathbf{x}_i)^{3/2} = o(1)$.

Condition 5. The prior density $\pi(\boldsymbol{\beta})$ satisfies

$$\inf_{\boldsymbol{\beta}\in N_n(2\delta_n)} \pi(\boldsymbol{\beta}) \geqslant c_1 \quad \text{and} \quad \sup_{\boldsymbol{\beta}\in\mathbb{R}^d} \pi(\boldsymbol{\beta}) \leqslant c_2$$

for some positive constants c_1 and c_2 and a sequence of positive numbers δ_n satisfying $\delta_n^{-1} = o\{\log^{-1/2}(n)\}$ and $\max_{\boldsymbol{\beta}\in N_n(2\delta_n)} \max[|\lambda_{\min}\{\mathbf{V}_n(\boldsymbol{\beta}) - \mathbf{V}_n\}|, |\lambda_{\max}\{\mathbf{V}_n(\boldsymbol{\beta}) - \mathbf{V}_n\}|] = o(1)$, and $\lambda_{\max}(\mathbf{V}_n)$ is of a polynomial order n.

Condition 6. $\sup_{\boldsymbol{\delta}} \max_{j,k,l} |\partial^3 \eta_n(\boldsymbol{\beta}_{n,0} + n^{1/2}\mathbf{C}_n^{-1}\boldsymbol{\delta})/\partial\delta_j\,\partial\delta_k\,\partial\delta_l| = o(n^{3/2})$ and $E\|\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})\|^{3+\alpha} = O(1)$ for some $\alpha > 0$, where $\mathbf{C}_n = \mathbf{B}_n^{1/2}\mathbf{A}_n$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_d)^{\mathrm{T}}$.

Condition 7. There exists some $L > 0$ such that, for any $\delta > 0$, the functions $n^{-1}\,\mathbf{A}_n(\boldsymbol{\beta})$,

$$n^{-1}\mathbf{X}^{\mathrm{T}} \operatorname{diag}\{|\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})|\}\mathbf{X}$$

and

$$n^{-1}\mathbf{X}^{\mathrm{T}} \operatorname{diag}[\{\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})\}\circ\{\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})\}]\mathbf{X}$$

are Lipschitz with a Lipschitz constant L in the neighbourhood $N_n(\delta)$. In addition, $n^{-1/2}\Sigma_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\{Y_i - E(Y_i)\}^2$ has bounded second central moment and $P\{\max_{i=1}^n |Y_i - E(Y_i)| > K_n\} \to 0$ as $n \to \infty$ for some $K_n = o(n^{1/2})$.

Condition 1 contains some basic smoothness and regularity conditions to ensure the uniqueness of the QMLE $\hat{\boldsymbol{\beta}}_n$ by a standard strict concavity argument and the uniqueness of the best misspecified GLM $F_n(\cdot, \boldsymbol{\beta}_{n,0})$ under the KL divergence in theorem 5.

Condition 2 is for establishing the consistency of the QMLE $\hat{\boldsymbol{\beta}}_n$ in theorem 6. The first part of condition 2 requires that the smallest eigenvalue of $\mathbf{X}^{\mathrm{T}}\operatorname{cov}(\mathbf{Y})\mathbf{X}$ diverges as n increases. When $\operatorname{var}(Y_i)$s are bounded away from 0 and ∞, this assumption amounts to the usual assumption on the design matrix $\mathbf{X}$ in linear regression for consistency, i.e. $\lambda_{\min}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \to \infty$ as $n \to \infty$. Intuitively, the second part of condition 2 means that the covariance structures that are given by the misspecified GLMs in a shrinking neighbourhood of the best working model $F_n(\cdot, \boldsymbol{\beta}_{n,0})$ cannot be too far from the covariance structure under the true model G_n. This requirement is analogous to that in importance sampling, i.e. the support of the sampling distribution must cover that of the target distribution. When the model is correctly specified, we have $\mathbf{V}_n = \mathbf{V}_n(\boldsymbol{\beta}_{n,0}) = I_d$ since $\mathbf{A}_n = \mathbf{B}_n$ by the equivalence of the Hessian and outer product forms for the Fisher information matrix.

Conditions 3 and 4 facilitate the derivation of the asymptotic normality of the QMLE $\hat{\boldsymbol{\beta}}_n$ in theorem 7, where the former is on the continuity of the matrix-valued function $\tilde{\mathbf{V}}_n(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ in a shrinking neighbourhood of $\boldsymbol{\beta}_{n,0}$ and the latter is a typical moment condition for proving asymptotic normality by using Lyapunov's theorem.

The first part of condition 5 is a mild condition on the prior distribution on $\boldsymbol{\beta}$ requiring that its prior density is bounded and locally bounded away from zero in a shrinking neighbourhood of $\boldsymbol{\beta}_{n,0}$ and facilitates the derivation of the asymptotic expansion of the Bayes factor in misspecified models in theorems 1 and 2. See Section 3.1 for more discussion on this condition. The second part of condition 5 ensures that the integral $E_{\mu_M}\{U_n(\boldsymbol{\beta})^n \mathbf{1}_{\tilde{N}_n^{\mathrm{c}}(\delta_n)}\}$ with the prior distribution outside the neighbourhood $\tilde{N}_n(\delta_n)$ is negligible in approximating the log-marginal-likelihood $S(\mathbf{y}, \mathfrak{M}; F_n)$.

Condition 6 contains some regularity conditions for rigorously deriving the asymptotic expansion of the KL divergence principle in misspecified models in theorem 3. It is also needed for deriving the asymptotic expansion of the log-prior-probability in theorem 2.

Condition 7 collects additional regularity conditions to ensure that the simple estimator $\hat{\mathbf{H}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n$, with $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ given in equations (29) and (30) respectively, is a consistent estimator of $\mathbf{H}_n = I_d$ when the model is correctly specified. The Lipschitz conditions on the three matrix-valued functions in the asymptotically shrinking neighbourhood $N_n(\delta)$ are mild and reasonable. The moment and tail conditions on the response variable Y_i are standard.

4.2. Asymptotic properties of the quasi-maximum-likelihood estimator

For completeness, we briefly present here some asymptotic properties of the QMLE in misspecified GLMs with deterministic design matrices, which serve as the foundation for deriving asymptotic expansions of model selection principles. The consistency and asymptotic normality of maximum likelihood estimation in correctly specified GLMs were studied in Fahrmeir and Kaufmann (1985). A general theory of maximum likelihood estimation of misspecified models was presented in White (1982), who studied the case of IID observations from a general distribution and used the KL divergence as a measure of model misspecification. We generalize those results to misspecified GLMs with deterministic design matrices, where the observations may no longer be IID.

Theorem 5 (parameter identifiability). Under condition 1, the KL divergence $I\{g_n; f_n(\cdot, \boldsymbol{\beta})\}$ has a unique global minimum at $\boldsymbol{\beta}_{n,0} \in \mathbb{R}^d$, which solves the equation

$$\mathbf{X}^{\mathrm{T}}\{E(\mathbf{Y}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})\} = \mathbf{0}. \tag{31}$$

Theorem 6 (consistency). Under conditions 1 and 2, the QMLE $\hat{\boldsymbol{\beta}}_n$ satisfies $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0} = o_P(1)$.

Theorem 7 (asymptotic normality). Under conditions 1–4, the QMLE $\hat{\boldsymbol{\beta}}_n$ satisfies $\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \to^{\mathcal{D}} N(\mathbf{0}, I_d)$, where $\mathbf{C}_n = \mathbf{B}_n^{-1/2}\mathbf{A}_n$.

When the model is correctly specified, we have $\mathbf{C}_n = \mathbf{B}_n^{-1/2}\mathbf{A}_n = \mathbf{A}_n^{1/2}$ since $\mathbf{A}_n = \mathbf{B}_n$. Thus, in this case, the consistency and asymptotic normality of the QMLE $\hat{\boldsymbol{\beta}}_n$ in theorems 6 and 7 become the conventional asymptotic theory of the maximum likelihood estimator. To simplify the technical presentation, the above asymptotic normality is for fixed dimensionality d. With more delicate analysis, one can show the asymptotic normality for diverging dimensionality d, which is not the focus of the current paper. See, for example, the technical analysis in Fan and Lv (2011) for penalized maximum likelihood estimation, where the dimensionality is allowed to grow non-polynomially with the sample size.

5. Numerical examples

In this section, we illustrate the performance of model selection criteria GAIC, GBIC and GBIC$_p$ in relatively small samples, with the bootstrap estimator for $\mathrm{tr}(\mathbf{H}_n)$ and the simple estimator for $\log|\mathbf{H}_n|$ that were introduced in Section 3.4. These studies represent some first attempts for obtaining insights into the effect of incorporating model misspecification for model selection; the estimation of $\mathrm{tr}(\mathbf{H}_n)$ and $\log|\mathbf{H}_n|$ still requires rigorous theoretical investigations. Model selection with limited sample size is frequently encountered in practice and the dimension of candidate models can be large owing to the large number of covariates. One example in Section 1 is for the case of correctly specified models, and five other examples are for the case of misspecified models. In particular, in Sections 5.1.2 and 5.1.3, and 5.2.3, we consider high dimensional linear and logistic regression with interaction respectively, with Section 5.1.3 focusing on the case of moderate effects. Measures of prediction and variable selection are used to

evaluate the model selected. We include the classical AIC and BIC as benchmark model selection criteria. With reduced dimensionality given by the set of covariates in the model selected, one can perform more delicate analysis in practice.

5.1. Linear models

We consider here three linear models, starting with a correctly specified model of moderate size and proceeding to high dimensional sparse linear models with interaction terms. We observed that GBIC and GBIC$_p$ outperformed other information criteria in all settings, often by a large margin when the sample size is small.

5.1.1. Best subset linear regression

We simulated 100 data sets from the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{32}$$

where $\mathbf{y}$ is an n-dimensional response vector, $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is a p-dimensional regression coefficient vector and the error vector $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ is independent of $\mathbf{X}$. For each simulated data set, the rows of $\mathbf{X}$ were sampled as IID copies from $N(\mathbf{0}, \Sigma_0)$ with $\Sigma_0 = (0.5^{|i-j|})_{i,j=1,\ldots,p}$. To apply best subset regression, we chose $p = 6$ and set $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_0 = (1, -1.25, 0.75, 0, 0, 0)^{\mathrm{T}}$. We considered $n = 20$, 40, 80, and $\sigma = 0.25$ and $\sigma = 0.5$. For each data set, we first applied best subset regression to build all possible submodels of $\{1, \ldots, 6\}$, each of which numbers represents the corresponding covariate, and then selected the final model by using AIC, BIC, GAIC, GBIC and GBIC$_p$.

Table 1 summarizes the comparison results of the frequency of estimated model size with 3* representing the true underlying sparse model $\mathrm{supp}(\boldsymbol{\beta}_0) = \{1, 2, 3\}$ for compactness. The performance of all information criteria generally improves as the sample size increases. AIC tended to select a larger model than the true model even for large sample sizes, whereas BIC performed well when the sample size is large but also tended to select a larger model than the true model when the sample size is small. GAIC and GBIC improved over AIC and BIC respectively,

Table 1. Frequency of estimated model size for the model in Section 5.1.1 over 100 simulations for various n and σ†

σ	Criterion	*Frequencies for the following model sizes ($n = 20$):*					*Frequencies for the following model sizes ($n = 40$):*					*Frequencies for the following model sizes ($n = 80$):*				
		2	*3**	*4*	*5*	*6*	*2*	*3**	*4*	*5*	*6*	*2*	*3**	*4*	*5*	*6*
0.25	AIC	0	55	32	12	1	0	55	36	7	2	0	71	19	9	1
	BIC	0	68	24	7	1	0	81	15	4	0	0	94	5	1	0
	GAIC	0	79	19	1	1	0	65	27	7	1	0	74	19	6	1
	GBIC	0	78	18	4	0	0	83	15	2	0	0	95	5	0	0
	GBIC$_p$	0	94	5	1	0	0	90	9	1	0	0	97	3	0	0
0.5	AIC	0	51	34	10	5	0	54	33	11	2	0	59	36	3	2
	BIC	0	65	28	6	1	0	80	16	4	0	0	85	15	0	0
	GAIC	2	72	21	5	0	0	69	21	10	0	0	63	34	1	2
	GBIC	0	73	22	5	0	0	87	11	2	0	0	86	14	0	0
	GBIC$_p$	1	94	5	0	0	0	93	7	0	0	0	98	2	0	0

†3* corresponds to the true model $\{1, 2, 3\}$.

across all sample sizes tested and especially when the sample size is moderate, indicating the effectiveness of the second-order terms even in correctly specified models. When the sample size became larger, GBIC tended to perform better than GAIC. GBIC_p outperformed all other model selection criteria.

5.1.2. High dimensional sparse linear regression with interaction

To evaluate how each information criterion fares in high dimensional incorrectly specified settings, we simulated 100 data sets from the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{x}_{p+1} + \boldsymbol{\varepsilon}, \tag{33}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ design matrix, $\mathbf{x}_{p+1} = \mathbf{x}_1 \circ \mathbf{x}_2$ is an interaction term which is the product of the first two covariates and the rest are the same as in equation (32). To make variable selection easily interpretable in misspecified models, we consider the case that all covariates are independent of each other and thus set $\Sigma_0 = I_p$. We chose $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_0 = (1, -1.25, 0.75, -0.95, 1.5, 0, \ldots, 0)^{\mathrm{T}}$, $n = 80$ and $\sigma = 0.25$, and considered $p = 50$, 100, 200. Although the data were generated from model (33), we fitted the linear regression model (32) without interaction. This is an example of misspecified models. In view of model (33), the true model involves only the first five covariates in a non-linear form. Since the other covariates are independent of those five covariates, the oracle working model is $\text{supp}(\boldsymbol{\beta}_0) = \{1, \ldots, 5\}$, which indicates the linear regression model (32) with the first five covariates. Clearly it is unrealistic to implement best subset regression because of its computational complexity. Therefore, we used the smooth integration of counting and absolute integration regularization method SICA in Lv and Fan (2009) and implemented it with the iterative co-ordinate ascent algorithm ICA in Fan and Lv (2011). For each data set, we first applied the SICA method to build a sequence of sparse models and then selected the final model by using the same five model selection criteria.

To compare the selected models by the information criteria, we consider the oracle estimate based on the oracle working model $\{1, \ldots, 5\}$ as the benchmark and use both measures of prediction and variable selection. The first performance measure is the prediction error defined as $E(Y - \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2$ with $\hat{\boldsymbol{\beta}}$ an estimate and $(\mathbf{x}^{\mathrm{T}}, Y)$ an independent observation. The second and third measures are the numbers of false positive and false negative results respectively. Here, a false positive result means a selected covariate outside the oracle working model and a false negative result means a missed covariate in the oracle model. The fourth measure is the probability of recovering exactly the oracle working model, and the fifth measure is the inclusion probability of including all covariates in the oracle working model. The former characterizes the model selection consistency property, whereas the latter characterizes the sure screening property (Fan and Lv, 2008). In the calculation of the expectation in the prediction error, an independent test sample of size 10 000 was generated for approximating the expectation.

Table 2 summarizes the comparison results of the 25th, 50th and 75th percentiles of the first three performance measures as well as the selection probability and inclusion probability of the model estimated. The conclusions are similar to those in Section 5.1.1. As the dimensionality p increases, all information criteria tend to select a larger model and the prediction performance deteriorates. GAIC performed better than AIC, and GBIC and GBIC_p performed better than BIC in terms of prediction and variable selection. In particular, the model that was selected by GBIC_p mimics closely the oracle estimate. We observe that, although all methods can have a good sure screening property, the probability of model selection consistency generally decreases with dimensionality. It is interesting that GAIC tended to perform better than GBIC when the dimensionality became larger, which can be understood as an effect of decreasing effective sample size due to the noise accumulation that is associated with high dimensionality.

Table 2. Percentiles of prediction error, false positive results FP and false negative results FN as well as the selection probability SP and inclusion probability IP of the SICA estimate for the model in Section 5.1.2 over 100 simulations for various p

p	Criterion	Prediction error			FP (FN)			SP	IP
		25th	50th	75th	25th	50th	75th		
50	AIC	1.3095	1.4433	1.6337	4 (0)	6 (0)	8 (0)	0	1
	BIC	1.1822	1.2661	1.3924	0 (0)	1 (0)	4 (0)	0.35	0.99
	GAIC	1.1843	1.2697	1.3820	0 (0)	2 (0)	4 (0)	0.27	0.99
	GBIC	1.1598	1.2357	1.3401	0 (0)	1 (0)	2 (0)	0.48	0.99
	GBIC_p	1.1350	1.1873	1.2749	0 (0)	0 (0)	0 (0)	0.79	0.99
	Oracle	1.1253	1.1600	1.2201	0 (0)	0 (0)	0 (0)	1	1
100	AIC	1.4579	1.6807	1.9342	7 (0)	10 (0)	13 (0)	0	1
	BIC	1.3123	1.5229	1.7922	2.5 (0)	7 (0)	10.5 (0)	0.11	1
	GAIC	1.2038	1.2733	1.4182	0 (0)	1 (0)	2.5 (0)	0.37	1
	GBIC	1.2193	1.3731	1.6458	1 (0)	3 (0)	7 (0)	0.23	1
	GBIC_p	1.1253	1.2076	1.2745	0 (0)	0 (0)	1 (0)	0.69	1
	Oracle	1.1105	1.1639	1.2318	0 (0)	0 (0)	0 (0)	1	1
200	AIC	1.6678	1.8637	2.1852	10 (0)	14 (0)	16 (0)	0	1
	BIC	1.6598	1.8637	2.1852	9.5 (0)	13.5 (0)	16 (0)	0	1
	GAIC	1.1860	1.3045	1.4507	0 (0)	1 (0)	3 (0)	0.27	0.99
	GBIC	1.5115	1.7947	2.1321	7 (0)	11 (0)	15 (0)	0.03	1
	GBIC_p	1.1561	1.2317	1.3614	0 (0)	1 (0)	2 (0)	0.43	0.99
	Oracle	1.1201	1.1661	1.2167	0 (0)	0 (0)	0 (0)	1	1

5.1.3. *High dimensional linear regression with interaction and moderate effects*

We now consider the case of high dimensional linear regression with interaction and moderate effects. We simulated 100 data sets from model (33) with the same setting as in Section 5.1.2, except that we set $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_0 = (1, -1.25, 0.75, -0.95, 1.5, 0.5, -0.5, 0.5, -0.5, 0.5, \ldots, 0)^{\mathrm{T}}$. Compared with the model in Section 5.1.2, the five additional main linear effects are moderately strong. For each data set, we fitted the misspecified linear regression model (32) without interaction. As argued in Section 5.1.2, the oracle working model is $\text{supp}(\boldsymbol{\beta}_0) = \{1, \ldots, 10\}$. We applied the same method to build a sequence of sparse candidate models and used the same performance measures to compare the five model selection criteria as in Section 5.1.2. Table 3 summarizes the comparison results of the 25th, 50th and 75th percentiles of the first three performance measures as well as the selection probability and inclusion probability of the model estimated. The conclusions are similar to those in Section 5.1.2. We observe that, owing to the appearance of many moderate effects, all information criteria have less probability of sure screening especially when the dimensionality is large, meaning that some covariates in the oracle working model can be missed.

5.2. Non-linear models

In this section, we examine behaviours of the five information criteria in non-linear models. They include a polynomial model, a single-index model and a high dimensional logistic regression model with interaction. We again observed that the new information criteria GBIC and GBIC_p outperformed the classical AIC and BIC in all settings.

5.2.1. *Polynomial regression with heteroscedasticity*

Since each of the criteria GAIC, GBIC and GBIC_p includes a second-order term that is

Table 3. Percentiles of prediction error, false positive results FP and false negative results FN as well as the selection probability SP and inclusion probability IP of the SICA estimate for the model in Section 5.1.3 over 100 simulations for various p

p	*Criterion*	*Prediction error*			*FP (FN)*			*SP*	*IP*
		25th	*50th*	*75th*	*25th*	*50th*	*75th*		
50	AIC	1.3743	1.5164	1.7514	3 (0)	5 (0)	7 (0)	0.01	0.94
	BIC	1.2867	1.4375	1.6424	0 (0)	2 (0)	4 (0)	0.22	0.79
	GAIC	1.2427	1.3988	1.6279	0 (0)	0 (0)	1 (1)	0.36	0.61
	GBIC	1.2586	1.3859	1.5947	0 (0)	1 (0)	3 (0)	0.32	0.78
	GBIC_p	1.2359	1.3793	1.6951	0 (0)	0 (0)	1 (1)	0.38	0.54
	Oracle	1.1723	1.2481	1.3116	0 (0)	0 (0)	0 (0)	1	1
100	AIC	1.5881	1.7814	2.0827	6 (0)	9 (0)	11 (0)	0	0.85
	BIC	1.5572	1.7290	2.0359	5 (0)	8 (0)	11 (0)	0.04	0.80
	GAIC	1.2877	1.4812	1.8517	0 (0)	0 (1)	1 (2)	0.29	0.42
	GBIC	1.3751	1.6393	1.8202	1 (0)	4 (0)	8 (1)	0.15	0.74
	GBIC_p	1.2877	1.4626	1.9309	0 (0)	0 (1)	1 (2)	0.29	0.42
	Oracle	1.1837	1.2405	1.3155	0 (0)	0 (0)	0 (0)	1	1
200	AIC	1.7916	2.0573	2.4881	9.5 (0)	12 (0)	15 (0)	0	0.78
	BIC	1.7735	2.0465	2.4881	9 (0)	12 (0)	15 (0)	0	0.78
	GAIC	1.3964	1.8119	2.2214	0 (0.5)	0 (2)	1 (3)	0.21	0.25
	GBIC	1.7128	2.0465	2.4881	9 (0)	12 (0)	15 (0)	0.01	0.77
	GBIC_p	1.3683	1.6990	2.1528	0 (0)	0 (1)	1 (3)	0.22	0.27
	Oracle	1.1855	1.2367	1.3168	0 (0)	0 (0)	0 (0)	1	1

related to the variance of the response, we consider the cubic polynomial regression model with heteroscedastic variance

$$y = 1 + 5x - 2x^2 + 1.55x^3 + |x|^{1/2}\varepsilon, \tag{34}$$

where y is the response, $x \sim N(0, 1)$ is the covariate and the error $\varepsilon \sim N(0, \sigma^2)$ is independent of x. We simulated 100 data sets from this model, each of which contains n IID observations, with $n = 20, 40, 80$, and $\sigma = 0.25$ and $\sigma = 0.5$. Although the data were generated from model (34), we fitted the polynomial regression model with constant variance which is incorrectly specified. For each data set, we first fitted the polynomial regression model with constant variance of order 1 up to 6 and then selected the final model by using the same five model selection criteria. Table 4 summarizes the comparison results of the frequency of estimated order of polynomial regression model. The conclusions are similar to those in Section 5.1.1. The improved performance of GAIC, GBIC and GBIC_p over AIC and BIC indicates the usefulness of the second-order terms in capturing the difference between the estimated error variance and the apparent residual variance.

5.2.2. Non-linear regression

As another example of misspecified models, we consider a single-index model in which the response depends on a linear combination of covariates through a non-linear link function. We simulated 100 data sets from the non-linear regression model

$$\mathbf{y} = \mathbf{f}(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \tag{35}$$

where $\mathbf{y}$ is an n-dimensional response vector, $\mathbf{f}(\mathbf{z}) = (f(z_1), \ldots, f(z_n))^{\mathrm{T}}$ with $f(z) = z^3/(0.5 + z^2)$ and $\mathbf{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$, $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is a p-dimensional regression coefficient

Table 4. Frequency of estimated order for the model in Section 5.2.1 over 100 simulations for various n and σ

σ	Criterion	*Frequencies for the following model sizes ($n=20$):*					*Frequencies for the following model sizes ($n=40$):*					*Frequencies for the following model sizes ($n=80$):*				
		2	*3*	*4*	*5*	*6*	*2*	*3*	*4*	*5*	*6*	*2*	*3*	*4*	*5*	*6*
0.25	AIC	0	48	17	19	16	0	47	23	14	16	0	54	17	12	17
	BIC	0	62	13	14	11	0	63	21	9	7	0	81	10	4	5
	GAIC	5	78	5	2	4	3	88	3	4	2	0	97	2	1	0
	GBIC	0	97	3	0	0	0	92	5	1	2	0	91	5	2	2
	GBIC_p	3	94	0	0	0	1	98	1	0	0	0	99	1	0	0
0.5	AIC	0	38	16	23	23	0	40	16	20	24	0	47	13	21	19
	BIC	0	55	15	18	12	0	64	14	12	10	0	81	9	5	5
	GAIC	12	72	1	1	1	5	89	4	0	0	1	98	1	0	0
	GBIC	0	97	2	1	0	0	88	5	4	3	0	93	4	2	1
	GBIC_p	8	82	0	0	0	3	95	1	0	0	0	100	0	0	0

Table 5. Frequency of estimated model size for the model in Section 5.2.2 over 100 simulations for various n and a†

σ	Criterion	*Frequencies for the following model sizes ($n=20$):*					*Frequencies for the following model sizes ($n=40$):*					*Frequencies for the following model sizes ($n=80$):*				
		2	*3**	*4*	*5*	*6*	*2*	*3**	*4*	*5*	*6*	*2*	*3**	*4*	*5*	*6*
0.25	AIC	0	51	41	8	0	0	60	32	7	1	0	64	32	3	1
	BIC	0	71	27	2	0	0	82	16	2	0	0	90	10	0	0
	GAIC	0	81	16	3	0	0	74	22	3	1	0	75	22	2	1
	GBIC	0	79	20	1	0	0	83	15	2	0	0	91	9	0	0
	GBIC_p	0	94	6	0	0	0	91	9	0	0	0	96	4	0	0
0.5	AIC	0	52	36	11	1	0	54	33	12	1	0	51	40	9	0
	BIC	0	66	28	5	1	0	77	21	2	0	0	93	6	1	0
	GAIC	3	73	22	1	0	0	67	27	6	0	0	63	30	7	0
	GBIC	0	77	20	3	0	0	83	17	0	0	0	94	6	0	0
	GBIC_p	1	90	9	0	0	0	90	10	0	0	0	99	1	0	0

†3* corresponds to the oracle working model $\{1, 2, 3\}$.

vector and the error vector $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ is independent of $\mathbf{X}$. The rest of the setting is the same as that in Section 5.1.1 except that, as in Section 5.1.2, we consider the case that all covariates are independent of each other and thus set $\Sigma_0 = I_p$. Although the data were generated from model (35), we fitted the linear regression model (32). Therefore, the family of working models excludes the true model owing to the non-linearity. As argued in Section 5.1.2, the oracle working model is $\text{supp}(\beta_0) = \{1, 2, 3\}$, which indicates linear regression model (32) with the first three covariates. For each data set, we first applied best subset linear regression to build all possible submodels of $\{1, \ldots, 6\}$ and then selected the final model by using the same five model selection criteria. Table 5 summarizes the comparison results of the frequency of estimated model size with 3* representing the oracle working model $\{1, 2, 3\}$ for compactness. The conclusions are similar to those in Section 5.1.2.

Table 6. Percentiles of prediction error, false positive results FP and false negative results FN as well as the selection probability SP and inclusion probability IP of the SICA estimate for the model in Section 5.2.3 over 100 simulations

Criterion	*Prediction error*			*FP (FN)*			*SP*	*IP*
	25th	*50th*	*75th*	*25th*	*50th*	*75th*		
AIC	0.1411	0.1512	0.1623	15 (0)	17 (0)	19 (0)	0	0.98
BIC	0.1330	0.1468	0.1591	12.5 (0)	15 (0)	18 (0)	0	0.98
GAIC	0.1148	0.1206	0.1284	9 (0)	11 (0)	14 (0)	0	0.99
GBIC	0.1017	0.1063	0.1155	1 (0)	2 (0)	5.5 (0)	0.23	0.96
GBIC_p	0.1010	0.1048	0.1094	0 (0)	1 (0)	2 (0)	0.39	0.94
Oracle	0.0923	0.0944	0.0968	0 (0)	0 (0)	0 (0)	1	1

5.2.3. High dimensional logistic regression with interaction

We now consider high dimensional logistic regression with interaction. We simulated 100 data sets from the logistic regression model (2) with interaction and an n-dimensional parameter vector

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + 2\mathbf{x}_{p+1}, \tag{36}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is an $n \times p$ design matrix, $\mathbf{x}_{p+1} = \mathbf{x}_1 \circ \mathbf{x}_2$ is an interaction term and the rest are the same as in model (33). For each data set, the n-dimensional response vector $\mathbf{y}$ was sampled from the Bernoulli distribution with success probability vector $(\exp(\theta_1)/\{1 + \exp(\theta_1)\}, \ldots, \exp(\theta_n)/\{1 + \exp(\theta_n)\})^{\mathrm{T}}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$ is given in equation (36). As in Section 5.1.2, we consider the case that all covariates are independent of each other and thus set $\Sigma_0 = I_p$. We chose $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_0 = (2.5, -1.9, 2.8, -2.2, 3, 0, \ldots, 0)^{\mathrm{T}}$ and considered $(n, p) = (200, 1000)$. Although the data were generated from model (2) with parameter vector (36), we fitted the logistic regression model without interaction. This provides another example of misspecified models. As argued in Section 5.1.2, the oracle working model is $\text{supp}(\boldsymbol{\beta}_0) = \{1, \ldots, 5\}$ which indicates logistic regression model (2) with the first five covariates. As in Section 5.1.2, for each data set, we first applied the SICA method (Lv and Fan, 2009) implemented with the ICA algorithm (Fan and Lv, 2011) to build a sequence of sparse models and then selected the final model by using the same five model selection criteria.

We use the same five performance measures as in Section 5.1.2. The prediction error is defined as $E[Y - \exp(\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}})/\{1 + \exp(\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\}]^2$ with $\hat{\boldsymbol{\beta}}$ an estimate and $(\mathbf{x}^{\mathrm{T}}, Y)$ an independent observation. Table 6 summarizes the comparison results of the 25th, 50th and 75th percentiles of the first three performance measures as well as the selection probability and inclusion probability of the model estimated. The conclusions are similar to those in Section 5.1.2. GAIC, GBIC and GBIC_p improved over AIC and BIC in terms of prediction and variable selection.

6. Discussions

We have considered the problem of model selection in misspecified models under two well-known model selection principles: the Bayesian principle and the KL divergence principle. The novel asymptotic expansions of the two principles in misspecified GLMs lead to GBIC and GAIC. In particular, a specific form of prior probabilities motivated by the KL divergence principle leads to GBIC_p, which has a natural decomposition into the negative maximum quasi-log-likelihood

and two penalties on model dimensionality and model misspecification. Numerical studies have demonstrated the advantage of the new methods for model selection in both correctly specified and misspecified models.

When the model is misspecified, the posterior distribution of the parameter may not reflect the true uncertainty of the estimation. In particular, the QMLE $\hat{\boldsymbol{\beta}}_n$ can have an asymptotic covariance matrix that is different from the posterior covariance matrix of $\boldsymbol{\beta}$. When this difference grows larger as the sample size n increases, its effect on model selection can be quite significant. To mitigate this issue, we have considered an asymptotic framework that allows the prior distribution to vary with sample size, and we introduced a local constraint on the prior density preventing it from giving diminishingly low probability to a shrinking region containing the theoretically best model.

Our technical analysis reveals that the covariance contrast matrix $\mathbf{H}_n = \mathbf{A}_n^{-1}\mathbf{B}_n$ between the covariance structures in the misspecified model and in the true model plays a pivotal role in our model selection framework. Specifically, effective estimates of its trace and log-determinant are needed for implementing the model selection criteria. Through numerical studies, we found that estimation of these two quantities exhibits different behaviour. We considered the simple estimator for $\log|\mathbf{H}_n|$ and the bootstrap estimator for $\mathrm{tr}(\mathbf{H}_n)$; these estimators may not be the best possible and the estimation of both quantities $\mathrm{tr}(\mathbf{H}_n)$ and $\log|\mathbf{H}_n|$ needs rigorous theoretical studies, especially in high dimensions. Some other possible ways of estimating $\mathbf{H}_n$ were discussed in Section 3.4. The problem of model selection in misspecified models is challenging and our studies can be regarded as the first step towards understanding the effect of explicitly incorporating model misspecification for model selection.

So far we have explored the expression of GBIC_p taking an additive form of the three terms goodness of fit, model complexity and model misspecification. Possible extensions of GBIC_p include introducing other weights and considering non-additive forms. We considered the classical asymptotic setting of fixed dimensionality in our asymptotic expansions for technical simplicity. It would be interesting to extend all the results to high dimensions as well as more general model settings, with the dimensionality diverging with the sample size. It also remains open to characterize the optimality of these model selection criteria in misspecified models. These problems are beyond the scope of the current paper and will be interesting topics for future research.

Acknowledgements

Lv's research was supported by National Science Foundation 'CAREER' award DMS-0955316 and grant DMS-0806030 and 2008 Zumberge individual award from the University of Southern California's James H. Zumberge faculty research and innovation fund. Liu's research was supported by National Science Foundation grant DMS-0706989 and National Institutes of Health grant R01-HG02518-02. We sincerely thank the Joint Editor, an Associate Editor and a referee for their constructive comments that significantly improved the paper.

Appendix A: Proofs

The proofs of theorem 5 and propositions 1 and 2 are omitted here and are given in Lv and Liu (2010).

A.1. Proof of theorem 1

We condition on the event $\tilde{\mathcal{Q}}_n = \{\hat{\boldsymbol{\beta}}_n \in N_n(\delta_n)\}$, where $N_n(\delta_n) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\mathbf{B}_n^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n,0})\| \leqslant \delta_n\}$. By the proof of consistency of the QMLE $\hat{\boldsymbol{\beta}}_n$ in theorem 6, we have

$$P(\tilde{Q}_n) \geqslant 1 - \frac{4d}{c^2\delta_n^2} \to 1$$

since $\delta_n \to \infty$ by condition 5. It is easy to see that $l_n^*(\mathbf{y}, \boldsymbol{\beta})$ defined in equation (17) is a smooth concave function on $\mathbb{R}^d$ with its maximum 0 attained at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_n$,

$$\partial^2 l_n^*(\mathbf{y}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}^2 = -\mathbf{A}_n(\boldsymbol{\beta}),$$

and $U_n(\boldsymbol{\beta}) = \exp\{n^{-1} l_n^*(\mathbf{y}, \boldsymbol{\beta})\}$ takes values between 0 and 1. Thus, by Taylor's theorem, expanding $l_n^*(\mathbf{y}, \cdot)$ around $\hat{\boldsymbol{\beta}}_n$ gives, for any $\boldsymbol{\beta}$ in the neighbourhood $\tilde{N}_n(\delta_n) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\mathbf{B}_n^{1/2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)\| \leqslant \delta_n\}$,

$$l_n^*(\mathbf{y}, \boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)^{\mathrm{T}} \frac{\partial^2 l_n^*(\mathbf{y}, \boldsymbol{\beta}_*)}{\partial\boldsymbol{\beta}^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n) = -\frac{n}{2}\boldsymbol{\delta}^{\mathrm{T}} \mathbf{V}_n(\boldsymbol{\beta}_*)\boldsymbol{\delta}, \tag{37}$$

where $\boldsymbol{\beta}_*$ lies on the line segment joining $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_n$, $\boldsymbol{\delta} = n^{-1/2}\mathbf{B}_n^{1/2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)$ and $\mathbf{V}_n(\boldsymbol{\beta}) = \mathbf{B}_n^{-1/2}\mathbf{A}_n(\boldsymbol{\beta})\mathbf{B}_n^{-1/2}$. Clearly, the neighbourhood $\tilde{N}_n(\delta_n)$ defined above is a convex set containing $\hat{\boldsymbol{\beta}}_n$ and thus $\boldsymbol{\beta}_* \in \tilde{N}_n(\delta_n) \subset N_n(2\delta_n)$. Let

$$\rho_n(\delta_n) = \max_{\boldsymbol{\beta} \in N_n(2\delta_n)} \max[|\lambda_{\min}\{\mathbf{V}_n(\boldsymbol{\beta}) - \mathbf{V}_n\}|, |\lambda_{\max}\{\mathbf{V}_n(\boldsymbol{\beta}) - \mathbf{V}_n\}|]$$

with $\mathbf{V}_n = \mathbf{V}_n(\boldsymbol{\beta}_{n,0})$. By conditions 5 and 2, $\rho_n(\delta_n) \leqslant \lambda_{\min}(\mathbf{V}_n)/2$ for sufficiently large n. Thus it follows from equation (37) that

$$q_1(\boldsymbol{\beta})\mathbf{1}_{\tilde{N}_n(\delta_n)}(\boldsymbol{\beta}) \leqslant -n^{-1} l_n^*(\mathbf{y}, \boldsymbol{\beta})\mathbf{1}_{\tilde{N}_n(\delta_n)}(\boldsymbol{\beta}) \leqslant q_2(\boldsymbol{\beta})\mathbf{1}_{\tilde{N}_n(\delta_n)}(\boldsymbol{\beta}), \tag{38}$$

where $q_1(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\{\mathbf{V}_n - \rho_n(\delta_n)I_d\}\boldsymbol{\delta}$ and $q_2(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\{\mathbf{V}_n + \rho_n(\delta_n)I_d\}\boldsymbol{\delta}$. This entails

$$E_{\mu_{\mathfrak{M}}}\{\exp(-nq_2)\mathbf{1}_{\tilde{N}_n(\delta_n)}\} \leqslant E_{\mu_{\mathfrak{M}}}\{U_n(\boldsymbol{\beta})^n \mathbf{1}_{\tilde{N}_n(\delta_n)}\} \leqslant E_{\mu_{\mathfrak{M}}}\{\exp(-nq_1)\mathbf{1}_{\tilde{N}_n(\delta_n)}\}. \tag{39}$$

We shall see later that inequality (39) is the key step in deriving the asymptotic expansion of $\log[E_{\mu_{\mathfrak{M}}}\{U_n(\boldsymbol{\beta})^n\}]$ in equation (18).

We list some auxiliary results in the following two lemmas, whose proofs are similar to those given in Lv and Liu (2010).

Lemma 1. Under condition 5 in Section 4.1, we have

$$c_1 \int \exp(-nq_j)\mathbf{1}_{\tilde{N}_n(\delta_n)}\,\mathrm{d}\mu_0 \leqslant E_{\mu_{\mathfrak{M}}}\exp(-nq_j)\mathbf{1}_{\tilde{N}_n(\delta_n)}\} \leqslant c_2 \int \exp(-nq_j)\mathbf{1}_{\tilde{N}_n(\delta_n)}\,\mathrm{d}\mu_0, \qquad j = 1, 2. \tag{40}$$

Lemma 2. It holds that

$$E_{\mu_{\mathfrak{M}}}\{U_n(\boldsymbol{\beta})^n \mathbf{1}_{\tilde{N}_n^c(\delta_n)}\} \leqslant \exp(-\kappa_n\delta_n^2), \tag{41}$$

$$\int \exp(-nq_1)\,\mathrm{d}\mu_0 = \left(\frac{2\pi}{n}\right)^{d/2} |\mathbf{V}_n - \rho_n(\delta_n)I_d|^{-1/2}, \tag{42}$$

$$\int \exp(-nq_2)\,\mathrm{d}\mu_0 = \left(\frac{2\pi}{n}\right)^{d/2} |\mathbf{V}_n + \rho_n(\delta_n)I_d|^{-1/2}, \tag{43}$$

$$\int \exp(-nq_j)\mathbf{1}_{\tilde{N}_n^c(\delta_n)}\,\mathrm{d}\mu_0 \leqslant \left(\frac{2\pi}{n\kappa_n}\right)^{d/2} \exp\left\{-\frac{1}{2}\kappa_n\delta_n^2 + \frac{d}{2} + \frac{d}{2}\log(\kappa_n\delta_n^2 d^{-1})\right\}, \qquad j = 1, 2, \tag{44}$$

where $\kappa_n = \lambda_{\min}(\mathbf{V}_n)/2$ and it is assumed that $\kappa_n\delta_n^2 > d$ in expression (44).

Now we are ready to obtain the asymptotic expansion of $\log E_{\mu_{\mathfrak{M}}}\{U_n(\boldsymbol{\beta})^n\}$ in equation (18). By conditions 5 and 2, $\lambda_{\min}^{-1}(\mathbf{V}_n)\delta_n^{-2} = o\{\log^{-1}(n)\}$, which along with inequality (41) shows that the term $E_{\mu_{\mathfrak{M}}}\{U_n(\boldsymbol{\beta})^n \mathbf{1}_{\tilde{N}_n^c(\delta_n)}\}$ converges to 0 at a rate faster than any polynomial rate in n. Similarly, in view of inequality (44), we can show that the two terms $\int \exp(-nq_j)\mathbf{1}_{\tilde{N}_n^c(\delta_n)}\,\mathrm{d}\mu_0$ with $q = 1, 2$ converge to 0 at a rate faster than any polynomial rate in n. It follows from $\rho_n(\delta_n) = o\{\lambda_{\min}(\mathbf{V}_n)\}$ that

$$|\mathbf{V}_n \pm \rho_n(\delta_n) I_d|^{-1/2} = |\mathbf{V}_n|^{-1/2} |I_d \pm \rho_n(\delta_n)\mathbf{V}_n^{-1}|^{-1/2}$$
$$= |\mathbf{V}_n|^{-1/2}[1 + O\{\rho_n(\delta_n)\operatorname{tr}(\mathbf{V}_n^{-1})\}]$$
$$= |\mathbf{V}_n|^{-1/2}[1 + O\{\rho_n(\delta_n) d \lambda_{\min}^{-1}(\mathbf{V}_n)\}] = |\mathbf{V}_n|^{-1/2}\{1 + o(1)\}.$$

Since $\lambda_{\max}(\mathbf{V}_n)$ is assumed to be of a polynomial order of n in condition 5, combining these results with inequality (39) and lemmas 1 and 2 yields

$$\log[E_{\mu_{\mathfrak{M}}}\{U_n(\boldsymbol{\delta})^n\}] = \log\left[\left(\frac{2\pi}{n}\right)^{d/2} |\mathbf{V}_n|^{-1/2}\{1 + o(1)\}\right] + \log(c_n)$$
$$= -\frac{\log(n)}{2} d + \frac{1}{2}\log|\mathbf{A}_n^{-1}\mathbf{B}_n| + \frac{\log(2\pi)}{2} d + \log(c_n) + o(1),$$

where $c_n \in [c_1, c_2]$. This together with equation (18) proves the conclusion on the event $\tilde{Q}_n$.

A.2. Proof of theorem 2

By equation (21) and the first part of the proof of theorem 3, we have

$$\log(\alpha_{\mathfrak{M}}) = -\tfrac{1}{2}\operatorname{tr}(\mathbf{H}_n) + \log(C) + o(1),$$

which along with theorem 1 leads to the desired asymptotic expansion of the Bayes factor.

A.3. Proof of theorem 3

It suffices to prove

$$E\{\eta_n(\hat{\boldsymbol{\beta}}_n)\} = \eta_n(\boldsymbol{\beta}_{n,0}) - \tfrac{1}{2}\operatorname{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n) + o(1) \tag{45}$$

and

$$\eta_n(\boldsymbol{\beta}_{n,0}) = E\{l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n)\} - \tfrac{1}{2}\operatorname{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n) + o(1). \tag{46}$$

We shall prove equations (45) and (46) in separate steps. Define

$$\tilde{l}_n(\mathbf{y}, \boldsymbol{\beta}) = l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n + n^{1/2}\mathbf{C}_n^{-1}\boldsymbol{\beta})$$

and

$$\tilde{\eta}_n(\boldsymbol{\beta}) = \eta_n(\boldsymbol{\beta}_{n,0} + n^{1/2}\mathbf{C}_n^{-1}\boldsymbol{\beta}),$$

where $\mathbf{C}_n = \mathbf{B}_n^{-1/2}\mathbf{A}_n$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$.

(a) We first prove equation (45). The key idea is to do a second-order Taylor series expansion of $\tilde{\eta}_n(\boldsymbol{\beta})$ around $\mathbf{0}$ and to retain the Lagrange remainder term. By theorem 5, η_n attains its maximum at $\boldsymbol{\beta}_{n,0}$. Thus we derive

$$\eta_n(\hat{\boldsymbol{\beta}}_n) = \tilde{\eta}_n\{n^{-1/2}\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})\} = \tilde{\eta}_n(\mathbf{0}) - \frac{n}{2}\mathbf{v}^{\mathrm{T}}(\mathbf{C}_n^{-1}\mathbf{A}_n\mathbf{C}_n^{-1})\mathbf{v} - \frac{1}{6}\sum_{j,k,l}\frac{\partial^3\tilde{\eta}_n(\boldsymbol{\beta}_*)}{\partial\beta_j\,\partial\beta_k\,\partial\beta_l} v_j v_k v_l$$
$$= \eta_n(\boldsymbol{\beta}_{n,0}) - \frac{n}{2}\mathbf{v}^{\mathrm{T}}(\mathbf{B}_n^{1/2}\mathbf{A}_n^{-1}\mathbf{B}_n^{1/2})\mathbf{v} - \frac{1}{6}\sum_{j,k,l}\frac{\partial^3\tilde{\eta}_n(\boldsymbol{\beta}_*)}{\partial\beta_j\,\partial\beta_k\,\partial\beta_l} v_j v_k v_l,$$

where $\mathbf{v} = (v_1, \ldots, v_d)^{\mathrm{T}} = n^{-1/2}\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})$ and $\boldsymbol{\beta}_*$ lies on the line segment joining $\mathbf{v}$ and $\mathbf{0}$. By condition 6,

$$|\partial^3\tilde{\eta}_n(\boldsymbol{\beta}_*)/\partial\beta_j\,\partial\beta_k\,\partial\beta_l| = o(n^{3/2}).$$

Since $\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \to^{\mathcal{D}} N(\mathbf{0}, I_d)$ by theorem 7 and $E\|\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})\|^{3+\alpha} = O(1)$ for some $\alpha > 0$ by condition 6, theorem 6.2 in DasGupta (2008) applies to show that, for any j, k and l,

$$E(n v_j v_k) \to \delta_{jk},$$
$$E(n^{3/2}|v_j v_k v_l|) = O(1),$$

where δ_{jk} denotes the Kronecker delta. By condition 2, we have $\lambda_{\min}(\mathbf{V}_n) \geqslant c$ for some positive constant c, which entails that

$$\begin{aligned}\operatorname{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n) &= \operatorname{tr}(\mathbf{B}_n^{1/2}\mathbf{A}_n^{-1}\mathbf{B}_n^{1/2}) \leqslant d\lambda_{\max}(\mathbf{B}_n^{1/2}\mathbf{A}_n^{-1}\mathbf{B}_n^{1/2}) \\ &= d\lambda_{\min}^{-1}(\mathbf{V}_n) = O(1).\end{aligned}$$

Combining these results yields

$$\begin{aligned}E\{\eta_n(\hat{\boldsymbol{\beta}}_n)\} &= \eta_n(\boldsymbol{\beta}_{n,0}) - \left\{\tfrac{1}{2}\operatorname{tr}(\mathbf{B}_n^{1/2}\mathbf{A}_n^{-1}\mathbf{B}_n^{1/2}) + o(1)\right\} + \tfrac{1}{6}\sum_{j,k,l} o(n^{3/2})\, O(n^{-3/2}) \\ &= \eta_n(\boldsymbol{\beta}_{n,0}) - \tfrac{1}{2}\operatorname{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n) + o(1).\end{aligned}$$

(b) We now prove equation (46). The key idea is to represent $l_n(\mathbf{y}, \boldsymbol{\beta}_{n,0})$ by using a second-order Taylor series expansion of $l_n(\mathbf{y}, \cdot)$ around $\hat{\boldsymbol{\beta}}_n$ and retaining the Lagrange remainder term. It is easy to see that the difference between $\eta_n(\boldsymbol{\beta})$ and $l_n(\mathbf{y}, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, which implies that each pair of their second or higher order partial derivatives agree. Since $l_n(\mathbf{y}, \cdot)$ attains its maximum at $\hat{\boldsymbol{\beta}}_n$, we derive

$$\begin{aligned}l_n(\mathbf{y}, \boldsymbol{\beta}_{n,0}) &= \tilde{l}_n(\mathbf{y}, \mathbf{v}) = \tilde{l}_n(\mathbf{0}) - \frac{1}{2}\mathbf{v}^{\mathrm{T}}\frac{\partial^2 \tilde{l}_n(\mathbf{y}, \mathbf{0})}{\partial \boldsymbol{\beta}^2}\mathbf{v} - \frac{1}{6}\sum_{j,k,l}\frac{\partial^3 \tilde{l}_n(\mathbf{y}, \boldsymbol{\beta}_*)}{\partial\beta_j\,\partial\beta_k\,\partial\beta_l} v_j v_k v_l \\ &= l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \frac{1}{2}\mathbf{v}^{\mathrm{T}}\frac{\partial^2 \tilde{\eta}_n(-\mathbf{v})}{\partial \boldsymbol{\beta}^2}\mathbf{v} - \frac{1}{6}\Sigma_{j,k,l}\frac{\partial^3 \tilde{\eta}_n(\boldsymbol{\beta}_{**})}{\partial\beta_j\partial\beta_k\partial\beta_l} v_j v_k v_l,\end{aligned}$$

where $\mathbf{v} = (v_1, \ldots, v_d)^{\mathrm{T}} = n^{-1/2}\mathbf{C}_n(\boldsymbol{\beta}_{n,0} - \hat{\boldsymbol{\beta}}_n)$, $\boldsymbol{\beta}_*$ lies on the line segment joining $\mathbf{v}$ and $\mathbf{0}$, and $\boldsymbol{\beta}_{**} = \boldsymbol{\beta}_* - \mathbf{v}$.

It follows from $\sup_{\beta}\max_{j,k,l}|\partial^3\tilde{\eta}_n(\boldsymbol{\beta})/\partial\beta_j\,\partial\beta_k\,\partial\beta_l| = o(n^{3/2})$ in condition 6 that

$$|\partial^3\tilde{\eta}_n(\boldsymbol{\beta}_{**})/\partial\beta_j\,\partial\beta_k\,\partial\beta_l| = o(n^{3/2})$$

and $\partial^2\tilde{\eta}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^2$ is Lipschitz with Lipschitz constant $o(n^{3/2})$. Thus

$$\begin{aligned}\mathbf{v}^{\mathrm{T}}\{\partial^2\tilde{\eta}_n(-\mathbf{v})/\partial\boldsymbol{\beta}^2\}\mathbf{v} &= \mathbf{v}^{\mathrm{T}}\{\partial^2\tilde{\eta}_n(\mathbf{0})/\partial\boldsymbol{\beta}^2\}\mathbf{v} + o(n^{3/2})\|\mathbf{v}\|^3 \\ &= n\mathbf{v}^{\mathrm{T}}(\mathbf{C}_n^{-1}\mathbf{A}_n\mathbf{C}_n^{-1})\mathbf{v} + o(n^{3/2})\|\mathbf{v}\|^3 \\ &= n\mathbf{v}^{\mathrm{T}}(\mathbf{B}_n^{1/2}\mathbf{A}_n^{-1}\mathbf{B}_n^{1/2})\mathbf{v} + o(n^{3/2})\|\mathbf{v}\|^3.\end{aligned}$$

Since $\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \to^{\mathcal{D}} N(\mathbf{0}, I_d)$ by theorem 7 and $E\|\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})\|^{3+\alpha} = O(1)$ for some $\alpha > 0$ by condition 6, theorem 6.2 in DasGupta (2008) applies to show that, for any j and k,

$$\begin{aligned}E(nv_jv_k) &\to \delta_{jk}, \\ E(n^{3/2}\|\mathbf{v}\|^3) &= O(1).\end{aligned}$$

These along with $\operatorname{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n) = O(1)$ yield

$$\begin{aligned}\eta_n(\boldsymbol{\beta}_{n,0}) &= E\{l_n(\mathbf{y}, \boldsymbol{\beta}_{n,0})\} = E\{l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n)\} - \left\{\tfrac{1}{2}\operatorname{tr}(\mathbf{B}_n^{1/2}\mathbf{A}_n^{-1}\mathbf{B}_n^{1/2}) + o(1)\right\} + o(n^{3/2})\, O(n^{-3/2}) \\ &= E\{l_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n)\} - \tfrac{1}{2}\operatorname{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n) + o(1),\end{aligned}$$

which completes the proof.

A.4. Proof of theorem 4

Since the model is correctly specified, i.e. $G_n = F_n(\cdot, \boldsymbol{\beta}_{n,0})$, we have the identity $\mathbf{A}_n = \mathbf{B}_n$, yielding $\mathbf{H}_n = \mathbf{A}_n^{-1}\mathbf{B}_n = I_d$. Note that, by assumption, the smallest and largest eigenvalues of $n^{-1}\mathbf{B}_n$ are bounded away from 0 and ∞. To prove the consistency of the estimator $\hat{\mathbf{H}}_n = \hat{\mathbf{A}}_n^{-1}\hat{\mathbf{B}}_n$, it suffices to show that

$$n^{-1}\hat{\mathbf{A}}_n = n^{-1}\mathbf{B}_n + o_P(1)$$

and

$$n^{-1}\hat{\mathbf{B}}_n = n^{-1}\mathbf{B}_n + o_P(1).$$

By the proof of theorem 6, we have $\mathbf{B}_n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) = O_P(1)$, which along with the assumption that $\lambda_{\min}(n^{-1}\mathbf{B}_n)$ is bounded away from 0 gives

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_{n,0} + O_P(n^{-1/2}).$$

By condition 7, $n^{-1}\mathbf{A}_n(\boldsymbol{\beta})$ is Lipschitz in an asymptotically shrinking neighbourhood of $\boldsymbol{\beta}_{n,0}$ given by $N_n(\delta) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\mathbf{B}_n^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n,0})\| \leqslant \delta\}$. It follows from the proof of theorem 6 that

$$P\{\hat{\boldsymbol{\beta}}_n \in N_n(\delta)\} \to 1 \qquad \text{as } n \to \infty.$$

Combining this result with the Lipschitz continuity of $n^{-1}\mathbf{A}_n(\boldsymbol{\beta})$ and the consistency of $\hat{\boldsymbol{\beta}}_n$ yields

$$n^{-1}\hat{\mathbf{A}}_n = n^{-1}\mathbf{B}_n + o_P(1).$$

It remains to prove that $n^{-1}\hat{\mathbf{B}}_n = n^{-1}\mathbf{B}_n + o_P(1)$. To analyse $n^{-1}\hat{\mathbf{B}}_n$, we decompose it into three terms

$$n^{-1}\hat{\mathbf{B}}_n = n^{-1}\mathbf{X}^{\mathrm{T}}\operatorname{diag}[\{\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\} \circ \{\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\}]\mathbf{X} \equiv \mathbf{G}_1 + \mathbf{G}_2 + \mathbf{G}_3,$$

with

$$\begin{aligned}
\mathbf{G}_1 &= n^{-1}\mathbf{X}^{\mathrm{T}}\operatorname{diag}[\{\mathbf{y} - E(\mathbf{y})\} \circ \{\mathbf{y} - E(\mathbf{y})\}]\mathbf{X}, \\
\mathbf{G}_2 &= -2n^{-1}\mathbf{X}^{\mathrm{T}}\operatorname{diag}[\{\mathbf{y} - E(\mathbf{y})\} \circ \{\boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})\}]\mathbf{X}, \\
\mathbf{G}_3 &= n^{-1}\mathbf{X}^{\mathrm{T}}\operatorname{diag}[\{\boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})\} \circ \{\boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})\}]\mathbf{X},
\end{aligned}$$

since $E(\mathbf{y}) = \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})$ by equation (3). Since

$$\mathbf{G}_1 = n^{-1}\sum_{i=1}^{n}[\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\{Y_i - E(Y_i)^2\}]$$

is the mean of n independent random variables and has asymptotically vanishing second central moment by condition 7, it follows from the law of large numbers that

$$\begin{aligned}
\mathbf{G}_1 &= E(\mathbf{G}_1) + o_P(1) \\
&= n^{-1}\mathbf{B}_n + o_P(1).
\end{aligned}$$

By condition 7, $n^{-1}\mathbf{X}^{\mathrm{T}}\operatorname{diag}[\{\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})\} \circ \{\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})\}]\mathbf{X}$ is Lipschitz in the neighbourhood $N_n(\delta)$. Thus, similarly to for $n^{-1}\hat{\mathbf{A}}_n$, we can show that

$$\mathbf{G}_3 = o_P(1).$$

We finally prove $\mathbf{G}_2 = o_P(1)$. By condition 7, $Y_i - E(Y_i)$ are uniformly bounded by $K_n = o(n^{1/2})$ with asymptotic probability 1, i.e.

$$P\{\max_{i=1}^{n}|Y_i - E(Y_i)| \leqslant K_n\} \to 1$$

as $n \to \infty$. Note that, on the above event, the largest absolute eigenvalue of the matrix $\mathbf{G}_2$ is bounded by the largest eigenvalue of

$$2n^{-1}\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{|\mathbf{y} - E(\mathbf{y})| \circ |\boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})|\}\mathbf{X} \leqslant 2K_n n^{-1}\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{|\boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})|\}\mathbf{X},$$

where the absolute value is understood componentwise and the inequality means the order for symmetric positive semidefinite matrices. By condition 7, $n^{-1}\mathbf{X}^{\mathrm{T}}\operatorname{diag}\{|\boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})|\}$ is Lipschitz in the neighbourhood $N_n(\delta)$. Thus, by $K_n = o(n^{1/2})$ and $\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_{n,0} + O_P(n^{-1/2})$, we have $G_2 = o_P(1)$. Combining these results concludes the proof.

A.5. Proof of theorem 6

It is easy to see that $N_n(\delta)$ is a convex set by its definition. Denote by

$$\partial N_n(\delta) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\mathbf{B}_n^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n,0})\| = \delta\}$$

the boundary of the closed set $N_n(\delta)$. We define an event

$$Q_n = \{l_n(\mathbf{y}, \boldsymbol{\beta}_{n,0}) > \max_{\boldsymbol{\beta} \in \partial N_n(\delta)} l_n(\mathbf{y}, \boldsymbol{\beta})\}.$$

We observe that, if the event Q_n occurs, the continuous function $l_n(\mathbf{y}, \cdot)$ has a local maximum in the interior of $N_n(\delta)$. Since the first part of condition 1 implies that $l_n(\mathbf{y}, \cdot)$ is strictly concave, this maximum must be located at $\hat{\boldsymbol{\beta}}_n$. This shows that

$$Q_n \subset \{\hat{\boldsymbol{\beta}}_n \in N_n(\delta)\}.$$

Next we construct a lower bound on $P(Q_n)$. By Taylor's theorem, we have, for any $\boldsymbol{\beta}$,

$$l_n(\mathbf{y},\boldsymbol{\beta}) - l_n(\mathbf{y},\boldsymbol{\beta}_{n,0}) = (\boldsymbol{\beta}-\boldsymbol{\beta}_{n,0})^{\mathrm{T}}\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0}) - \tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_{n,0})^{\mathrm{T}}\mathbf{A}_n(\boldsymbol{\beta}_*)(\boldsymbol{\beta}-\boldsymbol{\beta}_{n,0}),$$

where $\boldsymbol{\beta}_*$ lies on the line segment joining $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{n,0}$. We make a transformation of the variable by letting

$$\mathbf{u} = \delta^{-1}\mathbf{B}_n^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_{n,0}).$$

Then it follows that

$$l_n(\mathbf{y},\boldsymbol{\beta}) - l_n(\mathbf{y},\boldsymbol{\beta}_{n,0}) = \delta\mathbf{u}^{\mathrm{T}}\mathbf{B}_n^{-1/2}\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0}) - \delta^2\mathbf{u}^{\mathrm{T}}\,\mathbf{V}_n(\boldsymbol{\beta}_*)\mathbf{u}/2.$$

Note that $\boldsymbol{\beta}\in\partial N_n(\delta)$ if and only if $\|\mathbf{u}\|=1$, and that $\boldsymbol{\beta}\in\partial N_n(\delta)$ implies that $\boldsymbol{\beta}_*\in N_n(\delta)$ by the convexity of $N_n(\delta)$. Clearly

$$\max_{\|\mathbf{u}\|=1}\mathbf{u}^{\mathrm{T}}\mathbf{B}_n^{-1/2}\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0}) = \|\mathbf{B}_n^{-1/2}\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})\|$$

and, by the second part of condition 2, for n sufficiently large we have

$$\min_{\|\mathbf{u}\|=1}\mathbf{u}^{\mathrm{T}}\,\mathbf{V}_n(\boldsymbol{\beta}_*)\mathbf{u} \geqslant \min_{\boldsymbol{\beta}\in N_n(\delta)}\lambda_{\min}\{\mathbf{V}_n(\boldsymbol{\beta})\} \geqslant c.$$

Thus we have

$$\max_{\boldsymbol{\beta}\in\partial N_n(\delta)} l_n(\mathbf{y},\boldsymbol{\beta}) - l_n(\mathbf{y},\boldsymbol{\beta}_{n,0}) \leqslant \delta\{\|\mathbf{B}_n^{-1/2}\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})\| - c\delta/2\},$$

which along with Markov's inequality entails that

$$P(Q_n) \geqslant P\{\|\mathbf{B}_n^{-1/2}\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})\|^2 < c^2\delta^2/4\} \geqslant 1 - \frac{E\{\|\mathbf{B}_n^{-1/2}\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})\|^2\}}{c^2\delta^2/4}.$$

Observe that

$$\begin{aligned} E\{\|\mathbf{B}_n^{-1/2}\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})\|^2\} &= E[\mathrm{tr}\{\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})^{\mathrm{T}}\mathbf{B}_n^{-1}\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})\}] = E[\mathrm{tr}\{\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})^{\mathrm{T}}\mathbf{B}_n^{-1}\}] \\ &= \mathrm{tr}[E\{\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})\,\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})^{\mathrm{T}}\}\mathbf{B}_n^{-1}] = \mathrm{tr}(I_d) = d. \end{aligned}$$

For any given $\eta\in(0,1)$, letting $\delta = 2d^{1/2}/(c\eta^{1/2})$ thus makes

$$P(Q_n) \geqslant 1 - \frac{4d}{c^2\delta^2} = 1-\eta.$$

This, together with $Q_n \subset \{\hat{\boldsymbol{\beta}}_n \in N_n(\delta)\}$ and the first part of condition 2, proves that $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0} = o_P(1)$, which completes the proof.

A.6. Proof of theorem 7

We condition on the event $Q_n = \{l_n(\mathbf{y},\boldsymbol{\beta}_{n,0}) > \max_{\boldsymbol{\beta}\in\partial N_n(\delta)} l_n(\mathbf{y},\boldsymbol{\beta})\}$ defined in the proof of theorem 6, which has been shown to entail that $\hat{\boldsymbol{\beta}}_n \in N_n(\delta)$. By the mean value theorem applied componentwise and equation (31), we obtain

$$\begin{aligned} \mathbf{0} = \boldsymbol{\Psi}_n(\hat{\boldsymbol{\beta}}_n) &= \boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0}) - \tilde{\mathbf{A}}_n(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \\ &= \mathbf{X}^{\mathrm{T}}\{\mathbf{y} - E(\mathbf{y})\} - \tilde{\mathbf{A}}_n(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}), \end{aligned}$$

where each of $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d$ lies on the line segment joining $\hat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_{n,0}$. Let $\mathbf{C}_n = \mathbf{B}_n^{-1/2}\mathbf{A}_n$.

Then we have $\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) = \mathbf{u}_n + \mathbf{w}_n$, where $\mathbf{u}_n = -\mathbf{B}_n^{-1/2}\mathbf{X}^{\mathrm{T}}\{\mathbf{y} - E(\mathbf{y})\}$ and

$$\mathbf{w}_n = -\{\tilde{\mathbf{V}}_n(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_d) - \mathbf{V}_n\}\{\mathbf{B}_n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})\}.$$

By Slutsky's lemma and the proof of theorem 6, we see that, to show that $\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \to^{\mathscr{D}} N(\mathbf{0}, I_d)$, it suffices to prove that $\mathbf{u}_n \to^{\mathscr{D}} N(\mathbf{0}, I_d)$ and $\mathbf{w}_n = o_P(1)$.

The result $\mathbf{w}_n = o_P(1)$ follows from $Q_n \subset \{\hat{\boldsymbol{\beta}}_n \in N_n(\delta)\}$ and condition 3. Finally we show that $\mathbf{u}_n \to^{\mathscr{D}} N(\mathbf{0}, I_d)$. Fix an arbitrary unit vector $\mathbf{a}\in\mathbb{R}^d$. We consider the asymptotic distribution of the linear combination $v_n = \mathbf{a}^{\mathrm{T}}\mathbf{u}_n = -\mathbf{a}^{\mathrm{T}}\mathbf{B}_n^{-1/2}\mathbf{X}^{\mathrm{T}}\{\mathbf{y} - E(\mathbf{y})\} = \Sigma_{i=1}^n z_i$, where $z_i = -\mathbf{a}^{\mathrm{T}}\mathbf{B}_n^{-1/2}\mathbf{x}_i\{y_i - E(y_i)\}$, $i = 1,\ldots,n$. Clearly z_is are independent and have mean 0, and $\Sigma_{i=1}^n \mathrm{var}(z_i) = \mathbf{a}^{\mathrm{T}}\mathbf{B}_n^{-1/2}\mathbf{B}_n\mathbf{B}_n^{-1/2}\mathbf{a} = 1$. By condition 4, $E|y_i - E(y_i)|^3 \leqslant M$ for some positive constant M. Thus we derive

$$\sum_{i=1}^n E(|z_i|^3) = \sum_{i=1}^n |\mathbf{a}^{\mathrm{T}}\mathbf{B}_n^{-1/2}\mathbf{x}_i|^3\,E\{|y_i - E(y_i)|^3\} \leqslant M\sum_{i=1}^n |\mathbf{a}^{\mathrm{T}}\mathbf{B}_n^{-1/2}\mathbf{x}_i|^3$$

$$\leqslant M \sum_{i=1}^{n} \|\mathbf{a}\|^3 \|\mathbf{B}_n^{-1/2}\mathbf{x}_i\|^3 = M \sum_{i=1}^{n} (\mathbf{x}_i^{\mathrm{T}}\mathbf{B}_n^{-1}\mathbf{x}_i)^{3/2} \to 0,$$

where we used the Cauchy–Schwarz inequality and condition 4. Applying Lyapunov's theorem yields $\mathbf{a}^{\mathrm{T}}\mathbf{u}_n = \Sigma_{i=1}^{n} z_i \to^{\mathcal{D}} N(0,1)$. Since this asymptotic normality holds for any unit vector $\mathbf{a} \in \mathbb{R}^d$, we conclude that $\mathbf{u}_n \to^{\mathcal{D}} N(\mathbf{0}, I_d)$, which concludes the proof.

Appendix B: Three commonly used generalized linear models

In this section, we provide the formulae for matrices $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ in equations (29) and (30), when the working model F_n is chosen to be one of the three commonly used GLMs: the linear regression model, logistic regression model or Poisson regression model.

B.1. Linear regression

The linear regression model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{47}$$

where $\mathbf{y}$ is an n-dimensional response vector, $\mathbf{X}$ is an $n \times d$ design matrix, $\boldsymbol{\gamma}$ is a d-dimensional regression coefficient vector and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ is an n-dimensional error vector. Let $\boldsymbol{\beta} = \boldsymbol{\gamma}/\sigma^2$ and $b(\theta) = \sigma^2\theta^2/2$, $\theta \in \mathbb{R}$. Then the quasi-log-likelihood of the sample in equation (6) becomes

$$\begin{aligned} l_n(\mathbf{y}, \boldsymbol{\beta}) &= \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - \mathbf{1}^{\mathrm{T}}\mathbf{b}(\mathbf{X}\boldsymbol{\beta}) - \frac{\|\mathbf{y}\|_2^2}{2\sigma^2} - \frac{n}{2}\log(\sigma^2) - \frac{n\log(2\pi)}{2} \\ &= -\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}\|_2^2}{2\sigma^2} - \frac{n}{2}\log(\sigma^2) - \frac{n\log(2\pi)}{2}. \end{aligned}$$

Maximizing $l_n(\mathbf{y}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\gamma}$ yields the least squares estimator $\hat{\boldsymbol{\gamma}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$.

We estimate the error variance σ^2 by using the residual sum of squares, $\hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\gamma}}\|_2^2/(n-d)$. Then we obtain an estimate $\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\gamma}}/\hat{\sigma}^2$. Since $b'(\theta) = \sigma^2\theta$ and $b''(\theta) = \sigma^2$, we have

$$\begin{aligned} \hat{\mathbf{A}}_n &= \hat{\sigma}^2\mathbf{X}^{\mathrm{T}}\mathbf{X}, \\ \hat{\mathbf{B}}_n &= \mathbf{X}^{\mathrm{T}}\,\mathrm{diag}\{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\gamma}}) \circ (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\gamma}})\}\mathbf{X}, \end{aligned} \tag{48}$$

where '$\circ$' denotes the Hadamard (componentwise) product.

An interesting special case is when the true model is indeed linear but may involve a different set of covariates. Then $\mathbf{B}_n = \tau^2\mathbf{X}^{\mathrm{T}}\mathbf{X}$, where τ^2 is the true variance of the response variable Y if we had known the true model precisely. GBIC_p just involves an additional term penalizing the inflation of the error variance due to model misspecification.

B.2. Logistic regression

In the logistic regression model, $b(\theta) = \log\{1 + \exp(\theta)\}$, $\theta \in \mathbb{R}$, and the quasi-log-likelihood of the sample is $l_n(\mathbf{y}, \boldsymbol{\beta}) = \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - \mathbf{1}^{\mathrm{T}}\mathbf{b}(\mathbf{X}\boldsymbol{\beta})$ with the last constant term in equation (6) ignored, which attains its maximum at $\hat{\boldsymbol{\beta}}_n$. Since $b'(\theta) = \exp(\theta)/\{1 + \exp(\theta)\}$ and $b''(\theta) = \exp(\theta)/\{1 + \exp(\theta)\}^2$, matrices $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ are given in equations (29) and (30) with

$$\begin{aligned} \boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) &= (\exp(\theta_1)/\{1 + \exp(\theta_1)\}, \ldots, \exp(\theta_n)/\{1 + \exp(\theta_n)\})^{\mathrm{T}}, \\ \boldsymbol{\Sigma}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) &= \mathrm{diag}[\exp(\theta_1)/\{1 + \exp(\theta_1)\}^2, \ldots, \exp(\theta_n)/\{1 + \exp(\theta_n)\}^2], \end{aligned}$$

and $(\theta_1, \ldots, \theta_n)^{\mathrm{T}} = \mathbf{X}\hat{\boldsymbol{\beta}}_n$.

B.3. Poisson regression

In the Poisson regression model, $b(\theta) = \exp(\theta)$, $\theta \in \mathbb{R}$, and the quasi-log-likelihood of the sample is $l_n(\mathbf{y}, \boldsymbol{\beta}) = \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - \mathbf{1}^{\mathrm{T}}\mathbf{b}(\mathbf{X}\boldsymbol{\beta})$ with the last constant term in equation (6) ignored, which attains its maximum at $\hat{\boldsymbol{\beta}}_n$. Since $b'(\theta) = \exp(\theta)$ and $b''(\theta) = \exp(\theta)$, matrices $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ are given in equations (29) and (30) with $\boldsymbol{\mu}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) = (\exp(\theta_1), \ldots, \exp(\theta_n))^{\mathrm{T}}$ and $\boldsymbol{\Sigma}(\mathbf{X}\hat{\boldsymbol{\beta}}_n) = \mathrm{diag}\{\exp(\theta_1), \ldots, \exp(\theta_n)\}$, and $(\theta_1, \ldots, \theta_n)^{\mathrm{T}} = \mathbf{X}\hat{\boldsymbol{\beta}}_n$.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csàki), pp. 267–281. Budapest: Akademiai Kiàdo.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

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

Bozdogan, H. (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Bozdogan, H. (2000) Akaike's information criterion and recent developments in information complexity. *J. Math. Psychol.*, **44**, 62–91.

Burnham, K. P. and Anderson, D. R. (1998) *Model Selection and Inference: a Practical Information-theoretic Approach*. New York: Springer.

Casella, G., Girón, F. J., Martínez, M. L. and Moreno, E. (2009) Consistency of Bayesian procedures for variable selection. *Ann. Statist.*, **37**, 1207–1228.

Cavanaugh, J. E. and Neath, A. A. (1999) Generalizing the derivation of the Schwarz information criterion. *Communs Statist. Theor. Meth.*, **28**, 49–66.

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

DasGupta, A. (2008) *Asymptotic Theory of Statistics and Probability*. New York: Springer.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

Fahrmeir, L. and Kaufmann, H. (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, **13**, 342–368.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.

Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statist. Sin.*, **20**, 101–148.

Fan, J. and Lv, J. (2011) Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theor.*, **57**, 5467–5484.

Fan, Y. and Tang, C. (2013) Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc.* B, **75**, 531–552.

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

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*, 2nd edn. Boca Raton: Chapman and Hall–CRC.

Hall, P. (1990) Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Probab. Theor. Reltd Flds*, **85**, 449–467.

Hosking, J. R. M. (1980) Lagrange-multiplier tests of time-series models. *J. R. Statist. Soc.* B, **42**, 170–181.

Konishi, S. and Kitagawa, G. (1996) Generalised information criterion in model selection. *Biometrika*, **83**, 875–890.

Kullback, S. and Leibler, R. (1951) On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

Liu, W. and Yang, Y. (2011) Parametric or nonparametric?: a parametricness index for model selection. *Ann. Statist.*, **39**, 2074–2102.

Lv, J. and Fan, Y. (2009) A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, **37**, 3498–3528.

Lv, J. and Liu, J. S. (2010) Model selection principles in misspecified models. *arXiv Preprint math.ST/1005.5483*.

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

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

Shibata, R. (1989) Statistical aspects of model selection. In *From Data to Model* (ed. J. C. Willems), pp. 215–240. New York: Springer.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.

Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc.* B, **39**, 44–47.

Takeuchi, K. (1976) Distribution of information statistics and criteria for adequacy of models (in Japanese). *Math. Sci.*, **153**, 12–18.

Tian, L., Cai, T., Goetghebeur, E. and Wei, L. J. (2007) Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*, **94**, 297–311.

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

White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.

Yang, Y. and Barron, A. R. (1998) An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theor.*, **44**, 95–116.

Żak-Szatkowska, M. and Bogdan, M. (2011) Modified versions of Bayesian information criterion for sparse generalized linear models. *Computnl Statist. Data Anal.*, **55**, 2908–2924.