

Model Selection Principles in Misspecified Models

Jinchi Lv and Jun S. Liu *

University of Southern California and Harvard University

April 10, 2010

Abstract

Model selection is of fundamental importance to high dimensional modeling featured in many contemporary applications. Classical principles of model selection include the Kullback-Leibler divergence principle and the Bayesian principle, which lead to the Akaike information criterion and Bayesian information criterion when models are correctly specified. Yet model misspecification is unavoidable when we have no knowledge of the true model or when we have the correct family of distributions but miss some true predictor. In this paper, we propose a family of semi-Bayesian principles for model selection in misspecified models, which combine the strengths of the two well-known principles. We derive asymptotic expansions of the semi-Bayesian principles in misspecified generalized linear models, which give the new semi-Bayesian information criteria (SIC). A specific form of SIC admits a natural decomposition into 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 newly proposed SIC methodology for model selection in both correctly specified and misspecified models.

Running title: New Principles in Misspecified Models

Key words: Model selection; Model misspecification; Kullback-Leibler divergence principle; Bayesian principle; Semi-Bayesian principles; AIC; BIC; SIC

1 Introduction

High dimensional modeling is commonly encountered in many contemporary applications. The data we consider in this article is of the type $(y_i, x_{i1}, \dots, x_{ip})_{i=1}^n$, where the y_i 's are

*Jinchi Lv is Assistant Professor, Information and Operations Management Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA (e-mail: jinchilv@marshall.usc.edu). Jun S. Liu is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138, USA (e-mail: jliu@stat.harvard.edu). Lv's research was partially supported by NSF Grant DMS-0806030 and 2008 Zumberge Individual Award from USC's James H. Zumberge Faculty Research and Innovation Fund. Liu's research was partially supported by NSF Grant DMS-0706989 and NIH Grant R01-HG02518-02.

n independent observations of the response variable Y conditional on its covariates, or explanatory variables, $(x_{i1}, \dots, x_{ip})^T$. When the dimensionality (or the number of covariates) p is large compared to the sample size n , it is desirable to produce sparse models that involve subsets of predictors. With such models one can improve the prediction accuracy and enhance the model interpretability. See Fan and Lv (2010) for an overview of recent developments in high dimensional variable selection. A natural and fundamental problem that arises from such a task is 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 Schwartz (1978), respectively, when the models are correctly specified. The AIC and BIC have been proven to be powerful tools for model selection in various settings, see Burnham and Anderson (1998) for a book-length account of these developments. Stone (1977) shows heuristically the asymptotic equivalence of AIC and cross-validation under the true model, where a logarithmic assessment of the performance of a predicting density is used in the cross-validation. Bozdogan (1987) studies asymptotic properties of the AIC and BIC, showing that asymptotically AIC has a positive probability of overestimating the true dimension, while BIC is asymptotically consistent in estimating the true model. Hall (1990) compares AIC with KL cross-validation in the setting of histogram density estimation. Yang and Barron (1998) study the asymptotic property of model selection criteria related to the AIC and minimum description length principles in density estimation. Other works on model selection include the risk inflation criterion (Foster and George, 1994), generalized information criterion (Konishi and Kitagawa, 1996), Bayesian measure using deviance information criterion (Spiegelhalter *et al.*, 2002; Gelman *et al.*, 2004), model evaluation by using the absolute prediction error (Tian *et al.*, 2007), tuning parameter selection in penalization method (Wang, Li and Tsai, 2007), parametricness index (Liu and Yang, 2009), and many variations of the AIC and BIC (see, e.g., Bozdogan, 1987 and 2000).

A common strategy of parametric modeling is to choose *a priori* a family of distributions and then find a model in this family that best fits the data, often based on the maximum likelihood principle. It has been a common wisdom in statistics, however, that “all models are wrong, but some are more useful than others.” Thus, in the broad sense all models are misspecified. A direct generalization of the parametric modeling setting, which was targeted by AIC, BIC, and other model selection criteria, is to choose the best model among a set of families of models with different dimensional parameter spaces. In the linear model setting, for example, one is given a set of potential predictors but does not know which subset of the predictors are truly useful. Many approaches (see, e.g., Tibshirani, 1996; Fan and Li, 2001; Zou, 2006; Candes and Tao, 2007; Fan and Lv, 2008; Hall and Miller, 2009; Hall, Titterton and Xue, 2009; Lv and Fan, 2009) have been proposed in the literature

for variable selection. A common idea is to first construct a sequence of candidate linear models with different subsets of predictors and then choose the best one according to some model evaluation criterion such as AIC or BIC. However, even in this setting, the families of models may still not contain the true model, and neither AIC nor BIC explicitly account for such model misspecification. Explicitly estimating the discrepancy between the “best” fitted model in the “best” family and the true model can be potentially helpful.

In this paper, we derive asymptotic expansions of several model selection principles in misspecified generalized linear models (GLMs) (McCullagh and Nelder, 1989). The general idea applies to other model settings as well. The technical results lead us to propose the new semi-Bayesian information criteria (SIC). In particular, a specific form of SIC 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. Most conventional information criteria have been derived for independent and identically distributed (i.i.d.) observations. Our results can also be viewed as an extension of the classical approaches to non-i.i.d. settings.

The rest of the paper is organized as follows. In Section 2 we study the consistency and asymptotic normality of the quasi-maximum likelihood estimator in misspecified GLMs. We investigate the asymptotic expansions of the KL divergence principle and the Bayesian principle in Sections 3 and 4, respectively. In Section 5, we present the semi-Bayesian principles and the SIC methodology. We present several numerical examples to illustrate the finite sample performance of the proposed SIC methodology for model selection in misspecified models in Section 6. Section 7 provides some discussions of our results and their implications. All technical details are relegated to the Appendix. Some necessary formulas of SIC for three commonly used GLMs are also given in the Appendix.

2 Technical preparation

For completeness, we present here some asymptotic properties of the quasi-maximum likelihood estimator in misspecified GLMs with deterministic design matrices, which serve as the foundation for deriving asymptotic expansions of model selection principles in Sections 3–5. See White (1982) for a systematic treatment of the problem in the i.i.d. setting. Assume that the n -dimensional random vector of responses $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ has a true unknown distribution G_n with density function

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

where $\mathbf{y} = (y_1, \dots, y_n)^T$. 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.

2.1 The quasi-maximum likelihood estimator

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

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

where $\mathbf{z} = (z_1, \dots, z_n)^T$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T = \mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\beta} \in \mathbf{R}^d$, $b(\theta)$ is a convex function, μ_0 is the Lebesgue measure, and μ is some fixed measure on \mathbf{R} . Clearly $\{f_0(z, \theta) : \theta \in \mathbf{R}\}$ is a family of distributions in the regular exponential family and may not contain $g_{n,i}$'s. The following condition is common in the GLM setting.

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

For notational convenience, we define two vector-valued functions $\mathbf{b}(\boldsymbol{\theta}) = (b(\theta_1), \dots, b(\theta_n))^T$ and $\boldsymbol{\mu}(\boldsymbol{\theta}) = (b'(\theta_1), \dots, b'(\theta_n))^T$, and a matrix-valued function $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \text{diag}\{b''(\theta_1), \dots, b''(\theta_n)\}$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T$. For any n -dimensional random vector \mathbf{Z} with distribution $F_n(\cdot, \boldsymbol{\beta})$ given by (2), it holds that

$$E\mathbf{Z} = \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}) \quad \text{and} \quad \text{cov}(\mathbf{Z}) = \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta}). \quad (3)$$

The density function (2) can be rewritten as

$$f_n(\mathbf{z}, \boldsymbol{\beta}) = \exp[\mathbf{z}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta})] \prod_{i=1}^n \frac{d\mu}{d\mu_0}(z_i),$$

where $\frac{d\mu}{d\mu_0}$ denotes the Radon-Nikodym derivative. Given the observations \mathbf{y} and \mathbf{X} , this gives the quasi-log-likelihood function

$$\ell_n(\mathbf{y}, \boldsymbol{\beta}) = \log f_n(\mathbf{y}, \boldsymbol{\beta}) = \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta}) + \sum_{i=1}^n \log \frac{d\mu}{d\mu_0}(y_i). \quad (4)$$

The quasi-maximum likelihood estimator (QMLE) of the d -dimensional parameter vector $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}}_n = \arg \max_{\boldsymbol{\beta} \in \mathbf{R}^d} \ell_n(\mathbf{y}, \boldsymbol{\beta}). \quad (5)$$

The following proposition follows easily from a standard strict concavity argument.

Proposition 1. *Under Condition 1, the QMLE $\widehat{\boldsymbol{\beta}}_n$ is unique.*

Clearly the QMLE $\widehat{\beta}_n$ is the solution to the score equation

$$\Psi_n(\beta) \equiv \frac{\partial \ell_n(\mathbf{y}, \beta)}{\partial \beta} = \mathbf{X}^T [\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\beta)] = \mathbf{0}, \quad (6)$$

which becomes the normal equation $\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X} \beta$ in the linear regression model. The consistency and asymptotic normality of the maximum likelihood estimator (MLE) in correctly specified GLMs have been studied by Fahrmeir and Kaufmann (1985). A general theory of maximum likelihood estimation of misspecified models was presented by White (1982), who studied the case of i.i.d. 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 i.i.d.

2.2 Kullback-Leibler divergence and parameter identifiability

The KL divergence of density f from density g , which was introduced by Kullback and Leibler (1951), is defined as

$$I(g; f) = \int [\log g(u)] g(u) du - \int [\log f(u)] g(u) du,$$

which gives an upper bound on the squared Hellinger distance between distributions. The KL divergence of the posited GLM $F_n(\cdot, \beta)$ in (2) from the true model G_n is

$$\begin{aligned} I(g_n; f_n(\cdot, \beta)) &= \int [\log g_n(\mathbf{z})] g_n(\mathbf{z}) d\mathbf{z} - \int [\log f_n(\mathbf{z}, \beta)] g_n(\mathbf{z}) d\mathbf{z} \\ &= \sum_{i=1}^n \left\{ \int [\log g_{n,i}(z)] g_{n,i}(z) dz - \int [\log f_0(z, \theta_i)] g_{n,i}(z) dz \right\} = \sum_{i=1}^n I(g_{n,i}; f_0(\cdot, \theta_i)). \end{aligned} \quad (7)$$

Throughout the paper, the expectation and covariance are taken with respect to the true distribution G_n unless specified otherwise. We need the following basic regularity condition.

Condition 2. For each $1 \leq i \leq n$, $h_i(\theta) = E \log f_0(Y_i, \theta)$ is smooth in θ , and the differentiation and expectation are exchangeable so that $h'_i(\theta) = E \partial \log f_0(Y_i, \theta) / \partial \theta$ and $h''_i(\theta) = E \partial^2 \log f_0(Y_i, \theta) / \partial \theta^2$, which are also smooth in θ .

The following theorem shows that there is a unique distribution $F_n(\cdot, \beta_{n,0})$ in the family of misspecified GLMs in (2) that has the smallest KL divergence from the true model G_n .

Theorem 1. (Parameter identifiability). Assume that Conditions 1 and 2 hold. Then the KL divergence $I(g_n; f_n(\cdot, \beta))$ has a unique global minimum at $\beta_{n,0} \in \mathbf{R}^d$, which solves the equation

$$\mathbf{X}^T [E\mathbf{Y} - \boldsymbol{\mu}(\mathbf{X}\beta)] = \mathbf{0}. \quad (8)$$

This shows that the parameter identifiability problem has content in misspecified models. The uniqueness entails that $F_n(\cdot, \boldsymbol{\beta}_{n,0}) = G_n$ when the model is correctly specified, i.e., $G_n \in \{F_n(\cdot, \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbf{R}^d\}$. Since $\mathbf{X}^T \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0}) = \mathbf{X}^T E\mathbf{Y}$, we have

$$\text{cov}[\boldsymbol{\Psi}_n(\boldsymbol{\beta}_{n,0})] = \text{cov}(\mathbf{X}^T \mathbf{Y}) = \mathbf{X}^T \text{cov}(\mathbf{Y}) \mathbf{X} \equiv \mathbf{B}_n, \quad (9)$$

where $\text{cov}(\mathbf{Y}) = \text{diag}\{\text{var}(Y_1), \dots, \text{var}(Y_n)\}$ by the independence assumption. We introduce another matrix

$$\frac{\partial^2 I(g_n; f_n(\cdot, \boldsymbol{\beta}))}{\partial \boldsymbol{\beta}^2} = -\frac{\partial^2 \ell_n(\mathbf{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta}) \mathbf{X} \equiv \mathbf{A}_n(\boldsymbol{\beta}) \quad (10)$$

and define $\mathbf{A}_n = \mathbf{A}_n(\boldsymbol{\beta}_{n,0})$. In view of (3), $\mathbf{A}_n(\boldsymbol{\beta})$ is exactly the covariance matrix of $\mathbf{X}^T \mathbf{Y}$ when \mathbf{Y} has distribution $F_n(\cdot, \boldsymbol{\beta})$, and thus \mathbf{A}_n is the covariance matrix of $\mathbf{X}^T \mathbf{Y}$ under the best misspecified GLM $F_n(\cdot, \boldsymbol{\beta}_{n,0})$, whereas \mathbf{B}_n is the covariance matrix of $\mathbf{X}^T \mathbf{Y}$ under the true model G_n . It is known in classical maximum likelihood theory that $\mathbf{A}_n = \mathbf{B}_n$ when the model is correctly specified. These two matrices play a pivotal role in quasi-maximum likelihood estimation of misspecified models.

2.3 Consistency and asymptotic normality of QMLE

Early work on the consistency of estimators for the parameters of interest when the probability model is misspecified includes Berk (1966, 1970) and Huber (1967). Berk (1966, 1970) takes a Bayesian approach, and Huber (1967) provides conditions built on those of Wald (1949) for the consistency of maximum likelihood estimates. White (1982) uses the KL divergence as a measure of model misspecification and extends conditions on the asymptotic properties of maximum likelihood estimates given by LeCam (1953) to QMLE under misspecified models.

We need the following regularity conditions to establish the consistency of the QMLE $\hat{\boldsymbol{\beta}}_n$ in misspecified GLMs introduced in Section 2.1.

Condition 3. $\lambda_{\min}(\mathbf{B}_n) \rightarrow \infty$ as $n \rightarrow \infty$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue.

Condition 4. There exists some $c > 0$ such that for any $\delta > 0$, $\min_{\boldsymbol{\beta} \in N_n(\delta)} \lambda_{\min}[\mathbf{V}_n(\boldsymbol{\beta})] \geq c$ 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 $N_n(\delta) = \{\boldsymbol{\beta} \in \mathbf{R}^d : \|(n^{-1} \mathbf{B}_n)^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_{n,0})\| \leq n^{-1/2} \delta\}$ with $\|\cdot\|$ the Euclidean norm.

Condition 3 requires that the smallest eigenvalue of $\mathbf{X}^T \text{cov}(\mathbf{Y}) \mathbf{X}$ diverges as n increases. If $\text{var}(Y_i)$ are bounded away from 0 and ∞ , then it is equivalent to the usual assumption on the design matrix \mathbf{X} in linear regression for consistency, i.e., $\lambda_{\min}(\mathbf{X}^T \mathbf{X}) \rightarrow \infty$ as $n \rightarrow \infty$. Intuitively, Condition 4 means that the covariance structures given by the misspecified GLMs in a neighborhood of the best working model $F_n(\cdot, \boldsymbol{\beta}_{n,0})$ cannot be too far away 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 has to cover that of the target distribution. When the model is correctly specified, we have $\mathbf{V}_n(\boldsymbol{\beta}_{n,0}) = I_d$ since $\mathbf{A}_n = \mathbf{B}_n$ by the equivalence of the Hessian and outer product forms for the Fisher information matrix.

Theorem 2. (Consistency). *Under Conditions 1–4, the QMLE $\hat{\boldsymbol{\beta}}_n$ satisfies $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0} = o_P(1)$.*

To obtain the asymptotic normality of the QMLE $\hat{\boldsymbol{\beta}}_n$, we need two additional regularity conditions, where the first one is on the continuity of a matrix-valued function and the second one is a moment condition. For any $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d$, denote by $\tilde{\mathbf{A}}_n(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d)$ a $d \times d$ matrix with j -th row the corresponding row of $\mathbf{A}_n(\boldsymbol{\beta}_j)$ for each $j = 1, \dots, d$, and define matrix-valued function $\tilde{\mathbf{V}}_n(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d) = \mathbf{B}_n^{-1/2} \tilde{\mathbf{A}}_n(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d) \mathbf{B}_n^{-1/2}$.

Condition 5. *For any $\delta > 0$, $\max_{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d \in N_n(\delta)} \|\tilde{\mathbf{V}}_n(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d) - \mathbf{V}_n\| = o(1)$, 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 $\|\cdot\|$ denotes the matrix operator norm.*

Condition 6. $\max_{i=1}^n E|Y_i - EY_i|^3 = O(1)$ and $\sum_{i=1}^n (\mathbf{x}_i^T \mathbf{B}_n^{-1} \mathbf{x}_i)^{3/2} = o(1)$.

Theorem 3. (Asymptotic normality). *Under Conditions 1–6, the QMLE $\hat{\boldsymbol{\beta}}_n$ satisfies*

$$\mathbf{C}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_d),$$

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 2 and 3 become the conventional asymptotic theory of the MLE. The asymptotic normality of the QMLE $\hat{\boldsymbol{\beta}}_n$ in misspecified GLMs provides the theoretical foundation for proposing the new model selection methodology. 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, e.g., the technical analysis in Fan and Lv (2009) for penalized maximum likelihood estimator, where the dimensionality can grow non-polynomially with sample size.

Since the QMLE $\hat{\boldsymbol{\beta}}_n$ provides a consistent estimator of $\boldsymbol{\beta}_{n,0}$ in the best misspecified GLM $F_n(\cdot, \boldsymbol{\beta}_{n,0})$, natural estimates of matrices \mathbf{A}_n and \mathbf{B}_n are, respectively,

$$\hat{\mathbf{A}}_n = \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n) = \mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X} \hat{\boldsymbol{\beta}}_n) \mathbf{X} \quad (11)$$

and

$$\hat{\mathbf{B}}_n = \mathbf{X}^T \text{diag} \left\{ \left[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X} \hat{\boldsymbol{\beta}}_n) \right] \circ \left[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X} \hat{\boldsymbol{\beta}}_n) \right] \right\} \mathbf{X}, \quad (12)$$

where \circ denotes the Hadamard (componentwise) product. $\widehat{\mathbf{B}}_n$ is an unbiased estimator of \mathbf{B}_n when $E\mathbf{Y} = \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta}_{n,0})$. In practice, one can construct other estimates of \mathbf{B}_n by, e.g., using bootstrapping or treating the squared residual $[\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n)] \circ [\mathbf{y} - \boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n)]$ as a function of the corresponding fitted value $\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n)$, or some covariate, and doing some local smoothing. For simplicity, we use the estimate $\widehat{\mathbf{B}}_n$ defined above in all the numerical studies.

3 The Kullback-Leibler divergence principle and GAIC

After choosing a sequence of subsets $\{\mathfrak{M}_m : m = 1, \dots, M\}$ of the full model $\{1, \dots, p\}$, we can construct a sequence of QMLE's $\{\widehat{\boldsymbol{\beta}}_{n,m} : m = 1, \dots, M\}$ by fitting the GLM (2) as described in Section 2. A natural question is how to compare those fitted models. In this section we consider the KL divergence principle of model selection introduced by Akaike (1973, 1974), who studied the case of i.i.d. observations with correctly specified model.

3.1 Kullback-Leibler divergence principle of model selection

The QMLEs $\{\widehat{\boldsymbol{\beta}}_{n,m} : m = 1, \dots, M\}$ become the MLEs when the model is correctly specified. Akaike's principle of model selection is choosing the model \mathfrak{M}_{m_0} that minimizes the KL divergence $I(g_n; f_n(\cdot, \widehat{\boldsymbol{\beta}}_{n,m}))$ of the fitted model $F_n(\cdot, \widehat{\boldsymbol{\beta}}_{n,m})$ from the true model G_n , i.e.,

$$m_0 = \arg \min_{m \in \{1, \dots, M\}} I(g_n; f_n(\cdot, \widehat{\boldsymbol{\beta}}_{n,m})). \quad (13)$$

In view of (7), we have

$$I(g_n; f_n(\cdot, \widehat{\boldsymbol{\beta}}_{n,m})) = E \log g_n(\widetilde{\mathbf{Y}}) - \eta_n(\widehat{\boldsymbol{\beta}}_{n,m}), \quad (14)$$

where $\eta_n(\boldsymbol{\beta}) = E \ell_n(\widetilde{\mathbf{Y}}, \boldsymbol{\beta})$ and $\widetilde{\mathbf{Y}}$ is an independent copy of \mathbf{Y} . Thus

$$m_0 = \arg \max_{m \in \{1, \dots, M\}} \eta_n(\widehat{\boldsymbol{\beta}}_{n,m}) = \arg \max_{m \in \{1, \dots, M\}} E_{\widetilde{\mathbf{Y}}} \ell_n(\widetilde{\mathbf{Y}}, \widehat{\boldsymbol{\beta}}_{n,m}),$$

which shows that Akaike's principle of model selection is equivalent to choosing the model \mathfrak{M}_{m_0} that maximizes the expected log-likelihood with the expectation taken with respect to $\widetilde{\mathbf{Y}}$, an independent copy of \mathbf{Y} . Fix an arbitrary $\mathfrak{M} = \mathfrak{M}_m$ and $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\beta}}_{n,m}$. Using the asymptotic theory of MLE, Akaike (1973) shows for the case of i.i.d. observations that $\eta_n(\widehat{\boldsymbol{\beta}}_n)$ defined in (14) can be asymptotically expanded as $\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) - |\mathfrak{M}|$, which then leads to the seminal AIC for comparing competing models:

$$\text{AIC}(\mathbf{y}, \mathfrak{M}) = -2\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) + 2|\mathfrak{M}|. \quad (15)$$

The factor of 2 was historically included due to its connection with log-likelihood ratio tests.

3.2 GAIC in misspecified models

In this section, we derive a generalization of AIC in misspecified models. AIC has been studied by many researchers, mainly for the case of i.i.d observations. Takeuchi (1976) generalizes the derivation of AIC and obtains a term $\text{tr}(\mathbf{A}^{-1}\mathbf{B})$ 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 the cross-validation. In fact, $\text{tr}(\mathbf{A}^{-1}\mathbf{B})$ 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).

For simplicity, hereafter we drop the last term in the quasi-log-likelihood $\ell_n(\mathbf{y}, \boldsymbol{\beta})$ given by (4), which does not depend on $\boldsymbol{\beta}$, redefine it as

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

Fix any model \mathfrak{M}_m in the sequence of competing models. We drop the subscript m to simplify the presentation. In view of (14), for the purpose of model comparison, we need to estimate the quantity $\eta_n(\widehat{\boldsymbol{\beta}}_n)$. In practice, we have only a single sample \mathbf{y} . The maximum quasi-log-likelihood $\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n)$ tends to overestimate $\eta_n(\widehat{\boldsymbol{\beta}}_n)$ since we would use the same data twice, i.e., in both estimation and model evaluation. More specifically, $\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n)$ tends to overestimate $\ell_n(\mathbf{y}, \boldsymbol{\beta}_{n,0})$, which is an unbiased estimate of $\eta_n(\boldsymbol{\beta}_{n,0}) = E\ell_n(\widetilde{\mathbf{Y}}, \boldsymbol{\beta}_{n,0})$, and $\eta_n(\boldsymbol{\beta}_{n,0})$ tends to overestimate $\eta_n(\widehat{\boldsymbol{\beta}}_n)$ since $F_n(\cdot, \boldsymbol{\beta}_{n,0})$ is the best misspecified GLM under the KL divergence. We need the following regularity condition for deriving the asymptotic expansion of $E\eta_n(\widehat{\boldsymbol{\beta}}_n)$.

Condition 7. $\text{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n) = O(1)$, $\sup_{\boldsymbol{\delta}} \max_{j,k,l} |\partial^3 \eta_n(\boldsymbol{\beta}_{n,0} + n^{1/2}\mathbf{C}_n^{-1}\boldsymbol{\delta}) / \partial \delta_j \partial \delta_k \partial \delta_l| = o(n^{3/2})$, and for some $\delta > 0$, $E\|\mathbf{C}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})\|^{3+\delta} = O(1)$, where $\mathbf{C}_n = \mathbf{B}_n^{-1/2}\mathbf{A}_n$ and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_d)^T$.

Theorem 4. Under Conditions 1–7, we have

$$E\eta_n(\widehat{\boldsymbol{\beta}}_n) = E\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) - \text{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n) + o(1), \quad (16)$$

where both expectations are taken with respect to the true distribution G_n .

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

Definition 1. The GAIC of the competing model \mathfrak{M} is

$$\text{GAIC}(\mathbf{y}, \mathfrak{M}; F_n) = -2\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) + 2\text{tr}(\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n), \quad (17)$$

where $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ are estimates of \mathbf{A}_n and \mathbf{B}_n given in Section 2.3.

GAIC shares the common feature of generalizations of AIC. It incorporates the effects of model complexity and model misspecification in a single term. Since $\text{tr}(\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n) = \text{tr}(\widehat{\mathbf{B}}_n^{1/2}\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n^{1/2}) \geq 0$, the term added to the negative maximum quasi-log-likelihood in GAIC is indeed a penalty term. When the model is correctly specified, $\text{tr}(\mathbf{A}_n^{-1}\mathbf{B}_n) = \text{tr}(I_d) = d = |\mathfrak{M}|$ since $\mathbf{A}_n = \mathbf{B}_n$, which is the number of predictors that drive the GLM (2). Thus we would expect $\text{tr}(\widehat{\mathbf{A}}_n^{-1}\widehat{\mathbf{B}}_n) \approx \text{tr}(I_d) = |\mathfrak{M}|$, which reflects the penalty on model complexity as that in AIC. As mentioned before, many extensions of AIC under different settings share the same form as in (17). 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 the cross-validation uses the log-density assessment.

4 The Bayesian principle and GBIC

4.1 BIC and KL divergence

Suppose we have a set of competing models: $\{\mathfrak{M}_m : m = 1, \dots, M\}$. A popular Bayesian model selection procedure is to first put nonzero prior probability $\alpha_{\mathfrak{M}_m}$ on each model \mathfrak{M}_m , and then give 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 $\mathbf{R}^{\mathfrak{M}_m} = \mathbf{R}^{d_m}$ and locally bounded away from zero throughout the domain. The Bayesian principle of model selection is to choose the most probable model *a posteriori*. That is, to choose model \mathfrak{M}_{m_0} such that

$$m_0 = \arg \max_{m \in \{1, \dots, M\}} S(\mathbf{y}, \mathfrak{M}_m; F_n), \quad (18)$$

where

$$S(\mathbf{y}, \mathfrak{M}_m; F_n) = \log \int \alpha_{\mathfrak{M}_m} \exp[\ell_n(\mathbf{y}, \boldsymbol{\beta})] d\mu_{\mathfrak{M}_m}(\boldsymbol{\beta}) \quad (19)$$

with the log-likelihood $\ell_n(\mathbf{y}, \boldsymbol{\beta})$ as in (4) and the integral over \mathbf{R}^{d_m} . Fix an arbitrary $\mathfrak{M} = \mathfrak{M}_m$ and $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\beta}}_{n,m}$. Using the asymptotic theory of MLE, Schwartz (1978) shows for the case of i.i.d. observations with correctly specified model in a regular exponential family that $S(\mathbf{y}, \mathfrak{M}; F_n) \approx \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) - \frac{\log n}{2}|\mathfrak{M}| + \log \alpha_{\mathfrak{M}}$, where $\widehat{\boldsymbol{\beta}}_n$ is the MLE of $\boldsymbol{\beta}$. This expression allows Schwartz to introduce the seminal BIC for model selection:

$$\text{BIC}(\mathbf{y}, \mathfrak{M}) = -2\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) + (\log n)|\mathfrak{M}|, \quad (20)$$

where a factor of 2 is included to make it consistent with AIC in (15). Schwartz's original arguments were generalized later by Cavanaugh and Neath (1999) via weakening the assumptions.

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

$$\frac{d\nu_m}{d\mu_0}(\mathbf{z}) = \int \exp[\ell_n(\mathbf{z}, \boldsymbol{\beta})] d\mu_{\mathfrak{M}_m}(\boldsymbol{\beta}), \quad \mathbf{z} \in \mathbf{R}^n, \quad (21)$$

where μ_0 is the Lebesgue measure. Similar to (7), the KL divergence of ν_m from the true model G_n is

$$\begin{aligned} I\left(g_n; \frac{d\nu_m}{d\mu_0}\right) &= \int [\log g_n(\mathbf{z})] g_n(\mathbf{z}) d\mathbf{z} - \int \left[\log \frac{d\nu_m}{d\mu_0}(\mathbf{z}) \right] g_n(\mathbf{z}) d\mathbf{z} \\ &= E \log g_n(\mathbf{Y}) - E \log \int \exp[\ell_n(\mathbf{Y}, \boldsymbol{\beta})] d\mu_{\mathfrak{M}_m}(\boldsymbol{\beta}), \end{aligned} \quad (22)$$

where leads to the following proposition.

Proposition 2. a) For each $m \in \{1, \dots, M\}$, we have

$$ES(\mathbf{Y}, \mathfrak{M}_m; F_n) = -I\left(g_n; \frac{d\nu_m}{d\mu_0}\right) + \log \alpha_{\mathfrak{M}_m} + E \log g_n(\mathbf{Y}), \quad (23)$$

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

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

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

Proposition 2 holds regardless whether or not the true model is in the set of candidate models. We see also that for each $m \in \{1, \dots, M\}$, $-S(\mathbf{y}, \mathfrak{M}_m; F_n) + \log \alpha_{\mathfrak{M}_m}$ gives, up to a common additive constant, an unbiased estimate of $I\left(g_n; \frac{d\nu_m}{d\mu_0}\right)$. We also see 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.

4.2 GBIC in misspecified models

In this section, we derive the asymptotic expansion of the log-marginal likelihood $S(\mathbf{y}, \mathfrak{M}; F_n)$ in misspecified models. For notational convenience, we drop the subscript m . In the GLM (2), the parameter vector $\boldsymbol{\theta} \in \mathbf{R}^n$ for the response vector \mathbf{Y} is given by $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$. Under this model and some regularity conditions, as in Schwartz (1978) we can apply the classical Laplace approximation (i.e., Taylor-expanding the exponent to the second order and using the Gaussian integration to obtain necessary terms) to get that

$$\begin{aligned} S(\mathbf{y}, \mathfrak{M}; F_n) &= \ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + \log \int e^{-(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)^T \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n) (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)/2} d\boldsymbol{\beta} + C'_n \\ &= \ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \frac{\log n}{2} |\mathfrak{M}| - \frac{1}{2} \log |n^{-1} \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n)| + C_n, \end{aligned} \quad (24)$$

where $\widehat{\boldsymbol{\beta}}_n$ is the maximizer of $\ell_n(\mathbf{y}, \boldsymbol{\beta})$, $\mathbf{A}_n(\boldsymbol{\beta}) = -\partial^2 \ell_n(\mathbf{y}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^2 = \mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})\mathbf{X}$, and C'_n and C_n are both bounded in n . However, when the model is misspecified, it is no longer clear whether the C_n term is still bounded.

Note that the term $-\frac{1}{2} \log |n^{-1} \mathbf{A}_n(\widehat{\boldsymbol{\beta}}_n)|$ in the asymptotic expansion (24) depends on the scale of the design matrix \mathbf{X} . Thus this second order term is not very meaningful. In our derivation below, we correct this by rescaling the sequence of design matrices $\{\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T : n = 1, 2, \dots\}$ before we do the Bayesian inference. Let \mathbf{P}_n be a $d \times d$ nonsingular matrix. We apply a linear transformation to the design matrix, $\mathbf{X} \rightarrow \widetilde{\mathbf{X}} \equiv \mathbf{X}\mathbf{P}_n$. Although there are infinitely many choices of \mathbf{P}_n , we consider one such that $n^{-1/2} \widetilde{\mathbf{X}}^T \mathbf{Y}$ has the identity covariance matrix I_d under the true model G_n . Thus

$$I_d = \text{cov} \left(n^{-1/2} \widetilde{\mathbf{X}}^T \mathbf{Y} \right) = \mathbf{P}_n^T \text{cov} \left(n^{-1/2} \mathbf{X}^T \mathbf{Y} \right) \mathbf{P}_n = \mathbf{P}_n^T (n^{-1} \mathbf{B}_n) \mathbf{P}_n, \quad (25)$$

where $\mathbf{B}_n = \mathbf{X}^T \text{cov}(\mathbf{Y})\mathbf{X}$ is the covariance matrix of $\mathbf{X}^T \mathbf{Y}$ under the true model G_n . It is known that any solution \mathbf{P}_n to equation (25) admits a decomposition $\mathbf{P}_n = (n^{-1} \mathbf{B}_n)^{-1/2} \mathbf{Q}$, where \mathbf{Q} is some $d \times d$ orthogonal matrix. Thus, our choice of \mathbf{P}_n is unique up to an orthogonal matrix. As seen later, this is sufficient for deriving the asymptotic expansion of $S(\mathbf{y}, \mathfrak{M}; F_n)$ since only the determinant of \mathbf{P}_n comes into play. For simplicity, we take $\mathbf{P}_n = (n^{-1} \mathbf{B}_n)^{-1/2}$, whose inverse appeared in $N_n(\delta)$ in Condition 4 for establishing the consistency and asymptotic normality of the QMLE $\widehat{\boldsymbol{\beta}}_n$. In the specific case of linear regression with error variance τ^2 , \mathbf{P}_n becomes $\tau^{-1} (n^{-1} \mathbf{X}^T \mathbf{X})^{-1/2}$.

Let $\{\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{P}_n = \mathbf{X}(n^{-1} \mathbf{B}_n)^{-1/2} : n = 1, 2, \dots\}$ be the sequence of aligned design matrices. Then we have the parameter space $\{\widetilde{\boldsymbol{\beta}} : \widetilde{\boldsymbol{\beta}} \in \mathbf{R}^d\}$ and thus $\boldsymbol{\beta} = \mathbf{P}_n \widetilde{\boldsymbol{\beta}} = (n^{-1} \mathbf{B}_n)^{-1/2} \widetilde{\boldsymbol{\beta}}$. Hereafter the prior distribution $\mu_{\mathfrak{M}}$ is understood on the parameter $\widetilde{\boldsymbol{\beta}}$. Let μ_0 be the Lebesgue measure on \mathbf{R}^d , and $\pi(\widetilde{\boldsymbol{\beta}}) = \frac{d\mu_{\mathfrak{M}}}{d\mu_0}(\widetilde{\boldsymbol{\beta}})$ the prior density on $\widetilde{\boldsymbol{\beta}}$ for model \mathfrak{M} . Define $\boldsymbol{\delta} = (n^{-1} \mathbf{B}_n)^{1/2} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n)$. Then with a change of the variable $\widetilde{\boldsymbol{\beta}} = (n^{-1} \mathbf{B}_n)^{1/2} \boldsymbol{\beta} = \boldsymbol{\delta} + (n^{-1} \mathbf{B}_n)^{1/2} \widehat{\boldsymbol{\beta}}_n$, the prior density of $\mu_{\mathfrak{M}}$ can be written in terms of $\boldsymbol{\delta}$ as

$$d\mu_{\mathfrak{M}}(\widetilde{\boldsymbol{\beta}}) = \pi(\widetilde{\boldsymbol{\beta}}) d\widetilde{\boldsymbol{\beta}} = \pi \left(\boldsymbol{\delta} + n^{-1/2} \mathbf{B}_n^{1/2} \widehat{\boldsymbol{\beta}}_n \right) d\boldsymbol{\delta}. \quad (26)$$

For any $\boldsymbol{\delta} \in \mathbf{R}^d$, we define

$$\ell_n^*(\mathbf{y}, \boldsymbol{\delta}) = \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n + n^{1/2} \mathbf{B}_n^{-1/2} \boldsymbol{\delta}) - \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n), \quad (27)$$

which is the deviation of the quasi-log-likelihood from its maximum. It follows from (19), (26) and (27) that

$$S(\mathbf{y}, \mathfrak{M}; F_n) = \log \alpha_{\mathfrak{M}} + \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) + \log E_{\mu_{\mathfrak{M}}}[U_n(\boldsymbol{\delta})^n], \quad (28)$$

where $U_n(\boldsymbol{\delta}) = \exp[n^{-1} \ell_n^*(\mathbf{y}, \boldsymbol{\delta})]$. We need the following two regularity conditions to derive the asymptotic expansion of the above term $\log E_{\mu_{\mathfrak{M}}}[U_n(\boldsymbol{\delta})^n]$.

Condition 8. *There exist some $r_0, c_1, c_2 > 0$ such that the prior density $\pi(\tilde{\boldsymbol{\beta}})$ satisfies*

$$\inf_{\|\boldsymbol{\delta}\| \leq r_0} \pi\left(\boldsymbol{\delta} + n^{-1/2} \mathbf{B}_n^{1/2} \hat{\boldsymbol{\beta}}_n\right) \geq c_1 \quad \text{and} \quad \sup_{\boldsymbol{\delta} \in \mathbf{R}^d} \pi\left(\boldsymbol{\delta} + n^{-1/2} \mathbf{B}_n^{1/2} \hat{\boldsymbol{\beta}}_n\right) \leq c_2.$$

Moreover, $\|n^{-1/2} \mathbf{B}_n^{1/2} \hat{\boldsymbol{\beta}}_n\|$ is bounded away from 0 and ∞ .

Condition 9. $\lambda_{\max}(\mathbf{D}_n) = O(1)$, and there exists a sequence of $r_n > 0$ with $r_n = o(1)$ such that $\rho_n(r_n) \leq c_3 \lambda_{\min}(\mathbf{D}_n)$ and $\lambda_{\min}(\mathbf{D}_n)^{-1} r_n^{-2} = O(n^\alpha)$ for some $c_3, \alpha \in [0, 1)$, where $\mathbf{D}_n = \mathbf{B}_n^{-1/2} \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n) \mathbf{B}_n^{-1/2}$, $\rho_n(r) = \max_{\|\boldsymbol{\delta}\| \leq r} \max\{|\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\delta})]|, |\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\delta})]|\}$ with $\mathbf{F}_n(\boldsymbol{\delta}) = \mathbf{B}_n^{-1/2} [\mathbf{A}_n(\hat{\boldsymbol{\beta}}_n + \mathbf{B}_n^{-1/2} \boldsymbol{\delta}) - \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n)] \mathbf{B}_n^{-1/2}$, and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue.

In view of (26), Condition 8 requires that the prior density for $\tilde{\boldsymbol{\beta}}$ is bounded and locally bounded away from zero. Since the relative scale of $\tilde{\boldsymbol{\beta}}$ compared with the original parameter $\boldsymbol{\beta}$ is given by $\tilde{\boldsymbol{\beta}} = n^{-1/2} \mathbf{B}_n^{1/2} \boldsymbol{\beta}$, the condition that $\|n^{-1/2} \mathbf{B}_n^{1/2} \hat{\boldsymbol{\beta}}_n\| = [n^{-1} (\mathbf{X} \hat{\boldsymbol{\beta}}_n)^T \text{cov}(\mathbf{Y}) (\mathbf{X} \hat{\boldsymbol{\beta}}_n)]^{1/2}$ is bounded away from 0 and ∞ justifies the use of prior on parameter $\tilde{\boldsymbol{\beta}}$. The technical proof applies as well to the case when $\lambda_{\max}(\mathbf{D}_n)$ diverges as $n \rightarrow \infty$. Conditions 8 and 9 are sensible by noting that the QMLE $\hat{\boldsymbol{\beta}}_n$ is shown to have asymptotic normality under some regularity conditions in Theorem 3.

Theorem 5. *Under Conditions 1, 8 and 9, we have*

$$S(\mathbf{y}, \mathfrak{M}; F_n) = \ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \frac{\log n}{2} |\mathfrak{M}| + \frac{1}{2} \log |\hat{\mathbf{A}}_n^{-1} \mathbf{B}_n| + \log \alpha_{\mathfrak{M}} + R(\mathbf{y}, \mathfrak{M}; F_n), \quad (29)$$

where $\hat{\mathbf{A}}_n = \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n)$ and the remainder $R(\mathbf{y}, \mathfrak{M}; F_n)$ is bounded in n .

In Condition 9, $\log |\hat{\mathbf{A}}_n^{-1} \mathbf{B}_n| = \log |\mathbf{B}_n^{1/2} \hat{\mathbf{A}}_n^{-1} \mathbf{B}_n^{1/2}|$ is allowed to diverge as $n \rightarrow \infty$ since it is only assumed that $\lambda_{\max}(\mathbf{B}_n^{1/2} \hat{\mathbf{A}}_n^{-1} \mathbf{B}_n^{1/2}) = \lambda_{\min}(\mathbf{D}_n)^{-1} = O(n^\alpha r_n^2)$, where $\alpha \in [0, 1)$. As mentioned before, $\lambda_{\min}(\mathbf{B}_n^{1/2} \hat{\mathbf{A}}_n^{-1} \mathbf{B}_n^{1/2}) = \lambda_{\max}(\mathbf{D}_n)^{-1}$ can also be allowed to converge to zero as $n \rightarrow \infty$. Following the asymptotic expansion in the above theorem, we introduce the generalized BIC (GBIC) as follows.

Definition 2. *We define GBIC of the competing model \mathfrak{M} by*

$$\text{GBIC}(\mathbf{y}, \mathfrak{M}; F_n) = -2\ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) + (\log n) |\mathfrak{M}| - \log |\hat{\mathbf{A}}_n^{-1} \hat{\mathbf{B}}_n|, \quad (30)$$

where $\hat{\mathbf{A}}_n$ and $\hat{\mathbf{B}}_n$ are estimates of \mathbf{A}_n and \mathbf{B}_n given in Section 2.3.

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. However, generally the third term $-\log |\hat{\mathbf{A}}_n^{-1} \hat{\mathbf{B}}_n|$ in GBIC is not always nonnegative. 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 $\log |\hat{\mathbf{A}}_n^{-1} \hat{\mathbf{B}}_n| \approx \log |I_d| = 0$. In this sense, GBIC introduced above generalizes BIC.

5 The semi-Bayesian principles and SIC methodology

In this section we introduce a family of semi-Bayesian principles for model selection in misspecified models that bridge the two well-known principles: the KL divergence principle and the Bayesian principle for model selection, which have been considered in Sections 3 and 4. The semi-Bayesian principles give the new SIC methodology. Assume that we have a sequence of competing models indexed by subsets $\{\mathfrak{M}_m : m = 1, \dots, M\}$ of the full model $\{1, \dots, p\}$ of all the p covariates. We can construct a sequence of QMLE's $\{\widehat{\beta}_{n,m} : m = 1, \dots, M\}$ as in Section 3, and fit the GLM (2) using a Bayesian approach as in Section 4.

5.1 Semi-Bayesian principles of model selection

We define the semi-Bayesian principles of model selection as follows.

Definition 3. *The semi-Bayesian principle of model selection with index $\gamma \in [0, 1]$ is choosing the model \mathfrak{M}_{m_0} that minimizes*

$$D_\gamma(\widehat{\beta}_{n,m}; \mathfrak{M}_m) = (1 - \gamma)I(g_n; f_n(\cdot, \widehat{\beta}_{n,m})) - \gamma_* I(g_n; f_n(\cdot, \beta_{n,m,0})) + \gamma I\left(g_n; \frac{d\nu_m}{d\mu_0}\right), \quad (31)$$

where $\gamma_* = \gamma \wedge (1 - \gamma)$, $I(g_n; f_n(\cdot, \widehat{\beta}_{n,m}))$ is defined in (14), $I(g_n; f_n(\cdot, \beta_{n,m,0}))$ is the minimum KL divergence of the misspecified GLM $F_n(\cdot, \beta)$ from the true model G_n over $\beta \in \mathbf{R}^{|\mathfrak{M}_m|}$, and $I(g_n; \frac{d\nu_m}{d\mu_0})$ is defined in (22) with ν_m given by (21) the marginal distribution of the response vector \mathbf{Y} conditional on model \mathfrak{M}_m . That is,

$$m_0 = \arg \min_{m \in \{1, \dots, M\}} D_\gamma(\widehat{\beta}_{n,m}; \mathfrak{M}_m). \quad (32)$$

We see that the semi-Bayesian principle of model selection with index $\gamma = 0$ becomes the KL divergence principle of model selection, and that with index $\gamma = 1$ is equivalent to the Bayesian principle of model selection by its KL divergence interpretation given in Proposition 2. It is easy to see that $D_\gamma(\widehat{\beta}_{n,m}; \mathfrak{M}_m) \geq 0$ for any $\gamma \in [0, 1]$ since $I(g_n; f_n(\cdot, \widehat{\beta}_{n,m})) \geq I(g_n; f_n(\cdot, \beta_{n,m,0}))$ always holds. In particular, for the semi-Bayesian principle of model selection with index $\gamma = 1/2$, $2D_\gamma(\widehat{\beta}_{n,m}; \mathfrak{M}_m)$ is the sum of the excess KL divergence of the model relative to the minimum one and the KL divergence of the marginal distribution of the response from the true one.

The intuition is that when the model is misspecified, even the best marginal distribution of \mathbf{Y} can be far away from the true distribution G_n . Thus finding a model that compromises the terms in (31) can be desirable, which combines the strengths of the two well-known principles. For simplicity we drop the subscript m . The following theorem, which is proved using the theory established in Sections 2–4, gives the asymptotic expansion of $ED_\gamma(\widehat{\beta}_n; \mathfrak{M})$.

Theorem 6. *Under Conditions 1–9, we have*

$$ED_\gamma(\widehat{\boldsymbol{\beta}}_n; \mathfrak{M}) = -\gamma^* E\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) + \frac{\gamma \log n}{2} |\mathfrak{M}| - \frac{\gamma}{2} E \log |\widehat{\mathbf{A}}_n^{-1} \mathbf{B}_n| + \frac{\gamma^{**}}{2} \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) \\ + \gamma^* E \log g_n(\mathbf{y}) - \gamma ER(\mathbf{y}, \mathfrak{M}; F_n) + o(1), \quad (33)$$

where all expectations are taken with respect to the true distribution G_n , $\gamma^* = \gamma \vee (1 - \gamma)$, $\gamma^{**} = (2 - 3\gamma) \vee (1 - \gamma)$, $\widehat{\mathbf{A}}_n = \mathbf{A}_n(\widehat{\boldsymbol{\beta}}_n)$, and $R(\mathbf{y}, \mathfrak{M}; F_n)$ is given in (29).

5.2 The SIC methodology

The asymptotic expansions of the semi-Bayesian principles in misspecified GLMs give a family of semi-Bayesian information criteria (SIC). Following Theorem 6, we define the SIC methodology as follows.

Definition 4. *We define SIC with index $\gamma \in [0, 1]$ of the competing model \mathfrak{M} by*

$$\text{SIC}_\gamma(\mathbf{y}, \mathfrak{M}; F_n) = -2\gamma^* \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) + \gamma (\log n) |\mathfrak{M}| - \gamma \log |\widehat{\mathbf{H}}_n| + \gamma^{**} \text{tr}(\widehat{\mathbf{H}}_n), \quad (34)$$

where $\widehat{\mathbf{H}}_n = \widehat{\mathbf{A}}_n^{-1} \widehat{\mathbf{B}}_n$ with $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ estimates of \mathbf{A}_n and \mathbf{B}_n given in Section 2.3.

We see that SIC with index $\gamma = 0$ becomes GAIC defined in (17), and SIC with index $\gamma = 1$ becomes GBIC defined in (30). In particular, the semi-Bayesian principle of model selection with index $\gamma = 1/2$ gives SIC with index $\gamma = 1/2$. We show next that this specific form of SIC in misspecified models admits a natural decomposition of the form

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

where the first term is the negative maximum quasi-log-likelihood, and penalties on model complexity and model misspecification are both nonnegative. As mentioned before, neither GAIC nor GBIC has such a decomposition.

Proposition 3. *For any $d \times d$ symmetric positive definite matrices $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ given in (34), we have for $\gamma = 1/2$,*

$$\text{SIC}_{1/2}(\mathbf{y}, \mathfrak{M}; F_n) = -\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) + \frac{1 + \log n}{2} |\mathfrak{M}| + I \left[N(\mathbf{0}, \widehat{\mathbf{B}}_n); N(\mathbf{0}, \widehat{\mathbf{A}}_n) \right], \quad (35)$$

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

The penalization on model misspecification, i.e., the third term, in SIC with index $\gamma = 1/2$ has a natural interpretation. Under the true model G_n , we have

$$E(\mathbf{X}^T \mathbf{Y}) = \mathbf{X}^T E\mathbf{Y} \quad \text{and} \quad \text{cov}(\mathbf{X}^T \mathbf{Y}) = \mathbf{X}^T \text{cov}(\mathbf{Y}) \mathbf{X} = \mathbf{B}_n.$$

Under some regularity conditions, we can show that due to the central limit theorem the asymptotic distribution of $\mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu})$ is close to $N(\mathbf{0}, \mathbf{B}_n)$, where $\boldsymbol{\mu} = E\mathbf{Y}$. On the other hand, if we misspecify the true model as the GLM $F_n(\cdot, \boldsymbol{\beta})$, the QMLE $\widehat{\boldsymbol{\beta}}_n$ converges to $\boldsymbol{\beta}_{n,0}$ so that $F_n(\cdot, \boldsymbol{\beta}_{n,0})$ is the best model in the posited GLM family for approximating the true model G_n under the KL divergence. It follows from (3) and (8) that

$$E_{F_n}(\mathbf{X}^T \mathbf{Y}) = \mathbf{X}^T E_{F_n} \mathbf{Y} = \mathbf{X}^T \boldsymbol{\mu} \quad \text{and} \quad \text{cov}_{F_n}(\mathbf{X}^T \mathbf{Y}) = \mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X} \boldsymbol{\beta}_{n,0}) \mathbf{X} = \mathbf{A}_n,$$

where $F_n = F_n(\cdot, \boldsymbol{\beta}_{n,0})$ and $\boldsymbol{\mu} = E_{G_n} \mathbf{Y}$. We can again show that under some regularity conditions, the asymptotic distribution of $\mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu})$ is now close to $N(\mathbf{0}, \mathbf{A}_n)$. Thus the KL divergence of $N(\mathbf{0}, \mathbf{A}_n)$ from $N(\mathbf{0}, \mathbf{B}_n)$ gives a natural measure of model misspecification. This gives another justification for the penalization term on model misspecification in SIC with index $\gamma = 1/2$. When the model is correctly specified, we would expect $\widehat{\mathbf{A}}_n \approx \widehat{\mathbf{B}}_n$ since $\mathbf{A}_n = \mathbf{B}_n$ and, thus, $I[N(\mathbf{0}, \widehat{\mathbf{B}}_n); N(\mathbf{0}, \widehat{\mathbf{A}}_n)] \approx 0$. For simplicity, SIC in all the numerical studies implicitly refers to the specific form of SIC with index $\gamma = 1/2$.

6 Numerical examples

6.1 Polynomial regression

We simulated 100 independent datasets from the cubic polynomial model

$$y = 1 + 5x - 1.25x^2 + 0.55x^3 + \varepsilon,$$

where $x \sim N(0, 1)$, $\varepsilon \sim N(0, \sigma^2)$, and ε is independent of x . Each dataset contains n i.i.d. samples, where $n = 25, 50, 200$ and $\sigma = 0.5, 1, 2$, respectively. For each dataset, we fit the polynomial regression model of order from 1 up to 6 and used AIC, BIC, GAIC, GBIC, and SIC to select the order of polynomial regression. Table 1 summarizes the comparison results of the frequency of estimated order of polynomial regression model. As expected, AIC tended to select larger model than the true one. Interestingly, GAIC performed no better than AIC. BIC performed well when the sample size is large but suffered from smaller sample sizes. Almost all times both SIC and GBIC outperformed the other information criteria, while SIC performed the best by a small margin.

6.2 Best subset linear regression

We are curious to see how the new information criteria fare in the linear regression model selection scenarios. We simulated 100 datasets from the linear model

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \tag{36}$$

Table 1: Frequency of estimated order of polynomial regression model over 100 simulations for different n and σ

σ	Criterion	Model size ($n = 25$)					Model size ($n = 50$)					Model size ($n = 200$)				
		2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
0.5	AIC	0	68	13	8	11	0	72	16	6	6	0	73	17	3	7
	GAIC	0	34	20	19	27	0	55	18	10	17	0	35	19	22	24
	BIC	0	85	5	5	5	0	94	5	1	0	0	94	6	0	0
	GBIC	0	100	0	0	0	0	99	1	0	0	0	98	2	0	0
	SIC	0	100	0	0	0	0	99	1	0	0	0	99	1	0	0
1	AIC	0	72	15	8	5	0	72	13	9	6	0	80	11	5	4
	GAIC	0	37	25	15	23	0	55	14	17	14	0	47	19	13	21
	BIC	0	87	8	5	0	0	94	6	0	0	0	100	0	0	0
	GBIC	0	99	1	0	0	0	100	0	0	0	0	100	0	0	0
	SIC	0	99	1	0	0	0	100	0	0	0	0	100	0	0	0
2	AIC	1	75	10	10	4	0	81	5	8	6	0	79	13	4	4
	GAIC	0	34	20	20	26	0	56	11	10	23	0	46	21	14	19
	BIC	2	90	3	3	2	0	90	5	4	1	0	99	0	0	1
	GBIC	15	85	0	0	0	1	96	3	0	0	0	100	0	0	0
	SIC	8	92	0	0	0	0	97	3	0	0	0	100	0	0	0

where \mathbf{X} is $n \times p$ design matrix and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ independent of \mathbf{X} . For each simulated dataset, the rows of \mathbf{X} were sampled as i.i.d. copies from $N(\mathbf{0}, \Sigma_0)$ with $\Sigma_0 = (0.5^{|i-j|})_{i,j=1,\dots,p}$. We fixed $p = 6$ and $\boldsymbol{\beta}_0 = (1, -1.25, 0.75, 0, 0, 0)^T$ and considered $n = 50$ and 200 and $\sigma = 0.5$ and 1 , respectively. For each dataset, we applied the best subset regression and used AIC, BIC, GAIC, GBIC, and SIC to select the model size. Table 2 summarizes the comparison results of the frequency of estimated model size. Again AIC tended to select larger model than the true one, and it is surprising that GAIC performed no better than AIC. The performance of BIC deteriorated when the sample size is not large. Both SIC and GBIC outperformed the other information criteria, while SIC performed the best. The inclusion of second order term in both SIC and GBIC partially explains their good performance in small samples.

6.3 Linear regression with interaction

Here we test how each information criterion behaves when the true model is not among the family of candidate models for selection. We simulated 100 datasets from the following model, which adds an interaction term to the linear model (36) in Section 6.2:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + 0.5\mathbf{x}_{p+1} + \boldsymbol{\varepsilon},$$

Table 2: Frequency of estimated model size by best subset regression over 100 simulations for different n and σ

σ	Criterion	Model size ($n = 50$)					Model size ($n = 200$)				
		2	3	4	5	6	2	3	4	5	6
0.5	AIC	0	64	24	11	1	0	63	27	10	0
	GAIC	0	49	34	15	2	0	59	28	12	1
	BIC	0	87	11	2	0	0	97	2	1	0
	GBIC	0	92	6	2	0	0	97	2	1	0
	SIC	0	97	3	0	0	0	100	0	0	0
1	AIC	0	59	28	10	3	0	59	31	10	0
	GAIC	0	45	37	13	5	0	57	29	14	0
	BIC	0	89	11	0	0	0	95	5	0	0
	GBIC	0	90	10	0	0	0	95	5	0	0
	SIC	1	95	4	0	0	0	97	3	0	0

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ is $n \times p$ design matrix with $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$,

$$\mathbf{x}_{p+1} = (x_{1,p+1}, \dots, x_{n,p+1})^T \quad \text{with } x_{i,p+1} = x_{i1} \cdot x_{i2},$$

and $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ independent of \mathbf{X} . We kept the rest of the setting in Section 6.2 except that $\Sigma_0 = I_p$ and $n = 50, 200, 1000$, and pretended that the data were generated from model (36) with predictors X_1, \dots, X_6 . So the family of our working models does not include the true model due to the presence of the interaction term, 0.5 times the product of the first two predictors. For each dataset, we applied the best subset regression and used AIC, BIC, GAIC, GBIC, and SIC to select the model size. Table 3 summarizes the comparison results of the frequency of estimated model size. The conclusions are similar to those in Section 6.2. It is interesting to note that since the interaction term $X_1 X_2$ is uncorrelated with all the candidate variables X_1, \dots, X_6 , the ‘‘correct’’ number of predictors for this misspecified model should be 3.

6.4 Nonlinear regression

Here we simulate a single-index model, i.e., Y is dependent on a linear combination of \mathbf{x} through a nonlinear link function. Let

$$f(z) = \frac{z^2}{a+z} \quad \text{and} \quad \mathbf{f}(\mathbf{z}) = (f(z_1), \dots, f(z_n))^T, \quad \mathbf{z} = (z_1, \dots, z_n)^T. \quad (37)$$

We simulated 100 datasets from nonlinear regression model

$$\mathbf{y} = \mathbf{f}(\mathbf{X}\boldsymbol{\beta}_0) + \varepsilon,$$

Table 3: Frequency of estimated model size by best subset regression over 100 simulations for different n and σ

σ	Criterion	Model size ($n = 50$)					Model size ($n = 200$)					Model size ($n = 1000$)				
		2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
0.5	AIC	0	64	30	6	0	0	15	63	20	2	0	58	38	4	0
	GAIC	0	54	29	16	1	0	14	68	15	3	0	58	37	5	0
	BIC	0	88	12	0	0	0	61	39	0	0	0	99	1	0	0
	GBIC	0	90	10	0	0	0	65	35	0	0	0	99	1	0	0
	SIC	0	93	7	0	0	0	73	27	0	0	0	100	0	0	0
1	AIC	0	61	32	6	1	0	43	42	13	2	0	57	40	3	0
	GAIC	0	49	38	9	4	0	42	43	13	2	0	57	39	4	0
	BIC	0	85	14	1	0	0	87	12	1	0	0	99	1	0	0
	GBIC	0	87	12	1	0	0	88	12	0	0	0	99	1	0	0
	SIC	0	94	6	0	0	0	93	7	0	0	0	100	0	0	0

where \mathbf{X} is an $n \times p$ design matrix and $\varepsilon \sim N(\mathbf{0}, I_n)$ independent of \mathbf{X} . We set $a = 0.25, 0.5$ in (37) and kept the rest of the setting in Section 6.2 except that $\Sigma_0 = I_p$ and $n = 50, 150$, and 1000. We pretended that the data were generated from model (36). So the family of our working models certainly does not include the true model due to the nonlinearity. For each dataset, we applied the best subset regression and used AIC, BIC, GAIC, GBIC, and SIC to select the model size. Table 4 summarizes the comparison results of the frequency of estimated model size. The conclusions are similar to those in Section 6.3. We see as well that GAIC had no clear advantage over AIC in some of the settings.

6.5 Polynomial regression with heteroscedasticity

Since our new criteria add terms related to the variance of the response variable Y , we considered a cubic polynomial model with heteroscedastic variance

$$y = 1 + 5x - 1.25x^2 + 1.55x^3 + |x|^{1/2}\varepsilon,$$

where it is assumed that $\varepsilon \sim N(0, \sigma^2)$ and is independent of x . We kept the rest of the setting the same as in Section 6.1 and set $n = 50, 200, 5000$ and $\sigma = 0.5, 1$. Our model assumes that the data were generated from a polynomial model with constant variance. So the family of our working models does not include the true model due to the heteroscedasticity. For each dataset, we fit the polynomial regression model of order from 1 up to 6 and used AIC, BIC, GAIC, GBIC, and SIC to select the order of polynomial regression. Table 5 summarizes the comparison results of the frequency of estimated order of polynomial regression model. We see that GAIC performed no better than AIC, and both of them tended to select large

Table 4: Frequency of estimated model size by best subset regression over 100 simulations for different n and a

a	Criterion	Model size ($n = 50$)					Model size ($n = 150$)					Model size ($n = 1000$)				
		2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
0.25	AIC	0	58	33	7	2	0	42	43	13	2	0	58	31	11	0
	GAIC	0	52	31	13	4	0	56	33	10	1	0	56	33	11	0
	BIC	0	84	14	2	0	0	88	11	1	0	0	97	3	0	0
	GBIC	0	86	12	2	0	0	87	12	1	0	0	97	3	0	0
	SIC	0	95	5	0	0	0	92	8	0	0	0	98	2	0	0
0.5	AIC	0	57	29	11	3	0	59	36	5	0	0	88	12	0	0
	GAIC	0	54	32	10	4	0	52	35	13	0	0	100	0	0	0
	BIC	0	88	12	0	0	0	95	5	0	0	0	100	0	0	0
	GBIC	0	92	8	0	0	0	96	4	0	0	0	100	0	0	0
	SIC	0	97	3	0	0	0	99	1	0	0	0	100	0	0	0

model. SIC and GBIC performed significantly better than all other information criteria, indicating the usefulness of capturing the difference between the estimated error variance and the apparent residual variance.

7 Discussions

We have considered the problem of model selection in misspecified models and introduced a family of semi-Bayesian principles with indices $\gamma \in [0, 1]$ connecting the two well-known principles: the KL divergence principle and the Bayesian principle for model selection. The asymptotic expansions of the semi-Bayesian principles in misspecified GLMs give a family of semi-Bayesian information criteria (SIC) with indices $\gamma \in [0, 1]$. Considerations of the two well-known principles in misspecified GLMs lead to generalizations of AIC and BIC, the GAIC and GBIC. In particular, a specific form of SIC with index $\gamma = 1/2$ has a natural decomposition into the negative maximum quasi-log-likelihood and two penalties on model dimensionality and model misspecification, respectively. Numerical studies have demonstrated the advantage of the proposed SIC methodology in finite sample performance for model selection in both correctly specified and misspecified models.

When the model is misspecified, the posterior distribution of the parameter would not reflect the true uncertainty of the estimation. In particular, the QMLE $\hat{\beta}_n$ can have a different asymptotic covariance matrix than the posterior covariance matrix of β . When this difference grows larger as sample size n increases, its impact on model selection can be rather significant.

Table 5: Frequency of estimated order of polynomial regression model over 100 simulations for different n and σ

σ	Criterion	Model size ($n = 50$)					Model size ($n = 200$)					Model size ($n = 5000$)				
		2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
0.5	AIC	0	58	15	12	15	0	27	14	24	35	0	34	15	21	30
	GAIC	0	48	20	13	19	0	22	17	20	41	0	65	14	13	8
	BIC	0	76	14	4	6	0	72	17	2	9	0	87	8	5	0
	GBIC	0	91	5	2	2	0	78	16	2	4	0	87	8	5	0
	SIC	0	94	5	0	1	0	76	16	2	6	0	94	4	2	0
1	AIC	0	50	11	21	18	0	33	22	17	28	0	33	14	20	33
	GAIC	0	45	15	23	17	0	22	23	14	41	0	56	11	16	17
	BIC	0	76	11	8	5	0	75	15	3	7	0	81	13	5	1
	GBIC	0	89	11	0	0	0	83	9	3	5	0	79	13	5	3
	SIC	0	91	9	0	0	0	81	11	3	5	0	90	9	1	0

Our technical analysis has revealed that the contrast (i.e., $\mathbf{A}_n^{-1}\mathbf{B}_n$) between the covariance structure in the misspecified model and that in the true model plays a pivotal role for model selection in misspecified models. So it is important to construct accurate estimates of both covariance matrices in practice. Some suggestions have been discussed in Section 2.3. So far we have explored the expression of SIC with index $\gamma = 1/2$ taking an additive form for the three terms: goodness of fit, model complexity, and model misspecification. Possible extensions of this specific form of SIC include introducing other weights and considering non-additive forms. These problems are beyond the scope of this paper and will be interesting topics for future research.

A Proofs

A.1 Proof of Proposition 1

It suffices to show that the Hessian matrix of $-\ell_n(\mathbf{y}, \boldsymbol{\beta})$ is always positive definite. In view of (4), we have $-\partial^2 \ell_n(\mathbf{y}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^2 = \mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta}) \mathbf{X}$. Fix an arbitrary $\boldsymbol{\beta} \in \mathbf{R}^d$. By assumption the minimum diagonal element a of $\boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})$ is positive. Thus it follows easily that

$$\mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta}) \mathbf{X} \geq \mathbf{X}^T (aI_n) \mathbf{X} = a \mathbf{X}^T \mathbf{X} > 0,$$

since \mathbf{X} has full column rank d . This completes the proof.

A.2 Proof of Theorem 1

By (7) and Condition 2, it is easy to show that the Hessian matrix $\mathbf{A}_n(\boldsymbol{\beta})$ of $I(g_n; f_n(\cdot, \boldsymbol{\beta}))$ is $\mathbf{X}^T \boldsymbol{\Sigma}(\mathbf{X}\boldsymbol{\beta})\mathbf{X}$. So the proof of Proposition 1 applies to prove the uniqueness of the global minimizer of the KL divergence $I(g_n; f_n(\cdot, \boldsymbol{\beta}))$. It follows from (7) and Condition 2 that the minimizer solves the gradient equation

$$\partial I(g_n; f_n(\cdot, \boldsymbol{\beta})) / \partial \boldsymbol{\beta} = -\mathbf{X}^T [E\mathbf{Y} - \boldsymbol{\mu}(\mathbf{X}\boldsymbol{\beta})] = \mathbf{0},$$

which concludes the proof.

A.3 Proof of Theorem 2

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

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

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

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

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

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

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

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

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

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

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 $\boldsymbol{\beta}_* \in N_n(\delta)$ by the convexity of $N_n(\delta)$. Clearly

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

and by Condition 4, for n sufficiently large we have

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

Thus we have

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

which along with Markov's inequality entails that

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

Observe that

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

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

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

This together with $Q_n \subset \{ \widehat{\boldsymbol{\beta}}_n \in N_n(\delta) \}$ and Condition 3 proves $\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0} = o_P(1)$.

A.4 Proof of Theorem 3

We condition on the event Q_n defined in the proof of Theorem 2, which has been shown to entail that $\widehat{\boldsymbol{\beta}}_n \in N_n(\delta)$. By the mean value theorem applied componentwise and (8), we obtain

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

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

$$\mathbf{C}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) = \mathbf{u}_n + \mathbf{w}_n,$$

where $\mathbf{u}_n = -\mathbf{B}_n^{-1/2} \mathbf{X}^T(\mathbf{y} - E\mathbf{y})$ and

$$\mathbf{w}_n = - \left[\widetilde{\mathbf{V}}_n(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d) - \mathbf{V}_n \right] \left[\mathbf{B}_n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \right].$$

By Slutsky's lemma and the proof of Theorem 2, we see that to show

$$\mathbf{C}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_d),$$

it suffices to prove $\mathbf{u}_n \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_d)$ and $\mathbf{w}_n = o_P(1)$.

$\mathbf{w}_n = o_P(1)$ follows from $Q_n \subset \{\widehat{\boldsymbol{\beta}}_n \in N_n(\delta)\}$ and Condition 5. Finally we show that $\mathbf{u}_n \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_d)$. Fix an arbitrary unit vector $\mathbf{a} \in \mathbf{R}^d$. We consider the asymptotic distribution of the linear combination

$$v_n = \mathbf{a}^T \mathbf{u}_n = -\mathbf{a}^T \mathbf{B}_n^{-1/2} \mathbf{X}^T (\mathbf{y} - E\mathbf{y}) = \sum_{i=1}^n z_i,$$

where $z_i = -\mathbf{a}^T \mathbf{B}_n^{-1/2} \mathbf{x}_i (y_i - Ey_i)$, $i = 1, \dots, n$. Clearly z_i 's are independent and have mean 0, and

$$\sum_{i=1}^n \text{var}(z_i) = \mathbf{a}^T \mathbf{B}_n^{-1/2} \mathbf{B}_n \mathbf{B}_n^{-1/2} \mathbf{a} = 1.$$

By Condition 6, $E|y_i - Ey_i|^3 \leq M$ for some constant M . Thus we derive

$$\begin{aligned} \sum_{i=1}^n E|z_i|^3 &= \sum_{i=1}^n |\mathbf{a}^T \mathbf{B}_n^{-1/2} \mathbf{x}_i|^3 E|y_i - Ey_i|^3 \leq M \sum_{i=1}^n |\mathbf{a}^T \mathbf{B}_n^{-1/2} \mathbf{x}_i|^3 \\ &\leq M \sum_{i=1}^n \|\mathbf{a}\|^3 \|\mathbf{B}_n^{-1/2} \mathbf{x}_i\|^3 = M \sum_{i=1}^n (\mathbf{x}_i^T \mathbf{B}_n^{-1} \mathbf{x}_i)^{3/2} \rightarrow 0, \end{aligned}$$

where we used the Cauchy-Schwarz inequality and Condition 6. Applying Lyapunov's theorem yields

$$\mathbf{a}^T \mathbf{u}_n = \sum_{i=1}^n z_i \xrightarrow{\mathcal{D}} N(0, 1).$$

Since this asymptotic normality holds for any unit vector $\mathbf{a} \in \mathbf{R}^d$, we conclude that $\mathbf{u}_n \xrightarrow{\mathcal{D}} N(\mathbf{0}, I_d)$, which completes the proof.

A.5 Proof of Theorem 4

It suffices to prove

$$E\eta_n(\widehat{\boldsymbol{\beta}}_n) = \eta_n(\boldsymbol{\beta}_{n,0}) - \frac{1}{2} \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) + o(1) \quad (38)$$

and

$$\eta_n(\boldsymbol{\beta}_{n,0}) = E \left[\ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) \right] - \frac{1}{2} \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) + o(1). \quad (39)$$

We will prove (38) and (39) in separate steps. Define

$$\tilde{\ell}_n(\mathbf{y}, \boldsymbol{\beta}) = \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n + n^{1/2} \mathbf{C}_n^{-1} \boldsymbol{\beta}) \quad \text{and} \quad \tilde{\eta}_n(\boldsymbol{\beta}) = \eta_n(\boldsymbol{\beta}_{n,0} + n^{1/2} \mathbf{C}_n^{-1} \boldsymbol{\beta}),$$

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

1) We first prove (38). The key idea is to do a second order Taylor expansion of $\tilde{\eta}_n(\boldsymbol{\beta})$ around $\mathbf{0}$ and retain the Lagrange remainder term. By Theorem 1, η_n attains its maximum

at $\boldsymbol{\beta}_{n,0}$. Thus we derive

$$\begin{aligned}\eta_n(\widehat{\boldsymbol{\beta}}_n) &= \tilde{\eta}_n \left[n^{-1/2} \mathbf{C}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0}) \right] = \tilde{\eta}_n(\mathbf{0}) - \frac{n}{2} \mathbf{v}^T (\mathbf{C}_n^{-1} \mathbf{A}_n \mathbf{C}_n^{-1}) \mathbf{v} \\ &\quad - \frac{1}{6} \sum_{j,k,l} [\partial^3 \tilde{\eta}_n(\boldsymbol{\beta}_*) / \partial \beta_j \partial \beta_k \partial \beta_l] v_j v_k v_l \\ &= \eta_n(\boldsymbol{\beta}_{n,0}) - \frac{n}{2} \mathbf{v}^T (\mathbf{B}_n^{1/2} \mathbf{A}_n^{-1} \mathbf{B}_n^{1/2}) \mathbf{v} - \frac{1}{6} \sum_{j,k,l} [\partial^3 \tilde{\eta}_n(\boldsymbol{\beta}_*) / \partial \beta_j \partial \beta_k \partial \beta_l] v_j v_k v_l,\end{aligned}$$

where $\mathbf{v} = (v_1, \dots, v_d)^T = n^{-1/2} \mathbf{C}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n,0})$ and $\boldsymbol{\beta}_*$ lies on the line segment joining \mathbf{v} and $\mathbf{0}$. By Condition 7,

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

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

$$E(nv_j v_k) \rightarrow \delta_{jk} \quad \text{and} \quad E(n^{3/2}|v_j v_k v_l|) = O(1),$$

where δ_{jk} denotes the Kronecker delta. These along with $\text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) = O(1)$ by Condition 7 yield

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

2) We then prove (39). The key idea is to represent $\ell_n(\mathbf{y}, \boldsymbol{\beta}_{n,0})$ using a second order Taylor expansion of $\ell_n(\mathbf{y}, \cdot)$ around $\widehat{\boldsymbol{\beta}}$ and retaining the Lagrange remainder term. It is easy to see that the difference between $\eta_n(\boldsymbol{\beta})$ and $\ell_n(\mathbf{y}, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, which implies that each pair of their second or higher order partial derivatives agree. Since $\ell_n(\mathbf{y}, \cdot)$ attains its maximum at $\widehat{\boldsymbol{\beta}}_n$, we derive

$$\begin{aligned}\ell_n(\mathbf{y}, \boldsymbol{\beta}_{n,0}) &= \tilde{\ell}_n(\mathbf{y}, \mathbf{v}) = \tilde{\ell}_n(\mathbf{0}) - \frac{1}{2} \mathbf{v}^T \left[\partial^2 \tilde{\ell}_n(\mathbf{y}, \mathbf{0}) / \partial \boldsymbol{\beta}^2 \right] \mathbf{v} \\ &\quad - \frac{1}{6} \sum_{j,k,l} \left[\partial^3 \tilde{\ell}_n(\mathbf{y}, \boldsymbol{\beta}_*) / \partial \beta_j \partial \beta_k \partial \beta_l \right] v_j v_k v_l \\ &= \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) - \frac{1}{2} \mathbf{v}^T \left[\partial^2 \tilde{\eta}_n(-\mathbf{v}) / \partial \boldsymbol{\beta}^2 \right] \mathbf{v} - \frac{1}{6} \sum_{j,k,l} \left[\partial^3 \tilde{\eta}_n(\boldsymbol{\beta}_{**}) / \partial \beta_j \partial \beta_k \partial \beta_l \right] v_j v_k v_l,\end{aligned}$$

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

It follows from $\sup_{\boldsymbol{\beta}} \max_{j,k,l} |\partial^3 \tilde{\eta}_n(\boldsymbol{\beta}) / \partial \beta_j \partial \beta_k \partial \beta_l| = o(n^{3/2})$ in Condition 7 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

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

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

$$E(nv_j v_k) \rightarrow \delta_{jk} \quad \text{and} \quad E(n^{3/2}\|\mathbf{v}\|^3) = O(1)$$

These along with $\text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) = O(1)$ by Condition 7 yield

$$\begin{aligned} \eta_n(\boldsymbol{\beta}_{n,0}) &= E\ell_n(\mathbf{y}, \boldsymbol{\beta}_{n,0}) = E\ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \left[\frac{1}{2} \text{tr}(\mathbf{B}_n^{1/2} \mathbf{A}_n^{-1} \mathbf{B}_n^{1/2}) + o(1) \right] \\ &\quad + o(n^{3/2})O(n^{-3/2}) \\ &= E\ell_n(\mathbf{y}, \hat{\boldsymbol{\beta}}_n) - \frac{1}{2} \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) + o(1). \end{aligned}$$

This proves the conclusion.

A.6 Proof of Proposition 2

Part a) follows from (22) and the definition of the log-marginal likelihood $S(\mathbf{y}, \mathfrak{M}_m; F_n)$ in (19). Part b) is an easy consequence of part a).

A.7 Proof of Theorem 5

It is easy to see that $\ell_n^*(\mathbf{y}, \boldsymbol{\delta})$ defined in (27) is a smooth concave function on \mathbf{R}^d with its maximum 0 attained at $\boldsymbol{\delta} = \mathbf{0}$,

$$\partial^2 \ell_n^*(\mathbf{y}, \boldsymbol{\delta})/\partial \boldsymbol{\delta}^2 = -n\mathbf{B}_n^{-1/2} \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n + n^{1/2}\mathbf{B}_n^{-1/2}\boldsymbol{\delta})\mathbf{B}_n^{-1/2},$$

and $U_n(\boldsymbol{\delta}) = \exp[n^{-1}\ell_n^*(\mathbf{y}, \boldsymbol{\delta})]$ takes values between 0 and 1. Thus by Taylor's theorem, expanding $\ell_n^*(\mathbf{y}, \cdot)$ around $\mathbf{0}$ gives for any $\boldsymbol{\delta}$,

$$\ell_n^*(\mathbf{y}, \boldsymbol{\delta}) = -\frac{1}{2}\boldsymbol{\delta}^T [\partial^2 \ell_n^*(\mathbf{y}, \boldsymbol{\delta}_*)/\partial \boldsymbol{\delta}^2] \boldsymbol{\delta}, \quad (40)$$

where $\boldsymbol{\delta}_*$ lies on the line segment joining $\boldsymbol{\delta}$ and $\mathbf{0}$. Let $\mathbf{D}_n = \mathbf{B}_n^{-1/2} \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n)\mathbf{B}_n^{-1/2}$ and for $r \in (0, \infty)$,

$$\rho_n(r) = \max_{\|\boldsymbol{\delta}\| \leq r} \max \{ |\lambda_{\min}[\mathbf{F}_n(\boldsymbol{\delta})]|, |\lambda_{\max}[\mathbf{F}_n(\boldsymbol{\delta})]| \},$$

where $\mathbf{F}_n(\boldsymbol{\delta}) = \mathbf{B}_n^{-1/2} [\mathbf{A}_n(\hat{\boldsymbol{\beta}}_n + \mathbf{B}_n^{-1/2}\boldsymbol{\delta}) - \mathbf{A}_n(\hat{\boldsymbol{\beta}}_n)] \mathbf{B}_n^{-1/2}$.

Fix a sequence of $r_n \in (0, \infty)$ satisfying $\rho_n(r_n) < \lambda_{\min}(\mathbf{D}_n)$ and define a sequence of closed balls

$$B_{r_n} = \{\boldsymbol{\delta} \in \mathbf{R}^d : \|\boldsymbol{\delta}\| \leq r_n\}.$$

It follows from (40) that

$$q_1(\boldsymbol{\delta})1_{B_{r_n}}(\boldsymbol{\delta}) \leq -n^{-1}\ell_n^*(\mathbf{y}, \boldsymbol{\delta})1_{B_{r_n}}(\boldsymbol{\delta}) \leq q_2(\boldsymbol{\delta})1_{B_{r_n}}(\boldsymbol{\delta}), \quad (41)$$

where

$$q_1(\boldsymbol{\delta}) = \frac{1}{2}\boldsymbol{\delta}^T [\mathbf{D}_n - \rho_n(r_n)I_d] \boldsymbol{\delta} \quad \text{and} \quad q_2(\boldsymbol{\delta}) = \frac{1}{2}\boldsymbol{\delta}^T [\mathbf{D}_n + \rho_n(r_n)I_d] \boldsymbol{\delta}.$$

This entails

$$E_{\mu_{\mathfrak{M}}} (e^{-nq_2} 1_{B_{r_n}}) \leq E_{\mu_{\mathfrak{M}}} [U_n(\boldsymbol{\delta})^n 1_{B_{r_n}}] \leq E_{\mu_{\mathfrak{M}}} (e^{-nq_1} 1_{B_{r_n}}). \quad (42)$$

We will see later that inequality (42) is the key step in deriving the asymptotic expansion of $\log E_{\mu_{\mathfrak{M}}}[U_n(\boldsymbol{\delta})^n]$ in (28).

We list some auxiliary results in the following two lemmas, whose proofs will be given separately.

Lemma 1. *Under Condition 8 and provided that $r_n \leq r_0$, we have*

$$c_1 \int e^{-nq_j} 1_{B_{r_n}} d\mu_0 \leq E_{\mu_{\mathfrak{M}}} (e^{-nq_j} 1_{B_{r_n}}) \leq c_2 \int e^{-nq_j} 1_{B_{r_n}} d\mu_0, \quad j = 1, 2. \quad (43)$$

Lemma 2. *It holds that*

$$E_{\mu_{\mathfrak{M}}} [U_n(\boldsymbol{\delta})^n 1_{B_{r_n}^c}] \leq \exp(-n\kappa_n r_n^2), \quad (44)$$

$$\int e^{-nq_1} d\mu_0 = \left(\frac{2\pi}{n}\right)^{d/2} |\mathbf{D}_n - \rho_n(r_n)I_d|^{-1/2}, \quad (45)$$

$$\int e^{-nq_2} d\mu_0 = \left(\frac{2\pi}{n}\right)^{d/2} |\mathbf{D}_n + \rho_n(r_n)I_d|^{-1/2}, \quad (46)$$

$$\int e^{-nq_j} 1_{B_{r_n}^c} d\mu_0 \leq \left(\frac{2\pi}{n\kappa_n}\right)^{d/2} \exp\left[-\frac{1}{2}n\kappa_n r_n^2 + \frac{d}{2} + \frac{d}{2} \log(n\kappa_n r_n^2 d^{-1})\right], \quad (47)$$

where $\kappa_n = \lambda_{\min}(\mathbf{D}_n) - \rho_n(r_n)$, $j = 1, 2$, and it is assumed that $n\kappa_n r_n^2 > d$ in (47).

Now we are ready to obtain the asymptotic expansion of $\log E_{\mu_{\mathfrak{M}}}[U_n(\boldsymbol{\delta})^n]$ in (28). We pick a sequence of positive numbers $r_n \rightarrow 0$ and work under Conditions 8 and 9. It follows from Condition 9 that

$$(1 - c_3)\lambda_{\min}(\mathbf{D}_n) \leq \kappa_n = \lambda_{\min}(\mathbf{D}_n) - \rho_n(r_n) \leq \lambda_{\min}(\mathbf{D}_n).$$

Thus in view of $r_n = o(1)$ and $\lambda_{\min}(\mathbf{D}_n)^{-1}r_n^{-2} = O(n^\alpha)$, we have

$$k_n^{-1} = O(n^\alpha r_n^2) = o(n^\alpha) \quad \text{and} \quad n^{-(1-\alpha)}(n\kappa_n r_n^2) \rightarrow \infty,$$

where $1 - \alpha \in (0, 1]$. Note that $\lambda_{\max}(\mathbf{D}_n) = O(1)$. Since an exponential rate of convergence to zero is faster than a polynomial rate, by Lemmas 1 and 2 we can claim that for $j = 1, 2$,

$$\log E_{\mu_{\mathfrak{M}}} (e^{-nq_j} 1_{B_{r_n}}) = \log \left[\left(\frac{2\pi}{n} \right)^{d/2} |\mathbf{D}_n|^{-1/2} \right] + O(1),$$

which along with (42) and (44) yields

$$\begin{aligned} \log E_{\mu_{\mathfrak{M}}}[U_n(\boldsymbol{\delta})^n] &= \log \left[\left(\frac{2\pi}{n} \right)^{d/2} |\mathbf{D}_n|^{-1/2} \right] + O(1) \\ &= -\frac{\log n}{2}d + \frac{1}{2} \log |\mathbf{A}_n(\widehat{\boldsymbol{\beta}}_n)^{-1} \mathbf{B}_n| + O(1). \end{aligned}$$

This together with (28) proves the conclusion for fixed \mathbf{y} .

A.8 Proof of Lemma 1

Since $r_n \leq r_0$, by (26) and Condition 8 we have for any $\boldsymbol{\delta} \in B_{r_n}$,

$$c_1 \leq \frac{d\mu_{\mathfrak{M}}}{d\mu_0}(\boldsymbol{\delta}) \leq c_2,$$

which immediately gives inequality (43).

A.9 Proof of Lemma 2

In light of the choice of r_n and the definition of $\rho_n(r_n)$, it is easy to show that

$$U_n(\boldsymbol{\delta})^n 1_{B_{r_n}^c} \leq \exp \{ -n [\lambda_{\min}(\mathbf{D}_n) - \rho_n(r_n)] r_n^2 \},$$

which yields inequality (44) since $\mu_{\mathfrak{M}}$ is a probability distribution. Equations (45) and (46) can be obtained by using the multivariate Gaussian integral. It remains to prove (47).

By the definition of functions q_1 and q_2 , we observe that

$$q_j(\boldsymbol{\delta}) \geq \frac{1}{2} [\lambda_{\min}(\mathbf{D}_n) - \rho_n(r_n)] \|\boldsymbol{\delta}\|^2, \quad \boldsymbol{\delta} \in B_{r_n}^c, j = 1, 2.$$

Let $\kappa_n = \lambda_{\min}(\mathbf{D}_n) - \rho_n(r_n)$ and $\mathbf{F} = n\kappa_n I_d$. Then for $j = 1, 2$

$$\begin{aligned} \int e^{-nq_j} 1_{B_{r_n}^c} d\mu_0 &\leq \int e^{-\frac{1}{2} \boldsymbol{\delta}^T \mathbf{F} \boldsymbol{\delta}} 1_{B_{r_n}^c} d\mu_0 \\ &= \left(\frac{2\pi}{n\kappa_n} \right)^{d/2} P \left[\|(n\kappa_n)^{-1/2} \mathbf{Z}\|^2 > r_n^2 \right], \end{aligned} \quad (48)$$

where $\mathbf{Z} \sim N(\mathbf{0}, I_d)$. Assume $n\kappa_n r_n^2 > d$. Pick $\varepsilon_n \in (0, \infty)$ in a way such that $d(1 + \varepsilon_n) = n\kappa_n r_n^2$. Then by the deviation bound in Lemma 3(a) of Fan and Lv (2008), we have

$$P \left[\|(n\kappa_n)^{-1/2} \mathbf{Z}\|^2 > r_n^2 \right] = P \left(d^{-1} \|\mathbf{Z}\|^2 > 1 + \varepsilon_n \right) \leq e^{-A\varepsilon_n d},$$

where $A_{\varepsilon_n} = [\varepsilon_n - \log(1 + \varepsilon_n)]/2$. Thus

$$P \left[\|(n\kappa_n)^{-1/2} \mathbf{Z}\|^2 > r_n^2 \right] \leq \exp \left[-\frac{1}{2}n\kappa_n r_n^2 + \frac{d}{2} + \frac{d}{2} \log (n\kappa_n r_n^2 d^{-1}) \right].$$

This along with (48) concludes the proof.

A.10 Proof of Theorem 6

In view of (31), we have for $\gamma \in [0, 1/2]$,

$$\begin{aligned} D_\gamma(\widehat{\boldsymbol{\beta}}_{n,m}; \mathfrak{M}_m) &= (1 - 2\gamma)I(g_n; f_n(\cdot, \widehat{\boldsymbol{\beta}}_{n,m})) + \gamma \left[I(g_n; f_n(\cdot, \widehat{\boldsymbol{\beta}}_{n,m})) - I(g_n; f_n(\cdot, \boldsymbol{\beta}_{n,m,0})) \right] \\ &\quad + \gamma I \left(g_n; \frac{d\nu_m}{d\mu_0} \right), \end{aligned}$$

and for $\gamma \in [1/2, 1]$,

$$D_\gamma(\widehat{\boldsymbol{\beta}}_{n,m}; \mathfrak{M}_m) = (1 - \gamma) \left[I(g_n; f_n(\cdot, \widehat{\boldsymbol{\beta}}_{n,m})) - I(g_n; f_n(\cdot, \boldsymbol{\beta}_{n,m,0})) \right] + \gamma I \left(g_n; \frac{d\nu_m}{d\mu_0} \right).$$

We drop the subscript m for simplicity. It follows from (14) and (16) in Theorem 4 that

$$I(g_n; f_n(\cdot, \widehat{\boldsymbol{\beta}}_n)) = E \log g_n(\mathbf{y}) - E \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) + \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) + o(1). \quad (49)$$

By (14), we have

$$I(g_n; f_n(\cdot, \widehat{\boldsymbol{\beta}}_n)) - I(g_n; f_n(\cdot, \boldsymbol{\beta}_{n,0})) = \eta_n(\boldsymbol{\beta}_{n,0}) - \eta_n(\widehat{\boldsymbol{\beta}}_n).$$

This along with (38) in the proof of Theorem 4 entails

$$E \left[I(g_n; f_n(\cdot, \widehat{\boldsymbol{\beta}}_n)) - I(g_n; f_n(\cdot, \boldsymbol{\beta}_{n,0})) \right] = \frac{1}{2} \text{tr}(\mathbf{A}_n^{-1} \mathbf{B}_n) + o(1). \quad (50)$$

It follows from (23) and (29) in Theorem 5 that

$$\begin{aligned} EI \left(g_n; \frac{d\nu_m}{d\mu_0} \right) &= -E \ell_n(\mathbf{y}, \widehat{\boldsymbol{\beta}}_n) + \frac{\log n}{2} |\mathfrak{M}| - \frac{1}{2} E \log |\widehat{\mathbf{A}}_n^{-1} \mathbf{B}_n| \\ &\quad + E \log g_n(\mathbf{y}) - ER(\mathbf{y}, \mathfrak{M}; F_n), \end{aligned} \quad (51)$$

where $\widehat{\mathbf{A}}_n = \mathbf{A}_n(\widehat{\boldsymbol{\beta}}_n)$. Therefore combining (49)–(51) together with the above representations of $D_\gamma(\widehat{\boldsymbol{\beta}}_n; \mathfrak{M})$ yields (33), which completes the proof.

A.11 Proof of Proposition 3

The decomposition in (35) follows from the definition of $\text{SIC}_\gamma(\mathbf{y}, \mathfrak{M}; F_n)$ with $\gamma = 1/2$ in (34) and a fact that the KL divergence of a d -variate Gaussian distribution $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ from a d -variate Gaussian distribution $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ is given by

$$I[N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1); N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)] = \frac{1}{2} \left[\text{tr}(\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1) - \log |\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1| + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - d \right].$$

B Three commonly used GLMs

Linear regression model, logistic regression model, and Poisson regression model are three commonly used GLMs. In this section, we give the formulas for $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ given in Section 2.3, when F_n is chosen to be one of them.

B.1 Linear regression

This model assumes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (52)$$

where \mathbf{X} is an $n \times d$ design matrix and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$. Let $\boldsymbol{\beta} = \boldsymbol{\gamma}/\sigma^2$ and $b(\theta) = \sigma^2 \theta^2/2$, $\theta \in \mathbf{R}$. Then the quasi-log-likelihood of the sample defined in (4) becomes

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

Maximizing $\ell_n(\mathbf{y}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\gamma}$ yields the least squares estimator $\widehat{\boldsymbol{\gamma}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. We estimate σ^2 by using the residual sum of squares (RSS), $\widehat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}\|^2}{n-d}$. Then we obtain an estimate $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\gamma}}/\widehat{\sigma}^2$. Since $b'(\theta) = \sigma^2 \theta$ and $b''(\theta) = \sigma^2$, we have

$$\widehat{\mathbf{A}}_n = \widehat{\sigma}^2 \mathbf{X}^T \mathbf{X} \quad \text{and} \quad \widehat{\mathbf{B}}_n = \mathbf{X}^T \text{diag}\{[\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}] \circ [\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}]\} \mathbf{X}, \quad (53)$$

where \circ denotes the Hadamard (componentwise) product. An interesting specific case is when the true model is indeed linear, but may involve a different set of covariates. Then $\mathbf{B}_n = \tau^2 \mathbf{X}^T \mathbf{X}$, where τ^2 is the true variance of Y had we known the model precisely. The SIC with index $\gamma = 1/2$ just involves an additional term penalizing the inflation of the error variance due to model misspecification.

B.2 Logistic regression

In this model, $b(\theta) = \log(1 + e^\theta)$, $\theta \in \mathbf{R}$ and the quasi-log-likelihood of the sample is $\ell_n(\mathbf{y}, \boldsymbol{\beta}) = \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta})$, where we ignored the last constant term in (4). By definition $\ell_n(\mathbf{y}, \cdot)$ attains its maximum at $\widehat{\boldsymbol{\beta}}_n$. Since $b'(\theta) = \frac{e^\theta}{1+e^\theta}$ and $b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2}$, matrices $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ are given in (11) and (12) with $\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n) = (\frac{e^{\theta_1}}{1+e^{\theta_1}}, \dots, \frac{e^{\theta_n}}{1+e^{\theta_n}})^T$, $\boldsymbol{\Sigma}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n) = \text{diag}\{\frac{e^{\theta_1}}{(1+e^{\theta_1})^2}, \dots, \frac{e^{\theta_n}}{(1+e^{\theta_n})^2}\}$, and $(\theta_1, \dots, \theta_n)^T = \mathbf{X}\widehat{\boldsymbol{\beta}}_n$.

B.3 Poisson regression

In this model, $b(\theta) = e^\theta$, $\theta \in \mathbf{R}$ and the quasi-log-likelihood of the sample is $\ell_n(\mathbf{y}, \boldsymbol{\beta}) = \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \mathbf{b}(\mathbf{X}\boldsymbol{\beta})$, where we ignored the last constant term in (4). By definition $\ell_n(\mathbf{y}, \cdot)$

attains its maximum at $\widehat{\boldsymbol{\beta}}_n$. Since $b'(\theta) = e^\theta$ and $b''(\theta) = e^\theta$, matrices $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ are given in (11) and (12) with $\boldsymbol{\mu}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n) = (e^{\theta_1}, \dots, e^{\theta_n})^T$ and $\boldsymbol{\Sigma}(\mathbf{X}\widehat{\boldsymbol{\beta}}_n) = \text{diag}\{e^{\theta_1}, \dots, e^{\theta_n}\}$, and $(\theta_1, \dots, \theta_n)^T = \mathbf{X}\widehat{\boldsymbol{\beta}}_n$.

References

- [1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (eds. B. N. Petrov and F. Csaki), Akademiai Kiado, Budapest, 267–281.
- [2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Auto. Control* **19**, 716–723.
- [3] Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics* **37**, 51–58.
- [4] Berk, R. H. (1970). Consistency a Posteriori. *Annals of Mathematical Statistics* **41**, 894–906.
- [5] Bozdogan, H. (1987). Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345–370.
- [6] Bozdogan, H. (2000). Akaike’s information criterion and recent developments in information complexity. *J. Math. Psychol.* **44**, 62–91.
- [7] Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer, New York.
- [8] Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313–2404.
- [9] Cavanaugh, J. E. and Neath, A. A. (1999). Generalizing the derivation of the Schwarz information criterion. *Commun. Statist. - Theory and Methods* **28**, 49–66.
- [10] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. New York: Springer-Verlag.
- [11] Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13**, 342–368.
- [12] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

- [13] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849–911.
- [14] Fan, J. and Lv, J. (2009). Non-concave penalized likelihood with NP-dimensionality. *Manuscript*.
- [15] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica* **20**, 101–148.
- [16] Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–1975.
- [17] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd ed.). Chapman & Hall/CRC.
- [18] Hall, P. (1990). Akaike’s information criterion and Kullback-Leibler loss for histogram density estimation. *Probab. Theory Relat. Fields* **85**, 449–467.
- [19] Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.*, to appear.
- [20] Hall, P., Titterington, D. M. and Xue, J.-H. (2009). Tilting methods for assessing the influence of components in a classifier. *J. Roy. Statist. Soc. Ser. B*, to appear.
- [21] Hosking, J. R. M. (1980). Lagrange-multiplier tests of time-series models. *J. Roy. Statist. Soc. Ser. B* **42**, 170–181.
- [22] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, University of California Press, Berkeley.
- [23] Konishi, S. and G. Kitagawa, G. (1996). Generalised information criterion in model selection. *Biometrika* **83**, 875–890.
- [24] Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.
- [25] LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes’ estimates. *University of California Publications in Statistics* **1**, 277–330.
- [26] Liu, W. and Yang, Y. (2009). Parametric or nonparametric? A parametricness index for model selection. *Manuscript*.
- [27] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.

- [28] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.
- [29] Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.
- [30] Shibata, R. (1989). Statistical aspects of model selection. In *From Data to Model* (Ed. J. C. Willems), 215–240. New York: Springer-Verlag.
- [31] 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. Roy. Statist. Soc. Ser. B* **64**, 583–639.
- [32] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike’s criterion. *J. Roy. Statist. Soc. Ser. B* **39**, 44–47.
- [33] Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences* **153**, 12–18 (in Japanese).
- [34] 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.
- [35] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.
- [36] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **60**, 595–603.
- [37] Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.
- [38] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
- [39] Yang, Y. and Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory* **44**, 95–116.
- [40] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.